Machine Learning Methods in the Inclinometers Readings Anomaly Detection Issue on the Example of Tailings Storage Facility

Wioletta Koperska, Maria Stachowiak, Bartosz Jachnik, Pawel Stefaniak, Bartlomiej Bursa, Pawel Stefanek

Machine learning methods in the inclinometers readings anomaly detection issue on the example of tailings storage facility

Wioletta Koperska[1][0000-0002-5882-362X], Maria Stachowiak[1][0000-0001-7501-3437], Bartosz Jachnik[1][0000-0002-7050-4373], Paweł Stefaniak[1][0000-0002-1772-5740], Bartłomiej Bursa[2][0000-0001-8076-7006] and Paweł Stefanek[3][0000-0003-3357-0053]

[1] KGHM CUPRUM Research and Development Centre Ltd, gen. W. Sikorskiego Street 2-8, 53-659 Wroclaw, Poland
[2] GEOTEKO Serwis Ltd., Wałbrzyska Street14/16, 02-739 Warszawa, Poland
[3] KGHM Polska Miedź S.A., M. Skłodowskiej-Curie 48, 59-301 Lubin, Poland

Abstract. Measurement of structure deformation is one of the two most important elements in assessing the current operating condition of a hydro-technical facility, which is especially important when the object is under constant expansion. This is the case of KGHM's Zelazny Most tailing dam which is the largest tailings storage facility (TSF) in Europe. The considerable size of the facility entails a very complex monitoring system consisting of numerous inclinometers, piezometers, seismic stations, geodetic benchmarks, etc. Interpretation of data from such an extensive system requires a certain degree of automation. It is not possible to perform a real-time complete data analysis through human resources, despite several teams responsible for supervision and maintenance of the TSF. The detection of anomalous events is one of the objectives of the monitoring process. This problem concerns, among others, the readings of the inclinometers responsible for the measurement of surface displacements, necessary in the assessment of tailing dam stability. The article presents methods of finding anomalies on the inclinometer with the use of machine learning techniques, which significantly simplifies the process of identifying attention-requiring areas. The effectiveness of the algorithms was tested on data samples from various measurement points. The best method will be to build learning-based supervised classifiers in the decision-making process of the TSF stability.

Keywords: inclinometers, data mining, DBSCAN.

1 Introduction

Tailings Storage Facility (TSF) is one of the largest geotechnical facilities made up of earth embankments built to store uneconomical ore and water from the mining process. An example of a large-scale embankment dam is the Syncrude Mildred Lake Tailings Dyke in Alberta (Canada), the length of which is about 18 km and the height varies from 40 to 88 m. It is the largest earth structure in the world by volume of fill. Historic structural damage often resulted in serious catastrophes with large financial losses and a serious threat to the local community and environment. Therefore, facilities of this type are expected to maintain the highest possible safety indicators and the lowest possible environmental impact. For this reason, these facilities develop advanced monitoring systems covering a wide range of sensors in the field of geotechnical, hydrological, geodetic survey, and seismic networks. Additionally, weather conditions are monitored on an ongoing basis, visual inspections in the field are performed, and satellite data are analyzed. Tracking TSF activity parameters is laborious and time-consuming. Existing measurement networks generate huge amounts of data, which are usually analyzed by several teams of employees. Unfortunately, a complete analysis of the collected data is not possible using human resources. To meet the current expectations of the TSF area, an international consortium was formed and the Illumineation project was launched [www.illumineation-h2020.eu]. One of the goals of the Illumineation project is to develop an Internet of Things platform for monitoring the TSF structures and Big Data analytics using machine learning (ML) methods to support engineers in data analysis. The project assumed the development of a sensor network with low-cost sensors and advanced algorithms to improve the efficiency of data processing, track TSF stability parameters, detect and diagnose potential anomalies and identify potential threats, including estimating the impact of TSF on the environment and the local community. Finally, developed technology will be able to "self-learn" and anticipate potential threats and their potential consequences in advance.

One of the critical tasks is the analysis of displacement data for estimating the deformation of the inclinometer pipe. The inclinometers are used for monitoring horizontal displacements using a probe passing along the pipe. The probe contains a gravity sensor that allows measuring inclination with respect to the vertical. The pipe is usually installed in a borehole or fill. The typical applications of the inclinometers include: determining shear zones in the ground, monitoring the extent and rate of horizontal displacement, monitoring of deflection of bulkheads, piles, or retaining walls. Fig. 1 shows a typical inclinometer body. After the installation the probe is lowered to the bottom the readings are made as the probe is raised incrementally to the top of the pipe, providing data for the determination of initial pipe alignment. The difference between initial and subsequent readings allows the calculation of absolute horizontal deformation at any point along the inclinometer pipe.

