







Point Process Analysis applied to fire outbreaks in São Paulo

Vinicius Crispim Tavares Monteiro^{1 *} 
Manoel Alves de Oliveira² 
Ricardo Alves de Olinda³ 
Lucas Cardoso Pereira⁴ 

^{1 2 3 4 5} Departamento de Estatística, State University of Paraíba, Campina Grande, Brazil
Emails: ¹ vinicius.monteiro@aluno.uepb.edu.br; ² manoel.alves.oliveira@aluno.uepb.edu.br;
³ prof_ricardo@servidor.uepb.edu.br; ⁴ lucascardoso@servidor.uepb.edu.br
* corresponding author

How to cite this paper: Monteiro, V. C. T., de Oliveira, M. A., de Olinda, R. A. Pereira, L. C. (2024). Point Process Analysis applied to fire outbreaks in São Paulo. *Socioeconomic Analytics*, 2(1), 164-174. <https://doi.org/10.51359/2965-4661.2024.265082>

RESEARCH ARTICLE

Socioeconomic Analytics
<https://periodicos.ufpe.br/revistas/SECAN/>
ISSN Online: 2965-4661

Submitted on: 25.11.2024.
Accepted on: 15.12.2024.
Published on: 27.12.2024.

Copyright © 2024 by author(s).

This work is licensed under the Creative Commons Attribution International License CC BY-NC-ND 4.0
<http://creativecommons.org/licenses/by-nc-nd/4.0/deed.en>



Abstract

This study applied spatial analysis techniques to point processes to investigate the distribution of fire outbreaks in the state of São Paulo during the months of June to September 2024. Using data from the Burning Program of the National Institute for Space Research (INPE), descriptive and inferential analyses were performed, including Ripley's K Function and the Kolmogorov-Smirnov Test, with the aim of testing the hypothesis of Complete Spatial Randomness (CSR). The results indicated a pattern of spatial clustering in short distances, rejecting the hypothesis of randomness. Predictive models based on Poisson processes were adjusted, highlighting the most vulnerable areas, especially in the Atlantic Forest and Cerrado biomes. This work reinforces the importance of spatial statistics as an essential tool for identifying patterns and planning mitigation strategies, contributing to environmental preservation and combating fire outbreaks.

Keywords

Point Process Analysis, Spatial Statistics, Fires, Hypothesis Tests, São Paulo.

1. Introduction

Fire outbreaks are occurrences that have a significant impact on ecosystems, the economy and human health. They can be caused due to natural factors, such as lightning in dry forests, but are often the result of human activities, including agriculture, cattle ranching, and deforestation. These fires disrupt the rainfall cycle and contribute to long-term climate change by affecting temperature and precipitation (IBAMA, 2024).

Fire monitoring and control measures are essential for environmental preservation. Agencies such as the National Institute for Space Research (*Instituto Nacional de Pesquisas Espaciais* - INPE), in Brazil, use satellites to detect hot spots and fires in real time, providing important data for the performance of brigades and the application of environmental policies (AMBIENTE, 2024).

Spatial statistics is an area of statistics dedicated to the study of data that has an associated geographic location. This field is applied in various areas, such as ecology, epidemiology, geology, urban planning, among others. Spatial statistics works with the analysis of patterns, distribution, and spatial dependence on data, and is fundamental to understand phenomena that occur in a continuous or discrete space (CRESSIE, 1993).

Within spatial statistics there are point processes, which are used to model the distribution of points in a given region of space. These events may represent, for example, the location of fire outbreaks, the occurrence of a disease in individuals, or the distribution of species in a habitat. A point process is a collection of events that occur in an area or volume and can be described by its intensity and its spatial dependence properties (DIGGLE, 2013).

When analyzing point processes, some hypotheses are formulated to describe the distribution of points in space. The Complete Spatial Randomness (Completa Aleatoriedade Espacial - CSR) hypothesis assumes that events are randomly distributed in a space, so that any point has the same probability of occurrence at any location in the study area. An example of a process that accepts this hypothesis is the homogeneous Poisson process. CSR is often used as a point of comparison to determine if there is an underlying structure in the data (BADDELEY, RUBAK, & TURNER, 2015).

