A Concept of Time Windows Length Selection in Stream Databases in the Context of Sensor Networks Monitoring

Monika Chuchro, Michał Lupa, Anna Pięta, Adam Piórkowski, and Andrzej Leśniak

Department of Geoinformatics and Applied Computer Science
AGH University of Science and Technology
al. Mickiewicza 30, 30-059 Cracow, Poland
{chuchro,apieta}@geol.agh.edu.pl
{mlupa,pioro,lesniak}@agh.edu.pl
http://www.geoinf.agh.edu.pl

Abstract. Monitoring systems are a source of large amounts of data. These streams of data flow down as information which, in the case of sensor networks is often associated with the measurement of the selected physical signals. Processing of these data is a non-trivial issue, because accurate calculations often require dedicated solutions and large computing power.
In the case of flood embankment monitoring systems the essence of the calculation is the analysis of time series in terms of similarities to herald danger scenarios. This analysis also includes data series from neighboring sensors, which increases the difficulty of the calculation. This paper proposes the concept of a data analysis system, allowing for dynamic evaluation of the embankments[1].

Keywords: time series, time windows, stream database, embankments, flood

1 Introduction

Integrated embankment monitoring systems have been created in a few countries, such as in the Netherlands [2, 12]. In Poland, in cooperation with the AGH University of Science and Technology in Cracow and Cracow companies SWECO

[1] This is the manuscript of:

M. Chuchro, M. Lupa, A. Pieta, A. Piorkowski, A. Lesniak: A Concept of Time Windows Length Selection in Stream Databases in the Context of Sensor Networks Monitoring. In: New Trends in Database and Information Systems II, Springer, AISC, Vol. 312, 2015, pp 173-183.

Hydroprojekt Ltd and Neosentio, project ISMOP (Computer Monitoring System for Flood Embankments) has been developed. This was founded under the NCBiR project (National Centre for Research and Development, Poland), which involves the creation of a monitoring system of static and dynamic behavior of the embankments that works in real time [17, 19].

The basic problem of flood embankment is to control their condition, both during the flood season and during exposure to the interruption of the embankment. Visual observation of the shaft during the flood does not answer whether the section of the shaft has lost stability, and if so, how long it can effectively resist the flood. The aim of the research is to provide answers to the question of whether the measurement of technical parameters inside the shaft, such as temperature, pore pressure and humidity, can allow us to estimate the probability of damage to the embankment. An additional difficulty in giving a clear answer to this question is heterogeneous building embankments - the most common material that is near the bund.

1.1 Sensor network for embankments monitoring system

This article concerns the difficult issue of assessment of flood embankment stability. This is a complex problem due to the amount of data processed from sensors located in the experimental embankment and the calculations required to assess flood embankment stability.

In order to facilitate the evaluation, an experimental embankment was divided into sections with a length of 1km. In each section of the embankment, there are 1,000 sensors arranged in 5 layers. A cross section of the experimental embankment is shown in Fig. 1. Each sensor measures parameters which can influence the condition of the embankment. These include: temperature, pressure, pore water, humidity, stress, strain and electrical conductivity. A sensor network is thus created [17, 13]. Due to the advances in the electronic design, sensor network use IPv6 [3, 18] for communication purposes. It enables possibility to expose its functionality as services [20] and make them coherent with the server side part of the system. Depending on the depth of the sensor in the embankment, the average measured values may vary slightly. These differences in measured values are caused by weather conditions and hydrogeological sensors located in the subsurface layers. All measurements are carried out with a time

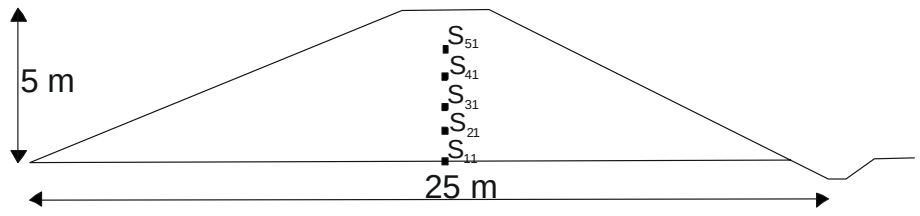

Fig. 1. The schema of experimental flood embankments with location of sensors (S_{nm}).