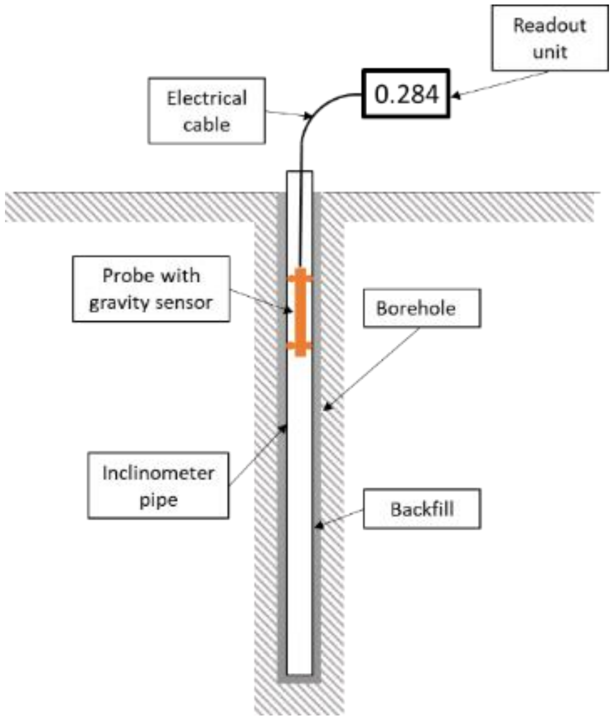

Fig. 1. Inclinometer body.

The TSF structures cause a large increase of stress in the ground and it may result in the formation of shear zones. The shear zones are the areas where the shear strength of the ground material is lower than usual. Stability analyses must detect them using the data obtained by the inclinometers.

From an analytical point of view, this problem comes down to the task of anomaly detection. The problem of detecting anomalies in signals is very well recognized in the literature in many areas. It is especially popular in the task of tracking airport security, detecting fraud (e.g. banking) or cyber-attacks, and technical diagnostics of machines and processes. The greatest challenge is to obtain a very high accuracy of anomaly detection (the lowest possible level of false alarms). Depending on the case, the effectiveness of the detection method may depend on the informativeness of the input signals [Sawicki, *et al.* 2015], the extraction of the robust and effective feature [Ye, 2008, Wodecki, *et al.* 2018], the signal-noise separation technique [Ye, 2018] or a feature classification method [Ahrens, *et al.* 2019].

The article presents methods of detection of anomalies in the inclinometer readings for the needs of TSF stability monitoring. In the beginning, the characteristics of the investigated object based on the example of Zelazny Most TSF were presented. Then, the measurement network and the description of the basic procedures related to the supervision and maintenance of the facility were described. Next, the input data was presented together with a short statistical description. In the next step, the authors described the methodology of the algorithms proposed for the anomaly detection task.

Finally, the article ended with a summary, which presents the further direction of the algorithm development assumed in the Illumineation project.

2 Description of the research object and problem

In the Illumineation project, the main object of research as a test site for developed technology is the Zelazny Most TSF (ZM TSF), the largest reservoir of post-flotation tailings in Europe. Zelazny Most storages waste from mining activities of all KGHM underground copper ore mines located in SW Poland. It is a huge hydrotechnical structure, which covers an area of almost 1,600 ha, and its circumference exceeds 14 km. The height of the dams above the original ground level ranges from 35 m in the southern part to 70 m in the eastern part. The total amount of tailings from the mines in Lubin, Rudna, and Polkowice stored each year at the disposal reaches approximately 30 million tons. The complex monitoring equipment of the reservoir measures all aspects, from geodetic monitoring or water level measurements in piezometers to seismic stations. There are over 40,000 measurement points in Zelazny Most within the developed monitoring network: a geotechnical network, a hydrological network, a geodetic network, and a network of seismic sensors. In total, the network consists of around 2,900 measuring devices and sensors. Field studies, sampling for laboratory tests, and geophysical research are carried out here in cooperation with many national and global research centers [Stefanek, et al. 2017].

Fig. 2. Zelazny Most Tailings Storage Facility located in SW Poland [Stefanek, et al. 2017].

The ZM TSF is located in a complex geological environment. From the ground surface downwards, the foundation soils consist of Pleistocene deposits, including silty lake clays and out-wash sands, rare sandy gravel inclusions, and silty sands. These are underlain by thick layers of freshwater, medium- to high-plasticity Pliocene clays, which incorporate thin, brown coal and sand strata. The Pliocene deposits overlie Triassic strata, which include beds of halite, below which the copper ore body is encountered

[Jamiolkowski, 2014]. The Pliocene deposits contain high-plasticity slikensided clay with very low shear strength. The shear zones occur mostly in those layers. These shear zones are taken into account in stability analyses by modeling the weakened zones with lower shear strength. Hence it is crucial to detect the shear zone in order to obtain an accurate Factor of Safety. Unfortunately, the detection of shear zones is not an easy task. The geotechnical engineer must analyze many factors i.e. inclinometer data, ground conditions, groundwater conditions, etc. Therefore to analyze the inclinometer data more thoroughly there is a need for an algorithm that can learn from the engineers their expertise.

