Rapport de travail - partie quantitative

Zhivko Taushanov

July 11, 2019

Contents

Introduction

Dataset

Two datasets are available for this study at the moment: one containing the hospitalizations in the Valais hospital and one of the emergency admissions. The former plays a central role in this project and will be used most often in this document.

The hospitalizations data set contains distinct variables, most of which are measured twice: at the admission and at the discharge from the hospital. The total number of variables is then 174. After selecting only the population if interest, i.e. individuals aged 65 or more and living at home before the hospitalization, we finally obtain a sample of 36'792 hospitalizations. All observations have been collected between 2015 and 2018.

These variables are not completely independent and may be regrouped in several groups according to the dimension they are measuring as shown in figure 1. To begin we will analyze only the condition of the individuals **entering** the hospital.

The major groups of information can be split on: somatic/physical condition, psychological condition, number of medicines, diagnose(s), interventions and information on the medical course. Furthermore the precise medications will also be investigated.

Besides these most obvious distinctions between the variables, other underlying subgroups might also be present within these groups. This will be the subject of a complementary analysis within some groups. Therefore we will verify the presence of an interpretable **clustering of the variables** within a group before clustering the individuals.

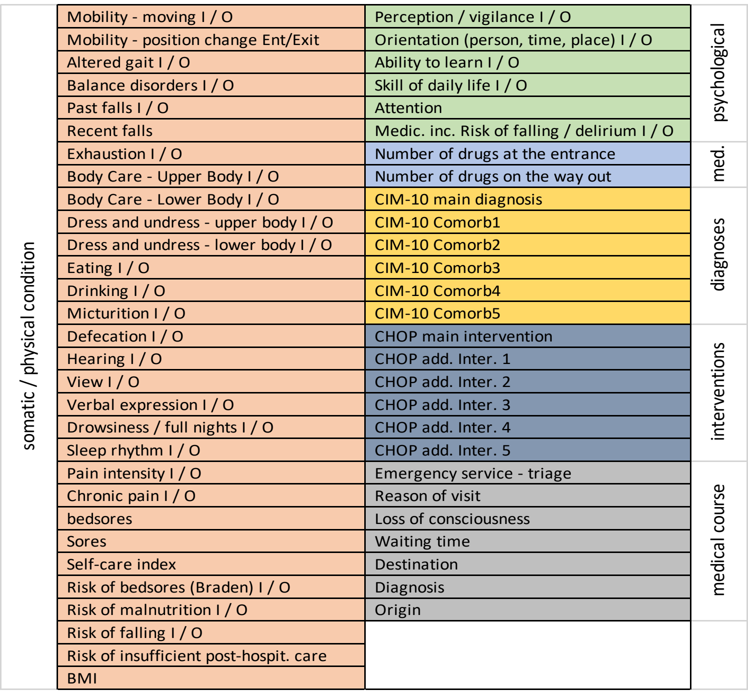
Clustering approach

The **large number of variables** in the data set makes it difficult to investigate the relations between the different factors and the risk of critical health events. Therefore the possibility to put all variables in the same model may be not an optimal choice of modeling if we consider the multi-dimensionality problem and the dependence between the variables.

An **alternative approach** will be considered in this study. Here we will make use of the important information provided by the experts in healthcare, that is the presence of clear groups within the set of **variables**.

For the cases when this grouping is not very clear, we may rely on the expert's decision. However this is not always sufficient and we also need to employ statistical methods to cluster the variables. The results of these methods will be compared to the experts opinion and will serve as a **validation tool** in order to limit a possible bias from the experts point of view or to propose a solution to an unclear relation. Both methods should pe performed independently.

A hierarchical cluster analysis using the R package "ClustOfVar" is suggested in this paper. As each statistical analysis, its result should not be accepted as they appear, but should be taken as suggestions or questions instead.



