PREDICTIVE MODEL AND METHODOLOGY FOR OPTICAL TELECOMMUNICATIONS INFRASTRUCTURE

*V. Douda, M. Jánešová**

Abstract: In this article a predictive model and a novel methodology of processing the data measured in the physical model of an optical telecommunications infrastructure is presented. The task is motivated by practical use of the results of experiments in the environment of the telecommunications network. We present an original predictive model and methodology, reflecting the specifics of examined infrastructure. The probabilistic prediction of the occurrence of emergencies is calculated via cluster analysis techniques used in Bayesian approach in the n-dimensional data space. The predictive model is experimentally verified on real data. Results of experiments are interpreted for practical use in real environment of the telecommunications infrastructure.

Key words: *cluster analysis, crisis management, optical transport network*

Received: November 20, 2015 **DOI:** 10.14311/NNW.2016.26.020
Revised and accepted: May 12, 2016

1. Introduction

Nowadays, many companies are not prepared to loss of ICT Services. It's only after the service outage that people realize, how much they are dependent on the perfectly functioning telecommunications infrastructure.

After such events, a retrospective analysis takes place to find out what has happened. This eventually leads to a proposal, and sometimes even to the implementation of a backup solution. In the ideal world, everyone should be able to avoid the risk, or could always prepare a backup plan. Avoiding risk is not possible at all times and preparing a backup solution is usually associated with expensive investment. Since the implementation of such solutions is usually associated with the investment, implementation of these solutions is not frequently executed. Absence of back-up solution can therefore eventually become more expensive.

Company management must therefore have sufficient information to consider investment decisions. They must be able to assess whether the imminent loss may threaten the stability of the company. Contingency and disaster recovery planning for the construction of optical infrastructure [1, 15] requires a long-term strategy.

*Vladimír Douda – Corresponding author, Mária Jánešová, Czech Technical University in Prague, Faculty of Transportation Sciences, Konviktská 20, CZ-110 00 Praha 1, Czech Republic, E-mail: vladimir.douda@seznam.cz, janesova@fd.cvut.cz

One of the common approaches is to build a robust fibre infrastructure, i.e. building more independent routes between major nodes. But it is important to compare the higher cost of investment, which is to be spent for building a backup optical transmission routes, with imminent risk [9]. In our research, we focused on prediction and quantification of the risks involved. We analysed the events that happened on a physical model of optical infrastructure, and predicted the conditions for such an event that could happen again, or what will be its financial impact. This information is then extremely valuable for the operator of this infrastructure because it can better prepare in advance for imminent scenario. In our work, we concider probabilistic prediction of different modes of behaviour [6]. Described instance is called a mixture model [3, 12, 14] in many cases used for estimation of mixture parameters [5]. From mentioned non-linear dynamic real system [10] the occurrence of emergencies is calculated via cluster analysis using mixture models clustering techniques [2] used in Bayesian approach in the ndimensional data space and especially those used in the area of data mining [16, 17]. We also analysed the role of the reliability of the telecommunications infrastructure during the preparation and planning of preventive measures based on statistical probability and prediction of emergencies. We suggest how efficiently and in advance plan network resilience [11, 13] and how proactively respond to the increased likelihood of an emergency of specific parameters.

The paper is organized in the following way. Section 2 introduces experimental data mining. In the Section 3.1 proposed algorithm of designed methodology is described and in this section we also provide an illustrative experiment. It demonstrates classification abilities of the proposed clustering algorithm. Section 3.2 contains an evaluation of the experiment.

2. Experimental data

The data for our experiments were collected in the Czech Republic in the city of Prague. For the purpose of our research, we selected the backbone of an optical transmission network, as one of the key parts of the telecommunications infrastructure of every major provider of ICT services. For this research, we picked a portion of the optical infrastructure of a local telecommunications operator namely T-Mobile Czech Republic as. This portion of the optical infrastructure is a comprehensive subset of the optical transport networks, specifically located in Prague. More precisely, we have analysed the impact of priority 1 (highest severity) incidents on data services provided by this infrastructure in the years 2006–2013. Examining the qualitative parameters generated by the physical model, we gained information about the outages that occurred during that period. During the first iteration, we used parameters such as: the number of days since the last event, the amount of financial loss and the type of incident that caused the incident in the first place. In the next part of the research, we have experimentally verified the other causal variables. Due to the use of computerized statistical processing of data by programming mathematical language SCILAB [18], we transferred all attributes into numbers. Data are available on special demand.