3 Analysis of the inclinometer changes

3.1 Data description

The data includes measurement values for the displacement of the inclinometer from the original state. Measurements were taken up to twice a year from the beginning of the establishment. For each inclinometer, the displacement is measured every 0.5 m of the rod. The displacement is measured in millimeters and its position in the ground is given in meters above sea level. Exemplary data are presented in Fig. 3. The problem of shear zones discussed in the article, visible at 50 m above sea level, is also present-ed. The problem became visible from the measurement on day 26/02/2010.

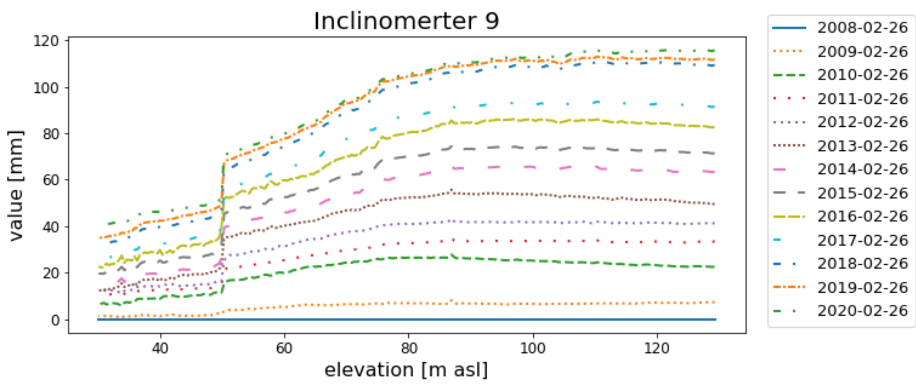

Fig. 3. Value of displacement on the inclinometer over time.

Inclinometers differ in length, location level, and place of implementation as can be seen in Fig. 4. The influence of the substrate on which the inclinometer is located, not considered in this paper. Currently, the surface in which the inclinometer is located is treated as a homogeneous body. It may have a significant impact on the size of the occurred shearing. The factor may be a great addition to the later development of the algorithm.

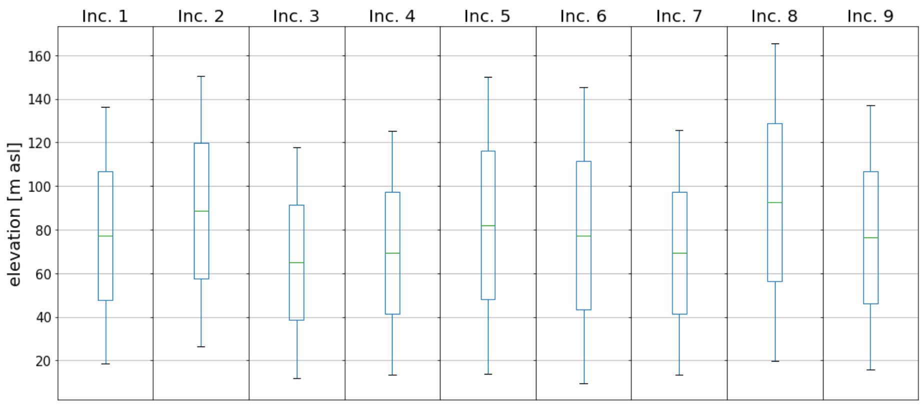

Fig. 4. Elevation's boxplots for each inclinometer.

Shear zones are characterized by a sudden jump to the next level of the inclinometer value. This relationship is clearly visible in the histogram. Fig. 5 shows the histograms for the values from the last measurement for 3 inclinometers. It can be seen that they are multi-modal. Local maximums can correspond to the value levels between the shear zones. This shows that it is possible to group values according to the adopted level. Additionally, it indicates that density clustering should be a good choice.

