# Analysis and Mining of Big Spatio-temporal Data

Aleksandra Stojnev[1] and Dragan Stojanović[2]

*Abstract* – **Spatio-temporal data analysis and mining has become a critical component of the Big spatial data strategy for most organizations in various domains. The interdisciplinary nature of spatiotemporal data mining means that its techniques must be developed with awareness of the underlying application domains. During the process of data mining, it is important to adjust datasets to the mining task that is being performed, to take into account different temporal and spatial models while mining, and to validate or discard relationships mined from the data. In this paper we give a brief overview of spatio-temporal data analysis and mining process, as well as its demonstration on telecom and OSM data.**

*Keywords* – **Big spatio-temporal data, Spatio-temporal data mining, Telecom data.**

## I. Introduction

Extensive growth in spatio-temporal data volumes and the rapid development of tools and technologies for managing big data have resulted in emerging of different methods for retrieving non-trivial information and useful knowledge from stored data. The analysis and mining of spatio-temporal data include any formal technique that examines the objects by analyzing their topology, geometric, geographic and temporal characteristics. Recent trends in IT have led to a proliferation of studies whose focus is the areas where mining and analysis of spatial and spatio-temporal data can be of crucial importance [1]. Consequently, as businesses rely on information retrieved by data mining process, a group of disciplines oriented to finding solutions for spatial and spatio-temporal data analysis and mining has emerged. For example, it is beneficial to extract different relations from the volume of communications in given area at specific time, as it can lead to better understanding of human behavior and thus provide valuable information to business analysts. These relations can include identification of popular places in the area, patterns in human movements or correlations between communication volume and popularity of a place or an event [2].

In this paper we introduce one solution for handling big volumes of spatio-temporal data so as to extract specific information. Specifically, with spatially and temporally aggregated telecom data, and data from OpenStreetMap dataset, we manage to identify and visualize popular areas, patterns in human movements and relations between these two. Additionally, we manage to find correlation of popular events and telecom activity.

[1]Aleksandra Stojnev is with the Faculty of Electronic Engineering at University of Niš, Aleksandra Medvedeva 14, 18000 Niš, Serbia, E-mail: aleksandra@elfak.rs

[2]Dragan Stojanović is with the Faculty of Electronic Engineering at University of Niš, Aleksandra Medvedeva 14, 18000 Niš, Serbia, E-mail: dragan.stojanovic@elfak.ni.ac.rs

The paper is organized as follows. The Section II surveys related work in both spatio-temporal data mining, and pattern extraction. Section III gives a brief overview of systems for spatio-temporal data analysis and mining. An example of spatio-temporal mining regarding area of Milan, Italy and November and December, 2013 is presented in Section IV. Section V concludes the paper and outlines direction for further research.

## II. Related Work

Specificity of spatio-temporal domain imposes the need for the integration of both spatial and temporal attributes in the process of data analysis, resulting in the development of specialized techniques and algorithms for spatio-temporal data mining. For example, the authors in [3] provide in-depth description of big spatio-temporal modeling and the analysis of data that have form of time-series with numerical values attached. The accent is on combination of interactive visualization techniques and statistical and machine-learning methods. In the same vein, in [4] another research regarding interactive visual analysis and trajectory events exploration regarding spatial, temporal and event aspect is presented.

Authors in [5] presented a framework for pattern detection using historical data and creation of efficient movement index. Furthermore, in [6], Distance-based Bayesian inference Spatial Association Index and Spatio-logical inference for associative analysis, are presented, that can lead to identification of relations between real-world events and model similar events that are described with highly variable values. Spatio-temporal data clustering is presented in [7], defining herd pattern that can describe development of a herd and behavior of animals in it. A large number of papers addresses traffic prediction problem in order to improve navigation, traffic regulation, urban planning and similar. In [8], a framework that can learn in real-time and predict traffic by mapping current state to the trained model is presented and evaluated. Furthermore, in [9], taxi routes are analyzed, and a model for visual query creation is presented.

There are a large number of published studies that deals with literature review and classification of spatio-temporal data analysis and mining techniques [10-12]. One systematic study of different approaches in spatio-temporal data mining regarding desired results is presented in [13]. In all the studies reviewed here, spatio-temporal data mining is recognized as highly important area of development, both for academic and industrial purposes.

