A Big Data processing strategy for hybrid interpretation of flood embankment multisensor data

Monika Chuchro, Anna Franczyk, Maciej Dwornik, Andrzej Leśniak

AGH University of Science and Technology, Faculty of Geology, Geophysics and Environmental Protection; al. A. Mickiewicza 30, 30-059 Krakow, Poland; e-mail: chuchro@geol.agh.edu.pl

Received: 15 April 2016; accepted: 21 September 2016

Abstract**:** The assessment of flood embankments is a key component of a country's comprehensive flood protection. Proper and early information on the possible instability of a flood embankment can make it possible to take preventative action. The assessment method proposed by the ISMOP project is based on a strategy of processing huge data sets (Big Data). The detection of flood embankment anomalies can take two analysis paths. The first involves the computation of numerical models and comparing them with real data measured on a flood embankment. This is the path of model-driven analysis. The second solution is data-driven, meaning time series are analysed in order to detect deviations from average values.

Flood embankments are assessed based on the results of model-driven and data-driven analyses and information from preprocessing. An alarm is triggered if a critical value is exceeded in one or both paths of analysis. Tests on synthetic data demonstrate the high efficiency of the chosen methods for assessing the state of flood embankments.

Keywords**:** Flood embankment, anomaly detection, numerical modelling, Big Data, flood embankment stability assessment

INTRODUCTION

Floods in Poland usually cause significant damage in the areas they affect with the most dangerous effects being in towns and industrial areas. The amount of damage depends on the scale of flooding and the catastrophic floods that happened in Poland in 1997 and 2010, for example, caused much more serious damage than the local floods that happen after limited heavy rainfall, mainly in mountain areas.

Dealing with local and regional flooding requires flood-warning systems. Systems that monitor the water level in rivers and the condition of river embankments are the most common and can be found in many countries. These systems can measure a single physical and mechanical parameter inside the embankment or a set of such parameters. The most common example of a simple monitoring system is an analogue or digital water gauge. However, integrated monitoring systems not only allow the measurement, but also the transmission and recording of experimental data in a dedicated database or data-warehouse while some systems also do data processing and interpretation. Such systems have been developed, for example, in Denmark and the Netherlands, and are now being constructed in Poland as part of the Computer System of River Embankment Monitoring (Polish acronym – ISMOP) project, financed by NCBiR (Baliś et al. 2011, Krzhizhanovskaya et al. 2011, Pyayt et al. 2012,

http://dx.doi.org/10.7494/geol.2016.42.3.269 ISSN 2299-8004 | e-ISSN 2353-0790 | Wydawnictwa AGH

Pengel et al. 2013, Piórkowski & Leśniak 2014, ismop.edu.pl).

The aim of the ISMOP project is the construction and implementation of an integrated embankment monitoring system for high water level conditions. To perform field experiments, a full-size artificial soil embankment was constructed. Inside this embankment, a system for continuous, passive or active temperature and thermal conductivity measurements using optical fibres was mounted. Furthermore, as well as other sensors, over one thousand thermal sensors were deployed in the embankment, allowing monitoring of dynamic water filtration processes and stability loss. In combination with dedicated software, a telemetric system allows data transmission in real time to a database where it is aggregated and analysed to detect dangerous changes in embankment stability.

The recognition of dynamic physical processes inside a flooded embankment is a complex task. Data from a huge number of sensors of different types has to be interpreted and the results compared with the known, internal structure of the embankment and its physical and mechanical parameters. Numerical modelling is an effective method for comparing knowledge about the studied object and the results of measurements. In our case, the mutual numerical simulation of the temperature, water pressure in pores, and stresses inside the embankment was performed. After comparison of measurements with modelling results, anomalous behaviour of temperature changes and filtration can be detected. Anomalies may be an indicators of ongoing filtration processes that may result in embankment failure.

This article presents a method for measured data interpretation linked with numerical modelling and integrated in real time with transmission. The method implemented in the ISMOP project as part of a computer system that performs data gathering, processing, and interpretation is now being tested during field experiments.

EXPERIMENTAL FLOOD EMBANKMENT

An experimental embankment in the shape of a stadium containing a water reservoir was built in Czernichów near Kraków (southern Poland) (Fig. 1). The height of the flood embankment is 4.5 m above terrain level and can be filled with water up to 4 m. The water reservoir has a length of about 150 m (N-S direction) and the width varies from 8 m at the bottom to 28 m at the top (W-E direction).