3. Experiment

3.1 Proposed methodology

During our experiments, we examined interdependence of nine variables as a variable parameters of specific incidents described in the Section 2.

Mentioned variables are:

- number of days since the last event
- the type of an extraordinary event
- financial loss
- city district
- quarter of the calendar year
- month of the calendar year
- season of the calendar year
- daily temperature
- wind speed.

The methodology will be presented using three selected variable parameters. In ndimensional space (three dimensional in this case) defined by the examined variables (days since the last event, the type of an extraordinary event, financial loss), we visualized the points where each point represents one incident, see Fig. 1. The base shows the dependence of variables "financial loss" and influence of "extraordinary event". The length on how high the point is located above this base represents the period of time since the last incident. This period of time from the last occurrence is represented by the length of the connector of point and base.

During the experiments, we noticed that the points in the space illustrated in Fig. 1 are not randomly distributed in space, but forms clusters. That led us to the idea to focus on these clusters and found the rules under which the clusters are formed and what they represent. Whether the formation of these clusters has any interpretation of the cluster location in the space remains to be discussed. There is a reference to advanced statistical methods and its applications, materials for PhD students [8], where this method can be found. Furthermore, we have engaged an cluster analysis method, while looking for a given set of input data set clusters and their centers and assessed the competence of individual values to the corresponding clusters to minimize the error function [7]. Based on the input data (input vectors), we analysed the data in three-dimensional space illustrated in the Fig. 1 from the perspective of formation of clusters. By this method, we were looking for working modes of a system. It was necessary to scale data to the zero mean and unit variance, at the beginning of the calculations, in order to appropriately display the findings graphically and consequently analyze it. The Fig. 1 illustrates the projection of points in a normalized threedimensional space. During our research we frequently examined also projection into two orthogonal planes - the plan and

