RESULTS OF APPLYING STANDARDIZED METHODS OF HYDROGRAPHIC DATA FOR STATIONS 74162 - SON TAY - VIET NAM

Hoang Quy Nhan[1], Do Thi Lan[1] and Nguyen Xuan Hoai[2]
[1]Faculty of Environment, Thai Nguyen University of Agriculture and Forestry (TUAF),
Thai Nguyen University. 10, Quyet Thang Commune, Thai Nguyen City, Vietnam
Email: hoangquynhan@tuaf.edu.vn
[2]AI Academy Vietnam (AIAV),
Inside Trong Dong Building, 389 Hoang Quoc Viet, Cau Giay, Hanoi, Vietnam,
Email: nxhoai@aiacademy.edu.vn

KEY WORDS: Artificial Intelligence, Missing values, Outliers, Time series, Water level.

ABSTRACT: Climate change has recently caused severe impacts. Particularly, flood has destroyed crops, houses, roads, ... resulting in highly vulnerable situations of the local people. Forecasting water level on Red River is an important task for flood warning. Currently, forecasting science along with the development of information technology, artificial intelligence and remote sensing has opened a new research direction. Water level data at river stations in Red river - Viet Nam are collected by automatic monitoring with frequency of collection depending on the time of year. These data need to be cleaned to eliminate outliers, missing values; standardized form of time series ... In our research, the authors will indicate the current status of water level data collected at the station 74162 - Son Tay. Our hybrid models are built and tested using big datasets from hydrological stations, namely, 74162 - Son Tay - Hanoi (both with collecteddata from 2011 to 2019). These are actual data, provided by the National Center for Hydrometeorological Forecasting. Based on the current status of this data set, experimental methods of Data processing to replace missing values with the method of interpolation and normalization of data in time series form shall be carried out with time spaced 3 hours apart. Theexperimental results show the effectiveness of the new approach in that the combinedmodel deeplearning of Artificial Intelligence. When there is complete data, ensuring the completeness and reliability will be the decisive factor to the accuracy of the prediction and forecast models.

1. INTRODUCTION

Water level data collected from monitoring stations on the river can be done through manual monitoring (directly recording the value of the measured element on the monitoring device) or automatic monitoring (recording the value of the measured element on the monitoring device). value of the factor measured by automatic equipment and transmitted to the user according to the needs) [1]. Currently, water level monitoring on river systems mainly still uses manual monitoring, the observer will record the value on the water level gauge and then send this data to the center for storage handle. Due to many subjective and objective factors, the process of recording values and sending monitoring data to the center is wrong, confused, and lost compared to the actual value. Moreover, depending on the time and season of the year, the water level monitoring regime is also different, maybe only 2 times/day (7h, 19h), 4 times/day (1h, 7h, 13h, 19h) or 8 times/day (1h, 4h, 7h, 10h, 13h, 16h, 19h, 21h). At the time of the dry season, or the beginning of the flood season when the daily water level amplitude is small; but can be increased to 12 times/day (1h, 3h, 5h, 7h, 9h, 11h, 13h, 15h, 17h, 19h, 21h, 23h), or 24 times/day (0h, 1h, 2h,..., 22h, 23h)… Applied in flood season when the water level changes during the day [1]. Therefore, the collected data is interrupted and discontinuous, the time of data collection is different depending on the season of the year, basin characteristics, rainfall characteristics, flood time... These are the data. recorded and stored over time, but not Time series data. Therefore, it is not possible to apply time series forecasting models such as MA, ARMA, ARIMA, PARMA, GARMA... or other models. Other machine learning and deep learning models in building water level prediction models at monitoring stations, serving flood warnings or other related problems [2-4]. It can be seen that currently collected and stored water level monitoring data are raw data, these data need to be normalized and cleaned (Data preparation) before using for any purpose. Whatever the purpose, this is a mandatory and indispensable step [5,6]. The results of many studies have shown that 80% of the time, effort and resources of a data science project is in data preparation. In the following sections of the article, the authors will learn about the collection method and current status of hydrological data at station 74162 – Son Tay in the 9-year period from January 1, 2011 to the end of December 31, 2019, thereby determining the necessary data normalization methods, suitable for this data set. The authors use libraries and programming techniques to build modules to remove outliers, missing data points and normalize water level data in time series form. The data preprocessing methods applied to station 74162 will serve as the basis for other hydrological monitoring stations on the Red River system in general.

2. DATA COLLECTION METHOD AND CURRENT STATUS HYDROLOGICAL DATA

2.1 Water level data collection method

Water level data at hydrological monitoring stations on the Red River in general and station 74162 in particular are collected by manual monitoring method. Every day, at the specified time, the observer will directly record the water level value on the monitoring device and then send this value to the Hydrometeorological and Information Center for storage, processing, and recovery. service for specific purposes. Figure 1 shows the locations of some stations on the Red River system, including station 74162 - Son Tay.