The western wing of the embankment has a symmetrical shape with a slope height to width proportion of 1:2. The eastern part has an asymmetrical shape: the wing on the water side has a proportion of 1:2.5 and the wing on the air side has a proportion of 1:2. The eastern embankment also has three zones, built from a more permeable material than other parts of the construction. The geometry and materials used in the construction are typical of existing flood embankments in Poland (Borys 2007).

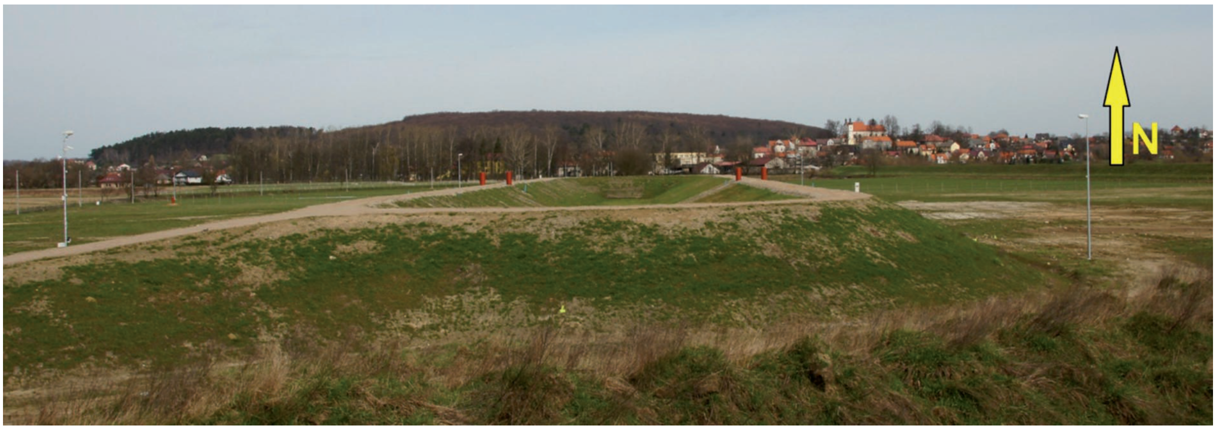


Fig. 1. *Photograph of the experimental embankment*

The embankment was fitted with different sensors for monitoring the state of the embankment. Thermal properties were measured using two loops of optical fibres with an accuracy of 0.3°C, and few stationary sensors located in three W-E profiles (marked as profile 1-1, 2-2 and 3-3 in Figure 2A) with an accuracy of 0.1°C. Pore pressure, piezometers and strain sensors are also located in these three profiles. The location of these sensors for the northern profile is presented in Figure 2B.

DATA ANALYSIS STRATEGY

For a comprehensive assessment of flood embankment state, two different analysis paths were proposed for data from sensors, based on Big Data assumptions. A characteristic feature of both data analysis approaches is the ability to parallelise calculations in order to shorten analysis time. The chosen analysis methods are able to analyse data from sensors of the selected flood embankment parts.

The first approach to the analysis of time series recorded by sensors placed in the flood embankment applies numerical modelling. This is a model-driven approach, as it uses numerical models. Numerical models include the structure of a flood embankment and the material from which it is formed, and simulate various processes in the structure during rising and falling water levels. The created models are compared with real data in order to select the most similar group of numerical models and analyse differences between models and real data.

The second approach to data analysis involves comparing time series from a particular sensor for a period in which no deviations from average values were observed with a period in which changes might have occurred (this is usually the last couple of days registered). The goal is to detect anomalies in the data. This is a data-driven approach to data analysis, as it focuses solely on the interpretation of measurement data.

In the preliminary stage of the process, a module for capturing errors from data acquisition and extreme values is also used. The preprocessing module also prepares data for analysis. The final effect of both approaches and the information from the pre-processing step is the assessment and visualisation of flood embankment state around the analysed sensors.