To test the hypothesis of Complete Spatial Randomness (Completa Aleatoriedade Espacial - CSR) in point processes, several functions can be used. These functions evaluate whether the points are randomly distributed in space or whether there is some kind of underlying structure, such as agglomeration or dispersion of the data. Some of the most used functions are: Ripley's K function; Nearest Neighbor G function; Function F of Empty Space (BADDELEY, RUBAK, & TURNER, 2015).

Based on this, the objective of this work is to carry out a spatial analysis in specific processes based on the behavior of fire outbreaks that occurred in the state of São Paulo between the months of June and September of the year 2024. As specific objectives we have: To test the hypothesis of Complete Spatial Randomness (CSR) and to investigate whether the data correspond to a homogeneous Poisson process; verify whether the distribution of fire outbreaks in the state follows a random, systematic or grouping pattern; adjust a model for

predicting fire outbreaks in the state; to ascertain which regions of the state were most affected by fires during the study period.

2. Materials and Methods

The database under study comes from the Queimadas Program of the National Institute for Space Research (*Instituto Nacional de Pesquisas Espaciais* - INPE), a page that offers real-time data on fire outbreaks in Brazil (INPE, 2024). It consists of 12 variables and 4819 observations, and corresponds to the fires that occurred in the state of São Paulo between the months of June and September 2024, was requested on the website in csv format and received by email in a compressed folder. Table 1 shows the variables contained in the database and their respective classifications.

Table 1: Table of variables and their classifications

Variables	Classifications
DataHora	Continuous quantitative
Satelite	Nominal qualitative
Pais	Nominal qualitative
Estado	Nominal qualitative
Municipio	Nominal qualitative
Bioma	Nominal qualitative
DiaSemChuva	Discrete quantitative
Precipitação	Continuous quantitative
RiscoFogo	Continuous quantitative
Latitude	Continuous quantitative
Longitude	Continuous quantitative
FRP	Continuous quantitative

Source: Prepared by the author (2024).

To perform the analysis, some descriptive statistics techniques will be applied in order to observe the behavior of fire outbreaks in the state of São Paulo. Then, spatial statistics methods will be used to evaluate and understand, through maps, the occurrences of fire in the state, as well as to ascertain possible indications of clustering in the distribution of the outbreaks. In addition, non-parametric tests will be applied to confirm the existence of any pattern in the distribution of fires, as well as a model adjustment for the prediction of fire occurrences during the study period. It is worth mentioning that all the analysis was carried out through the free R Studio software, the list of commands is available at request.

2.1 Kernel Intensity Estimator

The Kernel Intensity Estimator is an essential tool in spatial statistics to estimate the intensity of points over an area, i.e., to measure the “density” of events in different parts of space. It is a smoothing technique that calculates the density of points around each location, providing a continuous estimate of the intensity of the process at any point in space. This method

involves choosing a kernel function and a bandwidth that determines the radius of influence of each point on the surrounding area (DIGGLE, 2013).

From the concepts presented, suppose that u_1, \dots, u_n are locations of n observed events in a region A and that u represents a generic location whose value we want to estimate. The intensity estimator is computed from the m events $\{u_i, \dots, u_{i+m-1}\}$ contained in a radius of size τ around u and the distance d between the position and the i th sample, from functions whose general form is (CÂMARA & CARVALHO, 2004):

$$\hat{\lambda}_\tau(u) = \frac{1}{\tau^2} \sum_{i=1}^n k\left(\frac{d(u_i, u)}{\tau}\right), d(u_i, u) \leq \tau. \quad (1)$$

This estimator is called kernel estimator and its basic parameters are: (a) a radius of influence ($\tau \geq 0$) which defines the vicinity of the point to be interpolated and controls the “smoothing” of the generated surface; (b) an estimation function with properties of smoothing the phenomenon. The radius of influence defines the area centered on the point of view u which indicates how many events u_i contribute to the estimation of the intensity function λ (CÂMARA & CARVALHO, 2004).

2.2 Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov (KS) Test is a statistical test that can be applied in the analysis of point processes to compare the observed distribution of events with a theoretical distribution, such as the hypothesis of Complete Spatial Randomness (CSR). It measures the difference between the observed and expected cumulative distributions and provides a statistic that indicates whether this difference is statistically significant. If the value of the test is high, the hypothesis that the distribution of the points follows the theoretical pattern is rejected (BADDELEY, RUBAK, & TURNER, 2015).