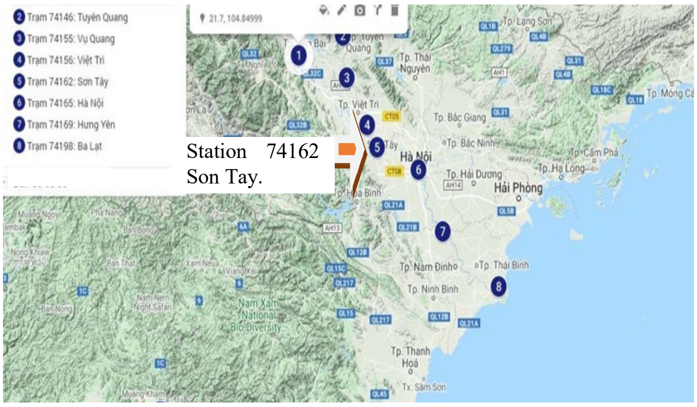
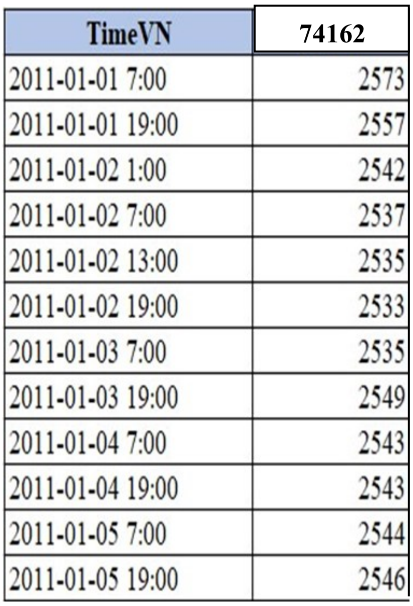


Figure 1: Location of station 74162 on the map

The water level monitoring regime must ensure to reflect the process of water level evolution fully, objectively and be feasible [1]. According to TCVN 12636-2:2019 with manual monitoring there are 8 modes. With station 74162, follow the monitoring modes from 1 to 6 depending on the specific conditions according to the season, the flood.... After recorded data, it will be sent to store in the database of the Center for Hydro-meteorological Information and Data. To facilitate the analysis, we have retrieved the hydrological data stored in MongoDB and split to get the data for the last 9 years (2011 - 2019); The data is then stored in a .CSV (Comma Separated Values) file named Data_waterlevel_74162, including the TimeVN attribute: Indicates the time of monitoring the water level in the format YYYY-MM-DD hh:mm; and attribute 74162: The monitoring value of the water level (Water level) of station 74162 corresponds to the time of monitoring, in cm. Figure 2 illustrates 12 the first row of data in the data set.



TimeVN	74162
2011-01-01 7:00	2573
2011-01-01 19:00	2557
2011-01-02 1:00	2542
2011-01-02 7:00	2537
2011-01-02 13:00	2535
2011-01-02 19:00	2533
2011-01-03 7:00	2535
2011-01-03 19:00	2549
2011-01-04 7:00	2543
2011-01-04 19:00	2543
2011-01-05 7:00	2544
2011-01-05 19:00	2546

Figure 2: File structure Data_waterlevel_74162.csv

2.2 Water level data mining at station 74162

Before introducing methods for processing and normalizing hydrological data for station 74162, it is necessary to explore and understand the details of the current status of these data. Table 1 shows the most general parameters of the monitoring data set.

Table 1. Statistics of monitoring parameters at station 74162

Parameter	Value
Starttime	2011-01-01 7:00 AM
Endtime	2019-12-31 7:00 PM
Total number of monitoring points	26 586 points
Mean water level	2668 cm
Standard Deviation (std)	176 cm
Minimum water level (min)	1 cm
Highest water level (max)	3312cm

Figure 3 shows a statistical chart of the number of monitoring points by year, through which we can see that the highest number of monitoring times varies from year to year, the highest is 2017 with 3635 monitoring times, the lowest is 2011 with 2002 time points. The difference is up to 1633 observation data points. Figure 4 shows the statistics of the number of monitoring points by month, we can see that the frequency of monitoring water level data changes from month to month of the year, high frequency in the period from May to October every year, the highest concentration is in July and August; It also reflects the general weather of the area when this period is in the flood season and the peak of rain and flood mainly falls in July and August.

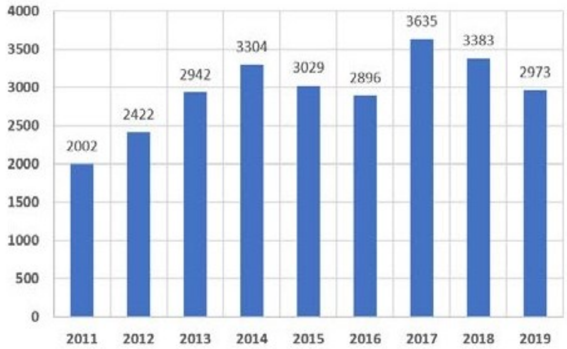


Figure 3: Statistical chart of the number of monitoring points by year

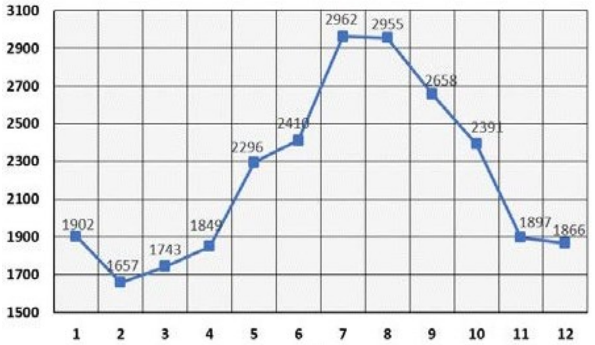


Figure 4: Statistical chart of the number of monitoring points by month

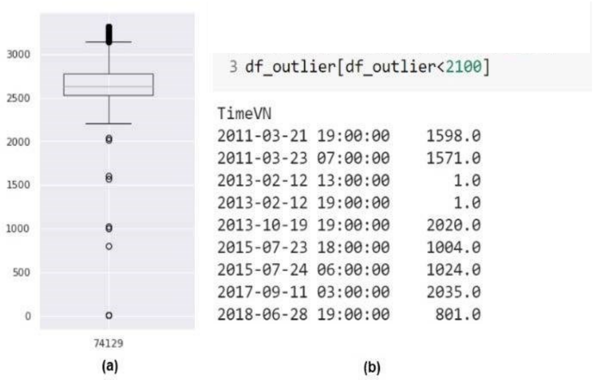
3. NORMALIZATION OF HYDROLOGICAL DATA OF STATION 74162

