Sentiment Analysis in Social Medias for Threats Prediction of Natural Extreme Events

S

Marilyn Minicucci Ibañez
National Instituto of Spatial Research, Federal Institute of São Paulo, Brazil

Reinado Roberto Rosa
Lab for Computing and Applied Mathematics, Brazil

Lamartine Nogueira Frutuoso Guimarães
National Instituto of Spatial Research, Brazil

INTRODUCTION

The end of the twentieth century was marked by the advent of the internet and consequently this favored those massive volumes of information, from the most different fields of knowledge, to circulate through the World Wide Web – WWW (Castells, 2003). The disclosure and sharing of this information by society introduced new understandings of how this volume of data could be used to generate value for the most diverse areas of knowledge and thus bring benefits to society. Among the most varied domains of information circulating on the WWW, natural extreme events deserve attention for a more detailed study and understanding of their causes, consequences, and possible prevention.

An extreme event is characterized by a sequence of small events generated by human emotions or some reaction of nature that can evolve into a larger event reaching up to a catastrophic event (Rosa et al., 2019) (Clauset, 2018) (Ibanez et al. 2022). The natural extreme event model considered is deforestation, as it has a great influence on the life of society (Santos et al., 2017). Due to its complexity, a multidisciplinary solution would assist in understanding its evolution and possible prevention of this natural extreme event model.

Based on the context presented, this work proposes a multidisciplinary solution that considers the threats of droughts and fires in the Brazilian Amazon region as the evolution of deforestation. To carry out the threat analysis, used as case studies for the natural extreme event data collected from social media, such as newspapers and magazines. Considered the social media of large national circulation, about the occurrence of droughts, fires, and deforestation in the years 2015, 2016, 2017, 2018, 2019 and 2020. The collection of this information is carried out using Google's web search engine (Google, 2016) which performs a search for news related to the topics addressed about drought, burning, and deforestation threats. Each collected news is stored and grouped considering the increasing order of its publication date (Ibanez et al., 2022).

The news collected from social media are processed using data science and machine learning techniques that allow identifying some nature reaction present in a text document. As per the context of the chapter, the reaction identified in the analyzed news texts is the threat of the natural extreme event addressed. The machine learning technique used for the identification of the threat in the news is sentiment analysis, being applied in the chapter with the Natural Language Processing technique, which performs

DOI: 10.4018/978-1-6684-7366-5.ch046

text analysis (Ibanez et al., 2022). Sentiment analysis makes it possible to identify how similar a text is to a given context, using a base text with words referring to a domain (Bird et al., 2009), in this case, the threat of the extreme event. Thus, for each news story collected, the percentage of threat existing in its text is analyzed and identified (Ibanez et al., 2022).

The result of this collection is a threat portfolio with time series with the threat degree referring to these extreme events. This portfolio is used as input for the P-model algorithm (Rosa et al., 2019) to generate a time series with endogenous characteristics. This time series model is characterized by considering only elements that exert some internal influence on the analyzed domain (Rosa et al., 2019) (Sornette, 2006). The generated endogenous series is used as input to a system developed with machine learning, more specifically deep learning, for the creation of the extreme event prediction model. The prediction of the evolution of the threat of occurrence of drought, fire, and deforestation is being carried out for a future period of three months (Ibanez et al., 2022).

The deep learning technique is being used due to its properties of automatically extracting features and nonlinear correlations existing in the data (Goodfellow, et al., 2016). Deep learning concepts are applied using a neural network framework that uses Long Short-Term Memory – LSTM recursive neural networks in a system developed using the TensorFlow Keras deep learning API (Chollet, 2015). LSTMs are suitable for classifying, processing, and predicting time series due to the unknown length delays of some series (Hochreiter, Schmidhuber, 1997).

Validation of the prediction is performed using the statistical tool, DTW (Dynamic time warping) that compares the predicted period with the first three months of the subsequent year. Proof of the endogeneity pattern is performed by an analysis of the endogenous time series generated by the threats using the statistical tools of mean, variance, standard deviation, skewness, and kurtosis.

BACKGROUND

This section presents the study of state of the art related to sentiment analysis and prediction of natural extreme events considering the droughts, fires, and deforestation.

The article Rainfall prediction for Manaus, Amazonas with artificial neural networks (Lima; Guedes, 2015) addressed the problem of rainfall prediction in Manaus using multilayer artificial neural networks. The input data were obtained from an automatic weather station during the years from 1970 to 2015. The performance factor considered was the normalized root-mean-square error. According to the observed results, a feedforward neural network with 2 hidden layers with 10 neurons each performed best in solving the issue. In the work, it was also observed that the use of recurrent neural networks had no influence on the performance gain of the problem addressed.

In Predicting amazon fires for policymaking (Morello et al., 2016) presented contributions with public policies for fighting fires in the Amazon. The work focused on identifying at the municipal scale the main variables for predicting fire occurrences. As a result, a data panel unprecedented in the literature was built from satellite images and socioeconomic data, covering the years 2008, 2010 and 2012. The work concluded that of the 41 potential fire predictors evaluated, only 9 were significant at a tolerable level of uncertainty, comprising deforested areas, pasture and forest areas, indigenous lands, temperature, and soil texture.

The paper Integrating remotely sensed fires for predicting deforestation for redd (Armenteras et al., 2017) presented work in which it addresses that the United Nations Reducing Emissions from Deforestation and Forest Degradation (REDD+) program leaves a gap in decision-making about REDD+ interventions.

This gap reference by failing to systematically include fires in its data. The paper addressed this critical knowledge gap in two ways. First, it reviewed REDD+ projects and programs to assess the inclusion of fires in monitoring, reporting, and verification (MRV) systems. Second, it modeled the relationship between fire and forest for a pilot site in Colombia using near-real-time (NRT) fire monitoring data derived from Moderate Resolution Imaging Spectroradiometer (MODIS). The model-based deforestation predictions performed better than the official REDD early warning system. The AUC of the model for 2013 and 2014 was 0.81, compared to 0.52 for the early warning system in 2013 and 0.68 in 2014. This demonstrated that NRT fire monitoring is a powerful tool to predict locations of forest deforestation.

The publication Deforestation prediction using neural networks and satellite imagery in a spatial information system (Ahmadi, 2018) investigated the spatial distribution of deforestation using artificial neural networks and satellite imagery. The article presented the modeling of land cover changes (forests) to predict deforestation using an artificial neural network Multilayer Perceptron due to its significant potential for developing complex nonlinear models. The procedure involved image registration and error correction, image classification, preparation of deforestation maps, layer determination, and design of a multilayer neural network to predict deforestation. The satellite images for this study are from a region of Hong Kong captured from 2012 to 2016. The results of the study strongly suggested that the neural network approach to predict deforestation can be used, and its results showed the areas that were destroyed during the research period. It was pointed out that due to climatic, economic, and political factors the prediction of deforested areas is difficult to accurately predict.

The paper Topic modeling and sentiment analysis of global climate change tweets (Dahal et al., 2019) addressed the use of social media data for the verification of climate change in a certain location. In the work, data from the social network Twitter was used with geotags that made it possible to identify the location, date, and time of the messages. To perform the data classification, it used the techniques of Natural Language Processing, such as: sentiment analysis and Latent Dirichlet Allocation (LDA). As a result, it is observed that sentiment analysis indicates that the overall discussion is negative, especially when users are reacting to extreme weather or political events. Topic modeling shows that the different topics of climate change discussion are diverse, but some topics are more prevalent than others.

The paper Deforestation prediction using time series and LSTM (Gao, 2019) used time series techniques and LSTM neural network, to predict future deforestation of the Amazon rainforest. The calculations revealed an urgent need to act to prevent further deterioration. It was discussed in the paper that the help of advanced AI techniques will reduce deforestation efficiently, improving the environment in the foreseeable future.

