Data Analytics for a Changing Climate: Feature Engineering for the forecast of hydrometeorological events

Caique S. Noboa[1], Daniel Pigatto[1], Elaiz M. Buffon[2], Luiz Gomes-Jr[1]

[1]UTFPR - Curitiba-PR - Brazil

[2]Unioeste - Francisco Beltrão - PR - Brazil

lcjunior@utfpr.edu.br

Abstract. *Floods are becoming increasingly frequent, and consequently, the number of people and infrastructure affected by these events has increased. It is essential to have accurate models for the prediction of such hydrometeorological events, improving preparedness and decision making for damage reduction. The goal of this work was to determine the variables (features) that contribute the most to predicting hydrometeorological events. Feature engineering techniques were used to understand which factors are most helpful in predicting floods. The features were composed based on data from rain gauges, altimetry, and location of rivers and lakes. It was observed that the variables that had the greatest impact on improving the model were the data from rain gauges and altitude data. The predictive model proposed is part of a larger system being developed in the context of Smart Cities called ICARUS. The system is aimed at improving response time and up-time of critical infrastructure during extreme events.*

1. Introduction

The number of people affected by hydrometeorological events has been increasing in recent years, even considering non-extreme rainfall events [Lohmann 2011]. Predicting these events requires data, and the lack of data is one of the main challenges. The absence of data (cartographic, meteorological, hydrological) at suitable scales for local studies has been one of the obstacles to conducting research related to understanding the dynamics of extreme events in Brazil [Lohmann 2011].

To address similar problems, precipitation data collected from rain gauges are commonly used. For the city of Curitiba, there are already other works using data collected by the Municipal Civil Defense, which registers the time and location of flood events [Buffon and de Sousa 2018]. Fernandez and Splendore [Fernandez and Splendore 2021] proposed the ICARUS (***I**ntegrated **C**risis **A**wareness and **R**esource **U**tilization for **S**martcities*) system, a platform that integrates detection of extreme events and management of Smart City infrastructure. The system allows faster response times in the face of such events. The work presented in this paper is part of the ICARUS system.

A significant challenge in the context of extreme event detection is identifying which variables are most relevant for predicting these events. The field of *Feature Engineering* involves comparing variables to select the best ones that can influence the final model's outcome.

The aim of this paper is to implement a predictive model for hydrometeorological events in the city of Curitiba, using data collected by rain gauges, altimetry, and information about rivers and lakes in the region. Feature engineering techniques are applied to understand the factors that most influence the prediction of these events. The knowledge generated in this work can contribute not only to future research related to the dynamics of these phenomena in Brazil but also to the development and improvement of existing models.

2. Fundamentals and Related Work

2.1. Floods, Inundations, and Flooding

Based on the work by Buffon [Buffon 2020], we adopt the following concepts: floods, inundation, and flooding, which will be used in this work. Therefore, it is essential to differentiate them. Floods, also known as high waters, according to Noronha [Noronha 2021], "involve an increase in the water level in the drainage channels of rivers, streams, and reservoirs". They occur during the rainy season without overflowing their banks, making them a natural process. Inundations refer to events with water overflowing from rivers. Meanwhile, flooding events occur when there is insufficient drainage in an area, primarily in urban settings, and are not necessarily linked to floods or inundations.

Buffon [Buffon 2020] provides a theoretical foundation that presents different types of flooding, structural and marginal, and explains their differences. This study serves as the basis for choosing the most influential variables in the occurrence of floods and flooding, as well as the theoretical foundation for this topic. From the thesis, two promising variables were identified: data recorded by rain gauges and altimetry data.

2.2. Decision Trees and Random Forests

A popular technique for supervised learning is the decision tree. It consists of a hierarchical structure similar to a flowchart, with each internal node representing a test of a specific feature or data attribute [Quinlan 1986]. The results of these tests branch the tree until the final prediction result is found at the leaf node. Since decision trees are easy to understand and interpret, they are often used for problems with few independent variables. However, they can quickly become complex when many attributes are involved.

