The use of artificial intelligence in the instrumentalization of disaster classifiers

https://doi.org/10.56238/Connexpemultidisdevolpfut-008

Samir Batista Fernandes
State Secretariat of Civil Defense of Rio de Janeiro. Deputy Director of the Scientific and Technological Institute in Civil Defense (ICTDEC). Master in Security and Civil Defense. R. Elpídio Boa Morte – Praça da Bandeira – 20270-170, Rio de Janeiro – RJ, Brazil;

Wagner dos Anjos Carvalho
Assistant Professor, coordinator of the post-graduation Industry 4.0 and professor of the MBA courses in Management and Educational Technologies and Controllership and Finance of the Mackenzie Rio Presbyterian College. Coordinator of the postgraduate course in Cost Engineering, Calculating Engineering and Artificial Intelligence for Accounting at Unyleya College. Professor of the graduate program in Occupational Safety Engineering and Project Management at UNIVERSO. Professor of the MBA in Project Management (ESTÁCIO). Professor of the MBA in Strategic Management of Marketing and Sales (ESTÁCIO). Professor at the Brazilian Institute of Municipal Administration (IBAM).

ABSTRACT
Disasters are events that claim thousands of lives, increasing in both quantity and intensity. For this reason, work that assists in understanding disasters has the potential to save lives and assist in best practices in policy planning aimed at reducing disaster risks. Thus, the research aimed to group disasters by classes using artificial intelligence considering the variables selected in the platform of the integrated disaster information system [S2ID] of the municipalities of Rio de Janeiro. To implement the algorithm, the work was structured in four sections, in which the first was contextualized, a brief problematization of the absence of disaster classifiers using artificial intelligence in the Brazilian system, followed by the second section, in which the materials and methods implemented in the research were presented. Then the results were presented and discussed in the third section and we concluded with the final considerations in the fourth section. At the end of the research we were able to create classes of disasters from quantitative variables that can help in understanding the intensity of disasters and allow the Civil Defense, the agency responsible for disaster risk reduction management, to develop better strategies in its activities.

Keywords: Civil Defense, Principal component analysis, Disaster intensity.

1 INTRODUCTION

(KOBIYAMA, MENDONCA, et al. , 2006) (DOURADO, ARRAES e SILVA, 2012) Disasters are events that produce damage and damage, as well as compromising local response capacity. In the State of Rio de Janeiro, disasters increase in frequency and intensity, in addition to claiming countless lives. Disasters are recorded in the integrated disaster information system – [S2ID], in which data are entered into digital processes with the dual purpose of validating the recognition of the disaster and the possibility of obtaining instruments for the recovery of the affected area. (FERNANDES, 2022)

Therefore, in the localities where there is a declaration of a disaster, it is considered that there were measurable variables such as: amount of rainfall, number of deaths, number of people displaced, number of requests for emergency surveys, a monetary value that represents the damages and losses, etc.. (CHMUTINA e BOSHER, 2017)

The quantification of the disaster is measured in the platform [S2ID] and later the recognition processes in Brazil are regulated by normative rites being evaluated according to the interpretation of a human being, through supporting documentation of the data entered. However, there is a limitation of the human being's ability to identify patterns and make inferences from a large database.(CABENA, 1997)

Research that combines past data analysis, the use of techniques with artificial intelligence and the search for non-trivial patterns that can categorize classes of disasters in a database are advantageous and constitute an excellent tool since the human being would have difficulty in making inferences in large volume of data or complex analyses. (FAYYAD, SHAPIRO e SMYTH, 1996)

Thus, it is increasingly necessary to study the variables that are related to the elements that trigger disasters and increasingly allow society to protect itself, in addition to developing more effective public policies. Research involving disasters represents a relevant aspect of society and has the potential to help save lives, but when it comes to the topic of disaster, many fields of knowledge launch their vision from a specific science. That is, geotechnics discusses the disaster from the perspective of the hard sciences, sociology and related sciences approaches the theme of disasters from the perspective of the social field and so with all sciences. The science of disasters is a relatively new area of knowledge that appropriates other fields to build its advances. (SOUSA, SANTOS e SOUZA, 2020)

Although the topic is so important, because it is a subject that involves human protection, the classification of disasters in [S2ID] only became required by law after 2022 according to the response of the National Secretariat of Protection and Civil Defense in a public demonstration during this research. As the data of interest and that are entered in the [S2ID] are, for the most part, quantitative metrics, the hypothesis is that there is the possibility of elaborating classes of disasters. For Azevedo the problem to be solved can be researched, as long as it has a clear, feasible objective and results attainable through a scientific process. (2018)

Thus, the objective of the research is to group disasters by classes using artificial intelligence considering the variables selected in the platform [S2ID] of the municipalities of Rio de Janeiro. The research will contribute to a disaster classification technique and may help in the planning of better public policies aimed at reducing disaster risks. Therefore, this research is aimed at professionals who work in disaster risk reduction management and other employees who work in the protection and civil defense system.

The work is structured in four sections in which the first is contextualized a brief problematization of the absence of disaster classifiers using artificial intelligence in the Brazilian system, followed by the second section in which the materials and methods implemented in the research are presented. Then the results are presented and discussed in the third section and concluded with the final considerations in the fourth section.

2 IMPLEMENTATION OF "MACHINE LEARNING" ALGORITHM

In this topic will be addressed the materials and methods used for the implementation of "Machine learning" algorithm starting with an object of study, the choice of a database, the materials and the choice of methods used (tests, techniques and analyses).

2.1 OBJECT OF STUDY

The State of Rio de Janeiro is divided into 92 municipalities, has an area of approximately 43,780 km² and is a region prone to natural disasters such as landslides and floods. The territorial extension of the state and the large number of municipalities make the region especially vulnerable to these events.

Disasters are recorded on the federal government platform [S2ID] for registration and analysis for recognition with the state or federal government. The process of disaster analysis is carried out through documentary analysis and compliance with normative rites.

2.2 MATERIAL USED

The research was conducted using the database of the [S2ID] from the Brazilian federal government. To access the database it was necessary to request the federal government, through a public demonstration. Disasters with the following characteristics were selected:

- Recognized by the federal government typified according to the Brazilian Classification and Codification of Disasters – [COBRADE] as: (1) Convective storm or heavy rains, (2) cold fronts or convergence zones, (3) Landslides, (4) floods and (5) floods;
- Damages and losses related to rainfall in its territory;
- Information on the intensity of rainfall in 24 hours and
- Registered in the database [S2ID], through the disaster information form – FIDE and in the technical opinion.

