CLASSIFICATION AND IDENTIFICATION OF THE CAUSES OF ABSENTEEISM IN A PUBLIC TRANSPORT COMPANY USING CLUSTER ANALYSIS AND PRINCIPAL COMPONENTS TECHNIQUES

Laryssa Ribeiro CALCAGNOTO[1]
Tiago Viana Flor SANTANA[2]
Rodrigo Rosseto PESCIM[2]

■ ABSTRACT: Absenteeism is the practice or custom of an employee to be absent from workplace. Its causes are diverse and may affect the workers income as well as to cause operational disruption, stress the administration and also financial losses for the company. Cluster analysis is a multivariate tool that can be used to determine groups in the sense that each group has its own characteristics in terms of the observed variables. In this sense, that technique can be used as a support to show which characteristics may contribute to absenteeism. We use the Ward hierarchical algorithm to build the clusters and to compare the groups the Kruskal-Wallis nonparametric test is adopted. Finally, a study on the strength of association among the variables is developed using Spearman's correlation and for the relationship among those variables related to absence and social aspects, we use the principal component analysis. Moreover, the study indicates the possibility to determine three heterogeneous groups in the company and to show characteristics in those groups which are potential factors that cause absenteeism to a greater or lower extent.

■ KEYWORDS: Absenteeism, Cluster analysis, Kruskal-Wallis test, Principal component analysis, Spearman's Correlation

[1]Universidade Estadual de Londrina - UEL, Caixa Postal 10.011, CEP: 86.057-970, Londrina, PR, Brasil. E-mail: *laryssacalcagnoto@gmail.com*
[2]Universidade Estadual de Londrina - UEL, Departamento de Estatística, Caixa Postal 10.011, CEP: 86.057-970, Londrina, PR, Brasil. E-mail: *tiagodesantana@uel.br; rrpescim@uel.br*

1 Introduction

Absenteeism is a term used to link the absence and the delay of an employee at work. Several studies have appeared in recent years to explain the reasons for that practice which causes harm to companies. It is well-known that their reasons are diverse such as medical consultations or those dependents, marriage, vacation and even low motivation to work as explained Penatti *et al.* (2006).

Some absences are expected by the companiesand it can be treated in advance such as vacations and licenses. However the vast majority of abstentions, the companies are not informed until the day in question. In this context, the absenteeism is considering an old problem and very well-known to the company administrators and its causes are relative and depend on a lot on the context experienced by each company. In general, absenteeism affects workers income, causes operational disruption, stresses management and causes financial losses for the company (ALMEIDA and NASCIMENTO, 2015).

The aim of this paper is to identify and understand the causes generated by absenteeism within a transport company from the north of state of Paraná in order to prevent the possible damages. We can note that the characterization of absenteeism and is not unanimous. For Calais and Zanelatto (2011) absenteeism is directly related to the stress of the employee which may be caused by the company or not. In the other hand, Penatti *et al.* (2006) apud Lee and Eriksen summarizes absenteeism as inversely proportional to job satisfaction, that is

$$\text{Absenteeism} = \frac{1}{\text{Satisfaction}}. \tag{1}$$

According to Penatti *et al.* (2006), an index was developed to control the abstentions by adding the periods of absence and delays of the employee during the workday, called absenteeism index and defined as

$$\text{Index} = \frac{\text{Certified absences} \ + \ \text{Not attested absences} \ + \ \cdots \ + \ \text{absence variables}}{\text{Working time}}. \tag{2}$$

Calais and Zanelatto (2011) and Penatti *et al.* (2006) showed that the absenteeism index is associated directly to the human resources area and to minimize it, they suggest implementations of programs such as workplace improvements and safety programs. For the transport company, the loss is associated to the relocation of staff since the absenteeism is hardly reported in advance. That relocation of employees has been provoked an economic loss and operational for the company.

The remainder of the paper is outlined as follows: In section 2, a review of the cluster and principal component analysis is provided. We introduce a brief description of the transport company data set in Section 3. The results given in Section 4 reveal the usefulness of the multivariate techniques for analyzing real data. Concluding remarks are addressed in Section 5.

2 Multivariate techniques

Cluster analysis is a multivariate statistical tool used for construction and classification of groups according to the features of each observation in order to obtain heterogeneous groups among themselves and the observations within each group are considering homogeneous, as exposed by Chatfield and Collins (2013); Mingoti (2005) and Johnson and Wichern (2007).

The techniques for building the clusters are divided into two types: hierarchical and non-hierarchical. The hierarchical techniques are agglomerative or divisive type. The agglomeratives start from single observations and mergers are performed up to all observations are in the same group while the divisive type does the opposite process. These techniques intend to determine k groups with remote characteristics (MINGOTI, 2005).

On the other hand, the non-hierarchical technique assumes that the number k of groups is already determined, for example, in the k-means method (FERREIRA, 2008) that divides the observations into k groups and then uses an algorithm to propose in which group the observation belongs to.

For the formation of groups, the Ward agglomerative hierarchical algorithm is presented by Mingoti (2005) and Johnson and Wichern (2007). We can note that in the first stage each observation is considered a cluster of unit size, totaling n groups, and in the end process there is only a single cluster for all the observations. The number of groups desired is described by g $(1 < g < n)$ and represents the natural division of the observations. For each step, the Ward algorithm combines the two clusters that result in the smallest value of

$$SSR = \sum_{i=1}^{g_k} SS_i,$$

where g_k is the number of groups in that step k, $SS_i = \sum_{j=1}^{n_i} \left(\boldsymbol{X}_{ij} - \bar{\boldsymbol{X}}_{i\cdot}\right)^T \left(\boldsymbol{X}_{ij} - \bar{\boldsymbol{X}}_{i\cdot}\right)$ is the sum of squared for the ith cluster in step k, $\boldsymbol{X}_{ij}$ is the jth observation for the ith cluster and $\overline{\boldsymbol{X}}_{i\cdot}$ is the mean of the ith *cluster*. In each step of the Ward algorithm the behavior of the merge level is studied and the number

$$d_{q,r} = \left[\frac{n_q n_r}{n_q + n_r}\right] \left(\bar{\boldsymbol{X}}_{q\cdot} - \bar{\boldsymbol{X}}_{r\cdot}\right)^T \left(\bar{\boldsymbol{X}}_{q\cdot} - \bar{\boldsymbol{X}}_{r\cdot}\right),$$

where $\bar{\boldsymbol{X}}_{q\cdot}$ and $\bar{\boldsymbol{X}}_{r\cdot}$ are the means and q and r are the sizes of qth and rth clusters, respectively.