3.1 Detect and process outliers

Water level data at station 74162 is collected by manual monitoring method, so in the process of data recording and transmission to the storage center due to subjective and objective reasons, errors may occur. cause data to be erroneous or anomalous. These data points are called outliers. An outlier is a data point that is significantly different from the rest of the data set. Outliers are often viewed as special data patterns, far removed from most other data in the data set [7]. There are many methods to detect outliers such as: Extreme Value Analysis; Probabilistic and Statistical Models; Linear Models; Proximity - based Models; The models are based on information theory (Information Theoretical Models) [7,8,9]. The collected water level data is one-way data, so a simple and effective way to detect these outliers is to use extreme value analysis. Two effective methods for detecting extreme values include Z-Scores and Box-plot plots [10].

In the experimental content for station 74162, the authors use Python programming language, combined with a number of open source libraries to support analysis, processing and visualization including: Pandas, Numpy and Matplotlib, the entire source code is written on the Google Colab system. To detect outliers for the observed water level data set, the authors use Box-plot chart. Box-plot plots are used to measure the dispersion tendency and identify outliers of the data set [10].

Figure 5(a) is a Box-plot plot of the data set. Data points outside the lowest horizontal line in the Box-plot are considered as left outliers. Figure 5(b) lists a list of 9 monitoring points with minimum values in the data set that are far from most other points. To be able to confirm these are outlier data points? As well as coming up with a suitable treatment plan for these points, we need to perform verification. In the following section, the authors perform verification for 2 data points with outliers recorded at 7:00 p.m. on March 21, 2011 and 7:00 a.m. on March 23, 2011, outliers verification for other points will be performed similarly.



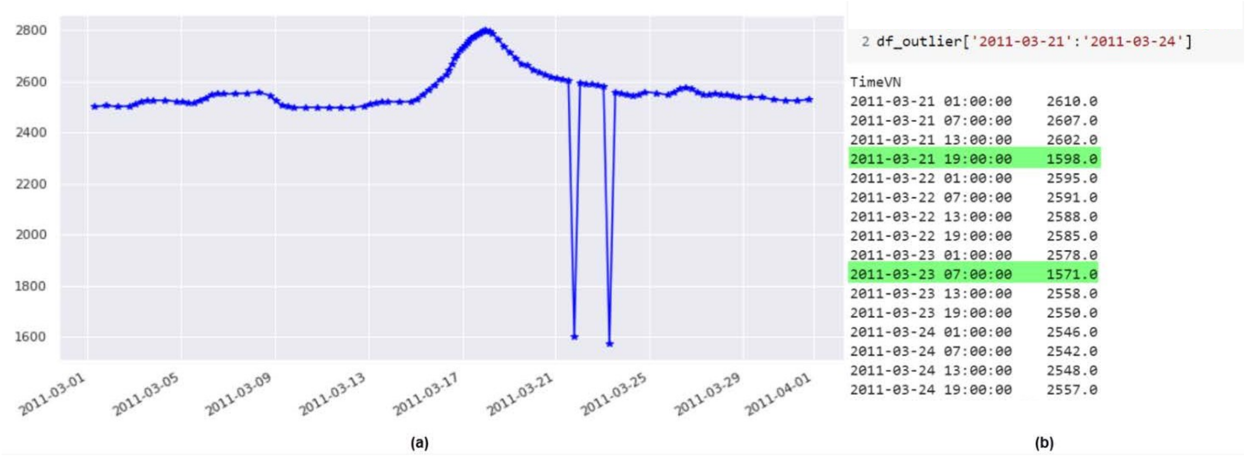
```
3 df_outlier[df_outlier<2100]

TimeVN
2011-03-21 19:00:00    1598.0
2011-03-23 07:00:00    1571.0
2013-02-12 13:00:00       1.0
2013-02-12 19:00:00       1.0
2013-10-19 19:00:00    2020.0
2015-07-23 18:00:00    1004.0
2015-07-24 06:00:00    1024.0
2017-09-11 03:00:00    2035.0
2018-06-28 19:00:00     801.0
```

Figure 5: The box-plot plot of the data set (a); List of monitoring points for external consideration left (b)