From the above criterion, ninety-six (96) processes with records on disasters in [S2ID] were selected. From the analysis of the processes, variables of interest were selected and recorded in a new database, in the form of a spreadsheet, as shown in table 1 below:

Table 1 - Variables selected for algorithm implementation (continues)

Variable	Meaning of the variable
Municipality	Name of the federative entity that suffered the disaster
Population	Absolute amount of municipal population declared in the lawsuit
Area	Municipal territorial extension
Annual budget	Total projected revenue for the annual fiscal year
Annual Revenue	Total revenue received for the annual fiscal year
Damages and Losses	Monetary value ascertained as a result of the effects of the disaster
Affected	Total population impacted by the disaster
Variable	**Meaning of the variable**
Rainfall	Amount of rain in mm in the period of 24 hours
Population density	Ratio between the population and total area of the municipality
Capacity to invest in disaster response	Ratio of annual revenue to disaster-related damages
Financial projection capacity	Ratio of annual revenue to annual budget
Population density of those affected	Ratio of those affected by disasters to the total population

Source: The Authors (2023)

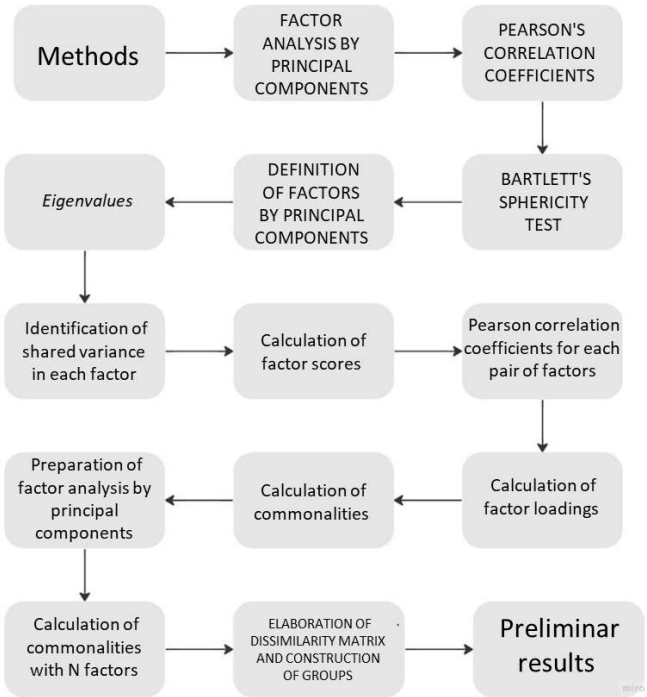
2.3 LIMITATIONS OF THE STUDY

Processes with the following were excluded from the database: (1) absence of data in the fields of interests; (2) processes whose disasters were not related to rainfall and (3) disasters of a gradual nature, that is, those that occur slowly and can be measured in days or months.

2.4 METHODOLOGY USED

The methodology below in Figure 1 was used, following the steps suggested by Fávero and Belfiore aiming at the implementation of the use of an algorithm for the classification of disasters. 2017

Figure 1 - Methodology used in disaster classification



Source: The Authors (2023)

2.5 CLUSTER ANALYSIS

Cluster analysis, also known as cluster analysis or cluster analysis, is a set of useful exploratory techniques to verify the existence of similar behaviors between observations and create homogeneous groups in relation to certain variables. These techniques are not predictive for other observations outside the initial sample, allow comparing the measure of similarity between observations within the same group and have the limitation of the inclusion of new observations, that is, new variables require a reapplication of modeling. To carry out the research was selected the method of factor analysis by principal components that had great contribution of Karl Pearson. (KING, 2014) (FÁVERO e BELFIORE, 2017) (ANDERSON, 2003; LOVRIC e ARSHAM, 2011)

2.6 PRINCIPAL COMPONENT FACTOR ANALYSIS

Factor analysis is an exploratory multivariate technique that seeks to establish new variables, called factors, from the grouping of original variables that have high correlation coefficients. This technique is useful to reduce the size of the data and identify relationships between the original variables, but it does not have a predictive character for other observations not present in the sample. (FÁVERO e BELFIORE, 2017)

Principal component factor analysis is the most widely used method in dimensionality reduction, as it allows to determine another set of variables resulting from the linear combination of the first set of variables. Principal component factor analysis has four main objectives: (1) structural reduction; (2) verification of the validity of previously established constructs; (3) elaboration of "rankings"; and (4) extraction of orthogonal factors for later use in confirmatory multivariate techniques that require the absence of multicollinearity. The extracted factors can be used as explanatory variables of other variables, such as in confirmatory multivariate models, such as multiple regression. In summary, factor analysis is a useful technique to identify relationships between original variables, reduce the size of the data, and create variables from the grouping of original variables with high correlation coefficients. (RENCHER e CHRISTENSEN, 2012) (FÁVERO e BELFIORE, 2017)

2.7 PEARSON CORRELATION COEFFICIENTS FOR EACH PAIR OF VARIABLES

From a baif of data that presents n observations (in the search were called processes inserted in the [S2ID]), and for each observation i ($i=1,\ldots, n$) values corresponding to each of the K metric variables X, as shown in table 2, it was possible to correlate each process analyzed as an observation. For each process, the variables described in table 1 were used.

Table 2 - General model of a database for the elaboration of factor analysis

Note i	X1i	X2i	...	ski
1	X11	X21		Xk1
2	X12	X22		Xk2
3	X13	X23	...	Xk3
⋮	⋮	⋮		⋮
n	X1n	X2n		Xkn

Source: FÁVERO and BELFIORE (2017)

Given a process p and q distinct in the database, one can represent a degree of similarity between them by the following expression:

$$\rho_{pq} = \frac{\sum_{j=1}^{k}\left(X_{jp} - \overline{X_p}\right) \times \left(X_{jq} - \overline{X_q}\right)}{\sqrt{\sum_{j=1}^{k}\left(X_{jp} - \overline{X_p}\right)^2} \times \sqrt{\sum_{j=1}^{k}\left(X_{jq} - \overline{X_q}\right)^2}} \quad (1)$$

where ρ_{pq} is the value of the Pearson coefficient between two distinct processes in the database and $\bar{X}_p$ and $\bar{X}_q$ represent the mean of each of the rows in the database. The above expression allows us to analyze the similarity between the rows of the database and the inline behavior of the observations for the set of variables.

