sbbd:42

Precipitation Nowcasting using Data Augmentation

Eduardo Bezerra[1], Augusto Fonseca[1], Adriano Cabo[1], Fabio Porto[2], Mariza Ferro[3]

[1]Federal Center for Technological Education of Rio de Janeiro - CEFET/RJ

[2]National Laboratory for Scientific Computing - LNCC

[3]Federal Fluminense University - UFF

ebezerra@cefet-rj.br,

{augusto.fonseca,adriano.cabo}@eic.cefet-rj.br,

fporto@lncc.br,

mariza@ic.uff.br

Abstract. *This paper proposes a simple data augmentation technique specifically designed to mitigate the data unbalancing problem in precipitation nowcasting. We consider the existence of one or more observational systems, each one comprised of a set of (either weather or rain gauge) stations. We use simulated data coming from the ERA5 numerical model to complement precipitation observations made by rain gauge stations, and use the resulting synthetic observations to augment data for a given weather station. We present preliminary results training a machine learning model using this data augmentation technique. These results show that the technique can be useful to improve the predictive performance of the resulting forecasting model.*

1. Introduction

Precipitation nowcasting, the prediction of short-term rainfall patterns, is a vital component of weather forecasting systems. Accurate nowcasting helps in various applications, such as flood management, agriculture, transportation, and emergency preparedness. However, one of the challenges in developing reliable precipitation nowcasting models is the inherent data imbalance between precipitation and non-precipitation instances. Data imbalance occurs when the number of samples in one class significantly outweighs the number of samples in another class, leading to biased model training and reduced prediction accuracy. In the context of precipitation nowcasting, the majority of instances in the dataset typically represent non-precipitation events, making it difficult for the model to learn the complex patterns associated with precipitation accurately. Consequently, this data imbalance issue poses a significant obstacle to the development of robust and reliable nowcasting models.

To address this challenge, this paper proposes a simple data augmentation technique designed to mitigate the data unbalancing problem in precipitation nowcasting. Data augmentation refers to the process of artificially expanding the dataset by applying various transformations or modifications to existing samples, effectively increasing the diversity and quantity of training instances. The proposed data augmentation technique leverages the intrinsic properties of precipitation patterns to generate realistic and diverse synthetic precipitation instances. More specifically, we use simulated data from the ERA5 numerical model [Hersbach et al., 2020] to complement precipitation observations made by rain gauge stations, and use the resulting synthetic observations to augment data for a given weather station. We present preliminary results training a machine learning model using this data augmentation technique using the AlertaRio observational system (see Figure 1). These results show that the technique can be useful to improve the predictive performance of the resulting forecasting model.

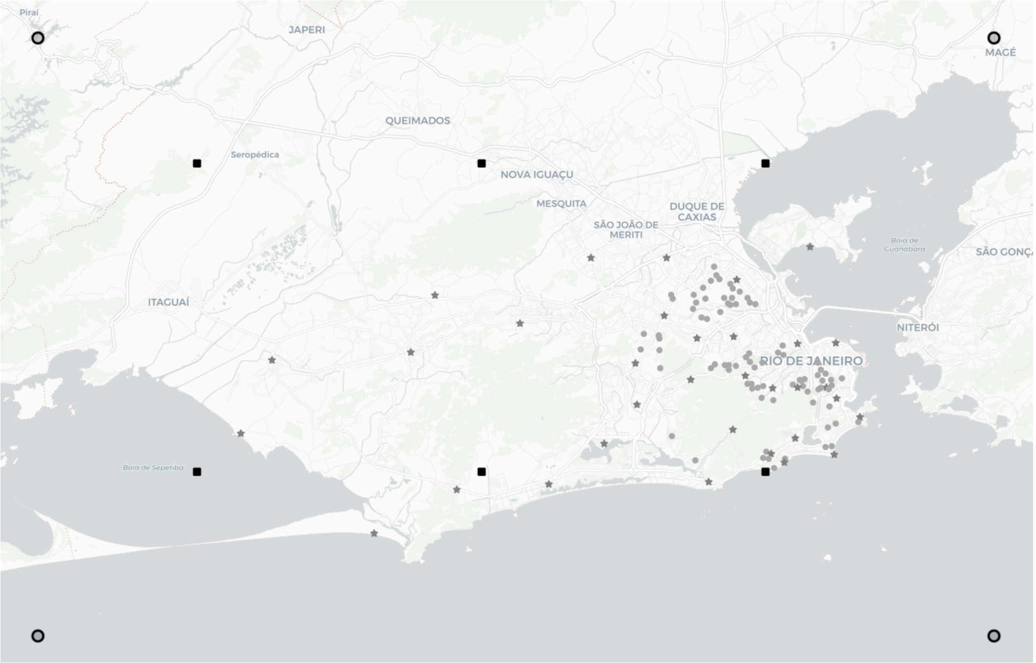

Figure 1. Map of our region of interest (mostly the metropolitan area of Rio de Janeiro city), depicting locations of ERA5 simulated data (black-squared dots, along with rain gauges and meteorological stations from AlertaRio and Sirenes observational systems (star-shaped and circle dots, respectively).