step of 15 minutes, giving 540 000 observations every hour (15×1000×9×4), and almost 13 million per day on a 1km length of the experimental embankment. Taking into account the length of the flood embankment in the municipalities, the huge amount of incoming data is a big challenge. The first problem that must be solved is the collection and storage of data. The second problem is assessment of the state of the experimental embankment. This state assessment includes a comparison of data flowing from the sensors with the previously created numerical models. Also the state assessment contains the visualization of models in a short period of time, shorter than getting the new batch of data from sensors.

Real data flowing from the sensors can be written as a multivariate time series of moments of time step of 15 min (1) [21, 14].

$$\begin{gathered} Y(t) = \{y_a, y_b, ..., y_i\} \\ t = 1, 2, ..., \tau. \end{gathered} \tag{1}$$

where:

- t - time step,
- y_a - observation of one parameter on the time step t,
- i - measured parameters $(a, ..., i)$.

The time series of a single sensor includes, in addition to the seven parameters, a timestamp and the coordinates of the point in space where the sensor is located. In addition to the data from the sensors, weather data from a weather station located in close proximity to the experimental flood embankment are processed and collected.

Assessment of the embankment stability is based on a comparison of data flowing from the sensor with dynamic numerical models of the embankment stability performed in the Flac [9]. Flac is a numerical modeling code for advanced geotechnical analysis, used in project for dynamical modeling. The concept used in Flac which is based on Lagrangian analysis is applied in many field (e.g. in obtaining a fluid-flow-based mechanical model for prediction of probability distribution [1]). Simulations in this environment are conducted on a numerical model of the embankment with the same parameters as the experimental embankment. In addition to the initial conditions, external conditions are used to simulate real phenomena occurring on and within the experimental embankment and in its proximity. Due to the possibly infinite number of opportunities to model the initial and external parameters, team modeling in Flac have to uniformly sampled space of possible initial and external parameters. The final result is a dynamic simulation of the embankment stability under the influence of different values of various external factors.

The simulated evaluation of the embankment stability should be consistent with the actual experimental embankment state. To facilitate searching the Flac models database, a treelike structure was created in which the first divisions concerned the initial conditions, and further divisions covered external conditions.

2 A Concept of Flood Embankment Condition Assessment

Due to the nature of the data and analysis requirements, no solution available is able to properly evaluate the condition of the experimental flood embankment. The number of observations in 1 km of the embankment, and the requirement that the assessment be realized in not more than 15 minutes, exclude a relational database. Daily, seasonal and annual cyclicality hinder the creation of generalized models, which is why correct space sampled of possible solutions will be a big challenge, especially for air temperature and humidity. On the other hand a strong dependence of autocorrelation means that the best solution seems to be to analyze the data into time windows using stream databases.

The general process of flood embankment condition assessment is shown in Fig. 2. Data flowing from the 1,000 sensors are evaluated for correctness. In the case of abnormalities or anomalies an error message is sent to the system administrator. An important step is the preprocessing of data, which includes examining the similarity of the values measured by the sensors in the moment τ and $\tau - 1$. If the module assessment of the shaft is not running, and the

