Big Mobility Data Analytics

Recent Advances and Open Problems

Mahmoud Sakr[1,2], Cyril Ray[3] and Chiara Renso[4]

[1]Engineering School Of Brussels, Université Libre de Bruxelles, Belgium.
[2]Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt.
[3]Naval Academy Research Institute (IRENav) and Arts et Métiers Institute of Technology, France.
[4]ISTI Institute of CNR, Italy.

Contributing authors: mahmoud.sakr@ulb.com; cyril.ray@ecole-navale.fr; chiara.renso@isti.cnr.it;

1 Introduction

Nowadays, we have the means to collect, store and process mobility data of an unprecedented quantity, quality and timeliness. This is mainly due to the wide spread of GPS-equipped devices, including new generation smartphones and many connected objects. As ubiquitous computing pervades our society, mobility data represents a very useful source of information. Movement traces left behind, especially when combined with societal data, can aid spatial epidemiologists, transportation engineers, urban planners, and eco-scientists towards decision making in a wide spectrum of applications, such as COVID contact tracing, traffic engineering and risk management.

Along the last two decades the research community has passed from spatial to spatio-temporal and, then, to mobility data [1]. So, what's next? Not perhaps, but for sure it is the mobility-aware integrated Big Data analytics. It is now that the Big Data technological infrastructure matures. It is now that the interest of the research community has come to a peak about all issues related to the Big Data Value Analytics (BDVA) reference model: data management, data processing, data analytics, data visualization and user interaction. The emergence of artificial intelligence also brings new ideas, possibilities, and opportunities to cope differently with mobility data. So, it is now the time for the field of Big Mobility Data Analytics to follow the trend.

There is a clear need to foster the exchange of new ideas on multidisciplinary real-world problems, discussion on proposals about innovative solutions, and identify emerging opportunities for further research in the area of big mobility data analytics, covering all layers of the Big Data Value Analytics reference model, namely data management, data processing, data analytics, and data visualization and user interaction. The range of real-world problems is large and necessitate to bridge the gap between researchers and big data stakeholders, including experts from critical domains, such as indoor, urban / maritime / aviation transportation, human complex networks, etc.

2 Big Mobility Data Analytics

This special issue further builds on previous efforts by the editorial team and others from the mobility data science community to foster the exchange of new ideas on multidisciplinary real-world problems, discuss proposals about innovative solutions. It is also the opportunity to identify emerging opportunities for further research in the area of big mobility data analytics, such as deep learning on mobility data, edge computing, visual analytics, etc. Specifically, this special issue section follows up on the BMDA@EDBT 2021 workshop on Big Mobility Data Analytics, co-located with EDBT 2021 – 23rd-26th March 2021, Nicosia, Cyprus. This special issue is a continuation of the GeoInformatica Special Issues on Big Mobility Data Analytics (BDMA 2019, 2020)[2, 3], and on the series of BMDA@EDBT workshop proceedings[1].

Besides the introduction of the papers accepted in this special issue, we take the opportunity to also highlight the recent developments in mobility data analysis from the last editions of BMDA. We comment on how mobility data management and analysis used to be studied in the transportation domain, and in niche research communities and how they are evolving. Nowadays the abundance of location tracking technologies creates huge amounts of data, and opens possibilities for applications in almost every aspect of our modern lives. We survey the 'Big Mobility Data' challenges research in all data science from several directions including processing platforms and architectures, data management, analysis algorithms, visualization, etc. We do so by commenting on the preceding BMDA publications and give an overall reflection on how the topic is evolving.

Below we first summarize the contribution of five original papers presented in this special issue, then we present a short survey on the previous BMDA papers to draw some research directions and trends.

The focus of this special issue is on three aspects: (1) predictive analysis (Tritsarolis *et al.*, Nanni *et al.*, Raffaetà *et al.*), (2) data enrichment (El

[1] https://www.datastories.org/bmda/

Hafyani *et al.*), and on (3) scalable complex event processing over event streams (Ntoulias *et al.*).

