

# The application of selected statistical tests in the detection and removal of outliers in water engineering data based on the example of piezometric measurements at the Dobczyce dam over the period 2012-2016

Stanisław Lach<sup>1,\*</sup>

<sup>1</sup>AGH University of Science and Technology, Faculty of Mining Surveying and Environmental Engineering, Department of Environmental Management and Protection, 30 Mickiewicza Av., 30-059 Krakow, Poland

**Abstract.** Due to their size, water construction belongs to the largest and heaviest engineering structures. Ensuring the safe operation of such facilities requires continuous monitoring. Among the basic forms of monitoring in such facilities one should list continuous seasonal piezometric measurements, which are obligatory elements of general control measurements aimed at ensuring safety when using the facility. The latter is directly related to guaranteeing the safety of people living and working in an area exposed to destruction in the event of a possible disaster involving the building. From the perspective of increasing the safety of the hydrotechnical facility, optimal conditions occur when the water levels in piezometers oscillate around a constant value as this signalizes that filtration processes in the body and the surface of a dam are stable. Various factors may disturb measurements of water levels in open piezometers or water pressure in closed piezometers. These factors may take the form of systematic, random or obvious errors. Thus, before analyzing this type of data, the largest errors (outliers) should be removed from the sample as they could significantly affect the outcomes of analysis and lead to a false interpretation. In such a situation, it is necessary to apply respective statistical tests, which allow verification of whether a particular portion of data may be treated as a set of outliers at a given significance level  $\alpha$ . In this work the following statistical tests were used to identify and remove outliers: Q-Dixon test, Grubbs test, Hampel test, Rosner test, Iglewicz and Hoaglin test, Tietjen-Moore test and quartile test. The scope of the empirical analysis is focused on piezometric measurements at the Dobczyce dam over the period 2012-2016.

## 1 Introduction

In the World Register of Dams, maintained by the International Commission on Large Dams (ICOLD), Poland, with its 69 large dams ranks 16<sup>th</sup> out of 35 European countries. In Poland,

---

\* Corresponding author: [slach@agh.edu.pl](mailto:slach@agh.edu.pl)

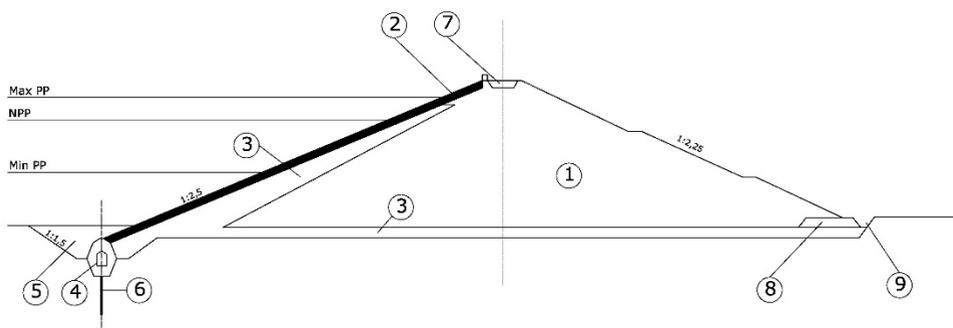
there are over 3 500 facilities permanently impounding water, including 327 dams, 2 284 weirs, 130 sluice locks and 383 hydroelectric plants. These facilities operate in changing hydrological and meteorological conditions, and they are constantly exposed to such phenomena as: lightning, intense atmospheric precipitation, landslides, the icing of water and floods. With time, impounding structures are subjected to damage posing a threat to human life, health, property and the environment. This applies to approximately 30% of Polish water impounding structures which have been operating for over 50 years. According to the ICOLD's opinion, such a long service life results in a more frequent occurrence of damage and an increased probability of failures [1].

Ensuring the safety of operation of impounding structures requires performing systematic control measurements [2]. The basic forms of monitoring dams include, e.g. piezometric measurements, which allow for measurements of water levels in open piezometers, or measurements of water pressure in closed piezometers [3]. These measurements enable controlling seepage through an impounding structure, and thus assessing a structure's performance [4, 5].

