Flood forecast by a deep learning method at Son Tay station on the Red river of Vietnam

Hoang Quy Nhan[1,*], Do Thi Lan[1], Khuat Thi Thanh Huyen[2], Nguyen Xuan Hoai[3], Ngo Van Manh[4]
[1]Faculty of Environment Thai Nguyen University of Agriculture and Forestry
[2]GeoInformatics Research Centre, TUAF
[3]AI Academy Vietnam, Ha noi
[4]Viet Nam Meteorological and Hydrological Administration
[*]Email: hoangquynhan@tuaf.edu.com, Tel: +84 985707811

Abstract

The Red River is located on Vietnamese territory with a total length of 510km, originating from Yunnan Province, China flowing through Vietnam, and discharging into the Southeast Asian Sea. This river makes an important contribution to ensuring food security and agricultural exports in Vietnam. However, climate change has recently caused severe impacts in the research area. Particularly, flood has destroyed houses, crops, roads, and basic infrastructure, resulting in highly vulnerable situations of the local people. This study aimed to forecast the Red River flow at Son Tay Station by using deep learning (DP) as a modeling tool and validating the accuracy of the model in comparison with the actual flow. A Keras DP model was formulated to simulate flows at a certain location in the river based on flows at upstream locations. Different procedures were applied to predict flooding by the Keras DP. Statistic data from Son Tay station between 2010 and 2018 were used to predict the possibility of flooding at Son Tay Station. Results of this study has proven that the Keras DP provides a reliable means of detecting flood hazards in the Red River delta.

Key words: *Deep Learning, Keras, Flood forecast, Red River, Son Tay*

1. Introduction

Currently, society has a constant development of science and technology, typically as the birth of the Internet and telephone devices, supercomputers, it has brought great changes in all areas of life. In the era of Internet of Things (IoT), Industry 4.0 and Artificial Intelligence with the application and extensive integration of mobile devices such as mobile phones, automobiles, and industrial machinery contribute to the creation and transformation of mobile devices. Data collection work is constantly generated every second, the term Big Data is used to refer to huge, largely unstructured datasets, collected from various sources. In particular, Big Data contains the number of valuable information. If successfully extracted, it will be a big help for business, scientific research, predicting weather and climate change, as well as the disease has potential to be arise and even determine the real-time traffic conditions. Outliers are observations that have extreme values relative to other observations observed under the same conditions. Observations may be outliers because of a single large or small value of one variable or because of an unusual combination of values of two or more variables [2]. Regarding floods on the Red River in this study, the unusually high water level values have been carefully studied and provided data for prediction rules by machine learning and deep learning

2. Subjects and research method

2.1 Research Subjects

- Flood history on the Red River,

- Flood value analysis and treatment of unusually high points of water level,
- Application of artificial intelligence to predict abnormal values (Outliers).

2.2. Research Methods

2.2.1. Method of collecting secondary data

The flood water level data was provided by the General Department of Meteorology and Hydrology. To consider whether that data is anomalous data or not, we have to consider the data on a time domain of 1 day, 1 week, 1 month, 1 year, between 2008 and 2018 [1, 4]. Because the data on that day may be anomalous data, it may not be abnormal in a week or month, not only when it detects anomalous data in a time domain. The abnormal data may be due to measurement errors, the data should be calibrated to fit the actual meteorological factors.

2.2.2. Method to determine the size and area of Red River.

The Red River is located on Vietnamese territory with a total length of 510km, originating from Yunnan Province, China flowing through Vietnam, and discharging into the Southeast Asian Sea [4]. This study aimed to forecast the Red River flow at Son Tay Station in Viet Nam using an deep learning (DP) as a modeling tool and validated the accuracy of the model against actual flow. Son Tay station is an automatic measurement of hydrological values.

2.2.3. Some methods of handling and forecasting water levels are unusually high

We look at dataset then the Water Level Value column on the rivers. Machine learning algorithms are very sensitive to the scope and distribution of attribute values. Data exceptions can damage and deceive the training process, resulting in longer training times, less accurate models, and ultimately worse results [5].