From multiple aspect trajectories to predictive analysis (Raffaetà et al.)

In this paper authors combine trajectories from fishing vessels in the Adriatic Sea and fish catch reports to study the fishing activities in the Northern Adriatic Sea. Authors build, implement and analyze this combined dataset presenting all the phases of the database creation with MobilityDB, starting from the raw data and proceeding through data exploration, data cleaning, trajectory reconstruction and semantic enrichment. Several analyses have been performed on the resulting spatio-temporal database, with the goal of mapping the fishing activities on some key species, highlighting all the interesting information and inferring new knowledge that will be useful for fishery management. Furthermore, authors investigate the use of machine learning methods for predicting the Catch Per Unit Effort (CPUE), an indicator of the fishing resources exploitation in order to drive specific policy design. A variety of prediction methods, taking as input the data in the database and environmental factors such as sea temperature, waves height and Clorophilla, are put at work in order to assess their prediction ability in this field.

City Indicators for Geographical Transfer Learning: An Application to Crash Prediction (Nanni et al.)

This paper proposes a solution for the problem of predicting the crash risk of a car driver in the long term. This is an interesting problem since it requires a good knowledge of the driver movement behaviour and the geographical context surrounding his or her movements. This is a problem that has many interesting applications, from coaching high-risk drivers to, more in general, optimize the insurance market. The problem is formulated as a data driven approach where user mobility is represented as a network. Interesting is the representation of the areas where the drive moves with city indicators that capture different aspects of the city based on human mobility using mobility traces and road networks. Authors therefore develop a geographical transfer learning approach for the crash risk task such that they can build predictive models for another area where labeled data is not available. The method is empirically tested over real datasets to show the superiority of their solution.

Learning the Micro-environment from Rich Trajectories (El Hafyani et al.)

Mobile Crowd Sensing (MCS) has emerged as a valuable solution to collect multiple and heterogeneous time-dependent location data cross-related with contextual data including environmental data (e.g., pollution, weather), transportation mode (e.g., bus, pedestrian) and human activities (e.g., sport, work, shopping). The paper entitled “Learning the Micro-environment from Rich Trajectories in the context of Mobile Crowd Sensing Application to Air Quality

Monitoring” considers MCS mobility data as Multivariate Time Series to analyse in search for individual exposure to pollution. Considering air quality, and thus individual exposure to pollution, strongly depends on micro-environments (e.g., being in a car, in a park, at home…) where persons can evolve, the paper focuses on the feasibility of recognizing the human’s micro-environment during urban mobilities.

Based on a set of wearable sensors carried by users, the paper proposes an approach for automatically learning and predicting that micro-environment from users’ trajectories enriched with environmental data. Given a set of annotated trajectories enriched with contextual information, authors trained a model where the trajectory segments are the predictors, and the annotations provided by users constitute the class labels. The work relies on the multi-view learning paradigm. In this approach, learning a model is based on the different views of mobility data and each data source is considered independently (i.e., first-level learner on univariate time series) before being fused with others (i.e., meta-learner weighting the predictions of the first-level learners).

The work has been experimented and evaluated through data collected during three campaigns with more than one hundred participants, each gathering location, time, and measurements of particulate matter, nitrogen dioxide, black carbon, as well as relative humidity and temperature.

Predicting Co-movement Patterns in Mobility Data (Tritsarolis et al.)

In this paper, authors address the problem of predicting collective behavioural patterns of movement in what is called Online Prediction of Co-movement Patterns. Given a look-ahead time interval, the goal is to predict the clusters of moving objects that are anticipated to be shaped after the look ahead time interval. The architecture consists of an offline and an online layer. The offline layer uses a Future Location Prediction-offline model whereas at the online layer receives the streaming GPS locations, predict the next objects’ location (Future Location Prediction-online module), and discover evolving clusters at each timeslice.