This test is used to verify whether an observed distribution of points differs significantly from the expected distribution under CSR, and is useful for validating or rejecting the hypothesis of spatial randomness. The test statistic is:

$$D = \sup_x |F_n(x) - F(x)|, \quad (2)$$

where D is the largest difference between the observed and theoretical cumulative distributions, $F_n(x)$ is the function of empirical distribution of the observed data and $F(x)$ is the cumulative distribution function of the theoretical hypothesis (under CSR). The value of D is compared with critical values to decide whether to reject the null hypothesis (BADDELEY, RUBAK, & TURNER, 2015).

2.3 Function K of Ripley

The Function K of Ripley it is widely used to evaluate spatial patterns and test the CSR hypothesis in point processes. It calculates the average number of points found within a given distance h from each point in the process, comparing it with the expected number of points under the hypothesis of complete randomness (BADDELEY, RUBAK, &

TURNER, 2015).

For a completely random process, Function K must grow linearly with h^2 . Deviations above this line indicate a crowding trend (excess of points within the distance h), while deviations below indicate repulsion. The K -Function is useful at various scales, as it allows you to assess the presence of spatial patterns at different distance levels (DIGGLE, 2013). It is defined as:

$$\lambda K(h) = E(\# \text{ events contained at a distance } h \text{ from an arbitrary event})$$

where $\#$ is associated with the number of events, $E(\)$ is the estimation operator, and λ is the intensity or average number of events per unit area, assumed to be constant in the region (CÂMARA & CARVALHO, 2004).

An estimate of $K(h)$ is:

$$\hat{K}(h) = \frac{A}{n^2} \sum_i^n \sum_{j, i \neq j}^n \frac{I_h(d_{ij})}{w_{ij}}, \quad (3)$$

where A is the area of the region, n is the number of events observed, $I_h(d_{ij})$ is an indicator function whose value is 1 if $(d_{ij}) \leq h$ and 0 otherwise, and w_{ij} is the proportion of the circumference of the circle centered on event i that is within the region (correction due to edge effect) (CÂMARA & CARVALHO, 2004).

2.4 Modelling

Point models of intensity are a class of models in spatial statistics that focus on describing the intensity of a point process in different regions of space. The intensity $\lambda(s)$ is a function that indicates the expected density of points at a location s , and is a central element in modeling spatial patterns. These models are widely used to analyze the spatial distribution of events such as fire outbreaks, disease outbreaks, or crime in a city, where the intensity reflects the probability of events occurring in different parts of space (DIGGLE, 2013).

The intensity $\lambda(s)$ is defined as the expected rate of events per unit area at the point s , given by:

$$\lambda(s) = \lim_{\Delta s \rightarrow 0} \frac{E[N(\Delta s)]}{|\Delta s|}, \quad (4)$$

where $N(\Delta s)$ is the number of points in the region Δs , $|\Delta s|$ is the area of the region Δs and $E[N(\Delta s)]$ is the expected value of events in Δs . If $\lambda(s)$ is constant throughout the region of interest, the process is said to be homogeneous, otherwise it is inhomogeneous (BADDELEY, RUBAK, & TURNER, 2015).

There are several point models that are used in spatial statistics to describe and analyze

patterns of points in a geographic space, the most used being: Poisson process; Agglomeration Models; Repulsion Models; Specific processes scheduled. These models use as parameters and estimates the functions of intensity and spatial dependence (ILLIAN, PENTTINEN, STOYAN, & STOYAN, 2008).

The Poisson Process is the simplest model and assumes Complete Spatial Randomness (CSR), where the points are independent and have a constant probability of occurring anywhere in the study region. Some of its properties is that there is no interaction between the points (neither attraction nor repulsion) and the intensity λ is constant in the homogeneous case ($\lambda(s) = \lambda$). In the inhomogeneous case, the intensity is allowed $\lambda(s)$ vary spatially, being modeled as a function of spatial covariates $X(s)$:

$$\lambda(s) = \exp(\beta_0 + \beta_1 X_1(s) + \beta_2 X_2(s) + \dots + \beta_p X_p(s)),$$

where β_i are the coefficients associated with the covariates (ILLIAN, PENTTINEN, STOYAN, & STOYAN, 2008).