somatic / physical condition	Mobility - moving I / O	Perception / vigilance I / O	psychological
	Mobility - position change Ent/Exit	Orientation (person, time, place) I / O	
	Altered gait I / O	Ability to learn I / O	
	Balance disorders I / O	Skill of daily life I / O	
	Past falls I / O	Attention	
	Recent falls	Medic. inc. Risk of falling / delirium I / O	
	Exhaustion I / O	Number of drugs at the entrance	med.
	Body Care - Upper Body I / O	Number of drugs on the way out	
	Body Care - Lower Body I / O	CIM-10 main diagnosis	diagnoses
	Dress and undress - upper body I / O	CIM-10 Comorb1	
	Dress and undress - lower body I / O	CIM-10 Comorb2	
	Eating I / O	CIM-10 Comorb3	
	Drinking I / O	CIM-10 Comorb4	
	Micturition I / O	CIM-10 Comorb5	
	Defecation I / O	CHOP main intervention	interventions
	Hearing I / O	CHOP add. Inter. 1	
	View I / O	CHOP add. Inter. 2	
	Verbal expression I / O	CHOP add. Inter. 3	
	Drowsiness / full nights I / O	CHOP add. Inter. 4	
	Sleep rhythm I / O	CHOP add. Inter. 5	
	Pain intensity I / O	Emergency service - triage	medical course
	Chronic pain I / O	Reason of visit	
	bedsores	Loss of consciousness	
	Sores	Waiting time	
	Self-care index	Destination	
	Risk of bedsores (Braden) I / O	Diagnosis	
	Risk of malnutrition I / O	Origin	
	Risk of falling I / O		
	Risk of insufficient post-hospit. care		
	BMI		

Figure 1: Structure of the hospitalization variables

When the final set of groups is defined, we will use statistical models to cluster the **individuals** within each group. This will provide one variable from each group, that indicates the type of characteristics that the individual displayed by his answers. For example, if we separate the individuals on three groups according to their psychological indicators, we might obtain a variable indicating that a person belongs to a group with noticeable, small or no psychological issues. This type of aggregated variables will be used in the final analysis of the risk factors.

Further analyses and tests

The approach described above will also be compared to the more typical method of feature selection. A series of regression analyses and tests will follow both approaches to understand which characteristics are the most important risk factors for occurrence of critical health events such as hospitalization, early death etc.

Longitudinal perspective

A longitudinal analyses may complement the research if the data allows (to be continued when we receive the identifiers).

Chapter 1

Cluster analysis

1.1 Introduction and clustering methods

1.1.1 Methods of clustering of Mixed variables data

A large variety of clustering methods exist in the literature. However the majority are focused on either continuous or nominal data alone. There exist a limited number of techniques and strategies to incorporate both variables types in the same clustering partition (add all the formulas and references later):

- Distance measure. The idea is to be able to create a measure of the distance between individuals (or sequences) that includes nominal and continuous variables. The **Gower distance** is the most used such measure and is defined as: (formulas)

  However because it uses the range of continuous variables to determine the distance and assumes that nominal variables have a distance of either 0 or 1, it may under-estimate the impact of the continuous variables (which reaches 1 much less often than in the nominal variables case). Furthermore, the weights are also arbitrarily selected, however they define the contribution of each data type to the global distance (see **??** for more detailed examples). As all measure distances, Gower should be used as input for clustering methods, such as k-means for instance, to provide clustering results.

- k-means is another algorithm mainly used for continuous variables. Several other implementations, such as the R package KAMILA, integrate different types of variables together. In this particular case, it uses the probabilities of a multinomial distribution for the discrete variables. The continuous variables distribution is estimated by univariate Kernel Densities. The probabilities resulting from the both distribution types are added together to obtain a measure of how close an observation is to the center of each cluster. (formulas)

- k-medoids is a more robust version of k-means. The difference is that in k-medoids a real data points are selected as centers of the clusters, whereas in k-means the centers are the computed averages. The R package PAM is a popular implementation of this approach.

- Normal-Multivariate mixture models are another although a bit more complex but very flexible and useful alternative (to detail with formulas)

- The standard method for clustering of factor variables is the **Multiple Correspondence Analysis (MCA)**. This model is implemented in the R packages ”FactoMineR” and ”PCAmix”. It splits all factors into multiple binary variables. Usually the principle components obtained by MCA are then clustered by a **kmeans** algorithm. (details and formulas)

In our analysis we tried several different clustering methods. However in the displayed results we most often used the following procedure to cluster the variables:

1. Typically one factor analysis type of model is used (such as MCA, PCA, or other depending on the data type).
2. Then the most important factors are selected. In this case we prefer to select larger number of components if it is necessary in order to keep larger part of the variation of the data. We keep in mind that our aim in this stage is to obtain an accurate clustering, rather than to reduce the dimensionality (this will be done using the final cluster partition).
3. At the end these factors are considered as variables and serve as input of an k-means clustering algorithm.
4. The number of clusters is then selected using the Silhouette statistic, but also by considering the interpretability of the resulting partition.