The accuracy of the proposed solution the paper proposes a co-movement pattern similarity measure, which facilitates the comparison between the predicted clusters and the actual ones. The clusters’ evolution through time (survive, split, etc.) is computed and compared with the cluster evolution predicted by their approach. The experimental study uses two real-world mobility datasets from the maritime and urban domain, respectively, to demonstrate the effectiveness of the proposed solution.

Online fleet monitoring with scalable event recognition and forecasting (Ntoulias et al.)

A lot of meaningful events, sometimes hidden, can be extracted from mobility data. This paper focuses on a specific mobility analysis sector which is the fleet management where organisations need to manage thousands or even millions of commercial vehicles or vessels, detect dangerous situations (e.g., collisions or malfunctions) and optimise their behaviour. It poses the question of how to perform both complex event recognition and complex event forecasting on such data. For this task it is crucial to have a monitoring system which is highly efficient and scalable, reporting any such situations or opportunities as soon as they appear automatically detecting complex situations, possibly involving multiple moving objects and requiring extensive background knowledge.

A key point of the approach relies on a formalism that allows analysts to define complex patterns ensuring unambiguous semantics. Authors highlight the fact that combining and finding a tradeoff between expressive power and scalability is a significant challenge. The complex event processing and forecasting relies on symbolic automata and regular expressions as a computational model for pattern detection. Regarding forecasting, Markov chains (variable-order Markov model) are used for deriving a probabilistic description of a deterministic symbolic automaton. This formalism allows analysts to define complex patterns in a user-friendly manner while maintaining unambiguous semantics and avoiding ad hoc constructs.

Authors implemented their event processing using the popular Apache Flink, a distributed processing engine and Apache Kafka, a messaging platform to connect mobility data stream sources. The implementation is designed to run in cluster environments and perform parallel in memory computations.

The solution was experimented and evaluated on a on real data stream of maritime trajectories for which they have defined a set mobility pattern (e.g., high speed near coast, loitering, trawling...) constructed with the help of domain experts in order to detect critical situations for vessels at sea.

3 Recent advances in Mobility Data Analytics

3.1 The Different Thematic of Mobility Data

The research in mobility covers a wide range of application domains. Multiple thematic for the data can be seen in the literature, including urban, maritime, aviation, logistics, indoor, and involving both real, synthetic or pseudo-synthetic data. Although a large part of research is focusing on data understanding through analysis and analytics, there are also essential works that build analysis tools for spatio-temporal data in general.

Urban mobility is the most recurring thematic in the previous BMDA papers. Intelligent Transportation Systems (ITS) depend on mobility data analysis for building location-based services, situational awareness and predictions, to improve the resource planning. Building models for traffic forecasting enables policies that can reduce fuel consumption, reduce the commute time, and improve the comfort of city-users (residents, commuters, and visitors). Two levels of traffic analysis can thus be identified: (1) mining the historical trajectories for building long term traffic policies [4], and (2) short term traffic monitoring and forecasting [5, 6] for enabling the traffic controllers to take real-time actions to avoid/alleviate congestion. The main challenges in the latter case are the data volume and the real-time forecasting requirements.

Road freight transport, while essential, is a main generator of congestion (economic impact), and pollutant emissions (environmental impact). Therefore, governments are putting in-place policies to track the monitor truck movements, and to analyze them to enhance or regulate mobilities this data. In Belgium, traffic data is gathered for Heavy-Goods Vehicles (HGV) by Bruxelles Mobilité, the public administration responsible for mobility-related equipment and infrastructure in the Brussels Capital Region (BCR). A daily average of 19 GB of truck tracks are continuously collected. In [7] a distributed architecture is presented to enable the timely processing of this data, enabling urban planner to perform network-scale analysis and forecasting in near real-time.