The Random Forest algorithm combines several independent decision trees to make the final classification [Breiman 2001]. Each tree is built from a random sample of data and only a subset of the available variables (randomly selected). Thus, each tree has different strengths and weaknesses when classifying examples. By combining these trees into a final model, individual weaknesses are compensated for by more robust trees. This process produces a model that is less sensitive to variations in the training set and can generalize correctly to new test sets [A and M 2002].

2.3. Related Work

Buffon and Souza [Buffon and de Sousa 2018] developed an analysis of flood and rainfall data in Curitiba to find interrelationships of the phenomena. Their work presents cases of recorded flooding without the occurrence of rain, which may indicate potential errors and inconsistencies. The article explains issues related to data preprocessing in this context and identifies areas in Curitiba with the highest flood risk. The results show trends in

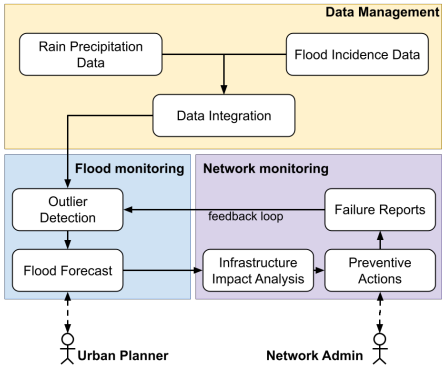

Figure 1. Architecture of the ICARUS System

the data, but the authors recommend the analysis of secondary data sources to enrich the modeling.

Lohmann [Lohmann 2011] employed logistic regression-based models for probabilistic flood prediction. This study was also conducted in Curitiba and utilized data from meteorological radar, satellite, and rain gauges. The model achieved highly positive results, with an F-Score of 0.8704, using higher resolution and more precise data.

Finally, the work presented in [Wu et al. 2010] demonstrates an approach for anomaly detection in spatial and precipitation data. The authors propose an algorithm called ”Outstretch” that is capable of identifying the top-k outliers using spatial scan statistics [Kulldorff 1997]. The authors' approach aligns with the work proposed here. However, the work differs from the proposed one as it evaluates the accuracy of the algorithm with data related to the behavior of the El Niño Southern Oscillation (ENSO) phenomenon, which is a much more atypical and persistent phenomenon compared to the types of natural disasters considered in this paper.

3. ICARUS System

The work presented in this paper is part of the ICARUS (***I**ntegrated **C**risis **A**wareness and **R**esource **U**tilization for **S**martcities*) System [Fernandez and Splendore 2021]. The name given to this systems aims to represent the awareness that a smart city has regarding the crises it faces, being integrated with a resource management and utilization module.

The general objective of the system is to develop an outlier-based model to predict natural disaster events. The model works in conjunction with the communication infrastructure management system to mitigate issues [Matisziw and Murray 2009]. This approach allows for prevention and contingency measures to avoid future communication failures. The integration of all these components is referred to as the ICARUS system, which includes the outlier detection model and the network infrastructure manager (Figure 1). The input to this platform is pluviometric data, a history of floods and inundations, and geographic data such as latitude and longitude, with the output being an action taken to mitigate a network connection failure.

The architecture of the system is shown in Figure 1, describing the stages of the communication flow and how the interaction and integration between the models are performed. The first stage shown in the diagram displays the management and acquisition of

Figure 2. Map discretized into hexagons of Curitiba.

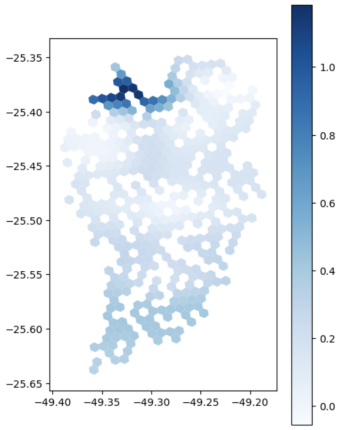

Figure 3. Rain Gauge Data, February 20, 2018.

all the data used. The second and third stages describe how the outlier detection model interacts with the infrastructure simulation model and vice versa.

