Statistical analysis of the risk of fires in the Pantanal in the first five months of 2022

Carlos Rogerio da Silva Ruiz[1], Roger Sedlacek Fibiger[2]

[1]Faculdade de Computação – Universidade Federal do Mato Grosso do Sul (UFMS)
79070-900 – Campo Grande – MS – Brazil

carlos.ruiz@ufms.br, roger@fibiger.com.br

Abstract. *This paper performs a statistical analysis to verify the impact of fire risk on factors such as number of days without rain in the region, accumulated value precipitation on the day until the moment detection of the outbreak and Fire Radioactive Power (FRP) made available by the Instituto National Space Research Institute (INPE), Corumbá-MS city, from January to May 2022.*

1. Introduction

Fires in the pantanal biome have increased significantly in recent years. The National Institute for Space Research (INPE) has developed a research, development and product innovation program, and technological geoservices development processes for monitoring risk and active development in research, its extension and gravity development, using Remote Sensing techniques, Geoprocessing and Numerical Modeling.

In order to develop an analysis of fire risk data provided by the INPE monitoring system considering the period from January 2022 to May 2022, Corumbá-MS city, a period of experiments will be presented below to verify a hypothesis of significance no impact of the fire rain in the region, the days of the fire rain in the region, the collection value without the focus collection and the Fire Radioactive Power (FRP) compared through comparison to the selected risk test.

2. Methodology

To carry out the analysis proposed in this document, data from the satellite monitoring system called BDQueimadas [Banco de Dados de queimadas 2022] which provides information captured via satellite and presents the fire risk that represents the condition of the fire risk observed using meteorological data from the last 120 days. The data made available by INPE are publicly disclosed, free of charge. Altogether, ten satellites are used that have optical sensors operating in the 4um thermal-medium range, which INPE is able to receive. The images of the polar satellites, the AVHRR/3 of NOAA-18 and 19, METOP-B and C, the MODIS of NASA TERRA and AQUA and the VIIRS of the NPP-Suomi and NOAA-20 and the images of the geostationary satellites, GOES -16 and MSG-3.

2.1. System Data BDQueimadas

With the universe of information disclosed in the program, a set of data was generated filtered by the Pantanal biome, in Corumbá-MS city, between January 01, 2022 and May 31, 2022, disregarding samples showing an error in the value of the accumulated precipitation on the day, the number of days without rain and the Fire Risk Value predicted for the day of detection of the outbreak (value equal to -999). Samples without the Fire Radiative Power (FRP) value were also disregarded. As information on temperature, relative air humidity, vegetation types were not available, they are not considered in the tests. Also not considered at the moment are the days and times and satellite used. Thus, the factors are days without rain, precipitation value and Fire Radiative Power (FRP) measured in MW (megawatts), with a total of 6197 samples collected, between-subject, as they are carried out under different conditions.

Fire risk is classified into five levels on the following scales from 0 to 1:

- Minimum, below 0.15;
- Low, from 0.15 to 0.4;
- Medium, from 0.4 to 0.7;
- High, from 0.7 to 0.95 e;
- Critical, above 0.95

The FRP is obtained by multiplying the radiance by the pixel area, resulting in units of Watts or MegaWatts, the same FRP value can correspond either to a small burn at high temperature, or to a more extensive burn at a lower temperature. Therefore, neither FRP can give an accurate measure of the real size or temperature of vegetation burning – even combining information at different wavelengths.

2.2. Basic Concepts

After identifying the factors, levels and the response variable, the models that each factor were analyzed in the response variable using the graphs.

To verify the degree of dispersion of the data set, the standard deviation and the mean of each factor were calculated according to the fire risk level.

