УДК 004.048
МРНТИ 20.19.01

EXTRACTING HIDDEN FEATURES OF HUMAN MOBILITY AND PREDICTING INFLOW AND OUTFLOW OF BIKE SHARING STATIONS

E.S. SEITBEKOVA[1,2], B.K. ASILBEKOV[1], A.B. KULJABEKOV[1], I.K. BEISEMBETOV[1]

[1]Satbayev University
[2]Kazakh-British Technical University

Abstract: *Huge amounts of spatial-temporal data are generated daily from all kinds of citywide infrastructures. Understanding and predicting accurately such a large amount of data could benefit many real world applications.*

This paper provides an analysis of human mobility data in an urban area using the amount of available bikes in the stations of the bicycle sharing program. Based on data sampled from the operator's website, it is possible to detect temporal and geographic mobility patterns within the city. These patterns are applied to predict the number of available bikes for any station some hours ahead. Our methodology first identifies and quantifies the latent characteristics of different spatial environments and temporal factors through tensor factorization. Our hypothesis is that the patterns of spatial-temporal activities are highly dependent on or caused by these latent spatial-temporal features. We model this hidden dependent relationship as a Gaussian process, which can be viewed as a distribution over the possible functions to predict human mobility.

Keywords: *human mobility, spatio-temporal, hidden features, tensor, Tucker decomposition, Gaussian process, prediction*

АДАМДАРДЫҢ ШАПШАҢДЫЛЫҒЫНЫҢ ЖАСЫРЫН ЕРЕКШЕЛІКТЕРІН АНЫҚТАУ ЖӘНЕ ВЕЛОСИПЕД БӨЛІСУ СТАНЦИЯЛАРЫНЫҢ ВЕЛОСИПЕД АҒЫНЫН БОЛЖАУ

Аңдатпа: *Кеңістік-уақыттық деректердің үлкен көлемі күнделікті барлық қалалық инфрақұрылымдардан жасалады. Осындай үлкен көлемді мәліметтерді түсіну және болжау көптеген нақты әлемдік қосымшаларға пайда әкелуі мүмкін.*

Бұл мақала велосипед бөлу станциясының деректерін пайдалана отырып, қалалық жерлердегі адамдардың мобильділігі туралы деректерді талдайды. Оператордың веб-сайтынан алынған мәліметтер негізінде қалада кеңістік-уақыттық ұтқырлық анықталуы мүмкін. Осы схемалар бірнеше сағат бұрын кез келген станция үшін қолжетімді велосипедтердің санын болжау үшін пайдаланылады. Біздің әдістемеміз алдымен тензорлы факторизация арқылы түрлі кеңістік-уақыттық факторлардың жасырын ерекшеліктерін анықтайды және сандық түрде айқындайды. Ал гипотеза кеңістік-уақыттық белсенділіктің үлгісі жасырын ерекшеліктерге тәуелді. Бұл тәуелділікті Гаусс үдерісі ретінде модельдейміз.

Түйінді сөздер: *кеңістік-уақыттық сипаттамалары, жасырын ерекшеліктері, тензоры, Такердің ыдырауы, Гаусс үдерісі, болжау*

ВЫЯВЛЕНИЕ СКРЫТЫХ ХАРАКТЕРИСТИК МОБИЛЬНОСТИ ЛЮДЕЙ И ПРОГНОЗИРОВАНИЕ ПРИТОКА/ОТТОКА ВЕЛОСИПЕДОВ НА СТАНЦИЯХ СОВМЕСТНОГО ИСПОЛЬЗОВАНИЯ

Аннотация: *Огромное количество пространственно-временных данных генерируется из всех типов городской инфраструктуры. Точное понимание и прогнозирование такого большого объема данных может принести пользу многим реальным приложениям.*

В этой статье представлен анализ данных о мобильности людей в городских районах с использованием данных со станции совместного использования велосипедов. На основе данных, взятых с веб-сайта оператора, можно определить временную и географическую мобильность в пределах города. Эти схемы используются для прогнозирования количества доступных велосипедов для любой станции на несколько часов вперед. Наша методология сначала идентифицирует и количественно определяет скрытые характеристики различных пространственных сред и временных факторов посредством тензорной факторизации. Гипотеза авторов состоит в том, что закономерности пространственно-временной активности сильно зависят от этих скрытых пространственно-временных особенностей. Мы моделируем это как зависимые отношения, как гауссовский процесс, который можно рассматривать как распределение.