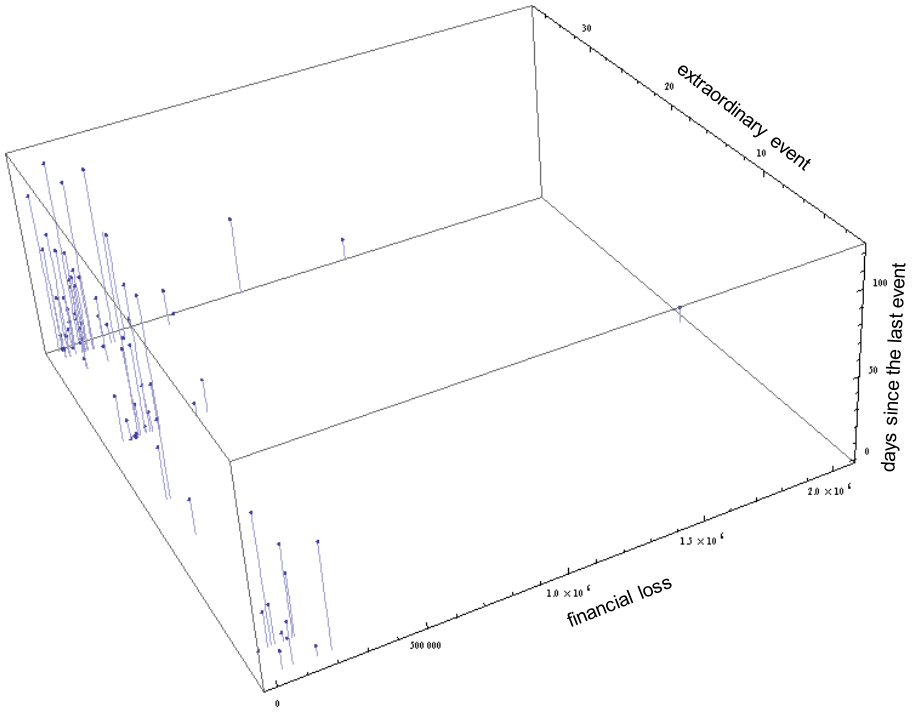

Fig. 1 *Events in the examined space.*

front view. This view is used in our experiments to establish initial values of the classification algorithm using a visual algorithm. Entering the appropriate coordinates for these initial cluster center position is very important because, in this position, at the beginning of the calculation of the estimate, the program sets the initial location of centers of Gaussian curves and in the case of an incorrect initial setup, algorithm classified incorrectly. Algorithm did not approached the actual center position of the cluster, as calculated parameters for values which are too distant from the center of the Gaussian curves show irrelevant and practically zero values. As a next step, we iteratively examined the measured points one by one and calculated which of the clusters have to be included. At the same time, we moved the centre of the selected Gaussian curve closer to that point. During classification, a correction of coordinates of cluster centers also takes place. The coordinates of each point used to correct the position of the three Gaussian curves in the same way as point would always slide the same way as all three Gaussian curves. We solved this by first counting how "deep" every point is located in every spatial Gaussian curve. We used these values as weights in a calculation (for the most distant bell curve is the smallest value, and vice versa for the next weight value is greatest). During the experiments, we also verified that after estimating the position of the Gaussian bell curves, covariances have to keep a solid line (narrow and constant) and not to assess them. Wider Gaussian had caused a significant distortion of the results.

Following is the initialization and normalization part of the algorithm, during which the initial positions of components are prepared. From the measured data (prior information) it is prepared the initial position of the components at the time of estimating initiation. In the next part of the algorithm, estimates are created. This is an iterative loop that is composed of three parts. Calculation of weights, update of statistics and calculation of the point estimates. In the last part of the algorithm results are processed.

In Fig. 2, development of individual X, Y and Z coordinates of the cluster centres in a standardized threedimensional space is illustrated after an analysis of the measured values. Coordinates of the cluster centre are altered at each step of the algorithm specified until these co-ordinates converge near to the co-ordinates of real centre of the cluster. From Fig. 2, it can be visually verified that the centers of the first and third cluster has been set correctly compared with the coordinates of the second cluster. Center coordinates of a second cluster has also been considerably clarified during an algorithm and the position of the center with each computed value was shifted. As the coordinates of centers, especially in the second cluster, have not been selected properly, it was necessary to better determine the initial coordinates, or the final output coordinates of the first run of the algorithm used as inputs and the algorithm run again with these values. By this method, not only the centers coordinates of the clusters in three-dimensional space were determined, but also the distribution of individual points into clusters. Fig. 3 graphically shows the relevance to the individual clusters.

In Fig. 3, you can clearly recognize the existence of clusters, but in my experiments, we prefer to use the view in the three-dimensional space. Example of threedimensional view is shown in Fig. 4.

In Fig. 4, one of the possible three-dimensional views is illustrated. X axis represents "days since the last event", Y axis type of "extraordinary event" and Z axis "financial loss" of the event. As to reveal two major clusters (shown by crosses and circles) and a minor cluster which is represented by squares. Calculation therefore suggests that in this setting, the analyzed system has two working modes. At the beginning of this chapter, we have performed an analysis using linear regression. Values from "extraordinary event" and "financial loss" were used to predict when the next extraordinary event will occur. The analysis was conducted over the entire defined area. Variables in this entire area behave non-linearly. We also confirmed the existence of clusters and it is clear that the analyzed system has several operating modes. In further research, we analyzed within the identified clusters and using the regression analysis method, we checked, whether these clusters variables behave linearly and whether it is possible to predict their future values. From Fig. 5, it was obvious that the predicted values follow the original values in the first (time units 1–14), as well as in the third cluster (time units 72–77). In the second cluster, it was evident that the predicted data do not follow the original one. For this particular job, the first and third cluster variables proved to behave linearly. Area of the second cluster is to be subjected to a deeper analysis. The reason may be that we did not exactly specify the initiation point of the second cluster, or there might have been multiple clusters.

