Spatiotemporal Extreme Event Prediction over the Indo-Gangatic Plain using Machine Learning

Project Report

Name:	**Jaydev Singh Rao**
Registration No./Roll No.:	19147
Institute/University Name:	IISER Bhopal
Program/Stream:	EECS
Problem Release date:	February 02, 2022
Date of Submission:	April 24, 2022

1 Introduction

In the given project we aim to predict visibility using machine learning methods. The given dataset consists of weather dataset collected over different locations in India for different times. Following are some of its major features:

- The recordings range from the year 1942 to the year 2021.
- There are a total of 934807 different recordings. Each recording has 122 features which describe the weather of the location at that particular time as well as some other information about the collected data.
- The target values are the visibility in kilometers corresponding to each kilometers. In the dataset, these fall in the range of 0 km to 100 km with average value of about 3.6 km (fig. 1).

The main aim of our findings is to highlight the importance of interpolation techniques for preparation of Spatiotemporal data for machine learning purposes.

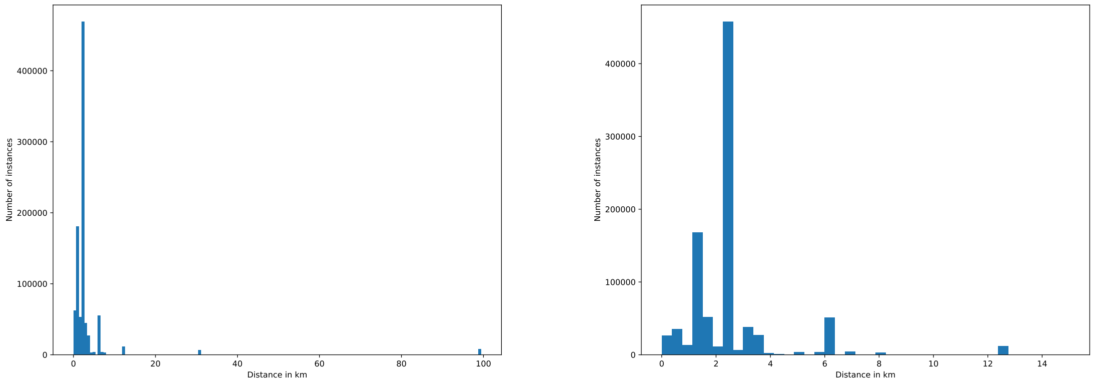

Figure 1: Histogram of visibility values in the dataset.

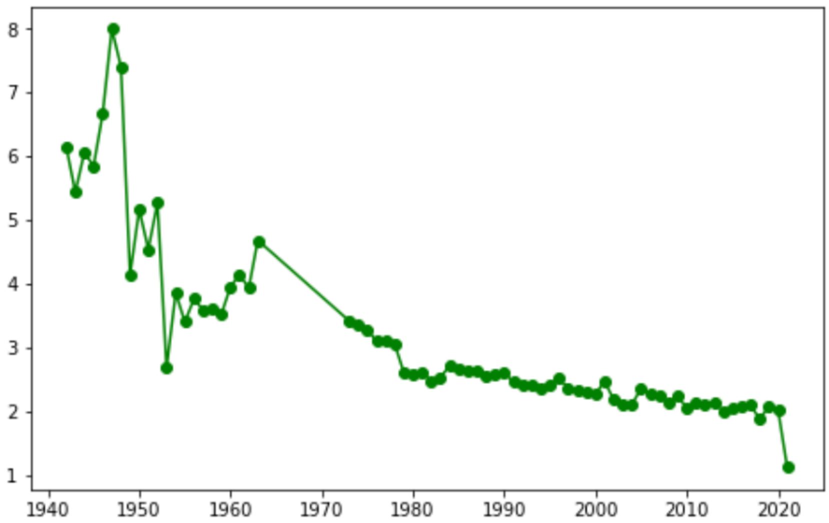

Figure 2: Trend of target over the years.