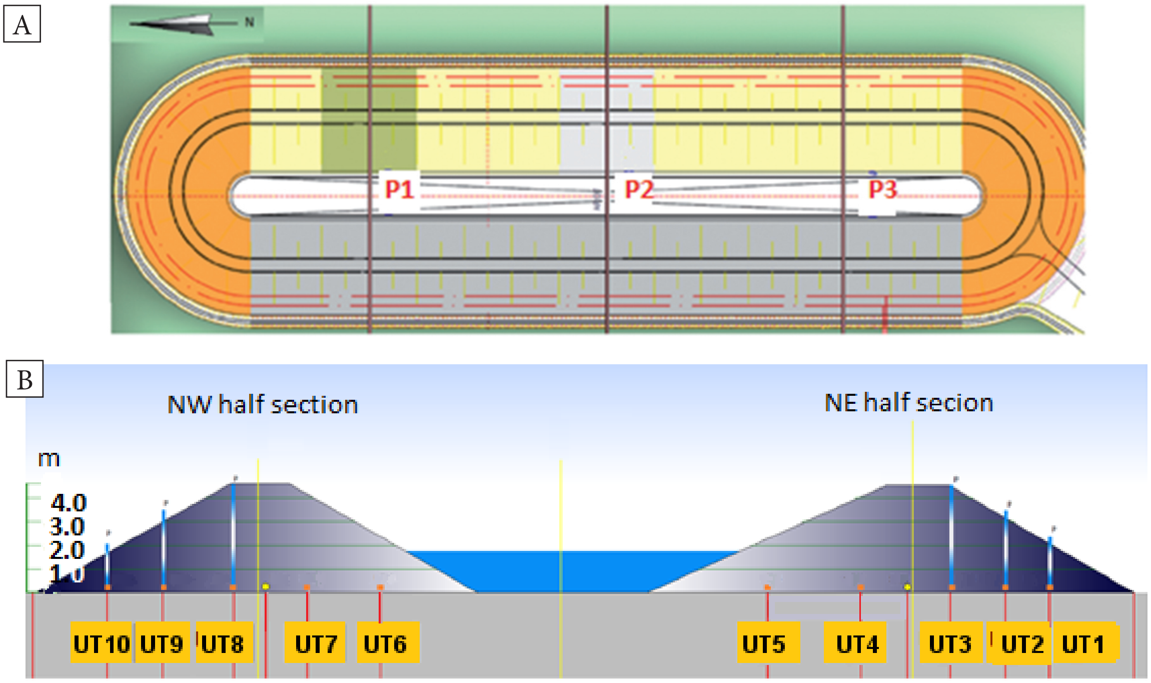


Fig. 2. *Scheme of the flood embankment and location of sensors on profile 1 (on the north part of flood embankment) (Zieliński & Chrost 2015)*

The data analysis is a multi-stage process and takes place in two modules, taking into account both the model-driven and data-driven approaches. Temperature and pore pressure sensor measurements are wirelessly transmitted to a Raw Data Repository. This data requires preprocessing, after which it is saved to a Measurements Database that is subsequently available for analysis, unlike the Raw Data Repository. The Raw Data Repository is a database of raw data that is only accessible to a small number of applications and processes and access to this database is read only.

The Measurements Database is used by both data analysis modules at the same time. The model-driven module also uses a third database that stores the results of numerical modelling. The general flow of flood embankment assessment is shown in Figure 3.

Preprocessing

The data cleaning and preparation for further analysis stages is an important component of any analysis and data modelling process. The function of this module is to remove erroneous values, for example from malfunctioning sensors, and extreme values from data downloaded from the Raw Data Repository (Kotu & Deshpande 2015). If there is missing data for a given point in time, linear interpolation is performed based on the values of the nearest sensors. In addition, the aim of the preprocessing module is to create a new timestamp variable. This variable stores information about the time parameter measurements, rounded to one common value for all measurements made in a given period (less than 1 minute). Transformed data is written to the tables in the Measurements Database, where it is available to other analysis modules.

Numerical models

The analysis of embankment stability can be supported by numerical modelling. Such models allow the examination and understanding of underlying physical processes that most influence embankment stability and failure and are crucial at the real experiment planning stage. A detailed study required the modelling of the thermal, filtration and mechanical processes that occur concurrently and affect the embankment state. Analysis using coupling techniques involving the deformation generated by filtration and mechanical processes is able to explore the mutual influence of these physical phenomena. Two mechanical effects may be observed in such cases.