Due to the appropriate interpretation of the calculated results in practice, we also analyzed the distribution of individual clusters and locations of their centers

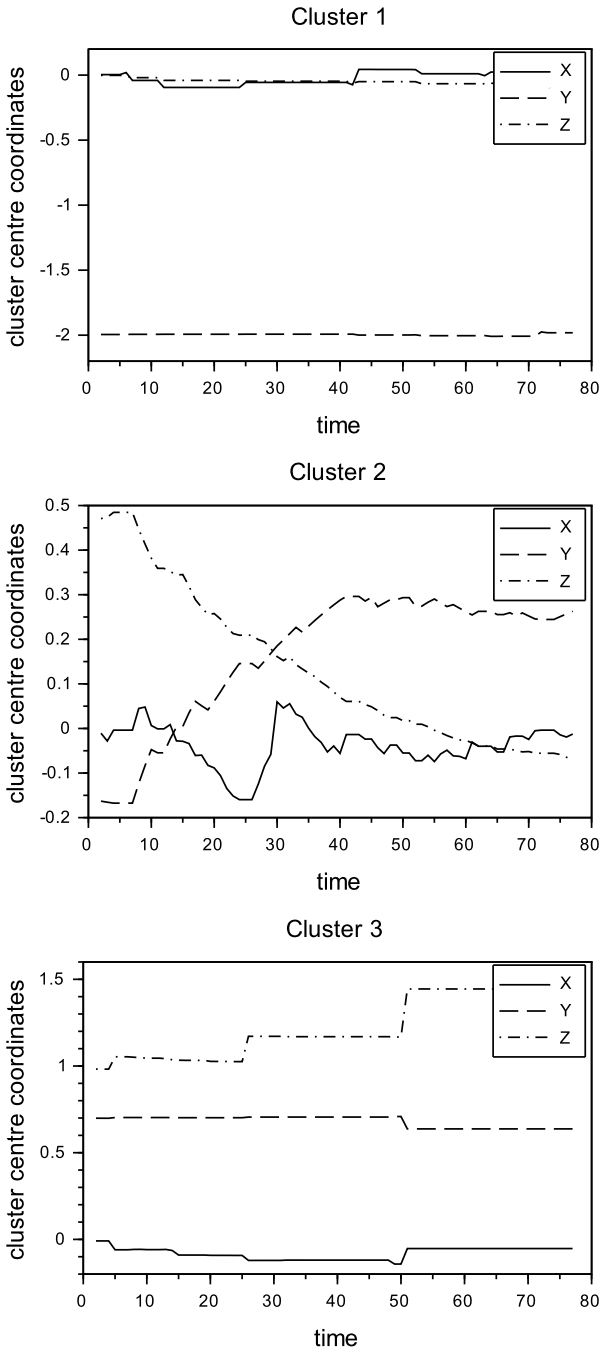

Fig. 2 *Development of cluster centers.*

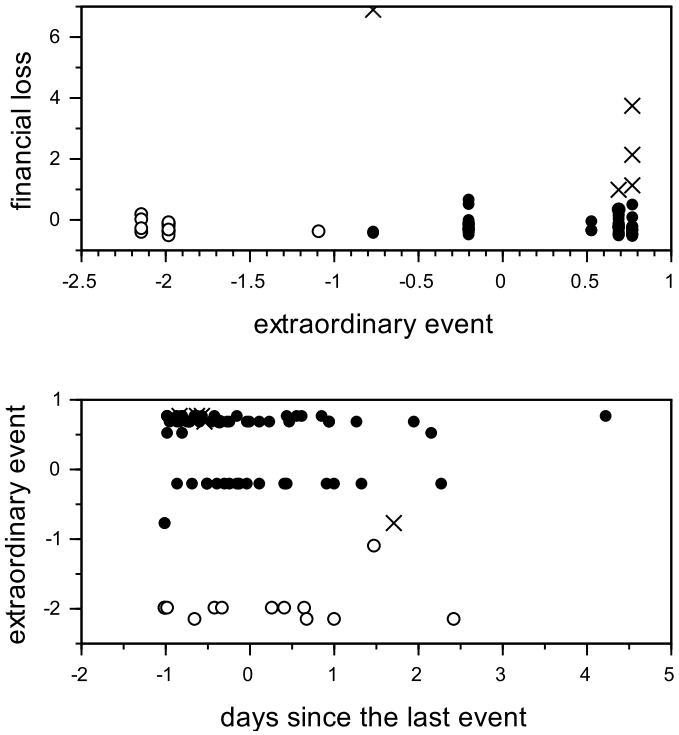

Fig. 3 *Clusters in the two-dimensional space, each single point represents one incident.*

