Exploratory analysis of the fire statistics using automatic time series decomposition

Dr. (Tech.), prof. M.M. Tatur
The Educational Establishment “Belarusian State University of Informatics and Radioelectronics”
Minsk, Belarus
tatur@bsuir.by

Ph.D. (Tech.), assoc. prof. A.G. Ivanitskiy
State Educational Establishment “University of Civil Protection of the Ministry for Emergency Situations of the Republic of Belarus”, Minsk, Belarus
a.ivanitski@gmail.com

V.M. Prorovsky
The Educational Establishment “Belarusian State University of Informatics and Radioelectronics”
Minsk, Belarus
slawapro@gmail.com

Miroslav Kvassay
Faculty of Management Science and Informatics, University of Zilina
Zilina, Slovakia

2021 International Conference on Information and Digital Technologies (IDT) | 978-1-6654-3692-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/IDT52577.2021.9497578

***Abstract*— The aim of the work was to determine the possibility of using software based on machine learning technologies at the stage of exploratory analysis of data on the situation with fires in settlements, to assess the possibility of using the results obtained earlier by other researchers. Exploratory analysis of data and comparative assessment with results from other sources was carried out.**

***Keywords*— emergency situation, data mining, exploratory analysis, time series, forecasting, fire statistics**

I. INTRODUCTION

Exploratory data analysis is an important step in data mining. The researcher uses various methods of data transformation and visualization, which can be rather time-consuming. Currently, software products based on machine learning technologies are available for these purposes.

The main algorithm of this study is close to the SEMMA methodology (Sample, Explore, Modify, Model, and Assess) [1], which defines the stages of data mining.

The Ministry for Emergency Situations of the Republic of Belarus performs accounting of fires and their consequences, which is the basis for state statistical accounting of fires. A man-made fire is an emergency, defined as an uncontrolled combustion outside a special fireplace, resulting in damage and losses [2]. The above analysis does not cover minor fires (ignitions), which do not entail material damages. The studies in this article mainly contain the result of the application of stages corresponding to stages 1-4 of SEMMA.

II. DATASETS

The initial set of data was obtained from the database of the software package “Accounting for Emergency Situations” [3] and respectively processed (grouped by calendar days for the period from 2011 to 2020 and divided into additional data sets, such as: the total number of fires (“a”), fires in cities and urban areas (“b”), fires in rural settlements (“c”), fires in apartment buildings (“d”), the number of fires in single-family houses, summer cottages, outbuildings, yards and adjacent territories (“e”).

III. ANOMALIES DETECTION

Modification stage included detection of anomalies - identification of rare data, events or observations that are unusual due to their significant difference from the rest of the data. As a rule, such anomalous data indicate an existing problem [4]. Despite the large number of studies on the detection of anomalies, not all methods are applicable to fire data due to the peculiarities of the seasonal and trend components. Therefore, the S-ESD (Seasonal Extreme Studentized Deviate) methods were used for automatic detection of data anomalies [5]. Statistical learning methods are used to detect anomalies. Decomposition by seasons is applied to filter trends and seasonal components of a time series and then robust median and median absolute deviation (MAD) statistics is considered to accurately detect anomalies and distinguish them from seasonal outliers. To exclude anomalous values, the outlier is removed or replaced with the nearest neighbors.

Reviewing anomalies in the course of data mining can lead to beneficial and previously unknown findings. For example, on July 17, 2016, an abnormal number of fires untypical for this time of the year was recorded - 41 cases (usually about 20). A detailed study proved that 23 of them were due to lightning strikes, and strong winds made the conditions even more complicated (strong wind is a registered meteorological emergency, the wind passes over the territory of three regions). A few days before the weather has changed to hot and dry, the air temperature reached 30 °C. Therefore, a rare simultaneous impact of several dangerous meteorological phenomena has led to a sharp increase in the number of fires.

Data visualization during preliminary study (Fig. 1) enabled researches to put forward a hypothesis about the signs of probable time series (trend and seasonal fluctuations).