- Single and multivariate method:

Single variable method: This method searches for data points of extreme value on a variable.
Multivariate method: Here search for unusual combinations on all variables.

To find abnormal data that compares the intervals of contiguous data and then averages over a time domain with 1 result a, on average, we make a comparison vs that result if any data result that is abnormal data.

2.2.4. Use some tools of deep learning

Use some tools of deep learning to analyze and process predicting abnormal values in flood data on the Red River. NumPy is an acronym for "Numeric Python" or "Numerical Python" [2]. It is an open source extension module for Python, providing fast compilation functions for mathematical and numeric operations. Moreover, NumPy enriches Python programming language with powerful data structures to efficiently calculate multidimensional arrays and matrices. The implementation is even aimed at the giant matrix and array. Besides the module provides a large library of advanced math functions to operate on matrices and arrays. Numpy can be downloaded from the website: http://www.numpy.org.

Basically this is the data division algorithm that includes n initial observations into the cluster K so that homogeneous between the observations in the group is as high as possible. In other words, this algorithm groups the observations into K different clusters so that the

difference between the observations in each cluster is the lowest. That difference can be a characteristic or a certain set of attributes (often called attributes) of observations [3].

The homogeneity (or difference between observations) is quantified by the sum of the "distances" between observations in a sub-cluster that we mentioned above and will be optimal when sum This is the smallest possible.

We will use k-Means Clustering in the hydrological problem to predict the height of the water level. Our goal is to predict the water level above sea level. Based on the data available from the past, we will predict and compare reality with a certain scale.

3. Research results and discussion

3.1. Flood on the Red River

The Red River has a very large average flow of water, up to 2,640 m^3 / s (at the river mouth) with a total volume of 83.5 billion m^3, but the water flow is unevenly distributed. In the dry season, the flow decreases to about 700 m^3 / s, but in the peak of the rainy season it can reach 30,000 m^3 / s.

Red river water in the flood season has red-pink color due to the sediment that it carries, this is also the origin of its name. The sediment load of the Red River is very large, averaging about 100 million tons per year, ie nearly 1.5 kg of sediment per cubic meter of water.

Red River plays an important part in living life as well as in production. The silt helps make the field more fertile and at the same time, to build and expand the delta in the coastal areas of Thai Binh and Nam Dinh provinces. The fish source of the Red River has provided considerable seed for freshwater fish farming in the Northern Delta.

Due to the large amount of silt, the river bed is always filled, causing flooding to occur frequently, so long ago, both sides of the river were covered with large and small dykes to avoid flooding. In the past 10 years. For instance, In 2008: Flooded in a large area with high water column value due to continuous rain with great intensity from the night of October 30, 2008 onwards. In 2010: Rain and big flood cause at least 46 deaths and 21 missing people.

3.2. Detection, analysis and data processing abnormalities

We notice in the data when extracting to the screen, the column of date, time, water flow is adjacent to each other, we use Excel software to handle separating the text into separate columns (Figure 1).

	A	B	C	D	E	F	G
1	DataDate,DataHour/DataMinute/Flag/StationID/WaterLevelValue						
2	2011-05-02T17:00:00.000Z/01/00/1/74169/70.0						
3	2011-05-04T17:00:00.000Z/07/00/1/74169/80.0						
4	2011-05-07T17:00:00.000Z/13/00/1/74169/106.0						
5	2011-05-14T17:00:00.000Z/19/00/0/74169/124.0						
6	2011-05-19T17:00:00.000Z/01/00/1/74169/211.0						
7	2011-06-03T17:00:00.000Z/07/00/1/74169/132.0						
8	2011-06-05T17:00:00.000Z/07/00/1/74169/156.0						
9	2011-06-09T17:00:00.000Z/19/00/1/74169/97.0						
10	2011-06-10T17:00:00.000Z/19/00/1/74169/95.0						
11	2011-06-12T17:00:00.000Z/13/00/2/74169/119.0						
12	2011-06-23T17:00:00.000Z/07/00/1/74169/139.0						
13	2011-07-06T17:00:00.000Z/01/00/0/74169/226.0						
14	2011-05-01T17:00:00.000Z/07/00/1/74169/39.0						
15	2011-05-05T17:00:00.000Z/07/00/1/74169/105.0						
16	2011-05-05T17:00:00.000Z/19/00/2/74169/131.0						
17	2011-05-11T17:00:00.000Z/01/00/2/74169/131.0						
18	2011-05-18T17:00:00.000Z/13/00/1/74169/169.0						
19	2011-05-31T17:00:00.000Z/07/00/1/74169/103.0						
20	2011-06-03T17:00:00.000Z/01/00/1/74169/161.0						
21	2011-06-07T17:00:00.000Z/19/00/2/74169/84.0						

