telmar

Telmar Cluster Analysis

User manual

Version 9.6

February 2019

Contents

Introduction

In Brian S. Everitt's text book "Cluster Analysis" in the introduction, there is an excellent description of clustering: -

"One of the most basic abilities of living creatures involves the grouping of similar objects to produce a classification. Early man for example, must have been able to realize that many individual objects shared certain properties such as being edible or poisonous or ferocious and so on.

Animals are named as cats, dogs, horses etc., and such a name collects individuals into groups. Naming is classifying.

The book quotes Linnaeus 1737 in his work Genera Plantarum

"All the real knowledge which we possess, depends on methods by which we distinguish the similar from the dissimilar. The greater number of natural distinctions this method comprehends the clearer becomes our idea of things."

The classification of animals and plants has clearly played an important role in the fields of biology and zoology particularly as a basis for Darwin's theory of evolution."

When a company wants to know which individuals to target, they will turn to cluster analysis, to create groups of consumers based on similarities and dissimilarities. Once the company determines which groups they wish to target, they can develop marketing strategies and relevant messages according to the needs and interests of the individual groups.

Cluster analysis is a statistical tool which allows us to separate people into several mutually exclusive groups (clusters). The aim is that the objects (people) in each cluster are ***similar*** to each other, but are ***different*** from objects that are in other clusters. This manual will show you how Telmar Clustering works and how it could be applied to solve typical marketing related problems. (See the section 11 appendix for more technical detail of cluster analysis).

Telmar's Cluster analysis was built to make the grouping of objects (e.g. consumers) as user friendly as possible. It does not require special statistical literacy, with the exception of the basic concepts like average and variance. The program guides the user through the process without too much technical detail.

This manual is divided in two parts. The first 9 sections describe how to use the program, with an example. The remaining sections (10+) describe the algorithms, parameters and defaults which are used, and why.

Part 1 (user guide)

1. Types of questions that can be used in clustering

Historically, the typical use of cluster analysis has been to create target segmentations based on attitudinal questions. Likert scale answers are very suitable for clustering. However, Telmar's new cluster tool can support other types of questions as well.

The types of input that cluster can now handle are described below

a) Likert scale

Likert scale questions are questions about the level of agreement to a statement (generally they have a range of agreement for example from 1 to 5). This type of question as described above will be weighted by Telmar by applying values from low to high (1 to 5).

An example of a Likert scale question is

- Definitely agree
- Tend to agree
- Neither
- Tend to disagree
- Definitely disagree

b) Volumetrics

These are numerical values traditionally known in marketing as volumetrics

Examples of volumetrics:

- How much time was spent with different activities
- How much money spent on different items
- Number of visits
- Amount of product consumed

c) Binary questions

Binary questions have only two mutually exclusive answer options e.g:-

- agree or disagree
- yes or no.

Examples are:

- Do you drink beer (yes/no)
- People who answered either definitely or tend to agree to an attitudinal question
- Read Time magazines (yes)

2. Example segmentation using an advertising campaign

The strategy

To illustrate the use of cluster, we will work through an example campaign using the Telmar screens.

The client is a furniture chain called La-z-boy, who provide comparatively inexpensive furniture. We are planning a marketing campaign to attract new customers to their stores. Our campaign will emphasize bargains and value for money. We will use the USA MRI survey for this purpose.

Our strategy or aim is to find segments or groups of furniture buyers (our target) who are especially receptive to advertising and who are interested in bargains. When we find these segments, we can add them back to the Telmar data set for analysis within SurveyTime where we can analyze each cluster group by looking at their consumer behavior, interests and media consumption and develop a strategy for targeting "similar" people.

The inputs for the cluster

The thought process and input selection for running a cluster analysis starts in Telmar's tabulation program - SurveyTime.

The choice of input variables are important for obtaining good clusters. Most clients use Telmar's *Correspondence Analysis* program to quickly create a shortlist of good cluster inputs. There is a How do I Guide that explains how to choose the best questions (inputs). This document is available on eTelmar, or it can be emailed to you via your Telmar help desk representative.

The cluster inputs for our example campaign are:

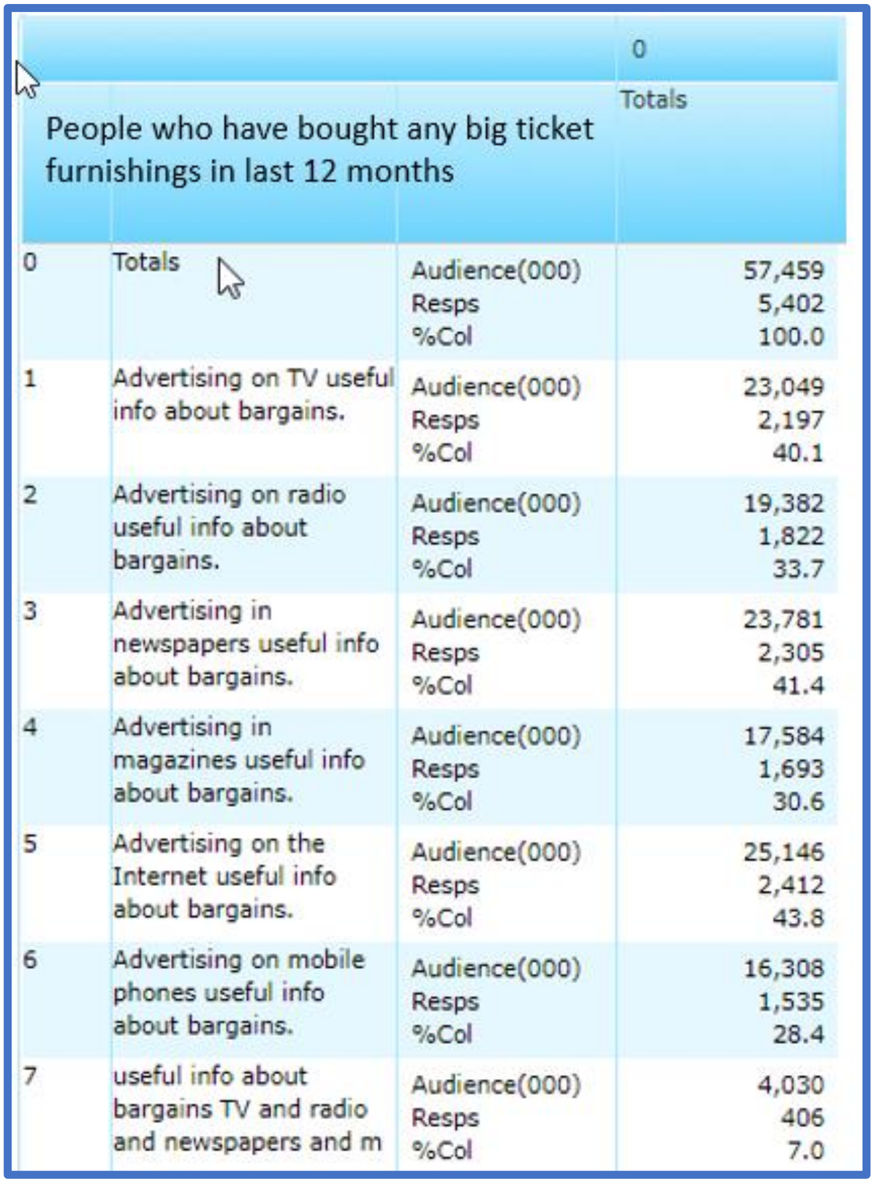
- the target audience to be clustered (i.e. the audience that you wish to break out into groups – this is entered as the base). In this example we are using "furniture buyers"
- the variables (or questions) to be used for the clustering, these are entered as rows.

Note: In SurveyTime, if you input any columns a warning message will appear.

In the report shown to the right, rows 1 to 6 are questions about advertising and bargains, and will be used as inputs for our clustering example.

You can see that in row 7 that the people who find all six media channels useful for finding bargains is only 7% of all shoppers.

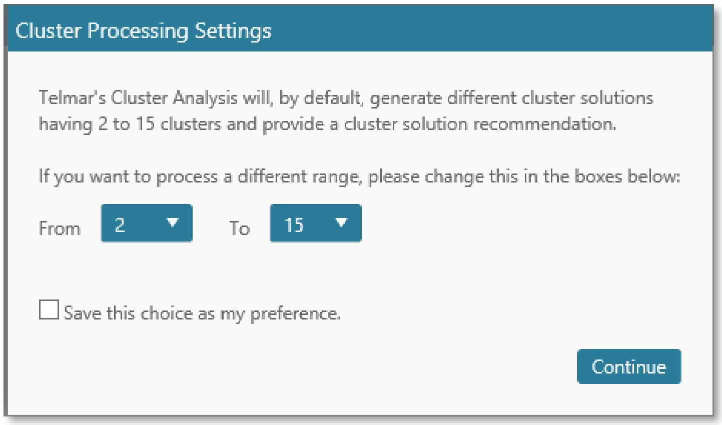
Clustering respondents using 6 media channels (the first 6 rows) will allow us to identify, group and target shoppers, who agree with different combinations of the six questions.



People who have bought any big ticket furnishings in last 12 months			0 Totals
0	Totals	Audience(000) Resps %Col	57,459 5,402 100.0
1	Advertising on TV useful info about bargains.	Audience(000) Resps %Col	23,049 2,197 40.1
2	Advertising on radio useful info about bargains.	Audience(000) Resps %Col	19,382 1,822 33.7
3	Advertising in newspapers useful info about bargains.	Audience(000) Resps %Col	23,781 2,305 41.4
4	Advertising in magazines useful info about bargains.	Audience(000) Resps %Col	17,584 1,693 30.6
5	Advertising on the Internet useful info about bargains.	Audience(000) Resps %Col	25,146 2,412 43.8
6	Advertising on mobile phones useful info about bargains.	Audience(000) Resps %Col	16,308 1,535 28.4
7	useful info about bargains TV and radio and newspapers and m	Audience(000) Resps %Col	4,030 406 7.0

Number of clusters

Once the user has input their chosen target and variables in SurveyTime, the user clicks on the Cluster+ icon and progresses to the next section. The user is asked how many groups they would like to see as shown in the screen below.



Cluster Processing Settings

Telmar's Cluster Analysis will, by default, generate different cluster solutions having 2 to 15 clusters and provide a cluster solution recommendation.

If you want to process a different range, please change this in the boxes below:

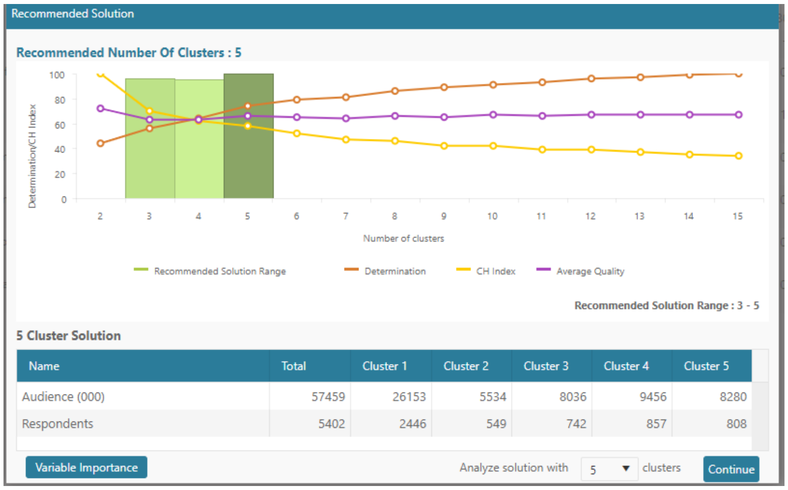
From 2 To 15

□ Save this choice as my preference.

The program default is to process a range of between 2 and 15 cluster solutions. For most purposes 8 or 10 is a suitable maximum. The user can change their default and “Save their preferred minimum and maximum (by clicking the save this choice as my preference box). This can also be changed in the preferences at any time.

Recommended solution

A screen, as shown below, will appear when you select “continue”. It shows the recommended number of clusters and two criteria of quality of clustering. These quality criteria are explained in the next section (with further details in the technical appendix).