2.8 BARLETT'S SPHERICITY TEST

Bartlett's Sphericity Test is a statistical test used to compare a correlation matrix with an identity matrix ρI of the same size. The test is used to assess the suitability of the data for factor analysis. To Favero and Belfiores "(2017, p. 386)if the differences between the corresponding values outside the main diagonal of each matrix are not statistically different from O, at a given level of significance, we can consider that the extraction of the factors will not be adequate". Therefore, Bartlett's Sphericity test tests the null hypothesis that the correlation matrix is an identity matrix, which means that there are no correlations between the variables. A significant result indicates that the data are suitable for factor analysis. (LOVRIC e ARSHAM, 2011)

3 DEFINITION OF FACTORS BY PRINCIPAL COMPONENTS

Once adequate for factor analysis has been identified, it is necessary to employ the factors that partially carry a characteristic of the original variables in groupings of variables. Below we will see the factors that helped in the determination of the groupings of variables.

- *Eigenvalues* - Eigenvalues are numerical measures that indicate the relative importance of each principal component in representing the variability of the data in a principal component factor analysis. They represent the variance explained by each component and are used to determine how many components should be retained in the analysis. The higher

the eigenvalue, the more important the contribution of that component to the data structure.(HAIR JR., BLACK, W.C., *et al.*, 2019) (JOHNSON e WICHERN, 2007)

- *Identification of the shared variance in each factor* The identification of the shared variance in each factor in the factor analysis by principal components is essential to understand the contribution of each factor to the total variation of the data. Shared variation is measured by eigenvalues, which indicate the amount of total data variation explained by each factor. The higher the eigenvalue, the greater the amount of variation explained by the corresponding factor. Thus, the identification of the shared variance in each factor allows to select the most relevant factors for the explanation of the data,.(TABACHNICK e FIDELL, 2013) (HAIR JR., BLACK, W.C., *et al.*, 2019)

It is common to use factors greater than 1 to represent the principal factors in principal component factor analysis because this allows you to retain more total variance from the original data set. For example, if only one main factor is defined, it may not be sufficient to represent all the variation of the data, resulting in a loss of important information. Therefore, by increasing the number of key factors, you can capture more variance and relevant information from the original dataset. (JOLLIFFE, 2002)

- *Calculation of factorial scores* – The calculation of factorial scores is an important step in factor analysis by principal components, as it allows to obtain numerical values for each individual (or object) in relation to each factor. These scores represent each individual's score in relation to the importance or influence of each factor in the data set. The scores are calculated by combining the values of the original variables with the weights of the corresponding factors. These weights are called "loadings" and indicate the importance of each variable for each factor. The result is a matrix of factorial scores that can be used for further analysis, such as visualizing the data in scatterplots or comparing groups of individuals in relation to each factor. (JOLLIFFE, 2002)
- *Pearson correlation coefficients for each pair of factors* Pearson's correlation coefficients for each pair of factors in principal component factor analysis are important for understanding the relationship between the factors and how they relate to the original variables. These coefficients measure the correlation between the factorial scores of each pair of factors and range from -1 to 1. A positive coefficient indicates a positive correlation between the factors, while a negative coefficient indicates a negative correlation. A coefficient close to zero indicates a weak or non-existent correlation. The interpretation of Pearson's correlation coefficients is fundamental for the interpretation of the results of the factor analysis by principal components,. (JOHNSON e WICHERN, 2007) (TABACHNICK e FIDELL, 2013)

- *Calculation of factor loadings* – Factor loadings are the coefficients that show the relationship between each original variable and each factor extracted in the factor analysis by principal components. They indicate how much each variable contributes to the definition of each factor. The calculation of factor loadings is performed by means of the correlation between each original variable and each factor, weighted by the inverse of the square root of the corresponding eigenvalues. The higher the factor load of a variable in a factor, the greater its contribution to the definition of that factor. (HAIR JR., BLACK, W.C., *et al.*, 2019; TABACHNICK e FIDELL, 2013)
- *Calculation of commonalities* – The calculation of commonalities in factor analysis by principal components is important to evaluate the amount of variance shared between the original variables and the extracted factors. Commonalities represent the proportion of the total variance of each original variable that can be explained by the factors. This measure is calculated by adding the squares of the factor loadings of each variable and can range from 0 to 1. When commonalities are high, this indicates that the original variables are well represented by the extracted factors. Otherwise, you may need to reconsider the factorial model or add more factors.(JOLLIFFE, 2002) (HAIR JR., BLACK, W.C., *et al.*, 2019)
- *Elaboration of the Factor Analysis by Principal Components* – In Principal Component Factor Analysis, it is common to extract factors only with eigenvalues greater than 1. This approach is taken because, by extracting only the most important factors, it is possible to retain most of the total variance of the original dataset. It is then possible to rotate the factors to obtain a clearer and more interpretable structure. The interpretation of the extracted factors is performed by means of the factor loadings, which represent the correlation between each original variable and each extracted factor.(HU e BENTLER, 1999) (BORGES, ESTEVES, *et al.*, 2018)
- *Calculation of commonalities with N factors* - The calculation of commonalities with N factors is performed from the sum of squares of the factor loads squared for each variable in each of the N factors. The commonalities represent the amount of total variance of each variable that can be explained by the factors extracted. When N is equal to the total number of variables, the commonalities correspond to the variances of the variables. Commonalities are important to assess the adequacy of the factorial model and identify which variables contribute more to the explanation of the factors. (FÁVERO e BELFIORE, 2017) (ENAP, 2019)

3.1 ELABORATION OF DISSIMILARITY MATRIX AND CONSTRUCTION OF CLUSTERS

Once the steps in the construction of the factors by principal components were completed, the next step was to construct the matrix of dissimilarities that is used in hierarchical "cluster" analysis to group similar objects into "clusters". In principal component factor analysis, the matrix of dissimilarities is calculated based on the factors extracted and the factor loadings of the objects. This approach allows the incorporation of the dependency structure between the variables in the grouping process, resulting in more accurate analyses and more reliable interpretations. (BORG e GROENEN, 2005) (INDHUMATHI e SATHIYABAMA, 2010)

The hierarchical grouping created in the factor analysis by principal components contains groups of disasters that present similar characteristics to each other with statistical relevance. One-factor analysis of variance (ANOVA) is a statistical technique used to compare the mean of one group with the mean of another group. It is used to test whether the difference in the means of two or more groups is significant or whether it can be attributed to chance. The output of an ANOVA includes the F statistic, which is used to assess the significance of the difference between the means of the groups, the p-value, which indicates whether the difference is statistically significant, and the ANOVA table, which shows the contribution of each source of variation in the difference between the means of the groups.