To compare homogeneous and inhomogeneous Poisson models, it is necessary to use the Akaike Information Criterion (AIC), widely used in the selection of statistical models, including point models. It balances the quality of the model's fit with its complexity, penalizing the use of many parameters. The AIC is calculated by the formula $AIC = -2 \log(L) + 2k$, where L is the verisimilitude of the adjusted model and k is the number of parameters of the model. Models with lower AIC values are preferable, as they indicate a good balance between fit and simplicity (AKAIKE, 1974).

3. Application

The state of São Paulo is composed of the characteristic plant formations of the Cerrado, which corresponds to 32.7% of the state area, and the Atlantic Forest, equivalent to 67.3%. From Figure 1, it can be seen that these biomes were affected in an equivalent way by fire outbreaks, however, it can also be observed that there is a slight difference between them, which makes the Atlantic Forest the biome most affected by fires in the state. It is worth noting that the Atlantic Forest and the Cerrado are considered hotspots (environments with high biodiversity, and highly threatened by human action in nature), as they have suffered great loss of habitats, so they have a high risk of disappearing.

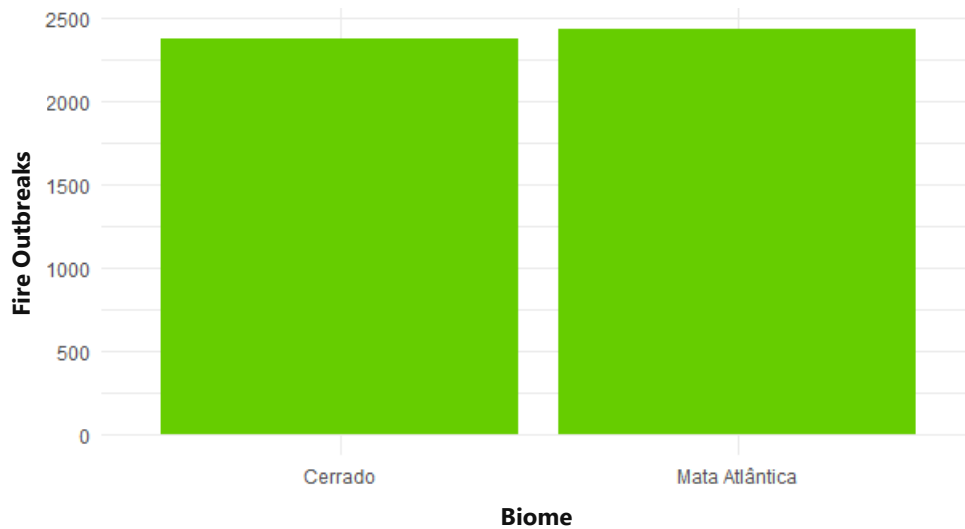
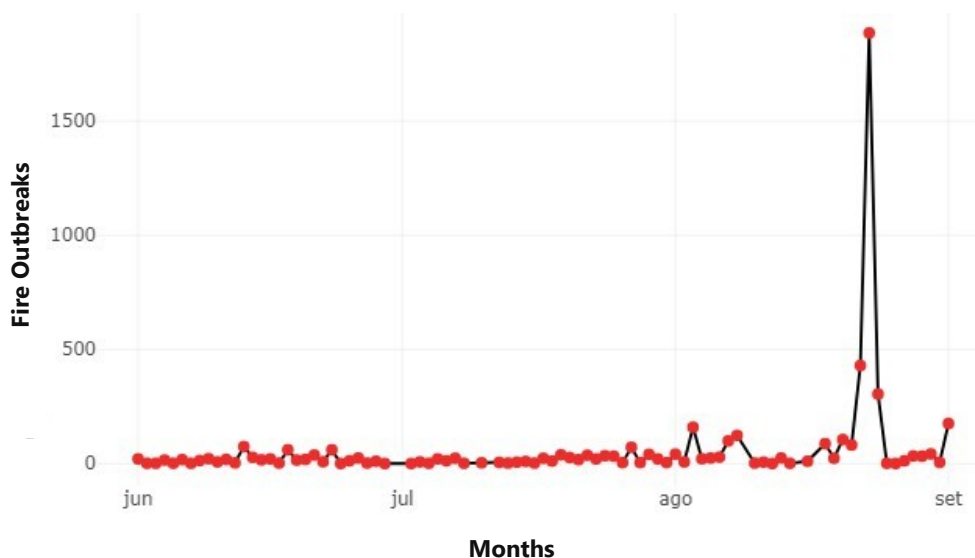
Figure 1: Bar chart for fire outbreaks in the biomes of the state of São Paulo.