Recommended Solution

Recommended Number Of Clusters : 5

Recommended Solution Range : 3 - 5

5 Cluster Solution

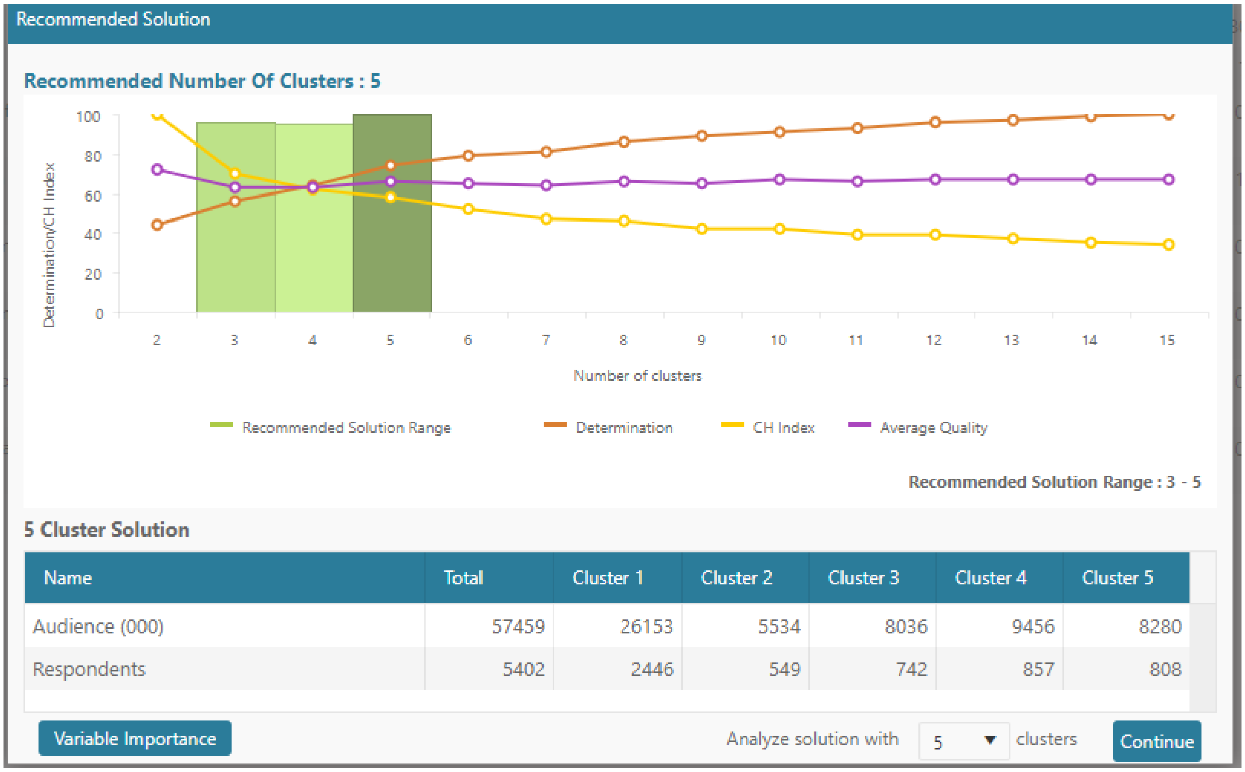
Name	Total	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Audience (000)	57459	26153	5534	8036	9456	8280
Respondents	5402	2446	549	742	857	808

Analyze solution with 5 clusters

The user can also review the “***Variables Importance***” report shown in the bottom left hand corner of the screen above). If specific variable(s) or row(s) are important, the client may wish to check where (on which cluster solution) their row(s) have the highest determination score (across all the cluster solutions). This option is also accessible from the results screen (under variables management). It is described in more detail later on in this guide.

The quality criteria

The quality criteria metrics are called “Determination” and “CH index”. Technical details about these quality metrics are explained in more detail in Appendix 2 of this guide. The higher the value of either criteria, the better the solution. They are both shown on the same scale, from 1 to 100%. This chart helps the user decide the best number of clusters. The recommended number is shown by the program (5 clusters in this case). This can be complimented with a ***recommended range***, where the system highlights not just one but several values.



5 Cluster Solution

Name	Total	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Audience (000)	57459	26153	5534	8036	9456	8280
Respondents	5402	2446	549	742	857	808

The criteria of quality are.

a) Determination

.

The D measure is an important part of understanding the cluster solution.

With any variable one can calculate the variation – a standard statistic which explains whether respondents tend to have similar replies (low variation) or very different replies (high variation). This measure can be calculated not only for the whole group but also for each individual variable. The cluster determination is an average of the determinations for all variables.

So lets say we have 4 clusters. The variance of the whole group is 100 and the variance of the 4 clusters are 10, 15, 15 and 20 – a total of 60

So the proportion of the variance ‘within cluster’ is 60/100 and the proportion ‘between clusters’ is the remainder is (100-60)/100 = 0.4

So in this case D=0.4

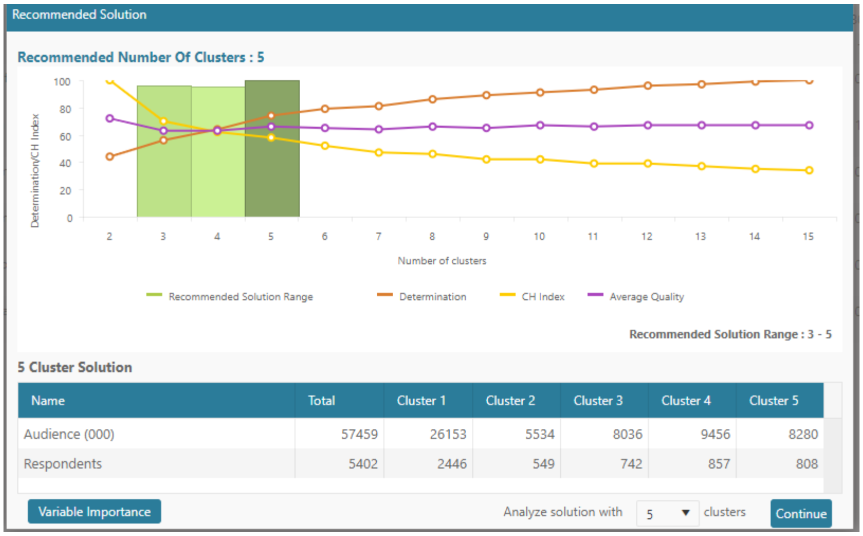
As to what is a high score that depends on the individual data set but getting anything over 0.5, or 50% would normally be good.

. The higher the D, the better the clusters separate the different respondents.
The disadvantage of D is, that it doesn't capture situations where clusters exist, yet each variable by itself cannot make a good separation.

Thus, D will be low, while cluster structure is very good. For that reason, a high determination value is always good a low value isn't necessarily bad (as a good data structure may still be there). The next criteria (CH) tries to overcome this problem.

b) Calinski-Harabasz (CH)

CH index exploits the same idea of the differences between variances "within and between" as D does, but does it in such a way that it works better in multidimensional situations . The CH score should be high and viewed in tandem with the Determination Score, rather than alone for a particular variable



Recommended Number Of Clusters : 5

Recommended Solution Range : 3 - 5

5 Cluster Solution

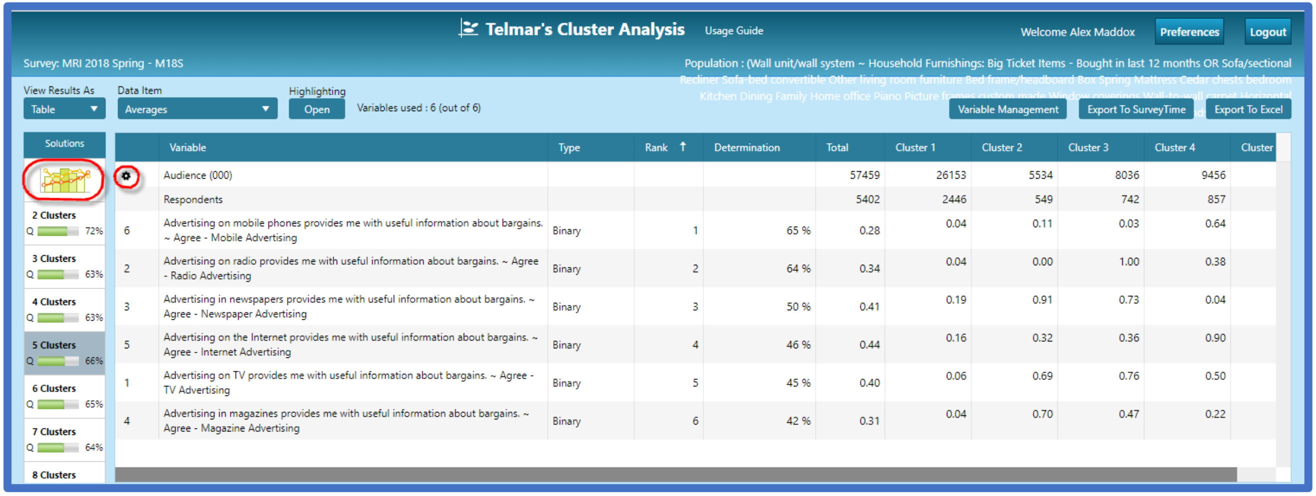
Name	Total	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Audience (000)	57459	26153	5534	8036	9456	8280
Respondents	5402	2446	549	742	857	808

Analyze solution with 5 ▾ clusters

When the user has reviewed the recommended solutions as shown above, the user can select which solution to analyze in more detail and click ***continue,*** this leads to the main cluster results table.

3. Main cluster results table

Several things can be easily manipulated within the table as detailed below:



Survey: MRI 2018 Spring - M18S

Population : (Wall unit/wall system ~ Household Furnishings: Big Ticket Items - Bought in last 12 months OR Sofa/sectional Recliner Sofa-bed convertible Other living room furniture Bed frame/headboard Box Spring Mattress Cedar chests bedroom Kitchen Dining Family Home office Piano Picture frames custom made Window coverings Wall-to-wall carpet Horizontal

View Results As: Table

Data Item: Averages

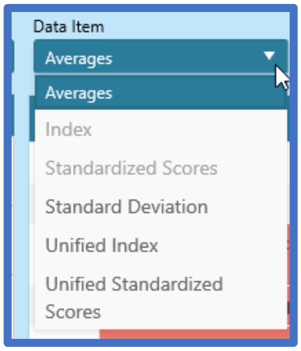
Highlighting: Open

Variables used : 6 (out of 6)

Solutions		Q
2 Clusters	Q	72%
3 Clusters	Q	63%
4 Clusters	Q	63%
5 Clusters	Q	66%
6 Clusters	Q	65%
7 Clusters	Q	64%
8 Clusters		

	Variable	Type	Rank ↑	Determination	Total	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster
⚙	Audience (000)				57459	26153	5534	8036	9456	
	Respondents				5402	2446	549	742	857	
6	Advertising on mobile phones provides me with useful information about bargains. ~ Agree - Mobile Advertising	Binary	1	65 %	0.28	0.04	0.11	0.03	0.64	
2	Advertising on radio provides me with useful information about bargains. ~ Agree - Radio Advertising	Binary	2	64 %	0.34	0.04	0.00	1.00	0.38	
3	Advertising in newspapers provides me with useful information about bargains. ~ Agree - Newspaper Advertising	Binary	3	50 %	0.41	0.19	0.91	0.73	0.04	
5	Advertising on the Internet provides me with useful information about bargains. ~ Agree - Internet Advertising	Binary	4	46 %	0.44	0.16	0.32	0.36	0.90	
1	Advertising on TV provides me with useful information about bargains. ~ Agree - TV Advertising	Binary	5	45 %	0.40	0.06	0.69	0.76	0.50	
4	Advertising in magazines provides me with useful information about bargains. ~ Agree - Magazine Advertising	Binary	6	42 %	0.31	0.04	0.70	0.47	0.22	

The Data Item option is a drop-down selection list in the upper left-hand corner of the screen that allows you to select your preferred statistic(s) for interpretation.



The options are:

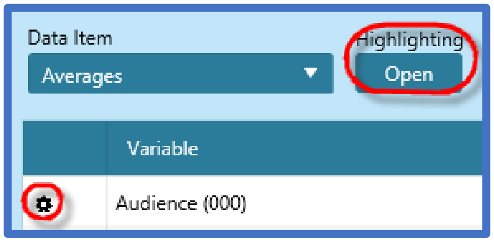
- Averages
- Standardized scores
- Index
- Standard deviation

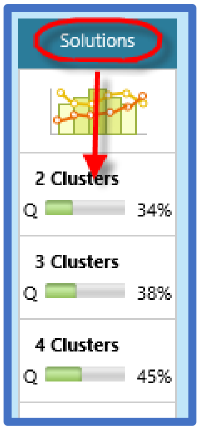
When Likert agree=1 type questions are used, the following 2 options are available

- Unified standardized scores
- Unified index

Highlighting allows the user to highlight either the row descriptions (variables) and/or the values for the data items according to their importance.

Variable Options Cog There is a cog (circled) which, if clicked, will allow you to deselect variables e.g. deselect variables with a low determination. Having deselected some variables, you can re-cluster.





