PATTERNING THE IMPOSSIBLE

Carolina Bellani (M20170098), Gonçalo Passão (M20170450) and Sofia Jerónimo (M20170070)
Nova Information Management School, Universidade Nova de Lisboa, Campus de Campolide, 1070-312 Lisboa, Portugal

KEYWORDS: Absenteeism, PCA, MCA, Clustering

Introduction

Absenteeism can be considered as absence of an employee in the workplace during a short (for being late) or long period of time.

The absenteeism can happen for various reasons that can be related with the work environment and/or organization. Concerning the environment, most of the reasons are associated with the noise, inappropriate lightening, extreme temperatures, vibration and/or hygiene. Associated with the work organization, it can be linked to the task, time schedule, overload or underload work, participation and envolvement in the tasks, interpersonal relationships, work rhythm and pressure, impossibility to get a promotion, lack of a work plan and salary (1).

It is extremely important to identify and study deeply the reason(s) that are causing the absenteeism inside a company. Therefore, identify **possible patterns of the employees and work conditions** can help understand which are the problems inside the company and consequently, improve the quality of life at work and productivity of the employees.

Objectives

We studied absenteeism of 34 employees working in a brazilian courier company.

The absenteeism was registered throughout 3 consecutive years (July 2007 to July 2010).

The main objectives were to understand the **pattern of the absenteeism** concerning the characteristics of the employees and households, the workload, number of hours absent and frequency, distance and costs of transportation until work, and disciplinary failure frequency. We expect that the most absent people would be the ones having more health problems, higher number of household and pets, overload or underload work and with higher failures.

Methods

We used the dataset 'Absenteeism at work' provided from UCI Machine Learning repository.

All data preparation and statistical analysis were performed in RStudio (2), please check the QR code to have a more detailed information about our analysis.

We have 667 observations concerning each absent day. For each observation, we have the ID of the employee, the reason for absence (originally 27 reasons but rearranged into 1-10 categories), the absent time and average workload (both in hours), % hit target, week day, month and season. For each ID, we have the following personal information: transportation distance (in km) and expenses (in reais), service time in years, age, education (high school, graduate and postgraduate), number of children (0-4), number of pets (0-8), weight (in Kg), height (in cm) and Body Mass Index (in Kg/m^2).

We created new features concerning the frequency of observations per ID for absenteeism (2-112) and for disciplinary failure (0-6). We also created the feature First start which concerns the age that each employee started to work in the company.

We identified and removed outliers (4,1%). Afterwards, we checked Pearson and Spearman correlation coefficients for the quantitative features and to test dependency between qualitative features, we did Pearson Chi-square test. We created a new data set only with the Unique ID.

To study the **pattern of the absenteeism**, we did a **multiple correspondence analysis** (MCA) and a **clustering analysis** using **principal components analysis** (PCA). We tested if our data was appropriate for factor analysis, but the overall MSA was 0.51. The performed clustering techniques were hierarchical (Ward, Average, Single, Complete and Centroid linkages) and partitional (K-means and K-medoids) with several numbers of clusters. The chosen distances were Spearman and Kendal.

Results

PCA

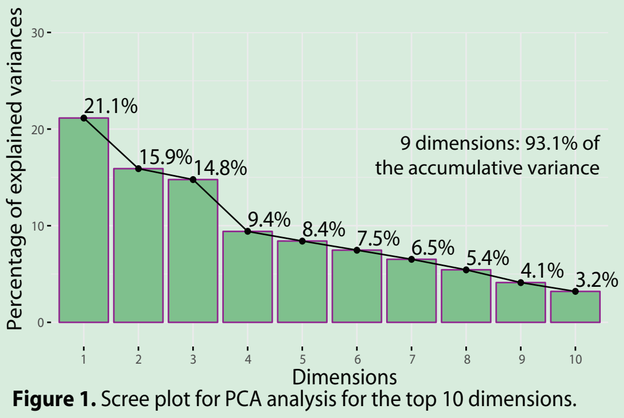

Figure 1. Scree plot for PCA analysis for the top 10 dimensions.