1.2 Psychological variables (green)

1.2.1 Data overview and strategies

All the **six psychological variables are ordinal**. However, together with many other variables in the data set, most often we will consider them as nominal in our analysis, because of the small number of modalities of each of these variables.

Some observations are excluded from the analysis because they contained only missing values. These are the first subjects in the data set and they have also been excluded from other analyses for the same reason.

The final sample for the following analyses contains 32'484 observations

1.2.2 Clustering of psychological variables

A hierarchical clustering method has been performed on the psychological variables in order to investigate any possible relation and presence of subgroups within these variables. The R package "ClustOfVar" has been used for this purpose.

The results do not suggest any clear interpretable structure within as illustrated by the dendrogram in figure 1.1. They indicate that only single variables clusters (singletons) may be separated one at a time to form separate and not very distinct clusters. This information does not provide any useful solution to our problem because obviously it does not make sense to cluster the individuals over one single variable. Therefore this result, combined with the small total number of variables (only 6), lead us to the conclusion that the six psychological variables should be considered together in the same individual clustering algorithm.

1.2.3 Clustering of individuals

Multiple correspondence analysis has been used to cluster the individuals according to their psychological state because all variables are categorical. Even though the first two principal components do not explain large part of the data (26%), we can observe the four most discriminant variables for the clustering (and the importance of their categories) on figure 1.2.

For further analysis we choose rather large number of principal components (9) because of the relatively low explanatory power (65% of the variance). After that we examined several different clustering partitions with respect to the number of clusters. Some particular groups and features can be systematically found in all the partitions. This allows us to make the following generalizations of the results, regardless the number of clusters:

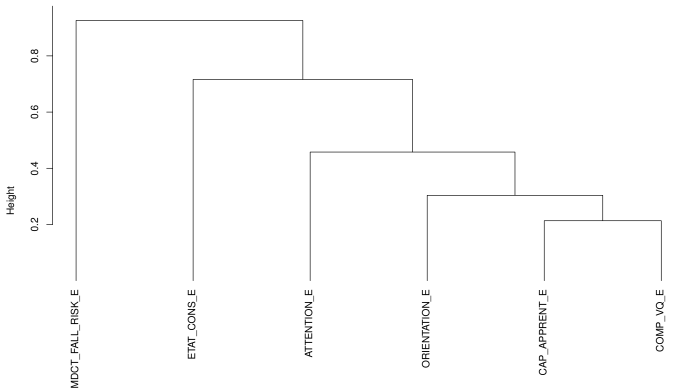


Figure 1.1: Dendrogram of psychological variables

- The majority of valid observations are displaying good condition in almost all of the variables. They are found in every clustering solution and form always the largest cluter.
- When increasing the number of clusters, the observations with average or ”bad” psychological condition are split and nuanced.
- One group of individuals with predominantly missing values have been excluded from the analysis.

The optimal number of clusters is determined here by the silhouette statistic on figure 1.3. This statistic measures how similar each observation is to its own cluster, compared to all other clusters. The results indicates that two or four clusters solution would be the most appropriate in terms of within and between cluster distances. These two solutions will be resumed in this section.

Two cluster soution

The two cluster solution is made of one dominant group of 29913 ”healthy” people and one small group of impatients in average and bad condition. On table 1.1 we observe that the two clusters are differently distributed over all 6 variables and the diagnoses (CIM). These differences are also highly significant. It is interesting to mention that much smaller part of the ”healthy” group has taken medications increasing the risk of falling or delirium, 15% vs 44% of group 2.

Two other variables (number of medications and primary diagnostic) are added to the analysis for sake of exploration. They do not participate in the clustering model. No difference is observed in the average number of medications, however the primary diagnosis appear to be different among the groups.

Four cluster solution

In the four clustering solution, the results are similar, except that we do not have a single ”unhealthy” group, but three clusters with different degree of health issues.