Ключевые слова: *мобильность людей, пространственно-временные характеристики, скрытые особенности, тензор, разложение Такера, гауссовский процесс, прогнозирование*

Introduction

Public bike sharing systems are becoming more and more popular in the past few years. A still growing list of cities which provides such service systems can be found at the Bike sharing world map [1]. The three big cities of Kazakhstan have such systems. They are: AlmatyBike, AstanaBike and ShymkentBike [2]. There are 200 stations in Almaty, 180 stations in Astana and 40 bike sharing stations in Shymkent. For optimal performance of such systems there must be (a) the possibility to find a bike when a user wants to start his/her journey and (b) the possibility to leave the bike in the user's destination. Without oversizing the system, there are basically two ways to solve these problems: Inform the user in advance about the best places to pick up or leave the bikes and improve the redistribution of bikes from full to empty stations.

In this study we aim to contribute to the solution of these problems via the analysis of cyclic mobility patterns which lead to short term predictions of the number of available bikes in the stations by prediction inflow and outflow between stations. Such predictions would allow us to improve the current web-service of AlmatyBike, AstanaBike and ShymkentBike and in turn increase users' satisfaction with the system. Once this type of information is available, users may use mobile devices to access it. Knowledge of those patterns could lead to an optimization of the bike sharing system itself, allowing the operator to predict shortage or overflow of bicycles in certain stations well in advance and adapt its redistribution schedule accordingly on the fly.

Furthermore, we intend to show that this type of data also allows us to infer the activity of city population as well as the spatial-temporal distribution of their displacements. Such knowledge may be interesting for city planners and may also represent a cheap way to compare the activity cycles between different cities.

Big Data trend brings great opportunities for tackling many real-world challenges. In this paper, we propose a novel methodology for prediction of spatial-temporal activities using latent spatial and temporal factors extracted from existing mobility datasets at a city level. Of spatial-temporal activities, we are interested in human mobility, especially the inflow and outflow of people in neighborhoods/areas during certain time periods. Understanding the inflow/outflow of people in urban environments spatially and temporally and predicting them correctly are essential to solve many real-world problems. Such as optimization bike sharing systems. To achieve these goals, we use spatial-temporal data, which has been obtained from New York Cities bike sharing systems open data. Website provides all historical and real time data about the number of bicycles available for the users in a certain moment in time in every one of the approximately 700 different stations, and information about all rent done by users from 2013 year.

The rest of the paper is organized as follows. We first review related work on the subject in

2 and give a more detailed description of the mobility spatio-temporal data in 2.1. Section 3 describes the tensor model of human mobility and the extraction of latent spatial and temporal features. Section 4 presents the prediction methodology through modeling the relationship between latent features and human mobility as a Gaussian process. Section 5 demonstrates the performance of our methodology through a series of experiments with the bike trips in New York. Finally, we present the conclusions in Section 6.

1. Related work

Given the importance of gaining a deeper understanding of many spatial-temporal activities, like human mobility, and predicting them accurately, related work in this area has been published in various fields, such as computer science, urban planning, sociology, and other areas. In this section, some of the works relevant to different aspects of mobility patterns are overviewed.

Some studies have visualized bike sharing systems activity, identifying trends, usually based on the performance analysis of connecting stations, observing the number of trips starting and ending at the station level [3] [4] [5]. The number of studies analyzes bike sharing systems imbalances caused by various levels of attractiveness and generation of station-level trips [6], often they provide efficient bike redistribution strategies [7] [8].

With a similar goal of introducing a more balanced systems, other studies modeled demand [9] or developed models that optimize the location of stations [10]. Signficant number of studies have recently focused on the GPS analysis of casual cyclists' routes [11]. A couple of studies have focused on the exploration of real bike routes. First, the route choice analysis performed by Khatri [12], based on approximately 12,000 trips collected through the Phoenix BSS bikes equipped with built-in GPS trackers, second , the research published by Wergin and Buehler [13], analysing 3596 trips obtained by introducing GPS trackers into 94 bikes in the Washington DC BSS in 2015, and the study to visualise the cycling flow derived from Madrid bike sharing system activity, obtained by processing over 250,000 GPS routes, and provide an analysis of how this flow is distributed across the urban street network at different moments [14].