## III. Systems for Analysis and Mining of Big Spatio-Temporal Data

Big spatio-temporal data mining process includes various preprocessing, analysis and postprocessing steps, as well as a

visualization of the identified results. Preprocessing steps include cleaning, integration, transformation, reduction and data protection. The analysis of big spatio-temporal data includes sorting, organizing or grouping large quantities of spatio-temporal data so as to identify the relevant information and knowledge. The main aim of spatial and spatio-temporal analysis is to detect relationships between objects, taking into account the area in which the entities are located, and their spatial relationships: topology, position and distance information. Different approaches in spatio-temporal data mining process can include multidimensional spatio-temporal data analysis, spatio-temporal characterization and topological relationships discovery, mining topological relationship patterns, spatio-temporal neighborhood and association rules, spatio-temporal classification, clustering, trend detection and prediction, outlier analysis, collocation pattern or episode discovery and discovery of movement and cascading patterns [13]. The knowledge obtained in the process of spatio-temporal data analysis and mining often requires additional processing: simplification, evaluation, visualization and documentation. Furthermore, the newly discovered knowledge can be interpreted and incorporated into existing systems, which could lead to potential conflicts with previously induced knowledge [14]. Important step in spatio-temporal data mining process is the efficient visualization of retrieved information, as it can improve general awareness of the targeted domain. Moreover, a large number of data mining techniques involve interactive analysis, which would not be possible if the data are not visualized properly.

Existing tools and libraries for handling spatial and temporal data are being adapted to meet the requirements of spatio-temporal data domain. Moreover, a vast number of dedicated systems for handling big data are emerging. For the purpose of this research we used Apache Spark[1] platform, which is specialized for distributed Big Data processing and analysis, and QGIS[2] for visualization of retrieved results.

Apache Spark is a fast and general cluster computing system for large-scale data processing. It provides high-level APIs in Scala, Java, Python, and R, and an optimized engine that supports general computation graphs for data analysis. It also supports a rich set of higher-level tools including Spark SQL for SQL and DataFrames, MLlib for machine learning, GraphX for graph processing, and Spark Streaming for stream processing. MLlib includes various learning algorithms such as classification, regression, clustering, and collaborative filtering, support for feature extraction, transformation, dimensionality reduction, and selection, and utility functions for linear algebra, statistics, data handling, etc.

QGIS is a multiplatform open source application for visualization, editing and analyzing spatial data. It has support for vector and raster data, and can be easily integrated with other GIS open source packages including PostGIS, GRASS GIS, MapServer and others. Functionality of the package can be extended by plugins that can be written in Python or C++.

[1] http://spark.apache.org/

[2] http://www.qgis.org

# IV. Analysis and Mining of Big Open Telecom and OSM Data

## A. Identified Challenges

In order to give an example of spatio-temporal data analysis and mining, we have identified several challenges that can be addressed using data mining techniques. Firstly, we want to identify popular places in one area, using telecom data. Given the fact that telecom interaction level is proportional with the number of individuals using their devices, we can assume that the areas that have the biggest telecom interaction are the ones that contain most popular places. If we create heat map for telecom activity, and overlay it with main tourist, amusement or business attractions, we can identify the most popular locations among them. Secondly, it is important to detect temporal patterns regarding the popularity of a given location. In such manner we can identify variations in telecom activities that are related to location and time of the day or week. Furthermore, it can be beneficial to identify patterns in human movements during the day, or create correlations with the location popularity and time of the day. In that way we can be able to detect what are the areas where traffic congestions are possible. At least, we want to isolate events that can influence crowd movements in order to create correlations between popular events and telecom activities. This analysis will provide an insight how particular events can impact trajectories of people in the area of interest. However, different datasets can have different spatial and temporal aggregation and in order to address identified challenges, an appropriate method for combining them must be found.

## B. Open Big Data Sets

For the purposes of this paper, we used both authoritative and VGI data. As an example of the former, Telecom Italia Open Big Data dataset is used [15], while Open Street Map[3] (OSM) data are used as an example of the latter.