We can observe that those variables do not follow normal distribution and for comparing those groups it is used the Kruskal-Wallis test, Hecke (2012) and Bewick *et al.* (2004). In the Kruskal-Wallis test the null hypothesis (H_0) considering that the groups come from the same population against the alternative hypothesis (H_1) considering the groups are originated from different populations. To perform this test, the set values of the g groups must be ordered and transformed into ranks and

considering a value of 1 for the lowest observed value, 2 for the second smallest and so on until N the highest value observed in the joint sample. The statistic is given by

$$H = \frac{\frac{12}{N(N+1)} \sum_{j=1}^{g} n_j \bar{R}_j^2 - 3(N+1)}{1 - \sum_{i=1}^{l} (t_i^3 - t_i)/(N^3 - N)} \tag{3}$$

where n_j, $j = 1, 2, ..., g$, represents the number of observations for the jth group, $N = \sum_{j=1}^{g} n_j$, $\bar{R}_j$ is the mean of the ranks for the jth group, l is the number of clusters with tied ranks and t_i is the number of ties for the ith group.

Assuming H_0 is true and for $g > 3$, $n_j > 5$, the H statistic has an approximate distribution chi-square distribution with $k-1$ degree of freedom. In every hypothesis there is always a risk associated to the decision to reject H_0 called the significance level of the test. In general, the level of significance is fixed at $\alpha = 5\%$. To decide in reject or not H_0 compares α with p-valor $= \mathbb{P}(H > h)$ where h is an estimate of 3. If p-valor $> \alpha$ then H_0 is rejected , otherwise the null hypothesis should not be rejected.

If the null hypothesis is rejected, then at least one group differs from the others, although the Kruskal-Wallis test does not identify which are the distinct groups, we can to test the difference between groups two by two by checking the validity of the inequality.

$$|\bar{R}_u - \bar{R}_v| \geq z_{\alpha/g(g-1)} \sqrt{\frac{N(N+1)}{12} \left(\frac{1}{n_u} + \frac{1}{n_v} \right)},$$

where the indexes u and v identify the groups and $z_{\alpha/k(k-1)}$ is the quantile of the standard normal distribution given by $\mathbb{P}(Z \geq z_{\alpha/k(k-1)}) = \alpha/k(k-1)$.

The study of the relationship between two different variables is performed using Spearman's rank correlation coefficient (since the variables are not normally distributed). The coefficient is a modification of the Pearson's coefficient in which the observed values of each variable are encoded using ranks. The Spearman coefficient can be expressed as

$$r_s = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N} (x_i - \bar{x})^2 \sum_{i=1}^{N} (y_i - \bar{y})^2}} \tag{4}$$

where x_i and y_i are the positions related to ith observation regarding the variables X and Y respectively and $\bar{x}$ and $\bar{y}$ are the X and Y mean ranks. The r_s coefficient is limited between -1 and 1. Values close to 1 for r_s indicate a strong positive association between X and Y, values of r_s close to -1 represent a strong negative association between the compared variables. On the other hand, if r_s is identically equal to zero or assumes values close to zero, it is said that there is no correlation between the variables or that the correlation is weak.

However, an observed value of r_s may be the result of chance due to sample randomness, and hence, it is necessary to perform hypothesis testing for correlation.

The hypotheses from the test are: H_0: "there is no association between X and Y" against H_1: "There is an association between X and Y". The statistics is given by

$$T = r_s \sqrt{\frac{N-2}{1-r_s^2}}$$

where the random variable T has t-student distribution with $N-2$ degrees of freedom.

In order to explain the distribution of the variance or covariance of the variables $\boldsymbol{X} = \left[X_1,\ X_2,\ \ldots,\ X_p\right]^T$, through linear combinations of variables $\boldsymbol{X}$, principal component analysis (PCA) can be used, where these p-variables are unrelated (MINGOTI, 2005). The principal component can be expressed as

$$\boldsymbol{Y} = \boldsymbol{O}^T \boldsymbol{X}$$

where $\boldsymbol{O}$ may be the eigenvalue of covariance matrix or data correlation. The variability explained by the principal component is given by

$$\frac{Var(Y_i)}{\text{Total variance}} = \frac{\lambda_i}{\sum \lambda_i}$$

where λ_i represents the eigenvalue associated with the variance of each principal component. So, it is possible to select a smaller number of components for the study.

3 Data set

The data were obtained from public transport company sector of the state of Paraná in 2016. Information was collected from 82 employees and 13 observed variables were identified as relevant to the problem of absenteeism. Below is the description of each variable:

1. **Certified absence***: time absent with presentation of medical certificate;
2. **Not attested absence***: time absent without justification;
3. **Delay***: time absent due to presentation after the determined time;
4. **Suspension***: time absent due to suspension;
5. **License***: time off due to license;
6. **Function**: classification of the workers position in the company as a driver or collector bus;
7. **Sex**: information of the employee's gender (male or female);
8. **Time**: accumulated service time of the employee since the hiring;

9. **Civil**: classification of marital status as married or single;
10. **Age**: age in full years;
11. **Instruction**: level of education classified in elementary or high school ;
12. **Pension**: indicative whether or not the employees pay a pension;
13. **Distance**: distance, in kilometers, from the employee's home to work.

For the descriptive study Figures 1 and 2 indicates that the mean age of employees is 45.3 years old and they reside an mean distance of 6.2 km from the company. Most of them are men, married, do not pay a pension and have completed high school. Almost 66% of them occupy a driver's position. Figure 2 shows that approximately 56% of the employees work at company at most 8 years, althogh there is a group that corresponding to 23% of the employees who worked for 15 to 18 years. The average and median length of stay of the employees at the company are 11 and 6 years, respectively.

For the study of the worker absences, the variable Absence* is obtained as a result of the sum of the first five variables listed above, given by

$$\text{Absence}_i^* = \text{Certified absence}_i^* + \text{Not attested absence}_i^* + \text{Delay}_i^* + \text{Suspension}_i^* + \text{License}_i^*$$

with $i = 1, 2, \ldots, 82$.

The descriptive results shows that 85% of the employees have up to 26 days continuous absences (Figure 2). Considering a daily workday of 8 hours that amount represents 79 days without attending the company. The average number of calendar days absent is 13 days (42 days of an 8-hour day).

The Spearman correlation test between the variables Time and Absence* presented $\hat{\rho} = 0.4375$ (p-value < 0.001) and therefore, there is a moderate and positive correlation between the employee's Time in the company and the total number of Absences obtained in the same period. However, it is to be expected that the longer an employee's time in the company the greater the chance of a high number of absences.