In order to verify if there is normality in the data, the *Anderson-Darling* test was applied, which allows the comparison of a large number of samples, and a great test power [Razali et al. 2011]. In order to attempt to normalize the data, the transformations of X Squared, cubic, Log (disregarding the values 0) and Box-Cox were applied. As it is not a normal distribution, it was not possible to use ANOVA analysis of variance for the data set, as there is a great risk of type I error. With independent and non-parametric samples, the *Kruskal-Wallis* test was used, which is an extension of the *Mann-Whitney* test [McKight and Najab 2010], in which *Wilcoxon* rank sum tests are performed in order to observe the median difference between groups for each factor. The *Conover-Iman* test was used to help analyze specific pairs of samples for stochastic dominance in post hoc tests [Conover 1999], as the *Kruskal-Wallis* only indicates that some sample has a significant difference, but does not tell which one. *Spearman's* correlation test is used because the measurements are not parametric. As for its interpretation, the recommendation of the power of correlation [Rumsey 2019] below was adopted:

- **Exactly –1.** A perfect downhill (negative) linear relationship;
- **–0.70.** A strong downhill (negative) linear relationship;
- **–0.50.** A moderate downhill (negative) relationship;
- **–0.30.** A weak downhill (negative) linear relationship;
- **0.** No linear relationship;

- **+0.30.** A weak uphill (positive) linear relationship;
- **+0.50.** A moderate uphill (positive) relationship;
- **+0.70.** A strong uphill (positive) linear relationship;
- **Exactly +1.** A perfect uphill (positive) linear relationship.

A linear regression prediction model is tested on the fit of the response variable. The multicollinearity test to verify that the model includes several factors that are correlated not only with its response variable, but also with each other. Large values are candidates that can be considered in model elimination.

3. Results

3.1. Descriptive analysis

In the period of the first five months of 2022, with a total of 6197 samples, the dataset was adjusted with the five Fire Risk levels (see Table 1). The mean (see Table 2) and the standard deviation (see Table 3) were calculated from the factors at each level.

Table 1. Summary (RiskFire)

daywithoutrain	precipitation	riskfire	frp	RiskFireLevel
Min. : 0.000	Min. : 0.000	Min. :0.0000	Min. : 0.0	Minimum : 642
1st Qu.: 1.000	1st Qu.: 0.000	1st Qu.:0.4000	1st Qu.: 3.5	Low : 879
Median : 3.000	Median : 0.000	Median :0.6000	Median : 8.4	Medium :2552
Mean : 4.379	Mean : 1.094	Mean :0.5594	Mean : 22.6	High :1210
3rd Qu.: 8.000	3rd Qu.: 1.220	3rd Qu.:0.8000	3rd Qu.: 19.9	Critical: 914
Max. : 16.000	Max. :30.400	Max. :1.0000	Max. :1425.3	

Table 2. Mean by level

	Minimum	Low	Medium	High	Critical
daywithoutrain	2.193146	5.293515	3.725313	5.001653	6.036105
precipitation	4.6919938	1.9950853	0.5864851	0.2950413	0.1744639
frp	11.00639	16.27838	24.72206	27.02777	25.06729

Table 3. Standard Deviation by level

	Minimum	Low	Medium	High	Critical
daywithoutrain	2.273105	3.010663	3.598556	3.942095	4.200846
precipitation	6.6983019	1.6775555	1.0578752	0.6549057	0.4585032
frp	18.96525	38.34785	65.12887	70.41697	58.33218

The histograms below show asymmetry in the frequency of occurrence of the Fire Risk values (Figure 1), of days without rain (Figure 2), precipitation values (Figure 3), and Fire Radiative Power (Figure 4), demonstrating a non- normality in the data, confirmed by the *Anderson-Darling* test, presenting a p-value smaller than 2.2e-16.