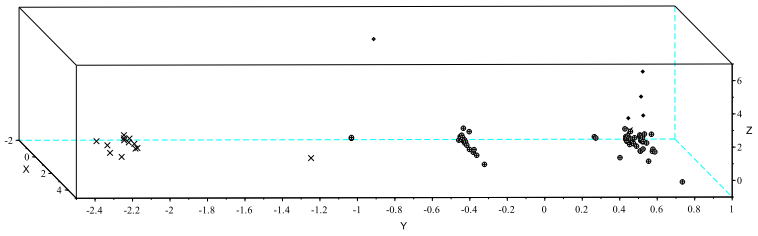

Fig. 4 *Clusters in the three-dimensional space, each single point represents one incident.*

in the original space. Finally, we pointed out, what the mathematical value from a practical point of view represents.

3.2 Experiment evaluation

We searched for a methodology and its practical applications which were verified via experiments. For more detailed information and a large number of experiments documented, see thesis [4]. During these experiments, we sequentially analysed the different variables that define the surveyed space and we interpreted appropri-

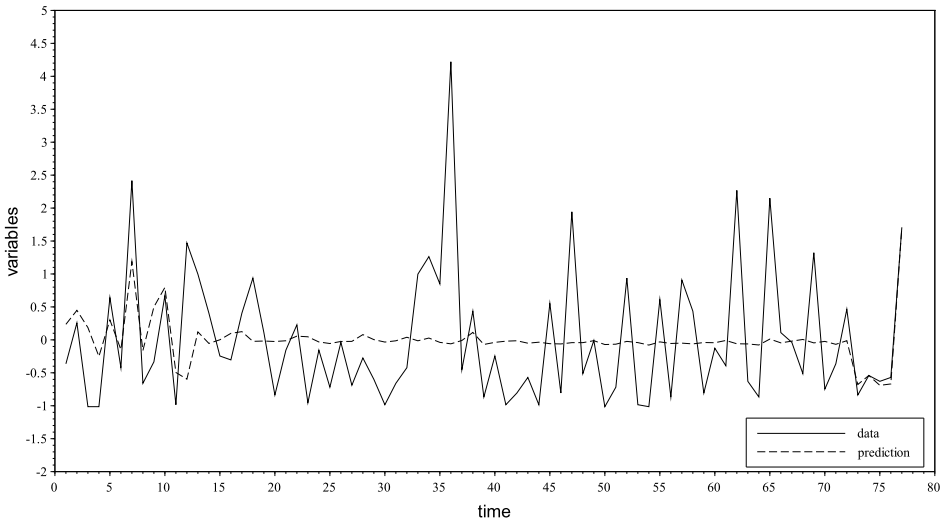

Fig. 5 *Original and predicted values.*

ately the causal events. For quantified experiments conclusions, we then verbally interpreted their impact on the analysed fibre optic infrastructure. On the basis of calculations and experiments, we reached the conclusion that the analysed optical infrastructure outages are most often caused by three types of emergencies namely "weather conditions", "criminal delicts" and "civil engineering work". Based on analysis conducted, the specific circumstances of the cases mentioned could be identified.

Emergencies caused by weather conditions compared to other emergencies have resulted in the lowest financial loss. In this context, we can also say that these kinds of incidents happen less frequently than events caused by construction work and comparable to events caused by criminal delicts. The occurrence of such emergencies is equally distributed throughout the various city districts. Another interesting finding is the fact that the greatest likelihood of emergencies caused by weather conditions are at the turn of the first and second quarters. Hence, in the early spring months around March and April. We also identified that these incidents occur most often at relatively lower temperatures, typically at temperatures of 10–13 °C. We also verified the hypothesis that the incident caused by weather conditions occurs more frequently at higher wind speeds, as we calculated that the highest probability of occurrence is at a wind speed of 6 m/s.

Emergencies caused by criminal delicts compared to other emergencies were seen to become less frequent. Survey results also indicate that the highest probabilities of emergencies caused by criminal delicts are located in city districts with a lower number. In practice, this means that incidents caused by criminal delicts have become more frequent in the city centre. Over the course of the whole year, most of these events occur early in the third quarter, around the summer months

of July and August. In comparison with the events caused by effects of the weather conditions, events caused by criminal delicts become more likely at higher temperatures and at the same time, at lower wind speeds. Particularly in this case, the temperature was around 15 °C and a wind speed of about 4 m/s. The financial impact of such incidents is deemed to be a higher than with outages caused by the weather conditions. Furthermore, the financial losses caused by criminal offenses were close to losses caused by the construction work.