To overcome this problem, the variable Absence is built as a result of the sum of the first five variables listed above and standardized by Time (variable 8), defined as

$$\text{Absence}_i = \frac{10000}{\text{Time}_i} \times (\text{Certified absence}_i^* + \text{Not attested absence}_i^* + \text{Delay}_i^* + \text{Suspension}_i^* + \text{License}_i^*)$$

where $i = 1, 2, \ldots, 82$ are indexes. Here, the variables 1, 2, 3, 4 and 5 (described above) are given by

$$\begin{aligned}
\text{Certified}_i &= \frac{10000}{\text{Time}_i} \times \text{Certified absence}_i^* \\
\text{Not attested absence}_i &= \frac{10000}{\text{Time}_i} \times \text{Not attested absence}_i^* \\
\text{Delay}_i &= \frac{10000}{\text{Time}_i} \times \text{Delay}_i^* \\
\text{Suspension}_i &= \frac{10000}{\text{Time}_i} \times \text{Suspension}_i^* \\
\text{License}_i &= \frac{10000}{\text{Time}_i} \times \text{License}_i^*
\end{aligned}$$

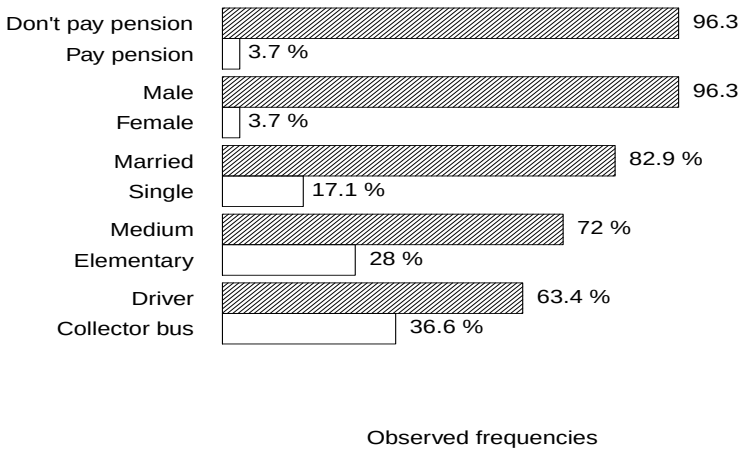


Figure 1 - Bar graph for social aspects.
Source: Authors' authorship.

The results of descriptive studies of the variables obtained are given in Table 1 and Figure 3. We can note that the variables indicate right-skewed and some extreme values which were not identified as outliers. It is also observed that the absences are largely due to Certified absences and Licenses.

Table 1 - Descriptive study for the variables Certified absence, Not attested absence, Delay, Suspension and License

Variables	Minimum	Median	Mean	Maximum	Coef. Var. (%)
Time (hours)	8766	52596	95463.89	280512	75.89
Absence	0.00	32.28	42.58	167.28	93.80
Certified	0.00	18.99	32.50	133.47	112.44
Not attested	0.00	1.71	3.46	31.94	157.89
Delay	0.00	0.00	0.12	1.13	189.19
Suspension	0.00	0.00	0.42	4.11	189.78
License	0.00	3.27	6.09	82.14	173.42

Source: Authors' authorship.

Table 2 indicates the correlations among standardized variables. We can note that there is a moderate interaction between Certified absences and Not attested absences indicating that the employees which present a large number of Certified absences tend to present a greater number of Not attested absences. Also, we observe that the absences due to Licenses and Delays show moderate association and positive. In this sense, the employees who requests more License tend to be late with higher frequency as well.

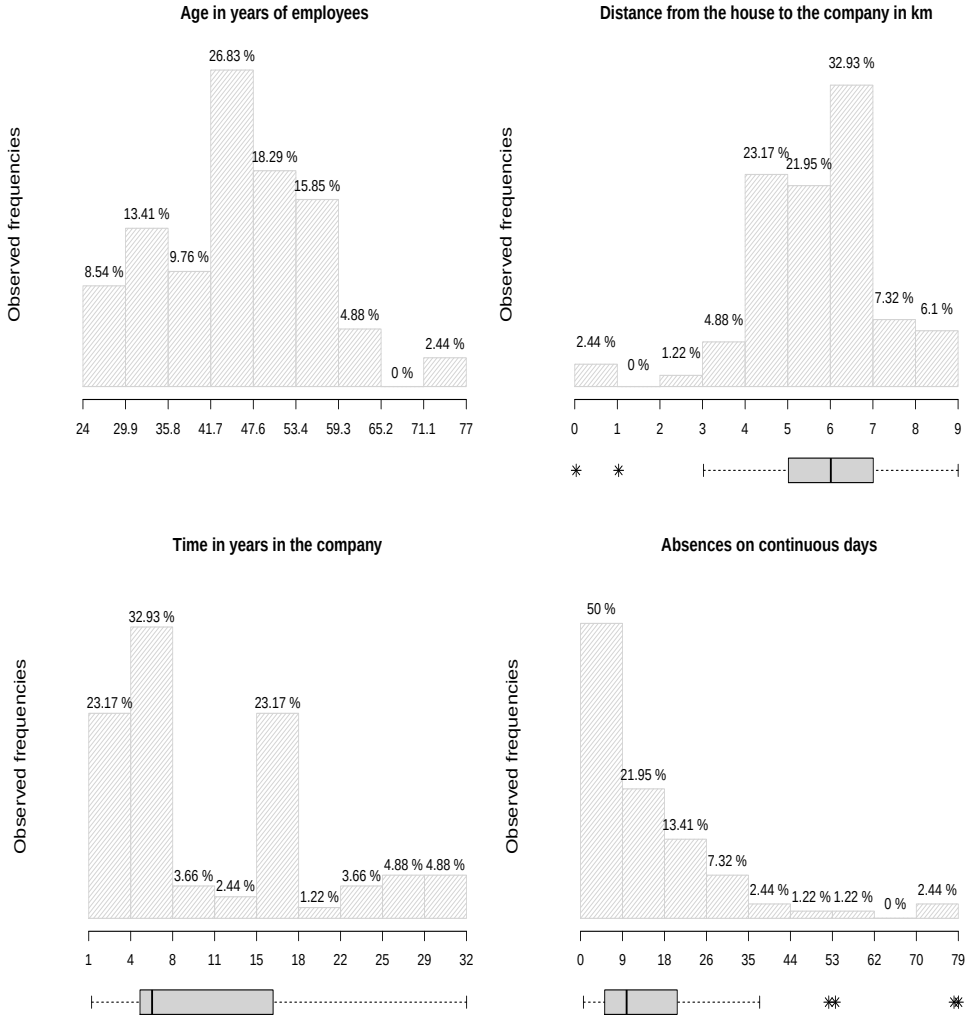


Figure 2 - Histogram for Age, Distance, Time in the company and total accumulated Absences company employees.
Source: Authors' authorship.