The rest of the paper is organized as follows. Section 2 describes the methodology and implementation details of our proposed data augmentation technique. Section 3 presents the experimental setup and results. Finally, Section 4 summarizes the contributions of this paper, highlights the limitations, and suggests future directions for research.

2. Method

We consider that a *weather station of interest* (WSoI) has been defined, located at a point in space for which we want to build a precipitation prediction model. This weather station can be any one that has historical data available. These data may include various meteorological variables. Without losing generality, to make the following description more concrete, let us consider that the information collected by the WSoI is as follows: temperature, barometric pressure, relative humidity of the air, wind (direction and speed) and precipitation. In addition to the WSoI, another input information for the proposed method is the list of historical observations for a set of neighboring stations. These neighboring stations can be either meteorological or rain gauges. In our data augmentation method, there is a processing mechanism to be applied in each of the two cases. In the first case (neighboring station is meteorological), we consider that this neighbor observes the same variables as the WSoI. In this case, the data augmentation procedure for the WSoI is relatively simple; the data observed by the neighboring station is fused with that of the WSoI, since both have the same set of measured variables. In the second case (neighboring station is rain gauge), the data from the numerical model ERA5 are used to complete the information about the variables that are not observed. Concretely, for a neighboring rain gauge station s, we determine the nearest simulated data point from ERA5 to s. We then use these simulated data to complement the precipitation observations made by the station s. The result is that, in an artificial way, we can make each rain gauge station have the same variables observed by a meteorological station.

By considering both cases described above, we have as a result a set of neighboring stations (with real/observed or artificial/synthetic data) to the WSoI whose data can be used to augment the training dataset for this station. Concretely, if the WSoI presents a historical series with m observations, and if the k neighboring stations have their respective historical series with $m_1, m_2, m_3, \ldots, m_k$ observations, then the historical series resulting from the data augmentation has $m + \sum_{i=1}^{k} m_i$ observations.

To build a precipitation prediction model for a WSoI, we apply the sliding window technique [Braverman, 2016] on its corresponding augmented historical series. The sliding windows technique is a method for segmenting a time series into fixed-size windows. Each window is then used as a training example for a machine learning model. The model is trained to predict the value of the next window based on the values of the previous windows. In the case of predicting precipitation, the time series is the augmented multivariate time series generated for the WSoI. We use the sliding windows technique to segment this time series into fixed-size windows. In this work, as we are interested in nowcasting scenarios, we experimented with building forecasting models to predict the precipitation value for the next hour based on the values of the observed variables for the previous 3 hours.

For training and validating the forecasting models, we first split the examples resulting from the sliding window technique into train, validation and test datasets. We took some measures to prevent data leakage issues [Montano et al., 2022]. First, we did the splitting using the proportions 70%, 20% and 10% for train, validation and test datasets, respectively, without shuffling the observations, to maintain the temporal order between datasets. Concretely, examples in the training dataset corresponds to observations that strictly precede in time those used in the test dataset.

To validate our data augmentation method, we performed experiments using a subset of stations from the AlertaRio system (see Table 1). As WSoI, we used the `guaratiba` (WSoI_1) and `sao_cristovao` (WSoI_2) stations. As rain gauge stations, we used the ones identified by `tijuca`, `tijuca_muda`, `saude`, and `grajau`. For the ERA5 simulated data, we used the Copernicus API[1] to retrieve data from January of 1997 to April of 2023. The points for which we retrieved ERA5 simulated data are depicted in Figure 1. Each of these points provides a hourly multivariate time series that are used in the fusion procedure with the rain gauge stations.

To build the forecasting models, we framed the problem as an ordinal classification

[1] https://scihub.copernicus.eu/twiki/do/view/SciHubWebPortal/APIHubDescription

Table 1. Stations used in the validation experiments and their corresponding Haversine distances (in Km) to WSoI$_1$ (guaratiba) and WSoI$_2$ (sao_cristovao).

Station	Operation Start Date	$\delta_{\mathbf{WSoI}_1}$	$\delta_{\mathbf{WSoI}_2}$
guaratiba	1997-01-01	0.00	41.84
sao_cristovao	2000-08-19	41.84	0.00
tijuca	1997-01-02	40.39	3.92
tijuca_muda	2011-02-07	38.27	4.59
saude	1997-01-01	45.04	3.46
grajau	1997-01-01	36.40	5.49

task. Concretely, we discretized the precipitation values (measured in mm/h) producing five levels, according to the following mapping: 0 → None; (0, 5] → Weak; (5, 25] → Moderate, (25,50] → Strong; (50, ∞] → Extreme.

For the learning algorithm, we experimented with a simple Artificial Neural Network architecture with two modules. Concretely, the architecture begins with a feature extraction module comprised of one Conv1D layer (16 kernels of size 3, stride = 1, padding = 3), followed by a ReLU layer, followed by one Dropout layer ($p = 0.5$) layer. The second and final module is comprised of a linear layer, followed by a ReLU layer, followed by a final layer with five sigmoid units. These five units in the last layer are meant to produce the encoded precipitation level (None, Weak, Moderate, Strong, Extreme). We trained each model for a maximum of 8,000 epochs using early stopping and a batch size of 1,024. The optimizer was Adam, with learning rate of 3×10^{-6}. To validate to predictive performance of each model, we report their confusion matrices.