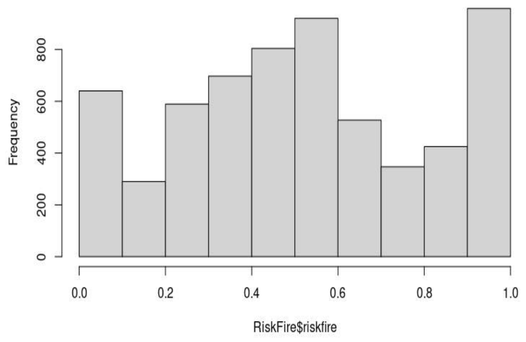

Figure 1. Frequency Risk Fire

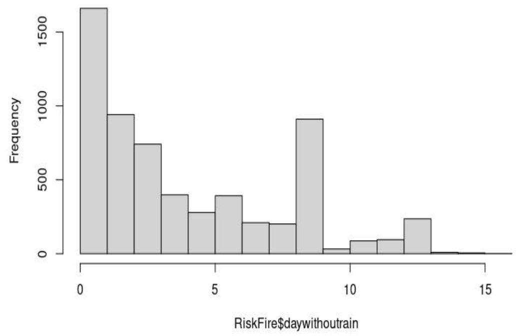

Figure 2. Frequency Day Without Rain

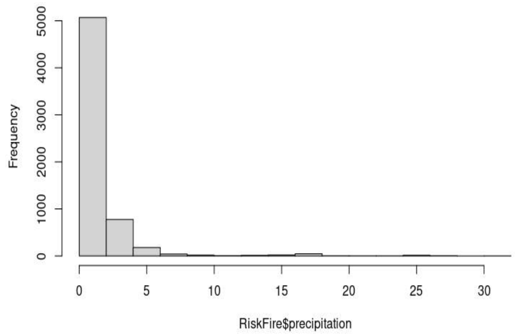

Figure 3. Frequency Precipitation

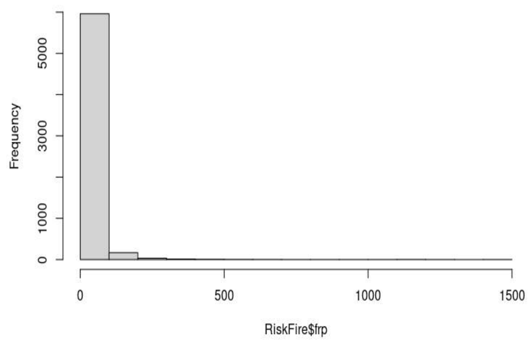

Figure 4. Frequency Fire Radioative Power

3.2. Statistical analysis

With the Boxplot graphs, it is possible to observe the variation of the data observed in the comparison between the Fire Risk levels with the other analyzed factors, ratifying the means provided (see Table 2) in addition to providing the visual location of the dispersion, symmetry, tails and outliers. All, except the risk fire value (Figure 9) show asymmetry in relation to the median indicated by the graph. Outliers appear in all of them, but with a considerable occurrence for precipitation (Figure 6) and Fire Radiation Power (Figure 7). For empirical reasons, where the hypothesis of the precipitation value could theoretically be linked to the number of days without rain, a comparison graph was generated (Figure 8), which shows the largest number of outliers. The Risk Fire value was also compared with the Risk Fire Levels (Figure 9), in which it demonstrated a proximity of symmetry in the medium and high level.