(**INCLUDE THE TABLE FOR 4 GROUPS**)

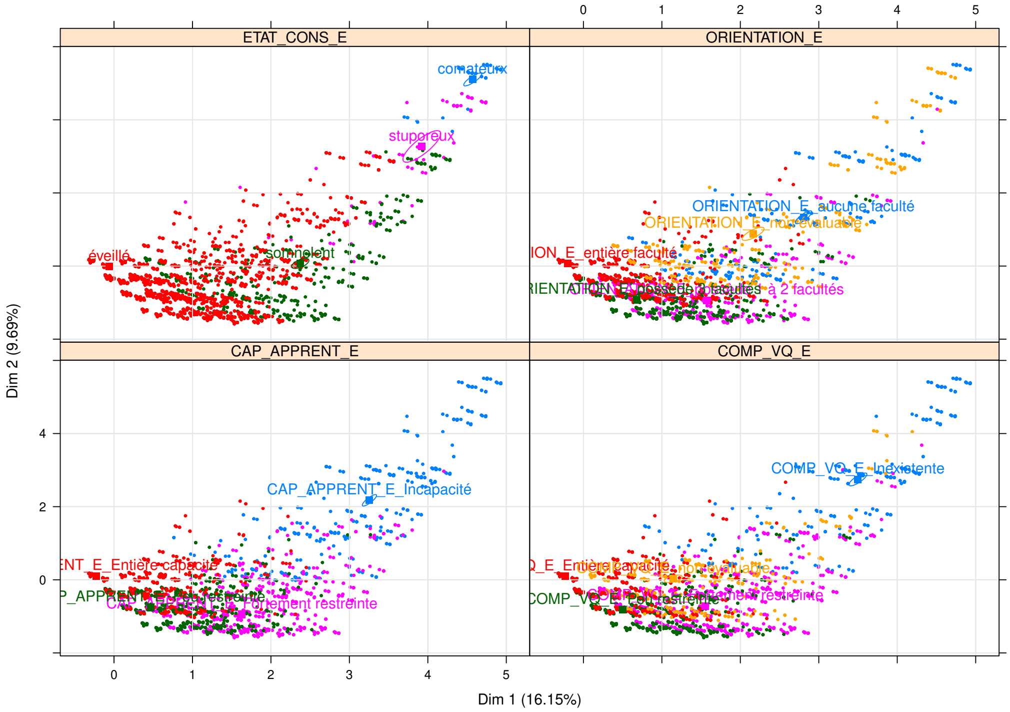


Figure 1.2: Dendrogram of psychological variables

1.3 Somatic/physiologic variables (orange)

1.3.1 Data overview and strategies

Note that several variables have modalities that do not correspond to these described in the list (see the variable description document "summaries age domicile"). These modalities have bean corrected but in an arbitrary manner. Therefore a discussion over all such corrections is necessary.

At least two of the variables from the list should be considered as continuous in this group (Braden risk of sores and risk of falling, probably *"Indice d'autosoins" and "risque de déficit de soins post-hospitalisation" may be also continuous*), therefore we dispose with **mixed data**, and will apply the corresponding model. Both continuous variables are finally present in the second sub-group.

1.3.2 Clustering of variables

The number of somatic variables is relatively large to perform a direct clustering on the individuals. Furthermore, the possible presence of similarities between the variables indicate that we must consider a split of these variables in multiple sub-groups.

The initial separation of the variables has been done according to the experts knowledge of the data. However the results from a statistical model for variable clustering have also been used in order to provide an external validation of the experts point of view. These results are summarized