Solutions allow you to select the number of clusters (vertical ribbon on the left)

Variable Management allows you to review the Variable Importance Report, review data treatment and re-visit the Cluster Solution Quality Chart

View results in table or graphical form (upper left corner)

Preferences in upper right-hand corner (allows the user to change defaults/settings)

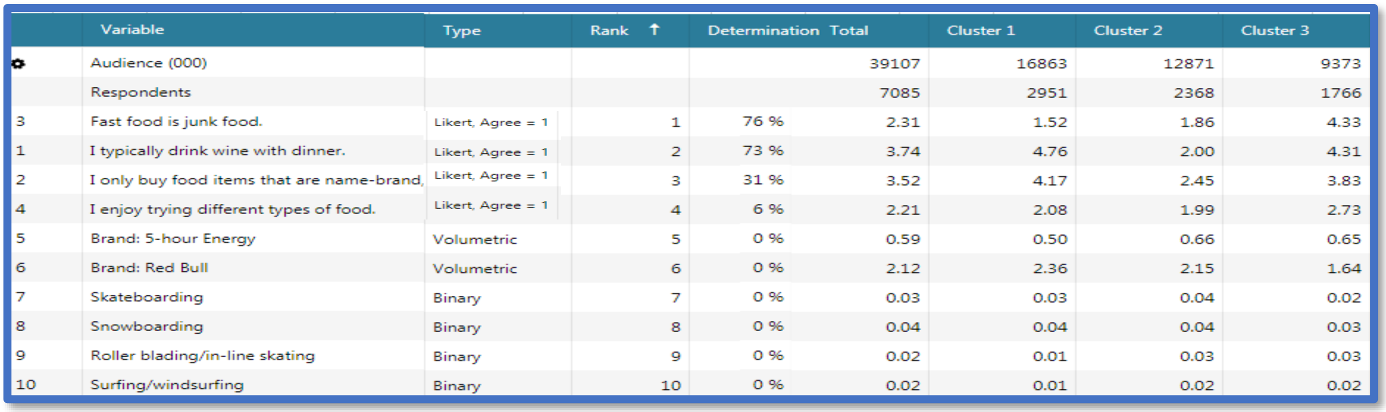
Export to Survey Time and Excel (upper right corner)

4. Explanation of the data items and question types

In order to explain how the statistics can be read for all variable types, we are showing a different target audience with a mixture of questions types.

Averages

The Averages within this example report have different meanings according to the question type. The target audience is people who buy energy drinks.



	Variable	Type	Rank ↑	Determination	Total	Cluster 1	Cluster 2	Cluster 3
⚙	Audience (000)				39107	16863	12871	9373
	Respondents				7085	2951	2368	1766
3	Fast food is junk food.	Likert, Agree = 1	1	76 %	2.31	1.52	1.86	4.33
1	I typically drink wine with dinner.	Likert, Agree = 1	2	73 %	3.74	4.76	2.00	4.31
2	I only buy food items that are name-brand,	Likert, Agree = 1	3	31 %	3.52	4.17	2.45	3.83
4	I enjoy trying different types of food.	Likert, Agree = 1	4	6 %	2.21	2.08	1.99	2.73
5	Brand: 5-hour Energy	Volumetric	5	0 %	0.59	0.50	0.66	0.65
6	Brand: Red Bull	Volumetric	6	0 %	2.12	2.36	2.15	1.64
7	Skateboarding	Binary	7	0 %	0.03	0.03	0.04	0.02
8	Snowboarding	Binary	8	0 %	0.04	0.04	0.04	0.03
9	Roller blading/in-line skating	Binary	9	0 %	0.02	0.01	0.03	0.03
10	Surfing/windsurfing	Binary	10	0 %	0.02	0.01	0.02	0.02

Interpretation of averages for Likert scale questions

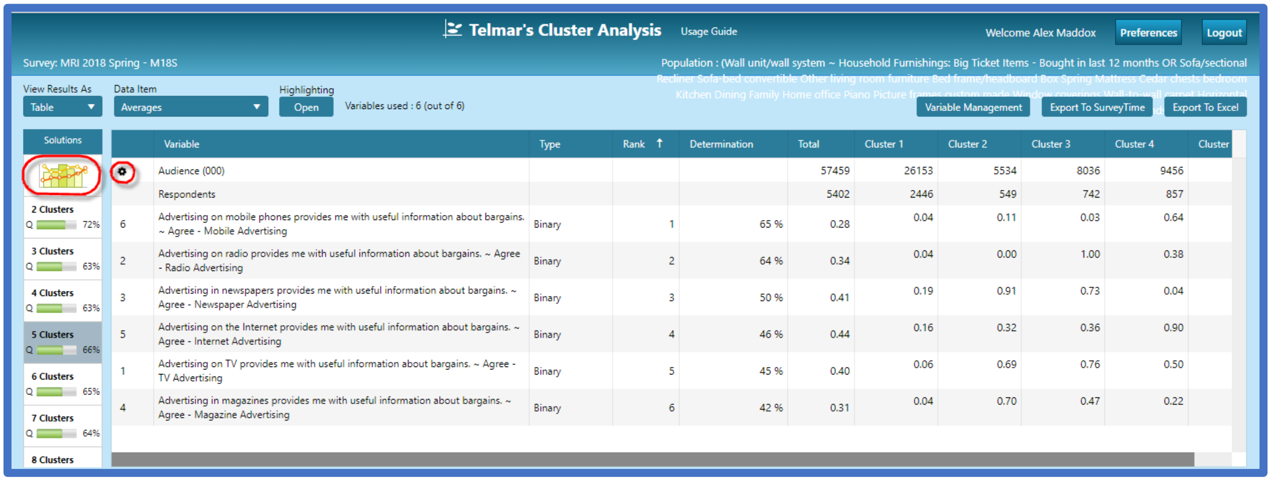
Where respondents answered their agreement on a scale from 1 to 5 with 1 meaning agree. As in the above report Likert agree=1. The lowest average value is the strongest agree average score. The score of 1.52 for “fast food is junk” means the average answer is in between “tend to agree” and “definitely agree”.

Interpretation of averages for Volumetric variables

These values or averages represent the volume or “*number of pieces bought*” amount spent pieces, etc.” A score of 2.12 represents the number of drinks.

Interpretation of averages for Binary variables

Binary variables have a value of 0 or 1 i.e. they are mutually exclusive values (yes or no). In the example shown below, the average value means the frequency of positive answers (yes’s or agrees).



Survey: MRI 2018 Spring - M18S

Population : (Wall unit/wall system ~ Household Furnishings: Big Ticket Items - Bought in last 12 months OR Sofa/sectional Recliner Sofa-bed convertible Other living room furniture Bed frame/headboard Box Spring Mattress Cedar chests bedroom Kitchen Dining Family Home office Piano Picture frames custom made Window coverings Wall-to-wall carpet Horizontal

View Results As: Table

Data Item: Averages

Highlighting: Open

Variables used : 6 (out of 6)

	Variable	Type	Rank ↑	Determination	Total	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster
	Audience (000)				57459	26153	5534	8036	9456	
	Respondents				5402	2446	549	742	857	
6	Advertising on mobile phones provides me with useful information about bargains. ~ Agree - Mobile Advertising	Binary	1	65 %	0.28	0.04	0.11	0.03	0.64	
2	Advertising on radio provides me with useful information about bargains. ~ Agree - Radio Advertising	Binary	2	64 %	0.34	0.04	0.00	1.00	0.38	
3	Advertising in newspapers provides me with useful information about bargains. ~ Agree - Newspaper Advertising	Binary	3	50 %	0.41	0.19	0.91	0.73	0.04	
5	Advertising on the Internet provides me with useful information about bargains. ~ Agree - Internet Advertising	Binary	4	46 %	0.44	0.16	0.32	0.36	0.90	
1	Advertising on TV provides me with useful information about bargains. ~ Agree - TV Advertising	Binary	5	45 %	0.40	0.06	0.69	0.76	0.50	
4	Advertising in magazines provides me with useful information about bargains. ~ Agree - Magazine Advertising	Binary	6	42 %	0.31	0.04	0.70	0.47	0.22	

Looking at the example screen above, 0.73 for cluster 3 for Newspapers means that 73% of the cluster 3 respondents agreed, that “*Advertising in newspapers provide me with useful information about bargains”.* The averages contain information about cluster differences, because data has no other sources of variation. Look, for example, at *Radio* in Cluster 2, it has no radio listeners, while cluster 3 has 100% and cluster+ is somewhere in between. As a result, this variable has a rather high Determination score of 64%. By contrast, Magazines averages (frequencies) are more similar to each other from cluster to cluster, and determination is 42%.

In the case of binary variables, the averages **can** be sufficient to understand what is going on.

Using the average data items, the values of the binary and Likert scale variables can be compared with each other, either within each variable type or amongst other variables of the same type. One can say, that frequency of snowboarding in cluster 1 is higher than in cluster 3, but one cannot compare this value (0.04) with, say, average rank of answers for the first question about junk food (2.31) as that is an average of the 1 to 5 scale of answers (with 1 being strongly agree).

Again, using the average data items with volumetrics, the question of comparison across numerical variables depends on whether they are comparable. You can compare similar things e.g. time with time, or number of visits, or even expenditure on similar items. Obviously, you cannot compare total amount of beer drunk with bottles of shampoo or money spent.

Standardized scores

The standardized score shows how far the average value in a cluster is located from the average value of the whole population (which is always equal to zero). This distance is measured in standard deviations. The value 0.5, for example, means, that average value in a cluster is higher than total average for half of the standard deviation; value -1 means, that the average value is lower than total average for one standard deviation. The higher values of standardized scores – the more distinguishable clusters are for the given variable. The closer to zero – the less difference between clusters is observed.

The **standardized scores** are calculated as follows:

Standardized score of X = (X-Average(X))/Standard deviation (X) (1)

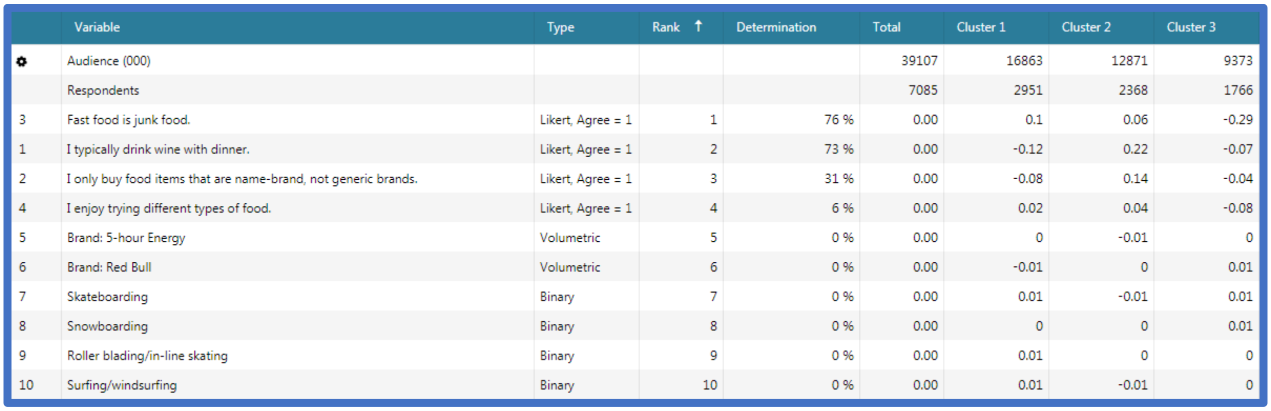
Standardized scores have several important features:

- Since standardized scores are measured for the entire population, the average value is always equal to 0
- The standard deviation of the Standardized scores is always equal to 1
- The standardized scores in each cluster are the weighted average of all respondents scores in the cluster.
- Typically, if data has a normal distribution, one may expect a standardized score higher/lower than +-1 in about 34% of cases; higher/lower than +-2 in about 95%, and higher/lower than +-3 in about 1%. In other words, +-3 is exceptional i.e. a very rare event. If certain cluster(s) have standardized scores higher than 3 it is due to extraordinarily high or low values for the given variable (or, possibly, an outlier).
- The above proportions roughly work for Likert scale variables
- For ***binary*** variables the situation is different, because, unlike Likert or volumetric scales, average and standard deviations here are not independent indicators, but are strictly determined by the frequency. The values given in the following table apply for the binary variable 0/1.

Frequency	1%	2%	5%	10%	20%	30%	50%
Max. Standardized score	10	7.0	4.4	3	2	1.53	1

It means, that if the frequency of a binary variable is, say, 10%, the standardized score in any cluster cannot exceed 3, even if all the respondents in the cluster have a value of 1 and everything else is 0; i.e., the cluster is as distinctive as possible, the average standardized score in this cluster will be 3.

For this reason, one should use caution when comparing standardized scores of Likert and binary variables. If one cluster has a standardized score for Likert scale variable of 3, it is an outstanding value. But 3 for a binary variable with a frequency of 2% means that in this cluster about half or fewer have values of 1 and the rest zero i.e. it is not something very special. However, the standardized scores retain their main feature: 3 still means, that value in this cluster **is larger** than the average value by 3 standard deviations, i.e. the cluster is not typical. Due to this complication, average frequency or Index for binary variables can be easier to understand than standardized scores (see below).



