Research Article

Lingkui Meng, Xiaobing Wei, Miao Yang, Yizhuo Meng, Yang Chen, Jianguo Cheng, and Wen Zhang*

A detection method for reservoir waterbodies vector data based on EGADS

https://doi.org/10.1515/geo-2020-0205
received October 06, 2019; accepted October 13, 2020

Abstract: Owing to the effects of camera, illumination, extraction algorithm defect, and other reasons, vector data for reservoir waterbodies extracted from remote sensing data may have quality issues, impacting the efficiency of data utilization in areas such as water resource management and reservoir monitoring. To efficiently detect abnormal data from massive vector products of reservoir waterbodies, a semi-automatic detection method for reservoir waterbody vector data is presented. The method has three phases. First, the original reservoir vector data are preprocessed to obtain the time series of the area of reservoir waterbodies. Second, data modeling with time series of reservoir waterbodies area data is done using the extensible generic anomaly detection system (EGADS) plug-in framework and time series modeling is conducted using the Olympic model. Third, data that have quality problems are identified with $K\sigma$ model was used to determine the outliers; thereby, the date of the outliers is detected. Results of accuracy verification show that the sensitivity and specificity of the method were 94.44 and 83.87%, respectively, showing its feasibility for use in anomaly detection in polygonal reservoir waterbody vector data with far greater efficiency than traditional manual inspection.

Keywords: EGADS plug-in framework, reservoir, vector data, time series analysis, anomaly detection

1 Introduction

Water conservancy information can serve as an important basis for management, forecasting, early warning, and emergency response related to water disasters [1,2]. Reservoirs are important man-made water conservancy facilities built for improved control and utilization of water resources. It is of great significance to understand the water volume change of reservoirs in real time [3]. There are a total of more than 98,000 reservoirs in China [4]. Remote sensing images can be used to obtain surface water information [5,6], and polygon vector data are its main product form. The quality management of the reservoir water vector data plays an important role in ensuring the provision of water conservancy information.

The commonly used water extraction algorithms based on multispectral images can be divided into four categories: the single-band threshold method [7], the multiband spectral relationship method [8,9], the water index method [10–12], and the remote sensing image segmentation method [13,14]. These methods can automatically and effectively extract water information from an image. Among them, water index methods (such as normalized difference water index [NDWI]) are easy to realize, efficient, and often used in engineering practice. However, due to cloud cover [15–17], shadow [9], image width [18], and defects in the algorithm, the results of water extraction appear as abnormal values. Therefore, it is necessary to further screen out the abnormal water extraction results.

In recent years, with the advent of the era of remote sensing big data, multitemporal remote sensing monitoring has become a research hotspot [19], and long

* **Corresponding author: Wen Zhang,** School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China, e-mail: Wen_zhang@whu.edu.cn
Lingkui Meng: School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China, e-mail: lkmeng@whu.edu.cn
Xiaobing Wei: School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China, e-mail: xiaobing_wei@whu.edu.cn
Miao Yang: School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China; Shanghai Baosight Software Co., Ltd, Shanghai 201900, China, e-mail: yanglosm@126.com
Yizhuo Meng: School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China, e-mail: yangguangmeng05@whu.edu.cn
Yang Chen: Wuhan Digital Engineering Institute, Wuhan 430070, China, e-mail: 46310402@qq.com
Jianguo Cheng: Information Center (Hydrology Monitor and Forecast Center), Ministry of Water Resources, Beijing 100053, China, e-mail: chengjg@mwr.gov.cn

time series monitoring has turned into one of the important applications of water conservancy remote sensing. Meanwhile, the water conservancy service platform has been developed for monitoring of geographical situation, which can realize the automatic production and the distribution of water conservancy products [20,21] and has led to the rapid increase of water monitoring products. However, the traditional vector data detection method needs to manually load the extracted vector data and the corresponding image data for visual evaluation, which consumes a lot of human and material resources [22,23]. Meanwhile, automatic anomaly detection of vector data products can improve the accuracy and update efficiency of water conservancy monitoring database.

Early work on automatic anomaly detection for vector data focused on testing spatial information, attribute data, and topological relations. Wei and Liu [24] proposed an extended model and the sorting method for the GIS real estate data, which allowed for the quality control of spatial information and attribute data and improved the efficiency of data testing. Popescu et al. [25] used the vector models of data representation in cadastral maps and developed a system for spatial location verification and attribute data constraint integration to assist in the automated management of cadastral information. Automated data quality inspection methods have been developed that are aimed at a diverse range of vector data in the geographical national conditions census and apply rule-based detection models for inspecting the positional accuracy (position edge) and attribute correctness of the vector data [26,27]. To solve the problem of attribute uncertainty in water vector data, Guo et al. [28] considered the data structure and topological relations synthetically and proposed an algorithm for automatic correction of linear water attributes. Gui et al. [29] noted that the quality of cadastral data includes location accuracy, attribute correctness, and topological consistency. In their study, verification and modification methods were designed for the topological inconsistency of cadastral spatial data, including node matching errors, cracks, and superposition. Zhang et al. [30] designed a topological consistency detection algorithm that integrated qualitative and quantitative location features of planar entities and allows for automatic detection of topological conflicts in polygonal vector data. In general, the current method of vector detection can achieve good automation for attribute data and topological relationships, but research on spatial information detection methods for specific objects is scarce. For vector data extracted from water, the accuracy of spatial information is the basis of water information monitoring. In this study, we focused on the accuracy of spatial range expression of water vector data and studied the automatic detection method for vector data products of massive time series remote sensing images.

Our detection method for massive vector data of reservoir waterbodies is based on the data analysis mechanism of big data platforms. Our method chooses the area properties to reflect the spatial characteristic and converts them into a time series set. The spatial variation law of vector data is analyzed using time series modeling. Outliers are identified by an anomaly detection algorithm. This method allowed semi-automatic inspection for polygon vector data and improved inspection efficiency. The main contributions of this study are as follows: (1) a proposed spatial information anomaly detection method for reservoir waterbody vector data; (2) a tool for automatic detection of massive vector data based on the EGADS plug-in framework; and (3) an evaluation of the application of this method to anomaly detection of water vector data in different reservoirs.

2 Detection method for polygon vector data of reservoir water supported by extensible generic anomaly detection system framework

2.1 Background

Anomaly detection can be simply described as follows: given a set of N data/objects, the process of finding data/objects that are significantly abnormal, isolated, or inconsistent with the general rules compared with the general data/objects [31]. There are three kinds of anomalies in time series: point anomalies, sequence anomalies, and pattern anomalies. Time series-oriented anomaly detection usually needs to achieve two things: time series analysis and effective abnormal data detection [31,32].

A time series is a column of ordered data recorded in a chronological order, which can be represented as the set $X = (x_1, x_2, \ldots, x_n)^T$. The result of water vector extraction for long time series observation is a form of time series. Time series analysis is used to express the trend of a time series through the mathematical theory and method and predict its trend [33]. Time series modeling reveals the development law of a time series from the perspective of sequence autocorrelation, which is an important method to explore the statistical characteristics (trend, seasonality, periodicity, stationarity, and autocorrelation) of a time series. It is often used in hydrological

forecasting, weather forecasting, market economic forecasting, and other fields [31,34,35].