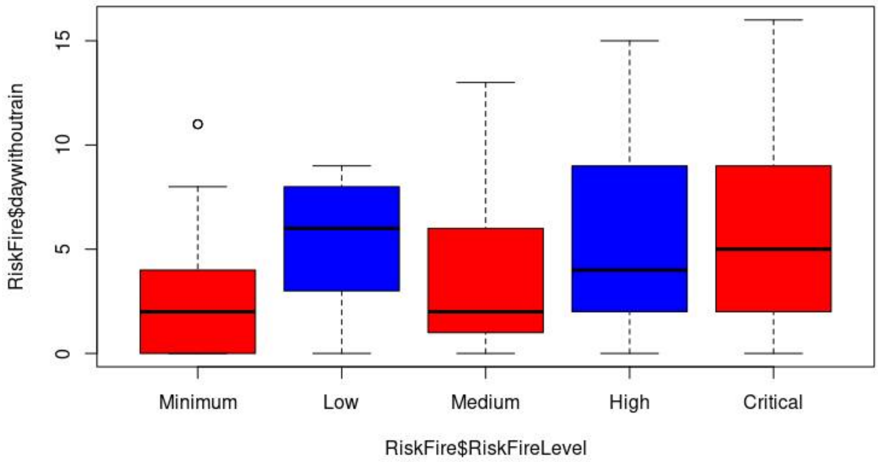

Figure 5. Comparison: Risk Fire Levels versus Days without rain

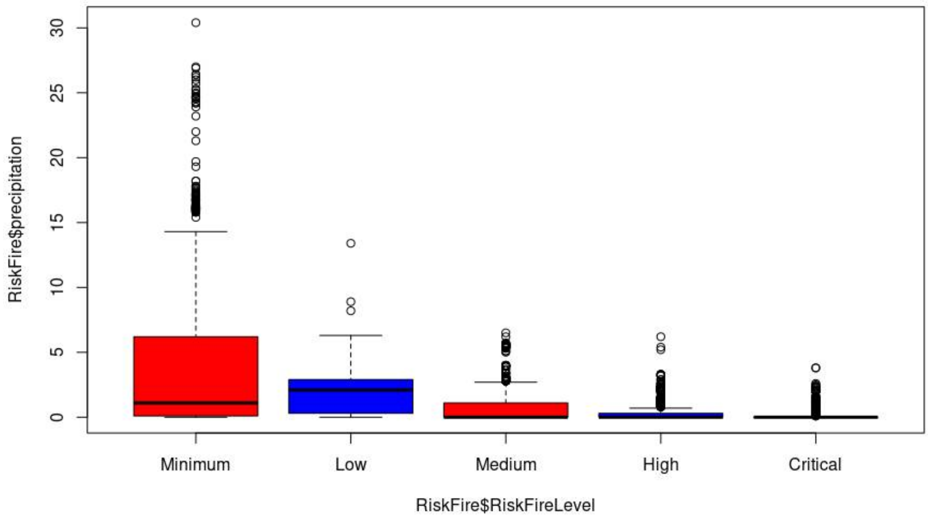

Figure 6. Comparison: Risk Fire Levels versus Precipitation

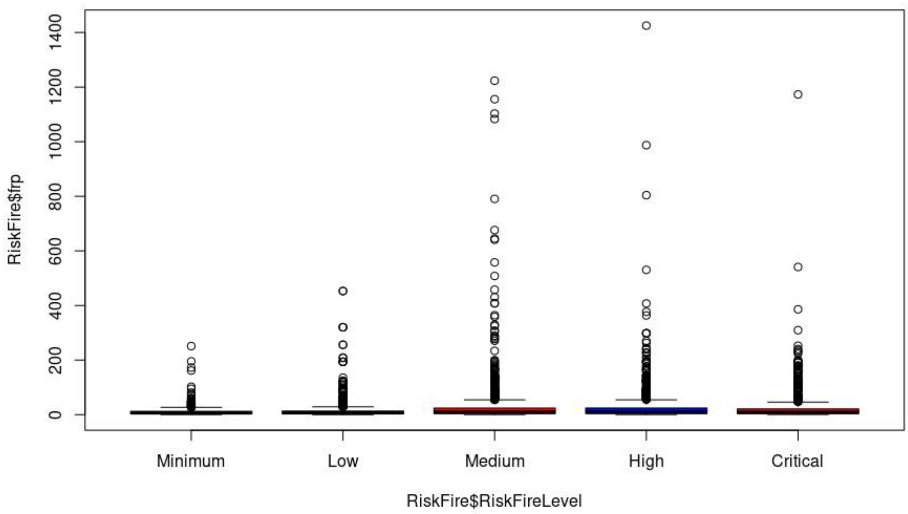

Figure 7. Comparison: Risk Fire Levels versus Fire Radiotative Power

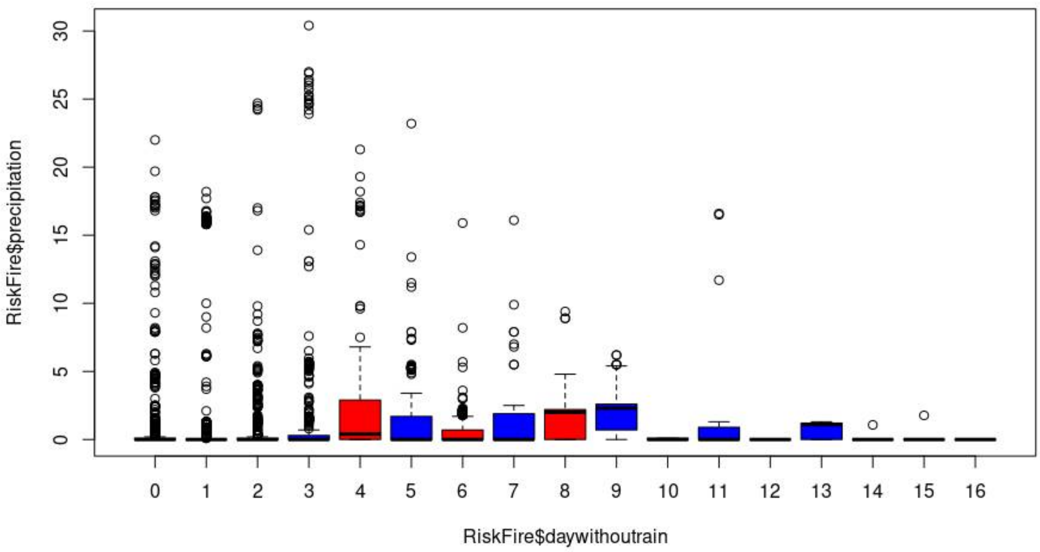

Figure 8. Comparison: Days without rain versus Precipitation

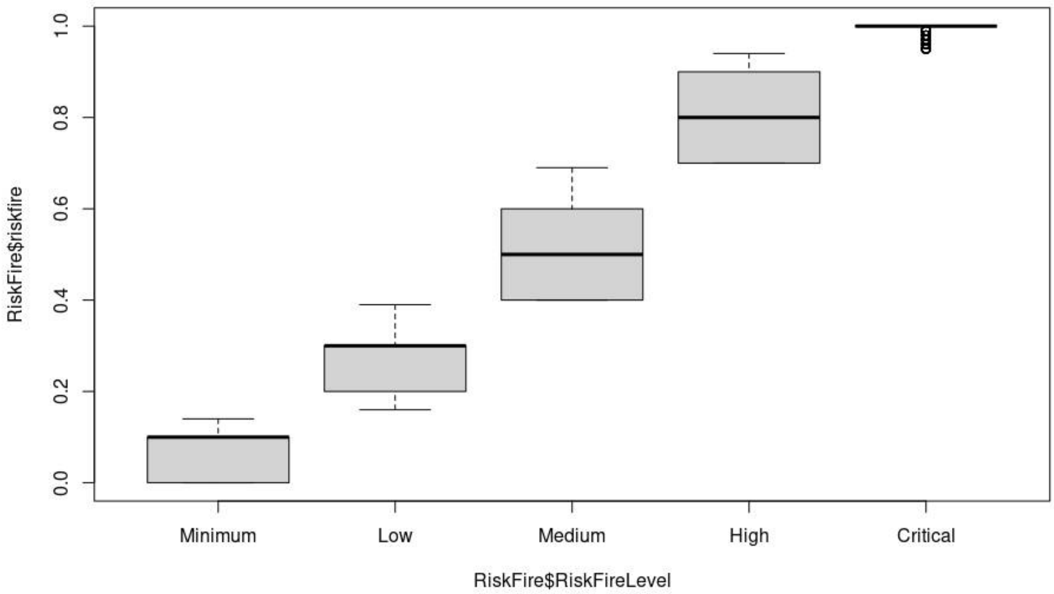

Figure 9. Comparison: Risk Fire Levels vesus Value Risk Fire

In an attempt to normalize the data, transformation methods of X Squared, cubic, Log (disregarding the values 0) and Box-Cox were used, but none showed satisfactory results, as the *Anderson-Darling* normality test resulted in all of them the smallest p-value 2.2e-16.

Using the non-parametric *Kruskal-Wallis* test, it was possible to compare samples from different groups, which demonstrated the significance of the Risk Fire Levels between the factors days without rain, precipitation and Fire Radiation Power, with a p-value lower than 2.2e- 16, rejecting the null hypothesis that the groups have the same median. To deepen and verify through the detailing of where the difference in the medians occurs, with the Conover-Iman test it is possible to carry out the identification.

In all factors, the differences are between the levels:

- Low and Minimum;
- Medium and (Minimum, Low);
- High and (Minimum, Low, Medium);
- Critical and (Minimum, Low, Medium, High).