Before analyzing the piezometric data, it is necessary to remove the outliers that may significantly affect the outcome of the analysis and result in a false assessment or interpretation of the analyzed phenomenon [6]. Then, it is necessary to apply some statistical tests, allowing accepting or rejecting a doubtful result at the predetermined significance level  $\alpha$  [5].

## 2 Material and methods

The Dobczyce dam is situated on 60.1 km of the Raba River, in the commune of Dobczyce, in Małopolska province. It was commissioned in 1986. On the right bank below the dam, there is a power plant put into operation in 1993. The main functions of the reservoir include: water supply for the city of Krakow, reduction of a flash flood wave and retention for energy purposes. The earthen dam was built of local materials (sand, gravel, pebble), with an asphalt concrete screen on the upstream slope, resting at the foot of the reinforced concrete control gallery (Fig. 1). The total length of the dam at its crown is 617.0 m, its maximum height is 30.6 m, and the width of the crown reaches 8.5 m. The slope of the upstream face is 1:2.5, and the slope of the downstream face is 1: 2.25.



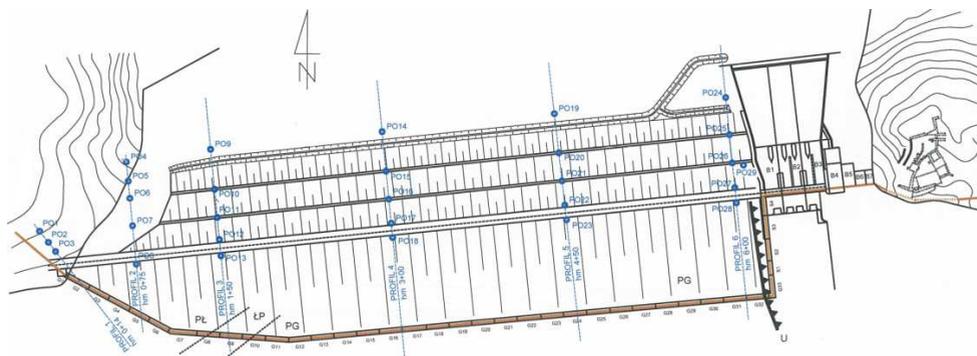
1. road; 2. asphalt concrete cutoff wall; 3. seepage layer; 4. clay sealing; 5. reinforced concrete control gallery; 6. core and foundation of the dam; 7. drainage; 8. drainage ditch

**Fig. 1.** Cross-section of the dam Dobczyce [7].

The dam is equipped with 30 open standpipe piezometers. Due to incomplete measuring sequences, seven piezometers (denoted as PO1, PO2, PO2a, PO6, PO8, PO18 and PO23)

were omitted in the analysis. Fig. 2 illustrates the distribution of the piezometers on the measuring cross-sections of the Dobczyce dam.

Measurements of changes in the water table level in the open standpipe piezometers, covering the period of 5 years (from 2012 to 2016), were analyzed for the Dobczyce dam. 121 piezometric measurements were carried out for each of the 23 open piezometers, which amounted to 2 783 observations. The piezometric data was made available by the Regional Water Management Authority in Krakow.



**Fig. 2.** Distribution of open piezometers on measuring cross-sections of the Dobczyce dam [8].

This study uses seven statistical tests to detect and remove outliers: Dixon’s Q-test, Grubbs’ test, Hampel’s test, Rosner’s test, Iglewicz and Hoaglin’s test, Tietjen-Moore’s test and the quartile test.