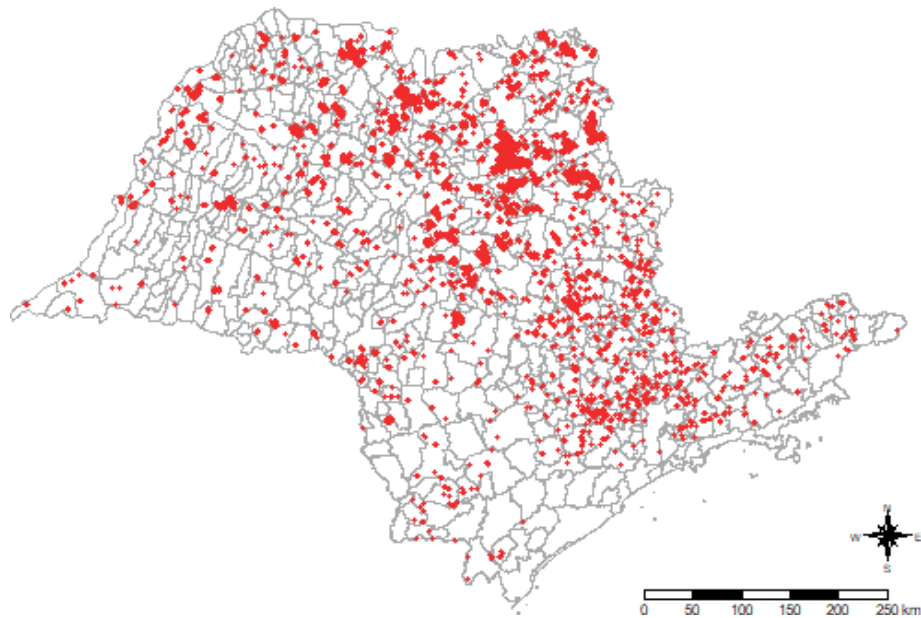
Figure 2 shows, through a time-series line graph, the number of fires that occurred during the study period, referring to the months of June to September 2024. It can be seen that in the first months (June and July) the number of fires in the state was low, with a constant occurrence of outbreaks over the days. By the end of August, a worrying increase in fire outbreaks can be observed based on the registration of more than 1500 occurrences in a single day. This increase may be linked to the dry season in the state, which lasts from May to August, with August being the month with the lowest number of days with precipitation.

Figure 2: Time series for the number of fire outbreaks during the period studied.

Based on the map in Figure 3, it is noted that the highest density of foci seems to be concentrated in the northeastern region of the state, possibly in areas of more susceptible vegetation, such as transition biomes between the cerrado and forests. In addition, the presence of outbreaks spread throughout almost the entire state is noticeable, however,

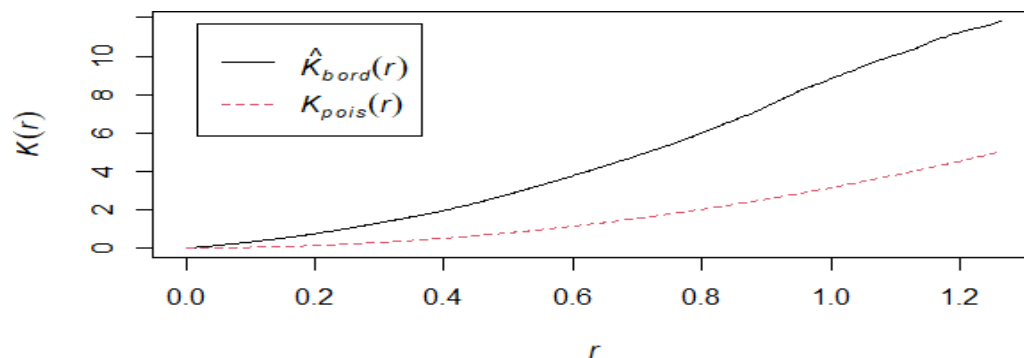
some areas have a lower incidence, such as the metropolitan region of São Paulo, which is more urbanized. Many of these occurrences may be related to agricultural practices, such as fires to clear land or expand arable areas.

Figure 3: Map for the distribution of fire outbreaks in the state of São Paulo.