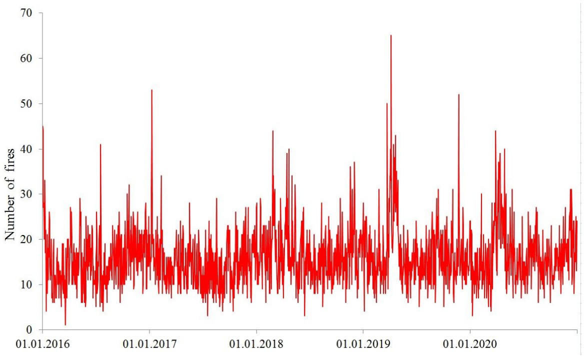

Figure 1. Five year fires chart

A time series is a sequence of observations, usually arranged by time. The main feature that distinguishes time series analysis from other types of statistical analysis is the importance of observations order. Time series analysis allows revealing the hidden patterns of a dataset, and forecasting methods provide information about the possible value of the studied indicator in the future. The main task of time series analysis, as a rule, is to define a model that describes the structure of the time series and can be further used for forecasting.

IV. Prophet features

Classical methods of time series forecasting based on statistical models require additional costs to engage the experts in forecasting, who should configure the model and adjust the parameters of the applied methods depending on the specific problem area. Customizing these methods requires a deep understanding of time series models and their operation. Many organizations can’t afford employing data scientists, and in most cases they lack resources for building complex forecasting platforms.

The choice of Prophet software framework as a forecasting tool is based on the studies of its estimated forecasting accuracy compared to other models [6, 7] (the current review does not cover the comparison with other models).

Prophet is designed to forecast the data using time series methods based on an additive model, in which non-linear trends correspond to yearly, monthly, weekly or daily seasonal fluctuations. It allows tracing of the holidays effect and other special days. The best forecasting results are achieved with time series that have strong seasonal effects and several seasons of historical data. Prophet is an open source software [8] developed by Facebook Core Data Science. The optimal parameters for the models are selected using machine learning methods expressed in the Stan probabilistic programming language, so the forecast can be obtained within a few seconds and in a fully automatic mode for disorganized data. Prophet is resistant to outliers, missing data and abrupt changes in time series, and includes many options for customizing and adjusting the forecasts based on available to user parameters for a specific area. The methodology is described in detail in [6]. It is based on the procedure for fitting additive (Generalized Additive Models, GAM) regression models of the following type:

$$y_t = g_t + s_t + h_t + e_t , \qquad (1)$$

where g_t and s_t are functions that approximate the trend of the series and seasonality (for example, yearly, monthly, weekly, etc.), h_t is a function reflecting the effects of holidays and other influencing events, e_t is normally distributed random disturbances.

The following methods are used to approximate the listed functions:

trend: piecewise linear regression or piecewise logistic growth curve;

yearly seasonality: partial sums of the Fourier series, ; the number of terms (order) determines the smoothness of the function;

weekly seasonality: presented as an indicator variable;

“holidays”: (public holidays and weekends - New Years, Christmas, etc., as well as other days when the properties of the time series can change significantly - sports or cultural events, natural phenomena, etc.) presented as indicator variables.

Estimation of the model’s parameters is performed using the principles of Bayesian statistics (either by the maximum a posteriori (MAP), or by Bayesian inference) [9].

V. Analysis of decomposed data

The performance of the model was evaluated using experimental software in the Python programming language, which provides loading and processing of initial data, setting up and training the model, building a forecast, visualizing the initial and resulting data, cross-validation and calculation of metrics, saving the results.

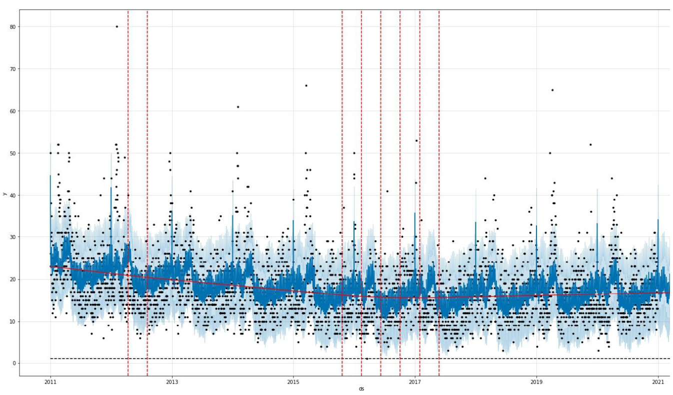

Figure 2. Extracting trend and chaingepoints

Let's train the model on the set “a” and draw the conclusion of the main components of the expanded time series.