Outliers are objects in the data set that deviate from most of the data. They are generated not because of random errors, but because of different mechanisms [36]. In practice, outliers often contain valuable information, and mining the hidden value in them is helpful for analyzing decision tasks. According to the existing studies, anomaly detection models can be divided into the following categories: the statistical-based methods [37,38] (establishing the data distribution model and screening anomalies through deviations); the distance-based method [39,40] (distance calculation of objects in data set through distance function and exception screening using a threshold); the density-based methods [41,42] (introducing local density and judging outlier degree of data by weight); and the method based on machine learning [43–45] (algorithms mainly based on the artificial neural network and the support vector machine). Studies have shown that the distance-based method and the machine learning algorithm tend to have high time complexity and low efficiency. Some literature state [31,32] that the time series anomaly detection based on prediction is the most concise and intuitive anomaly detection method, but this method relies on the prediction model and a reasonable threshold.

The extensible generic anomaly detection system (EGADS), which was designed by Yahoo, is an open-source extensible framework for automatic anomaly detection of massive time series data [46]. It consists of two main modules: the time series modeling module (TMM) and the anomaly detection module (ADM). In the EGADS plug-in framework, time series model and anomaly detection model can be inserted flexibly according to the specific application and data characteristics. As a java package, it can be easily integrated into the existing monitoring infrastructure. In this study, we used the EGADS framework as support to explore the anomaly mining method for vector data of reservoir waterbodies.

2.2 Vector detection of reservoir waterbodies based on EGADS

Facing abnormal phenomena, such as false and missed extractions and water level alarm in automatic water extraction, this study used the EGADS framework combined with the geometric characteristics of the water vector to design a detection method for vector data of reservoir waterbodies. First, rough inspection and processing are carried out on the water vector to find a preliminary solution to the most significant and obvious quality problems such as sawtooth, void, and irregular property structures and to obtain the initial area data for anomaly detection. Then, combining the EGADS framework plug-in outlier detection method, time series modeling is carried out on the area data of vector water to analyze and find the change rules of water data. The reliability of data is tested, based on this, using different anomaly detection models. Finally, abnormal data are found, and the alarm result is obtained. The specific research ideas are shown in Figure 1.

2.2.1 Rough inspection and processing of polygon vector data of reservoir waterbodies

The extraction products of reservoir waterbodies have some obvious and easy-to-handle quality problems, such as sawtooth, void, confusion, or abnormality of attribute information, which are common problems in water extraction

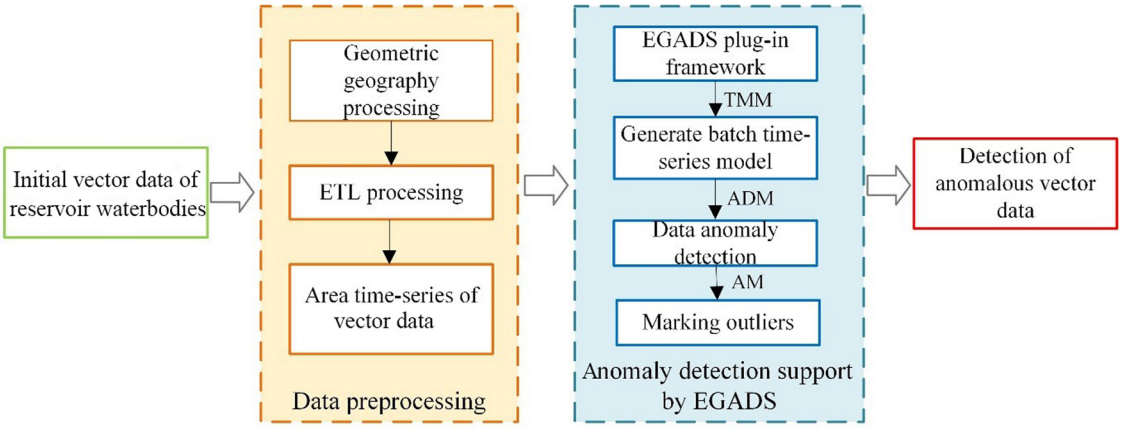

Figure 1: Detection flow chart for polygonal vector data of waterbodies in EGADS.

Table 1: Time series Models supported by TMM

Type	Model	Description
Simple window model	Olympic model (seasonal naive)	It belongs to the naive seasonal model. The predicted value is the smoothed average of the previous n periods
Exponential smoothing model	Simple exponential smoothing model Double exponential smoothing model Triple exponential smoothing model	These models are used to produce smoothed time series. The predicted value is the weighted sum of the previous value. Double and triple exponential smoothing variants add the trend and seasonality into the models
Moving Smoothing Model	Moving average model Weighted moving average model	The predicted value of point x is replaced by the average value of the adjacent points. Weighted moving average model adds weight factor
Regression model	Regression model Multiple linear regression model Polynomial regression model	They establish a relationship between one or more independent variables x and dependent variables y

[47,48]. These problems can be solved by the rough inspection and preprocessing of data, and the basic quality control of vector water body can be realized, which includes two parts: geographic geometry processing and ETL processing.

(1) Geographic geometry processing. First, given the initial reservoir waterbodies vector data, the gross errors were removed by using the filling and smoothing methods, which was caused by the hole and sawtooth. Then, as the results of the same waterbody may be extracted from different images on the same date, data fusion was carried out by setting the date and the same object judgment conditions to obtain complete vector data extraction results. The area of the fused vector data was calculated after projection.
(2) ETL processing. The table structure was designed in the Hive Data Warehouse to realize standardized data storage. After exporting the attribute data from the Geographic database, the data were extracted, transformed and loaded, and finally stored in the target database. It builds the foundation for follow-up modeling.

2.2.2 Time series modeling of polygonal vector data of reservoir waterbodies

Through rough inspection and processing, the obtained area information of water vector is a group of typical time series, and they could have trend, seasonal, and cyclical characteristics. In time series modeling, different modeling methods should be selected for time series with different regularities. Within the framework of EGADS, TMM provides nine time series models in common use, which are described in detail in Table 1.

Meanwhile, to realize the evaluation of the model, TMM provides five metrics to evaluate the time series model, and the description of each index is shown in Table 2.

Water is a type of ground object greatly affected by the pattern of water and heat distribution on the surface [49], which has obvious seasonal characteristics, so an appropriate time series data model with seasonal characteristics should be selected. This study compared various modeling methods and selected the optimal method to model and analyze the time series of the waterbody vector data.

2.2.3 Abnormal detection of polygonal vector data of reservoir waterbodies

In the time series modeling stage, given the area time series of waterbody vector, the prediction value of x_t at time t is obtained by modeling. TMM passes the original sequence and prediction results to ADM.

ADM calculates the deviation metric (DM) between the predicted value and the actual value, and analyzes the distribution characteristics of the deviation. When the error falls outside some fixed thresholds, the data are judged as an abnormal value, and an alarm is issued. Threshold reflects the sensitivity of the abnormal detection model, but it is difficult to determine the optimal thresholds artificially. The EGADS provides two kinds of algorithms, $K\sigma$ deviation and density distribution, which track a set of DMs by default and select the appropriate thresholds.

The ADM supports four anomaly detection models, which are described in detail in Table 3.