As shown in Figure 6(a), it can be seen immediately that the water level at Son Tay station in the period of March 2011 has 2 monitoring points with sudden changes in value. Figure 6 (b) shows the difference in water level of these two monitoring points compared to neighboring monitoring points; At 19:00 on March 21, 2011, the water level data recorded 1598cm while at the time of observation before it at 13:00 on March 21, 2011 it was 2602cm (the reduction difference between the two monitoring times is - 1004cm) and the time immediately after 1h on 22/03/2011 is 2595cm (the increasing difference between the two monitoring times is +997 cm). The level of sudden change also happened similar to the time at 7 am on 23/03/2011. March is the dry season period, according to the data, the monitoring regime is following mode 2 (6 hours every time at 1h, 7h, 13h, 19h), so it can be confirmed. are outliers, the recorded and stored data have completely deviated from the actual data. Outliers data points have a great influence on the accuracy of predictive models,

forecast. Therefore, it is imperative that they be detected and handled. The above has shown how to detect these points, the question is how to deal with these outliers? There are 3 methods used to handle outliers including: Remove lines containing outliers from the data set; Replace outliers with a more suitable one; Replace outliers with NULL (empty) values, considering this as a missing data point [11]. There is no general outlier data processing method that is applicable to all problems [12], so to choose the right method requires a deep understanding of the data set, the problem. problem solving, can use only one method and/or a combination of all three groups of methods above. And in fact, with hydrological data of station 74162, to process outliers, the authors have used all three of these methods in each specific case. In the case of outliers recorded at 7:00 p.m. on March 21, 2011 and at 7:00 p.m. on March 23, 2011, it can be seen that this outlier is caused by human subjective factors while recording and sending data. data to the storage center. This is the dry season month, the water level is tending to decrease and the intensity of change is low. The actual values in this case are 2598cm and 2571cm but have been skewed to 1598cm and 1571cm. Therefore, in this case, the treatment method will be used to replace the outlier value with a more suitable new value.

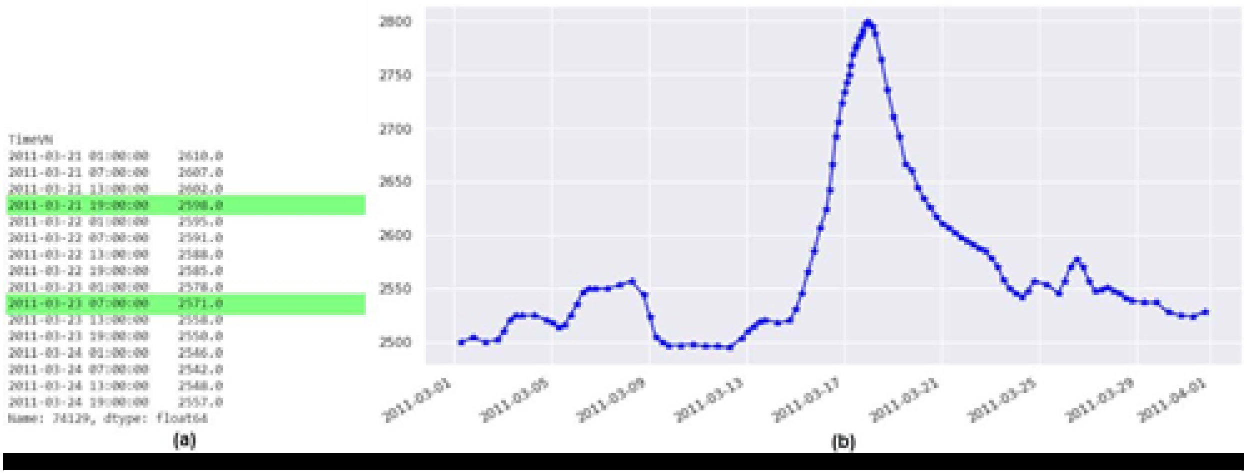


```
2 df_outlier['2011-03-21':'2011-03-24']

TimeVN
2011-03-21 01:00:00    2610.0
2011-03-21 07:00:00    2607.0
2011-03-21 13:00:00    2602.0
2011-03-21 19:00:00    1598.0
2011-03-22 01:00:00    2595.0
2011-03-22 07:00:00    2591.0
2011-03-22 13:00:00    2588.0
2011-03-22 19:00:00    2585.0
2011-03-23 01:00:00    2578.0
2011-03-23 07:00:00    1571.0
2011-03-23 13:00:00    2558.0
2011-03-23 19:00:00    2550.0
2011-03-24 01:00:00    2546.0
2011-03-24 07:00:00    2542.0
2011-03-24 13:00:00    2548.0
2011-03-24 19:00:00    2557.0
```

Figure 6. The chart shows the observed water level value of station 74162 during the period of March 2011 (a); List of monitoring times and recorded water level values from March 21 to March 24, 2011 (b).

Figure 7 illustrates the alternative method and the result after processing these two outliers. On the basis of the method and method described above, the verification and outliers will be performed for the entire data set. After this step, outliers in the hydrological data set of station 74162 have been processed.