	Variable	Type	Rank ↑	Determination	Total	Cluster 1	Cluster 2	Cluster 3
⚙	Audience (000)				39107	16863	12871	9373
	Respondents				7085	2951	2368	1766
3	Fast food is junk food.	Likert, Agree = 1	1	76 %	0.00	0.1	0.06	-0.29
1	I typically drink wine with dinner.	Likert, Agree = 1	2	73 %	0.00	-0.12	0.22	-0.07
2	I only buy food items that are name-brand, not generic brands.	Likert, Agree = 1	3	31 %	0.00	-0.08	0.14	-0.04
4	I enjoy trying different types of food.	Likert, Agree = 1	4	6 %	0.00	0.02	0.04	-0.08
5	Brand: 5-hour Energy	Volumetric	5	0 %	0.00	0	-0.01	0
6	Brand: Red Bull	Volumetric	6	0 %	0.00	-0.01	0	0.01
7	Skateboarding	Binary	7	0 %	0.00	0.01	-0.01	0.01
8	Snowboarding	Binary	8	0 %	0.00	0	0	0.01
9	Roller blading/in-line skating	Binary	9	0 %	0.00	0.01	0	0
10	Surfing/windsurfing	Binary	10	0 %	0.00	0.01	-0.01	0

Standardized scores using the same data as the averages table

The features of standardized scores allow us to see very quickly where the difference between values of the variable in clusters lies, regardless of the types of variables. Look at the above table. One may see, that all scores are very small, around zero – meaning, that cluster did not find groups of respondents sharply different from each other. What is important – even with a high level of determination, it does not guarantee, that scores will be large – yes, for the first variable determination = 76%, and scores are definitely larger (0.1, -0.29) than for the last one with determination=0 (-0.01, 0.00), but in both cases they are far even from one standard deviation. It implies an important rule: ***analyzing clustering results, one should look not only at the determination or CH, but also the actual difference in values between different clusters.*** A high value of the quality indicators (determination and CH) may in practice not mean much, when it comes to the average values in the groups.

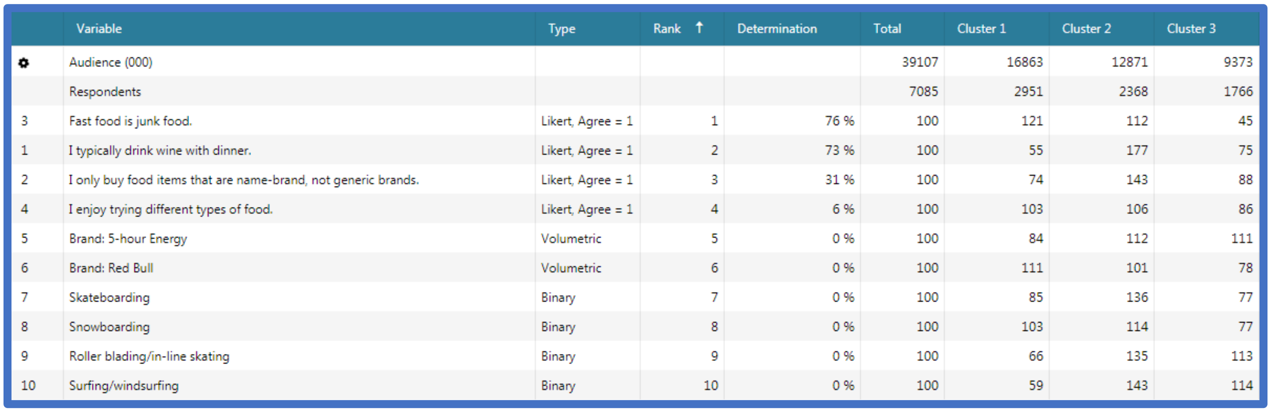
Index

The index allows comparison between different variables in a simpler form than standardized scores.

It is calculated as follows:

Index = 100 * (Average value in cluster / Average value in population for the given variable)

Look at the indices in the report below. What might have been a small difference for standardized scores may look more interesting here. For example, the index for surfers in cluster 1 is just 59, while in cluster 2 the index is much higher at 143. In cluster 1 therefore, surfers are 41% less likely than in the total population, whereas in cluster 2 surfers are 43% more likely than in the total population. In the standardized scores table, this difference was less clear. Technically, one may see that frequencies are twice as different, but the mixture of rounding error and the “small negative values” (0.01 vs 0.02) could prevent the majority of people noticing the difference. The index, contrary to that, reveals it quite clearly.



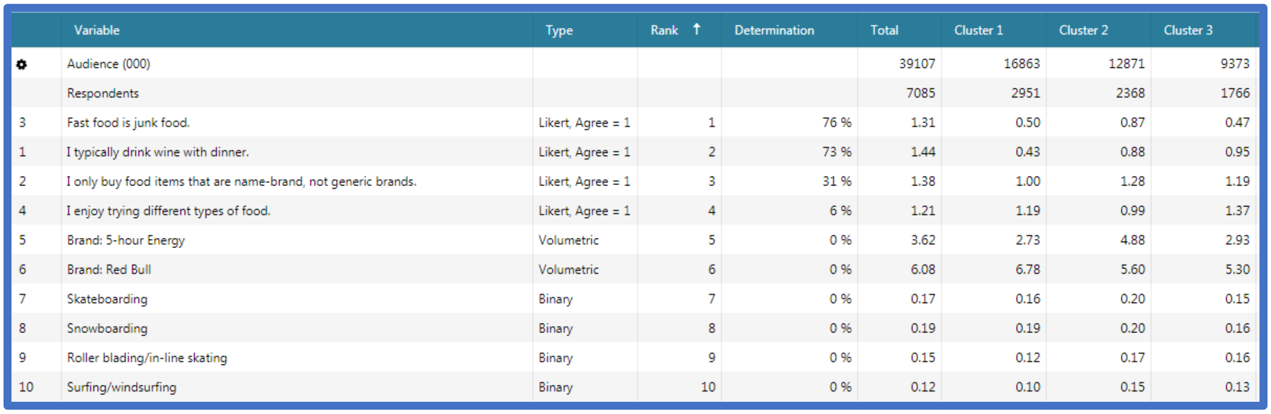
	Variable	Type	Rank ↑	Determination	Total	Cluster 1	Cluster 2	Cluster 3
⚙	Audience (000)				39107	16863	12871	9373
	Respondents				7085	2951	2368	1766
3	Fast food is junk food.	Likert, Agree = 1	1	76 %	100	121	112	45
1	I typically drink wine with dinner.	Likert, Agree = 1	2	73 %	100	55	177	75
2	I only buy food items that are name-brand, not generic brands.	Likert, Agree = 1	3	31 %	100	74	143	88
4	I enjoy trying different types of food.	Likert, Agree = 1	4	6 %	100	103	106	86
5	Brand: 5-hour Energy	Volumetric	5	0 %	100	84	112	111
6	Brand: Red Bull	Volumetric	6	0 %	100	111	101	78
7	Skateboarding	Binary	7	0 %	100	85	136	77
8	Snowboarding	Binary	8	0 %	100	103	114	77
9	Roller blading/in-line skating	Binary	9	0 %	100	66	135	113
10	Surfing/windsurfing	Binary	10	0 %	100	59	143	114

Index for the same data as the standardized scores and averages

Standard deviation

The standard deviation is a measure of variability – the higher its value in a cluster, the less compact this cluster is. The table below shows the standard deviations for the same data as before. Look at the statement “Fast food is junk food”, for example. Cluster 3 is denser than cluster 2 (almost double) - (0.47 vs 0.87). If one wants to target that cluster – the user will face a comparatively homogenous group of people, where all are concentrated around the average value in this cluster (which was 4.33 in the averages table). In fact, it means, that people in this cluster strongly disagree with the statement, and their deviation from the typical value (4.33) is small – just 0.47. If we recall the description of the standardized scores above – it is applicable here as well. Namely, roughly about 66% of all respondents in this cluster will lay in interval 4.33+-0.47, 95% - in interval 4.33+-2*0.47 and so on.

Bear in mind, that standard deviation has the same unit of measure as average value – so, they could be used for comparison only between clusters, not between variables (unless variables are measured in the same units).



	Variable	Type	Rank ↑	Determination	Total	Cluster 1	Cluster 2	Cluster 3
⚙	Audience (000)				39107	16863	12871	9373
	Respondents				7085	2951	2368	1766
3	Fast food is junk food.	Likert, Agree = 1	1	76 %	1.31	0.50	0.87	0.47
1	I typically drink wine with dinner.	Likert, Agree = 1	2	73 %	1.44	0.43	0.88	0.95
2	I only buy food items that are name-brand, not generic brands.	Likert, Agree = 1	3	31 %	1.38	1.00	1.28	1.19
4	I enjoy trying different types of food.	Likert, Agree = 1	4	6 %	1.21	1.19	0.99	1.37
5	Brand: 5-hour Energy	Volumetric	5	0 %	3.62	2.73	4.88	2.93
6	Brand: Red Bull	Volumetric	6	0 %	6.08	6.78	5.60	5.30
7	Skateboarding	Binary	7	0 %	0.17	0.16	0.20	0.15
8	Snowboarding	Binary	8	0 %	0.19	0.19	0.20	0.16
9	Roller blading/in-line skating	Binary	9	0 %	0.15	0.12	0.17	0.16
10	Surfing/windsurfing	Binary	10	0 %	0.12	0.10	0.15	0.13

Standard deviations

The unified index and unified standardized scores

In the majority of surveys that Telmar load, where there are Likert scale questions (5 scale attitudinal data), Telmar specially code them for clustering in the codebook. Historically Telmar have applied the value, as provided by the research company, with the majority of surveys using a value of 1 for definitely agree and 5 for definitely disagree. In order to make all types of variables go in the same direction in the reports, Telmar have created a new unified report.

Where Likert Agree is coded as 1 i.e. a low value for a positive answer (as shown in the Type column), there is a unified index and unified standardized scores report available. These reports convert (or inverse) the standardized scores and index, so that the values can be compared with the volumetric and binary values, where a higher value is more positive. To put it another way, the values are normalized so that they can all be read in the same direction with all of the higher values meaning that the respondents within a cluster are "more positive towards".

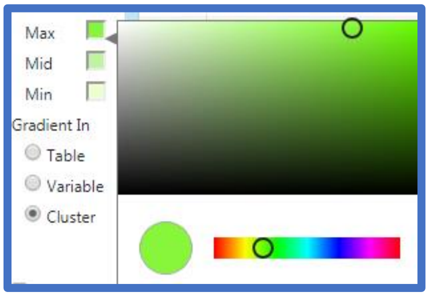
In **Unified mode** the other variable types (volumetric and binary) are not reversed. It is only the Likert scale variables that need reversing (or unifying).

5. Highlighting

The program provides ***highlighting*** options as an aid for interpretation. The tables shown in section 4, above can be easier to interpret with the aid of highlighting.

After clicking on the highlighting button, the options become available in the ribbon on the left part of the screen. It is divided in two sections, one for variables (the row titles) and one for data cells (the data item numbers). There are default colors for both, but they can be changed by the user.

Clicking the color buttons offers a two-ways pallet, as shown below:

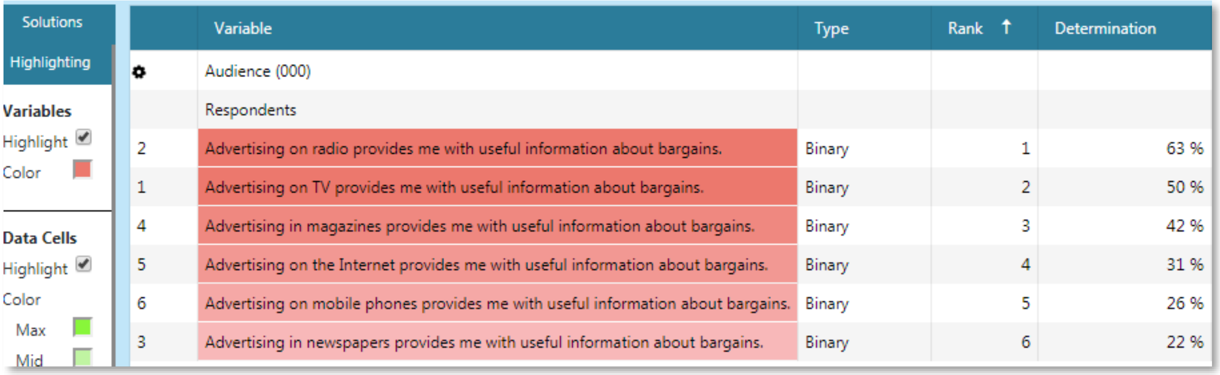


Max 🟩
Mid 🟩
Min □
Gradient In
○ Table
○ Variable
● Cluster

A pallet for manual change of the highlighting color

Highlighting variables

The higher the determination score, the more intense the background color, the lower the score, the less intensity and color. The default colors (again, let stick with one version, color, not two) for variables are shown below.