Table 2: Metrics for modeling in TMM

Model	Description	Calculation formula
Bias	The arithmetic mean of the errors	$\text{Bias} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \widehat{x_i})$
MAD	The mean absolute deviation	$\text{MAD} = \frac{1}{n}\sum_{i=1}^{n}\lvert x_i - \widehat{x_i}\rvert$
MAPE	The mean absolute percentage error	$\text{MAEP} = \frac{100\,\%}{n}\sum_{i=1}^{n}\left(\frac{\lvert x_i - \bar{x}_i\rvert}{x_i}\right)$
MSE	The mean square of the errors	$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \widehat{x_i})^2$
SAE	The sum of absolute errors	$\text{SAE} = \sum_{i=1}^{n}\lvert x_i - \widehat{x_i}\rvert$

Note: in the formula, the observed value of time i is x_i, the predicted value of time i is $\widehat{x_i}$.

After using the best time series model to obtain the fitting results, this study used the four anomaly detection models to analyze the waterbody vector data. We then analyzed the detection characteristics and the accuracy of the models and selected the best-fit anomaly detection model. Thus, the optimal anomaly detection model for the waterbody vector data was determined.

3 Data and preprocessing

3.1 Data

The Water conservancy Gaofen satellite product service and distribution subsystem is one of the operational subsystems of the Gaofen Water conservancy Remote sensing application demonstration system (phase 1), which is developed by the Ministry of Water Resources. The system supports the management, query and sharing of domestic Gaofen satellite images (GF-1, GF-2, etc.), as well as the extraction, management, and query of water vectors nationwide. Based on the water body extraction algorithm with the water body index and automatic threshold, it can automatically extract water nationwide based on domestic Gaofen series of remote sensing images and provide near real-time nationwide water conservancy production and query.

The experimental data are obtained from the national water vector data automatically extracted by this system. Since the series of Gaofen satellites was launched as early as April 2013 and the study started at the end of 2017, we chose a data time range covering the full four years: November 2013 to November 2017. Our study focused on the polygonal vector data of reservoir waterbodies. First-level large reservoirs are the largest in terms of storage capacity, water area, and project scale, and the social and economic benefits generated are relatively more important. Due to the larger coverage, they are more likely to be blocked and affected by the cloud, vegetation, mountain shadow, and other objects, resulting in inaccurate extraction results, which are more suitable for research. Therefore, in this study, the Miyun reservoir (located in Beijing), a first-level large reservoir, was selected to construct the detection method for the waterbody vector data, to test the detection effect of different models provided by EGADS in the actual sequence, and to obtain a suitable detection method for water reservoir vector data.

3.2 Data preprocessing

The data from the Miyun Reservoir from October 2, 2013, to October 5, 2017 were preprocessed, and we obtained 160 datasets, including the attribute information of reservoir name, reservoir type, time, and area. We plotted the scatter plot for the water area time series as shown in Figure 2.

The area data for the Miyun reservoir (Figure 2) range from [20,100] km^2, and there were obvious deviations from the surrounding points. So, anomaly detection of the waterbody vector data is very necessary. Waterbody vectors with areas of approximately 20, 30, 40, 50, 60, 70, 80, 90, and 100 km^2 were selected for observation, and typical data are shown in Figure 3.

There are two main reasons for the variation of the extraction areas of reservoir waterbody: (1) there is cloud

Table 3: Anomaly detection models supported by ADM

Model	Description
DBScan model	Density-based spatial clustering of applications with noise
$K\sigma$ model	The classic $K\sigma$ model, which is usually used for normal distribution data
AKDCPD model	Adaptive kernel density change-point detection
ELD model	Extreme low-density outlier detection

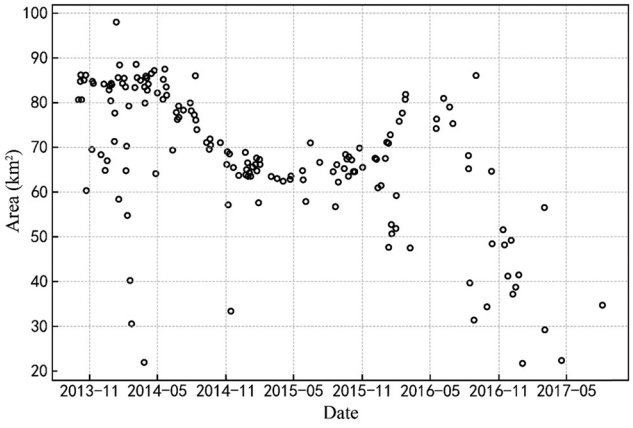

Figure 2: Scatter plot of area-time change of Miyun Reservoir from October 2013 to October 2017.

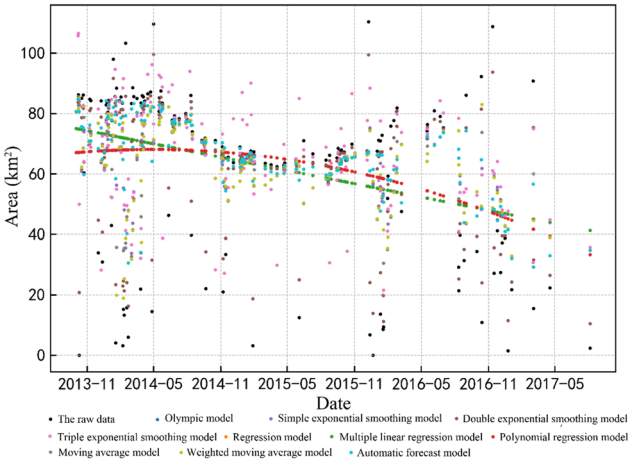

Figure 4: Scatter diagram of time series modeling results of Miyun Reservoir from October.

cover in the remote sensing image or the image does not cover the reservoir completely (Figure 3(a)–(e) and (i)). We need to identify and remove this type of abnormal data; (2) drought, flood, agricultural irrigation, and artificial storage cause the amount of reservoir water to increase or decrease, resulting in changes in the water area (Figure 3(f)–(h)). We need to detect the abnormal changes in the reservoir waterbody under these circumstances to improve water management, dispatching, and emergency response.