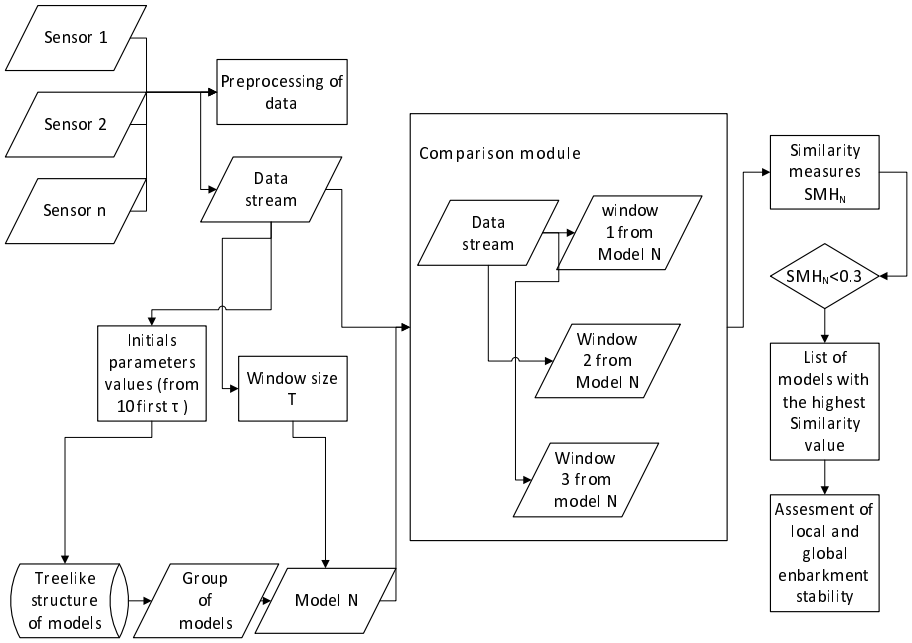

Fig. 2. The concept of flood embankment condition assessment.

difference between the measured values of the parameters for the individual sensor does not exceed 5%, the system goes into idle mode until time $\tau + 1$.

If the difference between the values of the parameters of the time τ and $\tau - 1$ (even of a single parameter or the sensor) is higher than 5%, that module of embankment stability assessment is enabled. Shutdown of assessment modules occur only when all of the scenarios created in Flac suggest "very good condition of experimental flood embankment" and the subsequent n measured values for all parameters did not show differences between them more than 5%. In the first step after starting the module assessment of the embankment stability, the last ten values for each sensor and each parameter are read from the database. With these values, average values are calculated, which we consider as initial values. Calculated initial values are compared with the initial values for the Flac scenarios. The group of scenarios with the highest degree of similarity are selected for further evaluation of the embankment stability.

As a measure of similarity between two multidimensional windows ranks time series considered one of several measures of mathematical and statistical. Initially, a typical measure of similarity like as Pearson's correlation coefficient is excluded because of the lack of fit to the nature of multivariate and nonlinear time series. Eight similarity and dissimilarity measures has been selected from many existing measures to test. In many fields it has been shown that the different similarity measures can lead to substantial differences in final results and the similarity measure performance can be influenced by dataset characteristic [11]. For testing were chosen: Jacard coefficient, Sorensen coefficient, Czekanowski coefficient, determination coefficient, average measure of the error, mean absolute percentage error (MAPE), average absolute error known as L_1 Norm (2), root mean square error (square L_2 Norm) [4, 8, 5].

$$L_1 = \frac{1}{n} \sum_{i=1}^{n} |Y_\tau - Y_p| \tag{2}$$

The selection of the optimal similarity measure to use was made on the basis of experiment. 100 pairs of series of 1000 observations were generated in terms discraibed in [16], corresponding to series of a single sensor (real data) and to Flac models. The "real" time series were composed of a slowly varying sinusoidal function with strong irregular components, which were Gaussian noise. Flac models deviated slightly from the real data time series, but the Flac model time series do not have Gaussian noise. Due to the nature of the modeled data the similarity measure should:

- properly evaluate similarity ranks in the case of one of the time series having a stochastic component (noise),
- evaluate the similarity in the case of linear and nonlinear dependencies,
- lower the value of similarity in the models which do not adapt to the extreme values, which may indicate deterioration of embankment stability,
- be easy to interpret.