The personal mobility of individuals is also a target for analyses including activity recognition, personalized routing, matching with ride-sharing, and crowd-sourcing. The paper [8] deals with the problem of providing route recommendations based on users' travel preferences. Instead of explicitly requiring users to indicate their preference, which can be complex, the history of user movement is used to imply her preferences. Privacy is a clear set-back for this kind of analysis, challenging the research to develop privacy-preserving algorithms.

The recent emergence of autonomous vehicles further increases the amount of mobility data. Beyond the challenges in terms of volumes and privacy it raises, falsifications and other alterations of the integrity of mobility data/sensors are becoming one of the most worrying subjects. [9] for instance summarizes few possible attacks on sensors embedded in autonomous vehicles which can affect data analytics on-board, driving or navigation. On the other hand, malversation of mobility data can be used to prevent tracking, hide an activity, alternate nominal function of location-based services. [10] for instance addresses the problem of integrity of maritime mobility data and proposes a methodology for maritime anomaly detection.

3.2 Mobility Data Models

While there is no widely agreed representation of mobility data tracks, a handful of models are recurring in the literature. Perhaps the most often used model is to represent a trajectory as a sequence of points. In this model, one point is at least a triple $< x, y, t >$ defining the spatial location of the moving object at the time instant t. Depending on the data and the analysis task, other attributes may be included, recording movement properties such as speed, heading, etc. The use of this model is motivated by the fact that the source data coming from location tracking sensors will natively arrive in this form. It can be in the form of an event stream [11], or in batch files csv, xml, etc, e.g., [12, 13]. This model is also straightforward to manage in relational databases and key-value stores, where every point is a tuple, or a key-value pair. Existing databases and spatial data libraries can thus be used for implementing the analysis [13, 14]. In [15] the proposed system enriches the individual location points with weather information. The points are processed independently from each other, where the system uses an external source storing weather data to enrich specific positions by weather attributes.

A common practice is to associate individual trajectory points with a trajectory identifier (or an object identifier) in order to perform analysis at the trajectory level. That is, the conceptual model is a set of trajectories, while the physical model is a set of points. A trajectory point p_i is pair of a location point 2D or 3D capturing the location of the moving object at a certain time instant. A raw trajectory τ_n is a sequence of trajectory points captured through time. This model allows to aggregate the analysis over at the trajectory points [4, 12, 14, 16, 17]. For instance, in [12] the individual points are annotated with the change in speed, course and heading to identify Search and Rescue (SAR) voyages, based on the observation that SAR voyages involve higher deltas for these properties.

For continuous movement, e.g., vehicles and, ships, the data model may involve interpolation between the observations [18] [19] [20]. Such a model tries to retain back the continuity of the data that is lost during the discrete observation process, i.e., in hertz. Certain analysis tasks, including proximity and similarity, require the continuous representation for the correctness of results.

For streaming location data, events are processed as they arrive, which simplifies the modelling perspective to individual points. Depending on the analysis, e.,g., CEP, the individual events are aggregated into the analysis model to incrementally extract information.The task of forecasting over time-evolving streams can assume that the stream is a time-series [5], and leverage the existing models such as ARIMA.

Data preparation, although important, is still not getting sufficient attention. Most of the cited works consider the cleaning as a secondary task, and perform basic deduplication and deletion of spikes. An often-needed preparation is to split the moving object trajectory into voyages [8, 12]. Thus, trajectory segmentation stands as an interesting analysis task in itself [4, 21].

3.3 Tools

There is a consensus in the mobility data communities that more work on systems for mobility data is needed. The literature is rich in terms of algorithms, specialized analyses, and prototypes. It remains difficult to integrate these works, since they base on different data models, specialized indexes, and different architecture requirements. There is lack of common data systems, similar to those existing for relational and spatial data.

Researchers thus work on adapting and extending existing big data tools for processing spatio-temporal data. In [22], for instance, the approach taken is to convert the spatio-temporal data into 1D, using an encoding based on space filling curves, then translate a limited set of spatio-temporal operations into the Spark operations: projection, selection and join. In contrast, works like [19] extend Spark with spatio-temporal indexing, partitioning, and query functionality. A similar line of work used to exist during the rise of Hadoop to extend it with spatio-temporal processing capabilities [23].