Figure 1. Basic Data Format

Processing data using Excel 2019 tool to separate text into columns (Figure 2).

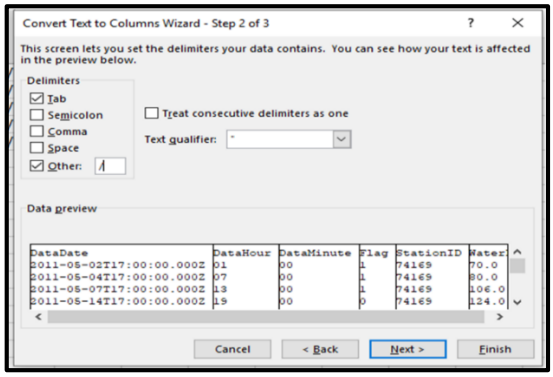
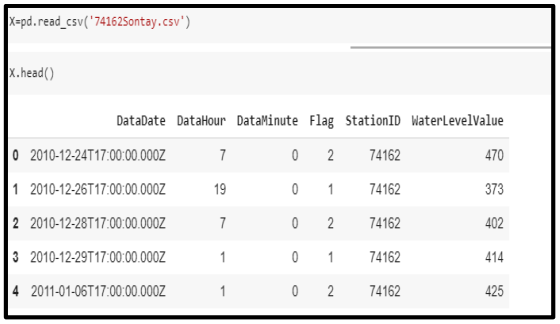


Figure 2.Analyze and process basic Data Form



```
X=pd.read_csv('74162Sontay.csv')
```

```
X.head()
```

	DataDate	DataHour	DataMinute	Flag	StationID	WaterLevelValue
0	2010-12-24T17:00:00.000Z	7	0	2	74162	470
1	2010-12-26T17:00:00.000Z	19	0	1	74162	373
2	2010-12-28T17:00:00.000Z	7	0	2	74162	402
3	2010-12-29T17:00:00.000Z	1	0	1	74162	414
4	2011-01-06T17:00:00.000Z	1	0	2	74162	425

Figure 3.Results of analysis and processing of Column Data Form

We see on the picture there are some unusual points compared to the usual data. Our mission is to identify and handle this unusual data (Figure 4).

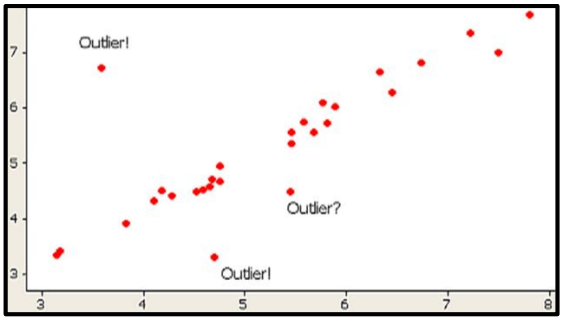


Figure 4.The data type has an abnormal value

The data consists of many characteristics (columns), and each property has different units and sizes. This affects the effectiveness of many algorithms, for example the execution time, the convergence process, or even the algorithm's accuracy. Therefore, it is often necessary to adjust the data so that the properties have the same data scaling. And often let properties have values within [0, 1]. The results will help many important algorithms in machine learning to use.