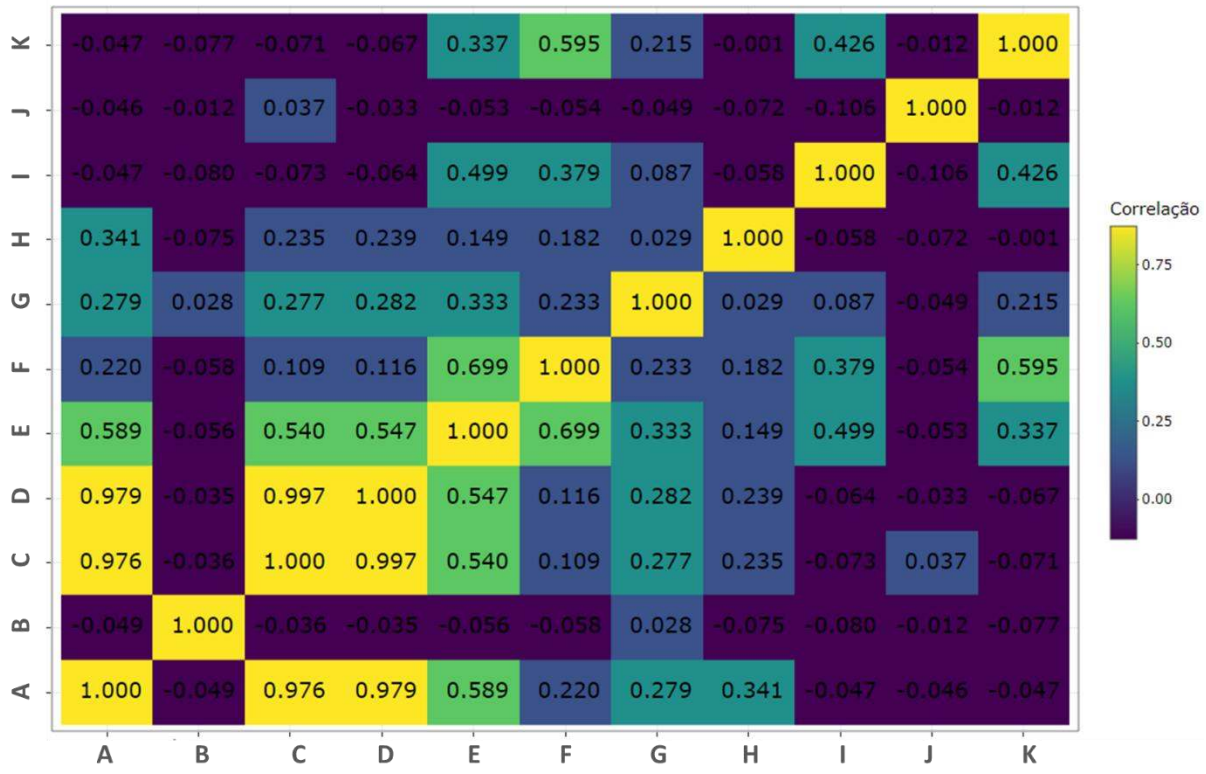
At the end of the whole process of creating clusters, univariate sensitivity analysis was used, which is the evaluation of the impact of variations in certain variables of a mathematical or statistical model on the results obtained, in which the highest and lowest values were taken from the samples.

In the next section the results of the methodology will be presented.

4 PRELIMINARY RESULTS

By identifying the degree of correlation between the pairs of variables, by means of Pearson's coefficient, it was possible to identify in which variables the correlation is higher or lower according to Figure 2. Variables such as: Population; Area; Annual Revenue; Annual budget; Damages and Losses and Affected carry the greatest correlations with each other.

Figure 2 - heat map of Pearson correlations between variables



Legend: A: Population; B: Area; C: Annual Revenue; D: Annual budget; E: Damages and Losses; F:Affected; G: Rainfall 24h; H: Population density; I: Capacity to invest in disaster response; J: Financial projection capacity; K: Population density of those affected;

Source: The Authors (2023)

Despite the low correlation between some variables, the Bartlett's sphericity presented the value statistically equal to O (zero), at a high level of significance, corresponding outside the main diagonal of each matrix. Thus, it can be considered that the extraction of the factors was adequate, so values can be extracted from the original variables, and the analysis of the factorial variable is appropriate.

Given the 11 (eleven) variables considered in the database, the analysis of the eigenvalues was performed with ten values of the main components – PC for further reduction obtaining the following results: PC1: 3,763; PC2: 2,268; PC3: 1,079; PC4; 1,027; PC5; 0.902; PC6: 0.826; PC7: 0.575; PC8: 0.447; PC9: 0.107; PC10: 0.011 and PC11: 0.

It is possible to observe in table 3 that PC1, PC2, PC3 and PC4 together have more than 70% of the accumulated variance and that by the latent root criterion or Kaiser's criterion CP greater than 1 (one) were considered, that is, PC1, PC2, PC3 and PC4. (FÁVERO e BELFIORE, 2017)

Table 3 - Calculation of factor loadings

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
Eigenvalues	3.763	2.268	1.079	1.027	0.902	0.826	0.575	0.447	0.107	0.011	0.000
Variance Prop.	0.342	0.206	0.097	0.093	0.082	0.075	0.052	0.040	0.009	0.001	0.000
Cumulative Variance Prop.	0.342	0.548	0.645	0.738	0.820	0.896	0.948	0.989	0.998	0.999	1.000

Source: The Authors (2023)

From the eigenvalues and the calculation of the factor load, it was possible to determine the factorial "scores". The "scores" allowed the achievement of Pearson's coefficient between the original variables and each of the factors. According to table 4, the process of "Standardized Loadings" ("Pattern Matrix") was performed and it was verified that 4 components would be sufficient for the representation of the variance of the original variables. That way we are left with PC1, PC2, PC3 and PC4. Redoing the Bartlett sphericity test for the new matrix created it was possible to verify that the value statistically equal to O (zero), at a high level of significance, corresponding outside the main diagonal of each matrix, which means that even with the original variance losses there was still a good statistical representation of these.