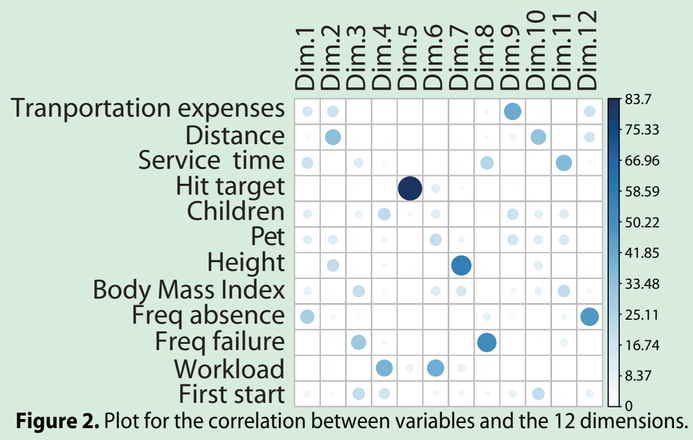

Figure 2. Plot for the correlation between variables and the 12 dimensions.

PCA clustering

K-medoids with Spearman Distance of the PC

Table 1. Centroids of the 15 variables for the 3 clusters and the significance test.

Variables	Cluster 1	Cluster 2	Cluster 3	P-value
Transportation expense	252.72	195.40	209.71	<2e-16
Distance	39.03	18.41	32.65	<2e-16
Service time	13.26	11.95	12.19	0.00317
Age	36.16	37.40	34.46	8.21e-06
Hit target	94.30	95.63	95.12	9.46e-06
Children	1.37	0.71	0.87	2.78e-11
Pet	0.89	0.27	1.14	1.63e-12
Weight	82.33	77.16	76.01	1.01e-07
Height	170.15	174.75	171.25	<2e-16
Absent hours	5.78	5.53	5.43	0.833
Body mass index	28.45	25.21	25.90	<2e-16
Freq.absence	54.38	29.46	60.37	<2e-16
Freq.failure	1.79	0.85	1.32	1.26e-12
Workload	4.16	4.33	5.17	<2e-16
First start	22.89	25.45	22.28	3.21e-14

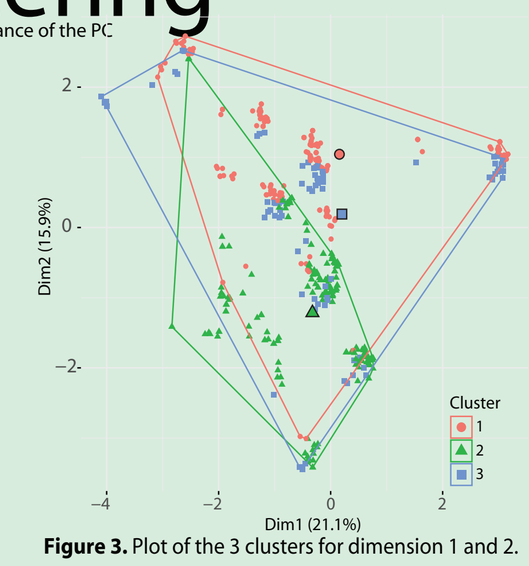

Figure 3. Plot of the 3 clusters for dimension 1 and 2.

Table 2. Frequency of each reason for the 3 clusters.

	Variables	Cluster 1	Cluster 2	Cluster 3
Reasons	**Accompanying person**	21	7	10
	Dental consultation	46	31	30
	Diagnosis, donation and vaccination	10	21	9
	Diseases	42	78	55
	Injury, poisoning	16	10	8
	Medical consultation	52	51	44
	Physiotheraphy	23	14	31
	Pregnancy, childbirth, perinatal complications	1	4	1
	Symptons and abnormal exams	7	9	4
	Unjustified	20	10	2
	Total	**238**	**235**	**194**

Table 3. Frequency of the absenteeism hour for the 3 clusters.