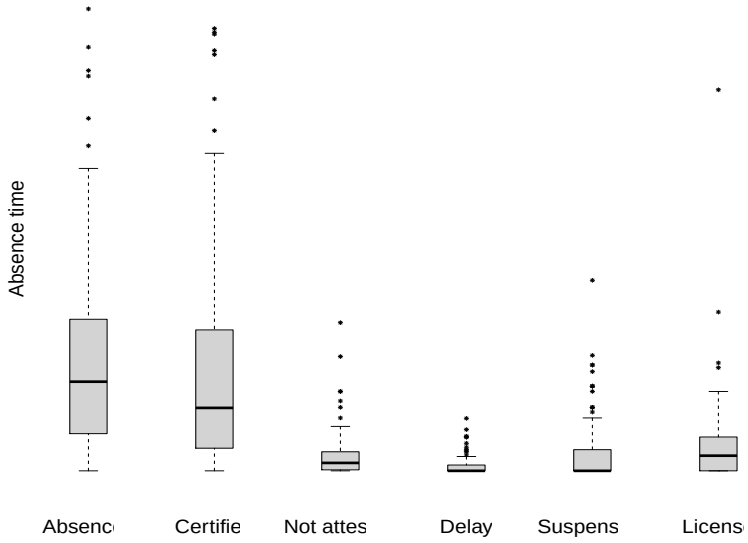


Figure 3 - Boxplot for the absence, Certified, Not attested, Delay, Suspension and License. Each boxplot is presented in a different scale for the purposes of visualization.
Source: Authors' authorship

Table 2 - Spearman correlation among the variables Certified, Not attested, Delay, Suspension and License

	Not attested	Delay	Suspension	License
Certified	0.40^{***}	0.14^{ns}	0.24^{**}	0.15^{ns}
Not Attested		0.25^{**}	0.24^{**}	0.31^{**}
Delay			0.21^{*}	0.42^{***}
Suspension				0.17^{ns}

^{ns} p−value > 0.1; * p−value < 0.05; ** p−value < 0.01; ** p−value < 0.001
Source: Authors' authorship

4 Results and discussion

In this section, we study the relationship among the variables related to the employees absence (Certified absence, Not attested absence, Delay, Suspension and License) with the variables associated with the social aspects (Function, Sex, Civil, Age, Instruction, Pension and Distance) using the principal component analysis. First of all, the complete data set was analyzed with the 82 observations considering only the variables related to the absence of the employee and after that we use the data only with the variables associated with the aspects was studied. In both situations the first principal component explained more than 90% of the data variability. The expressions obtained, expression (5), for the components are given by

$$\begin{cases} CP_1 &= 0.998 \text{ Certified} + 0.056 \text{ Not attested} + 0.002 \text{ Delay}+ \\ & \quad +0.005 \text{ Suspension} - 0.024 \text{ License} \\ CP_2 &= -0.005 \text{ Function} + 0.001 \text{ Sex} + 0.009 \text{ Civil} + 1.000 \text{ Age}- \\ & \quad -0.005 \text{ Instruction} - 0.001 \text{ Pension} + 0.001 \text{ Distance} \end{cases} \tag{5}$$

where the first principal component for the variables related to absence CP_1 explained 90.7% of the variability of the data while the first principal component for variables associated with social aspects, CP_2 explained 97.5%. The highest coefficient magnitude is related to Certified absences to CP_1 and Age to CP_2 indicating that those factors contribute strongly to the variation of the components. We can interpret CP_1 as an index associated with Certified absences and CP_2 as an index associated with Age of the employees since the remaining coefficients are proportionally smaller than the others.

The scatter plot for the two principal components is displayed in Figure 4, where it is possible to verify that there is a simultaneous behavior for the two components. However, obtaining these is not much informative since the same information is obtained by the dispersion graph of the Certified absences and Age variables, and thus,contribute little to explain the phenomenon of absenteeism in the company.

The analysis with the complete data presented so far, may not be adequate since it does not take variables together, and consequently, may mask relevant factors or even indicate results that are difficult to interpret. Even when the set of variables is taken into account (study of the principal components CP_1 and CP_2) the results are not satisfactory. This may be due to sample obtained from diverse populations that mask information about the phenomenon under study when the complete dataset is analyzed, which justifies the high variation coefficient.

The proposal, consequently, is to use multivariate techniques, in particular cluster analysis and principal components analysis, to separate the complete data set into groups more informative, taking into account factors associated with the absence of workers.

For the determination of the groups, the agglomerative hierarchical algorithm was considered Ward, Mingoti (2005) and Johnson and Wichern (2007) and the number of g groups was determined using the behavior of the fusion level at each step of Ward's algorithm. The graph of fusion level and the dendrogram are shown in Figure 5. The fusion level shows a jump more pronounced from step 79^{th} to 80^{th} of the algorithm, indicating a large increase in dissimilarity between the groups formed (from 237 to 515) and therefore, the algorithm was the step ended at 79^{th}, totaling $g = 3$ groups G1, G2 and G3, highlighted in Figure 5b.

Therefore, we obtained three groups called G1, G2 and G3 with $n = 53$, $n = 16$ and $n = 13$ observations respectively. The results for the mean values of some variables studied for each group and for the complete set ($n = 82$) is presented in Table 3. The table also shows the result for the group comparison test (Kruskal Wallis test), where equal

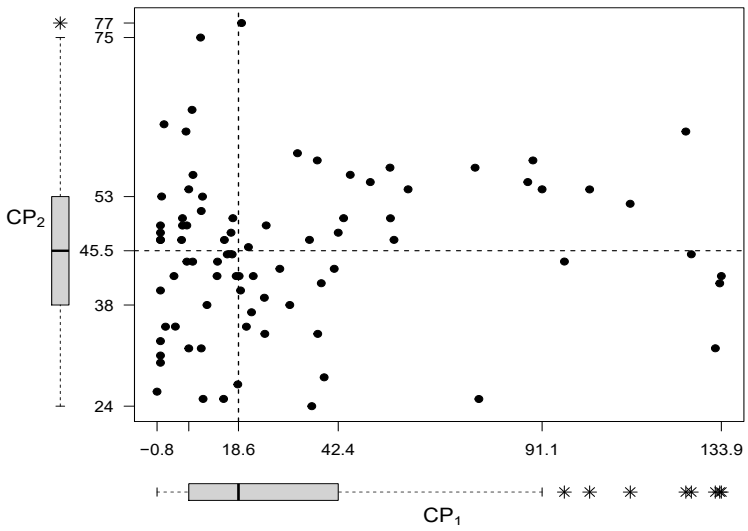


Figure 4 - Dispersion graph of the principal components CP_1 and CP_2.
Source: Authors' authorship.

letters represent statistical equality between different groups and letters identify statistical difference between the groups tested.