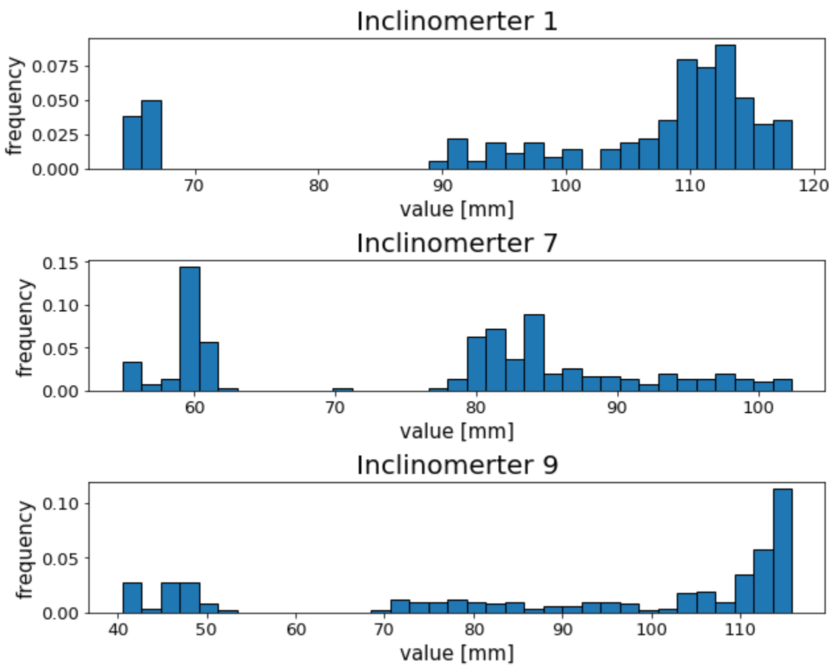

Fig. 5. Histograms of the values of the last measurement for the 3 selected inclinometers.

3.2 Shear zones detection

The shear zones are characterized by a quick shift in relation to the previous placement, which deepens over time. The remaining segments of the inclinometer change only slightly.

To find the elevations on which shear zones occurred, tools that emphasize rapid changes in the data were used. For all samples of each inclinometer owned, the values of the sliding standard deviation, the difference between successive values, and the value of the derivative were calculated.

As can be seen in Fig. 6, the shear areas are visible in all statics. They show jumps at the time of the shear. It is easiest to detect on the last measurement.

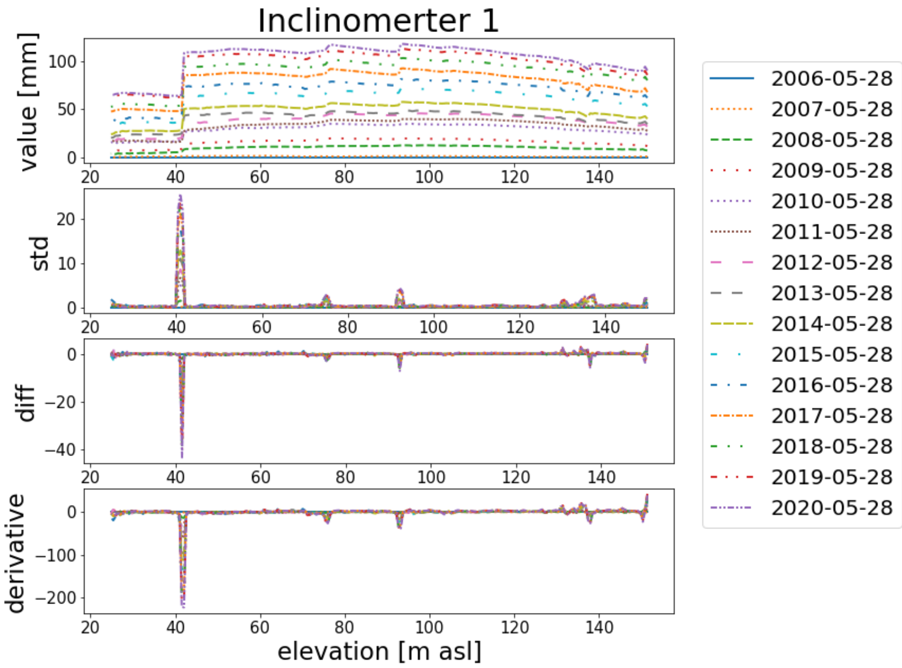

Fig. 6. Statistics values for inclinometer displacement at different elevation levels.

The simplest solution would be to use a threshold or ranges beyond which the data is treated as a shear zone. However, with the current assumptions, the differences between the sizes of the shear zone are so different that it is difficult to determine such ranges. Besides the usual statistical methods, clustering and classification algorithms are commonly used for these problems. The problem can be solved in two ways. With the help of supervised or unsupervised machine learning. The difference between them is that supervised learning needs to input a sample of data (called training data) into the training system. On their basis, the system searches for dependencies corresponding to the given data.

4 Supervised learning

Supervised learning often produces the best results, which is the algorithm that is best at situational awareness. However, as mentioned before, the most important aspect of this type of learning is having a training sample. In this case, it is information about the presence of a shear, which is shown in Fig. 7.

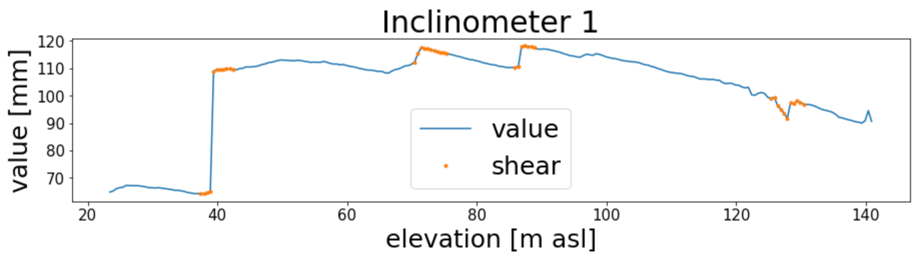

Fig. 7. Inclinometer 1 values with information about the occurrence of shearing.