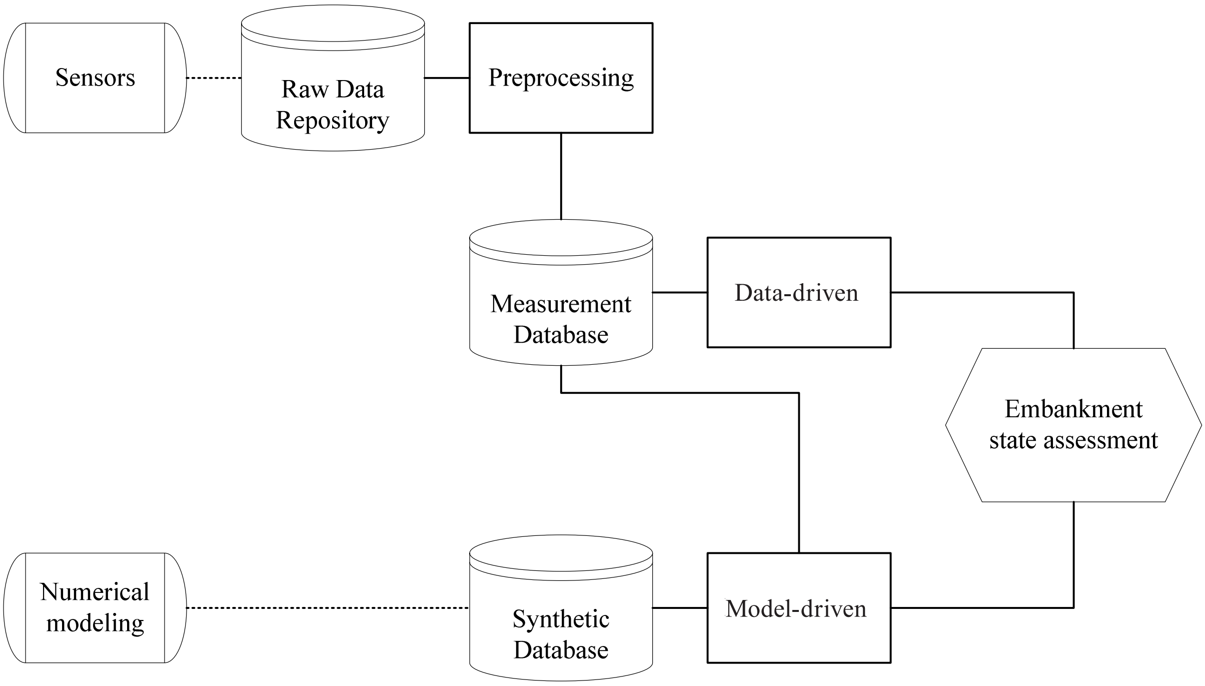


Fig. 3. *General scheme of flood embankment state assessment.*

On one hand, changes in pore pressure cause changes in effective stress that affects the soil and may induce plastic yield. On the other hand, the fluid that fills the pore spaces in the soils reacts to mechanical volume changes by a change in pore pressure. The thermal field is also coupled with the flow-mechanical calculation. Numerical modelling of embankment stability includes both forced convection, in which heat is carried by fluid motion, and free convection, for which fluid motion is caused by density differences due to temperature variations.

The modelling was performed using the FLAC two-dimensional explicit finite difference program that enables the performing of coupled mechanical-fluid flow-thermal processes (Itasca Consulting Group, 2011).

The numerical modelling was carried out within the framework of the quasi-static Biot theory (Biot 1955) that is widely used for modelling macroscopic behaviour of soils with porous flow (Melnikova et al. 2011). The simulation is based on single-phase Darcy flow in a porous medium.

The modelling of embankment stability was applied for several different flooding wave scenarios (Pięta et al. 2015). During each of the assumed scenarios, the values of the basic parameters were calculated and saved with the assumed time steps. As a result of the computation, parameters such as horizontal and vertical displacement, horizontal and vertical stress, horizontal and vertical fluid flow, temperature and pore pressure were saved for every point of the assumed computational node. The values of parameters important for embankment stability and changes during the assumed flooding process scenarios were analysed in order to assess the state of the embankment. This was achieved through the construction of a knowledge base of behaviours that might indicate embankment instability.

Model-driven analysis

Numerical modelling and information about phenomena inside the embankment due to external factors and changes in the riverbed can be used to assess the state of a flood embankment (Pyayt et al. 2015). The aim of the model-driven module is to perform a comparison between time series of measured parameters from sensors and numerical models. The general scheme of this module is shown in the figure below (Fig. 4).