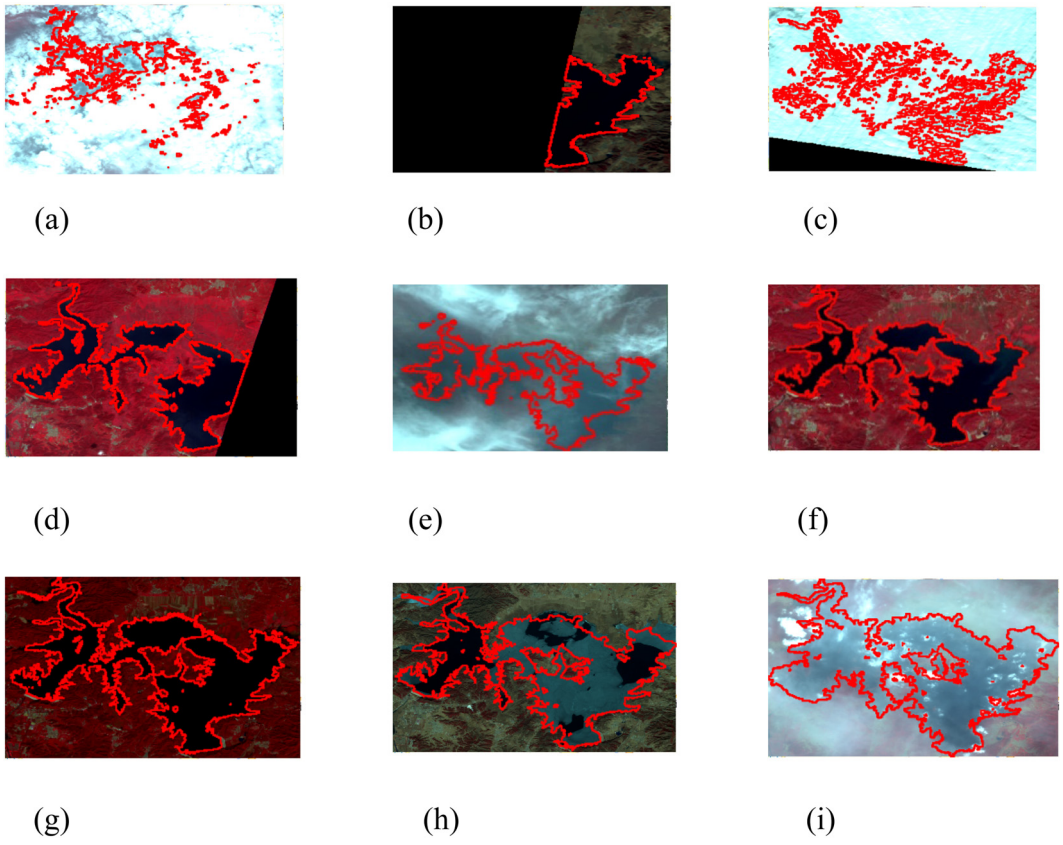

Figure 3: Water vector data for Miyun Reservoir (20–100 km^2). (a) 2014-9-21, (b) 2013-12-12, (c) 2014-1-6, (d) 2015-8-21, (e) 2013-10-23, (f) 2014-9-10, (g) 2013-10-2, (h) 2017-3-3, (i) 2014-5-1.

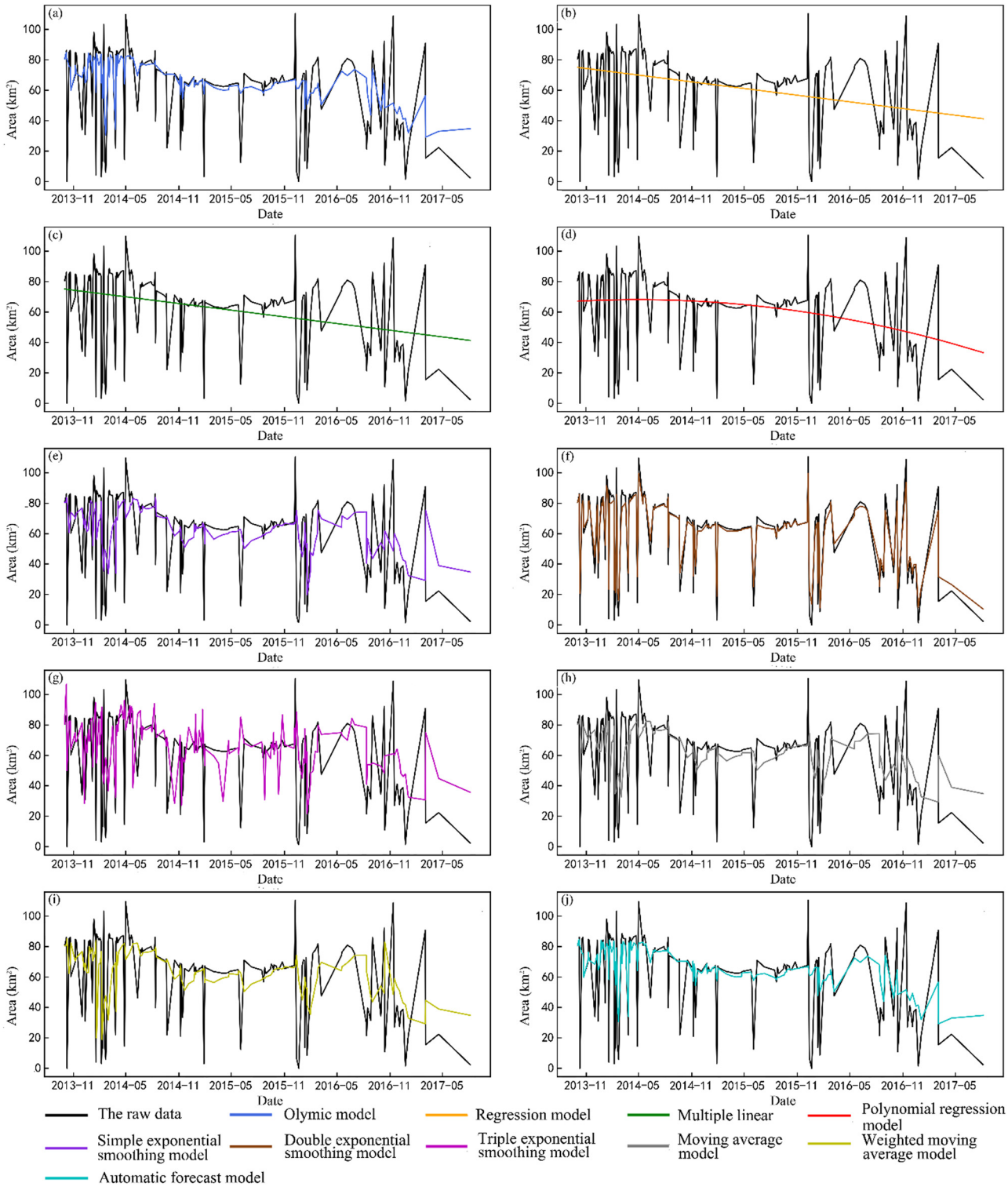

Figure 5: Sequence chart of time series modeling results of Miyun Reservoir from October 2013 to October 2017.

4 Results and analysis

4.1 Time series modeling

We used the area time series of the Miyun Reservoir for modeling using nine time series models of TMM. The results are shown in a scatter diagram (Figure 4) and sequence chart (Figure 5). The automatic forecast model is the optimal model chosen by EGADS from nine models according to the minimum deviation rule. However, this method has a single selection criterion, and it may not be applicable to the analysis of the waterbody vector data.

Table 4: Model performance on area time series of Miyun reservoir

Model	Bias	MAD	MAPE	MSE	SAE
Olympic model	−4.5676	12.3270	445.2518	433.5987	1972.3260
Simple exponential smoothing model	−1.5409	18.8335	427.7598	743.0800	3013.3656
Double exponential smoothing model	−0.3867	4.7088	105.9608	47.1614	753.4099
Triple exponential smoothing model	−4.5172	20.5639	360.7425	854.1170	3290.2209
Regression model	−1.1514	18.5161	406.8929	644.5372	2962.5733
Multiple linear regression model	−1.1514	18.5161	406.8929	644.5372	2962.5733
Polynomial regression model	−0.9466	18.4807	407.3938	634.2634	2956.9054
Moving average model	−1.2833	18.8168	432.2108	737.7722	3010.6924
Weighted moving average model	−1.0836	18.8511	437.1115	754.7399	3016.1811
Automatic forecast model	−4.567595	12.327037	445.251751	433.598741	1972.3260

Figure 4 shows the scatter diagram of the results of time series modeling for each model; EGADS judged that the reservoir automatic prediction model is the Olympic model. As only the rate of change of the reservoir area was considered in our experiment, the result of the multiple linear regression model is consistent with those of the regression model.