In order to more precisely describe considered phenomena, additional statistics have been calculated. In addition to the previously presented difference and deviation, the mean value, mean elevation, and kurtosis were added. Values were calculated for the sets containing readings from successive intervals of 5 m. These statistics will constitute predictors set, which we will refer to further as X. The binary response variable Y will indicate the shear occurrence (value equal to one) at a given depth.

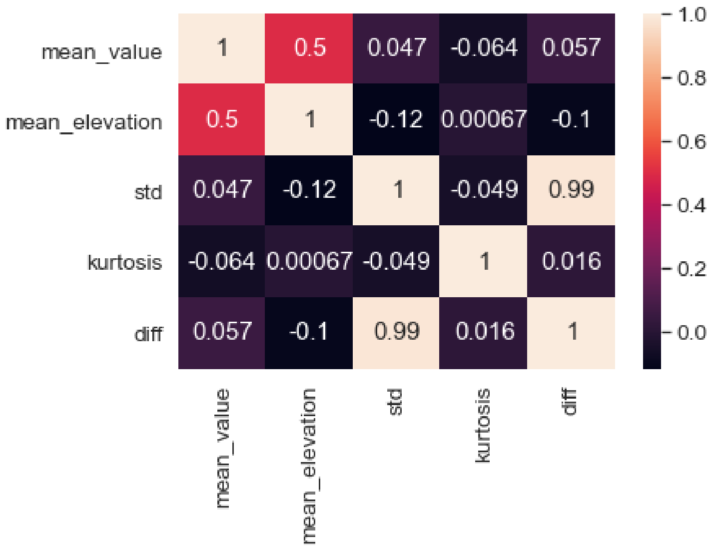

Fig. 8. Correlation matrix.

Fig. 8 shows the correlation matrix of selected variables. As previously shown, the deviation and the difference are strongly correlated with each other, so one of the variables can be rejected. Fig. 9 also shows that these two statistics split two considered classes most clearly.

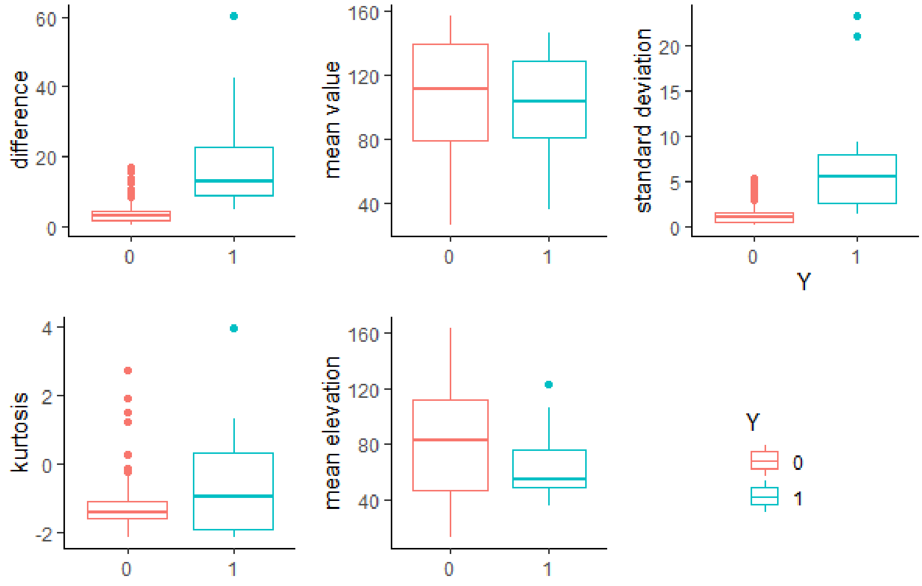

Fig. 9. Selected statistics among two of the considered classes. Y equal to 1 indicates the shear zone occurrence.

As it can be seen, the values of "difference" and "standard deviations" are considerably higher for the samples in which the shear zones have occurred. Yet, it can be seen, that there is a considerable number of samples with high levels of these statistics with Y equal to zero, indicating no shear zone occurrence in that area. The samples with shear occurrence have usually lower elevation values and have a much wider spread of kurtosis statistics.