The variables Population, Budget, Revenue, Damages and Losses carry high original variances in PC1. Affected, damages and losses and high population density variance in PC2. In general, most variables carry a high commonality, which allows you to build a model and understand the contribution of a given variable to the main components.

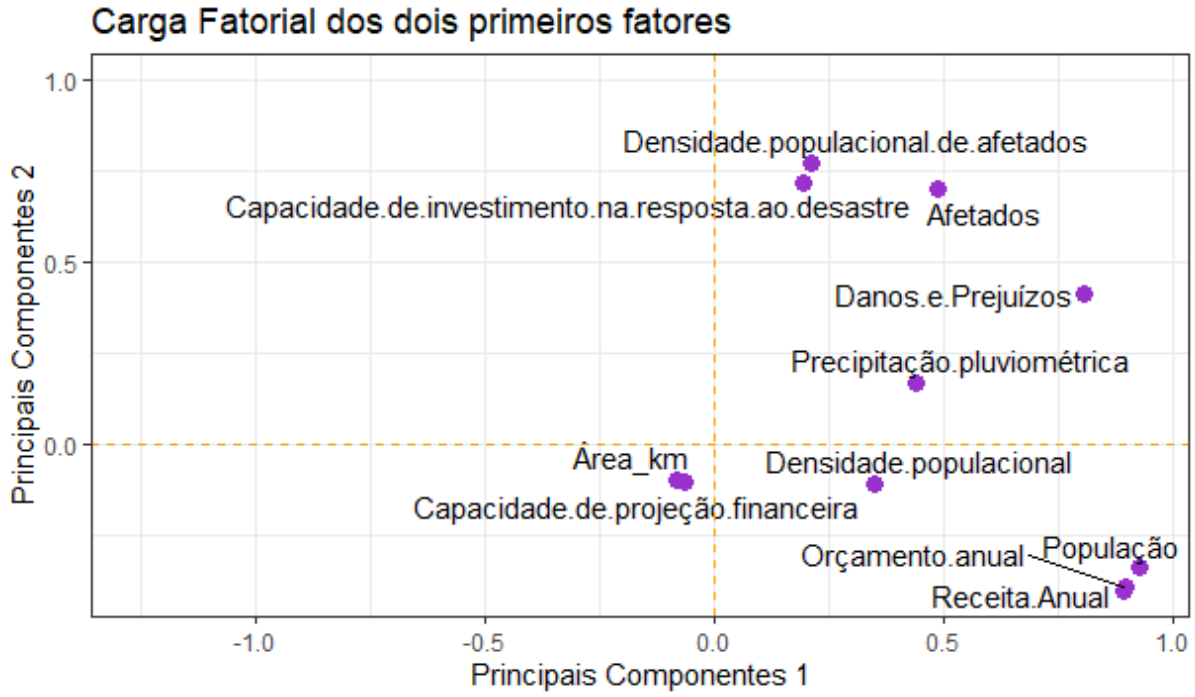
Table 4 – “Standardized Loadings” (“Pattern Matrix”) baseadas na matriz de correlação

	PC1	**PC2**	**PC3**	**PC4**	**Comunalidades**
Population	0.93	-0.34	-0.03	0.00	0.98
Area	-0.08	-0.10	0.69	-0.43	0.68
Annual budget	0.89	-0.40	0.07	0.08	0.97
Annual Revenue	0.90	-0.39	0.05	0.02	0.96
Damages and Losses	0.81	0.41	0.06	0.03	0.83
Affected	0.49	0.70	-0.06	0.02	0.73
Rainfall	0.44	0.17	0.40	-0.10	0.39
Population density	0.35	-0.11	-0.59	-0.17	0.52
Capacity to invest in disaster response	0.20	0.72	-0.01	-0.02	0.56
Financial projection capacity	-0.06	-0.10	0.26	0.88	0.86
Population density of those affected	0.21	0.77	0.03	0.10	0.65

Source: The Authors (2023)

In the elaboration visual of the factor load of the main factors, according to Figure 3, it is observed that population, budget and annual revenue are well grouped and have a positive correlation for high PC1. Another correlation is observed when grouping population density, disasters and affected, that is, a greater number of people affected in the disaster, increases the density of people affected in the territory and compromises the capacity to invest in the response.

Figure 3 - Factorial load of PC1 and PC2



Source: The Authors (2023)

In the analysis of variance of the factors (ANOVA) it was possible to verify that there is a greater variation outside the groups than inside the groups, this is possible to verify by the mean Sq of the clusters that are greater than the mean Sq of the residues according to the table 5 down. The F-value of the table 5 allowed us to verify that factor 1 has the most discriminating variable of the groups statistically and significantly. And since the Pr value of all factors are statistically equal to zero, the null hypothesis is rejected and we can conclude that at least one of the means of the groups is significantly different from the others. Thus, it is possible to group, with statistical significance, in groups of disasters that have similar characteristics to each other and that have differences outside the groups.