As far as Dixon’s Q-test is concerned, two variants were focused on: the tests denoted with the symbols N9 (verification of the hypothesis on a single outlier) and N13 (verification of the hypothesis on a pair of the largest or the smallest outliers). Dixon’s Q-test is used to detect whether there is a result encumbered with a gross error in a given data set. Using Q-Dixon’s test, only a single outlier or a pair of outliers can be removed from the analyzed data set each time [9]. Before Dixon’s Q-test is performed, a set of experimental results (statistical sample) should be arranged into a monotonically increasing sequence. For the N9 test, the value of the test statistic is expressed by the following Formula:

$$Q = \frac{x_n - x_{n-1}}{x_n - x_2} \tag{1}$$

For the N13 test, the value of the test statistic is calculated from the dependence:

$$Q = \frac{x_n - x_{n-2}}{x_n - x_3} \tag{2}$$

where:

$x_2, x_3, x_{n-2}, x_{n-1}, x_n$  denote the successive positions of the elements within the monotonically increasing sequence.

In order to reject the hypothesis of no outliers (variant N9) or a pair of outliers (variant N13), the value of the  $Q$  statistic is compared to the value read from the table of critical values  $Q_n$  of Dixon’s Q-test in the variant N9 or the variant N13 at the  $\alpha$  level of significance.

Before Grubbs’ test is performed, as it was in the case of Dixon’s Q-test, a set of experimental results should be arranged into a monotonically increasing sequence. The largest ( $x_{max}$ ) or the smallest ( $x_{min}$ ) value of the result in the analyzed sample may be encumbered with a gross error. Again, as it was in the case of Dixon’s Q-test, this test provides the possibility of detecting only one outlier, therefore, it should be repeated until no

more outliers are observed in the data set [10]. For the Grubbs test, the value of the test statistic is expressed by the Formula:

$$G_p = \frac{\max_{i=1, \dots, n} |x_i - \bar{x}|}{s} \quad (3)$$

where:

$\bar{x}$ ,  $s$  – denote the mean value and the standard deviation from the analyzed series of measurements, respectively.

The critical value of a two-sided Grubbs test for a given significance level  $\alpha$  can be read from the tables or calculated from the Formula:

$$G_{kr} = \frac{(n-1)}{\sqrt{n}} \cdot \sqrt{\frac{t_{(\alpha/(2n), n-2)}^2}{n-2+t_{(\alpha/(2n), n-2)}^2}} \quad (4)$$

As it appears from the Formula above, the critical value of the two-sided Grubbs' test for the significance level  $\alpha$  is calculated based on the critical  $t$ -Student distribution for the significance level  $\alpha/(2n)$  and the number of the degrees of freedom equal to  $n - 2$ , where  $n$  is the number of measurements in the series.

A large and unquestionable advantage of Hampel's test is the ease of its implementation, as there is no limit on the size of the data set being analyzed. The inference on the nature of the observation is subject to the assessment of the obtained results of the analysis based on specific formulas. Therefore, there is no need to read the critical value of the test statistic from special tables [11]. While performing this test, it is necessary to calculate the values of the median  $M_e$ , deviations  $r_i$  from the median, absolute values  $|r_i|$  and the deviation median  $Me_{|r_i|}$  in the analyzed data set, and then non-typical observations should be identified which satisfy the condition  $|r_i| \geq 4,5Me_{|r_i|}$ .

The Rosner test is used for samples with the size of  $n \geq 25$  observations with up to 10 outliers. It should be verified whether the analyzed set has a normal distribution (using e.g. Doornik-Hansen's, Shapiro-Wilk's, Lilliefors's or Jarque-Ber's tests). During the verification and distribution deviating from the normal distribution, it should be appropriately transformed, or the outlier-determining process should be carried out using a different test. The use of the test requires the detection of the maximum number of outliers  $r$ . Before the Rosner test is performed, a set of experimental results should be arranged into a monotonically increasing sequence. Then we calculate a series of test statistics by removing the datum (large or small) that is farthest from the mean and recomputing the test statistic according to the following equation [12]:

$$R_{i+1} = \frac{|x^{(i)} - \bar{x}^{(i)}|}{s^{(i)}} \quad (5)$$

where:

$$\bar{x}^{(i)} = \frac{1}{(n-i)} \sum_{j=1}^{n-i} x_j \quad (6)$$

$$s^{(i)} = \sqrt{\frac{1}{(n-i)} \sum_{j=1}^{n-i} (x_j - \bar{x}^{(i)})^2} \quad (7)$$

$y^{(i)}$  – the most distant value from the measurement from  $\bar{x}^{(i)}$ .