Figure 5 shows the characteristic curves of time series of sample and modeling results. The fitting effect of each model has the following characteristics: (1) the fitting curve of the Olympic model (Figure 5(a)) in the early stages of the data over a short time interval (before December 2015) agrees well with the actual area curve, but the fitting effect is ordinary when the time interval of later data is large. At the same time, the fitting curve of the model can reflect the seasonal variation of the performance data. (2) The fitting curves of the regression model, multivariate linear regression model, and polynomial regression model (Figure 5(b)–(d)) are smooth curves, which are very inconsistent with the actual area curve. (4) The fitting curve of the first exponential smoothing model (Figure 5(e)) is similar to that of the Olympic model as a whole. However, when the time interval is large, the sudden change in the predicted value is more obvious than that of the Olympic model. (5) The fitting curve of the double exponential smoothing model (Figure 5(f)) is very consistent with the actual curve, and it reveals the overfitting phenomenon. (6) The fitting effect of the triple exponential smoothing model is worse than that of the simple exponential smoothing model, and the sudden change in the predicted value is obvious (the difference between the predicted value and the observed value is significant around April 2015). (7) The fitting curve of the moving average model in Figure 5(h) is similar to that of the first exponential smoothing model; however, the sudden change slows down in February 2015, February 2015, and April 2017. The curve also reflects the seasonal variation in the data. (8) Compared with the moving average model, the sudden change in the predicted data slows down further in the fitted curve of the weighted moving average model shown in Figure 5(i) and reflects the seasonal variation. (9) The automatic forecast model selected by the system is the Olympic model. In summary, the time series model suitable for the Miyun Reservoir should be selected from the Olympic model and the weighted moving average model.

After time series modeling, the prediction results of each model were evaluated. These five metrics all reflect the fitting ability of the model. Smaller absolute values of the results indicate a higher fitting ability of the model, but they may correspond to the overfitting phenomenon. By using the five metrics provided by TMM, we calculated the metric results of each model as presented in Table 4.

The fitness of each model, from high to low, is as follows: double exponential smoothing model, Olympic model and automatic forecast model, Regression model and polynomial regression model, polynomial regression model, moving average model, simple exponential smoothing model, triple exponential smoothing model, and weighted moving average model.

In our data source, there are some abnormal vector data caused by clouds, incomplete image coverage of the study area, or other reasons. They deviate greatly from the general law, thus disturbing the trend of normal time series. Therefore, we need to further evaluate the time series model. The dispersion degree of the model results was analyzed by calculating for the indicators presented in Table 5. Concentrated model results show that the reservoir water storage was closer to a stable actual state, indicating the fitness of the model.

The dispersion degree of the corrected results of each model, from smallest to largest, is as follows: polynomial regression model, regression model and multiple linear regression model, Olympic model and automatic forecast

Table 5: Results of evaluation indexes for data correction

Model	Mode	Median	Average	Range	Sum of squares of mean deviation	Variance	Standard deviation	Coefficient of variation
Original data	67.5507	67.5507	61.7146	110.4154	113274.1519	712.4160	26.6911	0.4325
Olympic model	80.6221	66.2316	66.2822	55.5574	24184.6291	152.1046	12.3331	0.1861
Simple exponential smoothing model	80.6221	64.3946	63.2555	64.1325	26209.8704	164.8420	12.8391	0.2030
Double exponential smoothing model	—	65.9479	62.1013	89.1342	69446.1365	436.7682	20.8990	0.3365
Triple exponential smoothing model	—	67.1326	66.2318	85.0029	45514.1567	286.2526	16.9190	0.2555
Regression model	—	64.2221	62.8660	33.7570	12431.4155	78.1850	8.8422	0.1407
Multiple linear regression model	—	64.2221	62.8660	33.7570	12431.4155	78.1850	8.8422	0.1407
Polynomial regression model	—	66.6687	62.6612	34.8026	8888.0294	55.8996	7.4766	0.1193
Moving average model	80.6221	64.1643	62.9979	54.8733	26426.2731	166.2030	12.8920	0.2046
Weighted moving average model	80.6221	64.3562	62.7982	66.2429	29871.7775	187.8728	13.7067	0.2183
Automatic forecast model	80.6221	66.2316	66.2822	55.5574	24184.6291	152.1046	12.3331	0.1861

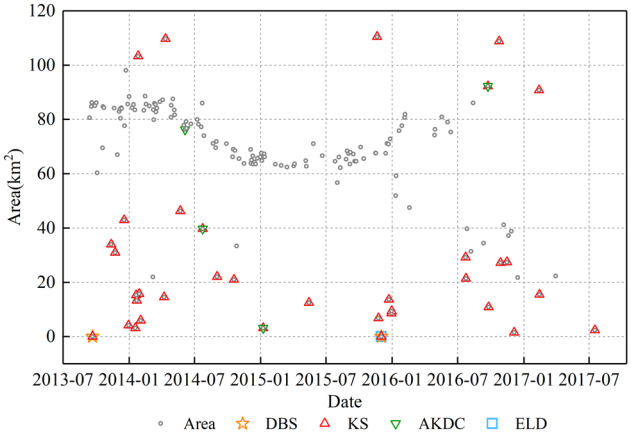

Figure 6: Abnormal area data for Miyun Reservoir from October 2013 to October 2017.

model, simple exponential smoothing model, moving average model, weighted moving average model, triple exponential smoothing model, double exponential smoothing model, and original data.

Considering the fitting curve characteristics, fitting degree, and the dispersion degree of the corrected results, the Olympic model is the most suitable time series model for the Miyun Reservoir. Meanwhile, the system's automatic prediction model was also the Olympic model, but the system needs to run all the models before selecting the best model. Considering the computer performance and resource consumption, it is appropriate to directly adopt the Olympic model.

4.2 Outlier detection

After using the Olympic model for time series modeling, we used the four kinds of anomaly detection models to detect outliers for the Miyun Reservoir area vector data. We marked the outliers with a different color in the scatter plots (Figure 6). The occurrence date of the abnormal area data is presented in Table 6.

The characteristics and detection results of each model were analyzed. The purpose of the density-based spatial clustering of applications with noise (DBScan) model is to find outliers that are free from clusters; it only found near-zero abnormal values in our experiments. The adaptive kernel density change-point detection (AKDCPD) model considers the change points testing problem; it mistakenly regarded the area data on June 24, 2014, as an outlier and failed to detect most of the other abnormal values. The extreme low-density (ELD) model regards data in low-density regions as outliers, and it had

Table 6: Occurrence date of abnormal area data detected

Model	DBScan model	*Kσ* model			AKDCPD model	ELD model
Occurrence date of abnormal area data	2013/10/10	2013/10/10	2014/5/1	2016/1/19	2014/6/24	2015/12/21
	2015/12/21	2013/12/1	2014/6/11	2016/8/11	2014/8/12	
		2013/12/12	2014/8/12	2016/8/12	2015/1/27	
		2014/1/6	2014/9/21	2016/10/12	2016/10/12	
		2014/1/18	2014/11/7	2016/10/14		
		2014/2/6	2015/1/27	2016/11/12		
		2014/2/8	2015/6/3	2016/11/16		
		2014/2/10	2015/12/9	2016/12/4		
		2014/2/14	2015/12/13	2016/12/24		
		2014/2/17	2015/12/21	2017/3/3		
		2014/2/21	2016/1/11	2017/3/4		
		2014/4/27	2016/1/18	2017/8/5		

the weakest anomaly detection effect in our experiment; only the smallest area was detected. The *Kσ* model effectively identified anomalies in high or low area values (all anomalies detected by other models were included). However, a few abnormal values had not been detected, and the reasons are explained in Section 4.4.2. In comparison, the optimal anomaly detection model for the Miyun Reservoir is *Kσ* model.