Table 5 - "Output" da ANOVA

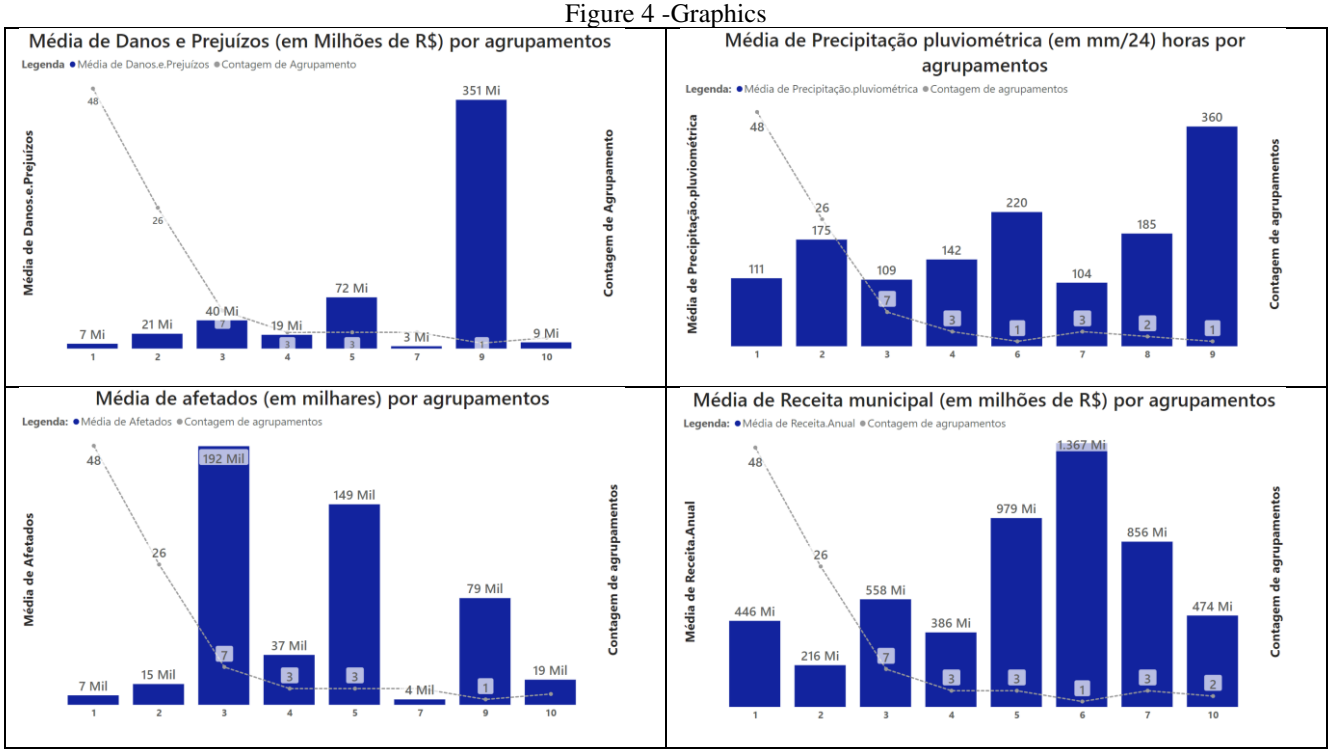
	Grouping	Soma Sq	Average Sq	F value	Pr(>F)
Factor 1	Groupings	88.83	9.870	137.6	<2e-16
	Waste	6.17	0.072		
Factor 2	Groupings	84.85	9.427	79.84	<2e-16
	Waste	10.15	0.0118		
Factor 3	Groupings	86.03	9.559	91.66	<2e-16
	Waste	8.97	0.104		
Factor 4	Groupings	52.32	5.813	11.71	<13e-12
	Waste	42.68	0.496		

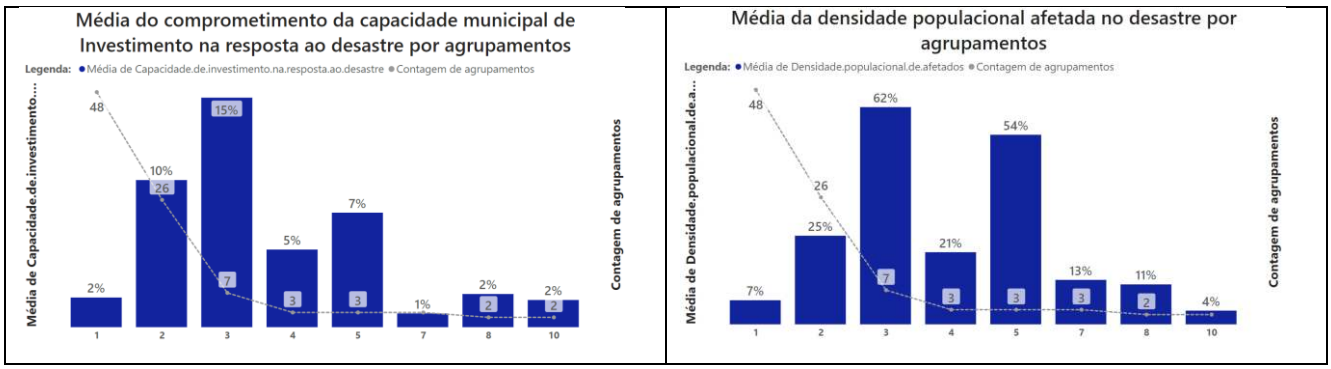
Source: The Authors (2023)

From the Validation of the above results, it was possible to create 10 (ten) groups of disasters classified with labels from 1 to 10 in a qualitative way and represented in Figure 4, using the sensitivity analysis, in which the classes with the highest and lowest value were removed. It should be noted that disaster classes 1, 2 and 3 have the largest number of disasters in which they total 81 disaster processes.

The average economic losses in classes 1, 2 and 3 are not the largest, but They are the classes that are most affected in their ability to invest in disaster response and have the lowest revenues. Group 3 has the highest average number of people affected in disasters, in terms of compromising the ability to invest in disaster response and in the population density affected, however the average rainfall for disasters in group 3 is small.

Figure 4 -Graphics





Source: The Authors (2023)

5 FINAL CONSIDERATIONS

Using the material and methodology used and presented the results, it was possible to group the disasters by classes, using artificial intelligence, considering the variables selected in the platform [S2ID] of the municipalities of Rio de Janeiro. The creation of disaster classes is of utmost importance to improve emergency prevention and response strategies. The classification by intensity allows a more precise and adequate analysis of the risks associated with each type of disaster, which can help in the definition of more effective preventive measures and in the mobilization of resources in crisis situations.

In this sense, the use of artificial intelligence to classify disasters can be extremely useful. Machine learning algorithms can analyze a large volume of data that humans would be unable to analyze to identify patterns and correlations that allow them to classify the event accurately and quickly.

In summary, the creation of disaster classes and the use of artificial intelligence to classify events can contribute significantly to improving risk and emergency management. These measures can save lives, preserve public and private assets and ensure the safety of the population in crisis situations. Future suggestions include works that correlate and make predictive analyses of municipal capacity in Brazil.

REFERENCES

ANDERSON, T.W. An Introduction to Multivariate Statistical Analysis. California: [S.n.], v. III, 2003.

BORG, I; GROENEN, P. J. Modern multidimensional scaling: Theory and applications. [S.l.]: Springer, 2005.

BORGES, Vinicius R. P. et al. Using Principal Component Analysis to support students' performance prediction and data analysis. VII Congresso Brasileiro de Informática na Educação (CBIE 2018), Brasilia, 2018.