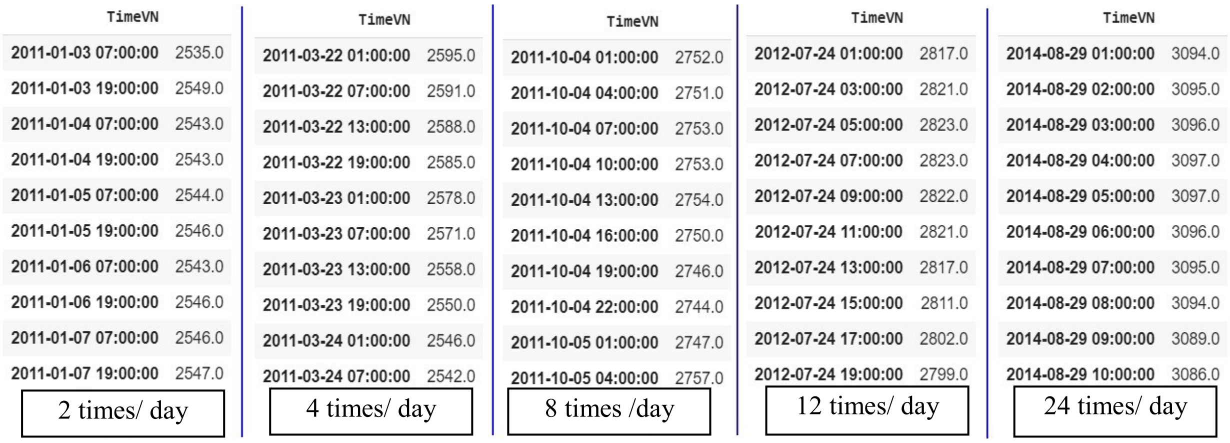


```
TimeVN
2011-03-21 01:00:00    2610.0
2011-03-21 07:00:00    2607.0
2011-03-21 13:00:00    2602.0
2011-03-21 19:00:00    2598.0
2011-03-22 01:00:00    2595.0
2011-03-22 07:00:00    2591.0
2011-03-22 13:00:00    2588.0
2011-03-22 19:00:00    2585.0
2011-03-23 01:00:00    2578.0
2011-03-23 07:00:00    2571.0
2011-03-23 13:00:00    2558.0
2011-03-23 19:00:00    2550.0
2011-03-24 01:00:00    2546.0
2011-03-24 07:00:00    2542.0
2011-03-24 13:00:00    2548.0
2011-03-24 19:00:00    2557.0
Name: 74129, dtype: float64
```

Figure 7: Handling outliers by replacing with new value (a); The graph shows the water level data in March 2011 after processing outliers (b).

3.2 Normalize data to the form of time series

Time series data is a series of data points measured at consecutive intervals, with equal distances between measurements [2]. Water level data of station 74162 was collected during the period from 1 a.m. on January 1, 2011 to 11 p.m. on December 31, 2019. However, as discussed in the problem section, the frequency of water level data collection varies greatly depending on the time of year, as well as on the intensity and severity of each flood. With station 74162, perform data collection in 6 different modes from mode 1 to mode 6. Figure 8 shows data collected at some time corresponding to different monitoring modes. These are the 2 peak months in the flood season, the monitoring regime is mainly according to regimes 5 and 6. Thus, it can be seen that hydrological observation data is collected according to specific time points in hours, but this is not time series data because the distance between observations is not evenly spaced, depending on the time series. depending on specific conditions (the dry season is much smaller than the flood season). Because it is not time series data, it is not possible to use time series forecasting models such as MA, ARMA, ARIMA…[4]. Therefore, it is necessary to normalize this data to the form of time series so that the above prediction and prediction models can be applied.



TimeVN	
2011-01-03 07:00:00	2535.0
2011-01-03 19:00:00	2549.0
2011-01-04 07:00:00	2543.0
2011-01-04 19:00:00	2543.0
2011-01-05 07:00:00	2544.0
2011-01-05 19:00:00	2546.0
2011-01-06 07:00:00	2543.0
2011-01-06 19:00:00	2546.0
2011-01-07 07:00:00	2546.0
2011-01-07 19:00:00	2547.0

2 times/ day

TimeVN	
2011-03-22 01:00:00	2595.0
2011-03-22 07:00:00	2591.0
2011-03-22 13:00:00	2588.0
2011-03-22 19:00:00	2585.0
2011-03-23 01:00:00	2578.0
2011-03-23 07:00:00	2571.0
2011-03-23 13:00:00	2558.0
2011-03-23 19:00:00	2550.0
2011-03-24 01:00:00	2546.0
2011-03-24 07:00:00	2542.0

4 times/ day

TimeVN	
2011-10-04 01:00:00	2752.0
2011-10-04 04:00:00	2751.0
2011-10-04 07:00:00	2753.0
2011-10-04 10:00:00	2753.0
2011-10-04 13:00:00	2754.0
2011-10-04 16:00:00	2750.0
2011-10-04 19:00:00	2746.0
2011-10-04 22:00:00	2744.0
2011-10-05 01:00:00	2747.0
2011-10-05 04:00:00	2757.0

8 times /day

TimeVN	
2012-07-24 01:00:00	2817.0
2012-07-24 03:00:00	2821.0
2012-07-24 05:00:00	2823.0
2012-07-24 07:00:00	2823.0
2012-07-24 09:00:00	2822.0
2012-07-24 11:00:00	2821.0
2012-07-24 13:00:00	2817.0
2012-07-24 15:00:00	2811.0
2012-07-24 17:00:00	2802.0
2012-07-24 19:00:00	2799.0

12 times/ day

TimeVN	
2014-08-29 01:00:00	3094.0
2014-08-29 02:00:00	3095.0
2014-08-29 03:00:00	3096.0
2014-08-29 04:00:00	3097.0
2014-08-29 05:00:00	3097.0
2014-08-29 06:00:00	3096.0
2014-08-29 07:00:00	3095.0
2014-08-29 08:00:00	3094.0
2014-08-29 09:00:00	3089.0
2014-08-29 10:00:00	3086.0