2.1 CitiBike NYC bike sharing system

The methodology described in this paper is tested on open data available on web site https://www.citibikenyc.com/system-data. There can be found data about all trips done by user of system CitiBike New York and annual reports per month from May 2013. CitiBike system start his operation in May 2013. On average, there are 43,604 rides per day, with each bike used 3.5 times per day. It has 8,629 annual members and 61,715 casual members signed up or renewed during the month. There are 757 active stations at the end of the month. The average bike fleet is 12,744 with 12,793 bikes in the fleet. Citi Bike staff rebalances on average 22,280 bicycles per month. The Service Delivery Department utilizes box trucks, vans, contracted trikes, articulated trikes ('bike trains'), valets, and member incentives ('Bike Angels') to redistribute bikes system-wide.

CitiBike system publish downloadable files of CitiBike trip data. The data includes: Trip Duration (seconds), Start Time and Date, Stop Time and Date, Start Station Name, End Station Name, Station ID, Station Lat/Long, Bike ID, User Type (Customer = 24-hour pass or 3-day pass user; Subscriber = Annual Member), Gender (Zero=unknown; 1=male; 2=female), Year of Birth. This data has been processed to remove trips that are taken by staff as they service and inspect the system, trips that are taken to/from any of "test" stations, and any trips that were below 60 seconds in length (potentially false starts or users trying to re-dock a bike to ensure its secure).

3. Tensor model and extracting latent features

Tensor is a array with 3 or more dimensions. Decompositions of a higher-order tensor can be used to extract and explain the properties among the tensor. Tensor decomposition is widely

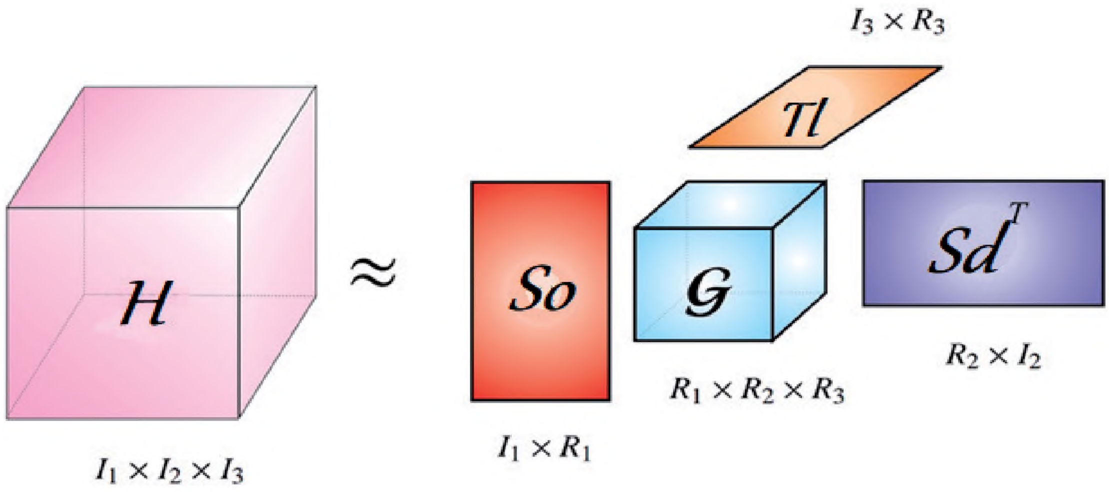

Figure 1. Tensor decomposition

used in computer vision, numerical analysis, data mining, neuroscience, graph analysis etc. [15]. In this paper, we propose to model human flow between different neighborhoods with a 3-dimensional tensor H ∈ $R^{N \times N \times L}$, as shown in Figure 1. The first dimension of the tensor H denotes N origin neighborhoods, the second dimension denotes N destination neighborhoods, and the third dimension denotes L time slots, respectively. Each entry of the tensor H(i, j, l) stores the average number of trips starting from neighborhood i to neighborhood j during time period l.

With this tensor model, we extract the latent spatial features of each origin neighborhood, destination neighborhood, and the latent temporal feature of each time slot through a Tucker decomposition. The Tucker decomposition can be thought of as the form of higher-order Principal Component Analysis (PCA). It decomposes a tensor into a core tensor multiplied by a matrix along each dimension [15]. In our case, we decompose the tensor H into three matrices

$\mathcal{S}o^{\ N \times P}$, $\mathcal{S}d^{\ N \times Q}$, $\mathcal{T}^{\ L \times R}$, and a core tensor $G^{P \times Q \times R}$, as shown in Figure 1. Mathematically, this relationship can be expressed as in Equation 1:

$$\mathcal{H} \approx G \times_1 \mathcal{S}_o \times_2 \mathcal{S}_d \times_3 \mathcal{T} \qquad (1)$$