CHMUTINA, Ksenia; BOSHER, Lee. Disaster Risk Reduction for the Built Environment. [S.l.]: [S.n.], 2017.

ENAP. Análise fatorial. Livro da Fundação Escola Nacional de Administração Pública, Brasília, 2019.

FÁVERO, Luiz Paulo Lopes; BELFIORE, Patrícia Prado. Manual de análise de dados: estatística e modelagem multivariada com excel, SPSS e stata. Rio de Janeiro: Elsevier, 2017.

FAYYAD, Usama ; SHAPIRO, Gregory Piatetsky; SMYTH, Padhraic. From Data Mining to Knowledge Discovery in Databases. AI Magazine. American Association for Artificial Intelligence, v. 17, 1996.

HAIR JR., J.F. et al. Multivariate Data Analysis. 7ª. ed. [S.l.]: Prentice Hall, 2019.

HU, Li-tze; BENTLER, Peter M. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling: A Multidisciplinary Journal, 1999.

INDHUMATHI, R; SATHIYABAMA, S. Reducing and Clustering high Dimensional Data through Principal Component Analysis. International Journal of Computer Applications , World, Dezembro 2010.

JOHNSON, Richard A. ; WICHERN, Dean W. Applied Multivariate Statistical Analysis. [S.l.]: Pearson, 2007.

JOLLIFFE, I. T. Principal Component Analysis. World: Springer, 2002.

KING, Ronald S. Cluster Analysis and Data Mining: An Introduction. [S.l.]: [S.n.], 2014.

LOVRIC, Miodrag ; ARSHAM, Houssein. Bartlett's Test. [S.l.]: [S.n.], 2011.

RENCHER, A. C.; CHRISTENSEN, W. F.. Methods of Multivariate Analysis. [S.l.]: ohn Wiley & Sons., 2012.

SOUSA, Hélio Marcus Damasceno; SANTOS, Daiana Ferreira de Deus; SOUZA, Gabriel Vieira Damasceno de. Disaster research: challenges and contributions to society. Journal of Environmental Analysis and Progress, 2020. 57-66.

TABACHNICK, B. G.; FIDELL, L. S. Using Multivariate Statistics. Boston: Pearson, 2013.

TABACHNICK, B. G.; FIDELL, L. S. Using multivariate statistics. [S.l.]: [S.n.], 2013.