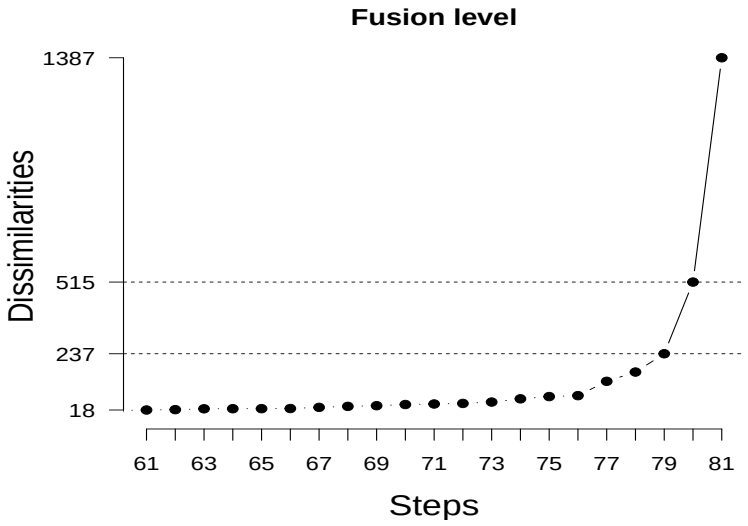
The Kruskall-wallis test show that there is no statistical difference between the groups regarding length of service at the company, Age of employees, Distance to company and absences due to Delay, Suspension and License. However, the groups difference in relation to the number of absences (in days), Certified absences and Not attested absences, shown in the Table (3). The group G1 with $n = 53$ employees had the shortest absences in contrast to the group G3 ($n = 13$) that had a significant number of absences. The group G2 statistically does not differ in absences from the group G3, but it has a lower index of Certified absences in comparison the group G3. In general, the absence, in large part, is due to the Certified absences. As for the social factors, the three groups behave similarly (Figure 7), with the exception of group G1, consisting only of employees with a medium level of education.

Table 3 - mean values for the variables of each group formed and the test result for comparison of groups

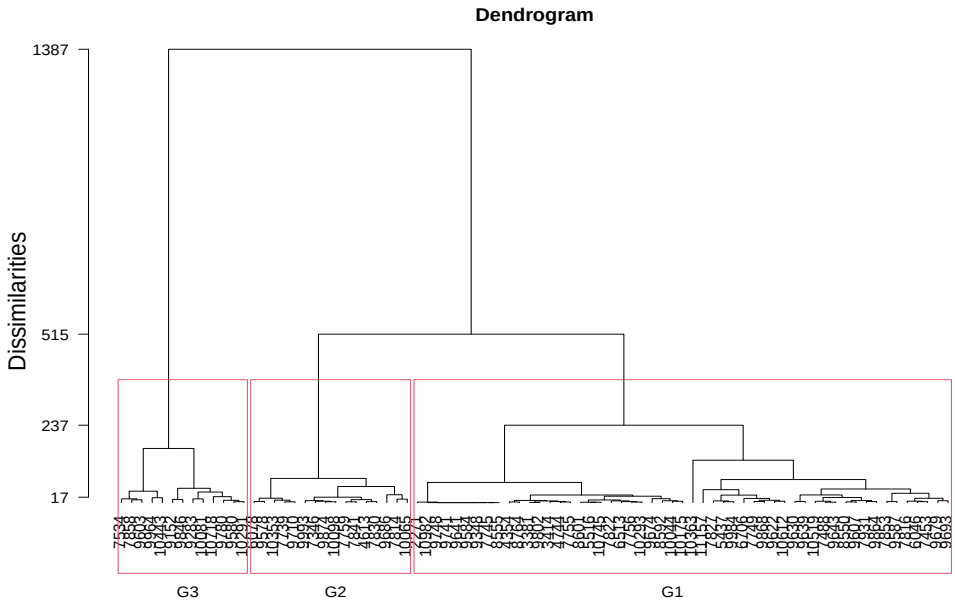
Groups	n	Absences (days)		Certified		Not attested	
G	82	14		32.5		3.5	
G1	53	8	b	10.9	c	2.4	b
G2	16	23	a	44.0	b	4.2	ab
G3	13	29	a	106.2	a	6.8	a

Source: Authors' authorship.

The study of the correlation between the variables associated with absence was carried out each group. The results can be viewed in Tables 4, 5 and 6. Likewise as



(a)



(b)

Figure 5 - (a) Graph of the behavior of the fusion level of Ward's algorithm.
(b) Dendrogram diagram with the hierarchical history of the groups formed, being the y-axis the dissimilarities and the x-axis the codifications of the employees.
Source: Authors' authorship

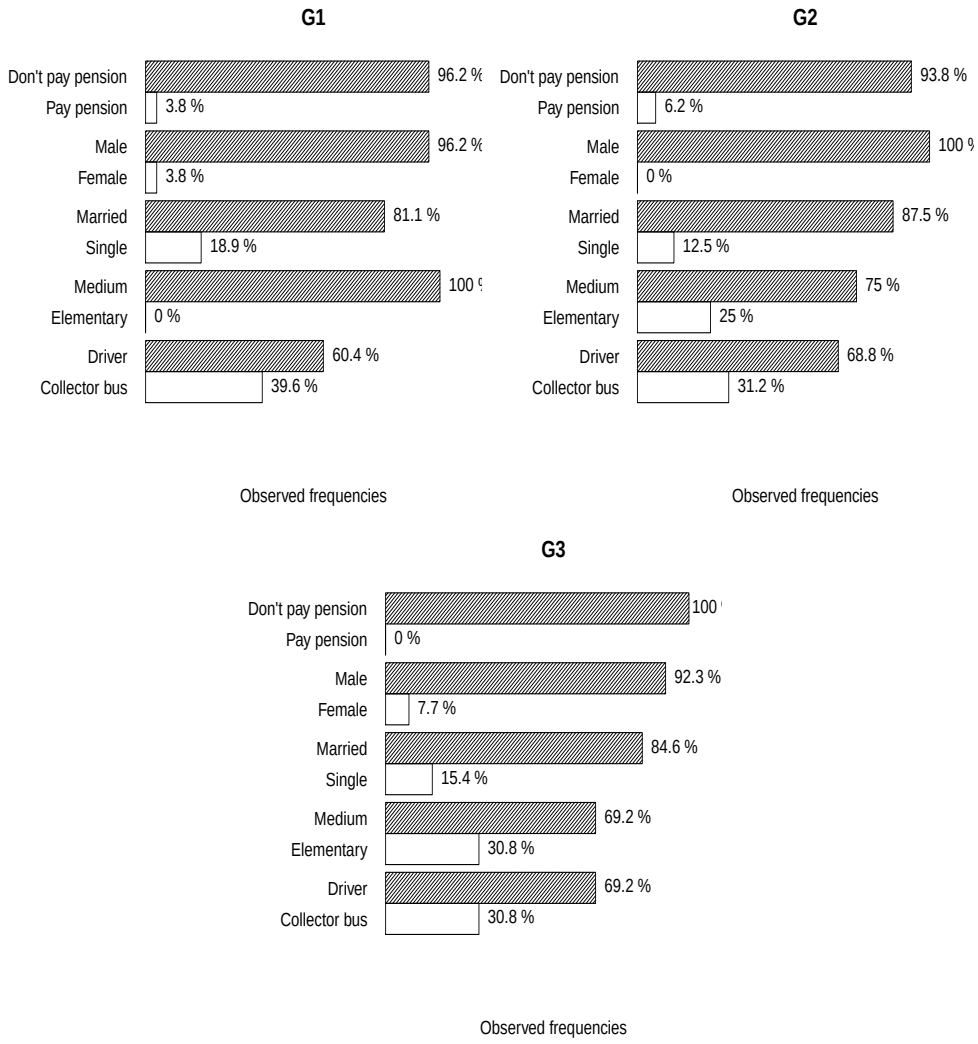


Figure 6 - Bar graph for social variables of each group.
Source: Authors' authorship.