Among a broad list of initiatives dealing with VGI, OpenStreetMap (OSM) is one of the most promising crowd-sourced projects. The collected spatial data are made publicly available and may thus be used for individual purposes as well. The data themselves are distributed under a license that guarantees freedom of use, but makes it mandatory that all derived data are distributed under the same. For the purpose of this paper, we used part of OSM data often referred to as Points of Interest (POI) data. POI is an object on a map or in a geodataset that occupies a particular point and has tags which describe the feature they represent. List of amenities inside Milan grid area (118,084 records) is retrieved from OSM using OSM plugin for QGIS. Set of POIs tagged with tourist or office tags is extracted from retrieved data.

Telecom Open Big Data is one of the most popular authoritative datasets. The data cover the period of two months (November and December 2013) and the area of Milan and Trent in Italy. Data are divided into several sets

[3] https://www.openstreetmap.org/

containing information about mobile telecommunications, Twitter activity, weather, published news and electricity consumption. The whole set is preprocessed and thus prepared for further use. For the purposes of this study only data pertaining to the telecom activity in Milan is used. Some of the data are aggregated using predefined Milano grid. Milano grid is a square grid (100 rows and 100 columns), which covers an area of Milan. Each cell network has an area of 235 meters. Telecommunications dataset is part of Open Big Data set and it provides information about the telecommunication activity over the city of Milano. The dataset is the result of a computation over the Call Detail Records (CDRs) generated by the Telecom Italia cellular network. CDRs log the user activity for billing purposes and network management. There are many types of CDRs, but for the generation of this dataset considered are those related to the received and sent SMS, incoming and outgoing calls and internet usage. By aggregating the aforementioned records, the level of interaction of the users with the mobile phone network is measured. This dataset has spatial aggregation equal to Milan grid and temporal aggregation in timeslots of ten minutes. It contains 319,896,289 records, each of which includes data for all telecommunication types, for a distinct 10 minute time slot, country code, and grid square.

### *C. Spatio-Temporal Analysis of Open Big Data*

In order to prepare telecom data for further analysis, Apache Spark platform is used. All the data are firstly uploaded to HDFS, and then loaded into DataFrame structure using SparkSQL component. Timestamps are converted to readable date time format. Temporal aggregation is performed on entire dataset, with value of one hour. Null values are ignored. For some challenges, the action itself is not important, but its time and location. For that case, dataset with summary of all actions in grid cell is created, again with two different spatial aggregations: per day and per hour.

First challenge is related to detection of popular places in city center of Milan. Telecommunications heat map is created to show telecom activity for every cell in grid. Fig. 1. shows this map, with darker blue shadows indicating higher activity.

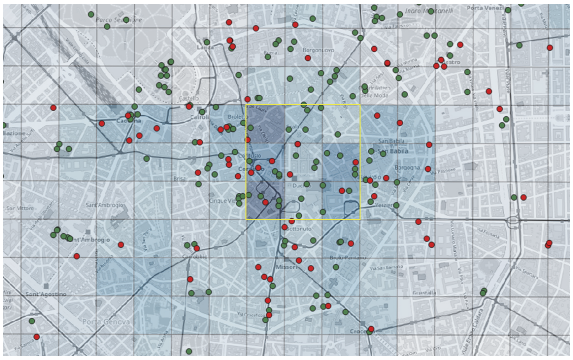

Fig. 1. Popular places

Visualization shows that highest activity level is near Duomo cathedral and the main square in Milan (yellow outline). Secondly, POIs that have tourism (green) or office (red) tags are shown on map. POIs that are located in the outlined area are various museums, hotels and sculptures, which are likely to be very popular among tourists. Second challenge is to find temporal distribution of telecom activity for before identified popular area. This distribution is shown in Fig. 2.

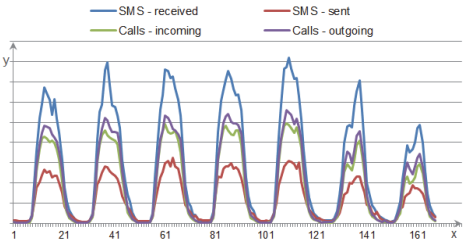


Fig. 2. Visualization of temporal distribution of telecom activity

Hour in week is shown on x-axis, starting from Monday, 00:00, while CDR activity is shown on y-axis. While analyzing the results of the preprocessed dataset, it was found that the user activity in city center, in terms of Telecom interaction, is higher during working days in the week. This can be substantiated by the fact that during the week a number of companies, firms and public institutions perform their daily tasks requiring the use of telecom services. The individual analysis of each working day found that the highest user activity is in the period from 9h in the morning until 19h in the evening. Over the weekend, the greatest activity ranges from 11h in the morning to 23h in the evening. Furthermore, by visualizing spatial distribution of telecom activity at particular time, it is possible to create correlations with the location popularity and time of the day. Visualization of telecom activity during the day is shown in Fig. 3.