The ever increasing volume of data (not only mobility data) has also favoured along last years the development of many state of art tools to support data pre-processing, streaming, analysis and visualisation. Amongst a wide range of tools, one can cite for instance the ones developed by the Apache software foundation which cover the whole chain of data processing. For instance, Apache Nifi provides an easy to use and reliable system to graphically process and distribute data. Apache Kafka is a messaging platform to connect data stream sources. Apache Flink provides a distributed processing engine and a support for event detection. Finally Apache Superset provides a data exploration and visualization platform. Many alternative also exist such as the popular ELK platform which aggregate Elasticsearch, Logstash and Kibana in order to provide more or less a similar processing chain as the one developed by Apache Foundation. Often connected to R or Python codes or notebooks it provides an easy way to handle and analyses mobility data.

A notable direction is the adoption of Jupiter notebooks and libraries such as GeoPandas, Tensorflow, Keras, Scikitlearn, Folium, and Matplotlib with mobility data. A tutorial on traffic forecasting using deep learning is presented in [6], using UBD data. The method employs a Direct LSTM encoder-decoder in Tensorflow and train it using matrix representing the traffic readings at different time instances. GeoPandas and Folium are respectively used to process and visualize the geolocated traffic input. The python stack also offers opportunities for visual data exploration and analytics [13]. If only for the wide user base, the research community should put attention to developing specialized spatio-temporal libraries and visualizations. Another challenge, not exclusive to mobility data, is how to scale this stack for big data applications.

MobilityDB [24, 25] is an open-source extension to the PostgreSQL database system, and its spatial extension PostGIS. It implements temporal types and spatio-temporal types and operators in the database. As such, one can store a complete spatio-temporal trajectory into an attribute, i.e., one tuple represents a full trajectory and other properties. The values between successive instants are interpolated using a linear function. MobilityDB provides a rich set of functions over its temporal and spatio-temporal types. Interestingly some of these functions return temporal types, i.e., values changing in time. In [18] MobilityBD is used to represent semantic trajectories.

4 Conclusions

The work presented in this special issue and its preceding BMDA publications address data challenges that are of importance in many domains of our nowadays world. Moving forward, we see a clear need to reach to the application owners, to have access to real data, and to tailor the solutions to real problems. Data quality is often bypassed. For responsible analysis, more effort needs to be done in data cleaning and preparation, ultimately aiming at generalized methods that will get widely accepted. Lastly, we need to empower the professionals who deal with mobility data with tools that natively implement sound spatio-temporal data models and functions.

References

[1] Mokbel, M., Sakr, M., Xiong, L., Züfle, A., Almeida, J., Anderson, T., Aref, W., Andrienko, G., Andrienko, N., Cao, Y., *et al.*: Mobility data science (Dagstuhl seminar 22021). In: Dagstuhl Reports, vol. 12 (2022). Schloss Dagstuhl-Leibniz-Zentrum für Informatik

[2] Palpanas, T., Pelekis, N., Theodoridis, Y.: Special issue on big mobility data analytics (BDMA 2019). GeoInformatica **25**(2) (2021)

[3] Pelekis, N., Renso, C., Theodoridis, Y., Zeitouni, K.: Special issue on big mobility data analytics (BDMA 2020). GeoInformatica **26**(3) (2022)

[4] Etemad, M., Júnior, A.S., Hoseyni, A., Rose, J., Matwin, S.: A trajectory segmentation algorithm based on interpolation-based change detection strategies. In: EDBT/ICDT Workshops (2019)

[5] Cheng, S., Lu, F.: Short-term traffic forecasting: A dynamic ST-KNN model considering spatial heterogeneity and temporal non-stationarity. In: EDBT/ICDT Workshops, pp. 133–140 (2018)