APPENDIX

ALGORITHM USING R

```
<- c("plotly", "tidyverse", "ggrepel", "knitr", "kableExtra", "reshape2", "PerformanceAnalytics",
"psych", "Hmisc", "readxl", "cluster", "factoextra")

if(sum(as.numeric(!pacotes %in% installed.packages())) != 0){
 < installer- packages[!packages %in% installed.packages()]
 for(i in 1:length(instalador)) {
  install.packages(instalador, dependencies = T)
  break()}
 sapply(pacotes, require, character = T)
} else {
 sapply(pacotes, require, character = T)
}

list.files()
desastres <- read.csv("dados_07042023.csv", sep = ",", dec = ".")

# Descriptive statistics
summary(disasters)
resumo_estatistico <- summary(desastres)

# Scatter and adjustment between the variables 'affected' and 'precipitation'
Disasters %>%
 ggplot() +
 geom_point(aes(x = affected, y = precipitacao_24h),
          color = "darkorchid",
          size = 3) +
 geom_smooth(aes(x = affected, y = precipitacao_24h),
           color = "orange",
           method = "loess",
           formula = y ~ x,
           if = FALSE,
           size = 1.3) +
 labs(x = "Affected",
     y = "Precipitation in 24h") +
 theme_bw()

# Pearson correlation coefficients for each pair of variables
rho <- rcorr(as.matrix(desastres[,2:12]), type="pearson")

correl <- rho$r # Matrix of correlations
sig_correl <- round(rho$P, 4) # Matrix with p-value of the coefficients

# Elaboration of a heat map of Pearson's correlations between variables
ggplotly(
 disasters[,2:12] %>%
  cor() %>%
  melt() %>%
  rename(Correlação = value) %>%
  ggplot() +
  geom_tile(aes(x = Var1, y = Var2, fill = Correlação)) +
```

```
    geom_text(aes(x = Var1, y = Var2, label = format(Correlação, digits = 1)),
              size = 5) +
    scale_fill_viridis_b() +
    labs(x = NULL, y = NULL) +
    theme_bw())

# Visualization of the distributions of variables, scatters, values of correlations
chart. Correlation(desastres[,2:12], histogram = TRUE, pch = "+")


### Elaboration of the factor analysis by principal components ###

# Bartlett's sphericity test
cortest.Bartlett(Disasters[,2:12])

# Elaboration of factor analysis by principal components
factorial <- main(disasters[,2:12],
                  nfactors = length(disasters[,2:12]),
                  rotate = "none",
                  scores = TRUE)

# Eigenvalues (autovalores)
eigenvalues <- round(fatorial$values, 5)
eigenvalues

round(sum(eigenvalues), 2)

# Identification of the shared variance in each factor
variancia_compartilhada <- as.data.frame(fatorial$Vaccounted) %>%
  slice(1:3)

rownames(variancia_compartilhada) <- c("Autovalores",
                                       "Variance Prop.",
                                       "Cumulative Variance Prop.")

# Variance shared by the original variables for the formation of each factor
round(variancia_compartilhada, 3) %>%
  kable() %>%
  kable_styling(bootstrap_options = "striped",
                full_width = FALSE,
                font_size = 20)

# Calculation of factorial scores
scores_fatoriais <- as.data.frame(fatorial$weights)
write.csv(scores_fatoriais, file = "D:/OneDrive - defesacivil.rj.gov.br/USP/n supervisionadas
exercicios PCA/scores_fatoriais.csv", row.names = TRUE)


# Visualization of factorial scores
round(scores_fatoriais, 3) %>%
  kable() %>%
  kable_styling(bootstrap_options = "striped",
                full_width = FALSE,
```

```
            font_size = 20)

# Calculation of the factors themselves
Factors <- as.data.frame(factorial$scores)

View(factors)

# Pearson correlation coefficients for each pair of factors (orthogonal)
rho <- rcorr(as.matrix(fatores), type="pearson")
round(rho$r, 4)

# Calculation of factor loadings
cargas_fatoriais <- as.data.frame(unclass(fatorial$loadings))

# Visualization of factor loadings
round(cargas_fatoriais, 3) %>%
  kable() %>%
  kable_styling(bootstrap_options = "striped",
                full_width = FALSE,
                font_size = 20)

# Calculation of commonalities
comunalidades <- as.data.frame(unclass(fatorial$communality)) %>%
  rename(comunalidades = 1)

# Visualization of commonalities (here they are equal to 1 for all variables)
# 10 factors were extracted in this first moment
round(comunalidades, 3) %>%
  kable() %>%
  kable_styling(bootstrap_options = "striped",
                full_width = FALSE,
                font_size = 20)


### Elaboration of Factor Analysis by Principal Components ###
### Factors extracted from eigenvalues greater than 1 ###

# Definition of the number of factors with eigenvalues greater than 1
k <- sum(eigenvalues > 1)
print(k)

# Elaboration of factor analysis by principal components without rotation
# With quantity 'k' of factors with eigenvalues greater than 1
Fatorial2 <- Main (disasters[,2:11],
                   nfactors = k,
                   rotate = "none",
                   scores = TRUE)
fatorial2
print(fatorial2)


#Cálculo of commonalities with only the first 'k' ('k' = 4)
comunalidades2 <- as.data.frame(unclass(fatorial2$communality)) %>%
```

```
  rename(comunalidades = 1)

# Visualization of commonalities with only the first 'k' ('k' = 4) factors
round(comunalidades2, 4) %>%
  kable() %>%
  kable_styling(bootstrap_options = "striped",
                full_width = FALSE,
                font_size = 20)

# Loading plot with the loads of the first two factors
cargas_fatoriais[, 1:2] %>%
  data.frame() %>%
  rownames_to_column("variables") %>%
  ggplot(aes(x = PC1, y = PC2, label = variables)) +
  geom_point(color = "darkorchid",
             size = 3) +
  geom_text_repel() +
  ggtitle("Factorial Load of the first two factors") +
  theme(plot.title = element_text(hjust = 0.5)) +
  geom_vline(aes(xintercept = 0), linetype = "dashed", color = "orange") +
  geom_hline(aes(yintercept = 0), linetype = "dashed", color = "orange") +
  expand_limits(x= c(-1.25, 0.25), y=c(-0.25, 1)) +
  labs(x = "Main Components 1",
       y = "Main Components 2") +
  theme_bw()

# has ended the factor analysis and now the cluster analysis will be performed
# Adding the extracted factors to the original database

# Cluster Analysis Using the 4 Factors
disasters <- bind_cols(disasters,
                       "fator_1" = factors$PC1,
                       "fator_2" = factors$PC2,
                       "fator_3" = factors$PC3,
                       "fator_4" = factors$PC4)

# Cluster Analysis Using the 4 Factors

# Analysis of factors (mean and standard deviation)
summary(disasters[,13:15])
SD(disasters[,13])
SD(disasters[,14])
SD(disasters[,15])

# Matrix of dissimilarities
matriz_D <- disasters[,13:15] %>%
  dist(method = "euclidean")
print(matriz_D)

# Elaboration of hierarchical clustering
cluster_hier <- agnes(x = matriz_D, method = "complete")
print(cluster_hier)
```

```
# Definition of the hierarchical scheme of agglomeration

# The distances to the combinations at each stage
coeficientes <- sort(cluster_hier$height, decreasing = FALSE)
Coefficients

Schema <- as.data.frame(cbind(cluster_hier$merge, coefficients))
names(schema) <- c("Cluster1", "Cluster2", "Coefficients")
scheme

# Dendrogram construction
dev.off()
fviz_dend(x = cluster_hier, show_labels = FALSE)

# Dendrogram with visualization of clusters (definition of 10 clusters)
fviz_dend(x = cluster_hier,
        h = 3.0,
        show_labels = FALSE,
        color_labels_by_k = F,
        rect = F,
        rect_fill = F,
        ggtheme = theme_bw())

res.pca <- PCA(USArrests, ncp = 3, graph = FALSE)

fviz_dend(res.hcpc, show_labels = FALSE)

res.hcpc <- HCPC(res.pca, graph = FALSE)

fviz_cluster(res.hcpc, geom = "point", main = "Factor map")

fviz_dend(cluster_hier,
        cex = 0.7, # Label size
        palette = "jco", # Color palette see ?ggpubr::ggpar
        rect = TRUE,
        rect_fill = TRUE, # Add rectangle around groups
        rect_border = "jco", # Rectangle color
        labels_track_height = 0.8 # Augment the room for labels)

# Creating categorical variable for cluster indication in the database
## The 'k' argument indicates the number of clusters
desastres$cluster_H <- factor(cutree(tree = cluster_hier, k = 10))

# One-factor analysis of variance (ANOVA). Interpretation of the output:

# ANOVA of the variable 'factor 1'
summary(anova_fator_1 <- aov(formula = fator_1 ~ cluster_H,
                        date = disasters))

# ANOVA of the variable 'factor 2'
summary(anova_fator_2 <- aov(formula = fator_2 ~ cluster_H,
                        date = disasters))
```

```
# ANOVA of the variable 'factor 3'
summary(anova_fator_3 <- aov(formula = fator_3 ~ cluster_H,
                        date = disasters))

# ANOVA of the variable 'factor 4'
summary(anova_fator_3 <- aov(formula = fator_4 ~ cluster_H,
                        date = disasters))

# 24h precipitation
group_by(disasters, cluster_H) %>%
  summarise(
    mean = mean(Precipitação.pluviométrica, na.rm = TRUE),
    sd = sd(Precipitation.rainfall, na.rm = TRUE),
    min = min(Precipitation.rainfall, na.rm = TRUE),
    max = max(Precipitação.pluviométrica, na.rm = TRUE),
    obs = n())

# affected
group_by(disasters, cluster_H) %>%
  summarise(
    mean = mean(Affected, na.rm = TRUE),
    sd = sd(Affected, na.rm = TRUE),
    min = min(Affected, na.rm = TRUE),
    max = max(Affected, na.rm = TRUE),
    obs = n())

# Damages and losses
group_by(disasters, cluster_H) %>%
  summarise(
    mean = mean(Danos.e.Prejuízos, na.rm = TRUE),
    sd = sd(Damages.e.Damages, na.rm = TRUE),
    min = min(Damages.e.Damages, na.rm = TRUE),
    max = max(Damage.e.Damages, na.rm = TRUE),
    obs = n())
```