4. Predicting using latent features

After the extraction of latent spatial-temporal features, we mathematically model the relationship between spatial-temporal activities such as human mobility and the extracted latent features for prediction. For this, we assume that people’s mobility is generated from a smooth and continuous process. This process has typical amplitude and variations in the function which takes place over spatial, temporal, and other characteristics. For example, to predict the volume of outflow xo i, in the neighborhood i during time period l (or the volume of inflow $x\iota$ i, l), we can model the relationship as below:

$$x_{o\,i,l} = g(\mathcal{S}_{o\,i,:}, \mathcal{T}_{l.:}, x_{o\,i,l-1}, \dots) \qquad (2)$$

$$x_{\iota i,l} = g(\mathcal{S}_{di.:}, \mathcal{T}_{l,:}, x_{\iota i,l-1}, \dots) \qquad (3)$$

Note that instead of relating this relationship to some specific models such as linear, quadratic, cubic, or even non-polynomial models, which may have numerous possibilities, we modeled this relationship as a free-form Gaussian process. One reason for using the Gaussian process is that for any spatial-temporal activity y (e.g., xo i,l) to be predicted, it will likely be generated by the same process and have similar values as the historical processes that share similar latent spatial-temporal features. We can take advantage of this relationship and use it for prediction. The Gaussian process is described properly in the work of Rasmussen [16].

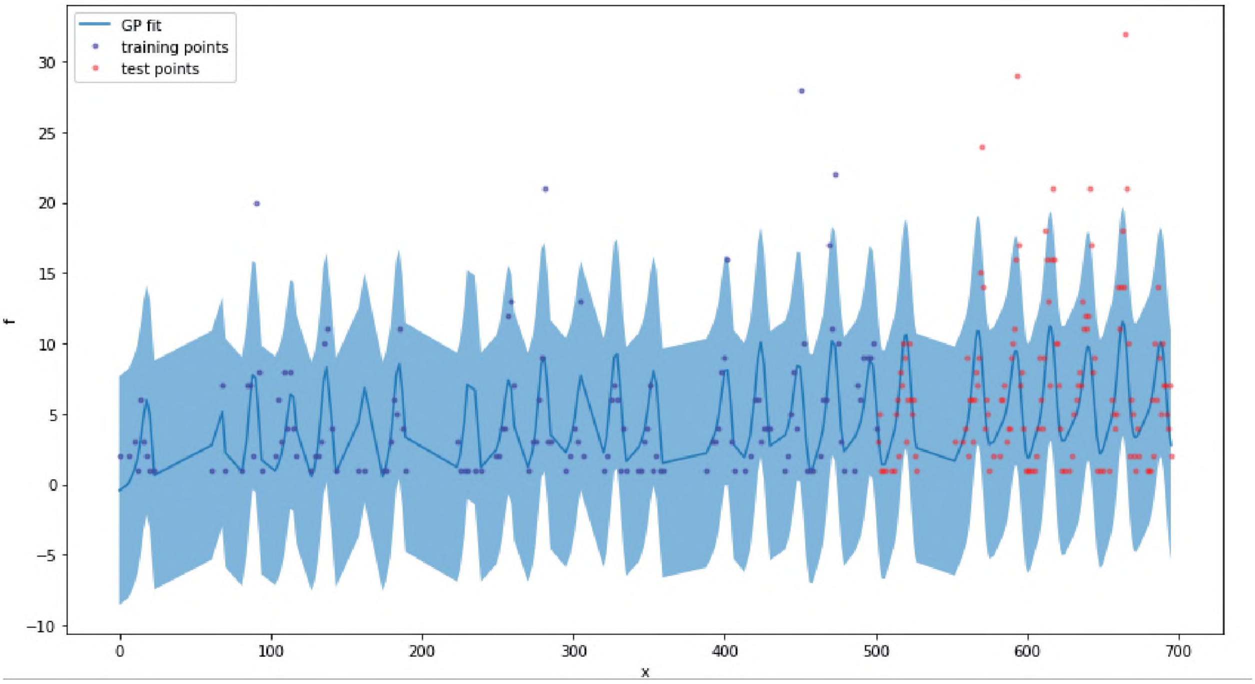

Figure 2. Results of prediction

5. Results

The mathematical model and all calculations are done in programming language python with help of libraries. Three-dimensional tensor was created with numpy library. Generated tensor is decomposed by Tucker decomposition algorithm using tensorly library, that is highly recommended for tensor learning in python. Gaussian regression was done by library GPy. In order to verify our assumptions, the data set was divided into two part: train data and test data. Taking a look at the training data, we can see a number of features that occur in the data. There is a clear periodic trend that is daily or weekly. We can use this prior information in our choice of kernel to give some meaning to the Gaussian process fit. In the Figure 2 you can see the prediction of outflow of one station for one month. In figure vertical axis is the number of bikes taken from that station, horizontal axis is the number of hours in month. For example, 0 is first hour (00:00-00:59) of first day of month, 39 is time period 14:00-14:59 of second day of month (39-24=15).

