Data Mining of Network Events With Space-Time Cube Application

Viktor Putrenko[1], Nataliia Pashynska[2], and Sergii Nazarenko[1]

[1]*World Data Center for Geoinformatics and Sustainable Development, Igor Sikorsky Kyiv Polytechnic Institute, Ukraine*
[2]*Taras Shevchenko National University of Kyiv, Ukraine*

At the moment, there is a problem of studying the detailed relationship of social phenomena and their attachment to the territory. The simple means of conducting such research is getting harder due to the large amount of data that needs to be processed. It is suggested to use the methodology of space-time cube construction as one of the types of data mining with spatial-temporal distribution. The use of this method on the example of information analysis from subscribers of one of the major mobile operator networks allows to carry out statistical analysis and to detect statistically significant spatio-temporal clusters in the data that can be used during data structuring in order to provide safety and react quickly to hazardous situations.

Keywords: data mining; spatial-temporal cube; national security

1 Introduction

Starting from 2013, Ukraine's national security becomes more and more important. Every day, citizens of Ukraine meet with the most diverse threats of natural, technological, social, and military character. Dangerous processes, extreme events, catastrophes, virtual, real terrorism, etc. – these are exactly the things to which public authorities need to react practically every day.

The modern world is extremely rich in the most diverse information, the vast arrays of which people collect, store, analyze, and on the basis of it tries to make forecasts and predictions. This is especially true for information that directly and indirectly affects human security. Equally important information for government bodies, whether public or private institutions, the correct analysis of which allows you to take a step in the right (progressive) direction. But operating with big data are very problematic.

When it comes to eliminating the consequences of a certain disaster, where the bill goes, sometimes, for a minute, is needed a quick and balanced solution. For example, international statistics show that the number of rescued after the earthquake directly depends on the beginning of rescue operations. If the saviours arrive in the earthquake zone in the first three hours, they can save up to 90% of the survivors, after six hours only 50%. Only the means of rapid response can reduce the number of victims by 20–30% (ESRI, 2017a; Rak et al., 2017). That is why it is important to get information and to give an answer immediately after an emergency and not only after a while and, moreover, to prevent an emergency in advance, which can lead to a significant number of victims.

V Putrenko, N Pashynska, and S Nazarenko (2018): *Data Mining of Network Events With Space-Time Cube Application*. In: R Westerholt, F-B Mocnik, and A Zipf (eds.), Proceedings of the 1st Workshop on Platial Analysis (PLATIAL'18), pp. 75–82

https://doi.org/10.5281/zenodo.1472753

PLATIAL'18 First Workshop on Platial Analysis (PLATIAL'18)
Heidelberg, Germany; 20–21 September 2018

2 Background

Recently, security issues become increasingly significant due to the increasing number of threats to ordinary people and the region or the country as a whole. One of the options for solving this issue is to study, analyze, and forecast the event by building a spatial-temporal cube. For the first time, the use of the space-time cube was proposed by Hägerstrand (1970) in the early 70's, whose possibilities he described in his work "What about people in regional science?". Despite the active development of geographic information systems (GIS), its use was limited. Only in the 2000s there are works on the use of the spatial-temporal cube in GIS. In the works of this period, new possibilities for using the spatial-temporal cube were presented using GIS, including earthquake surveys (Andrienko and Andrienko, 1999; Andrienko et al., 2003; Gatalsky et al., 2004; Kraak, 2003).

The next steps in using the spatial-temporal cube method were its application in the intelligent analysis of data of a variety of nature: crime analysis, infrastructure studies, animal behaviour analysis, human motion visualization, dependence studies on weather conditions changes over time, etc. (Baas, 2013; Cheng and Adepeju, 2013; Gonçalves et al., 2014; Hurcilava et al., 2013; Shapiro and Hall, 2017; Yusof et al., 2014). In the field of data mining, Ukraine is widely known for the Institute for Applied System Analysis NTUU "KPI", World Data Center for Geoinformatics and Sustainable Development (Petrenko, 2008; Putrenko, 2017; Putrenko and Pashynska, 2017; Zgurovsky et al., 2013; Zgurovsky and Pankratova, 2005).