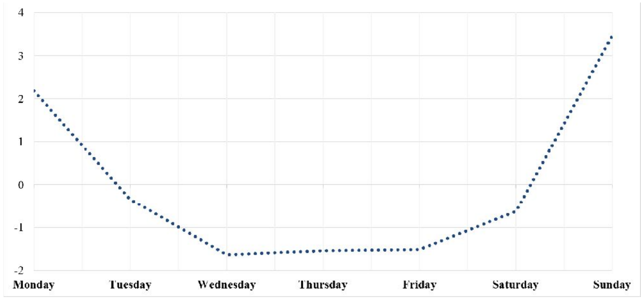

Figure 3. Weekly component

As is follows from the diagram (Fig. 2), the trend line (continuous red line) of the fire situation has several changepoints (vertical dashed lines). A slight fluctuation was noted in 2012. A total change in direction took place in 2016-2017. The similar section of the growing trend over the last five-year period in Russia is presented in [7], and suggests the presence of a common influencing factor for both states. In this case, it is possible to consider following hypotheses:

the consequences of the currency crisis in Russia in 2014-2015, which was caused by a rapid decline in world oil prices and led to a sharp weakening of the Russian ruble against foreign currencies, rise of inflation, a decrease in consumer demand, an economic recession, an increase in poverty and a decrease in real income of the population [10]. For example, in [11], it was suggested that the fire situation grew worse in connection with an economic recession and decline in the living standard of people;

changes due to climatic conditions. Studies of the influence of climatic conditions on the fire situation in Russia are given in [12].

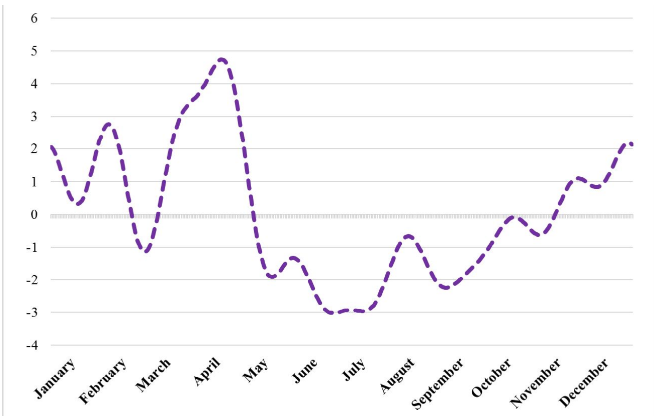

Figure 4. Yearly component

Assessment of the influence of days of the week on the frequency of fires (Fig. 3) is fully consistent with previously published data [12], [13]. There are more fires on weekends than on weekdays. Considering the fact that approximately 80% of fires occur in the residential sector [12], it seems reasonable to assume that there is a positive correlation between the time of stay at home and the frequency of fires. The model has correctly calculated and reflected this dependence.

The yearly component review (Fig. 4) provides a lot of additional information. The presence of a burst in April-May is a priori associated with a large number of dry vegetation fires within that period and is confirmed by other sources [7, 12]. At the same time, it is interesting to identify the features and conditions that have led to this burst. For this purpose we can use two additional data sets of time series “b” and “c”, which we obtained by grouping the initial data according to the type of settlement in which the fire occurred.

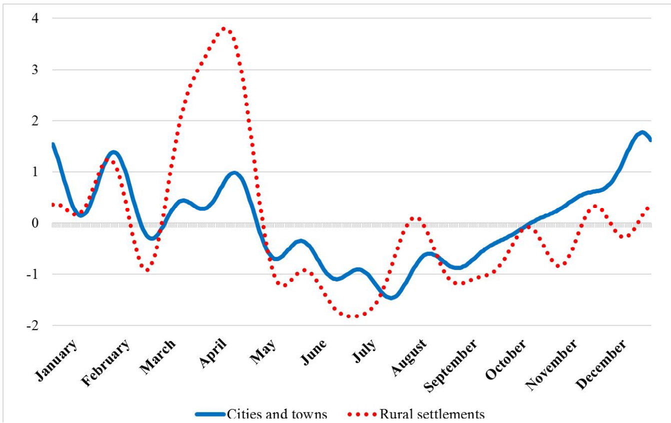

Figure 5. Yearly component by type of settlement

The spring avalanche-like increase in the number of fires (Fig. 5) was due to fires in rural settlements, enables us to continue research on the level of objects on where the fire occurred. We use the data sets “d” and “e”.