The paper Predicting the deforestation probability using the binary logistic regression, random forest, ensemble rotational forest, reptree: A case study at the gumani river basin, India (Saha et al., 2020) used the coupling of binary logistic regression (BLR), Random Forest (RF), ensemble rotational forest and reduced error pruning trees (RTF-REPTree) with DCF to find out the deforestation probability. In the work, Advanced Vegetation Index (AVI), Bare Soil Index (BSI), Shade Index (SI) and Scaled Vegetation Density (VD) derived from Landsat images were used as the main input parameters to identify the DCF. These deforestation probability models were validated using area under the curve (AUC), receiver operating characteristics (ROC), efficiency, true skill statistics (TSS) and Kappa coefficient. The validation result indicated that all the models like BLR (AUC = 0.874), RF (AUC = 0.886) and RTF-REPTree (AUC = 0.919) had good ability to evaluate the probability of deforestation, but among them, RTF-REPTree had the highest-level accuracy.

The article Sentiment Analysis in Tweets Related to the Deforestation of the Amazon Rainforest (Paes et al., 2022) presents an analysis of the feelings of Brazilian Twitter users related to the deforestation

of the Amazon Rainforest through Twitter text mining and seeks to understand how Brazilians opine and dialogue about the deforestation of the Amazon Rainforest. The results reveal that Brazilian users tend to react to events related to the deforestation of the Amazon Rainforest on Twitter and that, for the most part, users have negative feelings about the topic, reaching peaks of approximately 60% of tweets at any given time.

FOCUS OF THE ARTICLE

Analysis of Brazilian Social Media Using Data Science Concepts

Social media have long been characterized as one-way media models, such as magazines, newspapers, and radios. With the increased use of the internet, the concept of social media also came to encompass all applications and theories that are based on the internet, on Web 2.0, and that also allow the sharing of information among users (Ibanez et al., 2022). The particularities of the data shared by social media have brought the need for the use of new techniques from the field of data science, such as social media mining. This model of data mining uses the theories from various areas of knowledge, such as computing, machine learning, ethnography, sociology, statistics, mathematics, among others, for a complete analysis of the large amount of information generated by these media (Zafarani et al., 2014).

In social media mining, one collects information about individuals and entities to discover some pattern existing in the interaction between these two elements (Zafarani et al., 2014). In this chapter, we are using one-way social media, such as magazines and newspapers, of large circulation in Brazil and specific social media of the environmental area, such as (Folha, 1996), (Globo, 2006), (Butler, 1999), (Lima,1994), etc. This news was collected, in Portuguese language, and in a manual empirical way, from January 2015 to October 2020 about the occurrence of drought, fires, and deforestation in the Amazon Forest region.

To carry out this process, initially collected news that present evidence of threats related to the natural extreme event analyzed and that this threat can lead to a sequence of smaller events until an endogenous extreme event results. That is, a threat can lead to a smaller event, which would lead to an extreme event (Ibanez et al., 2022). This reflection is characterized, as presented in Table 1.

Table 1. Characterization of the term's threat, event, and extreme event for natural extreme events

Term	Description
threat	mention of the possibility of drought
event	burning caused by the threat
extreme event	deforestation generated by threat and event

Table 2. Sample of some words in Portuguese language about occurrence of drought, fires, and deforestation used in the news search manually

seca	pluviosidade	desmatamento
queimada	pouca	amazonia
baixa	chuva	queima

In this chapter, the discovery of patterns in the data collected from social media are identified through the data science process called Knowledge Discovery in Databases (KDD). The steps of KDD are presented below.

- **Selection:** definition of the target data using the selected database.
- **Pre-processing:** storing the data in a database.
- **Transformation:** processing the stored data.
- **Mining:** identifying patterns in the processed and stored data.
- **Interpretation or Evaluation:** generating knowledge about the mined data.

The application of the concepts of data science and the KDD process, in this chapter, began with the data selection and the definition of the natural extreme event theme. The data preprocessing began with the storage of the collected news by developing a table structure that has Data, URL, and the Degree of Threat Similarity (DTS) as columns. The Data and URL data were filled in with the collection process in the social media. Thus, for each theme analyzed, at least one news story per month was collected from each year of the period considered in the analysis. In the data transformation step of the KDD process, the news were organized in increasing date order, that is, considering day, month and year, with the oldest news at the top of the table. This table is stored in a .csv file and illustrated by Table 3 (Ibanez et al., 2022).

Table 3. Example of the organization of news about extreme natural events (occurrence of drought, fires, and deforestation in the Amazon) stored in the .csv file

Date	URL
02/03/17	https://jornal.usp.br/desmatamentochuvasamazonia/
13/03/17	https://www.dw.com/amazoniaciclodesmatamentoseca/
18/03/17	https://crisalida.eco.br/efeitossecaperdaamazonia/
20/03/17	https://www.pirelli.com/mudancaclimatica/
29/03/17	https://www.ecodebate.com.br/amazoniasubstituidagramineas/

The next steps of KDD, data mining, interpretation or evaluation, were developed, applying the concepts of natural language processing and sentiment analysis, and generated the Degree of Threat Similarity (DTS).

Evolution of Natural Extreme Events in Social Media Sentiment Analysis

In the stage of collecting news from social media, the URL (Uniform Resource Locator) information was considered for the real-time reading of the information contained in the HTML (HyperText Markup Language) page of the magazines and newspapers. The .html file is performed only reading the information contained within the paragraph tags < p >< /p >. The result of this process is a text containing the information on the news to be analyzed (Ibanez et al., 2022).

The treatment of the information contained in the text is performed using the concepts of natural language processing through the use of the tokenization step. In this step realized the elimination of symbols

and characters that have no representation of meaning for analysis is performed, such as !, ?, $, &, etc. NLP is an area of artificial intelligence that uses natural human language to perform human-computer interaction (Jackson & Mouliner, 2002) (Ibanez et al., 2022). After the tokenization of the information, the news analysis phase begins by applying the concepts of sentiment analysis. In this chapter, topic- or feature-based sentiment analysis is used to extract information from the selected news stories. According to Ibanez et al., 2022, this analysis model is based on the verification of existing characteristics related to sentiment about the subject. In this information extraction process, initially a base text is generated, in which the concept of the threat that one would like to identify in the analyzed texts is defined. This definition represents 100% of the threat about the evolution of droughts, fires, and deforestation. This base text is constructed empirically, considering the knowledge of the people in the working group about the subject. The same tokenization process performed in the news is applied in the base text for the elimination of symbols and characters without representation. Next, we present the base text about drought, fires, and deforestation in Portuguese language used to analyze the evolution of the extreme natural event drought, fires, and deforestation.

Excerpt in Portuguese language from the base text for the natural extreme event - occurrence of droughts, fires, and deforestation

---

A temporada de incêndios de 2020 na floresta amazônica pode ser muito mais rigorosa do que em 2019 e um dos motivos dessa piora são as mesmas condições climáticas que intensificam a temporada de furacões no hemisfério norte segundo pesquisadores Em agosto passado uma série de grandes incêndios provocados pelo homem na Amazônia lançou nuvens de fumaça sobre a cidade de São Paulo transformando o dia em noite e gerou protestos internacionais Mas embora esses incêndios tenham sido incomuns e alarmantes a situação poderia ter sido ainda pior se a Amazônia estivesse em época de estiagem Mas neste ano infelizmente condições mais secas do que a média são exatamente o que está previsto para o sul da Amazônia e um dos motivos é o aumento extraordinário de calor no Atlântico Tropical Norte a milhares de quilômetros de distância Esse calor oceânico também fez com que a temporada de furacões no Atlântico batesse recordes logo de início um prenúncio das previsões para uma temporada extraordinariamente tumultuada Algumas pesquisas sugerem existir uma relação causal entre os próprios furacões e os piores anos de incêndios na Amazônia embora isso seja assunto de maior debate.

With the base text created and standardized, the percentage of similarity of the new news is identified by comparing the two texts. The result of this process is the degree of threat similarity (DTS) that each news item represents in relation to the extreme event analyzed. This threat degree is calculated for each news story accessed, and this calculation is stored in the DTS field of the .csv file. The DTS value is presented with 14 decimal places for better accuracy of the result (Ibanez et al., 2022).

All steps of the sentiment analysis process were performed using the Python 3.7 language along with the NLTK (Natural Language toolKit) APIs (Bird et al., 2009) and Embedded Keras Tensorflow (Chollet, 2015) for the tokenization step. Reaction identification in the news was performed using the SpaCy (Industrial-Strength Natural Language) library. The following is a pseudocode of the GSA calculation process, which shows application of the mentioned libraries (Ibanez et al., 2022).