The measure of average absolute error (2) fulfills the abovementioned requirements, and this measure will be used in the project as a measure of similarity. Experimental embankment stability assessment runs iteratively. As the first group

of embankments dynamic models was selected, this group, which was calculated for the highest similarity measure for the initial parameters values. Generally speaking, the same number of observations are selected for the first sensor and the first dynamic model, counting the real time series from a sensor 1. Then, for both multidimensional time series a measure of similarity is calculated. The calculated value of similarity and the number of initial observations of the time window are written to a temporary table. In the next iteration, the time window is shifted to the right by one observation in the dynamic model and then the measure of similarity is calculated again. If the new measured value is higher than the similarity stored in a temporary table, these values are overwritten. If the iterations come to the last observation in the first dynamic model, the highest value of the similarity measure, together with the number of initial observations of the time window, the number of dynamic model and the result of the script is saved to the table with the best match scenarios, provided that the measure of similarity does not exceed the limit value. The predetermined threshold based on the similarity measure experiments is 0.3. After calculating the measure of similarity for the first dynamic model and real times series from the first sensor, similarity measures are calculated for all dynamic model parameters preselected by initial values of parameters.

The final of the dynamic model is stored in binary form, in which 0 represents break or destruction of the flood embankment and 1 is maintained stability of the embankment. The local assessment of embankment stability, called "state", for a single sensor positioned in the experimental embankment is calculated based on Bayesian probability from all chosen by similarity measue models finals and is stated as $P(C_k)$, where P is a Bayesian probablility and C_k is a sensor number [15, 24]. For a single sensor at time t can take one of five states chosen arbitrary:

- Alarm status - threat of break or damage to the embankment. The system sends a message indicating the status of the emergency to the designated administrative units along with the location of the threat (3).

$$P(C_k) < 0.2 \tag{3}$$

- Warning status - threat of break or destruction of the experimental embankment, but also with the possible improvement of the embankment condition. The system sends a message indicating the status of the emergency to the designated administrative units along with the location of the threat (4).

$$0.2 \leq P(C_k) < 0.4 \tag{4}$$

- Neutral status - depending on external conditions the experimental embankment may stabilize, or a real threat of break or destruction of the embankment may occur (5).

$$0.4 \leq P(C_k) < 0.6 \tag{5}$$

- Good status - the likelihood of disturbances of the state of the experimental embankment is low, however, some dynamic models indicate the possibility of destruction of the embankment in the future (6).

$$0.6 \leq P(C_k) < 0.8 \tag{6}$$

– Very good status - the likelihood of disturbances to the state of the shaft is very low (7).

$$P(C_k) \geq 0.8 \tag{7}$$

The final threshold values of $P(C_k)$ will be modified during the experiments on real flood embankment.

Global assessment of the stability state of the entire experimental embankment is calculated as the average value of local states calculated in parallel to each other. A special case is the occurrence of a warning or alarm condition even for a single sensor (9 parameters). The status of the warning or alarm for a single sensor automatically grants the same status to the global assessment. In the case of an output state indicating possible damage or break of the experimental embankment, it is possible to run predictive modules allowing the assessment of the future state of the embankment on the basis of current data. For each evaluation time reports are generated with graphical presentation of the results, which are sent to the control and administration unit.

3 The Issue of Time Window Length Selection

Data in the stream database are processed and modeled in the form of streams of data. A single stream S can be called a multicollection of elements in the form $< s, \tau >$ in which s is a tuple, and τ is the time of appearance of the item, or timestamp. A characteristic feature of streams is their possible unlimited length and repeatable values τ. One of the most important features of the stream database is the time of arrival of the data to the database (timestamp). The older the data is, the less that data is associated with the current data streaming to the database. For this reason and because of the computational complexity, for processing streaming data generally only the last n observations are used with sliding time windows [6, 7, 22]. The time window converts the data stream into a table. In the stream each tuple has a timestamp τ. The time window is a function that defines the life of the component in a table based on timestamp tuple. The arrival of a new tuple will update the current time of the operator in accordance with the time stamp of the tuple. Depending on the determining method, we distinguish agglomeration, sliding and sequencing time windows. The initial time window is in the agglomerative windows parameter, and the window size changes with each new input tuple. After updating, the current time window is increased by the unit of time. A sliding (sliding window) - parameter window is the size of T in units of time. Each new input tuple is inserted into the table. The next step is to update operator current time. Then the table is cleared of tuples whose timestamp is beyond the scope of the window. The sliding window allows the grouping of tuples from within T units of time. This can be expressed as (8):

$$R(\tau) = \{s | s, \tau' \in S \wedge (\tau' \leq \tau) \wedge (\tau' \geq max(\tau - T, 0)\}. \tag{8}$$

where:

– S - data stream,

- τ - current time operator,
- τ' - the newest tuple,
- T - window size.