The yearly component reflecting the data for multi-apartment (including multi-storey) buildings has a shape close to a straight line, and minor fluctuations are presumably associated with fires in one- and two-story buildings (Fig. 6). For an in-depth analysis, we could consider such parameters as the number of residential buildings of each type, as well and their residents.

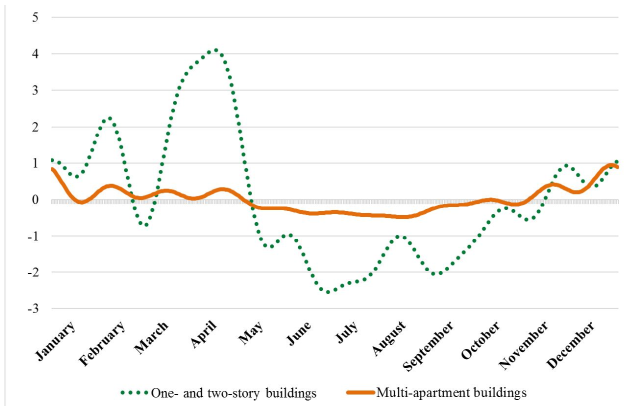

Figure 6. Yearly component by buildings type

978-1-6654-3692-2/21/$31.00 ©2021 IEEE

Let us review the diagram of influence (Fig. 7) of national holidays in Belarus, “standardly” used by the framework and the assessment of days based on statistical data. The highest score falls on January 1, the rest of the values are distributed almost evenly. Obviously, a more detailed study of this specificity is required.

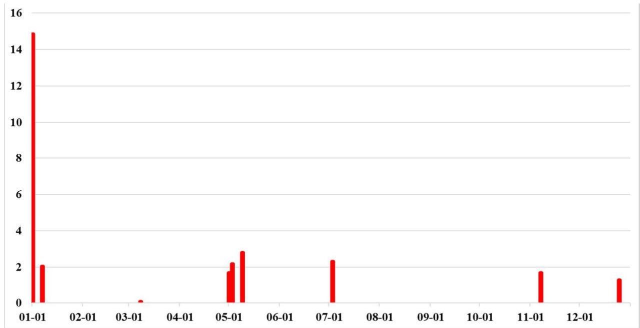

Figure 7. Holiday influence component

Taking into account the fact that the model supports a more accurate procedure for entering the information about the days on which social activity changes, it is obvious that the forecasting accuracy can be improved by setting the length of periods for consecutive holidays, as well as adding information about other anomalous dates, which include the shift of working.

Conclusions

1. The use of Prophet framework with additive models has enabled us perform exploratory data analysis, visualize the trend and its changepoints, weekly and yearly seasonal components, and determine the effect of holidays preset in the model by default on the fire situation.

2. Using the decomposition of the time series into independent separate series that reflect the aggregation of individual section of indicators, we reviewed the effect of avalanche-like increase in the number of fires in April and May.

3. Hypotheses of the reasons for the change in the direction of the trend line in 2016-2017 have been identified for further investigation.

4. A hypothesis was proposed that the model can be strengthened by improving the quality of the initial data on holidays, the total duration of weekend periods and adding other anomalous dates to the set.

5. In the process of detecting anomalies, previously undefined dependencies were revealed - an avalanche-like burst of fires under certain hazardous meteorological conditions.

Acknowledgment

This work was partially supported by the Slovak Research and Development Agency under the contract No. SK-FR-2019-0003 "Mathematical Models based on Boolean and Multiple-Valued Logics in Risk and Safety Analysis".

References

[1] Azevedo, A. KDD, SEMMA and CRISP-DM: A parallel overview [Electronic resource] / A. Azevedo, M. Santos // IADIS Multi Conference on Computer Science and Information Systems, Amsterdam, 22-27 July 2008 / Intern. Assoc. for Development if the Inform. Soc.; Associate Ed.: Luís Rodrigues and Patrícia Barbosa.– Amsterdam, 2008. – P. 182-185.