From Figure 4 that shows the graph of Ripley's K function, it can be seen that the observed function ($K_{bord}(r)$) is above the reference line of the random pattern ($K_{pois}(r)$). This indicates that the fire outbreaks are aggregated on small scales, that is, there are clusters of fires in proximity. In addition, it can be seen that as the distance increases, the black line remains above the red line, suggesting that the aggregation pattern is maintained at different spatial scales. Thus, the distribution of fire outbreaks is not random, as it shows a pattern of grouping.

Figure 4: Graph of Ripley's K function.



Based on Figure 5, it can be seen that the observed function ($G_{obs}(r)$) deviates considerably from the theoretical function ($G_{theo}(r)$), especially for short distances. This suggests that the spatial distribution of fire outbreaks does not follow a completely random pattern. In addition, it is noted that the black line ($G_{obs}(r)$) exceeds the confidence interval limits ($G_{hi}(r)$ and $G_{lo}(r)$), indicating that the hypothesis of Complete Spatial Randomness (CSR) should be rejected. With this, there is evidence to believe that fire outbreaks follow a pattern of spatial grouping, especially at short distances.

Figure 5: Graph of the Kolmogorov-Smirnov test for complete spatial randomness.

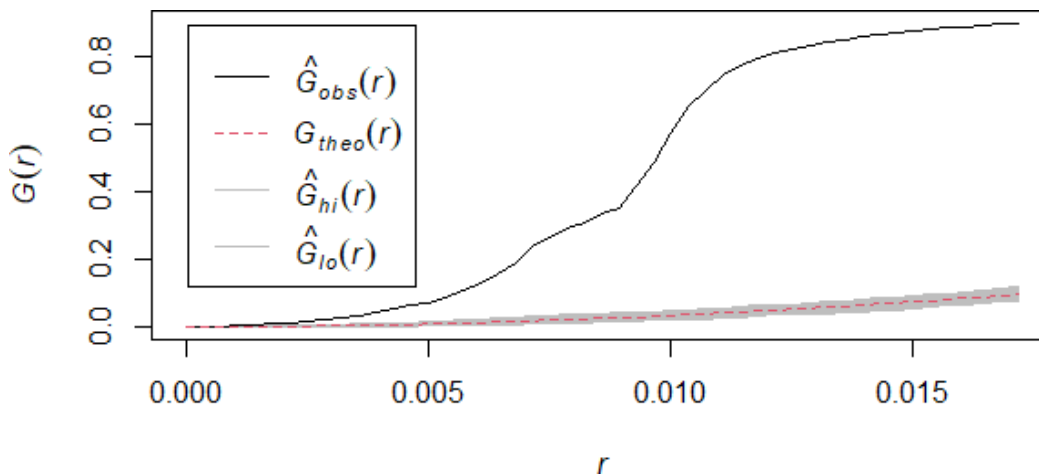


Table 2 shows that models 1 and 2 have similar ICAs, suggesting that both have almost identical performances in terms of fit and complexity. In addition, it can be observed that model 3 has the lowest AIC value, indicating that it presents the best fit among the models evaluated, despite being potentially more complex.

Table 2: AIC information criterion of the adjusted models

Model 1	Model 2	Model 3
-35572.290	-35572.580	-37413.860

Based on Table 3, it can be seen that the estimated coefficients for the intercept and y are statistically significant and have a substantial and precise impact on the dependent variable. The intercept represents the expected value of the number of fire outbreaks when all explanatory variables are equal to zero. In this case, when $y = 0$, an average of about 14,610 fires are expected.

The coefficient of y indicates that for each increase of one unit in the variable y , the expected number of fire outbreaks increases by 0.450. The positive relationship between y and fire outbreaks suggests that as y increases, fire outbreaks also tend to increase. This can be relevant for prevention planning, especially at critical times of the year, such as dry seasons.