Time-based sliding window contains all the elements at the time τ of the stream S, which appeared from time $\tau - T$ to time τ. Another type of window is the tuple-base sliding window, which is associated with a fixed number of tuples, and contains the N most recent tuples that have emerged to date τ. The last type of sliding window is the partitioned sliding window. The stream S is divided into subattributes with identical features. The next step is selection of N most recent tuples to time τ. Shifting / fixed window - window parameter is the size of T in units of time. The interval is defined as:

- Beginning marker: $i * T$
- End marker: $(i + 1) * T$

3.1 A Concept of Experimental Flood Embankments Condition Assessment in Time Windows

Assessment of the experimental embankment is carried out iteratively, in the time windows. Already, on the basis of preliminary findings, two types of time windows were excluded (partitioned sliding and shifting windows) as not fulfilling the requirements of the project. Two types of windows were chosen as the most promising, fitted simultaneously to the data coming from the sensors and the planned evaluation of the stability of experimental embankment. The first type of time window selected is the sliding window with a fixed number of observations of 100 tuples from a single sensor, with a time-resolution of 15 minutes (Fig. 3). The advantage of such a solution, based on streaming databases, is that it does not require intermediary relational database calculations and so performance problems that could result from this kind of data storage are avoided. In addition, fixed window size reduces the assessment time, by reducing the number of function calls which count the length of the time series of measurements from the sensor, and function calls choosing an adequate number of observations for the dynamic models. The disadvantage of fixed size time windows is the possibility of loss of part of the dynamic models, due to the loss of initial readings from the sensors. To prevent the loss of significant parts of the models it would be necessary, for each time window $\tau - T$, to calculate the new initial value and to research the treelike structure of dynamic models. This research, unfortunately, will increase the number of time-consuming calculations and increase hardware resource requirements. The second solution is to use the agglomerative time windows (Fig. 4). The disadvantage of this solution is an increase in the number of calculation and in the time of calculation. With each new moment, a time window increases by 9 parameter values for a single sensor, which gives 9 000 new observations for a 1 km stretch of the experimental shaft for a moment, and during the day up to 864 000 additional observations. Additionally, the computing time also increases because of selecting and counting invoke functions.

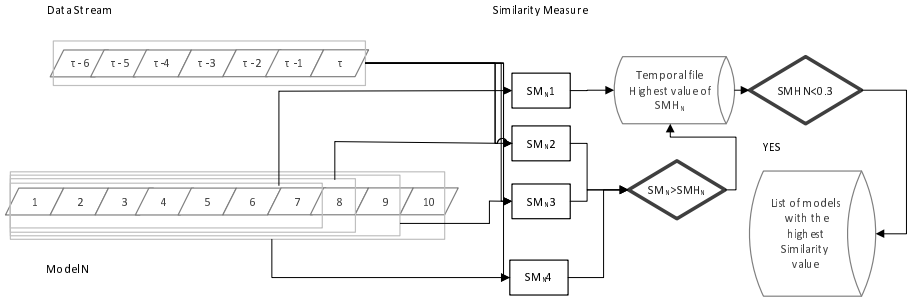

Fig. 3. Processing in agglomerative window.