	Variables	Cluster 1	Cluster 2	Cluster 3
Absent time	**1 hour**	25	30	31
	2 hours	35	64	58
	3-7 hours	77	54	45
	>8 hours	101	88	60
	Total	**238**	**235**	**194**

MCA

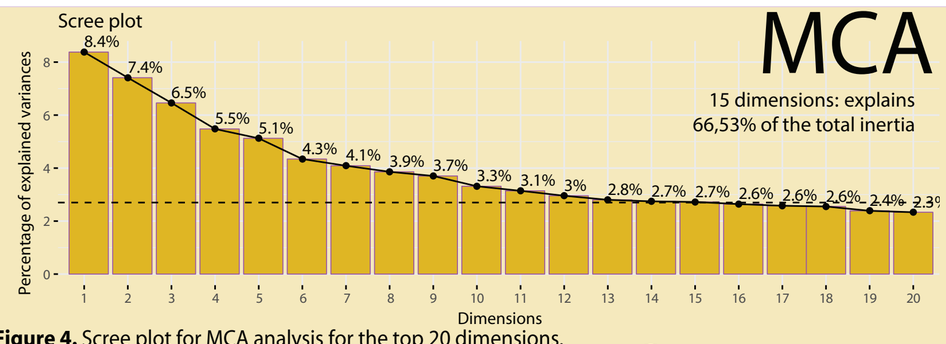

Figure 4. Scree plot for MCA analysis for the top 20 dimensions.