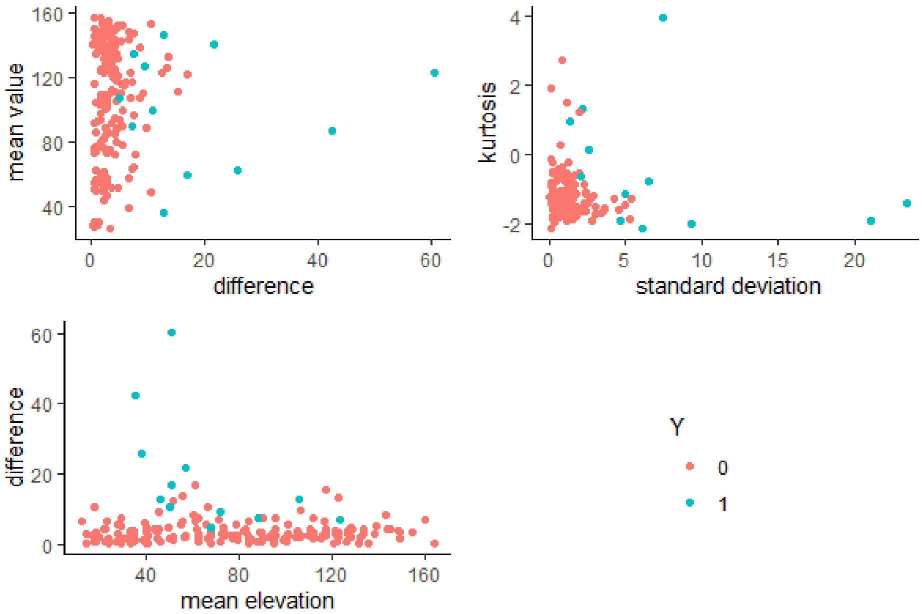

Fig. 10. Summary of individual statistics plotted against each other. The zones with shear occurrence were marked with blue dots.

In this approach, a set of different classifiers will be tested out:

- Logistic regression (LR): involves directly modeling the conditional probability of shear occurrence given the set of beforementioned statistics. In other words, in logistic regression, the conditional distribution of Y (shear occurrence) is being modeled given the set of predictors X.
- Linear Discriminant Analysis (LDA): in this approach, the distribution of X is modeled for each of the response classes, and is then converted using the Bayes formula. This approach assumes, that observations in each class are drawn from the Gaussian distribution.
- Quadratic Discriminant Analysis (QDA): unlike LDA, assumes that each distribution of the response classes has its own covariance matrix. This means, that QDA is significantly more flexible than LDA, and therefore will have a higher variance.
- K-nearest neighbors algorithm (KNN): being the non-parametric method will provide an even more flexible approach to the problem and should benefit the prediction if the assumptions of LR, LDA, and QDA will be not met.
- Neural Network (NN): which is often best suited for high-dimensionality problems.

5 Unsupervised learning

With unsupervised learning, there is no information about the occurrence of the shearing, therefore the problem must be described in a different way. The shear zones can be treated as anomalies in the data, an unexpected pattern that does not match the general behavior. The main idea behind this is to teach an algorithm to detect normally behaving data and then use that information to pinpoint points that do not meet these assumptions. The advantages are that in addition to detecting points outside certain thresholds (extreme values), it also detects those that do not occur frequently [Çelik et al., 2011]. In this case, due to the lack of a final number of clusters, some of the methods like the k-means algorithm cannot be used. DBScan meets all the assumptions. Moreover, in [Thang et al., 2011] it is shown that the application of the DBScan algorithm gives very good results in relation to the previously proposed methods.

The DBScan method allows dividing the samples into groups without prior declaration of the number of clusters. The number of groups is selected by the algorithm. The algorithm groups together points that are close to each other based on a distance measurement (usually Euclidean distance) and have a specified minimum number of points. At the same time, the method indicates points that could not be classified into any of the groups, such values can be understood as outliers - an anomaly occurring in the data, which is exactly what is needed to be highlighted to solve the problem.

The DBScan algorithm takes two initial parameters:

- ε - if the considered point is at ε distance from another point, the algorithm will distinguish these two points as neighbors. Otherwise, the point is considered an outlier.
- minPoints - the minimum number of points in the vicinity of the particular point that allows the area to be considered as dense.

The DBScan checks the surrounding of each point in the sample, and labels them as a:

- Core point if it meets the condition of the minimal number of points in its vicinity. These points form separate clusters.
- Border Point - this is a point that does not meet the density condition, however, there is a core point in its vicinity. It is part of the cluster and constitutes its border.
- Outlier - it is not in the position of the principal point and also does not satisfy the density condition itself.

Points parameterized by DBScan can create three different links between each other. Points are directly density-reachable when at least one of them is a core point and these two are within ε distance from each other. Points are density-reachable when there is a core point that both points are directly density-reachable with. Density-connected points can be located on the opposite sides of the cluster, as long as there is a point density reachable with them.

Given the above definitions, the DBScan algorithm can be described in the following steps of the algorithm (Fig. 11).