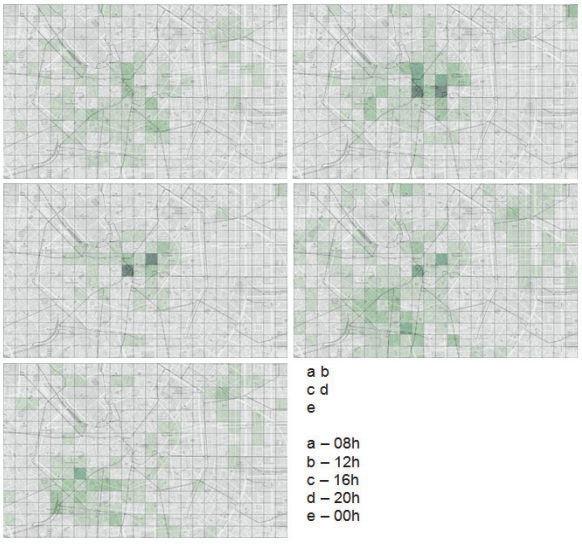


Fig. 3. Visualization of spatial distribution of telecom activity

Five snapshots of spatial distributions are observed, starting from 8h, with four-hour steps (positions a-e). Visual analysis of these snapshots shows that greatest interaction was concentrated in the city core of Milan, at midday. Popular

events can cause aberrations from normal telecom activity. X Factor finale was held on December, 12, in Milan, more precisely in Mediolanum Forum. Fig. 4. shows visualization of telecom activity distribution in that area, for December, 12.

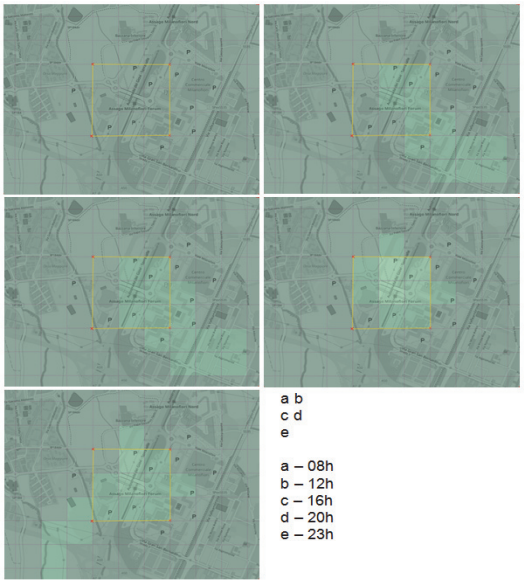


Fig. 4. Visualization of spatial distribution of telecom activity

Presented visualization shows how events can change human behavioral patterns. Lighter areas are the ones that have higher activity. Namely, increased activity in selected area is consistent with the event itself, as it starts at 20h, and peaks in spatial distribution of telecom interaction align both with location and with time of the event.

## V. Conclusion

Extraction of useful information and knowledge from large spatio-temporal datasets is not an easy task. The main aim of this research was to create an overview of all stages of spatio-temporal data mining process. Each data mining task, after acquiring of the data, includes preprocessing, processing and postprocessing steps. The most commonly used methods for every stage in spatio-temporal analysis and mining are listed. Furthermore, the demonstration of processing and analysis of authoritative (Telecom Open Big Data) and VGI data (OSM) is presented. As large amounts of data require Big Data technologies, Apache Spark is used for detection of useful information and mobility patterns, regarding temporal and spatial distribution of user activities. Four challenges related to spatio-temporal data mining are identified and addressed, including finding popular places in a city, identifying patterns in temporal and spatial distribution of telecom activity, and detecting correlations between popular events and distribution of telecom interaction. Further work will be focused on integrating different datasets and extracting useful knowledge from combined datasets. The analysis and mining of spatio-temporal data is, and will be the point of strong focus for further improvement. This research provides perfect starting point for extensive research in spatio-temporal mining.