Solutions
Highlighting

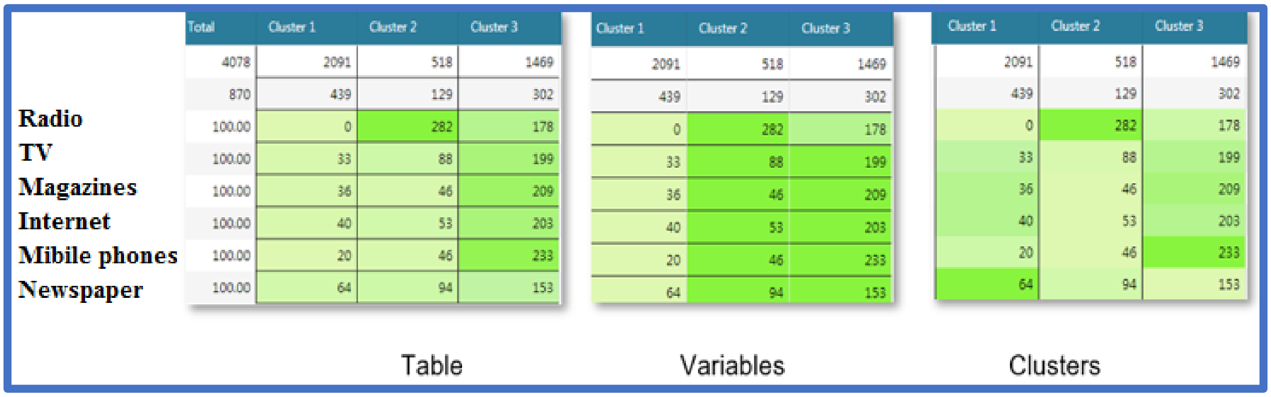
Variables
Highlight ☑
Color 🟥

Data Cells
Highlight ☑
Color
Max 🟩
Mid 🟩

	Variable	Type	Rank ↑	Determination
⚙	Audience (000)			
	Respondents			
2	Advertising on radio provides me with useful information about bargains.	Binary	1	63 %
1	Advertising on TV provides me with useful information about bargains.	Binary	2	50 %
4	Advertising in magazines provides me with useful information about bargains.	Binary	3	42 %
5	Advertising on the Internet provides me with useful information about bargains.	Binary	4	31 %
6	Advertising on mobile phones provides me with useful information about bargains.	Binary	5	26 %
3	Advertising in newspapers provides me with useful information about bargains.	Binary	6	22 %

Data cells highlighting

A useful feature of data cells highlighting is that it can be done in three different directions, as shown below based on data from the furniture example campaign. Default Index highlighting in three different modes as used for the furniture clusters



	Total	Cluster 1	Cluster 2	Cluster 3
	4078	2091	518	1469
	870	439	129	302
Radio	100.00	0	282	178
TV	100.00	33	88	199
Magazines	100.00	36	46	209
Internet	100.00	40	53	203
Mibile phones	100.00	20	46	233
Newspaper	100.00	64	94	153

Table

Cluster 1	Cluster 2	Cluster 3
2091	518	1469
439	129	302
0	282	178
33	88	199
36	46	209
40	53	203
20	46	233
64	94	153

Variables

Cluster 1	Cluster 2	Cluster 3
2091	518	1469
439	129	302
0	282	178
33	88	199
36	46	209
40	53	203
20	46	233
64	94	153

Clusters

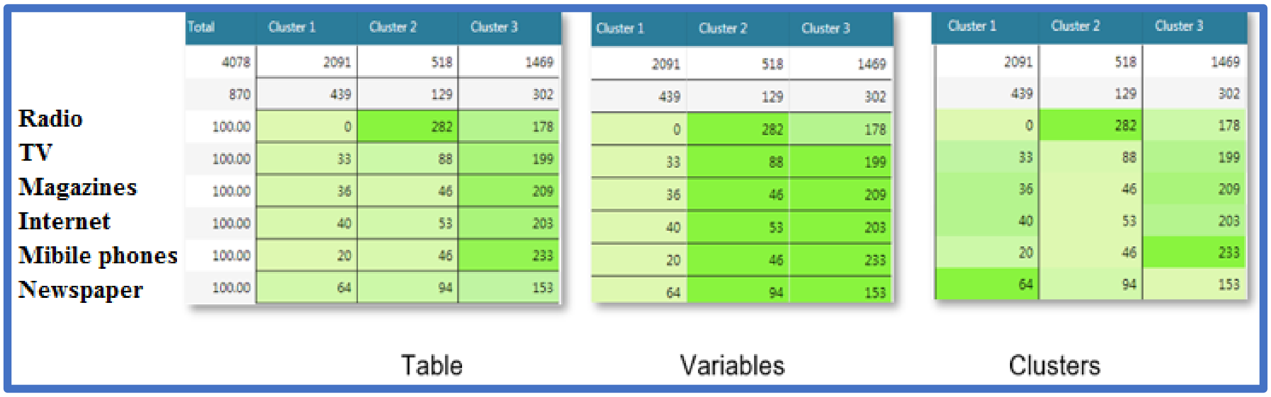
Table mode

In Table mode all cells of data are compared. The brightest color is assigned to the highest value in the table (282) and the color intensity decreases as the value reduces. It highlights the most interesting data cells in the table. Looking at the table, we see, that the concentration of radio adopters (row 1) is almost three times higher than average in cluster 2; the others are not extremely different.

Variables mode

Variables mode puts the brightest color in each row independently, which draws attention to the most important cells for each variable. Here, for example, it is immediately clear that cluster 1 contains very low indexes for all variables, - so there is no a cluster group worth advertising to. But cluster 3, by contrast, has all variables bright – very promising sign. And those two important conclusions are obtained in a matter of second, just glancing at the colored table!

Cluster mode



	Total	Cluster 1	Cluster 2	Cluster 3
	4078	2091	518	1469
	870	439	129	302
Radio	100.00	0	282	178
TV	100.00	33	88	199
Magazines	100.00	36	46	209
Internet	100.00	40	53	203
Mibile phones	100.00	20	46	233
Newspaper	100.00	64	94	153

Table

Cluster 1	Cluster 2	Cluster 3
2091	518	1469
439	129	302
0	282	178
33	88	199
36	46	209
40	53	203
20	46	233
64	94	153

Variables

Cluster 1	Cluster 2	Cluster 3
2091	518	1469
439	129	302
0	282	178
33	88	199
36	46	209
40	53	203
20	46	233
64	94	153

Clusters

In Cluster mode the user can see what is going on within each cluster. The highest value within each cluster represents the “cluster profile”. Cluster 2, for example, is something special – it has one bright cell for radio (row 1) and low indices (below 100) for other media. Obviously, the people in this cluster could only be reached by radio.

The lines on the grids show the direction of the value comparisons. Darker shading or highlighting is used to draw attention to the higher values. With horizontal lines the values are compared and shaded horizontally. With vertical lines, the values are compared and shaded vertically and in table mode, where you find grid lines all of the values are compared and the darker highlighting indicates the highest values.

All changes made by the user in highlighting can be saved as a preference by ticking the check box in the lower part of the panel – they will be used as a new default when the user next enters the program.

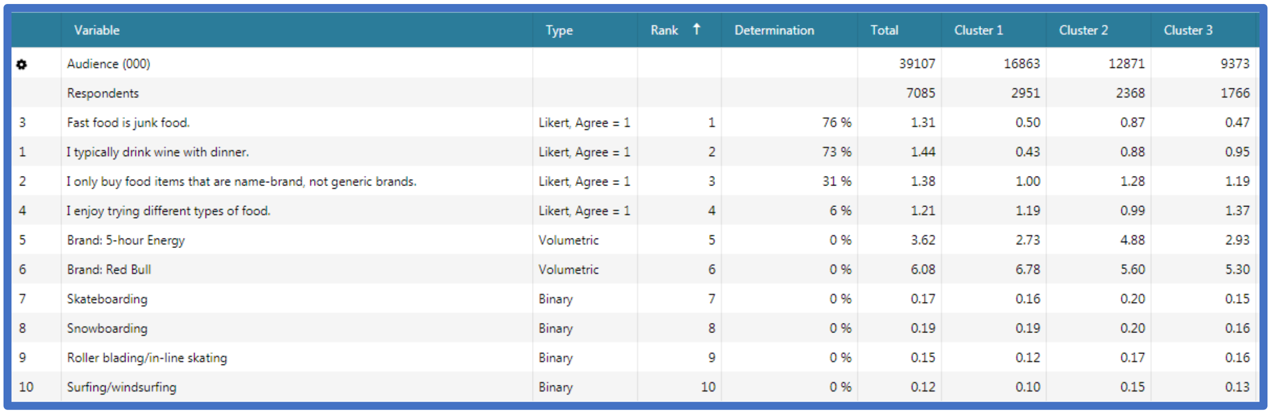
6. Variables management

As stated in our introduction, ***“the aim of cluster is to achieve groups of respondents that are similar to each other within groups and as different as possible from other groups”.*** In order to achieve this aim, the input is key.

Determination

The program provides a criteria evaluation aid, called a determination score to help the user understand the quality of the variable's contribution. This has already been explained in section 2 of this usage guide and more information can be found in the technical appendix. The variables, entered as rows, in SurveyTime can be evaluated within the cluster program. A ranked report with the variables sorted by their determination score can help the user can decide which variables to short-list e.g. whether to include variables with a high score and/or remove variables with a low score.

Look, for example, at the variables below. Just first three of them have a determination score over 30%, the fourth one has a determination score of 6% and the others have a zero score. How useful are those last ones in clustering? Are they needed at all? If the clustering algorithm makes such groups that are not distinguishable among clusters for those variables – maybe, it could do a better job without those variables?



	Variable	Type	Rank ↑	Determination	Total	Cluster 1	Cluster 2	Cluster 3
⚙	Audience (000)				39107	16863	12871	9373
	Respondents				7085	2951	2368	1766
3	Fast food is junk food.	Likert, Agree = 1	1	76 %	1.31	0.50	0.87	0.47
1	I typically drink wine with dinner.	Likert, Agree = 1	2	73 %	1.44	0.43	0.88	0.95
2	I only buy food items that are name-brand, not generic brands.	Likert, Agree = 1	3	31 %	1.38	1.00	1.28	1.19
4	I enjoy trying different types of food.	Likert, Agree = 1	4	6 %	1.21	1.19	0.99	1.37
5	Brand: 5-hour Energy	Volumetric	5	0 %	3.62	2.73	4.88	2.93
6	Brand: Red Bull	Volumetric	6	0 %	6.08	6.78	5.60	5.30
7	Skateboarding	Binary	7	0 %	0.17	0.16	0.20	0.15
8	Snowboarding	Binary	8	0 %	0.19	0.19	0.20	0.16
9	Roller blading/in-line skating	Binary	9	0 %	0.15	0.12	0.17	0.16
10	Surfing/windsurfing	Binary	10	0 %	0.12	0.10	0.15	0.13

Determination has to be viewed in the context of the overall result. We cannot assume that because the determination score is low, the variable should be removed. Look at fig 2 below for instance. Whichever variable, X or Y, we take off of fig 2, the structure will not be explained, only two variables together allow an explanation.

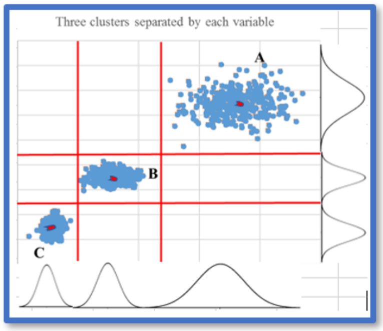


Fig 1. Three clusters separated by each variable independently

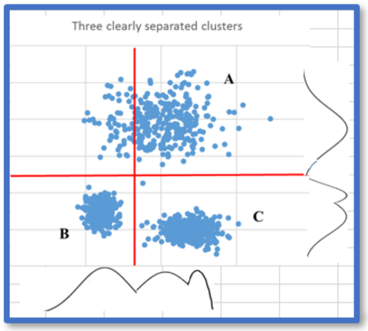
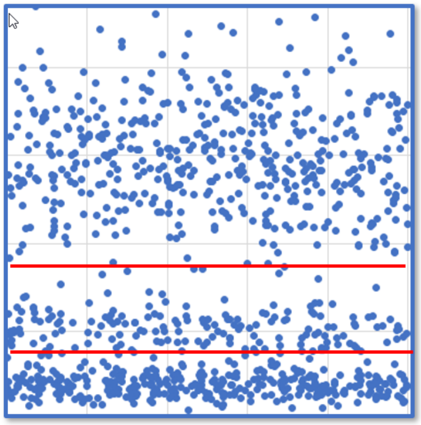


Fig.2. Clusters which cannot be separated by single variables

Let’s look at another example shown below. This provides a very different picture. Clearly, the distinction between groups is determined by Y variable only, while adding X (which has a determination of 0%) just makes it more complicated, because it adds a lot of noise to the data, when the distances between objects are calculated. Elimination of X would help to reveal the real structure (red lines separate three clusters).