24 times/ day

Figure 8: Water level monitoring modes at station 74129

The authors propose a way to normalize this data set to the form of time series as follows:
- Step 1: Determine the time interval t evenly spaced between observations. The parameter t is used as the basis to normalize the data to the form of time series with the observed time intervals t apart. With the hydrological data of station 74162, the parameter t is selected by the hour, which can be 1h, 2h, 3h... According to the statistics shown in the chart Figure 4, we see that in the time period from five from 2011 to 2019, the monitoring time focuses mainly on the time points of 1h, 4h, 7h, 10h, 13h, 16h, 19h, 22h of the day (> 2000 observations), other monitoring times in the rest of the day. day 0h, 2h, 3h, 5h, 6h, 8h, 9h, 11h, 12h, 14h, 15h, 17h, 18h, 20h, 21h, 23h with very few points (<600 observations); Therefore, with this dataset, we will choose the parameter t = 3, that is, we will normalize the collected hydrological data to the form of a time series with an equally spaced sampling interval of 3h (mode 3: 8 times/day).
- Step 2: Perform the filtering of the missing sampling times in the data set corresponding to the time interval t = 3h at 1h, 4h, 7h, 10h, 13h, 16h, 19h, 22h of the day . Figure 11 illustrates the source code that performs the statistics of the points and the list of points without data. According to statistics, if normalized in time series form according to mode 3, the data set lacks 4725 monitoring points. Insert these missing sampling times into the data file of station 74162 with NULL values (consider these as missing values). The time of non-observation has been added to the data set in step 2.
- Step 3: There are many methods of processing missing data, the content of step 3 will be detailed in section 3.3.
- Step 4: Normalize the data set to the form of time series; At the end of step 3, the hydrological data set of station 74162 has been processed with missing data. However, this data set also contains many other observation times besides the above 8 times corresponding to the monitoring periods according to regimes 4, 5 and 6. Therefore, the authors performed the extraction and filtering. data from the original set at positions 1h, 4h, 7h, 10h, 13h, 16h, 19h, 22h daily to obtain hydrological data in the form of time series with 3h interval.

3.3 Handling missing values in data set

Missing values is always an important and mandatory step in the data cleaning process [5]. Due to many subjective and objective reasons during data collection can lead to missing values. As described in section 3.2, in order to normalize the form of time series with 3h sampling interval, it is necessary to insert into the dataset these unobserved times with Null data (regard these as data points). missing values - missing values). Therefore, it is required to handle these missing data. There are many methods to handle missing data [13-14], these methods can be grouped into two main groups: Remove rows or columns of data containing missing values from the data set; Replace missing data points with a new, algorithm-specific value. With time series data, we cannot remove the missing data lines, but we can only use the second group of methods, which is to replace it with a new value. With time series data, the data points will have a relationship with the points before and behind it, and follow trends and seasons. There are 4 simple but effective solutions to deal with missing data for time series including:
- Replace the missing value with the previous value (Last observation carried forward - LOCF);
- Replace the missing value with the following value (Next observation carried backward - NOCB);
- Replace missing value with Linear interpolation;

- Replace missing values with spline interpolation.

With the characteristics of hydrological data of station 74162, the authors use the 3rd order Spline interpolation method to handle missing values. Spline interpolation is a method of constructing smooth curves that pass through n + 1 known data points (x_0, y_0),..., (x_n, y_n). In fact, go find one function f(x) such that f(x_i) = y_i for all i. We will define n polynomials of degree p_0,…., p_{n-1} such that f(x) = p_i(x) for every x in the interval [x_i, x_{i+1}] [15]. In fact, the authors use spline interpolation with Polynomial of degree 3 then pi(x) is defined as follows:

$$\mathbf{p_i(x) = a_i(x - x_i)^3 + b_i(x - x_i)^2 + c_i(x - x_i) + d_i \text{ [16]}}$$

Figure 9 illustrates the construction of quadratic curves (red line) passing through 14 known points (black dot).

Applying to hydrological data of station 74162, Figure 10a shows the first 10 data points in January 2012 containing missing value points at 4h, 10h, 16h, 22h (in Pandas the missing value is denoted as NaN) and a graph showing the observed water level values in January 2012 - Figure 10b.

Figure 11a is the result after processing the missing value with the 3rd order Spline interpolation method for the data points depicted in Figure 13a as well as the graph showing all the data of station 74162 in January 2012 including: including observation data and interpolated data for missing points (Figure 11b).