For the Iglewicz and Hoaglin test, the following value should be calculated [13]:

$$M_i = \frac{0.6745(x_i - \tilde{x})}{MAD} \tag{8}$$

where:

$\tilde{x}$  – median of the data set,

$MAD$  – median absolute deviation of the data set, calculated as:

$$MAD = median(|x_i - \tilde{x}|) \tag{9}$$

where:

$\tilde{x}$  – is the median of the data,

$|x|$  – is the absolute value of  $x$ .

Iglewicz and Hoaglin recommend that  $M_i$  with an absolute value of greater than 3.5 can be labeled as potential outliers.

The Tietjen-Moore test is a modification of the Grubbs' test, which allows to detect more than one outlier at a time (unlike the Grubbs' test, which allows to detect only one outlier at a time). The Tietjen-Moore test is a two-sided test that allows detecting  $m$  outliers in a data set, for which the normality of the distribution should be checked. For this test, the value of the test statistic is expressed by the following Formula [14]:

$$E_m = \frac{\sum_{i=1}^{n-m} (r_{(i)} - \bar{r}_{n-m})^2}{\sum_{i=1}^n (r_{(i)} - \bar{r})^2} \tag{10}$$

where:

$r_i = |x_i - \bar{x}|$  is the absolute deviation of  $x_i$  from the sample mean,

$\{r_{(i)}\}$  – are the values of  $r_i$  in ascending order ( $r_{(1)} < r_{(2)} \dots < r_{(n-1)} < r_{(n)}$ ),

$\bar{r}$  – is the mean of all the  $r_i$ ,

$\bar{r}_{n-m}$  – is the mean of the  $(n - m)$  lowest  $r_i$ .

In the Tietjen-Moore test, the value of the statistic  $E_m$  is compared with the limit value of the test statistic read from relevant tables at the significance level  $\alpha$ .

In the quartile test, in order to detect outliers, the value of the statistic is determined:

$$K = Q_3 - Q_1 \tag{11}$$

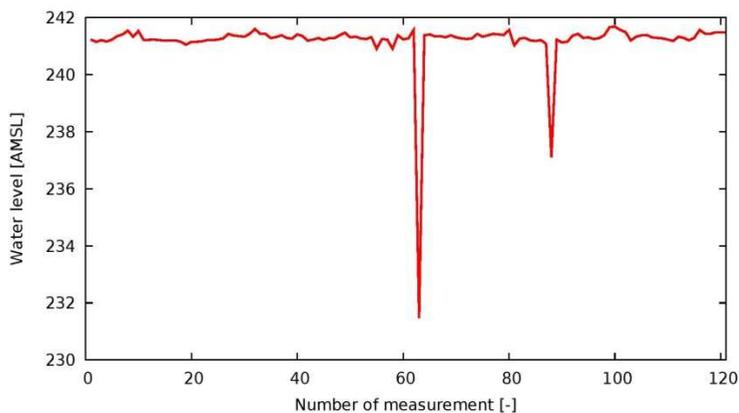
where  $Q_3$  is the value of the third quartile, and  $Q_1$  is the value of the first quartile. If the value of the analyzed quantity is lower than  $Q_1 - 1,5 \cdot K$  or higher than  $Q_1 + 1,5 \cdot K$ , it is considered to be an outlier [14].