Emergencies caused by civil engineering work in comparison with other emergencies resulted in the highest financial loss. Based on the experiments conducted, we can also confidently say that incidents caused by construction works are happening most frequently. Regarding the location, a greater likelihood of such an event is in downtown Prague, i.e. in urban areas with a lower number. Emergencies caused by construction work are most likely to occur at the end of the third quarter, in the early autumn. The events caused by construction works are happening usually at higher temperatures. Specifically, in this case, it most likely happens at temperatures around 16 °C and wind speed about 5 m/s.

They are also some very interesting findings of other experiments that have been carried out [4]. The outcomes of these experiments are not surprising, but confirm the fact that the proposed methodology processes the data properly and can also be used for cases where the outcome is impossible to predict easily in advance. For example, we have to explore whether it will be necessary to process a large amount of input data or if it is necessary to work with values which are difficult to imagine using common sense. An example might be the measured values of chromatic dispersion in single-mode optical fibres.

For example, during the experiment, we confirmed the hypothesis that emergencies were most likely to happen in winter, often at relatively low temperatures and also in the summer when the temperature is high. In the summer, emergencies also occur more frequently at a lower wind speed than in the beginning and end of the year. Another example is that during the summer months, incidents occur at a higher daily temperature than in the winter months. During the first half of the year, emergencies also occur at a higher wind speeds.

For such a defined task, some of the results were predictable. From experience, we verified that hanging optical links are more susceptible to break during strong winds, which is strongest in the spring when the temperature is usually lower.

Calculated results of experiments according to the proposed methodology were therefore possible to compare with practical experience and qualifications. We see some practical benefits for the telecommunication operator where these findings can be used during planning operating budgets, utilization of individual service teams, planning holidays, etc.. In practice, it is useful to know that using the proposed methodology, we can predict when and under what conditions a particular type of incident is most highly likely to occur and what its financial impact will be.

For example, during the spring months, incidents caused by the weather conditions have less financial impact, and incidents caused by criminal delicts are more frequent during the summer within the city centre, thereby having a greater financial impact. During autumn, incidents, caused by the construction work, are more probable in the city centre, which again have higher financial impact. It is impor-

tant to mention that for general practical use of my experimental conclusions, it is always necessary to take into account the specifics of the selected site. For example, there are a relatively higher percentage of suspension fibre optic cables in the city of Prague, which are more prone to outages caused by wind than cables buried in the ground. Another specific feature of Prague can also be increased construction activities or higher incidence caused by criminals.

4. Conclusions

In our research, we first defined part of the transmission infrastructure of telecommunications operator T-Mobile Czech Republic as. The Data was effectively measured at this fibre infrastructure. We then proposed a methodology to work with these measured values. We have identified that it is not a random system. We have confirmed the existence of clusters that represent the working modes of the system. We have also verified the fact that in some clusters variables, they behave linearly. In such cases, we are in clusters capable of using regression models to predict future values based on past and present values of selected variables. The correctness or viability of the calculations was verified by numerous experiments. Predicted values could then be transformed into the environment of infrastructure of any telecommunications operator to practically use the theoretical results of my experiments. When interpreting the results, it is necessary to take into account the specifics of selected sites and infrastructure. Research findings can be applied in practice, not only in preventing the occurrence of incidents, but also in the planning of operating budgets, the workload of individual service teams, planning holidays, etc. At first sight it may seem that the conclusions of some experiments are not surprising. But this only confirms that the proposed methodology processes the data correctly and it can also be used for those cases where the outcome is not possible to simply estimate in advance. There may be cases where it is necessary to process a large amount of numerical data, or where data are difficult to imagine.

Other benefit of the work conducted can be attributed to the fact that we proposed the methodology through which we can examine the situation in the individual working modes of the system, which can be also examined separately. In some cases, we can confidently proceed straight to the prediction. However, in the area of cluster variables where they do not behave linearly, we would like to conduct further research. In these instances, we would like to commit to some deeper analysis and continue with our future works and experiments.