Pseudo-code Calculation_DTS():

```
/*Reading of the data in the URL field in the .csv file*/
data_url = read_csv_file()
/*Loop for reading each news item with URL in the .csv*/ file
For i until tamanho(data_url):
/*Access the news URL*/
open_html_file = request.get(data_url[i])
/*Reading the contents of the .html file with the BS4 API*/
html_file = BS4(open_html_file.content)
/*Selecting the data between the <p><p/> tag*/
text = found_all('<p><p/>')
/*Elimination of meaningless characters using
the tokenization process*/
data = NLTK_tokenização(text)
/*Reading the news base text */
news_base = read_news_base()
/*Elimination of meaningless characters using
the tokenization process*/
news_base_data = NLTK_tokenização(news_base)
/*Transforms the text into a vector of words*/
nbd_nlp = nlp(news_base_data)
data_nlp = nlp(data)
/*Compares the word vectors for calculating the
similarity between texts*/
DTS = nbd_nlp.similarity(data_nlp)
/*Stores the DTS value in the .csv file*/
write_csv_file(DTS)
End For
End Pseudo-code
```

Table 4. Presents examples of the collected news stories with the Degree of Threat Similarity value (DTS) for each story

Date	URL	DTS
02/03/17	jornal.usp.br/desmatamentochuvasamazonia/	0.91
13/03/17	dw.com/amazoniaciclodesmatamentoseca/	0.85
18/03/17	crisalida.eco.br/efeitossecaperdaamazonia/	0.89
20/03/17	pirelli.com/mudancaclimatica/	0.89
29/03/17	ecodebate.com.br/amazoniasubstituidagramineas/	0.91

The degree of threat similarity (DTS) is organized in ascending date order, forming the threat time series used to create the endogenous time series, which will be presented in the next section.

Generating Endogenous Time Series Using the Degree of Threat Similarity

The time series generated by the degree of threat similarity (DTS) gave rise to the endogenous time series by applying it to the P-Model algorithm (Rosa et al., 2019). This algorithm creates an inhomogeneous cascade that is compatible with the energy dissipated by extreme events until the moment of its apex, in which the maximum energy dissipation takes place (Ibanez et al., 2022). The multiplicative cascade of the P-Model is represented by Equations 1 and 2 and is defined in Halsey et al. (1987).

Equation 1. *Representation of the Multiplicative Cascade of the P-Model*

$$\alpha = \frac{log_2 p_1 + (\omega - 1) log_2 p_2}{log_2 l_1 + (\omega - 1) log_2 l_2}$$

Equation 2. *Representation of the Multiplicative Waterfall of the P-Model (cont.)*

$$f(\alpha) = \frac{(\omega - 1) log_2 (\omega - 1) - \omega log_2 \omega}{log_2 l_1 + (\omega - 1) log_2 l_2}$$

where,

α - strength of the singularity
p_i – probability that some event occurs in the i-th fraction, for i - 1, 2.
ω – multiplicative weight given by 1 - (1 - 2p).
l_i – i-th fraction of an eddy of size L (Ibanez et al., 2022), for i = 1, 2.
f(α) - describes how the singularities are densely distributed

In practice, the P-Model algorithm uses three input parameters:

- *n – the* number of elements of the endogenous time series to be generated. The P-Model works best with values representing the power of 2.
- ***p –*** which represents the series category.
- β value (slope) – which is obtained by calculating the DFA of the threat time series generated by the DTS.

The method of DFA (Detrended fluctuation analysis) was designed by (Peng et al., 1994) to investigate long-range correlation in nonstationary series (Morariu et al., 2007). DFA, according to (Harvard, 2019), can be applied by considering the following steps.

- the time series, with N samples, to be analyzed is initially integrated
- the integrated time series is divided into bins of equal length, n
- in each box of length n, a least squares line is fitted to the data (representing the trend in that box)
- the y-coordinate of the straight-line segments is denoted by $y_n(k)$
- we decrease the integrated time series, y(k), subtracting the local trend $y_n(k)$, in each box

- the mean square root fluctuation of this integrated, trendless time series is calculated using Equation 3

Equation 3: *Formula of the Detrended Fluctuation Analysis (DFA)*

$$F(n) = \sqrt{\frac{1}{N}\sum_{k=1}^{N}\left[y(k) - y_n(k)\right]^2}$$

This calculation is repeated at all time scales (box sizes) to characterize the relationship between F(n), the mean fluctuation, as a function of box size. Typically, F(n) will increase with box size n. A linear relationship in a log-log plot indicates the presence of power-law (fractal) scaling. Under such conditions, fluctuations can be characterized by a scaling exponent, the slope of the line that relates log(F(n)) to log(n) (Harvard, 2019). In this chapter, the calculation of DFA is performed using the DFA() function from the nops library of the Python 3.7 programming language. The input of the function is the DTS threat series, generated by applying the sentiment analysis technique to the data collected from social media. The result of the DFA calculation is used as input to the P-Model algorithm for generating the endogenous threat series. For the construction of the endogenous series we used, besides the DFA value, the values of the number of points of the series, considering 16384 points, and the value of the series type, considering the parameter p = 0.60 representing the creation of an endogenous type series. The pseudocode with the idea of creating the endogenous threat time series used in this work is presented below.

Pseudo-code Series_Endogenous():

```
/*pmodel input variables declaration*/
var slope: decimal
var p: decimal
var dados_serie: string
var dados_pmodel: string
var tamanho_serie: inteiro
/*Defining the number of elements of the endogenous time series */
tamanho_serie = 16384
/*Reading threat time series data from the .csv file*/
dados_serie = decimal(read_csv_file())
/*Calculating the slope value using the DFA of the threat time series*/
slope = dfa(dados_serie)
/*Definition of the p parameter for endogenous time series*/
p = 0,60
/*Generating the endogenous time series using the P-Model algorithm*/
dados_pmodel = PModel(tamanho_serie, p, slope)
/*Storing the endogenous series in a .txt file*/
write_file(dados_pmodel)
End Pseudo-code
```

Statistical Techniques for the Validation of Endogenous Threat Time Series

To prove that the endogenous threat time series represents the threat variation contained in the threat time series generated by the GSA values, the statistical techniques of arithmetic mean, variance, standard deviation, skewness, and kurtosis were used.

The arithmetic mean is applied using the function mean from the SciPy library found in the Python 3.7 language. The arithmetic mean used by the mean function is represented by Equation 4.

Equation 4. *Formula of the arithmetic mean used by the mean function of the Python language*

$$\underline{X} = \sum \frac{X_i}{n}$$

where,

n – size of the data set
X_{ith} – ith value from the data set X.

The variance was applied using the variance function from the statistics library found in the Python 3.7 language. The variance function used by the variance function is represented by Equation 5.

Equation 5. *variance formula used by the statistic function of the Python language*

$$v = \sum \frac{\left(X_i - \underline{X}\right)^2}{n-1}$$

where,

n – size of the data set.
Xith – ith value from the data set X.
$\underline{X}$ - average of the values of X.

The standard deviation, Skewness s, is represented by Equation 6.

Equation 6. *Standard deviation formula used in the chapter for the analysis of endogenous time series*

$$s = \sqrt{v}$$

where,

v – variance of the data set

Asymmetry is a measure of the symmetry of a distribution. Asymmetry is also defined as the relative measure in two tails. Thus, the closer the skewness value approaches zero, the more the data set is

considered symmetric (Wheeler, 2011). Asymmetry is represented by Equation 7. Asymmetry is applied using the skew function from the SciPy library found in Python 3.7.

Equation 7. *Asymmetry formula used by the skew equation in the Python language*

$$a_3 = \sum \frac{\left(X_i - \underline{X}\right)^3}{ns^3}$$

n – size of the data set.
X_{ith} – ith value of X.
$\underline{X}$ – average of the values of X.
s – standard deviation of the data set.

The asymmetry parameter is also referred to as the third standardized central moment for the likelihood model, justifying the exponent 3 of the summation. According to Groeneveld (1991), asymmetry can be classified into:

- a3 > 0 - the distribution grout points to the right and the center of mass of the distribution is located on the left.
- a3 < 0 - the distribution grout is shifted to the left and the center of mass of the distribution is located to the right.
- a3 = 0 - distribution is approximately symmetric (to the third power of the deviation from the mean).