Variable X elimination helps to find clusters by variable Y

In general, there are several reasons to reduce the number of variables for clustering and work only with those which have high Determination:

- It may help to make better clustering (as in the above variable X elimination);
- It can help to prioritize variables, which are most important for the cluster i.e. bear some substantial information about the data;
- It makes data description easier
- High determination for each variable creates sharp “cut off” structure, when data could be described like cross-table (as on fig.1), rather than in more vague terms of clouds (as on fig. 2).

Whilst determination is certainly a useful criterion for evaluating variables, other criteria might also affect your choice of the short list of variables. For example, if you are exploring new products or new markets, you may wish to include criteria that will help influence the creation of the groups e.g. if you were segmenting the banking sector in order to launch the first online banking brand, financial questions might give a high determination score, but attitudes to internet, security, technology, even if they had a low determination score, might be important variable to include as inputs.

Determination is, strictly speaking, a heuristic to help guide variable selection. There is no optimal or perfect solution for showing you what the exact variables are that will provide the best cluster groups.

Variables with a low Determination are not always bad – the example on fig. 2 proves the opposite, but there are other considerations before making the final decision.

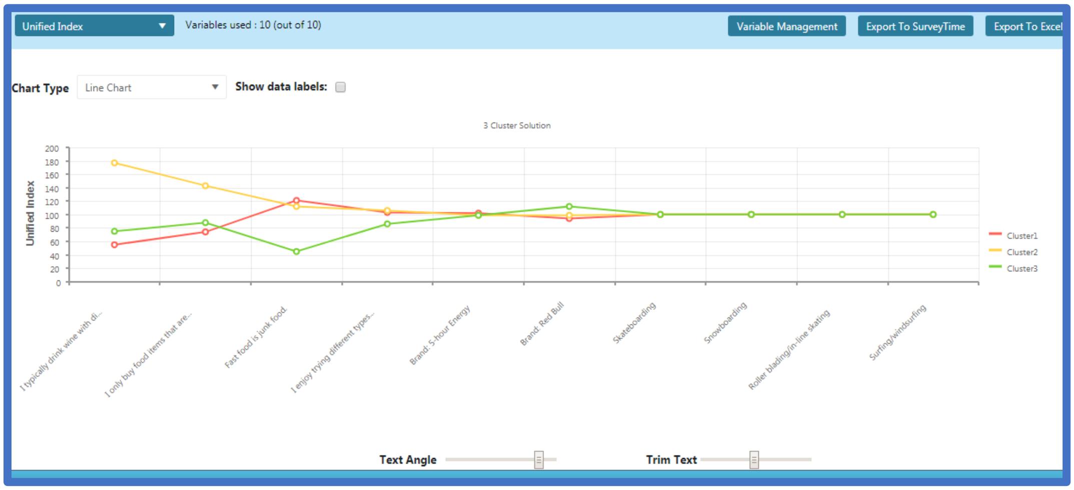
Within the “recommended solution screen” there is a ***Variables importance button.*** This option can also be found on the main table under **Variable Management** in a pop-up menu.

7. Graphical representation of results

The results of clustering have been shown as tables in the above examples, they can also be presented in graphical form. In the top left-hand corner of the Table screen there is a *View Results as Table / Chart* option. Within that two types of charts are offered: Line chart and Radar chart.

Line chart

A line chart for the unified index is shown below. The data is as before, using the data from the energy drinks with the different variable types. Looking at the line chart, it is clear that the indices are very different for the first variable (drinking wine) and undistinguishable for the last four variables. The user can change the angle of the **text** and also **trim** the titles. This is more important when the titles are long and with a large number of variables.



Unified Index

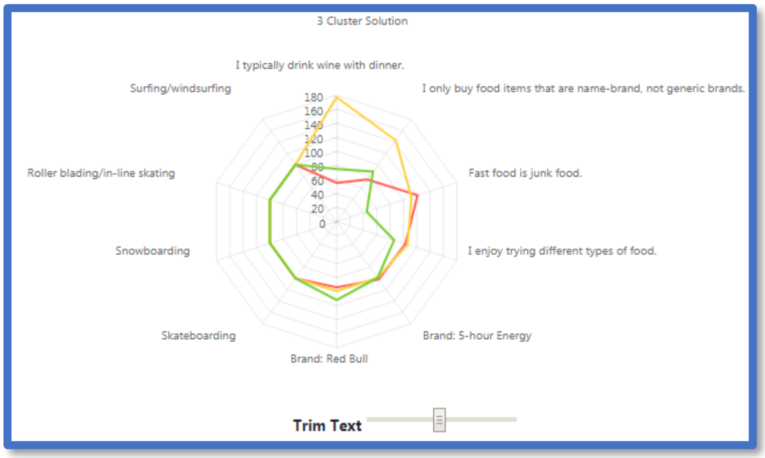
Variables used : 10 (out of 10)

Chart Type: Line Chart

Show data labels: □

Radar chart

There is also a radar chart for the same data. It can often be easier to interpret, than the line chart because it is more compact. The difference for the wine question is very pronounced for the second cluster, as well as the neutrality of cluster 3, where all indexes are close to 100.

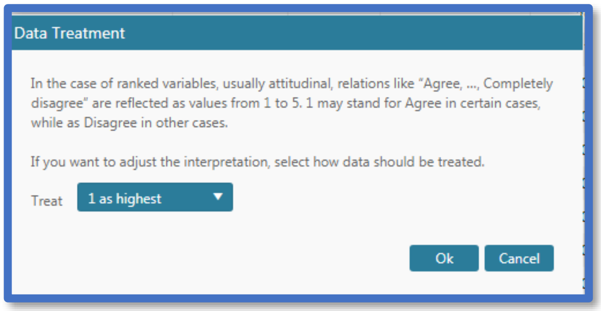


8. Other options on the main table

There are some other options which may help in some situations.

a) Data treatment

Data treatment (found in the Variable Management menu) offers the option of changing the value of the Likert scale agree questions (or variables). Data treatment can invert values for the Likert scale variables, changing the value from 1 as high to 5. Most of the time this is unnecessary. For more information call your local Telmar helpdesk.



Data Treatment

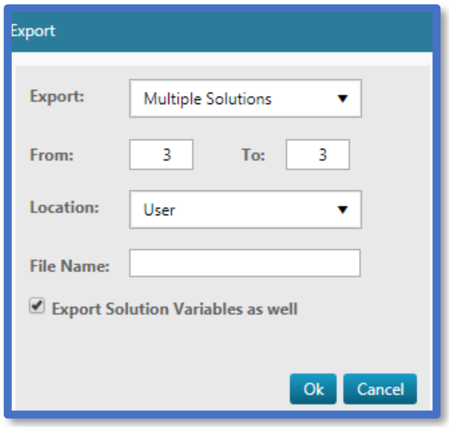
In the case of ranked variables, usually attitudinal, relations like "Agree, ..., Completely disagree" are reflected as values from 1 to 5. 1 may stand for Agree in certain cases, while as Disagree in other cases.

If you want to adjust the interpretation, select how data should be treated.

Treat 1 as highest

b) Export to SurveyTime

Export to Survey Time allows the user to export the clusters for further analysis in SurveyTime. The button offers different options for saving. *Current solution* saves the solution that is currently under review. Alternatively, a user can export a range (Multiple solutions). Once they have been exported, the results of clustering can be found in Own Codes in Survey Time. They can be saved for just the user, or, if saved on the company drive, they can be shared across the client's company. The variables, used for clustering, should also be saved, as they are a record of your final input. This is especially important, if you have removed variables, for example those with low determination, because you may need to list these when presenting the final clusters to your client.


Export

Export: Multiple Solutions

From: 3 To: 3

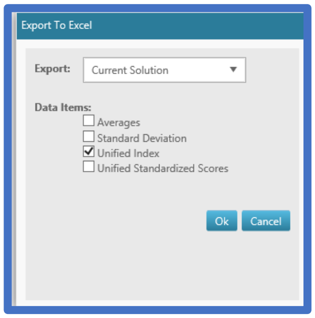
Location: User

File Name:

☑ Export Solution Variables as well

C) Export to Excel

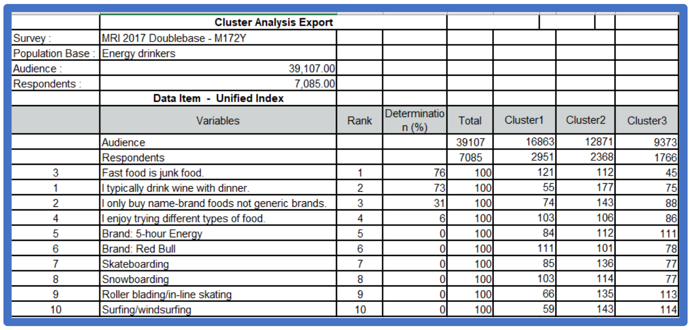
Export to Excel creates an excel file containing the tables selected by the user from a pop-up menu. There is an option to export the current solution or Multiple solutions.


Export To Excel

Export: Current Solution

Data Items:

□ Averages
□ Standard Deviation
☑ Unified Index
□ Unified Standardized Scores



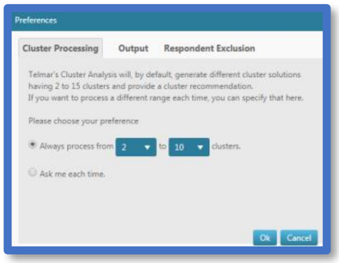
	Cluster Analysis Export						
Survey :	MRI 2017 Doublebase - M172Y						
Population Base :	Energy drinkers						
Audience :	39,107.00						
Respondents :	7,085.00						
	Data Item - Unified Index						
	Variables	Rank	Determination (%)	Total	Cluster1	Cluster2	Cluster3
	Audience			39107	16863	12871	9373
	Respondents			7085	2951	2368	1766
3	Fast food is junk food.	1	76	100	121	112	45
1	I typically drink wine with dinner.	2	73	100	55	177	75
2	I only buy name-brand foods not generic brands.	3	31	100	74	143	88
4	I enjoy trying different types of food.	4	6	100	103	106	86
5	Brand: 5-hour Energy	5	0	100	84	112	111
6	Brand: Red Bull	6	0	100	111	101	78
7	Skateboarding	7	0	100	85	136	77
8	Snowboarding	8	0	100	103	114	77
9	Roller blading/in-line skating	9	0	100	66	135	113
10	Surfing/windsurfing	10	0	100	59	143	114

9. Preferences

The Preferences button offers several different options:

a) Cluster processing

The tab *Cluster processing* offers the range for the number of clusters for the system to calculate. The narrower the range the faster calculations and the more interpretable the results, but in some cases people may need a larger number of solutions.


Preferences

Telmar's Cluster Analysis will, by default, generate different cluster solutions having 2 to 15 clusters and provide a cluster recommendation.
If you want to process a different range each time, you can specify that here.

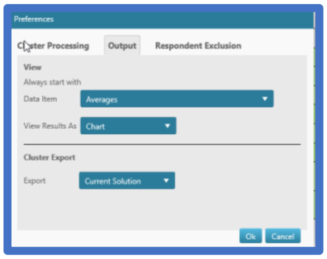
Please choose your preference

● Always process from 2 to 10 clusters.

○ Ask me each time.

b) Output

The *Output* tab defines the default view of reports e.g. show averages by default, view results as a table and export defaults e.g. current solution.



View

Always start with

Data Item: Averages

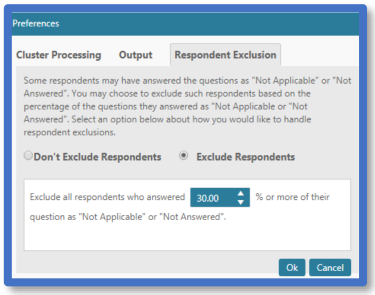
View Results As: Chart

Cluster Export

Export: Current Solution

c) Respondent exclusion

The *Respondent Exclusion* tab offers a special method of data treatment. Exclusion may be needed for some questions, where for example more than 30% of respondents did not answer the questions. This may be because they consider the questions not applicable to them e.g. "I always buy the brands my children prefer" might be irrelevant if someone does not have children. If there are large amount of not applicable or not stated answers, the program can create a group or cluster of people who have the same in common. The user can decide the % of respondents to exclude as shown here.