### 3 Study results

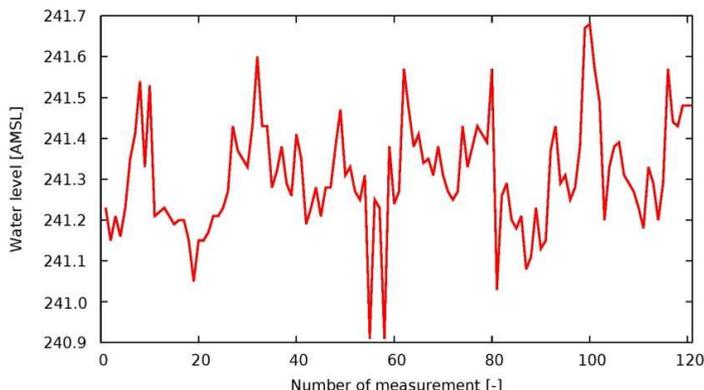
The obtained results are summarized in Table 1. Moreover, Fig. 3 presents a graph of changes in water levels in the PO25 piezometer of the Dobczyce dam before the detection and removal of the outliers, while Fig. 4 illustrates a graph of changes in water levels in the PO25 piezometer of the Dobczyce dam after the detection and removal of the outliers.

**Table 1.** Summary of the number of outliers for selected statistical tests for the Dobczyce dam in the years 2012-2016.

Denotation	Number of detected outlier observations for statistical tests:							
	Dixon's Q		Grubbs'	Hampel's	Rosner's	Iglewicz and Hoaglin's	Tietjen-Moore's	Quartile
	N9	N13						
PO3	0	0	0	0	0	0	0	0
PO4	0	0	0	0	0	0	0	0
PO5	0	0	0	23	0	0	0	23
PO7	1	2	1	1	1	1	1	1
PO9	2	4	2	3	3	3	2	4
PO10	0	0	1	5	5	5	1	5
PO11	0	2	2	4	2	2	2	4
PO12	0	2	2	2	2	2	2	2
PO13	1	2	1	1	1	1	1	5
PO14	0	0	1	3	2	2	1	10
PO15	0	0	1	7	2	2	1	9
PO16	0	4	2	5	5	5	2	6
PO17	1	2	1	1	1	1	1	0
PO19	2	4	2	2	2	2	2	2
PO20	0	0	0	0	0	0	0	0
PO21	0	0	7	10	9	9	7	10
PO22	1	2	1	1	1	1	1	1
PO24	1	2	0	32	0	0	0	0
PO25	2	2	2	2	2	2	2	6
PO26	0	0	0	0	0	0	0	0
PO27	0	0	0	0	0	0	0	1
PO28	2	4	2	2	2	2	2	3
PO29	0	0	0	2	0	0	0	2



**Fig. 3.** Graph of changes in water levels in the PO25 piezometer of Dobczyce dam before detection and removal of outliers.



**Fig. 4.** Graph of changes in water levels in the PO25 piezometer of Dobczyce dam after detection and removal of outliers.

## 4 Discussion and conclusions

Dealing with measurement results considered to be doubtful is one of the greatest problems that can be encountered during any data analysis. Such results are caused by a one-time influence of an important, interfering, yet transient reason, and only for some measurements. A single measurement result encumbered with this kind of error is usually an extreme value (minimum or maximum) of the ascending order set of results. In the case of a measurement series which includes the results of measurements performed under repeatability conditions, such an error is easy to detect and identify. This means that independent test results of the same units are captured using the same method, by the same observer, and using the same equipment, under the same conditions, and in relatively short time intervals.

The concept of an error, occurring in the scientific measurement, is closely linked to uncertainty, which is impossible to completely avoid, and is inseparably connected with the essence of performing a measurement using a given method. In this sense, errors do not characterize mistakes, which can be avoided if measurements are performed with greater diligence. Therefore, the size of errors should be minimized and a way to estimate their size should be found [15].