We will use the kmeans algorithm in the sklearn library to cluster the same data into a cluster.

3.3. Forecast results

The predictive structures are made on the adjusted data set, the water level of nx will be predicted by using data of four preceding results, which are nx-1; nx-2; nx-3; nx-4. The steps are replicated and cyclic until end of the data set. Then, the predicted results are compared

with the actual measurement results which measured at stations and calculate the deviation. Apply machine learning to change predictive structures so that the results are as close as possible to reality. [1] (Figure 5).

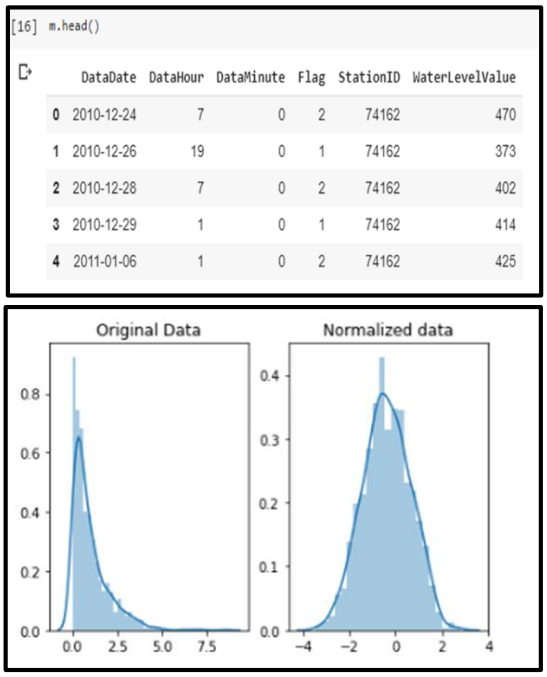
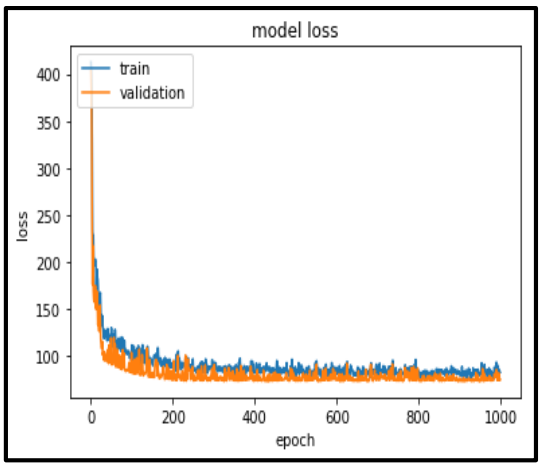


Figure 5.Forecasting results compared with real data

The blue line represented the set of predicted water level values, the orange line showed the actual measured water level values. The test set compared with the actual result set when undergoing the process of running the artificial intelligence model. The forecast results show that abnormal values are forecasted during the flood season in the period 2008 to 2018.At Son Tay station, the highest flood peak is in line with the forecast in 2018 with an accuracy of 96%.Using abnormal data to forecast floods by artificial intelligence on the river will open up a new research direction.



The biggest difficulty in the process of data analysis, we often observe observations whose value is very different from the value of other observations. Therefore, we must perform the initial pre-processing step. Preprocessing is the data preparation step that is often required

before performing machine learning algorithms, to help the algorithm be more efficient. In many problems, we have to use several different ways to handle data, to choose the most suitable way.

After simulating the processed data, we get the result. All water level data are considered to be visualized data by chart types.These data are predictive in the short term, in this study we take 4 forward values that correspond to 4 days to predict the next day's results.

```
[6]  import numpy as np
     import pandas as pd
     import seaborn as sns

[4]  df=pd.read_csv('./processing74162Sontay.csv')
```