3 Goal and Tasks

The goal of the work is to analyze the spatial-temporal regularities in the distribution of events in the Vodafone network based on the use of the methodology of space-time cube construction. Among the tasks are to study the methodology of using the space-time cube for the data mining of spatial-temporal data; the study of the application peculiarities of the space-time cube construction method for the analysis of space-time series of data generated by users of Vodafone telecommunication network; the use of building space-time cube for distribution analysis of spatial and temporal patterns of mobile data for the purpose of emergency response to natural and social emergencies.

4 Spatial-Time Cube

Spatial-time cubes are a three-dimensional visualization technology designed to simultaneously represent spatial and temporal characteristics of motion. Accordingly, trajectory points are displayed in three-dimensional space, where the vertical axis usually expresses time (Peuquet, 1994).

In the early 70's Hägerstrand (1970) proposed the use of a graphical approach to reflecting time as an addition to spatial measurements. He developed a three-dimensional diagram as a spatio-temporal cube, which allows to visually explore space-time events and interactions of processes. The cube's base reflects a flat geographic dimension, and the cube's height is time. Initially, the tool was designed to manually reproduce graphics.

The use of the space-time cube requires spatial and temporal data, for the purpose of analyzing certain events. Examples of such events include earthquakes, road accidents, cases of disease, and the observation of rare animals (Gatalsky et al., 2004). Hägerstrand proposed to apply the space-time cube to the data on the motion of objects on the changes of spatial sites with an anchor to time. In this paper, the authors propose to apply the concept of Hägerstrand to another type of data, namely, to analyze network events.

Using the spatial-temporal cube makes it possible to answer three questions of Puke (Andrienko et al., 2003), supplied to spatial-temporal data (ESRI, 2017b):

- "what, depending on when and where": description of objects or a set of objects that are present in a certain place or a set of locations for a certain time or time interval;
- "where, depending on when and what": a description of the location or set of locations occupied by a particular object or set of objects at a specific time or time interval; and

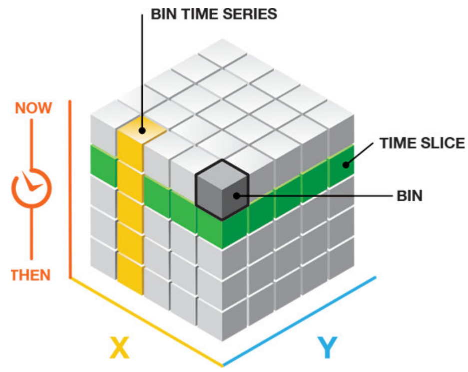

Figure 1: Bin display in the space-time cube (Source: ESRI ArcMap ESRI 2017b)

- "when, depending on where and what": a description of a specific time or interval of time when a particular object or set of objects occupied a particular place or set of locations.

The research uses the data provided by Vodafone, which has spatial and temporal bindings as well as some attribute information. The processed database has 1.5 million network events from the most diverse devices and from different subscribers. All events are concentrated practically in the western regions of Ukraine and has geographic coordinate system WGS84. The location coordinates are recorded in decimal degrees and determine the location of the nearest mobile communication station through which the message or call was sent. All data has been collected between July and August 2017. The authors use a set of tools for in-depth analysis of spatial and temporal regularities in the software ArcMap 10.5. This toolkit contains tools for analyzing data distribution and identifying patterns in the context of space-time.

The dataset structure has a combined set of attributes that characterize the nature of the communication event, location, feature of calls and devices, as well as subscriber preferences. The description of network events is the event type, which is divided into incoming-outgoing calls, SMS, and Internet usage. The location is described by the direction and coordinates of the signal receiving station. The peculiarities of network events include the tariff plan, the category of numbers, the amount of Internet traffic, the cost of use, and the type of device. Personal preferences of the client are presented in the form of three attributes describing the interests located in the first, second, and third place for the subscriber. Examples of such preferences are the categories of science, culture, tourism, travel, football, etc.