Analysis is performed separately for each flood embankment half section. A time series with 100 observations is downloaded from the Measurements Database for each sensor from the half section of flood embankment selected for analysis. The correlation matrix for the downloaded time series is evaluated in order to reduce the number of calculations. If the value of the Pearson correlation coefficient for two adjacent sensors is greater than a predetermined value (0.9), the data from these sensors is averaged.

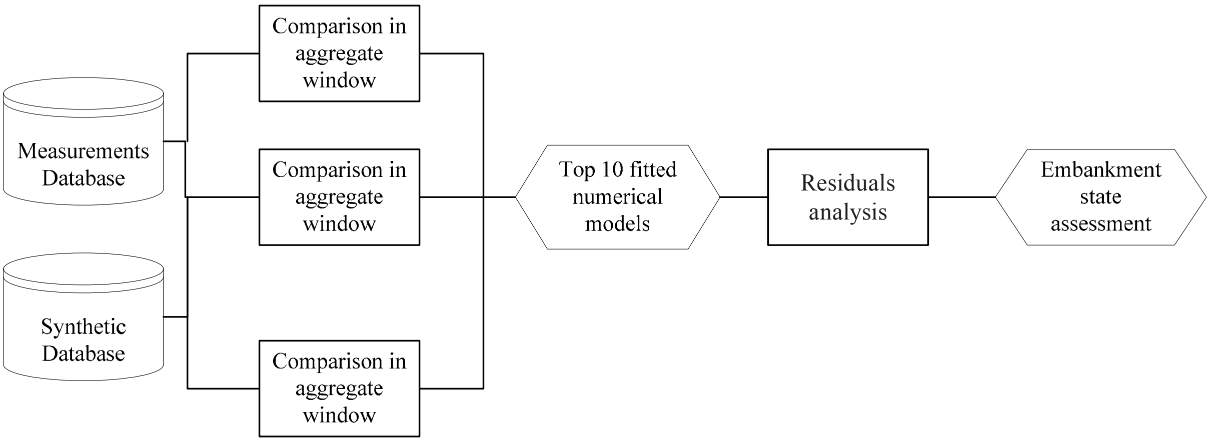

Measurements Database
Synthetic Database
Comparison in aggregate window
Comparison in aggregate window
Comparison in aggregate window
Top 10 fitted numerical models
Residuals analysis
Embankment state assessment

Fig. 4. *General scheme of model-driven module*

A time series are collected from all numerical models from the Synthetic Database for nodes corresponding to the position of chosen sensors (from the half section). The time series from nodes corresponding to the sensors from which data was averaged are also averaged. The most important stage of this module is to compare the time series from modelling with the real time series registered by sensors. The Mean Square Error (MSE) was selected as a measure that assesses the similarity between the time series from the sensors and the time series from the numerical model. The Mean Square Error is calculated in an aggregate window of 100 observations (Chuchro et al. 2015). Among the calculated values of MSE, the smallest value with the amount of offset for the numerical modelling is recorded. For each sensor in the half section, ten best numerical models are selected with the smallest MSE values that suit the comparable time series from the half section based on the MSE. In subsequent iterations of the model-driven module, the length of time series from sensors increases by 1 each time. Until MSE rises above the critical value, comparison occurs only between the pre-selected ten best-fit numerical models.

For ten best-fit numerical models, the differences (residuals) between this data and the time series from sensors are computed. The values and variances of the residuals are checked for increases in time that exceed the critical value designated in tests. If a model matches the first part of data but the similarity subsequently decreases, a phenomenon has probably affected the value observations.

Alternatively, the future state of the flood embankment can be assessed. The probability of maintaining stability can be calculated based on information about the last (resulting in either failure or a stable embankment) of ten best-fit numerical models selected as best suited to the time series of measurements.

Data-driven analysis

Another approach to assessing an embankment state is the analysis of data from sensors, aimed at detecting changes and deviations from standard values. Temperature and pore pressure time series are periodic, with very weak daily periodicity and explicit periodicity associated with the seasons. A characteristic feature of the analysed time series is the lack of a trend and the strong influence of irregular components related to weather conditions.

The aim of the analysis is to detect emerging anomalies at the end of the analysed time. Anomalies might be present as group values higher than the average values, a huge single value, or a growing trend. Such changes in the average level of the phenomenon, referred to as anomalies, are an indicator of unfavourable changes in the flood embankment that could indicate instability.