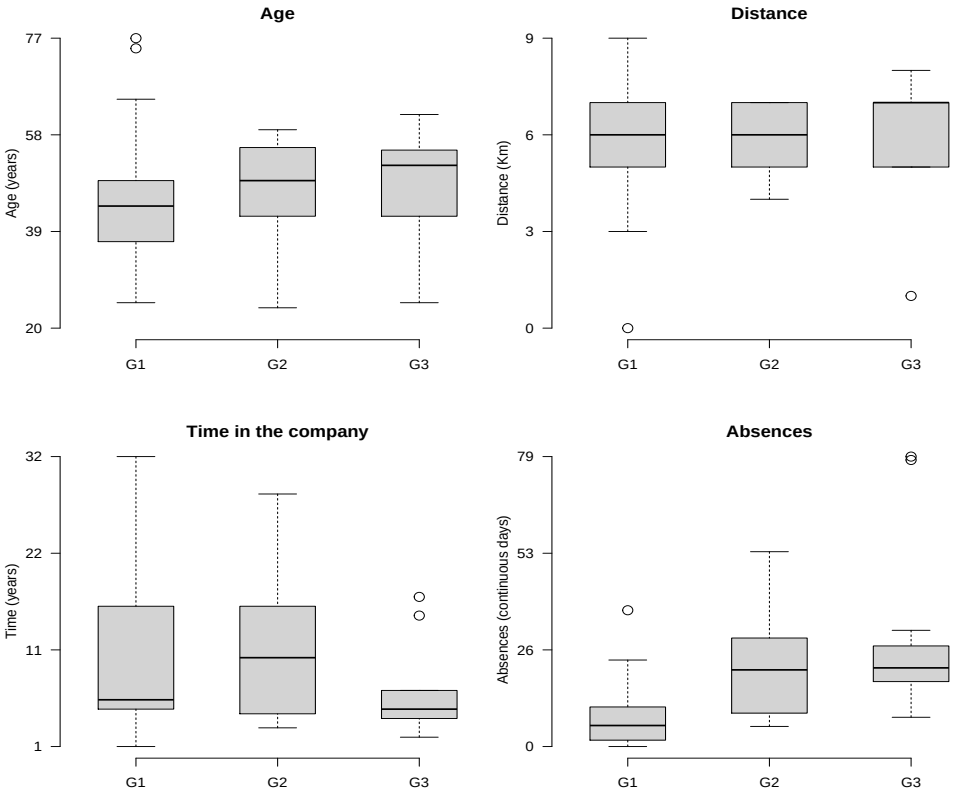


Figure 7 - Boxplot for the variables Age, Distance, time and Absences for each group.

Source: Authors' authorship.

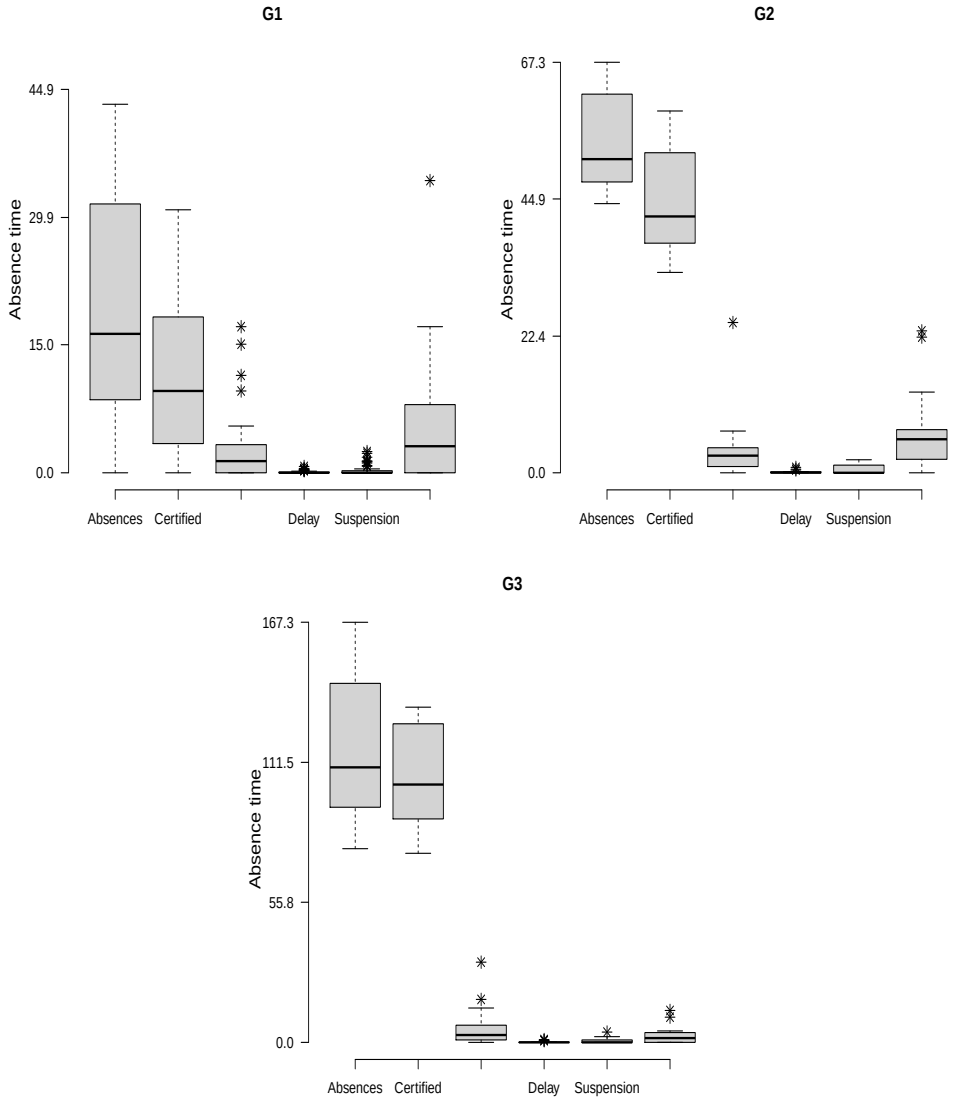


Figure 8 - Boxplot for the variables associated with absences within each group.
Source: Authors' authorship.

occurred with the complete data, Certified absences and Not attested absences presented moderate and positive correlation, $\hat{\rho}_1 = 0.42$ for group G1 and $\hat{\rho}_3 = 0.59$ for group G3, indicating that the increase in Certified absences may be associated with the increase, even though minor magnitude, in the Not attested absences for the employees of these groups.