In this chapter, it is considered that negative asymmetry may represent a possible increase in the chances of some extreme event occurring due to the characteristic of the left-tailed distribution, which is present in endogenous time series data.

The kurtosis statistics parameter is defined as a measure that combines the weights of the tails relative to the rest of the distribution (Wheeler, 2011). Kurtosis can be defined by Equation 8 (Wheeler, 2011). Kurtosis is applied using the kurtosis function from the SciPy library found in the Python 3.7 language.

Equation 8. *kurtosis formula used by the kurtosis library of the Python language*

$$a_4 = \sum \frac{\left(X_i - \underline{X}\right)^4}{ns^4}$$

where,

n – size of the data set.
X_{ith} – ith value of X.
$\underline{X}$ – average of the values of X.
s – standard deviation of the data set.

The kurtosis parameter is also referenced as the fourth standardized central moment for the probability model, justifying the exponent 4 of the summation. According to (Brown, 2020), a normal distribution, has a kurtosis equal to 3 and the excess kurtosis equals -3. Thus, (Hayes & James, 2021) presents the following classification for excess kurtosis.

- A normal distribution has kurtosis exactly 3 (excess kurtosis exactly 0).
- A distribution with kurtosis » 3 (excess » 0) is called mesocurtic.
- A distribution with kurtosis < 3 (excess < 0) is called a platykurtic distribution. Such a distribution has shorter and thinner tails, and its central peak is often lower and wider, compared with a normal distribution.
- A distribution with kurtosis < 3 (excess < 0) is called a platykurtic distribution. Such a distribution has shorter and thinner tails, and its central peak is typically lower and wider, compared with a normal distribution.
- A distribution with kurtosis > 3 (excess > 0) is called leptokurtic and has longer and thicker tails, and often its central peak is higher, and sharper compared to a normal distribution, resulting in a greater opportunity for extreme positive or negative events.

In this chapter, the concept of leptokurtic kurtosis is used as a parameter to verify the endogenous threat time series due to the characteristics of their high peaks and tails represent a higher probability of occurrence of an extreme event considering the threat variation found in the time series data.

The Dynamic time warping (DTW) algorithm is used to compare the prediction result of part of the endogenous time series with the original part of the endogenous time series. DTW is used to calculate the dissimilarity between two time series of the same dimension (Giusti;Batista, 2013). Thus, in this chapter, the value of DTW represents the degree of difference between the two parts of the series. This algorithm was used through the tsle-arn.metrics.dtw package (Tavernard et al., 2020) in Python 3.7 language. The formula representing the DTW calculation can be seen in Equation 9.

Equation 9. *Formula of the DTW algorithm used by the tsle-arn.metrics.dtw library of the Python language*

$$DTW\left(x,y\right)=\sqrt{\sum_{i,j\epsilon\,\pi}\left(X_i-Y_j\right)^2}$$

where,

X e Y – represent the series to be compared
i e j – represent the i,j positions in a matrix of each element of the series
π – represents the set of possible paths between the series.

SOLUTIONS AND RECOMMENDATIONS

The application of the methodology developed, generated as a result the prediction of endogenous threat time series for drought, fire, and deforestation data for the years 2015, 2016, 2017, 2018, 2019 and 2020, considering the period of three future months. During the prediction process of this natural extreme event, several auxiliary results were also obtained, such as:

1) A portfolio with the news collected from social media, for each year considered, that contains the information about news date, news URL and similarity value.
2) The threat time series that was generated by the similarity calculation values presented in the threat portfolio.
3) The endogenous threat time series that was generated by calculating the DFA using the threat time series and the P-Model algorithm.
4) The statistical analysis of the endogenous time series generated by applying the statistical metrics (mean, standard deviation, variance, skewness, and kurtosis) of the endogenous threat time series for the year 2020.
5) The comparison of the prediction of the three future months with the first three months of each subsequent year, using the DTW value calculation.

Table 5 presents one of the threat portfolios (Ibanez et al., 2022) constructed for the years 2015, 2016, 2017, 2018, 2019, and 2020 for generating the threat time series. The portfolio in Table 5 considers 43 news stories published and collected in the year 2020 about the natural extreme event of drought, fires, and deforestation and presents the news stories collected from January to October 2020.

Table 5. Threat portfolio of the extreme natural event of drought, fires and deforestation in 2020

Date	URL	DTS
10/01/2020	https://noticias.uol.com.br/meio-ambiente/ultimas-noticias/redacao/2020/01/10/amazonia-esta-ficando-mais-seca-e-mais-propensa-as-queimadas-aponta-estudo.htm	0.893964703
11/01/2020	https://brasil.elpais.com/ciencia/2020-01-11/mudanca-climatica-aumentara-os-incendios-na-floresta-amazonica.html	0.893577612
13/01/2020	https://www.jornaldocomercio.com/_conteudo/geral/2020/01/720377-amazonia-tem-risco-de-grandes-queimadas-em-2020.html	0.901696075
13/01/2020	https://amazonia.org.br/2020/01/clima-tornara-amazonia-duas-vezes-mais-inflamavel-neste-seculo/	0.909248673
30/01/2020	https://conexaoplaneta.com.br/blog/o-ar-antes-sempre-umido-esta-progressivamente-mais-seco-tornando-a-floresta-inflamavel-alerta-antonio-nobre/	0.901099567
09/02/2020	https://climainfo.org.br/2020/02/09/satelite-revela-degradacao-generalizada-da-floresta-amazonica/	0.903992437
15/02/2020	https://www.dw.com/pt-br/sob-a-sombra-da-viol%C3%AAncia-na-amaz%C3%B4nia/a-52372640	0.917194963
18/02/2020	https://noticias.uol.com.br/meio-ambiente/ultimas-noticias/bbc/2020/02/18/o-que-ameaca-a-amazonia-em-seus-9-paises.htm	0.701143246
21/02/2020	https://www.modefica.com.br/desmatamento-secas-inundacoes-como-5-paises-tem-enfrentado-o-colapso-climatico/#.X4rSjCVv-V4	0.891219484
08/03/2020	https://www.cnnbrasil.com.br/nacional/2020/03/08/quais-os-riscos-do-avanco-do-desmatamento-na-amazonia	0.909880856
11/03/2020	https://www.dw.com/pt-br/amaz%C3%B4nia-pode-entrar-em-colapso-em-50-anos-diz-estudo/a-52723760	0.915482883
12/03/2020	https://diariodovale.com.br/colunas/nasa-preve-o-fim-da-amazonia/	0.919621871
27/03/2020	https://noticias.uol.com.br/ultimas-noticias/reuters/2020/03/27/exclusivo-brasil-reduz-fiscalizacao-ambiental-em-meio-ao-surto-de-coronavirus.htm	0.896859998
20/04/2020	https://www.brasildefato.com.br/2020/04/20/covid-19-sera-cortina-de-fumaca-para-desmatamento-alerta-especialista-do-greenpeace	0.901451655
24/04/2020	https://jornal.usp.br/atualidades/queimadas-na-amazonia-tendem-a-ser-mais-intensas-este-ano/	0.897999033
30/04/2020	https://veja.abril.com.br/brasil/em-meio-a-pandemia-o-desmatamento-dispara-na-amazonia/	0.873539635