In the case of periodic measurements of the water table level in an open standpipe piezometer or of the water pressure in a closed standpipe piezometer, the observations are recorded for each piezometer each time. Therefore, there is no sample containing several results of the same observation for the same piezometer, only independent measurements are available. As a result, it is possible to identify some potential disturbances in changes in the water level in piezometers by comparing them with the images from the previous measurement periods. If the measurement and calculations are carried out by one observer who controls and pays attention to the measurement conditions and any disturbances that may occur, it is possible to remove the outlier. For piezometric measurements, it would be advisable and useful to repeat them. As a rule, however, the person who performs the calculations and analyses is only presented with the results of the measurements, without any additional information on the course of their performance.

The Dixon Q-test (variant N9) performed for 23 open standpipe piezometers in the Dobczyce dam in years 2012-2016, detected 13 outliers (0.47% of all results). The variant N13, on the other hand, detected 16 pairs of outliers (1.15% of all results). The Grubbs and Tietjen-Moore tests allowed identification of 28 outliers (1.01% of all results), while the Rosner, Iglewicz and Hoaglin tests demonstrated the existence of 32 such observations (1.44% of all measurements). The highest number of outliers was detected using Hampel's

test – 106 (3.81% of all results). The large number of outliers observed for the Hampel statistical test is related to the structure of this test. If there are a large number of observations with the same values in the data set when calculating the values of deviations  $r_i$  from the median value, the result is equal to 0. Then, the median value from the set of deviations is also equal to 0, so if the module of deviations  $r_i$  takes a value greater than 0, this observation is treated as an outlier [5].

Before the analysis of piezometric data, gross errors should be eliminated, which, even in the case of one outlier observation, may significantly affect its result and cause a false assessment or interpretation of the studied phenomenon. A single measurement result encumbered with a gross error is usually an extreme value (minimum or maximum) of the ascending set of results. Gross errors that may occur during various types of measurements (also in the case of piezometric measurements) are caused by many different factors. The most important reasons for their occurrence include: mistakes when reading or recording instrument readings – most frequent observer's errors (e.g. incorrect numbering of points or accidental swapping of two neighboring numbers), measuring equipment failures, improper use of measuring equipment, specificity of measurements related to the selection of an appropriate method, changed measurement conditions (e.g. adverse weather conditions, icing, etc.), improper performance of the measurement (data collection), storage or preparation for the analysis, mechanical damage to measurement points, and improper input of measurement data into the database.

The paper has been prepared within the scope of the AGH UST statutory research no. 11.11.150.008.

## References

1. A. Gruszecka. *Sta. Cont.* **5**, 102 (2016)
2. S. Lach. *Jou. Eco. Eng.* **19**, 150 (2018)
3. D. Mirosław-Świątek, M. Kembłowski, W. Jankowski. *Ann. of War. Unive. of Li. Scie.* **44**, 25 (2012)
4. Z. Kledyński. *Mod. Build. Eng.* **2**, 54 (2011)
5. S. Lach. *Using the modified scalar product approach for testing the direction of seepage through the earth-fill dams* (AGH University of Science and Technology Press, Krakow, 2016)
6. S. Lach, L. Opyrchał. *Jou. Wat. La. Dev.* **33**, 89 (2017)
7. A. Bąk, S. Lach, L. Opyrchał. *The selected calculations in water engineering* (AGH University of Science and Technology Press, Krakow, 2017)
8. *Materials of the Regional Water Management Authority in Krakow*
9. V. Barnett, T. Lewis. *Outliers in statistical data* (Wiley, New Jersey, 1994)
10. F.E. Grubbs. *Technom.* **11**, 1 (1969)
11. P.L. Davies. *Fres. Zeits. f. Anal. Che.* **331**, 513 (1988)
12. R.D. Gibbons. *Statistical methods for groundwater monitoring* (John Wiley & Sons, New York, 1994)
13. B. Iglewicz, D.C. Hoaglin. *How to detect and handle outliers* (ASQC Quality Press, Milwaukee, 1993)
14. R.R. Johnson. *Elementary statistics* (Duxbury Press, Boston, 1984)
15. P. Konieczka, J. Namieśnik, B. Zygmunt. *Evaluation and quality control of analytical measurement results* (Scientific and Technical Publishing House, Warsaw, 2014)