References

[1] BATES J. *Disaster recovery planning: Networks, telecommunications and data communications*. New York: McGraw-Hill, 1992.

[2] CUESTA-ALBERTOS J.A., MATRAN C., MAYO-IISCAR A. Robust estimation in the normal mixture model based on robust clustering. *Journal of the Royal Statistical Society Series B—Statistical Methodology*. 2008, 70(4), pp. 779–802, doi: 10.1111/j.1467-9868.2008.00657.x.

[3] DACUNHA D., GASSIAT E. The estimation of the order of a mixture model. *Bernoulli*. 1997, 3(3), pp. 279–299, doi: 10.2307/3318593.

[4] DOUDA V. *Modelling of Crisis Management and Business Continuity of Telecommunication Companies*. Prague, 2015. PhD thesis, Czech Technical Univesity in Prague.

[5] KÁRNÝ M., NAGY I., NOVOVIČOVÁ J. Mixed-data multi-modelling for fault detection and isolation. *International Journal of Adaptive Control and Signal Processing*. 2002, 16(1), pp. 61−83, doi: 10.1002/acs.672.

[6] MURRAY-SMITH R., JOHANSEN T. *Multiple Model Approaches to Modelling and Control*. London: Taylor & Francis, 1997.

[7] NAGY I., SUZDALEVA E., KÁRNÝ M., MLYNÁŘOVÁ T. Bayesian Estimation of Dynamic Finite Mixtures. *International Journal of Adaptive Control and Signal Processing*. 2011, 25(9), pp. 765−787, doi: 10.1002/acs.1239.

[8] NAGY I. *Advanced statistical methods and its applications*. Materials for PhD students. Available from: http://nagy.rudolfpohl.cz/Doktorandi/PhDtext.pdf

[9] PROCHÁZKOVÁ D. *Analýza a řízení rizik*. Praha: ČVUT, 2011. In Czech.

[10] SCHOENBERG J.R., CAMPBELL M. Distributed terrain estimation using a mixture-model based algorithm. In: *12th International Conference on Information Fusion*, Seattle, WA. IEEE, 2009, 1–4, pp. 960–967.

[11] STERBENZ J., ÇETINKAYA E., HAMEED M., JABBAR A., QIAN S., ROHRER J. Evaluation of network resilience, survivability, and disruption tolerance: analysis, topology generation, simulation, and experimentation. *Telecommunication Systems*. 2013, 52(2), pp. 705−736, doi: 10.1007/s11235-011-9573-6.

[12] TITTERINGTON D., SMITH A., MAKOV U. Statistical Analysis of Finite Mixtures. New Yourk: John Wiley, 1985.

[13] TOSHIKAZU S., ZUBAIR F., THUAN N., HIROKI N., MASATAKA N., FUMIYUKI A., NEI K., ATSUSHI T., TOMOAKI K., HIROMICHI K., SHIGEKI K. Disaster-resilient networking: a new vision based on movable and deployable resource units. *IEEE Network*. 2013, 27(4), pp. 40−46, doi: 10.1109/MNET.2013.6574664.

[14] WANG H., LUO B., ZHANG Q., WEI S. Estimation for the number of components in a mixture model using stepwise split-and-merge EM algorithm. *Pattern Recognition Letters*. 2004, 25(16), pp. 1799–1809, doi: 10.1016/j.patrec.2004.07.007.

[15] WROBEL L.A., WROBEL S.M. *Disaster recovery planning for communication and critical infrastructure*. Norwood: Artech House, 2009.

[16] XU X., JÄGER J., KRIEGEL H.P. A fast parallel clustering algorithm for large spatial databases. Data Mining and Knowledge Discovery. 1999, 3(3), pp. 263−290, doi: 10.1007/0-306-47011-X_3.

[17] ZHANG T., RAMAKRISHNAN R., LIVNY M. Birch: A new data clustering algorithm and its applications. *Data Mining and Knowledge Discovery*. 1997, 1(2), pp. 141−182, doi: 10.1023/A:1009783824328.

[18] SCILAB. Scilab 5.5.2 [software]. Available from: http://www.scilab.org