Anomaly detection in a time series of sensor measurements was achieved using methods based on Fast Fourier Transform (FFT) and frequency models (Welch 1967, Maslova et al. 2016). In anomaly detection, two time series of the same length and the same sampling step are analysed and recorded from one sensor in similar atmospheric conditions. One of the time series is a data set for the absence or presence of anomalies is known. The second time series is a data set that we want to test for the occurrence of anomalies. In order to detect anomalies, spectral density values are compared for the first model. In the second model, for which a frequency model is calculated, changes in adjustment are rated, expressed by determination coefficient (R^2), distribution of residuals, and changes in residuals variance (Fig. 5). If the difference between the assessed parameters exceeds the critical value in the analysed time series, an anomaly is detected. The second case is when comparable parameters are not significantly different from each other and an anomaly was detected in a series that was compared, then the analysed time series probably contains an anomaly.

Embankment state assessment

The assessment of the flood embankment is carried out for each sensor separately and then generalised to the whole analysed section of the embankment. In time moment *t*, for a single sensor an assessment consists of three groups of results:

- results of model-driven analysis: numbers of best-fit numerical models, information about residuals,
- results of data-driven analysis: information about detection of anomalies (with two methods), residuals analysis,
- information from the preprocessing module (extremal value detection).

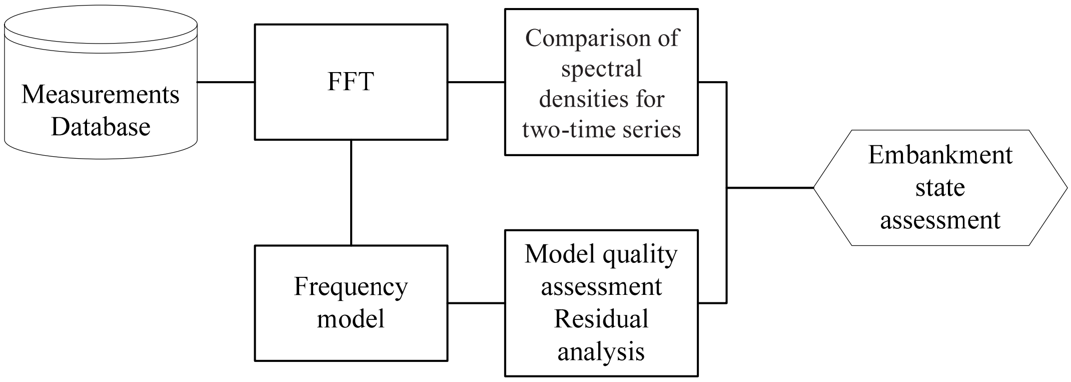

Measurements Database
FFT
Comparison of spectral densities for two-time series
Embankment state assessment
Frequency model
Model quality assessment Residual analysis

Fig. 5. *General scheme of the data-driven module*

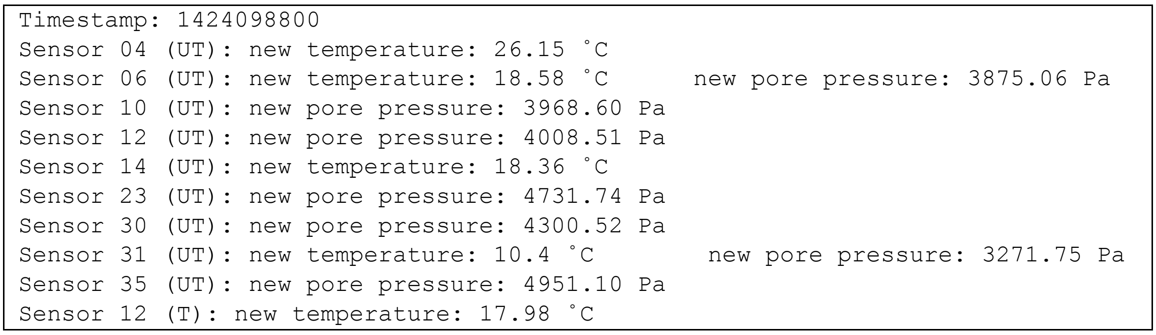
If the variability of residuals from period to period exceeds a critical value, or variance residues exceed the critical level, information about the instability of the flood embankment for the analysed sensor is generated. Likewise, if the data-driven module detects an anomaly or the variability of residuals from period to period exceeds a critical value, or variance residuals exceed the critical level, information about the instability of the flood embankment for the analysed sensor is generated. Similarly, if during the preprocessing of data, the measurement value from the sensor exceeds a critical value, the information is generated.