The aforementioned experiments show that in the framework of EGADS, combining the Olympic model and *Kσ* model into a single framework can effectively detect abnormal values in the area vector data for the Miyun Reservoir.

4.3 Application to other types of reservoirs

To further explore whether the optimal time series model and anomaly detection model are affected by reservoir scale in the framework of EGADS, we selected the Xishanwan Reservoir (second-level large; 99 datasets from November 8, 2013, to April 30, 2017), Dashuiyu Reservoir (third-level large; 157 datasets from October 2, 2013, to April 26, 2017), Xiagou Reservoir (fourth-level large; 83 datasets from May 2, 2014, to June 14, 2017), and Gulbeitou Reservoir (fifth-level large; 63 datasets from May 1, 2014, to May 13, 2017) for analysis and obtained the optimum models for waterbody vector data for each reservoir, which are presented as shown in Table 7.

Experiments show that although the fitting effect of each time series model is slightly different for the area data of reservoirs at different scales. Among the nine time series models, the Olympic model is the best in modeling performance for different scale reservoirs, in terms of seasonal variation, a high degree of fitting, and a small degree of dispersion. The *Kσ* model can effectively detect most of the abnormal values for all scales of reservoirs. Therefore, for reservoirs of different scales, the combination of the Olympic model and *Kσ* model in EGADS is suitable for detection of the abnormal values in the reservoir waterbody vector data.

4.4 Accuracy evaluation

4.4.1 Accuracy evaluation method

To verify the validity and accuracy of the method, we selected sensitivity and specificity as the performance measures. In our study, the experimental results were divided into four categories: in the first category, actual outliers are judged as anomalies, with a total of TY; in the second category, actual outliers were judged as qualified, with a total of FN; in the third category, the actual qualified value was judged as qualified, a total of TN; in the fourth category, the actual qualified value was judged as abnormal, a total of FY. The calculation formulae are as follows:

(1) Sensitivity: Refers to the percentage of correctly classified outliers among all predicted outliers by the classifier.

$$\text{Sensitivity} = \frac{\text{TY}}{\text{TY} + \text{FY}} \tag{1}$$

(2) Specificity: Refers to percentage of correctly classified qualified value among all predicted qualified value by the classifier.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FN}} \tag{2}$$

Table 7: Optimal Models selected for reservoirs at different scales

Reservoir scale	Name	Location	Time series model	Anomaly detection model
Second-level large	Xishanwan reservoir	Duolun County, Xilingol League, Inner Mongolia Autonomous Region	Olympic model	$K\sigma$ model
Third-level large	Dashuiyu reservoir	Huairou District, Beijing	Olympic model	$K\sigma$ model
Fourth-level large	Xiagou reservoir	Yiwu County, Hami District, Xinjiang Uygur Autonomous Region	Olympic model	$K\sigma$ model
Fifth-level large	Gulbeitou reservoir	Mengzhou City, Jiaozuo City, Henan Province	Olympic model	$K\sigma$ model

4.4.2 Model evaluation and analysis

We extracted the Miyun Reservoir vector data for the sample and interpreted it manually: 106 are qualified vector data and 54 are abnormal vector data. Comparing the results of detection with those of manual interpretation, we got TY = 34, FN = 20, TN = 104, FY = 2 in the outliers of the Miyun Reservoir detected by the Olympic model and $K\sigma$ model. By substituting them into formula (1) and formula (2), sensitivity = 94.44% and specificity = 83.87% were derived. The results show that for the Miyun Reservoir area vector data, the method exhibited a 94.44% accuracy rate for abnormal data detection, and the vector data accuracy is improved from 66.25%, before the inspection, to 83.87%. The missed abnormal vector data (FN) and false abnormal vector data (FY) of the reservoir are shown in Figure 7 and 8, respectively.

From Figures 2 and 7, the anomalous vector data of missing alarm can be divided into three categories: (1) when reservoir imaging was less affected by cloud coverage or the limitation of the width in the remote sensing image, the extracted vector changed slightly (Figure 7(a)–(c) and (e)–(h)). This category of abnormal vector data can be detected by judging the inclusion relation of the frame vector of the original image and the detected cloud with the general water vector or the vector centroid. (2) The extracted vector data were divided into many fragmentary vectors, but their total area did not change significantly (Figure 7(q) and (s)). This category of abnormal vector data can be detected by judging the size of each vector area, number, and offset of vector centroids. (3) The approximate time and concentrated values of abnormal data result in the larger or smaller fitting value in time series modeling, which lead to some abnormal data not being detected (Figure 7(i)–(p), (r) and (t)). This category of abnormal vector data can be detected by judging the offset of vector centroids. (4) Model defects caused part of abnormal vector data to be undetected (Figure 7(d) and (g)), which accounts for a small proportion of the total anomaly data. This category of abnormal vector data can be detected by judging the size of each vector area and offset of vector centroids.

From Figure 8, we found that the false alarm anomaly data were vector data with larger areas: (1) the waterbody vectors of the Miyun Reservoir (Figure 8(a)) could be extracted from two images from the same date, but one was inaccurate because of the excessive cloudiness in the image, which affected the area of the waterbody vectors after fusion. (2) The water vector (Figure 8(b)) was disturbed by thin clouds in the image, which resulted in a larger extraction result, but we could not rule out the