The creation of a spatio-temporal cube takes place by arranging point data of events in space and time in the form of a cubic structure. The base unit of the cube is the bin of space-time (Figure 1), which counts the number of points in time and each location using the Mann-Kendall statistics (ESRI, 2017b).

The spatial-temporal cube consists of rows, columns, and time steps, which together form the total number of bins in the cube. The rows and columns correspond to the placement of objects in the latitude and longitude plane. The cube height corresponds to the time period. If an event occurred for a certain period of time, it will be fixed in a certain bin with spatial-temporal characteristics.

5 Mann-Kendall Test

As input objects there can be only point classes that describe the events that have taken place. Such events may include network events, emergencies, trade operations, and other events that are time-consuming and space-based. In order to obtain valid data of distances calculations, rectangular coordinate systems with corresponding projections are used.

An important part of the operation of the tool is the analytical operations over the data bins used during the simulation. The basic set of operations is the definition of the general trend of data, which is calculated on the basis of time series. Using trend analysis allows to determine the positive or negative trends in the number of events. The trend analysis is based on Mann-Kendall's statistics.

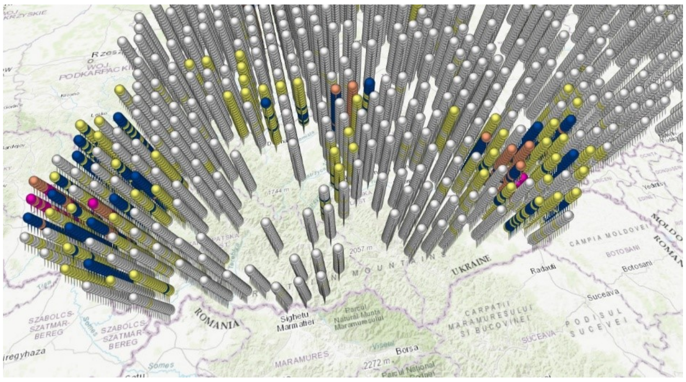

Figure 2: Three-dimensional visualization of space-time cube for western regions of Ukraine

The non-parametric Mann-Kendall test is commonly employed to detect monotonic trends in series of data. The null hypothesis H_0 is that the data come from a population with independent realizations and are identically distributed. The alternative hypothesis H_A is that the data follow a monotonic trend. The Mann-Kendall test statistic is calculated by

$$S = \sum_{k=1}^{n-1} \sum_{j=k+1}^{n} \operatorname{sgn}(X_j - X_k) \qquad \text{where} \qquad \operatorname{sgn}(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0 \end{cases}.$$

The mean of S is $E[S] = 0$ and the variance σ^2 is given by

$$\sigma^2 = \frac{1}{18}\left(n(n-1)(2n+5) - \sum_{j=1}^{p} t_j(t_j-1)(2t_j+5)\right)$$

where p is the number of the tied groups in the data set and t_j is the number of data points in the j-th tied group. The statistic S is approximately normal distributed provided that the following Z-transformation is employed:

$$Z = \begin{cases} \frac{S-1}{\sigma} & \text{if } S > 0 \\ 0 & \text{if } S = 0 \\ \frac{S+1}{\sigma} & \text{if } S < 0 \end{cases}$$

The statistic S is closely related to Kendall's τ as given by

$$\tau = \frac{S}{D} \qquad \text{where} \qquad D = \left(\frac{1}{2}n(n-1) - \frac{1}{2}\sum_{j=1}^{p} t_j(t_j-1)\right)^{1/2} \left(\frac{1}{2}n(n-1)\right)^{1/2}.$$

The resulting Vodafone data set is in the time period from June 1, 2017 to August 31, 2017. For the convenience of analysis, the authors used a 5-day time step. As a result, the tool built a cube with a height of 19 bins (Figure 2).

6 Hot Spots Analysis

Tool analysis of hot spots determines trends in the cluster of density of points (calculations) or fields of sums in a cube. The categories of cold and hot spots include the following characteristics (Anselin, 1995): new, consistent, growing, constant, declining, sporadic, and fluctuating historical (Figure 3).