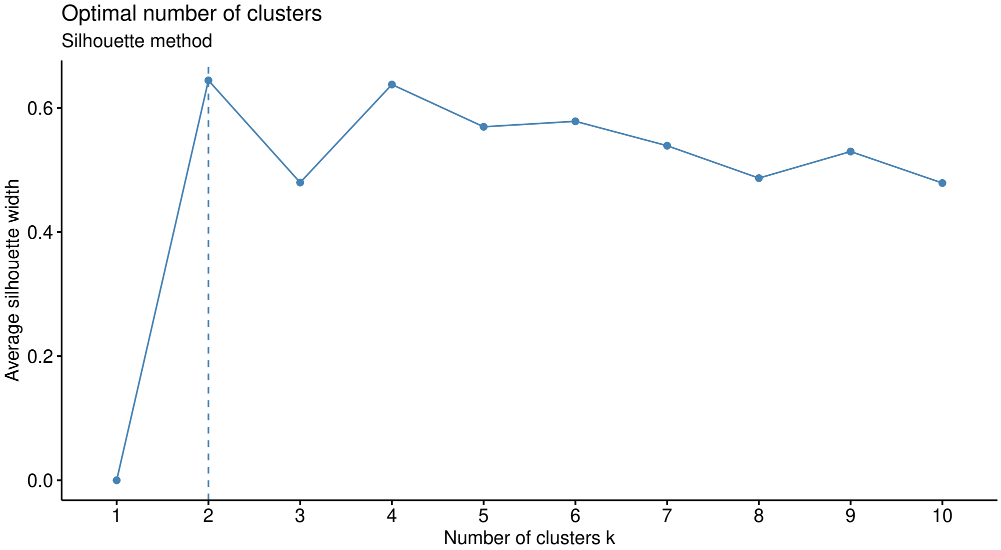


Figure 1.3: Silhouette statistic for choosing the number of clusters: two or four-cluster solution is suggested.

on figure 1.4. Even though they do not completely match the experts partition, we can observe that many of the variables can be found in the same cluster.

Initially four groups were formed: mobility, health difficulties, support for the daily life activities and other health risks.

As stated before, four groups of variable separation was the initial guess. However on table 1.3 we see that 3 of the variables in the last group ”other health risks” present an excessive number of missing values: bedsores, wounds and malnutrition risk.

First, this could be a reason for unreliable results from the variable clustering for these variables, which is a reason to ignore their place in the analysis presented earlier on figure 1.4. But most importantly, it is also a burden for any further clustering of the observations if we keep these variables. Therefore the only solution is to take them out of the analysis.

The two other variables from the group: Braden risk and risk of falling are not sufficient to create an entire group of clustering. Therefore they are attached to the group ”health difficulties” for the clustering of individuals. This leads to the following final three sub-groups of physiological/somatic variables displayed on table 1.4.

1.3.3 Clustering of individuals within the physiologic sub-groups

In this section, we will present the results of the 3 separate cluster partitions, one for each of the above-mentioned sub-groups.

Mobility (sub-group 1)

The optimal number of clusters n is unclear according to the silhouette statistic. It suggest rather similar and increasing values as n increases. Therefore we chose two cluster partition because this

Consciousness	comateurx	stuporeux	somnolent	éveillé		total
group 1	0.00	0.00	0.00	1.00		29913
group 2	0.02	0.03	0.19	0.76		2571
Orientation	aucune faculté	1 à 2 facultés	3 facultés	entière faculté	not measurable	
group 1	0.0	0.01	0.05	0.93	0.00	29913
group 2	0.2	0.34	0.20	0.12	0.14	2571
Learning capacity	Incapacity	severely reduced	slightly reduced	Full capacity		
group 1	0.00	0.01	0.09	0.90		29913
group 2	0.22	0.60	0.12	0.05		2571
Daily life skills	Inexistant	severely reduced	slightly reduced	Full capacity	not measurable	
group 1	0.00	0.01	0.08	0.90	0.01	29913
group 2	0.15	0.57	0.16	0.06	0.07	2571
Attention	perm. reduced	occas. reduced	not affected	not measurable		
group 1	0.01	0	0.98	0.01		29913
group 2	0.61	0	0.30	0.09		2571
Mdc incr. fall risk	yes	no				
group 1	0.15	0.85				29913
group 2	0.44	0.56				2571
Additional variables	(not included)					
Nbr of medications	0	1-3	4-5	6-9	10+	
group 1	0.57	0.12	0.09	0.13	0.09	29913
group 2	0.65	0.04	0.06	0.13	0.12	2571
	mean for gr.1	mean for gr.2				
	2.809748	2.846752				
CodeCim1 REC1	other	cancer	mental	sensory	systemes	
group 1	0.39	0.01	0.13	0.03	0.44	29913
group 2	0.32	0.01	0.08	0.04	0.55	2571

Table 1.1: Two clustering solution: distribution of the groups in all six psychological variables. All distributions are significantly different among clusters (χ^2-tests, p-values<0.01), except the mean number of medications.