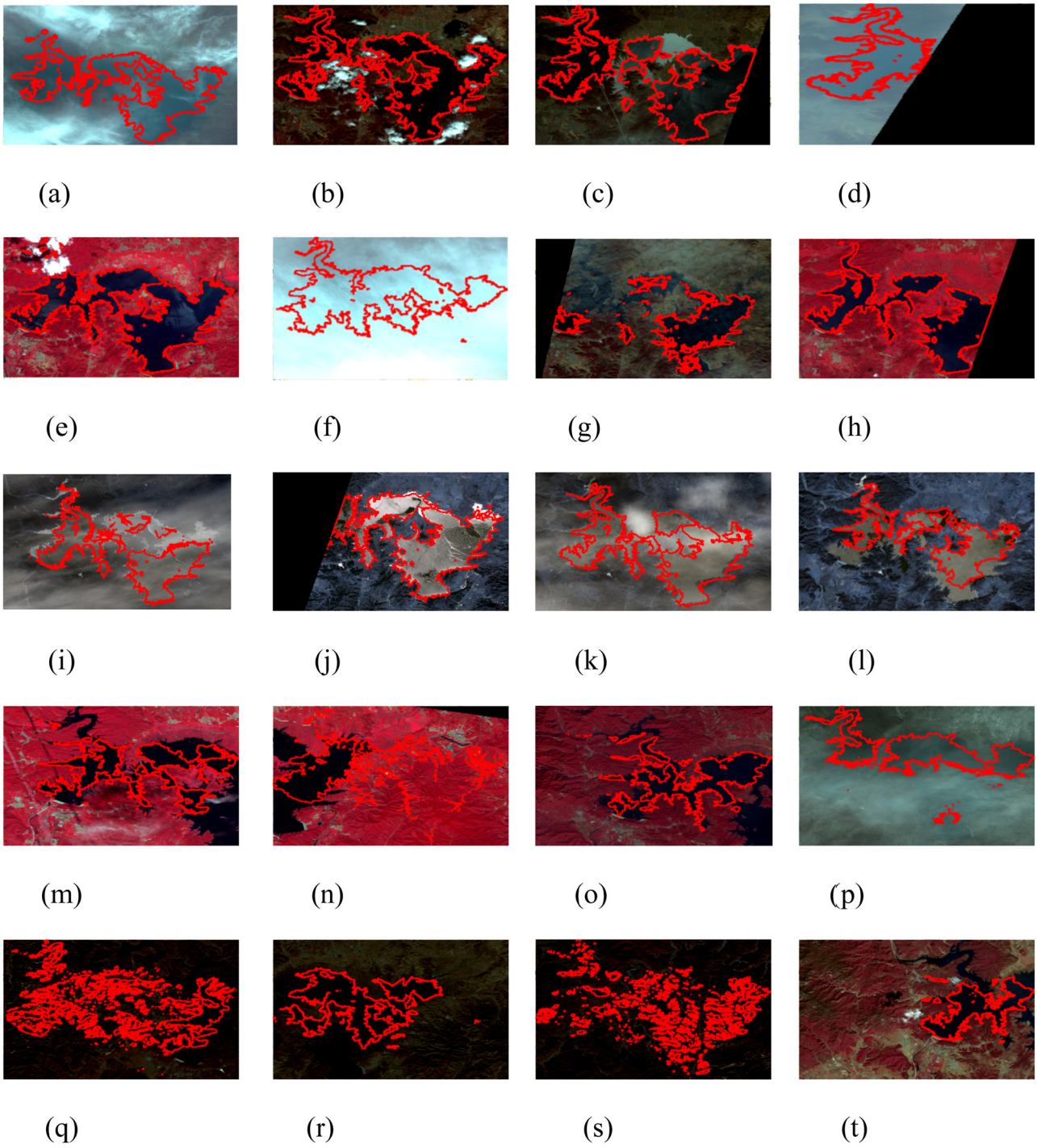

Figure 7: Missed alarm vector data. (a) 2013-10-23, (b) 2013-11-7, (c) 2014-1-7, (d) 2014-3-27, (e) 2014-6-28, (f) 2014-9-17, (g) 2014-11-14, (h) 2015-8-21, (i) 2016-1-30, (j) 2016-1-31, (k) 2016-2-25, (l) 2016-3-8, (m) 2016-8-15, (n) 2016-8-26, (o) 2016-9-30, (p) 2016-11-25, (q) 2016-12-8, (r) 2016-12-16, (s) 2017-1-3, (t) 2017-4-18.

increase in the water surface. To obtain more accurate anomaly detection results, we should also classify these two types of anomaly data as anomalies to some extent.

Accuracy evaluation results show that EGADS can detect 94.44% accuracy rate of reservoir vector data anomalies and improve the vector data accuracy from 66.25 to 83.87%. The errors in terms of undetected abnormal data are small, which can be determined by vector space position or manual detection. Therefore, the detection method had passed the qualitative evaluation and can

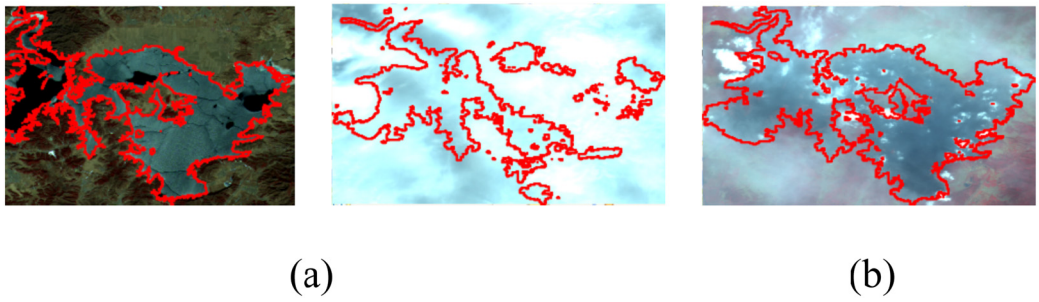

Figure 8: False alarm vector data. (a) 2014-2-14, (b) 2014-5-1.

accurately detect most of the reservoir vector data anomalies and combining it with vector space location inspection can further improve the detection accuracy.

5 Conclusions

With the continuous development of the automatic extraction ability of waterbody data, the volume of waterbody vector data products has increased rapidly, and the traditional manual inspection methods cannot meet the current demands. In this study, we utilized the fast calculation and scalability of the EGADS framework, and an anomaly detection method for reservoir waterbodies vector data was proposed. First, the area time series was obtained by preprocessing the vector data. Then, under the EGADS framework, the Olympic model was used to model the time series, and the variation rule for the water area was constructed. On this basis, the $K\sigma$ model was used to automatically detect the abnormal values based on the fitting values and the original data. Finally, the date for the abnormal values was obtained. Accuracy verification results show that the sensitivity and the specificity of the method were 94.44 and 83.87%, respectively. The accuracy of the vector data through EGADS detection increased from 66.25 to 83.87%.

The inspection method used in this study, the EGADS framework, allowed for the semi-automatic inspection of vector data of the reservoir water, which has practical applications in the development of water resources information monitoring technology and can even be used for climate, environmental changes, etc.

The missing outliers in this study can be checked in combination with the vector space position, so as to achieve higher accuracy. This problem can be used as the focus for future research on the application of the automatic anomaly detection method for reservoir waterbody vector data.

Acknowledgement: The financial support for this research was provided by the National Key Research and Development Program of China under Grant 2017YFC0405806. The authors thank all researchers and staff for providing the polygonal vector product of reservoir waterbodies from China Ministry of Water Resources, Hydrology Monitor and Forecast Center.

References

[1] Li JL, Cao LD, Pu RL. Progresses on monitoring and assessment of flood disaster in remote sensing. J Hydraulic Eng. 2014;45:253–60.
[2] Cai Y, Meng LK, Cheng JG. Satellite remote sensing water monitoring model and its application. 1st ed. China: Science Press; 2018.
[3] Yang T, Zhang Q, Chen YD, Tao X, Xu CY, Chen X. A spatial assessment of hydrologic alteration caused by dam construction in the middle and lower Yellow River, China. Hydrological Process. 2010;22:3829–43.
[4] Ministry of Water Resources of the PRC. Bulletin of the First National Water Conservancy Census. China Water Conservancy. 2013;7:64.
[5] Wu BF, Lu SL. Watershed remote sensing: methodology and a paradigm in Hai Basin. J Remote Sens. 2011;15:201–23.
[6] Paramate H, Supattra P. Entropy-based fusion of water indices and DSM derivatives for automatic water surfaces extraction and flood monitoring. ISPRS Int J Geo-Information. 2017;6:301–22.
[7] Work EA, Gilmer DS. Utilization of Satellite Data for Inventorying Prairie Ponds and Lakes. Photogrammetric Eng & Remote Sens. 1976;42(5):685–94.
[8] Du YY, Zhou CH. Automatically extracting remote sensing information for water bodies. J Remote Sens. 1998;2(4):264–9.
[9] Bi YY, Zhou CH. Automatically extracting remote sensing information for water bodies. J Remote Sens. 1998;2(4):264–9.
[10] Mcfeeters SK. The use of the normalized difference water index (NDWI) in the delineation of open water features. Int J Remote Sens. 1996;17(7):1425–32.
[11] Xu HQ. Modification of normalized difference water index (NDWI) to enhance open water features in remotely sensed imagery. Int J Remote Sens. 2006;27(14):3025–33.