The Hot Spot method calculates a statistic for each event in the data set. The final values of p (probability) and z-estimates (standard deviations) indicate in which region of the space clustered events with high or low values exist (Andrienko et al., 2003). The method analyzes each event in the context of the neighbouring geography of events. To be a statistically significant hot spot, the event

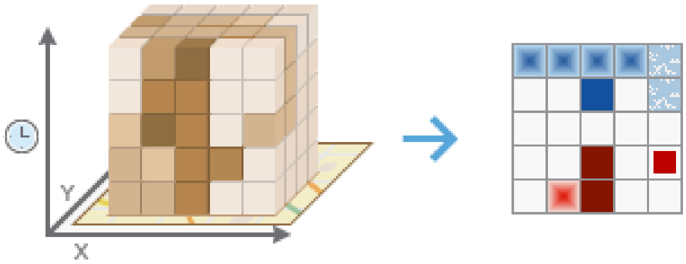

Figure 3: Transformation of a space-time cube for analysis through hot spots

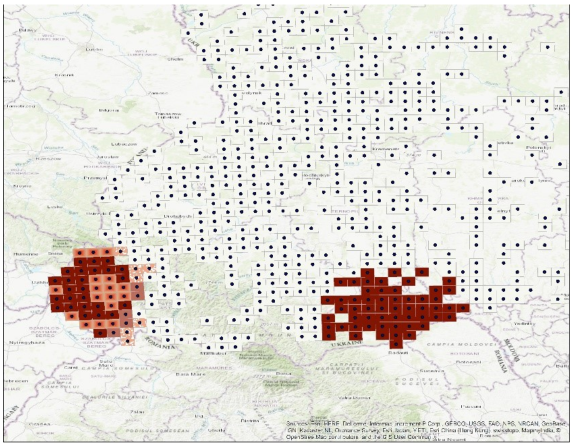

Figure 4: Map occurrence of hot spots and the resulting space-time analysis cube

must have a high value and needs to be surrounded by other approaches with also high values. Hot dots statistics uses the following formulae:

$$G_j^* = \frac{\sum_{j=1}^{n} w_{i,j} x_j - \bar{X} \sum_{j=1}^{n} w_{i,j}}{S \cdot \sqrt{\frac{n \sum_{j=1}^{n} w_{i,j}^2 - \left(\sum_{j=1}^{n} w_{i,j}\right)^2}{n-1}}} \quad \text{where} \quad \bar{X} = \frac{1}{n} \sum_{j=1}^{n} x_j \quad \text{and} \quad S = \sqrt{\frac{1}{n} \sum_{j=1}^{n} x_j^2 - \bar{x}^2}.$$

Here x_j is the attributive value for the event j; $w_{i,j}$ is the spatial weight between the events i and j; n is the total number of events; X is the mean of the arithmetic values of the course; and S is the dispersion.

The statistical value G_i gives each object in the set its own z-score. If the z-score has a positive value, then the probability of the intensity of the clustering of hot spots increases, which is proportional to the size of the positive estimate. Negative z-values are directly proportional to the intensity of clustering of low values and correspond to cold points.

Output objects are added to the table of contents and represent a summary of the spatial-temporal analysis for all the analyzed locations. In addition to creating a class of output objects, the summary analysis results are recorded in the results window (Figure 4).

7 Local Outlier Analysis

The analysis tools group includes the local outlier analysis tool, which allows to identify significant statistical data in both space and time. To determine statistically significant data outliers, the Anselin's Local Moran I statistics is the used statistic option, which calculates the value of each bin relative to its neighbours (Figure 5).

Based on the calculation of z and p-values of Anselin's Local Moran I statistics, each time series receives the coded value of belonging to a particular cluster with the corresponding statistical value.