Information about the instability of the flood embankment near the analysed sensor is generated if even one of the selected methods detects anomalies. As a global assessment of flood embankment state, information about anomalies in the analysed section is considered. If an anomaly is found or critical data values are exceed, then alarm information is sent.

TEST

Validation tests for stability evaluation algorithms were based on synthetic data. Data from numerical modelling with an additional irregular normal distribution component were used for the test. All modules presented in this paper were tested. The preprocessing module test was performed on a data set containing 33×500 observations. Each observation contained geographic coordinates, sensor type, the value of the measured parameters, the time to write to the database, and ID sensor. Each prepared file contained three missing pieces of data for measurement of temperature and pore pressure and seven incorrect values including the error code (9.99900). The output of the module is shown in Figure 6.

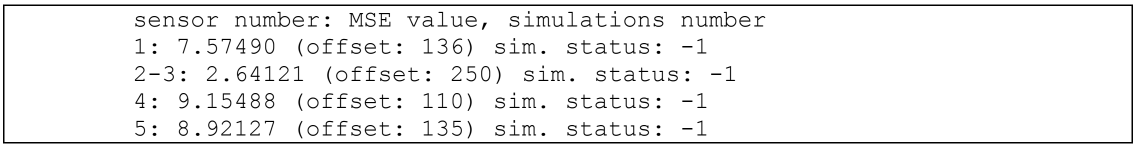
The model-driven module test was performed on a small sample of data containing only five time series (from one half-section, for example P1 sensors UT1:UT5).



```
Timestamp: 1424098800
Sensor 04 (UT): new temperature: 26.15 ˚C
Sensor 06 (UT): new temperature: 18.58 ˚C      new pore pressure: 3875.06 Pa
Sensor 10 (UT): new pore pressure: 3968.60 Pa
Sensor 12 (UT): new pore pressure: 4008.51 Pa
Sensor 14 (UT): new temperature: 18.36 ˚C
Sensor 23 (UT): new pore pressure: 4731.74 Pa
Sensor 30 (UT): new pore pressure: 4300.52 Pa
Sensor 31 (UT): new temperature: 10.4 ˚C        new pore pressure: 3271.75 Pa
Sensor 35 (UT): new pore pressure: 4951.10 Pa
Sensor 12 (T): new temperature: 17.98 ˚C
```

Fig. 6. *Preprocessing module test*

Each time series contained 100 observations and was created from the aforementioned numerical models, with the additional irregular component. Two groups of numerical modelling were also used in this test: sim status: –1 and sim status: –2. Time series from the half-section were compared against numerical models iteratively in time windows (number of iteration is called offset). The results of testing were the MSE values was the smallest and the corresponding offset for modelling. Values of residuals variance are shown only if they exceed critical values. The output from the model-driving module is presented below (Fig. 7) .



```
        sensor number: MSE value, simulations number
        1: 7.57490 (offset: 136) sim. status: -1
        2-3: 2.64121 (offset: 250) sim. status: -1
        4: 9.15488 (offset: 110) sim. status: -1
        5: 8.92127 (offset: 135) sim. status: -1
```

Fig. 7. *Model-driving module test*

The data-driven module was also tested on synthetic data to which an irregular component was added. The test was performed on 18 time series with and without anomalies. Six time series with a length of 192 observations were generated without anomalies. For each time series without anomalies, three similar time series were generated with anomalies. Each anomaly added to the time series has different parameters, values and time stamp. Both methods of anomaly detection are effective. Better results were obtained with anomaly detection method 2, where data were compared with frequency models. This method detected 17 of the 18 anomalies in the data: the problem was the growing anomaly (trend) with the following parameters:

$$T = (178 - 192) + 6.5 + n \cdot 0.2,$$

where $n = 1,..., 15$.

The first method did not recognise 3 of the 18 anomalies from the test data.

CONCLUSIONS

The article presents a comprehensive method for assessing the state of flood embankments. The proposed method is based on sensors mounted in a flood embankment that allow periodic and continuous monitoring of its condition. This method is suitable for the assessment of new or existing flood embankments. The initiation of flood embankment condition monitoring requires investment in sensors that measure temperature and pore pressure, the performing of numerical modelling, and the creation of IT facilities. However, the costs incurred can minimise losses in the detection of instabilities in the flood embankment.