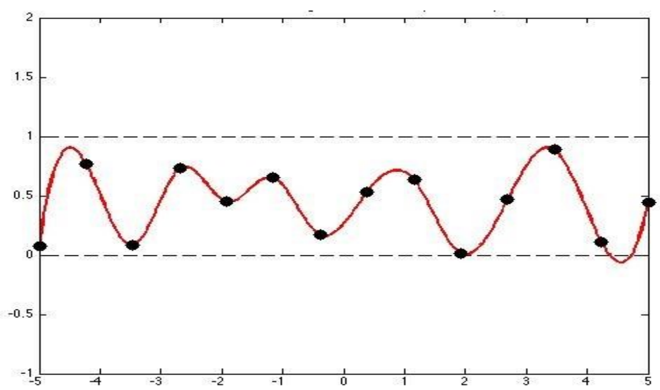
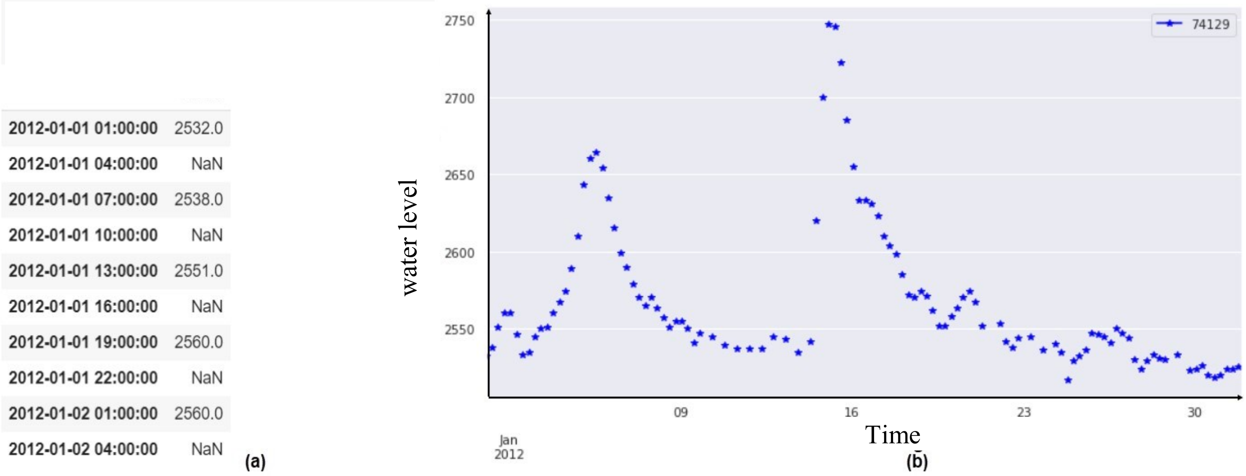
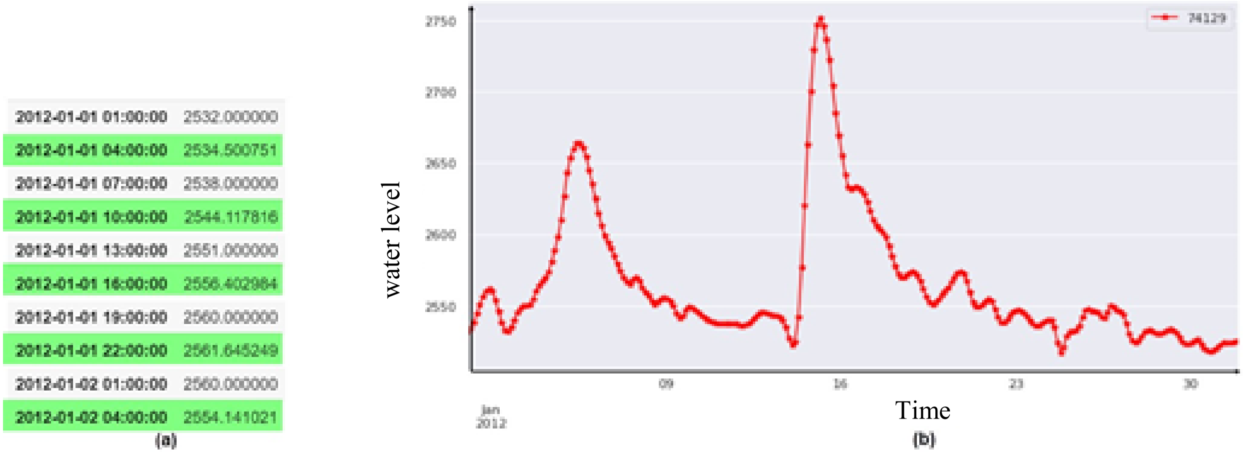


Figure 9: Interpolate 3rd order Spline over 14 known points



2012-01-01 01:00:00	2532.0
2012-01-01 04:00:00	NaN
2012-01-01 07:00:00	2538.0
2012-01-01 10:00:00	NaN
2012-01-01 13:00:00	2551.0
2012-01-01 16:00:00	NaN
2012-01-01 19:00:00	2560.0
2012-01-01 22:00:00	NaN
2012-01-02 01:00:00	2560.0
2012-01-02 04:00:00	NaN

Figure 10: Data before processing missing value (a) and Graph representing data in January 2012 (b)



2012-01-01 01:00:00	2532.000000
2012-01-01 04:00:00	2534.500751
2012-01-01 07:00:00	2538.000000
2012-01-01 10:00:00	2544.117816
2012-01-01 13:00:00	2551.000000
2012-01-01 16:00:00	2556.402984
2012-01-01 19:00:00	2560.000000
2012-01-01 22:00:00	2561.645249
2012-01-02 01:00:00	2560.000000
2012-01-02 04:00:00	2554.141021

Figure 11: Data after processing missing values by spline(a) interpolation method and Graph representing data after processing in January 2012(b)

4. RESULTS DATA NORMALIZATION STATION 74162

After performing the steps of preprocessing and normalizing the data presented in part 3, a new hydrological data set of 74162 - Son Tay station will be obtained, which is saved with the name Data_processed_74162.csv, this dataset also has the structure. The structure is similar to the original raw data set with 2 columns, TimeVN showing the monitoring time and column 74162 showing the water level value corresponding to each monitoring time. The data set after normalization has processed outliers, processed missing data points and returned to the form of time series with time interval t = 3h. Table 2 describes the main statistical characteristics and Figure 12 shows the Histogram of the station water level data set 74162 after normalization.