2 Methods

Data Cleaning

The data is very noisy and can not be directly used for model training. We performed some preprocessiong tasks for making the dataset useful for further applications:

- Most of the columns have all null values. We dropped such columns after which only 22 columns survived. Of these also, only one feature vector has no null values. This makes the dataset unusable without data filling techniques.
- Some features with numeric data types have some non-numeric noise in many entries. For dealing with such noise two methods were used. Some of the entries were of the form '11.2v' or '2.3s' etc., so from these we retrieved the numeric values. If no numeric values could be retrieved we replaced the entry with NaN which can be imputed later.
- Some outliers are present in the target variable. The typical range of the target is 0 to 100 km, while the mean is nearly 3.6 km (fig. 1). About 98% values lie in the range of 0 to 15 km, so the rest can be considered as outliers and hence are discarded.

Filling Missing Values

Traditionally, for filling missing values imputation techniques are used such as **simple imputation** in which null values are filled using global values of mean, median or mode (in case of discrete values)[1]. But the given dataset is unique in the sense that we are given with time and place of the particular observation. This motivated us to treat the dataset as a set of time series for different locations.

With this interpretation of the dataset we grouped the feature vectors into different locations and sorted then with time for each location and performed the following operations:

- **Yearly imputation**: Instead of finding global median for imputation, we find yearly value of median for different locations and fill the missing values using these values for corresponding year and locations. This methods can fill values in all the rows.
- **Interpolation**: We perform linear interpolation on the columns sorted with time for different locations. After this about 48% of the feature vectors have no null values (which is a great improvement over only 1 feature vector initially). This makes a large subset of the dataset directly usable without any other imputation techniques.

[1] https://scikit-learn.org/stable/modules/impute.html

Model Training and Hyperparameter Tuning

We are testing all of the above methods on the RandomForestRegressor from sklearn library[2]. This is a robust method that is suitable for our testing purpose. We then tune this model by grid search over the parameter spacee using GridSearchCV and report the performance of the best model that is found.

Github: *https://github.com/JaydevSR/Spatiotemporal_Extreme_Event_Prediction_using_ML*.

3 Evaluation

In order to compare different models and methods, we are using **mean squared error** and **mean absolute error**. We are comparing the performance of **random forest regression** model on the dataset after different operations:

Table 1: Performance Of Different Methods Using RFR

Method	Mean squared error	Mean absolute error
Global imputation	2.29	0.69
Yearly imputation	0.86	0.44
Pure Interpolation (only non-null rows used)	0.16	0.21
Interpolation with global imputation	0.68	0.35
Interpolation with yearly imputation	0.72	0.37

4 Analysis of Results

One of the important observations that was made is that simple imputation results is significant change in correlation between the target values and the features before and after simple imputation. Whereas after interpretation, the correlation drops only by a small amount, this indicates the importance of this methods in weather dataset where this correlation is important for predictions. See the below table where correlation of some of the features is reported (the correlation is given with the percentage of non-null values.)

Table 2: Correlation values at different steps

Feature	Initial	Global imputation	Interpolation
HourlyAltimeterSetting	-0.47 (23.4%)	-0.13 (100%)	-0.38 (48%)
HourlyPressureTendency	-0.072 (63.7%)	-0.032 (100%)	-0.33 (93.7%)
HourlyRelativeHumidity	-0.22 (98.0%)	-0.21 (100%)	-0.21 (99.9%)
HourlyStationPressure	0.30 (2.5%)	0.04 (100%)	-0.12 (86.8%)

In table 2 we can note the following things for different features:

1. HourlyAltimeterSetting: The drop in correlation after simple imputation is substantial. Although interpretation only doubles the number of non-null values (due to its limitations) the drop in correlation is small.
2. HourlyPressureTendency: The correlation drops for simple imputation whereas increases for interpolation. This can be explained by a linear trend that is present in the target values in the recent years (see fig. 2).
3. HourlyRelativeHumidity: The changes are not much in both cases as almost all rows are non-null.
4. HourlyStationPressure: The changes are arbitrary, this can be explained by the really small percentatge of non-null values initially. This implies that such features can not be used for training.

[2] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

5 Discussions and Conclusion

In the course of this project we have found that simple imputation techniques are not suitable for data filling in Spatiotemporal datasets. In such dataset interpolation techniques prove to be a better candidate for the data filling. Although such techniques can not completely fill all the missing values, they still make a significant fraction of dataset usable without significant drop in correlation between features and targets.

In future there is scope of further exploration on different interpolation techniques and their performance. We also note that the spatial and temporal information is important for weather datasets so many techniques used in time series analysis can also be important in Spatiotemporal dataset like autocorrelation functions, cross-correlation functions etc. This provides opportunity for further research in this direction.