4. Methodology

4.1. Data Acquisition and Processing

Rain Gauge Data: Hourly precipitation data provided by [Cemaden 2021] were used, which has eleven automatic rain gauge stations in the Curitiba region. The work [Fernandez and Splendore 2021] served as the basis for interpolating the data to obtain the climatic conditions of the midpoint, using the interpolation method called *Kriging*.

The work [Fernandez and Splendore 2021] also discretized the map of Curitiba into hexagons (Figure 2). Each hexagon has approximately 600 square meters of area. The same was done in this work due to the amount of data and the high processing time at a finer resolution. The base data was obtained from [Cemaden 2021].

With the intention capturing the impact of water retention in the soil, two new metrics were created:(i) 3-Day Rain Accumulation, and (ii) 7-Day Rain Accumulation, both using the precipitation values of the last days for the each hexagon.

Elevation Data: The elevation data (contour lines) were obtained from IPPUC (Curitiba's Planning Department)[1]. The provided files are in Shapefile format. Just like in [Fernandez and Splendore 2021], these data were discretized into the same hexagons (Figure 2), averaging the measurements of all points within that hexagon. The dataset was obtained from [de Curitiba 2022].

Hydrometric Data: As an input for another variable, data from lakes, lagoons, reservoirs, rivers, quarries, and floodplains were used. They were also obtained from IPPUC, also in Shapefile format.

Again, the data were discretized into hexagons, and each hexagon received a variable indicating whether it contained rivers, lakes, or quarries, and the number of each

[1]https://www.ippuc.org.br/

type. This processing was done using the spatial operation Join. Then, the total number of that hexagon was summed, creating the *hydro score* variable. A neighborhood variable was also created, which sums the *hydro score* of neighboring hexagons. The dataset was obtained from [de Curitiba 2022].

Flood Records: To measure the quality of the models, data on flood records registered by the Municipal Civil Defense were used. It is important to note that these records do not have high reliability because not all cases of flooding are reported and, therefore, not recorded. These records were also discretized into hexagons. The dataset was obtained from [IPPUC 2022].

4.2. Feature Engineering

In order to find the best features that detect flooding, this work used the Random Forest algorithm, along with the Cross Validation method, executed in isolated environments with each feature. They were implemented using the Python programming language and the libraries Pandas and Scikit-learn.

This technique was used on precipitation data, elevation data for the city of Curitiba, and hydrometric data. For each data set, more than one variable was created. In order to group related variables and also reduce the number of executions and comparisons, three groups of features were created: (i) Precipitation: Data collected from rain gauges; (ii) Elevation: Elevation data; and (iii) Hydrometric: Data from lakes, lagoons, reservoirs, rivers, quarries, and floodplains. For each group of features used in the model, the times to run the model were recorded, as well as the evaluation metrics: Accuracy, Precision, Recall, and F-Score. By comparing the results obtained by different models using the F-Score metric, it was possible to identify the best features for classifying the hydrometeorological events in question.

Next, models were created with the following combinations of feature groups: (i) Precipitation + Elevation; (ii) Precipitation + Elevation + Hydrometric. Then, the performance of the feature groups was compared to understand if any feature group could be removed while maintaining equal or superior results.

4.3. Modeling and Evaluation

The Random Forest Classifier algorithm was used to classify points as flooding or no flooding. The purpose of the work is not to use a robust model but a model that is simple to understand to compare the results of different features. The process involved dividing the collected data into training and testing sets (for performance evaluation). An execution was performed for each feature group in isolation, and several executions considering the union of features.

To assess the quality of the models, data on flooding records provided by the Municipal Civil Defense were used. The main metric used is the F-Score, which combines the values of precision and recall. Precision is the ratio of correctly predicted data to all predicted data, and recall is the ratio of correctly identified data to all data. The Cross Validation method was used in the executions with features to ensure a fair evaluation of the features.

5. Results

Based on the results obtained after running the Random Forest Classifier model, a comparative analysis was conducted between three feature groups: Precipitation, Elevation, and Hydrometric (Table 1).