Mobility	**Health difficulties**	**Daily life activ. support**	**Other health risks**
Movement	Exhaustion	Body care - upper body	Sores
Changing position	Hearing	Body care - lower body	Wounds
Altered gait	View	Dress and undress - upper b.	Malnutrition risk
Balance disorders	Verbal expression	Dress and undress - lower b.	Risk of falling
Past falls	Drowsiness Full night	Eating	Braden risk (of sores)
Recent falls	Sleep rithm	Drinking	
	Pain intensity	Micturition	
	Chronic pain	Defecation	

Table 1.2: Initial idea for sub-goups of physiological/somatic variables

variable	bedsores	wounds	Braden risk	malnutrition risk	risk of falling
missing values	98.6%	93.6%	0.3%	87.7%	44.9%

Table 1.3: Percentage of missing values in sub-group ”other health risks”

is also the best separation in terms of interpretability of the results and implies a clear difference between the groups.

Again in table 1.5 we see that roughly $\frac{2}{3}$ of the subjects have little or no mobility issues (group 2). The remaining individuals exhibit problems in at least one of the 6 dimensions. That number is rather large but not surprising if we consider the advanced age of the selected population.

The χ^2-tests confirm the clear difference between the groups among all variables.

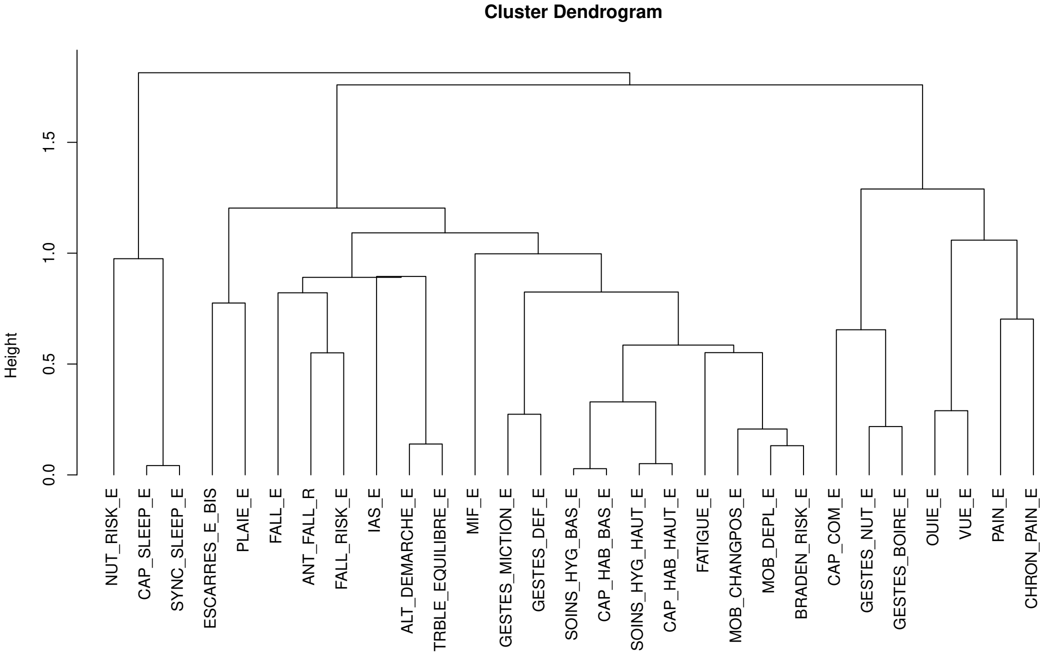


Figure 1.4: Dendrogram of physiological/somatic variables

Mobility	**Health difficulties**	**Daily life activities support**
Movement	Exhaustion	Body care - upper body
Changing position	Hearing	Body care - lower body
Altered gait	View	Dress and undress - upper body
Balance disorders	Verbal expression	Dress and undress - lower body
Past falls	Drowsiness Full night	Eating
Recent falls	Sleep rithm	Drinking
	Pain intensity	Micturition
	Chronic pain	Defecation
	Braden risk (of sores)	
	Risk of falling	

Table 1.4: Final sub-goups of physiological/somatic variables

Health difficulties (sub-group 2)

The objective of our analysis is clustering and not dimension reduction. Therefore it is worth taking into account larger number of principal components in the analysis in order to explain larger part of the variability of the data.