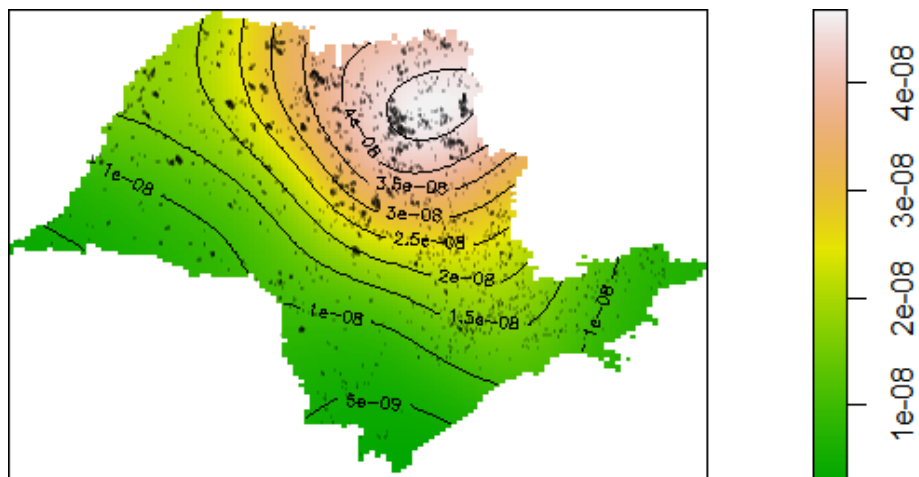
Table 3: Model 3 metrics

Parameter	Estimate	Standard Error	Lower 95% CI	Upper 95% CI	Z Value
Intercept	14.610	0.241	14.137	15.082	60.625***
y	0.450	0.011	0.428	0.472	40.437***

*** significant to 0.1%.

The contour map in Figure 6 shows the intensity of points per unit area, from which it can be seen that the highest concentration of fire outbreaks (areas in pink) is located in the northern region of the state. The green areas, which indicate lower fire intensity, cover the southwest and south of the state, regions possibly less affected by fire outbreaks. The isolines represent constant levels of focus intensity.

The closer these lines are to each other, the greater the variation in intensity in a smaller area, which indicates a more localized concentration of the foci. This spatial distribution may be related to factors such as: Dry climate or seasonality in certain areas; land use, such as the presence of agricultural or forest areas; proximity to urban regions, which may have a lower incidence due to less vegetation cover.

Figure 6: Contour map for the intensity of points per unit area.

4. Conclusion

Based on the objectives outlined, it is possible to conclude that the spatial analysis of the fire outbreaks that occurred in the state of São Paulo during the period from June to September 2024, provided important subsidies to understand the patterns of occurrence of these events.

Initially, the descriptive analysis showed that the Atlantic Forest and Cerrado biomes, highly threatened, were the most affected, with the northeastern region of the state being particularly vulnerable. In addition, statistical tests and spatial modeling confirmed that the distribution of fire outbreaks does not follow a random pattern, rejecting the hypothesis of

Complete Spatial Randomness (SRF).

The use of tools such as Ripley's K function and the Kolmogorov-Smirnov test showed a pattern of clustering over short distances, suggesting the presence of underlying factors that favor the concentration of fires in certain regions. The adjustment of the models indicated the ability to predict areas more prone to fires, a valuable contribution to mitigation strategies and allocation of combat resources.

Therefore, the study achieved its objectives by identifying distribution patterns, testing hypotheses, and adjusting predictive models, reinforcing the importance of spatial statistics as an essential tool for understanding and managing fire outbreaks. The findings can support more effective public policies and preventive actions in the context of environmental preservation.

References

1. Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6), 716-723.
2. Ambiente, M. D. (2024). *Monitoramento dos Focos de Queimadas*. Fonte: INPE: <https://queimadas.dgi.inpe.br/queimadas/portal>
3. Baddeley, A., Rubak, E., & Turner, R. (2015). *Spatial point patterns: methodology and applications with R*. CRC press. CÂMARA, G., & CARVALHO, M. S. (2004). *Análise espacial de eventos*. Brasília: EMBRAPA.
4. Cressie, N. (2015). *Statistics for spatial data*. John Wiley & Sons.
5. Diggle, P. J. (2013). *Statistical analysis of spatial and spatio-temporal point patterns*. CRC press.
6. IBAMA. (2024). *Queimadas e incêndios florestais*. Fonte: Governo Federal: <https://www.ibama.gov.br/queimadas>
7. Illian, J., Penttinen, A., Stoyan, H., & Stoyan, D. (2008). *Statistical analysis and modelling of spatial point patterns*. John Wiley & Sons.
8. INPE. (2024). *BDQUEIMADAS*. Fonte: terrabrasilis: <https://terrabrasilis.dpi.inpe.br/queimadas/bdqueimadas/#exportar-dados>