Group	Time	Accuracy	Precision	Recall	F-Score
Precipitation	6m 42s	0.999670	0.941733	0.711386	0.777140
Elevation	3m 5s	0.999720	0.982721	0.710657	0.786063
Hydrometric	2m 21s	0.999577	0.89978	0.576307	0.624638
Precipitation + Elevation	6m 13s	0.999740	0.989872	0.745748	0.824096
All	7m 28s	0.999712	0.988170	0.714212	0.790546

Table 1. Results of the Random Forest Classifier model execution with various features

Analyzing the features individually, we can see that the Hydrometric feature did not achieve a high F-Score compared to the others. On the other hand, the Elevation feature performed better in all metrics, except for Recall when compared to Precipitation. This was surprising because elevation does not vary with the date, and it was expected that Elevation alone would not achieve better results than the Precipitation group, which varies with the date.

Another point to discuss is the execution time of each feature. The Precipitation group was the slowest, which can be explained by the variation by date of the variable, being the only group with this feature. The Hydrometric group was faster, and this can be explained by the fact that there are many hexagons with a zero value for this group, reducing the required comparisons and increasing the speed of the model.

In the execution with all features, the results were better compared to the rounds with the features alone, but it was worse than the round using only Elevation and Precipitation. Therefore, we can say that the Hydrometric feature did not improve the results when using other features together. This may have happened due to the simplification of the Hydrometric Group, considering different types of water elements (lakes, lagoons, rivers, etc.) as the same element. Perhaps with an improvement in the collection of this variable, the result could have been better.

6. Conclusion

This work aimed to analyze variables that influence flood prediction in the city of Curitiba. The combination of Precipitation and Elevation features achieved the highest F-Score. The result of the work was better than expected, considering that the work [Fernandez and Splendore 2021] uses similar data at the same resolution, achieving an F-Score of 0.60. It was possible to verify that new features helped improve the model, and it was possible to verify that the Hydrometric data feature would be dispensable in a subsequent work.

As suggestions for future work, four points can be mentioned: (i) Improve data resolution and also the results; (ii) Explore a comparison between different machine learning models; (iii) Use more data from various sources, such as Lidar data for altimetry, which have higher resolution but are more computationally intensive; (iv) Apply the model to other cities with continuous records of floods and inundations and higher precision remote sensing data.

References

A, L. and M, W. (2002). Classification and regression by randomforest. *R News*, pages 18–22.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Buffon, E. A. M. (2020). *Inundações Em Áreas Urbanas: Proposição Conceitual-Metodológica E Sua Aplicação Na RMC – Região Metropolitana De Curitiba*.

Buffon, E. A. M. and de Sousa, M. S. (2018). *Proposta Metodológica Para Avaliação Dos Registros Secundários De Alagamentos: Uma Abordagem A Partir De Curitiba-Paraná, Brasil*.

Cemaden (2021). *Pluviômetros Automáticos – Cemaden.* Accessed: 2022-11-03.

de Curitiba, P. M. (2022). *Dados Geográficos de Curitiba*. Accessed: 2022-11-05.

Fernandez, H. G. and Splendore, P. R. (2021). *Sistema De Identificação Automática De Riscos Hidrometeorológicos Com Retroalimentação E Reestruturação Autônoma Da Infraestrutura De Comunicação*. Curitiba, Brasil.

IPPUC (2022). *Registros Alagamentos*. Accessed: 2022-11-07.

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics-Theory and methods*, 26(6):1481–1496.

Lohmann, M. (2011). *Regressão Logística E Redes Neurais Aplicadas À Previsão Probabilística De Alagamentos No Município De Curitiba, PR*. Curitiba, Brasil.

Matisziw, T. C. and Murray, A. T. (2009). Modeling s–t path availability to support disaster vulnerability assessment of network infrastructure. *Computers & Operations Research*, 36(1):16–26.

Noronha, G. (2021). *Enchentes – O que são, características, causas e impacto urbano.*

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1.

Wu, E., Liu, W., and Chawla, S. (2010). Spatio-temporal outlier detection in precipitation data. In *Knowledge Discovery from Sensor Data*, pages 115–133. Springer Berlin Heidelberg.