Movement	Incapacity	severely reduced	slightly reduced	full capacity	total
group 1	0.23	0.37	0.34	0.06	11328
group 2	0.02	0.01	0.16	0.82	21172
Changing position	Incapacity	severely reduced	slightly reduced	full capacity	
group 1	0.08	0.29	0.40	0.23	11329
group 2	0.00	0.00	0.05	0.95	21174
Altered gait	yes	no	not measurable		
group 1	0.56	0.09	0.35		11331
group 2	0.10	0.90	0.01		21172
Balance disorders	yes	no	not measurable		
group 1	0.42	0.21	0.37		11330
group 2	0.06	0.94	0.00		21172
Past falls	yes	no	not measurable		
group 1	0.33	0.59	0.08		11329
group 2	0.05	0.95	0.01		21170
Recent falls	yes	no			
group 1	0.11	0.89			9288
group 2	0.01	0.99			12925

Table 1.5: Two clustering solution of the ”mobility” subgroup. All distributions are significantly different among clusters (χ^2-tests, p-values<0.01).

The silhouette statistic suggests 2, 8 or 10 clusters . Our decision is to choose 2 cluster solution for two reasons, first it corresponds to the first and most pronounces peak in the graph 1.5, but it is also more easy to interpret. .

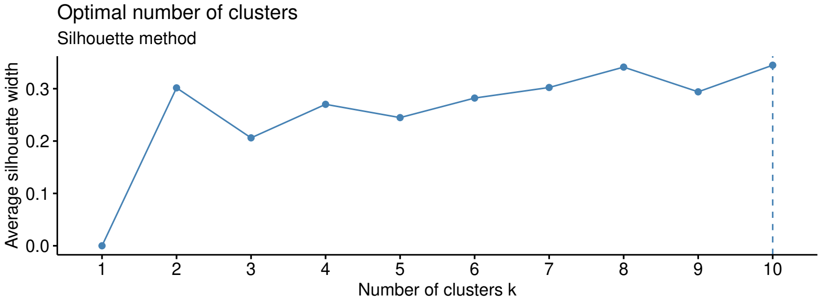


Figure 1.5: Sub-group ”Health difficulties”: silhouette statistic for choosing the number of clusters. Two or four-cluster solution is suggested.

Before adding the two continuous variables to this sub-group, a **three** cluster solution was the optimal solution, despite the excessively small size of one of the groups ($n_3 = 241$). However this group is the more distinct from the rest. It comprises impatient that were probably unconscious or in very bad condition. Concerning the two other large groups, the only clearly distinctive feature between them was the higher proportion belonging to the group ”not measurable” of the variables and therefore they could be merged together.

After adding both continuous variables to the analysis, we observe on figure 1.6 that both solutions are rather similar. The main difference is due to the rather large categories ”not measurable” in the variables Drowsiness and Sleep rhythm.

A possible solution to this problem is to **take these variables out of the analysis** and perform a new clustering. Note that both variables are not measurable for the same individuals, which biases the result of the clustering.

Exhaustion	no activitiy possible	some auton.	occas. act.	good phys./	not meas.		total
		a. w. recovery	possible	mental strenght			
group 1	0.01	0.12	0.21	0.65	0.00		24034
group 2	0.04	0.18	0.25	0.50	0.03		8458
Hearing	deafness	auditive	no auditive	not meas.			
		problems	problems				
group 1	0	0.1	0.90	0.00			24031
group 2	0	0.1	0.87	0.02			8460
View	blindness	visual	no visual	not meas.			
		problems	problems				
group 1	0	0.07	0.93	0.00			24032
group 2	0	0.08	0.88	0.03			8460
Verbal expression	Incapcity	Restricted	entire capacity				
group 1	0.00	0.03	0.96				24030
group 2	0.02	0.07	0.91				8461
Drowsiness	disturbed	no disturbation	not measurable				
group 1	0.15	0.84	0.01				24029
group 2	0.02	0.01	0.97				8459
Sleep rithm	modified	not modified	not measurable				
group 1	0.06	0.94	0.00				24025
group 2	0.02	0.02	0.96				8455
Pain intensity	Signs of pain	improbable	intense	meduim	slight	no pain	
	(3-d p.)	(3-d p.)	pain	pain	pain		
group 1	0	0	0.03	0.11	0.17	0.69	24017
group 2	0	0	0.03	0.11	0.17	0.69	8460
Chronic pain	yes	no	not meas.				
group 1	0.08	0.92	0.00				23998
group 2	0.07	0.87	0.05				8457
Continuous varibles							
Braden risk sores							
Welch 2 s. t-test:	mean gr.1	mean gr.2	95% conf. int.				
	21.1	19.9	(1.08; 1.23)				
Risk of falling							
Welch 2 s. t-test:	mean gr.1	mean gr.2	95% conf. int.				
	2.11	2.39	(-0.33;-0.24)				

Table 1.6: Two clustering solution of the "Health difficulties" subgroup. Nominal and continuous variables results. All distributions are significantly different among clusters (χ^2-tests, p-values<0.01).