For prediction accuracy measurements, we used the mean squared error (MSE) and mean absolute scaled error. MSE in average is 1919 and MASE is 0.38.

Conclusion

In this work, we proposed a new methodology for the prediction of spatial-temporal human mobility, especially the inflow and outflow of bikes from one station to another during certain time periods. Our methodology comprised two steps: (1) use of a 3D tensor to model human mobility and extract latent spatial and temporal features of different stations and time periods through tensor factorization; and (2) modeled relationship between mobility patterns and the extracted latent spatial and temporal features as a Gaussian process for prediction of human mobility. For validation of the proposed methodology, we experimented with New York City's bike trips. The results showed that our extracted latent features successfully distinguish between bike sharing stations with diverse unique characteristics.

This work has done under the project № AP05134776 «Location Analytics Techniques for Prediction of Mobility Patterns» of the Ministry of Education and Sciences of the Republic of Kazakhstan.

REFERENCES

1. Bike sharing world maps web site. URL: www.bikesharingmap.com
2. Kazakhstani bike sharing systems web sites. URL: https://velobike.kz/, https://almatybike.kz/, https://shymkentbike.kz/
3. Borgnat, P., Robardet, C., Abry, P., Flandrin, P., Rouquier, J-B., & Tremblay, N. (2013). A dynamical network view of Lyon ´ S V ´ Elo ´ v shared bicycle system. Dynamics on and of Complex Networks, 2, 267–284.
4. O'Brien, O., Cheshire, J., & Batty, M. (2014). Mining bicycle sharing data for generating insights into sustainable transport systems. Journal of Transport Geography, 34, 262–273. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0966692313001178
5. Zaltz Austwick, M., O'Brien, O., Strano, E., Viana, M., &Gomez-Gardenes, J. (2013). The structure of spatial networks and communities in bicycle sharing systems. PLoS ONE, 8(9), e74685, 1–17.
6. Goodman, A., & Cheshire, J. (2014). Inequalities in the London bicycle sharing system revisited: Impacts of extending the scheme to poorer areas but then doubling prices. Journal of Transport Geography. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0966692314000659
7. Lin, J. H., & Chou, T. C. (2012). A geo-aware and VRP-based public bicycle redistribution system. International Journal of Vehicular Technology, 2012, 1–14.
8. Raviv, T., Tzur, M., & Forma, I. A. (2013). Static repositioning in a bike-sharing system: Models and solution approaches. EURO Journal on Transportation and Logistics, 2(3), 187–229. Retrieved from http://link.springer.com/10.1007/s13676-012-0017-6
9. Systems, D. T. & Lackner, B. (2013). Modeling demand for bicycle sharing systems – Neighboring stations as a source for demand and a reason for structural breaks. TRB, 1–19.
10. García-Palomares, J. C., Gutiérrez, J., & Latorre, M. (2012). Optimizing the location of stations in bike-sharing programs: A GIS approach. Applied Geography, 35(1–2), 235–246.
11. Romanillos, G., Zaltz Austwick, M., Ettema, D., & De Kruijf, J. (2016). Big data and cycling. Transport Reviews, 36(1), 114–133.
12. Khatri, R. (2015). Modeling route choice of Utilitarian Bikeshare users from GPS Data. University of Tennessee. Retrieved from http://trace.tennessee.edu/cgi/viewcontent.cgi?article=4837&context=utk_gradthes
13. Wergin, J., & Buehler, R. 2018. Where do bikeshare bikes actually Go? An analysis of capital bikeshare trips using GPS data. Transportation Research Record, (January), 2662, 12–21.
14. Gustavo Romanillos, Borja Moya-Gómez, Martin Zaltz-Austwick & Patxi J.Lamíquiz-Daudén (2018) The pulse of the cycling city: visualising Madrid bike share system GPS routes and cycling flow, Journal of Maps, 14:1, 34-43, DOI: 10.1080/17445647.2018.1438932
15. Kolda, T. G., & Bader, B. W. (2009). Tensor decompositions and applications. SIAM Review, 51, 455–500.
16. Rasmussen, C. E. (2006). "Gaussian processes for machine learning."