## Acknowledgment

This paper has been realized as a part of the project "Studying climate change and its influence on the environment: impacts, adaptation and mitigation" (III 43007) financed by the Ministry of Education, Science and Technological Development of the Republic of Serbia for the period 2011-2017.

## References

[1] A. Stojnev, and D. Stojanović, "Sparkle – A Framework for Spatial Analysis on Apache Spark Platform", *IcETRAN 2016 Conference*, Zlatibor, 2016.

[2] A. Stojnev, and D. Stojanović, "Real-time Processing of Big Geospatial Data based on Spark Streaming", *SAUM 2016 Conference*, Niš, 2016.

[3] N. Andrienko, and G. Andrienko, "A Visual Analytics Framework for Spatio-Temporal Analysis and Modelling", *Data Mining and Knowledge Discovery*, pp.1-29, 2013.

[4] L. Cibulski, D. Gračanin, A. Diehl, R. Splechtna, M. Elshehaly, C. Delrieux, and K. Matković, "ITEA—Interactive Trajectories and Events Analysis: Exploring Sequences of Spatio-Temporal Events in Movement Data", *The Visual Computer*, vol. 32, no. 6-8, pp.847-857, 2016.

[5] N. Mamoulis, H. Cao, G. Kollios, M. Hadjieleftheriou, Y. Tao, and D. W. Cheung, "Mining, Indexing, and Querying Historical Spatiotemporal Data", *ACM SIGKDD 2004 Conference*, Seattle, WA, USA, 2004.

[6] R.F. Dos Santos, A. Boedihardjo, S. Shah, F. Chen, C. Lu, and N. Ramakrishnan, "The Big Data of Violent Events: Algorithms for Association Analysis Using Spatio-Temporal Storytelling", *GeoInformatica*, vol. 20, no. 4, pp. 879-921, 2016.

[7] Y. Huang, C. Chen, and P. Dong, "Modeling Herds and Their Evolvements from Trajectory Data", *Geographic Information Science*, vol. 5266, pp. 90-105, 2008.

[8] J. Xu, D. Deng, U. Demiryurek, C. Shahabi, and M. van der Schaar, "Mining the Situation: Spatiotemporal Traffic Prediction With Big Data", *Journal of Selected Topics in Signal Processing*, vol. 9, no. 4, pp. 702–715, 2015.

[9] N. Ferreira, J. Poco, H. T. Vo, J. Freire, and C. T. Silva, "Visual Exploration of Big Spatio-Temporal Urban Data: A Study of New York City Taxi Trips", *IEEE Trans. Vis. Comput. Graph.*, vol. 19, no. 12, pp. 2149-2158, 2013.

[10] T. Cheng, J. Haworth, B. Anbaroglu, G. Tanaksaranond, and J. Wang, "Spatiotemporal Data Mining", *Handbook of Regional Science*, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 1173-1193, 2014.

[11] R.R. Vatsavai, and B. Bhaduri, "Geospatial Analytics for Big Spatiotemporal Data: Algorithms, Applications, and Challenges", *InNSF Workshop on Big Data and Extreme-Scale Computing*, Charleston, SC, 2013.

[12] S. Shekhar, Z. Jiang, R.Y. Ali, E. Eftelioglu, X. Tang, V. Gunturi, and X. Zhou, "Spatiotemporal Data Mining: A Computational Perspective", *ISPRS International Journal of Geo-Information*, vol. 4, no. 4, pp. 2306-2338, 2015.

[13] K.V. Rao, A. Govardhan, and K. V.C. Rao "Spatiotemporal Data Mining: Issues, Tasks And Applications", *International Journal of Computer Science & Engineering Survey*, vol. 3, no. 1, pp. 39-52, 2012.

[14] I. Bruha, "Pre- and Post-processing in Machine Learning and Data Mining", *Lecture Notes in Computer Science*, pp. 258-266, 2001.

[15] Spazio Dati, "Dandelion API | Semantic Text Analytics as a Service", *Dandelion API*, Available: http://www.dandelion.eu. [Accessed: 10-Apr-2017].