We also note that for group G1 the correlation for absence due to License significant for all variables, indicating that the increase in employees of that License related group may imply an increase due to other factors, although the correlation coefficient is low. Group G2 presented an association moderate between absence due to Delay, Not attested absences and License. This suggests that employees who are often late tend to have more Not attested absences and Licenses.

Table 4 - Spearman correlation – Group G1

	Not attested	Delay	Suspension	License
Certified	0.42^{**}	0.05^{ns}	0.24^{*}	0.44^{**}
Not attested		0.09^{ns}	0.25^{*}	0.35^{**}
Delay			0.11^{ns}	0.43^{**}
Suspension				0.34^{*}

^{ns} p−valor > 0.1; * p−valor < 0.1; ** p−valor < 0.01; ** p−valor < 0.001

Source: Authors' authorship.

Table 5 - Spearman correlation – Group G2

	Not attested	Delay	Suspension	License
Certified	-0.41^{ns}	-0.25^{ns}	-0.07^{ns}	-0.41^{ns}
Not attested		0.51^{*}	-0.07^{ns}	0.38^{ns}
Delay			0.14^{ns}	0.49^{*}
Suspension				-0.18^{ns}

^{ns} p−valor > 0.1; * p−valor < 0.1; ** p−valor < 0.01; ** p−valor < 0.001

Source: Authors' authorship.

Table 6 - Spearman correlation – Group G3

	Not attested	Delay	Suspension	License
Certified	0.59^{*}	0.20^{ns}	0.07^{ns}	0.08^{ns}
Not attested		0.45^{ns}	0.43^{ns}	0.04^{ns}
Delay			0.39^{ns}	0.39^{ns}
Suspension				-0.10^{ns}

^{ns} p−valor > 0.1; * p−valor < 0.1; ** p−valor < 0.01; ** p−valor < 0.001

Source: Authors' authorship.

Finally, the study of principal components was carried out within each group obtained. For the construction of component CP_1 and CP_2 the data set was divided according to the variables associated with absenteeism (Certified absence, Not attested absence, Delay, Suspension and License) and according to social variables (Function of the employee in the company, Sex, Civil, Age, Instruction, Pension and Distance). The expressions of the principal components for each formed group are presented below.

For the components of group G1, expression (6), we observed that CP_1 (which explains 66.2% of variation of data) is dominated by License and Certified, whose coefficients are 0.95 and 0.3 respectively. And for CP_2 , which explains 97.4%, age is the variable that presents higher coefficient and therefore, CP_1 and CP_2 can be interpreted as being indices for "documented" absences and Age respectively. The scatter plot for these two components is shown in Figure 9a, where it can be seen that employees with higher Age are absent less, due to Certified absences and License than younger employees.

Group - G1

$$\begin{cases} CP_1 &= 0.3\ \text{Certified} + 0.04\ \text{Not attested} + 0.0009\ \text{Delay} + \\ & \quad +0.005\ \text{Suspension} + 0.95\ \text{License} \\ CP_2 &= -0.01\ \text{Function} + 0.001\ \text{Sex} + 0.005\ \text{Civil} + 1.0\ \text{Age} - \\ & \quad -0.009\ \text{Instruction} - 0.0009\ \text{Pension} - 0.0003\ \text{Distance} \end{cases} \tag{6}$$

For group G2, CP_1 explains 65.8% of the data variation and presents a contrast between Certified absences, Not attested absences and Licenses and CP_2 explains 98.6% of the variability data and is dominated by Age, expression (7). The plot in Figure 9b shows the prevalence Certified absences to older employees. Also a employee who has low value for CP_1 and CP_2 , that is, a young employee with high frequency of Not attested absences and Licenses. As a matter of curiosity the employee has 28 years old (mean Age of the group is 46.9 years), has been with the company for 5 years, is single, pays a pension, occupies the position of driver and has 168 hours in Certified absences, 108 hours in Not attested absences, 234 hours in Delays and 102 hours of License.

Group - G2

$$\begin{cases} CP_1 &= 0.7\ \text{Certified} - 0.04\ \text{Not attested} - 0.01\ \text{Delay} + \\ & \quad +0.003\ \text{Suspension} - 0.6\ \text{License} \\ CP_2 &= -0.001\ \text{Function} + 0.02\ \text{Civil} + 1.0\ \text{Age} + 0.01\ \text{Instruction} - \\ & \quad -0.00006\ \text{Pension} + 0.03\ \text{Distance}\ , \end{cases} \tag{7}$$

And finally for group G3, expression (8), CP_1 explains 86.9% of the data variation and is dominated for Certified absences and Not attested absences and CP_2 is again dominated by Age. In the graph for CP_1 versus CP_2 (Figure 9c) it is possible to observe that older people are absent less due to Certified and Not attested absences than younger people.

Group - G3

$$\begin{cases} CP_1 &= 0.97\ \text{Certified} - 0.2\ \text{Not attested} - 0.005\ \text{Delay} + \\ & \quad +0.004\ \text{Suspension} - 0.01\ \text{License} \\ CP_2 &= 0.02\ \text{Function} + 0.004\ \text{Sex} + 0.01\ \text{Civil} + 1.0\ \text{Age} - \\ & \quad -0.005\ \text{Instrution} - 0.02\ \text{Distance}\ , \end{cases} \tag{8}$$