The continuous variables have also a significant difference, but it is not a sufficient reason in terms of interpretability to keep this solution.

Daily life activities support (sub-group 3)

The Silhouette statistic is indecisive on figure 1.6, but the two cluster solution appears more appropriate and is our choice.

A brief look on the clusters in figure 1.7 is sufficient to spot the difference between groups. One large cluster of 27'233 observations is formed by mainly healthy individuals that have their full capacity on the majority of the variables. The smaller cluster 1 of 5'268 observations regroups the individuals who have at least one serious problem with their daily life activities. Overall the separation appears interesting for our aim of separating the observations. Once again the distributions of the clusters are significantly different over all variables.

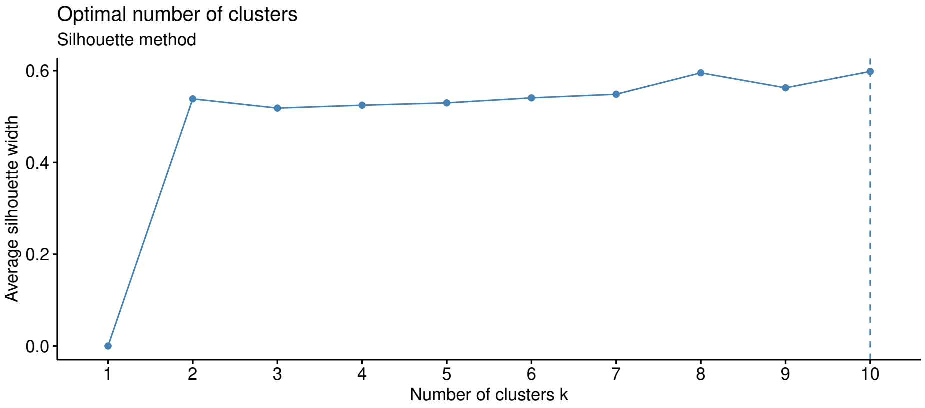


Figure 1.6: Sub-group 3 ”Daily life activities support”: silhouette statistic for choosing the number of clusters. Two, eight or ten-cluster solution is suggested. Two groups are chosen for sake of simplicity.

Body care - upper b.	incapacity	severely reduced	slightly reduced	full capacity	total
group 1	0.22	0.52	0.23	0.04	5268
group 2	0.00	0.00	0.23	0.76	27233
Body care - lower b.	incapacity	severely reduced	slightly reduced	full capacity	total
group 1	0.53	0.45	0.02	0.00	5268
group 2	0.01	0.09	0.25	0.65	27233
Dress and undress - upper	incapacity	severely reduced	slightly reduced	full capacity	total
group 1	0.26	0.50	0.21	0.03	5268
group 2	0.00	0.01	0.22	0.78	27234
Dress and undress - lower	incapacity	severely reduced	slightly reduced	full capacity	total
group 1	0.53	0.45	0.02	0.00	5268
group 2	0.01	0.08	0.24	0.67	27233
Eating	incapacity	severely reduced	slightly reduced	full capacity	total
group 1	0.13	0.13	0.29	0.45	5268
group 2	0.01	0.00	0.02	0.97	27232
Drinking	incapacity	severely reduced	slightly reduced	full capacity	total
group 1	0.09	0.08	0.18	0.65	5268
group 2	0.01	0.00	0.00	0.99	27229
Micturition	incapacity	severely reduced	slightly reduced	full capacity	total
group 1	0.31	0.24	0.21	0.23	5267
group 2	0.04	0.01	0.08	0.88	27224
Defecation	incapacity	severely reduced	slightly reduced	full capacity	total
group 1	0.17	0.28	0.19	0.36	5267
group 2	0.00	0.00	0.06	0.94	27227

Table 1.7: Two clustering solution of the ”Daily life activities support” subgroup. All distributions are significantly different among clusters (χ^2-tests, p-values<0.01).

Bibliography

[1] A semiparametric method for clustering mixed data. Foss. *Machine Learning*. 2016