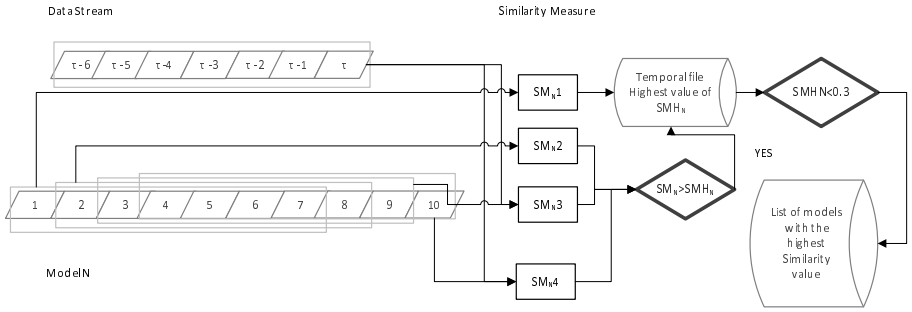

Fig. 4. Processing in sliding window.

In the classical approach to stream databases, the computationally less expensive solution is to introduce time windows with a fixed number of observations, despite the possibility of losing a valuable part of dynamic models of the stability flood embankment development. However, this solution is not optimal in terms of obtaining full information about the status of the flood embankment. To be able to take advantage of the benefits of the agglomerative window, a way to shorten the time of analysis is needed, which will be iteratively longer and longer with each new τ. Reducing the time may be accomplished by searching and reading tuples in the stream database, or reducing the number of parameters in the analysis. However the second solution reduces the possibility of a correct assessment of the embankment stability, the same as sliding windows. For this reason, we decided to use indexation of data dynamic model and in time series from sensors.

4 Conclusions and Further Work

This project on embankments monitoring is important due to the possibility of providing an early warning about embankment breach. This results in greater security in flood risk areas. It is necessary to establish and implement a dedicated system due to the lack of existing software that meets the requirements associated with the collection, analysis, and prediction of embankment condition. For this reason, the authors have proposed their own solutions, based on stream databases, where the information is evaluated by time series analysis tools.

The essence of time series analysis of data from the sensor network is the use of time windows, and the selection of the type and length of the windows, which will be adequate to the cyclicality periodic data (e.g. daily, monthly, yearly). This work also considers the selection of optimal time windows in the context of determining the similarity measure of time series and model data. The advantages and disadvantages of the proposed solution in relation to computational complexity were also presented. In future, the management of stream databases for discussed flood embankment will be considered [23, 10].

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

This work was partly support by the AGH - University of Science and Technology, Faculty of Geology, Geophysics and Environmental Protection, as a part of statutory project number No.11.11.140.032.

References

1. Augustyn, D.: Using the model of continuous dynamical system with viscous resistance forces for improving distribution prediction based on evolution of quantiles. In: Kozielski, S., Mrozek, D., Kasprowski, P., Malysiak-Mrozek, B., Kostrzewa, D. (eds.) Beyond Databases, Architectures, and Structures, Communications in Computer and Information Science, vol. 424, pp. 1–9. Springer International Publishing (2014)