[2] On the Approval of the State Statistical Reporting Form 1-os (fires) «Report on Fires (Except Forest Fires) and the Consequences of them» and Instructions for Filling it out : Regulation of the Nat. Statistic Comm. of the Republic of Belarus, 27th June, 2017 № 49 // Nat. juridical Internet-portal of the Republic of Belarus [Electronic resource]. – Minsk, 2021 – Mode of access: https://pravo.by/upload/docs/op/T21703807p_1501016400.pdf. – Date of access: 20.04.2021.

[3] Develop a Software Package for Collecting and Analyzing Information about Emergency Situations and their Consequences : sci. rep. (final) / Sci.-Res. Inst. of Fire Safety and Emergencies of the MES of the Republic of Belarus; V.M. Prorovsky, M.V. Hodin, N.D. Chistyakov, T.A. Kornacheva, O.E. Kozlova. – Minsk, 2017. – 54 p. – № ГР 20163551. – Deposited in BelISA 04.07.2018, № Д201828.

[4] Zimek, A. Outlier Detection / A. Zimek, E. Schubert // Encyclopedia of Database Systems. Living Edition / Springer. – New York, 2018. – P. 40. – DOI:10.1007/978-1-4899-7993-3_80719-1

[5] Hochenbaum, J. Automatic Anomaly Detection in the Cloud Via Statistical Learning [Electronic resource] / J. Hochenbaum, O.S. Vallis, A. Kejariwal. – 2017. – Mode of access : https://arxiv.org/pdf/1704.07706.pdf. – Date of access : 20.04.2021.

[6] Taylor, S.J. Forecasting at Scale [Electronic resource] / S.J. Taylor, B. Letham // The American Statistician. – 2018. – Vol. 72, № 1. – P. 37-45.

[7] Babenyshev, S.V. Time Series Forecasting Based on Machine Learning Methods for Ensuring Natural and Technosphere Safety / S.V. Babenyshev, O.S. Malyutin, E.N. Materov // Siberian Fire-Rescue Bull.– 2021. – Vol. 1 (20). – P. 75-83. DOI: 10.34987/vestnik.sibpsa.2021.20.1.013.

[8] Prophet: Automatic Forecasting Procedure [Electronic resource]. – GitHub, Inc., 2021. – Mode of access : https://github.com/facebook/prophet. – Date of access : 20.04.2021 (MIT License Copyright (c) Facebook, Inc. and its affiliates).

[9] Mastitsky, S.E. Time series analysis with R [Electronic resource] / S.E. Mastitsky. – 2020. – Mode of access: https://ranalytics.github.io/tsa-with-r/. – Date of access: 20.04.2021.

[10] Vinokurov, M.A. Economic Crisis in Russia 2014-2015: Reasons, Consequences and Ways Out of the Crisis / M.A. Vinokurov // Socio-economic and Legal Region Safety Problem: Materials of the Intern. Sci.-Res. Conf., Irkutsk, 19-21 Febr. 2015 / Baikal State Univ. of Econ. and Law. – Irkutsk, 2015. – P. 65-71.

[11] Fire Safety and Up-to-Date Directions for its Improvement / E.A. Serebrennikov, A.P. Chupriyan, N.P. Kopylov; Ed. U.L. Vorob’ev. – Moscow : VNIIPO, 2004. – 187 p.

[12] Tatur, M.M. On Forecasting Fire Situation Related to Technogenic Emergencies in the Republic of Belarus: Approaches and Problems / M.M. Tatur, A.G. Ivanitskiy, V.M. Prorovskiy // Bull. of the Univ. of Civil Protection of the MES of the Republic of Belarus. – 2020. – Vol. 4. – № 3. – P. 237-250. – DOI 10.33408/2519-237X.2020.4-3.237.

[13] Prorovskiy, V.M. Regularity of the Spatio-Temporal Allocation of the Fires on the settlements in the Republic of Belarus / V. M. Prorovskiy, M.V. Hodin // Actual Problems of Fire Safety : Collected thesis of the XXX Intern. Sci.-Res. Conf., Noginsk, 06-08th June 2018. – Noginsk: VNIIPO, 2018. – P. 108-110.