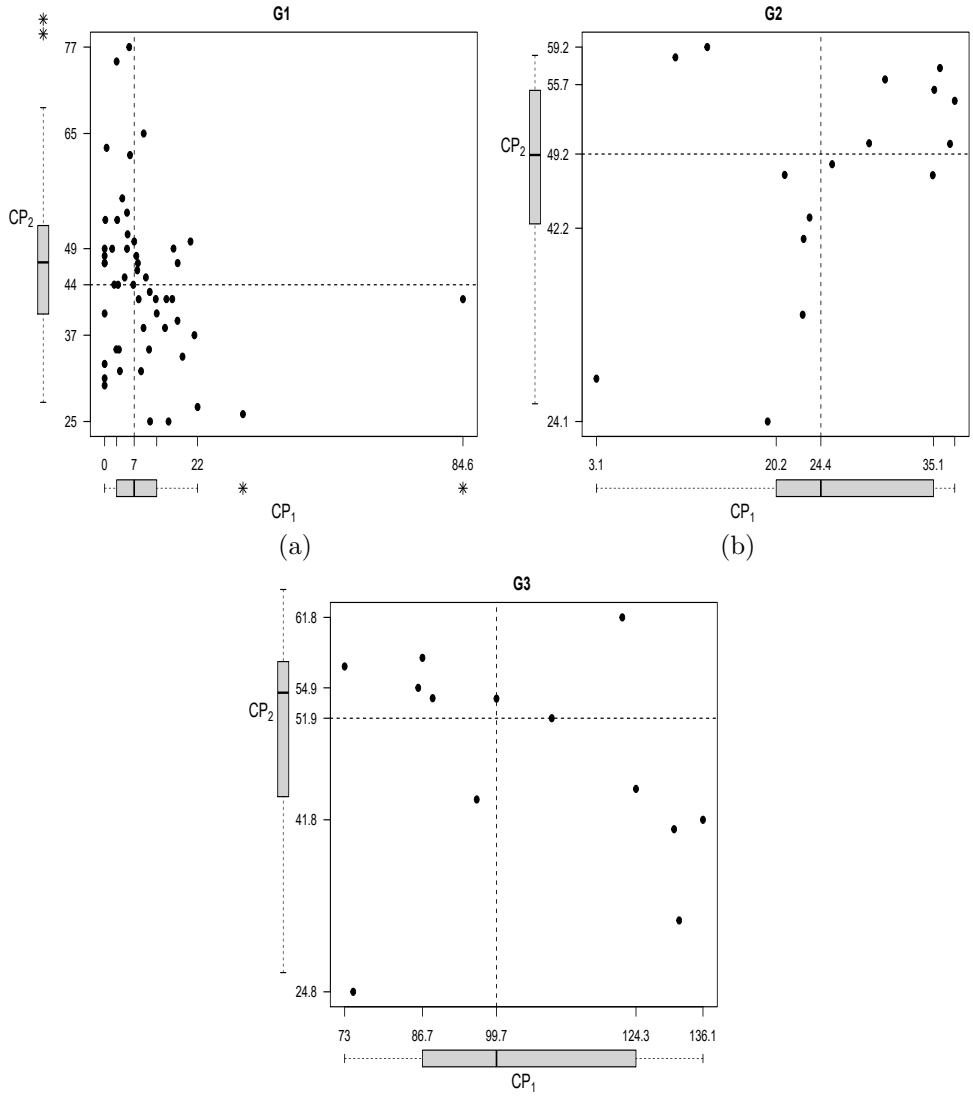


Figure 9 - Scatter plot for the components CP_1 and CP_2 of groups G1, G2 and G3.
Source: Authors' authorship.

Concluding remarks

In this paper, a cluster analysis was performed in which the existence of three distinct groups, these being G1, G2 and G3. Thus, it is concluded that the group G1 presents fewer absences in relation to groups G2 and G3, which do not differ statistically, however, the group G3 has a high number of attested absences. Regarding the social aspects, the three groups have similar characteristics with the exception of group G1, which presents a group of employees with complete high school education. The principal component analysis also highlighted the factors that contribute to positive or negative way for absences in the company, with Age being the main factorin the social aspect for absences. And related to absences, groups G1 and G2 are associated with Certificates and Licenses while the group G3 associates with Certificates and Not attested. For future work, a greater number of explanatory variables would be interesting regarding absence, such as number of children, history of illness, region housing, motivation and shift to study why these groups are absent and try to fix them.

CALCAGNOTO, L. R.; SANTANA, T. V. F.; PESCIM, R. R. Classificação e identificação das causas do absenteísmo em uma empresa de transporte publico utilizando técnicas de análise de *cluster* e componentes principais. *Rev. Bras. Biom.,* Lavras, v.39, n.1, p.25-44, 2021.

- *RESUMO: O absenteísmo é a prática ou costume de um colaborador de se ausentar de seu local de trabalho. Suas causas são diversas e afetam a renda do trabalhador, provoca transtornos operacionais, estressa a administração e causa prejuízos financeiros para empresa. A análise de cluster é uma ferramenta multivariada que pode ser utilizada para determinar grupos de modo que cada grupo apresente características próprias de acordo com as variáveis observadas. Assim, pode-se utilizar essa técnica como suporte para determinar as características que contribuem para o absenteísmo. O método para construção dos clusters utilizado foi o algorítimo hierárquico de Ward e para comparação dos grupos o teste não paramétrico de Kruskal-Wallis foi adotado. Por fim, um estudo sobre a força de associação entre as variáveis foi desenvolvido utilizando a correlação de Spearman e para a relação entre variáveis relacionadas a ausência e os aspectos sociais utilizou-se a análise de componentes principais. Através desse estudo foi possível determinar três grupos heterogêneos na empresa e evidenciar características nesses grupos que são potenciais fatores causadores do absenteísmo em maior ou menor grau.*

- *PALAVRAS-CHAVE: Análise de Clusters, Kruskal-Wallis, Correlação de Spearman, Componentes principais, Absenteísmo;*

References

ALMEIDA, D. R. O; NASCIMENTO, I. G; SILVA NETO, J. M.; ALMEIDA, A. G. B. Causas e desvantagens do absenteísmo: O Caso da Empresa Auto Center 24 Horas em Porto Velho. In: CONGRESSO NACIONAL DE EXCELÊNCIA EM GESTÃO. 2015. *Proceedings*, Rio de Janeiro:RJ, 2015.

BEWICK, V; CHEEK, L; BALL, J. Statistics review 10: Further nonparametric methods. *Critical Care*, v.8, n.196, 2004.

CALAIS, S. L.; ZANELATO, L. S.; Manejo de estresse e outros fatores em diferentes populações adultas. In: VALLE, TGM., and MELCHIORI, LE., *Saúde e desenvolvimento humano. São Paulo: Editora UNESP*, 2010. 217-236.

CHATFIELD, C.; COLLINS, A. J. *Introduction to multivariate analysis* , Springer, 2013. 246p.

FERREIRA, D. F. *Estatística multivariada* , 1 ed., Lavras: ed. UFLA, 2008. 662p.

HECKE, T. V. Power study of anova versus Kruskal-Wallis test. *Journal of statistics and management systems*, v.15, n.2-3, p.241-247, 2012.

JOHNSON, R. A; WICHERN, D. Wa. Do. *Applied multivariate statistical analysis*, 6 ed., Londres: Prentice Hall, 2007. 800p.

MANLY, B. J. F. *Métodos estatísticos Multivariados: uma introdução*, 3 ed., Porto Alegre: Bookman, 2008. 229p.

MINGOTI, S. A. *Análise de dados através de métodos de estatística multivariada: Uma abordagem aplicada*. Belo Horizonte: Editora UFMG, 2005. 297p.

PENATTI, I.; QUELHAS, O.; ZAGO, J. S. Absenteísmo: As consequências na gestão de pessoas. In: SIMPÓSIO DE EXCELÊNCIA EM GESTÃO E TECNOLOGIA. 2006. *Proceedings*, Resende:RJ, 2006.

QUICK,T. C.;LAPERTOSA, J. B. Análise do absenteísmo em usina siderúrgica. *Rev. Bras. Saúde Ocupacional*, v.10, n.40, p.62-67, 1982.

R Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018.

RStudio Team. *RStudio: Integrated Development for R*. RStudio, Inc., Boston, MA , 2016.

Received on 24.08.2020.

Approved after revised on 28.01.2021.