Some respondents may have answered the questions as "Not Applicable" or "Not Answered". You may choose to exclude such respondents based on the percentage of the questions they answered as "Not Applicable or "Not Answered". Select an option below about how you would like to handle respondent exclusions.

○ Don't Exclude Respondents ● Exclude Respondents

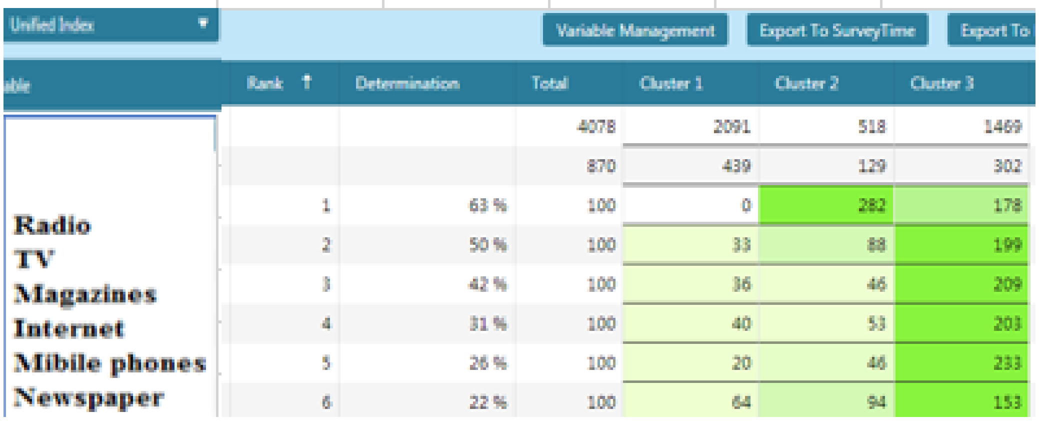
Exclude all respondents who answered 30.00 % or more of their question as "Not Applicable" or "Not Answered".

Part 2

10. Cluster analysis: a primer

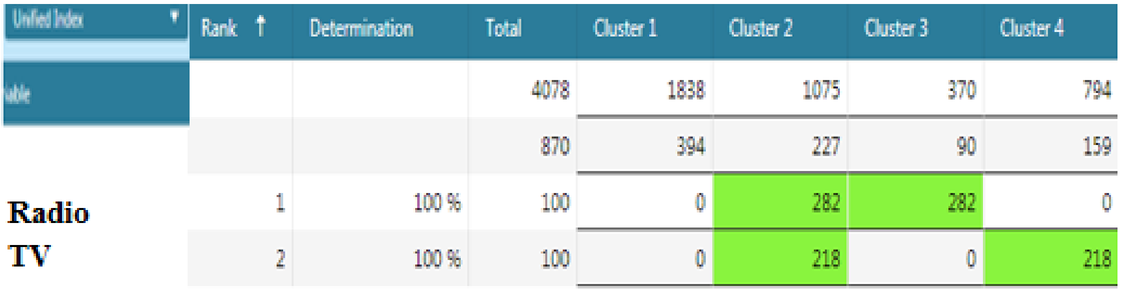
The main manual (Part 1) discuss the broad aspects of Cluster analysis and the functionality of Telmar's program. This Primer is designed to illustrate how creative implementation of cluster analysis could be performed.

As described in section 2 in Part 1, the goal of the furniture company La-z-boy was to find some prospective subsets of people to optimize its spending for media channels. It may not be rational to advertise on all six channels everywhere, if people are not receptive. One solution was presented with the furniture clusters (bargain hunters by media) from which it was identified that the first cluster was not perceptive to any advertising, while the second had a high concentration of *radio* lovers and the third had high indices for all media (as shown below):



able	Rank	Determination	Total	Cluster 1	Cluster 2	Cluster 3
			4078	2091	518	1469
			870	439	129	302
Radio	1	63 %	100	0	282	178
TV	2	50 %	100	33	88	199
Magazines	3	42 %	100	36	46	209
Internet	4	31 %	100	40	53	203
Mibile phones	5	26 %	100	20	46	233
Newspaper	6	22 %	100	64	94	153

If we follow general recommendations about using only variables with high determination (1.6) and select, say, just two top variables from the list, we get the following results:



able	Rank	Determination	Total	Cluster 1	Cluster 2	Cluster 3	Cluster 4
			4078	1838	1075	370	794
			870	394	227	90	159
Radio	1	100 %	100	0	282	282	0
TV	2	100 %	100	0	218	0	218

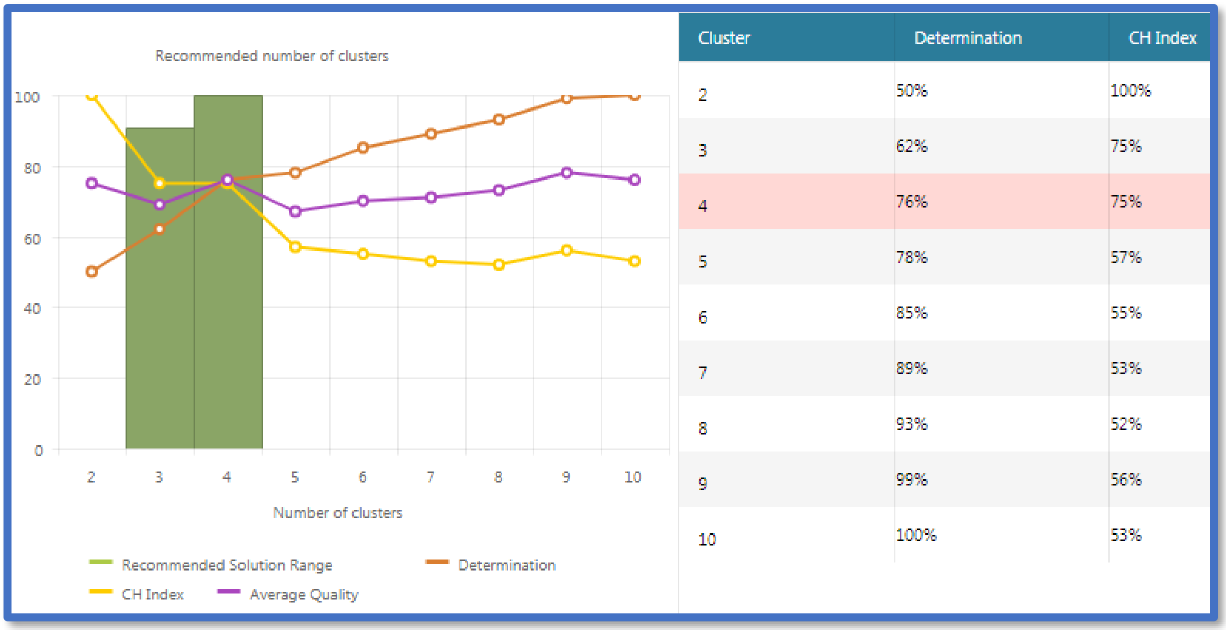
It looks very good – two determinations have maximal values, 100%; clusters are extremely distinct: cluster 2 is highly receptive for two media, 3 and 4 are exclusively receptive only for one of them. It is a very good way to make a very focused advertising campaign.

However, the problem with both of those solutions is that a high part of the population, about 50% in the first case and 45% in the second (the empty cluster 1), remain unaffected by any media. . Can it be improved?

Let’s review the advertising influence table again by looking at the indices, where highlighting is performed based on the highest value within each cluster. One can see, that the least concentrated media in Cluster 1 are radio and mobile phone. The mobile phone index is just 20 and it also has a low determination, just 26%. We could remove that variable and see if it affects the composition of Cluster 1.

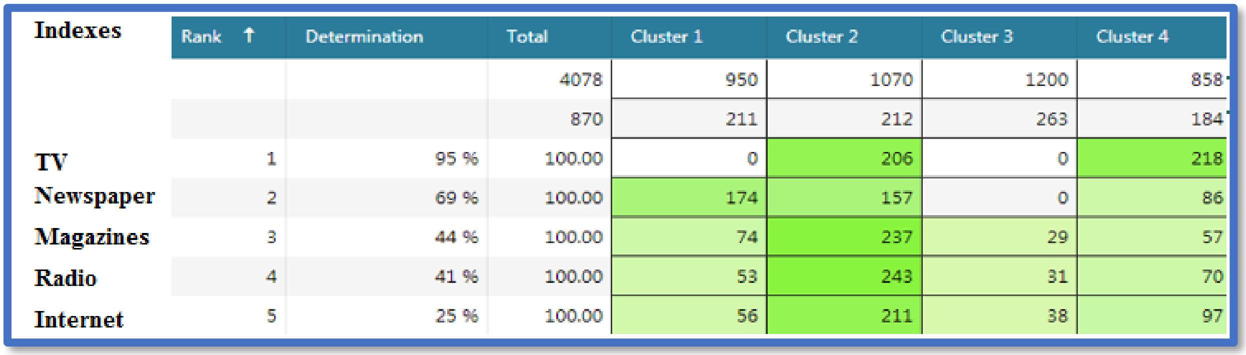
Clusters for six media variables

After removing the variable and re-running the cluster, the recommended solution is 4 clusters (see below), the criteria are in good shape. You can see, that CH after 4 clusters goes significantly down, whereas the average quality is actually stabilizing.



Cluster	Determination	CH Index
2	50%	100%
3	62%	75%
4	76%	75%
5	78%	57%
6	85%	55%
7	89%	53%
8	93%	52%
9	99%	56%
10	100%	53%

Criteria chart for five media variables



Indexes	Rank ↑	Determination	Total	Cluster 1	Cluster 2	Cluster 3	Cluster 4
			4078	950	1070	1200	858
			870	211	212	263	184
TV	1	95 %	100.00	0	206	0	218
Newspaper	2	69 %	100.00	174	157	0	86
Magazines	3	44 %	100.00	74	237	29	57
Radio	4	41 %	100.00	53	243	31	70
Internet	5	25 %	100.00	56	211	38	97

Five media, four clusters, small non-prospective cluster 3

This solution is different from the original one in two aspects: first, the number of the non-responsive people is sharply reduced (1,200 in cluster 3 vs 2,091 in Cluster 1 previously); second, the quality of the media variables and their order is significantly changed. If originally *radio* was most distinctive with D=63%, now *TV* is with D=95%, and radio is fourth in a row with D=41%. From a business standpoint, this solution could be very good: we can address about 4,000 people, using cluster 1 only *newspaper*; all media for cluster 2 and *TV* only for cluster 4.

This example illustrates couple of important principles:

- One can use some **external criteria** to make good clustering (in this case – elimination of the weakest variable for the original non-prospective cluster 1 we wanted to decrease), not only recommended minimal determination. They may be combined with recommended ones, of course.
- Importantly, if we do the same thing with *radio,* which has even smaller (zero!) value in cluster 1, the result will not be impressive at all – the same non-prospective cluster will remain
- When the number of variables and / or number of clusters is changed, **determination and CH may change in unpredictable way** – don't be surprised too much. **Experiment** instead – as we showed with elimination of *radio* and *mobile* variables.
- **Business priorities** should overcome statistical considerations in a case of conflict.

If the earlier solution using two variables could be preferred in one situation (when, say, prices namely for those two media are very good), the solution shown above is better in another situation, when the goal is to minimize the number of unreached targets.

11. Technical appendix

Many of statistical aspects of clustering have already been discussed – such issues as standardization of variables, discrimination and way of calculating of different data items do not need any additional comments. Some aspects though warrant more technical detail which is presented here.

General look at clustering problem, once again

To get an idea of what a good cluster looks like, we are going to examine fig. 1 and 2 (both repeated from Part 1). They show how some objects can be represented in a two-dimensional space. The X and Y axes could be Age and Income, Weight and Height, and so on. In both figures, one can easily see that the data is broken into three visually distinguished groups –the clusters.

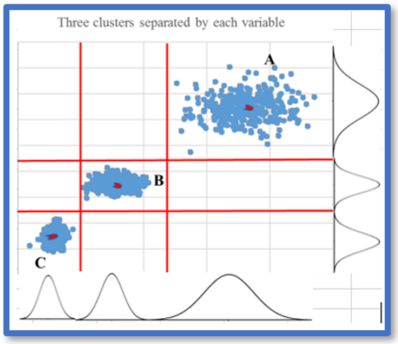


Fig 1. Three clusters separated by each variable independently

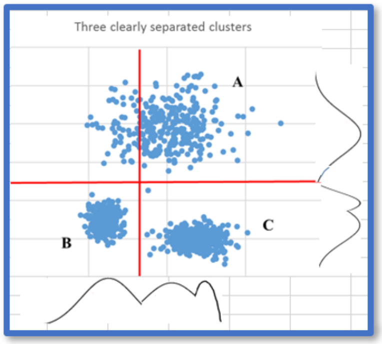


Fig.2. Clusters which cannot be separated by single variables

Clusters can have different features and characteristics, for example, Cluster C on fig. 1, is much more “compact” compared to cluster A. Cluster B is “prolongated”, while A on fig. 2 is more evenly distributed, but less condensed and “cloud-like”. Clusters can also be defined by mutual location: clusters on fig. 1 are easily separated either by X or by Y alone.