2. Balis, B., Kasztelnik, M., Bubak, M., Bartynski, T., Gubała, T., Nowakowski, P., Broekhuijsen, J.: The urbanflood common information space for early warning systems. Procedia Computer Science 4, 96–105 (2011)
3. Brzoza-Woch, R., Czekierda, L., Dlugopolski, J., Nawrocki, P., Psiuk, M., Szydlo, T., Zaborowski, W., Zielinski, K., Zmuda, D.: Implementation, Deployment and Governance of SOA Adaptive Systems. In: Ambroszkiewicz, S., Brzezinski, J., Cellary, W., Grzech, A., Zielinski, K. (eds.) Advanced SOA Tools and Applications, Studies in Computational Intelligence, vol. 499, pp. 261–323. Springer Berlin Heidelberg (2014)
4. Cha, S.H.: Comprehensive survey on distance/similarity measures between probability density functions. International Journal of Mathematical Models and Methods in Applied Sciences 1(4), 300–307 (2007)
5. Clements, M., Hendry, D.: Forecasting economic time series. Cambridge University Press (1998)
6. Golab, L., Özsu, M.T.: Issues in data stream management. ACM Sigmod Record 32(2), 5–14 (2003)
7. Golab, L., Özsu, M.T.: Processing sliding window multi-joins in continuous queries over data streams. In: Proceedings of the 29th international conference on Very large data bases-Volume 29. pp. 500–511. VLDB Endowment (2003)
8. Hamilton, J.D.: Time series analysis, vol. 2. Princeton University Press, Princeton (1994)
9. Itasca Consulting Group, Inc.: FLAC Fast Lagrangian Analysis of Continua and FLAC/Slope – User's Manual (2008)
10. Koudas, N., Ooi, B.C., Tan, K.L., Zhang, R.: Approximate nn queries on streams with guaranteed error/performance bounds. In: Proceedings of the Thirtieth international conference on Very large data bases-Volume 30. pp. 804–815. VLDB Endowment (2004)
11. Kozielski, M., Gruca, A.: Correlation of genes similarity measures based on go terms similarity and gene expression values. In: Czachorski, T., Kozielski, S., Stanczyk, U. (eds.) Man-Machine Interactions 2. Advances in Intelligent and Soft Computing, vol. 103, pp. 137–144. Springer Berlin Heidelberg (2011)
12. Krzhizhanovskaya, V.V., Shirshov, G., Melnikova, N., Belleman, R.G., Rusadi, F., Broekhuijsen, B., Gouldby, B., Lhomme, J., Balis, B., Bubak, M., et al.: Flood early warning system: design, implementation and computational modules. Procedia Computer Science 4, 106–115 (2011)
13. Li, J., Cai, Z., Li, J.: Data management in sensor networks. In: Wireless Sensor Networks and Applications, pp. 287–330. Springer (2008)
14. McGovern, A., Rosendahl, D.H., Brown, R.A., Droegemeier, K.K.: Identifying predictive multi-dimensional time series motifs: an application to severe weather prediction. Data Mining and Knowledge Discovery 22(1-2), 232–258 (2011)
15. McKenzie, C.R.: The accuracy of intuitive judgment strategies: Covariation assessment and bayesian inference. Cognitive Psychology 26(3), 209–239 (1994)
16. Pieta, A., Bala, J., Dwornik, M., Krawiec, K.: Stability of the levees in case of high level of the water. In: 14th SGEM Geoconference On Informatics, Geoinformatics And Remote Sensing – Conference Proceedings. vol. 1, pp. 809–815 (2014)
17. Piórkowski, A., Leśniak, A.: Using data stream management systems in the design of monitoring system for flood embankments. Studia Informatica 35(2), 297–310 (2014)
18. Szydlo, T., Gut, S., Puto, B.: Smart applications: Discovering and interacting with constrained resources ipv6 enabled devices. Przeglad Elektrotechniczny pp. 221–226 (06 2013)

19. Szydlo, T., Nawrocki, P., Brzoza-Woch, R., Zielinski, K.: Power aware mom for telemetry-oriented applications using gprs-enabled embedded devices - levee monitoring use case. In: Federated Conference on Computer Science and Information Systems (FedCSIS), 7-10 Sept. 2014, (in print) (2014)
20. Szydło, T., Suder, P., Bibro, J.: Message oriented communication for ipv6 enabled pervasive devices. Computer Science 14(4) (2013)
21. Vlachos, M., Hadjieleftheriou, M., Gunopulos, D., Keogh, E.: Indexing multidimensional time-series. The VLDB JournalThe International Journal on Very Large Data Bases 15(1), 1–20 (2006)
22. Wang, W., Li, J., Zhang, D., Guo, L.: Processing sliding window join aggregate in continuous queries over data streams. In: Advances in Databases and Information Systems. pp. 348–363. Springer (2004)
23. Zhang, R., Koudas, N., Ooi, B.C., Srivastava, D.: Multiple aggregations over data streams. In: Proceedings of the 2005 ACM SIGMOD international conference on Management of data. pp. 299–310. ACM (2005)
24. Zyphur, M.J.: Bayesian probability and statistics in management research: A new horizon. Journal of Management 39, 5–13 (2013)