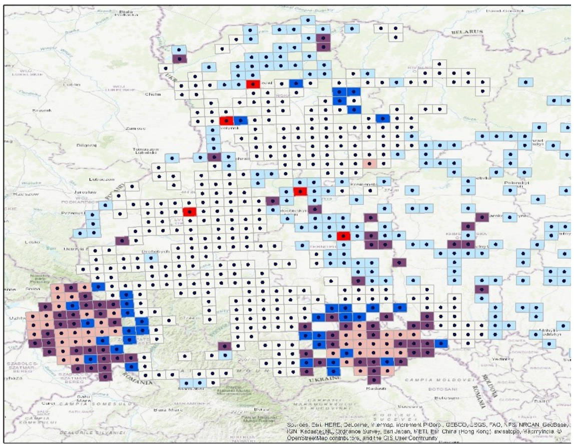

Figure 5: Map of local outliers, created as a result of the analysis of the space-time cube

The Local Moran's I statistic of spatial association is given as

$$I_i = \frac{x_i - \bar{X}}{S_i^2} \cdot \sum_{j=1, j\neq i}^{n} \omega_{i,j}(x_j - \bar{X})$$

where x_j is an attribute for feature i; $\bar{X}$ is the mean of the corresponding attribute; $\omega_{i,j}$ is the spatial weight between feature i and j; n the total number of features; and

$$S_i^2 = \frac{1}{n-1} \sum_{j=1, j\neq i}^{n} (x_j - \bar{X})^2.$$

The Z_{I_i}-score for the statistics are computed as

$$Z_{I_i} = \frac{I_i - E[I_i]}{\sqrt{V[I_i]}} \qquad \text{where} \qquad E[I_i] = \frac{1}{n-1} \sum_{j=1, j\neq i}^{n} \omega_{i,j} \qquad \text{and} \qquad V[I_i] = E[I_i^2] - E[I_i]^2$$

The presence of positive evaluations for I is a certificate that is adjacent to objects with similar values that may be part of a cluster. Negative values indicate the difference between the estimates of the object and its neighbours. In all cases, the value of p for the object must be so small such that the cluster is determined to be statistically significant.

To determine the belonging of the bin to the clusters, the rules of the conceptualization of spatial relationships are first defined, which determine the belonging of the bin to one of the clusters. Further, the values of bins are estimated in proximity to the centre of the cluster. Bins with high values of local emissions contain abnormal changes in the behaviour of users, which may have a different nature, both positive and negative. Together with the use of classifiers and social news dissemination channels, they can be identified and transmitted to relevant government agencies and services.

8 Conclusions

The toolkit for building a spatial-temporal cube provides a convenient visual interface for data mining of big data. The use of the spatial-temporal cube is practically possible in virtually all areas where it is necessary to analyze the behaviour of objects and events occurring with the change of location in space and time.

An example of the use of spatial-temporal analysis of data for events in mobile networks, e. g., of the Vodafone network, makes it possible to use the data more effectively, primarily for security purposes, which will be useful to governmental organizations for the rapid detection or prevention of dangerous situations, such as terrorism, extraordinary events, catastrophes, etc. In the future, using the spatial-temporal cube based on the data of mobile operators, it is possible to analyze the statistical emissions in the activity of subscribers in calls or connecting people to the Internet with an anchorage of a certain territory, which will allow to identify certain anomalies and respond accordingly.

ORCID

Viktor Putrenko https://orcid.org/0000-0002-0239-9241
Nataliia Pashynska https://orcid.org/0000-0002-0133-688X
Sergii Nazarenko https://orcid.org/0000-0003-3367-5875

References

Andrienko, Gennady and Andrienko, Natalia: *Interactive maps for visual data exploration*. International Journal Geographical Information Science, 13(4), 1999, 355–374. doi: 10.1080/136588199241247

Andrienko, Natalia; Andrienko, Gennady; and Gatalsky, Peter: *Exploratory spatio-temporal visualization: an analytical review*. Journal of Visual Languages and Computing, 14(6), 2003, 503–541. doi: 10.1016/S1045-926X(03)00046-6

Anselin, Luc: *Local indicators of spatial association – LISA*. Geographical Analysis, 27(2), 1995, 93–115