[6] Buroni, G., Bontempi, G., Determe, K.: A tutorial on network-wide multi-horizon traffic forecasting with deep learning. In: EDBT/ICDT Workshops (2021)

[7] Dillen, A., Buroni, G., Borgne, Y.-A.L., Determe, K., Bontempi, G.: MOBI-AID: A big data platform for real-time analysis of on board unit data. In: EDBT/ICDT Workshops (2020)

[8] de Oliveira e Silva, R.A., Cui, G., Rahimi, S.M., Wang, X.: Personalized route recommendation through historical travel behavior analysis. GeoInformatica **26**(3), 505–540 (2022)

[9] Kumar, A.D., Chebrolu, K.N.R., Vinayakumar, R., P, S.K.: A brief survey on autonomous vehicle possible attacks, exploits and vulnerabilities. CoRR (2018)

[10] Iphar, C., Ray, C., Napoli, A.: Data integrity assessment for maritime anomaly detection. Expert Systems with Applications, 113219 (2020)

[11] Qadah, E., Mock, M., Alevizos, E., Fuchs, G.: A distributed online learning approach for pattern prediction over movement event streams with apache flink. In: EDBT/ICDT Workshops (2018)

[12] Chatzikokolakis, K., Zissis, D., Spiliopoulos, G., Tserpes, K.: Mining vessel trajectory data for patterns of search and rescue. In: EDBT/ICDT Workshops, pp. 117–124 (2018)

[13] Graser, A.: Notebook-based visual analysis of large tracking datasets. In: EDBT/ICDT Workshops (2021)

[14] Makris, A., Tserpes, K., Spiliopoulos, G., Zissis, D., Anagnostopoulos, D.: Mongodb vs postgresql: A comparative study on performance aspects. GeoInformatica **25**(2), 243–268 (2021)

[15] Koutroumanis, N., Santipantakis, G.M., Glenis, A., Doulkeridis, C., Vouros, G.A.: Scalable enrichment of mobility data with weather information. Geoinformatica **25**(2), 291–309 (2021)

[16] Etemad, M., Júnior, A.S., Matwin, S.: On feature selection and evaluation of transportation mode prediction strategies (2019)

[17] Guidotti, R., Nanni, M., Sbolgi, F.: Data-driven location annotation for fleet mobility modeling. In: EDBT/ICDT Workshops (2020)

[18] Rovinelli, G., Matwin, S., Pranovi, F., Russo, E., Silvestri, C., Simeoni, M., Raffaetà, A.: Multiple aspect trajectories: a case study on fishing vessels in the northern adriatic sea. In: EDBT/ICDT Workshops (2021)

[19] Alamdari, O.I., Nanni, M., Trasarti, R., Pedreschi, D.: Towards in-memory sub-trajectory similarity search. In: EDBT/ICDT Workshops (2020)

[20] Abreu, F.H.O., Soares, A., Paulovich, F.V., Matwin, S.: Local anomaly detection in maritime traffic using visual analytics. In: EDBT/ICDT Workshops (2021)

[21] Wu, S., Zimányi, E., Sakr, M., Torp, K.: Semantic segmentation of AIS trajectories for detecting complete fishing activities. In: MDM 2022 (2022)

[22] Nikitopoulos, P., Vlachou, A., Doulkeridis, C., Vouros, G.A.: DiStRDF: Distributed spatio-temporal RDF queries on Spark. In: EDBT/ICDT Workshops (2018)

[23] Bakli, M.S., Sakr, M.A., Soliman, T.H.A.: A spatiotemporal algebra in hadoop for moving objects. Geo-spatial Information Science **21**(2), 102–114 (2018)

[24] Zimányi, E., Sakr, M., Lesuisse, A.: MobilityDB: A mobility database based on PostgreSQL and PostGIS (2020)

[25] Zimányi, E., Sakr, M., Lesuisse, A., Bakli, M.: MobilityDB: A mainstream moving object database system. SSTD '19 (2019)