Determination

The **determination** shows what percent of total variance can be explained by clustering alone.

Looking at the fig.1clusters A, B, and C are separated in such a way, that two variables, X or Y, describe them sufficiently. If you know **only** X, it is still enough to say that A, B and C are different groups; the same can be said for variable Y. As a result, both D(X) and D(Y) will be very high, and D = average (D(X), D(Y)) will be very high, too.

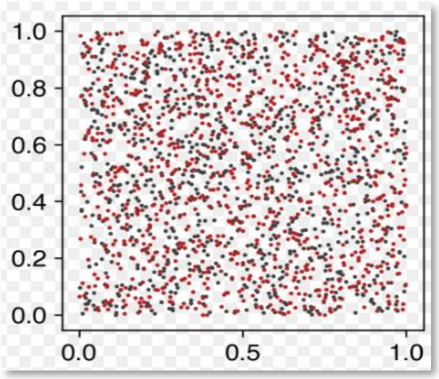
What, more precisely, does determination D(X) mean? The total variance of the X variable Var(X) is always equal to the sum of two variances: V(Within), which is the weighted average of variances in groups, and V(Between), which is the variance between the groups:

$$V(\text{Total}) = V(\text{Within}) + V(\text{Between}) \qquad (1)$$

The purpose of clustering is to separate groups in such a way that V(Between) is as large as possible. It would mean that each cluster is very far from each other - the groups are very distinguishable. It would automatically reduce V(Within), because V(total) is constant. So, the logical measure of the quality of the grouping is what percentage is V(Between) in the total variance, and it is exactly what we are talking about:

$$\text{Determination} = D = V(\text{Between})/V(\text{Total}) = 1- V(\text{Within})/V(\text{Total});\ 0<=D<=1 \qquad (2)$$

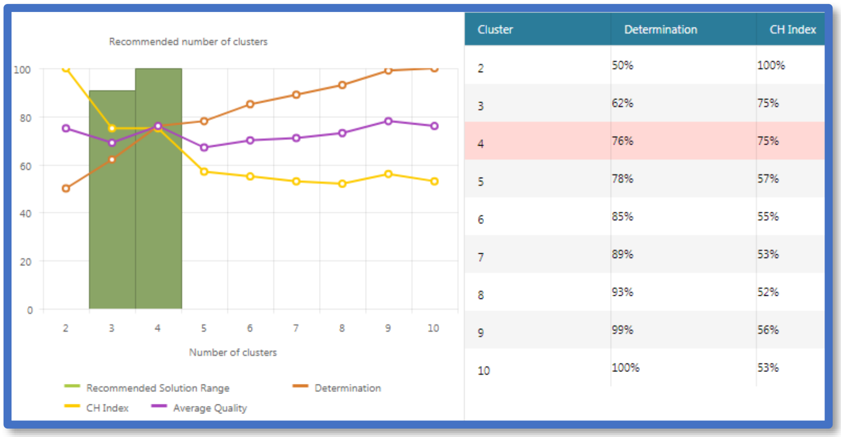
D could be zero, when clusters are maximally intermixed, undistinguishable from each other, like those shown below. And D could be 1 in extreme case, when whole variance is concentrated only in clusters, not anywhere in between – as if all clusters will be condensed into three single red points in center of three clusters on fig.1. This situation may sometimes appear, when, say, X is a dummy variable for Toyota (company) and Y is a dummy variable for Prius (car). Since no other company produces Prius, separation of data by only variables “company – brand” would yield D=1. Of course, both of those extreme situations are very rare in practice, and ***typically D varies from 5-90% or so.***



Homogeneous mixture of two clusters (red and black). D(X) = D(Y) = D = 0

Now let’s look at the chart on fig 2. It also has three very distinct clusters, but the key difference with the structure on fig.1 is that those clusters cannot be separated by each variable alone, the red lines cross other clusters in any position. This means that D for each variable, and for all of them, will be decreased. This is a **disadvantage of Determination** criterion: even when separation of clusters is easy and obvious, as on fig.2, D may give a false signal, that clustering is bad via its low value. The **advantage** of D is that it translates the language of multi-dimensional clustering into the understandable language of the one-dimensional variables. ***A high value of D is always good, while a low value is not necessarily a bad thing***.

The determination score shown on the chart below is the average value of determination for all the variables. In this example, it is relatively high, at 76 %.



Cluster	Determination	CH Index
2	50%	100%
3	62%	75%
4	76%	75%
5	78%	57%
6	85%	55%
7	89%	53%
8	93%	52%
9	99%	56%
10	100%	53%

Standardized scores

Standardized score is determined as follows:

St. score = (Average in a cluster – Average in the whole population)/St. dev in a population (3)

Average value of standardized scores is 0, standard deviation is 1.
For **binary variables** the standard deviation is not independent of the average value:

st. dev. = (average value*(1-averagevalue))^0.5)

So, a standardized score, respectively, is a simple function of the frequency of the binary variable:

Standardized score for value 1 of the binary variable = ((1-f)/f)^0.5, (4)

where f is a frequency of the binary variable. The score for value 0, respectively, will by the same with negative sign. The table below shows the maximum possible value of a standardized score for a binary variable of a given frequency. As this illustrates you cannot use standardized scores to compare binary variables of different frequency.

Frequency	1%	2%	5%	10%	20%	30%	50%
Max. Standardized score	10	7.0	4.4	3	2	1.53	1

K-means algorithm

The K-means clustering algorithm, which is the basis of the program, was proposed in 1967 and since that time remains one of the most popular in cluster analysis, which underlines its unique features and convenience in implementation (Mirkin 2012). It exists in several versions, one of which is used in Telmar. The basic procedure was taken from (Extreme Optimization Library, 2016). The main steps are the following.

1. Make all variables normalized (i.e. create standardized scores as in (3)) in such a way, that their average is zero and standard deviation is 1.

2. Select a number of ***seeds***, K, i.e. points in multidimensional data cube, which will be the ***centroids*** to the clusters, using ***K-means++*** procedure (see below).

3. Calculate ***distances*** from each data point to those centroids. We used the traditional Euclidian distance, i.e. square root from sum of squared differences of values of the given object (respondent) and value of the centroid by each variable.

4. Assign each object to the closest cluster; for equal distances select randomly any cluster.

5. Calculate new centroids as average for all objects in a given cluster and reassign the objects until the process of assigning stops (converges).

In this process the key thing is the original selection of the seeds. In fact, as was shown in literature, the whole procedure is a way to solve an optimization problem, where the sum of squared distances from centers of the clusters is minimized. But the exact solution of the problem is NP-hard, i.e. exponentially difficult and cannot be obtained directly. Respectively, K-means is kind of approximation, and efficiency of this approximation depends on those seeds.

There are many approaches to select original centers (random, by S. Lloyd, by E. Forgy, and others), which have been used for years. But the ***K-Means ++*** algorithm (Arthur and Vassilvitskii 2007) demonstrates excellent features and is used in our system. It is described here (https://en.wikipedia.org/wiki/K-means%2B%2B):

1. Choose one center uniformly at random from among the data points.
2. For each data point x, compute D(x), the distance between x and the nearest center that has already been chosen.
3. Choose one new data point at random as a new center, using a weighted probability distribution where a point x is chosen with probability proportional to D(x)2.
4. Repeat Steps 2 and 3 until k centers have been chosen.
5. Now that the initial centers have been chosen, proceed using standard k-means clustering.

There have been many experiments in the last ten years that show it works better than many other techniques. As stated above, as the first step the data point is selected at random. It means, if one repeats the procedure again, in theory, the result will be different. In practical implementation we often come back to the same setting, and different results for the same number of clusters will be, of course, confusing. For that reason, in our system we save the seed (another seed!) for the randomly generating numbers, used to determine the very first point in K-means++. It makes the random numbers to be generated exactly as they were first time and thus yields the same original seed for the first cluster.

Calinski-Harabasz criterion for number of clusters

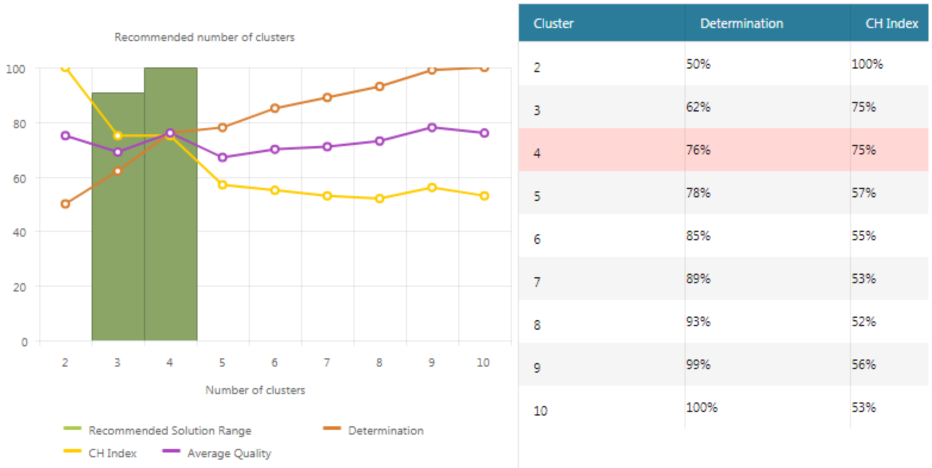
The CH index, as K-means algorithm, was also proposed many years ago (Calinski and Harabasz 1974), and according to authors themselves was in fact used for clustering purposes even earlier, in 1965, by Edwards and Cavalli-Sforza. The fact that it is in active use today also speaks about its high efficiency. The formula for this index:

$$CH(k) = B/(k-1) \,/\, W/(n-k), \qquad (5)$$

where n is number of data points, k is number of clusters; W - within cluster variation and B between cluster variation (see 18 a) for comments)

It is a form of the well-known F-criteria used for comparison of two variations with respective number of degrees of freedom, k-1 for variance between and n-k for variance within, applied in a multidimensional setting (originally F criteria is used in one dimension). The higher F (or CH, for this matter) – the more statistically different between and within variances and thus the better clustering is.

When k is growing, denominator for B is increasing (thus decreasing the first fraction) and for W – decreasing, i.e. increasing the second fraction. A total effect of larger k is for that reason decreasing the index (the smaller numerator is dividing for bigger denominator). But when k is growing, B is growing, generally, too (it reaches is maximum when k=n), and W is decreasing – the effect, we already observed when considered dynamics of determination (see 1.6), where the same B and W are at play. So, in CH, the way of calculating values represents the *punishment* (penalty) for too large number of clusters – and it makes the criterion have peaks, as observed on the same figures, where D is just increasing, as below.



Cluster	Determination	CH Index
2	50%	100%
3	62%	75%
4	76%	75%
5	78%	57%
6	85%	55%
7	89%	53%
8	93%	52%
9	99%	56%
10	100%	53%

Using CH and D together

To compensate for the drawbacks of D we use it in combination with the **Calinski-Harabasz (CH) Index.** CH Index will have a very high value for fig.2, i.e. it is less sensitive to overlapping cluster projections than D, but mostly reacts to the real structure of data. The **disadvantage** of the **CH Index** is that it does not have a theoretical maximum value (the minimum value is zero), whereas D does. In our charts the CH Index was normalized (the calculated values were divided by the CH maximum value) for ease of comparison and representation – so, it is always between 0 and 1, as D, but it has a different meaning, because maximal value of CH is not predetermined.

The **advantage** of CH Index is that it represents directly the multi-dimensional structure better than D. It makes ***CH less interpretable on a level of variables, but more useful for general judgement.*** In combination, both criteria will help a lot in determining the appropriate number of clusters.

The curve for D usually monotonically increases when the number of clusters increases. This is typical behavior when data does not have a very clear cluster structure. When clusters do have a clear structure and are easily separated by each variable, D will be also at a peak for that solution. At the same time, CH index shows where there is a best solution more often, and does it by showing a max value, equal 1 in our design.

References

B. Mirkin Clustering: A Data Recovery Approach, Chapman & Hall, 2nd Edition, 2012

ExtremeOptimizationLibrary,2016,https://www.extremeoptimization.com/Documentation/ Statistics/ Multivariate-Analysis/K-Means-Cluster-Analysis.aspx

Arthur, D.; Vassilvitskii, S. (2007). k-means++: the advantages of careful seeding *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics Philadelphia, PA, USA. pp. 1027–1035*

T. Caliński & J Harabasz (1974) A dendrite method for cluster analysis, *Communications in Statistics, 3:1, 1-27 To link to this article: http://dx.doi.org/10.1080/03610927408827101*