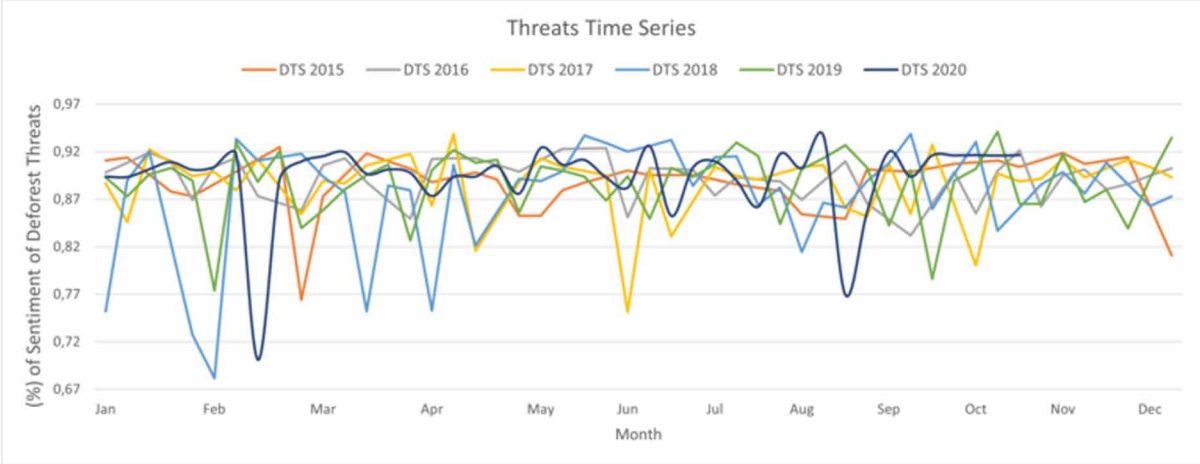
continues on following page

Table 5. Continued

Date	URL	DTS
30/04/2020	https://revistagloborural.globo.com/Noticias/Sustentabilidade/noticia/2020/04/instituto-de-pesquisa-ambiental-da-amazonia-alerta-para-risco-maior-de-queimadas-em-2020.html	0.893830352
01/05/2020	https://ciclovivo.com.br/planeta/meio-ambiente/desmate-aumenta-estacao-de-fogo/	0.893946057
04/05/2020	https://amazonia.org.br/2020/05/imagens-de-satelite-mostram-probabilidade-maior-de-seca-e-incendios-na-amazonia-este-ano/	0.904880914
26/05/2020	http://www.cemaden.gov.br/cientistas-alertam-para-a-contencao-das-queimadas-na-amazonia-e-o-colapso-do-sistema-de-saude-na-regiao/	0.876004602
27/05/2020	https://www.nytimes.com/pt/2020/05/27/opinion/international-world/amazonia-bolsonaro-quemas.html	0.924205117
04/06/2020	https://climainfo.org.br/2020/06/03/explosao-do-desmatamento-e-clima-mais-seco-devem-intensificar-queimadas-na-amazonia/	0.90649478
08/06/2020	https://www.dw.com/pt-br/desmatamento-e-fogo-podem-agravar-pandemia-na-amaz%C3%B4nia/a-53724198	0.911491545
15/06/2020	https://www.gazetadopovo.com.br/republica/queimadas-governo-plano-contingencia-amazonia/	0.893088718
21/06/2020	https://www.oeco.org.br/reportagens/destruicao-da-amazonia-pode-transforma-la-em-deserto-e-desencadear-pandemias/	0.88258079
25/06/2020	https://portalamazonia.com/noticias/cidades/acre-pode-ter-a-seca-mais-severa-dos-ultimos-anos-apontam-especialistas	0.926169753
09/07/2020	https://veja.abril.com.br/blog/impacto/a-projecao-da-nasa-para-a-temporada-de-queimadas-na-amazonia/	0.852539244
10/07/2020	https://isnportal.com.br/editorias/meio-ambiente/2020/07/10/nasa-alerta-para-risco-maior-de-incendios-na-amazonia-2020-esta-programado-para-ser-um-ano-perigoso-diz-cientista/	0.90379593
19/07/2020	https://www.nationalgeographicbrasil.com/meio-ambiente/2020/07/temporada-movimentada-de-furacoes-no-atlantico-pode-gerar-mais-incendios-na	0.910002761
28/07/2020	https://conexaoplaneta.com.br/blog/desmatamento-na-amazonia-seca-o-brasil-e-pode-levar-agronegocio-ao-colapso-alerta-relatorio-de-orgao-do-governo/	0.889712234
05/08/2020	https://projetocolabora.com.br/ods13/pouca-chuva-rios-secos-e-queimadas-assustam-pantanal/	0.862371353
07/08/2020	https://infoamazonia.org/pt/2020/08/portugues-avanco-do-desmatamento-e-tempo-seco-fazem-do-acre-o-mais-propicio-para-incendios-florestais/#!/map=51549&story=post-53586	0.917786164
14/08/2020	https://amazoniareal.com.br/amazonia-em-chamas-20-florestas-do-acre-podem-ser-mais-afetadas-por-incendios-diz-nasa-14-08-2020/	0.903044324
18/08/2020	https://noticias.uol.com.br/ultimas-noticias/afp/2020/08/18/desmatamento-e-impunidade-aticam-incendios-na-amazonia.htm	0.936240081
28/08/2020	https://g1.globo.com/am/amazonas/noticia/2020/08/28/viagens-de-barco-sao-suspensas-em-duas-cidades-do-alto-solimoes-no-am-por-conta-da-seca.ghtml	0.769455112
02/09/2020	https://acjornal.com/2020/09/28/sertao-acreano-seca-castiga-familias-da-floresta-estadual-do-antimary-que-caminham-horas-em-busca-de-agua/	0.857053481
09/09/2020	https://pp.nexojornal.com.br/perguntas-que-a-ciencia-ja-respondeu/2020/As-queimadas-na-Amaz%C3%B4nia-explicadas-em-10-pontos	0.920193429
19/09/2020	https://www.em.com.br/app/noticia/nacional/2020/09/15/interna_nacional,1185431/seca-e-o-fogo-criminoso-devastam-quase-todos-os-biomas-do-brasil.shtml	0.893926406
28/09/2020	https://www.tecmundo.com.br/ciencia/204415-amazonia-aproxima-ponto-nao-se-recuperar.htm	0.916403278
03/10/2020	https://www.anda.jor.br/2020/10/03/incendios-na-floresta-amazonica-sao-os-piores-em-uma-decada/	0.610168845
09/10/2020	https://amazonia.org.br/2020/10/combustivel-perfeito-desmatamento-provoca-incendios-no-territorio-indigena-do-xingu-mt/	0.916597098
09/10/2020	https://www.istoedinheiro.com.br/desmatamento-da-amazonia-cai-pelo-terceiro-mes-consecutivo-mas-queimadas-sobem/	0.91604095
13/10/2020	https://amazonia.org.br/2020/10/combustivel-perfeito-desmatamento-provoca-incendios-no-territorio-indigena-do-xingu-mt/	0.916597098

The degree of similarity of threats presented in the portfolio and generated by applying natural language processing and sentiment analysis techniques, generates the time series of threats that are presented in Figure 1. For the years 2015, 2016, 2017, 2018, and 2019, the data were collected in the period from January to December of each year. For the year 2020, the data was collected from January to October.

Figure 1. Threat time series for the years 2015, 2016, 2017, 2018, 2019, and 2020



For each threat time series generated by the collected news, the DFA value was calculated to generate the endogenous threat time series. Table 6 presents the values of the number of elements, the parameter *p* that defines the endogenous model, and the slope value or DFA calculation value.

Table 6. The endogenous time series were constructed according to the values of number of elements, parameter to represent the endogenous time series and slope or DFA value applied in the P-Model algorithm

Year	Elements Number	*p*	*Slope*
2015	16384	0.60	1.3082499264176222
2016	16384	0.60	1.3905561526142236
2017	16384	0.60	0.6792655564992033
2018	16384	0.60	0.8709826305556972
2019	16384	0.60	0.6708664045390897
2020	16384	0.60	0.5885749951561488

The endogenous series allowed the use of the deep learning techniques with the LSTM neural networks for the prediction of the future three-month period of threats. This prediction can be validated using the calculation of the DTW value by comparing the future three months of each year with the same period of the subsequent year. That is, the 2015 future prediction was compared with the same period in early 2016 and so on. In this way, it could be ensured that the 2020 future forecast presented information that

could be used to assist in the analysis of these natural extreme events. Figure 2 presents the endogenous threat series, the endogenous threat time series with the future prediction for a three-month period, and the DTW calculation with the overlap of the time series periods that were compared for each year analyzed. The prediction of the future of the endogenous threat time series is represented in red coloration at an approximate separation of the 3-month time interval from the predicted data.