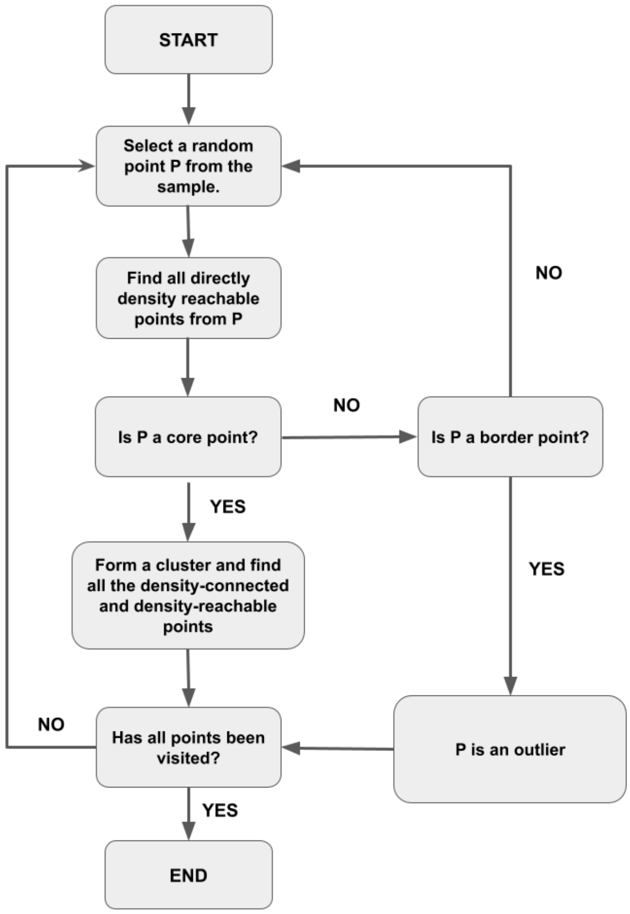

Fig. 11. Scheme of DBScan algorithm.

The algorithm should divide into groups the places where the distortion levels are within the norm and indicate between them moments of shear, which are anomalies not classified into any of the groups.

The following 3 feature vectors were used for clustering:

- elevation values,
- distortion values for the last (newest) sample,
- vector of distortion derivatives at each point (gradient) – it will mainly point to an anomaly (sudden jumps occur then).

6 Application to industrial data

After selecting the variables and assumptions, each of the selected methods was applied to the available data. It was applied to those inclinometers for which occurrence of shear zones was certain. For example, the results of clustering with DBScan for 3 inclinometers are presented in Fig. 12. The plots with the results (the first plot for each inclinometer) were compared with the plots with the manually marked shears (the second plot for each inclinometer).

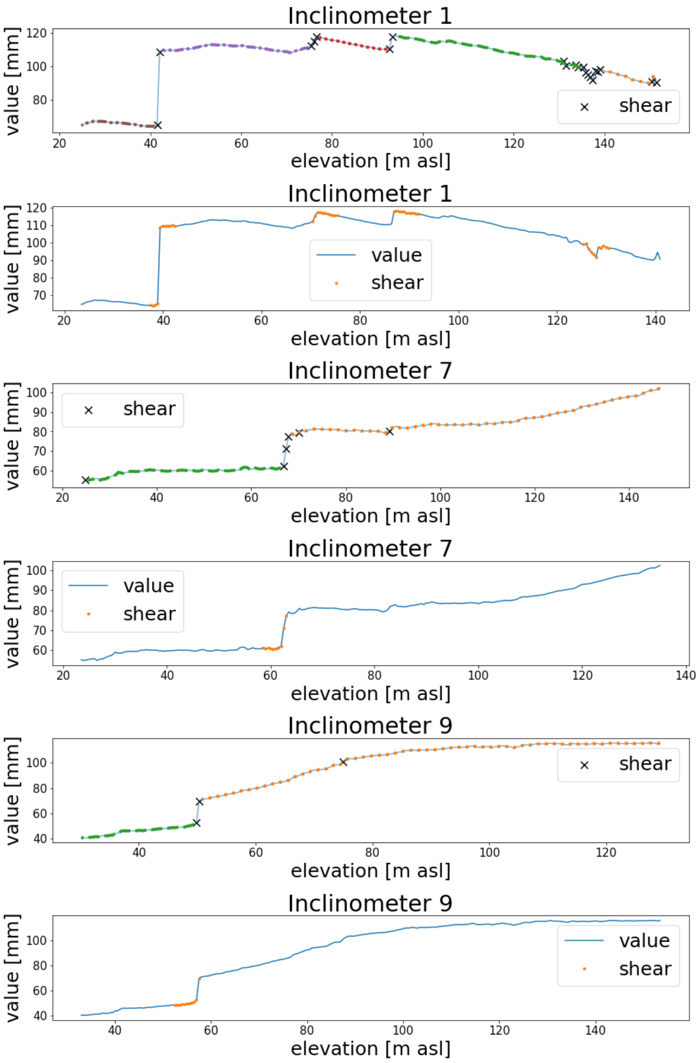

Fig. 12. Clustering result for the presented inclinometers.