MCA factor map

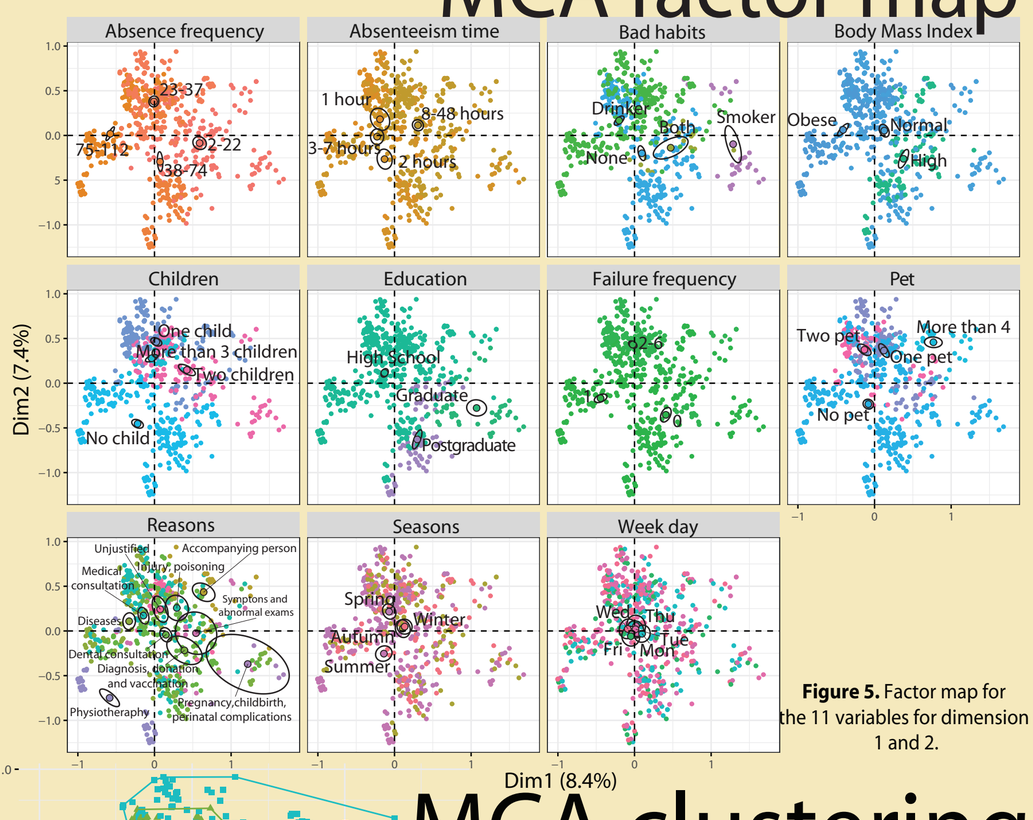

Figure 5. Factor map for the 11 variables for dimension 1 and 2.

MCA clustering

Hierarchical clustering

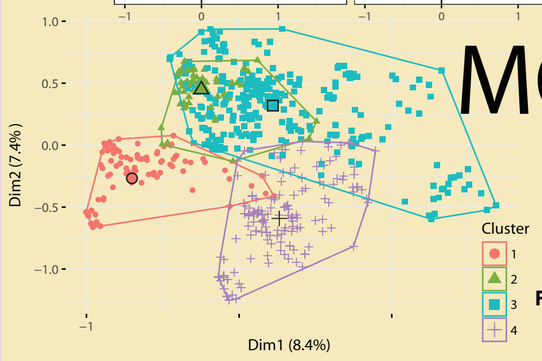

Figure 6. Plot of the 4 clusters for dimension 1 and 2.

Table 4. Frequency of the absenteeism hour for the 4 clusters.

	Variables	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Absent time	**1 hour**	21	13	34	18
	2 hours	34	26	47	50
	3-7 hours	49	26	61	39
	>8 hours	24	12	158	55
	Total	**128**	**77**	**300**	**162**

Table 5. Frequency of each reason for the 4 clusters.

	Variables	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Reasons	**Accompanying person**	0	1	32	5
	Dental consultation	39	10	45	13
	Diagnosis, donation and vaccination	5	6	18	11
	Diseases	23	16	79	57
	Injury, poisoning	2	3	25	4
	Medical consultation	18	36	60	33
	Physiotheraphy	38	4	0	26
	Pregnancy, childbirth, perinatal complications	0	0	5	1
	Symptons and abnormal exams	2	1	12	5
	Unjustified	1	0	24	7
	Total	**128**	**77**	**300**	**162**

Table 6. Centroids of the 15 variables for the 4 clusters

Variables	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Transportation expense	186.52	224.90	248.57	191.28
Distance	47.22	26.36	29.88	17.99
Service time	17.14	9.36	12.62	10.05
Age	37.79	28.40	38.47	34.04
Hit target	95.46	94.99	94.76	95.11
Children	0.00	1.01	1.85	0.19
Pet	0.00	2.00	1.01	0.23
Weight	85.73	69.51	79.75	75.44
Height	170.69	168.99	172.92	173.14
Absenteeism time	4.26	3.29	7.14	4.86
Body mass index	29.48	24.34	26.67	25.12
Freq.absence	96.88	72.35	26.18	35.50
Freq.failure	1.00	1.92	1.84	0.33
Workload	4.34	4.72	4.56	4.47

Conclusions

The lack of variability of the variables and/or the reduce number of employees used in this data sample do **not seem enough to describe properly the pattern of absenteeism** in terms of number of absenteeism hours and reasons of absence. A sample of different measures of **work environment,** for instance, survey concerning the noise, light or hygiene and/or associated with the **work organization** such as salary amount, last promotion date and survey about the relations inside the company could bring useful insights about this subject.

References

(1) Garcia, F. C. ; Silva, M. F. G. (2009). Causas do Absenteísmo nas Organizações: um estudo de caso em Unidades de Alimentação e Nutrição, FEA/USP, São Paulo

(2) RStudio Team (2016). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/.

Acknowledgements

We would like to thank the teacher Dr. Jorge Mendes for helping us and answering all questions throughout this project.

Script

NOVA IMS Information Management School