Table 2. Statistics of monitoring parameters at station 74162

Parameter	Value	
	Original data	**Processed data**
Starttime	2011-01-01 7:00 AM	2011-01-01 7:00 AM
Endtime	2019-12-31 7:00 PM	2019-12-31 7:00 PM
Total number of monitoring points	26 586 points	26 296 points
Mean water level	2668 cm	2631 cm
Standard Deviation (std)	176 cm	151 cm
Minimum water level (min)	1 cm	2406 cm
Highest water level (max)	3312cm	3394cm

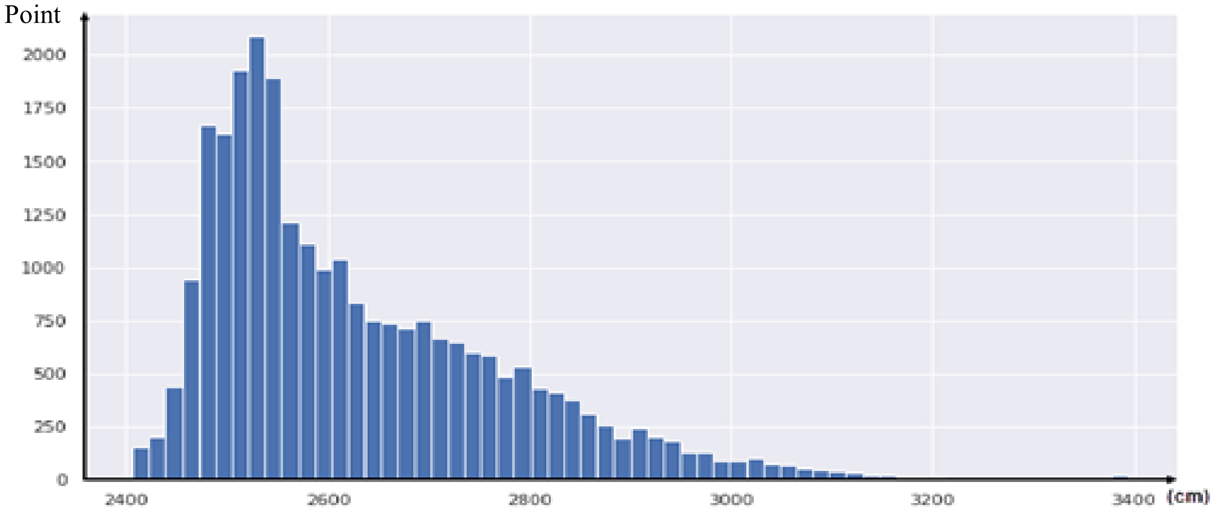


Figure 12 : Histogram of processed data set Data_processed_74162

The result obtained is a normalized data set, this data set can be used as input for predictive models, predicting time series such as MR, ARMA, ARIMA... or as data. input for machine learning models, deep learning.

5. CONCLUSION

The collected water level data are all raw data, need to be normalized and cleaned to remove outliers from the data set, outliers have a great influence on the accuracy of the water level. predictive models. Handling missing values is also a mandatory requirement in the data cleaning process, for each problem, a specific data type applies its own processing methods. At the same time, to be able to use time series forecasting models, the input data must be normalized to this form. The article analyzed in detail the collection method and current status of hydrological data of station 74162 - Son Tay, thereby standardizing this data by solving 3 main problems including: Detecting and processing. alien reason; Normalization of time series format; Handle missing values. The result after performing this whole process is a normalized and cleaned data set, which can be used as input for time series forecasting models, machine learning, deep learning. . The processing methods and techniques applied to station data 74162 can be used for other hydrological stations on the Red River system in general.

6. REFERENCES

References from Journals:

Ajao, I.O., Ibraheem, A.G., Ayoola, F.J. (2012), Cubic spline interpolation: A robust method of disaggregating annual data to quarterly series. Journal of Physical Sciens and Environmental Safety, 2 (1), 1-8.

Bonander, C., Strömberg, U. (2018), Methods to handle missing values and missing individuals. European Journal of Epidemiology, 34, 5-7.

Choi, J., Dekkers, O.M., le Cessie, S. (2018), A comparison of different methods to handle missing data in the context of propensity score analysis. European Journal of Epidemiology, 34 (1),23-36.

Dang Van Nam, Nong Thi Oanh, Ngo Van Manh, Nguyen Xuan Hoai, Nguyen Thi Hien (2020), Detection and processing of outliers for temperature data at 3h monitoring stations in Vietnam. Journal of Science and Technology Mining - Geology, 61 (1), 132-146.

Erdogan KAYA. Spline Interpolation Techniques. Journal of Technical Science and Technologies, 2 (1), 47-52.

Ranga Suri, N.N.R., Murty, N.M, Athithan, G. (2018), Outlier Detection: Techniques and Applications, IJCSI International Journal of Computer Science Issues, 9 (1), 307-323

Zhang, A., Song, S., Wang, J., Yu, P.S. (2017), Time series data cleaning: From anomaly detection to anomaly repairing. Proc. VLDB Endownment, 10 (10), 1046-1057.

References from Books:

Aggarwal, C.C. (2017), Outlier Analysis, Springer International Publishing AG, New York， Part 2, pp.454-459.

Box, G.E., Jenkins, G.M., Reinsel, G.C., Ljung, G.M. (2015), Time Series Analysis: Forecasting and Control. Hoboken, NJ, USA: Wiley.

Brockwell, P.J., Davis, R.A. (2016), Introduction to Time Series and Forecasting. Basel, Switzerland: Springer.

Munzer, T. (2014), Visualization Analysis and Design, CRC Press, 428 p.

National standard (2019), TCVN 12636-2:2019. Hydrometeorological observations-Part 2: Monitoring of water level and river water temperatur. Part 2, pp.4-9.

Shumway, R.H., Stoffer, D.S. (2017), Time Series Analysis and Its Applications: With R Examples. Cham, Switzerland: Springer, 562 p.

Song, S., Cao, Y., Wang, J. (2016), Cleaning timestamps with temporal constraints. Proc. PVLDB, 9 (10), 708-719.

Wang, X., Wang, C. (2019), Time Series Data Cleaning: A Survey, IEEE Access, 1866-1881.