With the *Spearman* correlation test, it suggests that precipitation has a stronger correlation with fire risk than the others, as the test result below shows:

Table 4. Spearman's Correlation

Factor	value	linear relationship
riskfire and daywithoutrain	0.1683654	weak uphill
riskfire and precipitation	-0.5012815	moderate downhill
riskfire and fire radiation power	0.1183027	weak uphill
daywithoutrain and precipitation	0.3059806	weak uphill
daywithoutrain and fire radiation power	-0.1006599	weak downhill
precipitation and fire radiation power	-0.0655393	weak downhill

With linear regression prediction, the best model is multilinear with interaction between factors, where the coefficient of determination (adjusted R squared) is 0.28 (largest possible), and the residual error is 0.2461. In the graph (Figure 10) it is possible to analyze the residual normality of the linear regression probability.

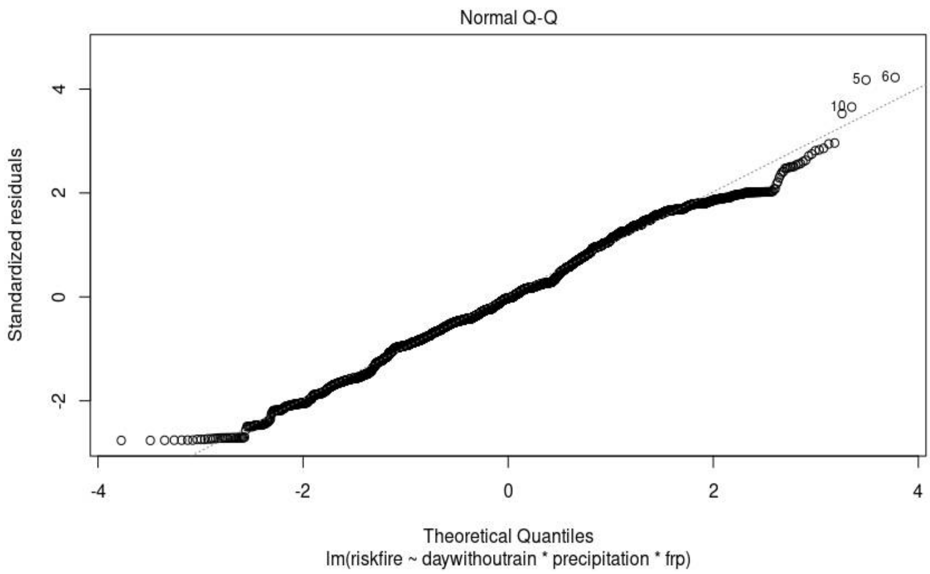

Figure 10. Linear regression: normality of residuals

Through the multicollinearity test of the generated model, the biggest candidate to be eliminated is the interaction between day without rain, precipitation and fire radiation power, with the Variation of Inflation Factor (VIF) of 8.293420.

4. Conclusion

Seeking to identify the main factors that influence the increase in risk fire, it is concluded that in the Pantanal region, specifically in Corumbá-MS city, the number of days without rain until the detection of the outbreak, precipitation accumulated value on the day until the moment of focus detection and Fire Radiotive Power are independent but have significance. Precipitation showed the highest correlation with risk fire levels, where the higher the precipitation value, the lower the risk fire. The number of days without rain showed a weak uphill correlation with precipitation. The best model discovered in the prediction of linear regression was the multilinear one with interactions between the factors, with lower residual error and higher coefficient of determination.

References

INPE. Banco de Dados de queimadas. Disponível em: http://www.inpe.br/queimadas/bdqueimadas. Acesso em: 13 de jun. 2022

Razali, N. M., Wah, Y. B., et al. (2011). Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. Journal of statistical modeling and analytics, 2(1):21–33.

McKight, P. E. and Najab, J. (2010). Kruskal-wallis test. The corsini encyclopedia of psychology, pages 1–1.

Conover, W. J. (1999) Practical Nonparametric Statistics *Wiley*, Hoboken, NJ. 3rd edition

Rumsey, D. J. (2019) How to interpret a correlation coefficient r, Dummies, (nd). Retrieved from https://www.dummies.com/education/math/statistics/how-to-interpret-a-correlation-coefficient-r/. Acesso em: 13 de jun. 2022