Figure 6. Results of test run of forecast data

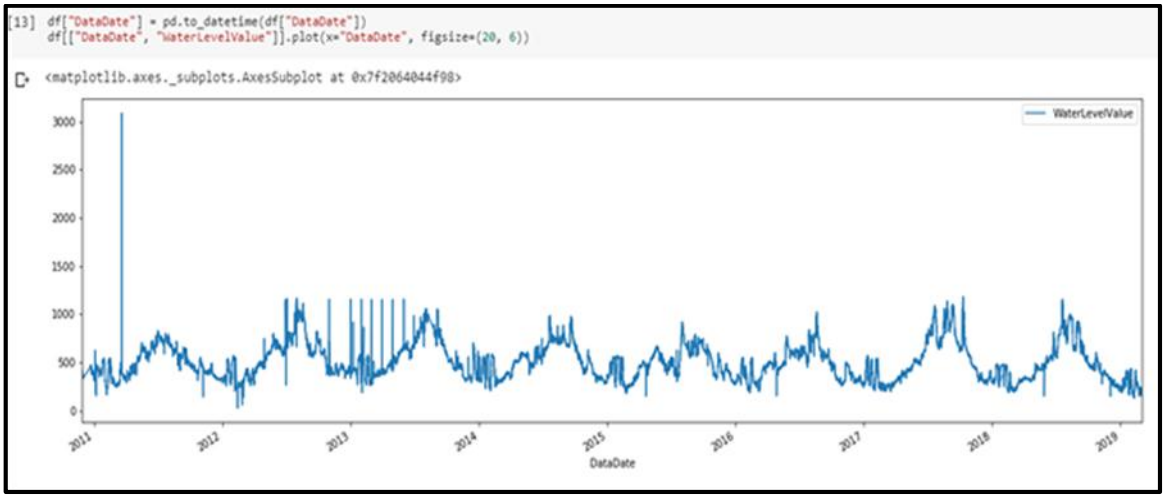


Figure 7.The graph of the value of the forecast results

Figure 7 showed that high water level values are often towards the end of years compared to actual data that coincide with climate events in Vietnam that is usually heavy rain and flooding from August to February of the following year.

The forecast results show that the overlap is quite accurate, the values of lesions are forecasted in a certain period of time, especially in the period of flood season through years.

4. Conclusions

The data extraction in bigdata has been very successful, helping a great deal in the study and prediction of changes in flood levels on the Red River in response to climate change between 2008 and 2018. However, Big data storage, resources, flow becomes a huge challenge. In addition, in the process of collecting and analyzing data, we often encounter observations whose values are very different from those of other observations, these are called abnormal values, those values. These are the values that the research interest.

The forecast results show that abnormal values are forecasted during the flood season in the period 2008 to 2018.At Son Tay station, the highest flood peak is in line with the forecast in 2018 with an accuracy of 96%.Using abnormal data to forecast floods by artificial intelligence on the river will open up a new research direction.

5. References

1.**Amir Ghaderi, Borhan M. Sanandaji*, Faezeh Ghaderi,** 2017, *Deep Forecast: Deep Learning-based Spatio-Temporal Forecasting*, International Conference on Machine Learning(ICML), Time Series Workshop, pp 52 – 55.

2. **Fernández, Á., González, A. M., Díaz, J., & Dorronsoro, J. R.** (2012, March). *Diffusion maps for the description of meteorological data. In International Conference on Hybrid Artificial Intelligence Systems*, pp. 276-287. Springer, Berlin, Heidelberg.

3. **Jagreet Kaur Gill**, 2018, *Introduction to Data Preparation and Preprocessing*, Data Science, Automation, Advanced Analytics and Artificial Intelligence, https://www.xenonstack.com/blog/data-preparation/.

4.**Ministry of Natural Resources and Environment (Monre),**2018, Protection of water resources in the Red River and Thai Binh Rivers: Urgent issues, Volume 1, Number 1, pp. 13 – 17. (Bảo vệ tài nguyên nước khu vực sông Hồng và sông Thái Bình: Những vấn đề cấp bách).

5.**R.Willink and B.D.Hall**, 2008, *A classical method for uncertainty analysis with multidimensional data*, Metrologia, Volume 39, Number 4, pp. 201 – 207. https://doi.org/10.1088/0026-1394/39/4/5