The proof of these results was performed using the same number of elements of 3276 from the beginning of the endogenous series of threats of the year 2016 with the calculation of the three-month future or 3276 elements for the year 2015. It was used in the comparison of these series excerpts, the similarity value calculation, DTW calculation, and the overlap of the predicted elements with the initial elements of the series of the year 2016. This procedure was performed for all years of collected data, except for 2020. The year of 2020 was compared using data from reliable sources, such as the portal of the TerraBrasilis program, from INPE (Instituto Nacional de Pesquisas Espaciais - INPE, 2019), with information about the focus of fires and deforestation.

Observing Figure 2 (a) shows the endogenous series for the year 2015 with a predicting of three months, in red color. The same figure shows, totally in black, the endogenous series for the year 2016 . The excerpts of the endogenous series of 2015 with three months of prediction is compared with begging, considering approximately the same period, of the endogenous series of the year 2016 and this representation is marked with orange square. In Figure 2 (a), also representing 2015 the result of the comparison of the excerpts of series with DTW presents a similarity of 67.61% and that this also demonstrates the good performance of the methodology developed.

For the comparison between years 2016 and 2017, 2017 and 2018, 2018 and 2019, and 2019 and 2020, was realized using the same process that was described for the year 2015.

For the year 2016, Figure 2 (b), it shows a similarity of 46.04%, which shows that the methodology had a satisfactory success in predicting these future points. For the year 2017, Figure 2 (c), the comparison shows a similarity between the series of 48.7%, which also indicates a satisfactory performance of the developed methodology. In the year 2018, Figure 2 (d), the similarity result presents a value of 63.99% between the sections of the series analyzed and this result also shows the good performance of the developed methodology. Finally, for the year 2019, Figure 2 (e), the similarity achieved with the comparison of the series excerpts was 71.17%. This value also demonstrates the good performance of the methodology with the deep learning architecture developed.

For the year 2020 in Figure 2 (f), consider the analysis of the results achieved for the predictions of the years 2015, 2016, 2017, 2018 and 2019. In this analysis of 2020, observe that the threat variation continues in the three months after the analysis period, as observed in the prediction of the future of the series. Thus, it is suggested that for the months of November, December 2020 and January 2021, the threats of drought, fires, and deforestation should continue. This suggestion can be observed in a bar graph that shows the TerraBrassilis program of the INPE (Instituto Nacional de Pesquisas Espaciais - INPE, 2019) data about Fire and Deforestation in 2020.

The predictions of future natural extreme events, presented an accuracy, in comparison with the respective successor year, ranging between 46% and 71%, which also shows the good performance of the methodology developed in the thesis. For the data of extreme natural events of the year 2020, the prediction of the future presented that the threats would be maintained. Thus, using the information from the TerraBrasilis program, it was shown that during the three-month period following the data analysis period, the number of fires and deforestation increased.

Figure 2. Endogenous threat time series, endogenous threat time series with the future forecast for the three-month period, and DTW comparison and calculation between the three-month period

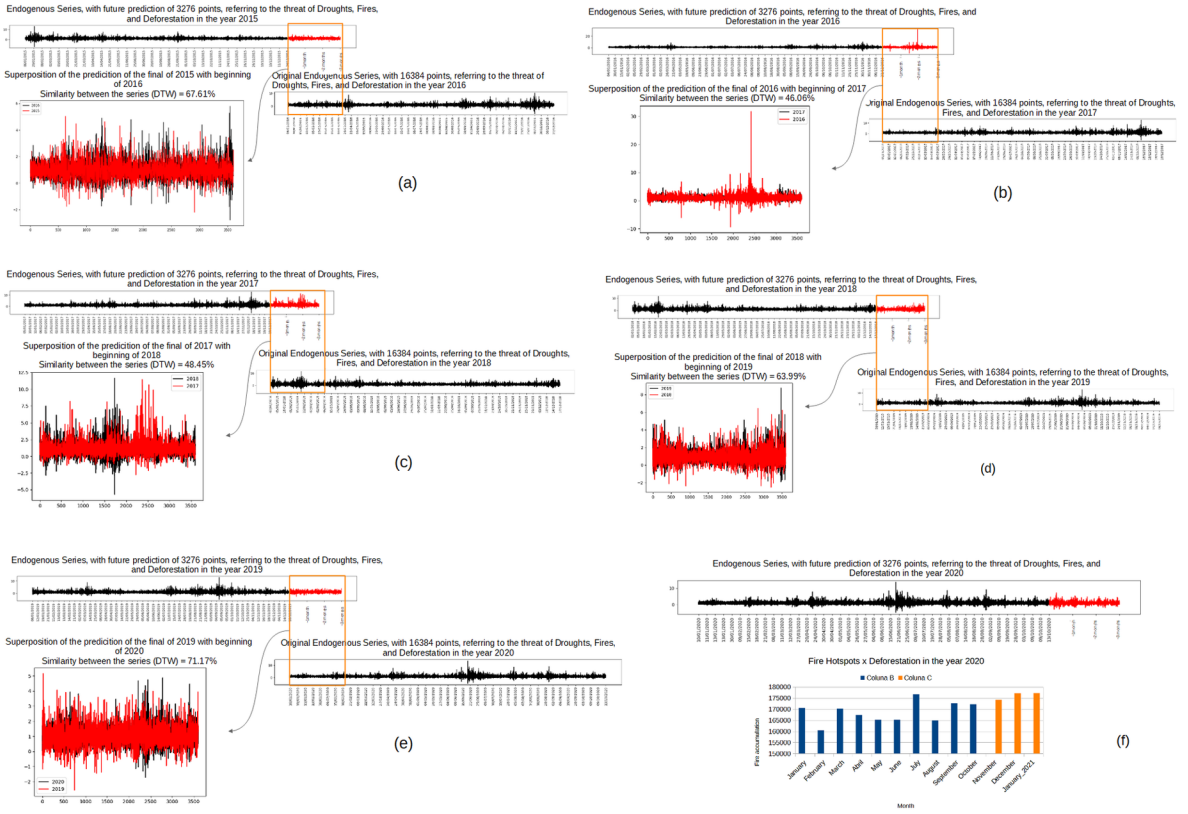


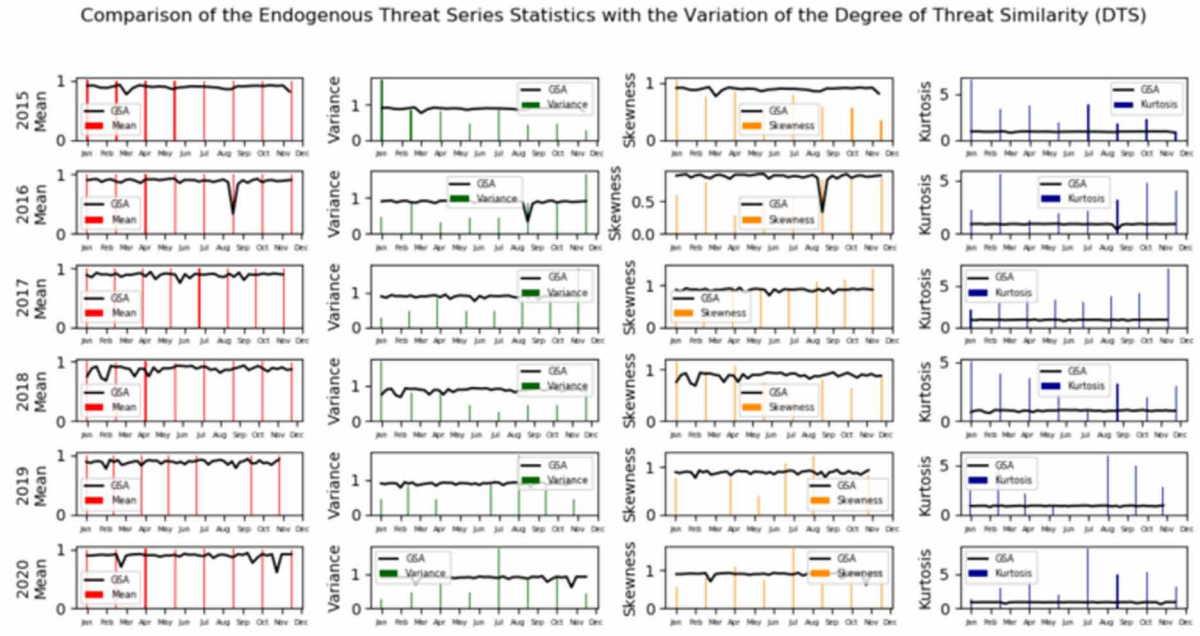
Table 7. Results of the similarities or accuracies achieved with the predictions of the data on the occurrence of drought, fires, and deforestation for the years 2015, 2016, 2017, 2018, 2019 and 2020

Year	Similarity of the future between series excerpts
2015	67.61%
2016	46.04%
2017	48.45%
2018	63.99%
2019	71.17%
2020	-

To verify whether the endogenous series performs a good representation of the threat variation present in the threat time series, statistical analysis is performed with the parameters of mean, variance, kurtosis and asymmetry. Figure 3 presents the result of this analysis.