[12] Ogilvie A, Belaud G, Massuel S, Mark M, Patrick LG, Roger C. Surface water monitoring in small water bodies:Potential and limits of multi-sensor Landsat time series. Hydrol Earth Syst Sci. 2018;22(8):1–35.
[13] Diao SJ, Liu CL, Zhang T, He P, Guo ZC, Tu JN. Extraction of remote sensing information for lake salinity level based on SVM:A case from Badain Jaran desert in Inner Mongolia. Remote Sens LResour. 2016;28(4):114–8.
[14] Feng QL, Liu JT, Gong JH. Urban flood mapping based on unmanned aerial vehicle remote sensing and random forest classifier: A case of Yuyao, China. Water. 2015;7:1437–55.
[15] Jain SK, Singh RD, Jain MK, Lohani AK. Delineation management of flood-prone areas using remote sensing techniques. Water Resour. 2005;19:333–47.
[16] Meng LK, Guo SX, Li S. Summary on extraction of waterbody from remote sensing image and flood monitoring. Water Resour Informatization. 2012;12:18–25.
[17] Tri A, Anoj S, Dong L. Evaluation of water indices for surface water extraction in a Landsat 8 scene of Nepal. Sensors. 2018;18:2580.
[18] Singh PP, Garg RD. Automatic road extraction from high resolution satellite image using adaptive global thresholding and morphological operations. J Indian Soc Remote Sens. 2013;41:631–40.
[19] Yang XL. Change detection of multi-temporal remote sensing image. PhD thesis. Xian: Xidian University; 2011
[20] Shi CC, Chen MS, Yang JY, Lv CX. Design on Nanjing water conservancy information resource sharing system integration. Computer Eng Softw. 2015;36(09):55–9.
[21] Gaofen application integrated information service sharing platform: Application case http://gaofenplatform.com/contents/83/521.html
[22] Fan DZ, Lei R, Zhang BM. The checking method of terrain feature vector data. J Inst Surveying Mapp. 2002;19:182–5.
[23] Zhang JS, Li GQ, Guo HT. A matchingmethod of remote sensingimage and GIS vectordata based on dynamic programmingand Hough transform. Eng Surveying Mapp. 2011;05:13–6.
[24] Wei XY, Liu C. Quality control and inspection method of landhouse GIS basic data. Mod Surveying Mapp. 2004;27:18–21.
[25] Popescu C, Balbo PP, Adrian S, Ciolac V. The analysis of the vector system of the cadastral maps for the creation of a gis project. Stat Med. 2010;26:2229–45.
[26] Chen F, Gong JH, Chen ZL, Meng Y. Design and implementation of quality inspection system for geographical condition census based on check rules. Bull Surveying Mapp. 2016;3:122–5.
[27] Zhang L. Research of quality inspection system for geographic condition monitoring vector data using arc engine. PhD thesis; Xian: Chang'an University; 2018
[28] Guo PP, Li CM, Yin Y, Wu PD. Automatic correction algorithm of water element attribute oriented to national geographic census. Bull Surveying Mapp. 2017;6:61–4.
[29] Gui D, Li G, Li C, Zhang C, Quality Check in Urban and Rural Cadastral Spatial Data Updating. The 8th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences Shanghai; 2008.
[30] Zhang QS, Wang YH, Liu XP. Topological conflict detection and consistency maintenance method in process of area entities incremental integration. Geomat Inf Sci Wuhan Univ. 2019;44:154–61.
[31] Yu YF, Zhu YL, Wan DS, Guan XZ. Time series outlier detection based on sliding window prediction. J Computer Application. 2014;34(08):2217–20.
[32] Chen B. Research on time series anomaly detection based on collaborative learning. PhD thesis. Xuzhou: China University of Mining and Technology; 2018
[33] Wang Y, Application of time series analysis. 4th Renmin University Press of China, China; 2015.
[34] Wang ML, Yu RL, Li J, Xia LZ, Li YC. Research on local environmental temperature change and forecast in the three gorges reservoir area based on time series analysis. Chin J Agrometeorology. 2018;39(01):9–17.
[35] Wang LN, Xu DR. Analysis of non-steady time series forecast for economy based on ARM a model. J Wuhan Univ Technol. 2004;28(1):133–6.
[36] Hawkins DM. Identification of Outliers. Berlin: Springer; 1980. p. 27–41.
[37] Zhang BW, He HC. Progress of temporal data mining research. Computer Sci. 2002;29(2):124–6.
[38] Chandola V, Banerjee A, Kumar V. Anomaly detection: A survey. Acm Comput Surv. 2009;41(3):1–58.
[39] Ramaswamy S, Rastogi R, Shim K. Efficient algorithms for mining outliers from large data sets. ACM SIGMOD Record. 2000;29(2):427–38.
[40] Chen Q, Hu GY, Lu W. Outlier detection for time series based on distance and DF-RLS. Computer Eng. 2012;38(12):32–5.
[41] Papadimitriou S, Kitagawa H, Gibbons PB, Faloutsos C. LOCI: fast outlier detection using the local correlation integral. Proceedings of the 19th International Conference on Data Engineering. Bangalore: 2002. p. 315–26.
[42] Meng FR, Yao YX, Chang YH, Yan QY. Uncertain continuous time series top-K anomaly detection method. Application Res Computers. 2014;31(3):765–8.
[43] Tian J, Gu H. Outlier one class support vector machines. J Electron Inf Technol. 2010;32(6):1284–8.
[44] Agyemang M, Barker K, Alhajj RS. Mining web content outliers using structure oriented weighting techniques and N-grams. Acm Symposium on Applied Computing. ACM; 2005. p. 482.
[45] Wang JW. Anomaly detection from time series data for decision support. PhD thesis. Anhui: University of Science and Technology of China; 2014
[46] Laptev N, Amizadeh S, Flint I. Generic and Scalable Framework for Automated Time series Anomaly Detection. KDD '15 Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney; 2015. p. 1939–47.
[47] Zhu L. Research on image processing method in geographical national conditions using the remote sensing. Master thesis. China: Xi'an University of Science and Technology; 2015.
[48] Zhao X, Wang P, Chen C, Jiang T, Yu Z, Guo B. Waterbody information extraction from remote-sensing images after disasters based on spectral information and characteristic knowledge. Int J Remote Sens. 2017;38:1404–22.
[49] Piao S, Ciais P, Huang Y, Shen Z, Peng S, Li J, et al. The impacts of climate change on water resources and agriculture in China. Nature. 2010;467:43–51.