Baas, Maarten: *Space-time cube analysis of animal behaviour*. Master's thesis, Wageningen University, 2013

Cheng, Tao and Adepeju, Monsuru: *Detecting emerging space-time crime patterns by prospective STSS*. Proceedings of the 12th International Conference on GeoComputation, 2013, 281–285

ESRI: *ArcGIS analysis workflows for public safety*. https://community.esri.com/docs/DOC-10345-arcgis-analysis-workflows-for-public-safety, 2017a. Retrieved 6 August 2018

—— *A quick tour of geoprocessing tool references*. http://desktop.arcgis.com/en/arcmap/10.3/tools/, 2017b. Retrieved 6 August 2018

Gatalsky, Peter; Andrienko, Natalia; and Andrienko, Gennady: *Interaction analysis of event data using space-time cube*. Proceedings of the Eighth International Conference on Information Visualization, 2004, 145–152. doi: 10.1109/IV.2004.1320137

Gonçalves, Tiago; Afonso, Ana Paula; and Martins, Bruno: *Visualizing human trajectories: comparing space-time cubes and static maps*. Proceedings of the 28th International BCS Human Computer Interaction Conference (HCI), 2014

Hurcilava, Khurtsilava Kostiantyn; Beleckij, Boris Alexandrovich; and Bespalov, Vyacheslav Petrovich: *On the question of space-time gis and some of their applications*. In: *Proceedings of the Conference on Decision Support Systems, Theory and Practice (DSS)*. 2013, 52–55

Hägerstrand, Torsten: *What about people in regional science?* Papers in Regional Science, 24, 1970, 7–24. doi: 10.1111/j.1435-5597.1970.tb01464.x

Kraak, Menno-Jan: *The space-time cube revisited from a geovisualization perspective*. In: *Proceedings of the 21st International Cartographic Conference (ICC)*. 2003, 1988–1996

Petrenko, Anatoly Ivanovich: *Grid and data mining for intellectual data processing*. System research and information technology, 4, 2008, 97–110

Peuquet, Donna J: *It's about time: a conceptual framework for the representation of temporal dynamics in geographic information systems*. Annals of the Association of American Geographers, 84(3), 1994, 441–461

Putrenko, Viktor: *Data mining of relationship in crowdsourcing projects and social activities of citizens*. Proceedings of the 1st Ukraine Conference on Electrical and Computer Engineering (UKRCON), 2017, 1060–1065. doi: 10.1109/UKRCON.2017.8100413

Putrenko, Viktor and Pashynska, Natalia: *Application of geoinformation modeling tools for intelligent data analysis of fire hazards*. Bulletin of National Technical University "KhPI", Series New Solutions in Modern Technologies, 7, 2017, 156–163

Rak, Jacek; Bay, John; Kotenko, Leonard, Igor Popyack; Skormin, Victor; and Szczypiorski, Krzysztof: *Computer network security*. Proceedings of the 7th International Conference on Mathematical Methods, Models, and Architectures for Computer Network Security (MMM-ACNS), 2017

Shapiro, Ben Rydal and Hall, Rogers P: *Interaction geography in a museum*. Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI), 2017, 2076–2083. doi: 10.1145/3027063.3053146

Yusof, Norhakim; Zurita-Milla, Raul; Kraak, Menno-Jan; and Retsios, Bas: *Mining frequent spatio-temporal patterns in wind speed and direction*. Proceedings of the 17th AGILE Conference on Geo-Information Science, 2014, 143–161. doi: 10.1007/978-3-319-03611-3_9

Zgurovsky, Mikhail Zacharovich; Boldak, Andrey Aleksandrovich; and Yefremov, Kostiantyn Viktorovich: *Intelligent analysis and the systemic adjustment of scientific data in interdisciplinary research*. Cybernetics and Systems Analysis, 49(4), 2013, 541–552. doi: 10.1007/s10559-013-9539-7

Zgurovsky, Mikhail Zacharovich and Pankratova, Natalia Dmytrievna: *System analysis: problems, methodology, applications*. Kyiv, Ukraine: Naukova Dumka, 2005