The following is the analysis of these parameters with their characteristics for each year of news collected.

Figure 3. Statistical analysis with the parameters of mean, variance, kurtosis and asymmetry of the endogenous threat time series for each year of data collection



The following is the analysis of these parameters with their characteristics for each year of news collected.

Year 2015

- The Mean parameter did not adequately represent the peaks and valleys of the threat series.
- The Variance parameter showed satisfactorily the variation of the threat's series, presenting all the peaks and valleys of the series.
- The Asymmetry parameter satisfactorily presented the variations of peaks and valleys of the threat series. Its values are always greater than zero, and its tails show a slight positive slope to the right.
- The kurtosis parameter, also presented satisfactorily the variation of threats, presenting all its peaks and valleys. The kurtosis parameter, at the end of the analyzed period, presented values greater than 3 or excess > 0, also defining a leptokurtic kurtosis. Thus, it is observed that the tails of the kurtosis are longer and fatter on the left. These characteristics suggest the possibility of the occurrence of some major event or extreme event, after the period analyzed.

Year 2016

- The Mean parameter did not adequately represent the peaks and valleys of the threat series.
- The Variance parameter satisfactorily demonstrated the threat variation of the threat series, presenting all the peaks and valleys of the series.
- The Asymmetry parameter presented the variations of peaks and valleys of the threat series. Its values are always greater than zero, and its tails show a slight negative slope to the left. These characteristics also suggest the possibility of some major event occurring.

- The kurtosis parameter, also presented satisfactorily the threat variation, presenting all its peaks and valleys. The kurtosis parameter, at the end of the analyzed period, presented values greater than 3 or excess > 0, also defining a leptokurtic kurtosis. Thus, it is observed that the tails of the kurtosis are longer and fatter on the left. These characteristics suggest the possibility of the occurrence of some major event or extreme event, after the period analyzed.

Year 2017

- The Mean parameter did not adequately represent the peaks and valleys of the threat series.
- The Variance parameter satisfactorily demonstrated the threat variation of the threat series, presenting all the peaks and valleys of the series.
- The Asymmetry parameter presented the variations of peaks and valleys of the threat series. Its values are always greater than zero, and its tails show a slight negative slope to the left. These characteristics also suggest the possibility of some major event occurring.
- The kurtosis parameter, also presented satisfactorily the threat variation, presenting all its peaks and valleys. The kurtosis parameter, at the end of the analyzed period, presented values greater than 3 or excess > 0, also defining a leptokurtic kurtosis. Thus, it is observed that the tails of the kurtosis are longer and fatter on the left. These characteristics suggest the possibility of the occurrence of some major event or extreme event, after the period analyzed.

Year 2018

- The Mean parameter did not adequately represent the peaks and valleys of the threat series.
- The Variance parameter satisfactorily demonstrated the variation of threats of the threat series, presenting all the peaks and valleys of the series in a satisfactory manner.
- The Asymmetry parameter satisfactorily presented the variations of peaks and valleys of the threat series. Its values are always greater than zero, and its tails show a slight positive slope to the right.
- The kurtosis parameter, also presented satisfactorily the variation of threats, presenting all its peaks and valleys. The kurtosis parameter, at the end of the analyzed period, presented values greater than 3 or excess > 0, also defining a leptokurtic kurtosis. Thus, it is observed that the tails of the kurtosis are longer and fatter on the left. These characteristics suggest the possibility of the occurrence of some major event or extreme event, after the period analyzed.

Year 2019

- The Mean parameter did not adequately represent the peaks and valleys of the threat series.
- The Variance parameter satisfactorily demonstrated the variation of threats of the threats series, presenting all the peaks and valleys of the series in a satisfactory manner.
- The Asymmetry parameter presented the variations of peaks and valleys of the threat series. Its values are always greater than zero, and its tails show a slight negative slope to the left. These characteristics also suggest the possibility of some major event occurring.
- The kurtosis parameter, also presented satisfactorily the threat variation, presenting all its peaks and valleys. The kurtosis parameter, at the end of the analyzed period, presented values greater than 3 or excess > 0, also defining a leptokurtic kurtosis. Thus, it is observed that the tails of the kurtosis are longer and fatter on the left. These characteristics suggest the possibility of the occurrence of some major event or extreme event, after the period analyzed.

Year 2020

- The Mean parameter, for this 2020 data, also, did not adequately represent the peaks and valleys of the threat series.
- The Variance parameter showed the threat variation of the threat series in a very satisfactory way, presenting definitely all the peaks and valleys of the series.
- The kurtosis parameter, also showed satisfactorily the variation of threats, presenting all its peaks and valleys. The kurtosis parameter, at the end of the analyzed period, presented values greater than 3 or excess > 0, also defining a leptokurtic kurtosis. Thus, it is observed that the kurtosis grids are longer and fatter on the left. These characteristics suggest the possibility of the occurrence of some major event or extreme event, after the period analyzed.
- The Asymmetry parameter presented the variations of peaks and valleys of the threat series. Although its values are always greater than zero, its tails present a slight negative slope to the left. These characteristics also suggest the possibility of some major event occurring.

The analysis of the parameters of mean, variance, kurtosis, and asymmetry, showed that the parameters of kurtosis and asymmetry, had a better response to suggest that the endogenous series would be representing the variations of threats and that this could suggest some event greater than the threats analyzed.

Considering the analysis of these parameters, for all years of data collected from the natural extreme events, although the variance presented in all analyses a good response for the representation of the peaks and valleys of the series, it was not found a relationship of the values of variance between the different populations or data models analyzed.

FUTURE RESEARCH DIRECTIONS

As future work, intend to expand the collection of data on fires and deforestation presented for natural extreme events so as not to need to use an interpolation thecnique. Intend to improve the application of sentiment analysis, considering more specific emotion information from the news such as sarcasm, irony, and the use of image data.

CONCLUSION

The advent of social media and networks has made it possible for the most diverse categories of information and opinions to circulate among society. One model of information that calls for a deeper analysis is news about natural extreme events. For the data regarding natural extreme events, analyzed information about the occurrence of drought, fires, and deforestation in the Amazon rainforest region for the years 2015, 2016, 2017, 2018, 2019, and 2020. As a result of the analysis of future threat of a possible increase in the number of occurrences of drought, fires and deforestation in the considered region, an accuracy between 46% and 71% was obtained for the prediction of 3276 points for the successive years of the years 2015, 2016, 2017, 2018, 2019 and 2020, as presented in Table 7. In the analysis of the year 2020, it can be seen that the threat of occurrence of drought, fires, and deforestation remains constant for the subsequent 3 months. This was proven, observing the data from the TerraBrasilis program of INPE. To verify the endogenous series and the prediction method generated, the parameters of mean, variance, kurtosis, and asymmetry were calculated. From this investigation, it was observed that the

kurtosis and asymmetry parameters suggested that the endogenous threat series would be representing the variation of threats and that this variation could indicate the occurrence of something greater than a threat. The mean parameter did not generate information that could lead to any analysis in this direction. The variance parameter, as far as it is concerned, presented in the analyses of all portfolios of extreme social and natural events analyzed, a good representation of the peaks and valleys of the threat series. However, considering all the populations or data models analyzed, this parameter did not allow generating a consistent relationship between this information.

REFERENCES

Ahmadi, V. (2018). *Deforestation prediction using neural networks and satellite imagery in a spatial information system*. https://arxiv.org/abs/1803.02489

Armenteras, D., Gibbes, C., Anaya, J. A., & Dávalos, L. M. (2017). Integrating remotely sensed fires for predicting deforestation for redd. *Ecological Applications*, *27*(4), 1294–1304. https://pubmed.ncbi.nlm.nih.gov/28208227/. doi:10.1002/eap.1522 PMID:28208227

Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with python:analyzing text with the natural language toolkit. O'Reilly Media.