As can be seen in the attached picture, the algorithm divided the signal into sections with normal behavior and determined all anomalies. All the shear zones have been detected. However, it is worth noting that the method does not only detect shearings, but also other unusual fragments of the signal.

Similar analyzes were performed for the remaining methods. The results were compiled in the form of a table (Table 1) with calculated performance measures. One of the goals for problem solving is to choose a method that selects false shearing more often than ignores true ones. It comes down to reduce the false negative error, so sensitivity must be maximized.

Table 1. Performance Measures for used algorithms.

Method	Accuracy	Precision	Sensitivity
LDA	0.94	0.89	0.32
QDA	0.94	0.67	0.71
LR	0.96	0.90	0.47
KNN	0.93	0.57	0.49
NN	0.93	0.88	0.67

From the point of view of assumptions, the best methods were obtained by QDA and NN. The methods are characterized by high accuracy and precision. QDA has the highest sensitivity among all the methods, NN has a slightly lower sensitivity, but it is clearly more precise. LDA, which was originally intended to be suboptimal, actually has the lowest sensitivity.

7 Summary

The article describes methods for detecting anomalies in the inclinometer time series resulting from shear planes. The tests were carried out in the Zelazny Most TSF plant. In order to obtain the required accuracy of detection, a multi-year monitoring database obtained from all measurement points was used. One of the main goals in designing the algorithm was to reduce the false negative error.

The article presents different methods from two types of machine learning: supervised and unsupervised learning. The methods use only data from inclinometers, which once again significantly extends the potential area of application far beyond the largest facilities with an extensive monitoring system.

The main goal of the algorithm is to support geoengineering responsible for assessing the stability of the tail dam in the area of detecting anomaly readings. In the next step, the algorithm will be used to build a training sample under the supervision of a domain expert. This, in turn, will be used to build tools supporting decision making and forecasting based on the fusion of data from various sources and Big Data analytics.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 869379.

References

1. Green, G. E., Mikkelsen, P. E.: Measurement of ground movement with inclinometers. In: Proceedings of Fourth International Geotechnical Seminar on Field Instrumentation and In-Situ Measurement, pp. 235-246, Singapore. (1986).
2. Konak, G., Onur, A. H., Karakus, D., Köse, H., Koca, Y., Yenice, H.: Slope stability analysis and slide monitoring by inclinometer readings: Part 2. Mining Technology 113, no. 3: 171-180, (2004).
3. Cała, M., Jakóbczyk, J., Cyran, K: Inclinometer monitoring system for stability analysis: the western slope of the Bełchatów field case study. Studia Geotechnica et Mechanica, 38(2), 3-13 (2016).
4. Sawicki, M., Zimroz, R., Wyłomańska, A., Obuchowski, J., Stefaniak, P., Żak, G.: An automatic procedure for multidimensional temperature signal analysis of a SCADA system with application to belt conveyor components. Procedia Earth and Planetary Science, 15, 781-790 (2015).
5. Ye, N.: Secure computer and network systems. (2008).
6. Wodecki, J., Stefaniak, P., Polak, M., Zimroz, R.: Unsupervised anomaly detection for conveyor temperature SCADA data. In: Advances in Condition Monitoring of Machinery in Non-Stationary Operations, pp. 361-369. Springer, Cham (2018).
7. Ye, N.: Analytical Techniques for Anomaly Detection Through Features, Signal-Noise Separation and Partial-Value Association. In: KDD 2017 Workshop on Anomaly Detection in Finance, pp. 20-32. PMLR (2018).
8. Ahrens, L., Ahrens, J., Schotten, H. D.: A machine-learning phase classification scheme for anomaly detection in signals with periodic characteristics. EURASIP Journal on Advances in Signal Processing, 2019(1), 27 (2019).
9. Stefanek, P., Engels, J., Wrzosek, K., Sobiesak, P., Zalewski, M.: Surface tailings disposal at the Żelazny Most TSF, today and into the future. In: Proceedings of the 20th International Seminar on Paste and Thickened Tailings, pp. 213-225. University of Science and Technology, Beijing (2017).
10. Jamialkowski, M.: Soil mechanics and the observational method: challenges at the Zelazny Most copper tailings disposal facility. Géotechnique, 64(8), 590-618 (2014).
11. Çelik, M., Dadaşer-Çelik, F., Dokuz, A. Ş.: Anomaly detection in temperature data using dbscan algorithm. 2011 international symposium on innovations in intelligent systems and applications. IEEE (2011).
12. Thang, T. M., Kim, J.: The anomaly detection by using dbscan clustering with multiple parameters. 2011 International Conference on Information Science and Applications. IEEE (2011).
13. Illumineation Homepage, www.illumineation-h2020.eu, last accessed 2021/02/24.