3. Results and discussion

We present our preliminary results in six tables. The initial three tables (Table 2, Table 3, Table 4) are related to experiments considering guaratiba as the WSoI. The remaining three tables (Table 5, Table 6, Table 7) report results using sao_cristovao as WSoI. A common behavior among all the models is that none of them was able to correctly predict the most extreme class, which corresponds to *severe thunderstorms* (at least 50 millimeters per hour). This is probably due to few cases of extreme events in the training dataset, as they are very rare. In fact, since the start of operation, in 1997, the guaratiba station measured only eight extreme events, and the sao_cristovao station did not report a single extreme event.

The results using just one neighboring station (Table 2/Table 3 and Table 5/Table 6) seem to present some evidence that augmenting data from observations of a neighbor station helps in improving the predictive performance for moderate and strong rains in the WSoI. However, the results using four neighboring stations are inconclusive and need further investigation.

4. Final remarks

Precipitation forecasting in a nowcasting setting is important in Meteorology. This work aimed to report preliminary experiments on a method to perform data augmentation in a scenario in which several rain gauge stations are present together with a few meteorological stations in a certain spatial region. We performed preliminary experiments to validate

Table 2. Confusion matrix for model trained using `guaratiba` station data only.

	None	Weak	Moderate	Strong	Extreme
true(↓)/pred(→)					
None	15487	108	6	0	0
Weak	956	329	37	0	0
Moderate	38	38	28	0	0
Strong	3	1	4	0	0
Extreme	0	0	3	0	0

Table 3. Confusion matrix for the model trained using `guaratiba` station augmented with data from `tijuca` station.

	None	Weak	Moderate	Strong	Extreme
true(↓)/pred(→)					
None	15443	152	6	0	0
Weak	790	492	40	0	0
Moderate	31	49	24	0	0
Strong	2	2	3	1	0
Extreme	0	0	2	1	0

this method using observations from a subset of AlertaRio's rain gauges and meteorological stations. The corresponding preliminary results we present in this paper confirm that the method has the potential to improve predictive performance in a precipitation forecasting scenario. However, the work presents some limitations that we plan to tackle in future work. A limitation of this work is that ERA5 is a global model, and as such, it does not have the resolution or accuracy to capture the small-scale features of the atmosphere that are important for nowcasting. To overcome this limitation, we plan to investigate fusion approaches of the gauge stations' observations with other data sources, such as radar and satellite. We also plan to expand our analysis to consider a larger subset of stations (in our preliminary experiments, we only used two meteorological stations and four rain gauge stations from the AlertaRio system; besides, there is another observational system named *Websirenes*, which is comprised of 83 gauge stations). Finally, we plan to investigate cleverer ways to choose the k most similar stations for a given WSoI, by considering

Table 4. Confusion matrix for the model trained using `guaratiba` station augmented with data from `tijuca`, `tijuca_muda`, `saude`, and `grajau` stations.

	None	Weak	Moderate	Strong	Extreme
true(↓)/pred(→)					
None	15453	143	5	0	0
Weak	794	496	32	0	0
Moderate	27	56	21	0	0
Strong	1	3	3	1	0
Extreme	0	0	2	1	0

Table 5. Confusion matrix for model trained using `sao_cristovao` station data only.

	None	Weak	Moderate	Strong	Extreme
true(↓)/pred(→)					
None	13418	150	16	0	0
Weak	581	345	98	0	0
Moderate	41	21	33	0	0
Strong	2	1	5	0	0
Extreme	0	0	0	0	0

Table 6. Confusion matrix for the model trained using `sao_cristovao` station augmented with data from `tijuca` station.

	None	Weak	Moderate	Strong	Extreme
true(↓)/pred(→)					
None	13358	209	16	1	0
Weak	451	477	91	5	0
Moderate	29	32	30	4	0
Strong	1	2	4	1	0
Extreme	0	0	0	0	0

the distance between them and relief information.

References

Vladimir Braverman. *Sliding Window Algorithms*, pages 2006–2011. Springer New York, New York, NY, 2016. ISBN 978-1-4939-2864-4. doi: 10.1007/978-1-4939-2864-4_797.

Hans Hersbach et al. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020. doi: https://doi.org/10.1002/qj.3803.

Herrera Montano et al. Survey of techniques on data leakage protection and methods to address the insider threat. *Cluster Computing*, 25(6):4289–4302, 2022. ISSN 1573-7543. doi: 10.1007/s10586-022-03668-2. RIa.

Table 7. Confusion matrix for the model trained using `sao_cristovao` station augmented with data from `tijuca`, `tijuca_muda`, `saude`, and `grajau` stations.

	None	Weak	Moderate	Strong	Extreme
true(↓)/pred(→)					
None	13343	225	16	0	0
Weak	437	504	82	1	0
Moderate	27	34	33	1	0
Strong	1	2	4	1	0
Extreme	0	0	0	0	0