Brown, S. (2020). *Measures of shape: Skewness and kurtosis*. https://brownmath.com/stat/shape.htm#Kurtosis

Butler, R. A. (1999). *Notícias ambientais - notícias sobre vida*. https://brasil.mongabay.com/

Castells, M. (2003). A galáxia da internet reflexões sobre a internet, os negócios e a sociedade. Zahar.

Chollet. (2015). *Keras: The Python deep learning library*. Retrieved from Keras: https://keras.io/

Clauset, A. (2018). Trends and fluctuations in the severity of interstate wars. *Social Sciences, 4*(2). https://advances.sciencemag.org/content/4/2/eaao3580

Dahal, B., Kumar, S. A. P., & Li, Z. (2019). Topic modeling and sentiment analysis of global climate change tweets. *Social Network Analysis and Mining, 9*(1). https://www.springerprofessional.de/en/topic-modeling-and-sentiment-analysis-of-global-climate-change-t/16789020

Folha, G. (1996). *Uol - o melhor conteúdo*. www.uol.com.br

Gao, S. (2019). Deforestation prediction using time series and lstm. 2019 International Conference on Information Technology and Computer Application (ITCA). Proceedings, 95–99. doi:10.1109/ITCA49981.2019.00029

Giusti, R., & Batista, G. E. A. P. A. (2013). An empirical comparison of dissimilarity measures for time series classification. In *Brazilian Conference on Intelligent Systems* (pp. 82–88). doi:10.1109/BRACIS.2013.22

Globo. (2006). *G1 - o portal de notícias da Globo*. https://g1.globo.com/

Goodfellow, I., Benbio, Y., & Courville, A. (2016). *Deep learning – adaptive computation and machine learning series*. MIT Press.

Google. (2015). *TensorFlow wide & deep learning tutorial*. https://www.tensorflow.org/tutorials/wide_and_deep$

Groeneveld, R. A. (1991, May). An influence function approach to describing the skewness of a distribution. *The American Statistician*, *45*(2), 97–102. https://www.jstor.org/stable/2684367

Halsey, T. C., Jensen, M. H., Kadanoff, L. P., Procaccia, I., & Shraiman, B. I. (1987). Fractal measures and their singularities: The characterization of strange sets. *Nuclear Physics B - Proceedings Supplement*, *2*, 501–511. doi:10.1016/0920-5632(87)90036-3

Harvard, U. (2019). *Detrended fluctuation analysis (DFA).* http://reylab.bidmc.harvard.edu/download/DFA/intro/

Hayes, A., & James, M. (2021). *Leptokurtic distributions*. https://www.investopedia.com/terms/l/leptokurtic.asp

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. doi:10.1162/neco.1997.9.8.1735 PMID:9377276

Ibanez, M. M., Rosa, R. R., & Guimaraes, L. N. F. (2022). Threat Emotion Analysis in Social Media: Considering Armed Conflicts as Social Extreme Events. In Handbook of Research on Opinion Mining and Text Analytics on Literary Works and Social Media. IGI-Global.

Instituto Nacional De Pesquisas Espaciais - INPE. (2019). *TerraBrasilis*. http://terrabrasilis.dpi.inpe.br/

Jackson, P., & Mouliner, I. (2002). *Natural language processing for online applications: Textretrieval, extraction and categorization*. John Benjamins B.V. doi:10.1075/nlp.5(1st)

Lima, D. (1994). *Instituto socioambiental (ISA).* https://www.socioambiental.org/pt-br/o-isa

Lima, P. M., & Guedes, E. B. (2015). Rainfall prediction for Manaus, Amazonas with artificial neural networks. *Latin America Congress on Computational Intelligence (LA-CCI) Proceedings*, 1–6. https://ieeexplore.ieee.org/document/7435934

Morariu, V. V., Buimaga-Iarinca, L., Vamos, C., & Soltuz, S. (2007). *Detrended Fluctuation Analysis of Autoregressive Processes*. Academic Press.

Morello, T. F., Ramos, R. M., Anderson, L. O., Rosan, T. M., & Steil, L. (2016). Predicting amazon fires for policy making. *Encontro Nacional de Economia - Economia Agrícola e do Meio Ambiente, 44*(11). https://www.anpec.org.br/encontro/2016/submissao/files_I/i11-3b68242e7c3a5a3a7f24ce256c5d517c.pdf

Newsbot. (2019). *Related news at the click of a button*. https://getnewsbot.com/

Paes, V., Araújo, D., Brito, K., & Andrade, E. (2022). Análise de Sentimento em Tweets Relacionados ao Desmatamento da Floresta Amazônica. In *Anais do XI Brazilian Workshop on Social Network Analysis and Mining* (pp. 61-72). Porto Alegre: SBC. 10.5753/brasnam.2022.222648

Peng, C. K., Buldyrev, S. V., Havtin, S., Simons, M., Stanley, H. E., & Goldbergrz, A. L. (1994). Mosaic organization of dna nucleotides. *Physical Review. E*, *49*(2), 1685–1689. doi:10.1103/PhysRevE.49.1685 PMID:9961383

Rosa, R. R., Neelakshi, J., Pinheiro, G. A. L. L., Barchi, P. H., & Shiguemori, H. (2019). Modeling social and geopolitical disasters as extreme events: a case study considering the complex dynamics of international armed; conflicts. In Towards mathematics, computers and environment: A disasters perspective (pp. 233–254). Academic Press.

Saha, S., Saha, M., Mukherjee, K., Arabameri, A., Thingo, P. T., & Paul, G. C. (2020). Predicting the deforestation probability using the binary logistic regression, random forest, ensemble rotational forest, reptree: A case study at the gumani river basin, india. *The Science of the Total Environment*, *730*, 139–197. doi:10.1016/j.scitotenv.2020.139197 PMID:32402979

Sornette, D. (2006). Endogenous versus exogenous origins of crises. In S. Albeverio, V. Jentsch, & H. Kantz (Eds.), *Extremes events in nature and society* (pp. 107–131). Springer. doi:10.1007/3-540-28611-X_5

Tavenard, R., Faouzi, J., Vandewiele, G., Divo, F., Androz, G., Holtz, C., Payne, M., Yurchak, R., Russwurm, M., Kolar, K., & Woods, E. (2020). Tslearn, a machine learning toolkit for time series data. *Journal of Machine Learning Research*, *21*(118), 1–6. https://jmlr.org/papers/v21/20-091.html

Wheeler, D. J. (2011). *Problems with skewness and kurtosis, part two*. https://www.qualitydigest.com/inside/quality-insider-article/problems-skewness-and-kurtosis-part-two-080111.html

Zafarani, R., Abbasi, M. A., & Liu, H. (2014). *Social media mining - An introduction*. Cambridge University Press. doi:10.1017/CBO9781139088510

ADDITIONAL READING

Pozzi, F. A., Fersini, E., Messina, E., & Liu, B. (2016). *Sentiment analysis in social networks*. Morgan Kaufmann.

Zhang, Y., Shang, L., & Jia, X. (2015). Sentiment analysis on microblogging by integrating text and image features. In T. Cao, E.-P. Lim, Z.-H. Zhou, T.-B. Ho, D. Cheung, & H. Motoda (Eds.), *Advances in knowledge discovery and data mining* (pp. 52–63). Springer. doi:10.1007/978-3-319-18032-8_5

KEY TERMS AND DEFINITIONS

Data Science: Collection, preparation, and analysis of a great amount of data.
Endogenous Events: Event that generates reaction based only on the domain of the event itself.
Event: burning caused by the threat.
Extreme Events: Deforestation generated by threat and event.
Natural Events: Event generated by a reaction of some phenomenon of nature.
Sentiment Analysis: Analysis to identify emotions in some kind data as text, video, sound, and image.
Similarity Degree: Degree of similarity between the collected news and text base generated.
Social Media: Place where public information is made available that can be collected and analyzed to extract some value´s type.
Threat: Mention of the possibility of drought.
Threat Time Series: Time series that represent the variation of degree of threat similarity.