Currently simulations are being carried on in the flood embankment in Czernichów.

In addition, tests of analysis modules on synthetic data and real data measured on experimental flood embankment are being performed.

This work is financed by the National Centre for Research and Development (NCBiR), Poland, project PBS1/B9/18/2013 (no. 180535).

REFERENCES

Baliś B., Kasztelnik M., Bubak M., Bartyński T., Gubała T., Nowakowski P. & Broekhuijsen J., 2011. The urbanflood common information space for early warning systems. *Procedia Computer Science*, 4, 96–105.

Biot M.A., 1955. Theory of elasticity and consolidation for a porous anisotropic solid. *Journal of Applied Physics*, 26, 2, 182–185.

Borys M., 2007. Przepisy i wymogi oraz aktualny stan obwałowań przeciwpowodziowych w Polsce. *Woda-Środowisko-Obszary Wiejskie*, 20, 7, 25–44.

Chuchro M., Lupa M. & Pięta A., 2015. A concept of time windows length selection in stream databases in the context of sensor networks monitoring. [in:] Bassiliades N. et al. (eds.), *New Trends in Database and Information Systems II*, Advances in Intelligent Systems and Computing, 312, Springer International Publishing, 173–183.

Itasca Consulting Group, I. 2011. *FLAC Fast Lagrangian Analysis of Continua and FLAC/Slope – User's Manual*.

Krzhizhanovskaya V.V., Shirshov G.S., Melnikova N.B., Belleman R.G., Rusadi F.I., Broekhuijsen B.J., Gouldby B.P., Lhomme J., Baliś B., Bubak M., Pyayt A.L., Mokhov I.I., Ozhigin A.V., Lang B. & Meijer R.J., 2011. Flood early warning system: design, implementation and computational modules. *Procedia Computer Science*, 4, 106–115.

Kotu V. & Deshpande B., 2015. *Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner*. Morgan Kaufmann.

Maslova I., Ticlavilca A.M. & McKee M., 2016. Adjusting wavelet-based multiresolution analysis boundary conditions for long-term streamflow forecasting. *Hydrological Processes*, 30, 1, 57–74.

Melnikova N.B., Shirshov G.S. & Krzhizhanovskaya V.V., 2011. Virtual Dike: multiscale simulation of dike stability. *Procedia Computer Science*, 4, 791–800.

Pengel B.E., Krzhizhanovskaya V.V., Melnikova N.B., Shirshov G.S., Koelewijn A.R., Pyayt A.L. & Mokhov I.I., 2013. Flood early warning system: sensors and internet. [in:] *Floods: From Risk to Opportunity*, IAHS Publication, 357, International Association of Hydrological Science, 445–453.

Pięta A. & Krawiec K., 2015. Random set method application to flood embankment stability modelling. *Procedia Computer Science*, 51, 2668–2677.

Piórkowski A. & Leśniak A., 2014. Using data stream management systems in the design of monitoring system for flood embankments. *Studia Informatica*, 35, 2, 297–310.

Pyayt A.L., Shevchenko D.V., Kozionov A.P., Mokhov I.I., Lang B., Krzhizhanovskaya V.V. & Sloot P.M.A., 2015. Combining Data-Driven Methods with Finite Element Analysis for Flood Early Warning Systems. *Procedia Computer Science*, 51, 2347–2356.

Pyayt A.L., Mokhov I.I, Kozionov A.P., Kusherbaeva V.T., Lang B., Krzhizhanovskaya V.V & Meijer R.J., 2012. Data-driven modelling for flood defence structure analysis. [in:] Klijn F. & Schweckendiek T. (eds.), *Comprehensive Flood Risk Management: Research for Policy and Practice*, CRC Press, Boca Raton, 77.

Welch P.D., 1967. The use of Fast Fourier Transform for the Estimation of Power Spectra: A method based on time averaging over short, modified periodograms. *IEEE Transactions on Audio and Electroacoustics*, 15, 2, 70–73.

Zieliński M. & Chrost A., 2015. *ASP Czernichów. Eksperymentalny wał przeciwpowodziowy – projekt, wykonanie w części dotyczącej poboru, transmisji i wizualizacji danych*. Internal documents of ISMOP project.