Determination of origins and destinations for an O-D matrix based on telecommunication activity records

Saša Dešić[1], Mia Filić[2], Renato Filjar[1]

[1]Ericsson Nikola Tesla, Krapinska 45, 10000 Zagreb, Croatia
[2]Faculty of Science, University of Zagreb, Croatia

Abstract - ***The origin-destination (O-D) matrix is a known indicator of individual and group mobility in transport science, an invaluable input for socio-economic activity assessment of the community, and a tool for strategic planning and policy developments. The abundance of mobility-related data allows for expansion of the ODM towards the description of the over-all socio-economic activity, of which the transport is just a component, thus rendering policy development and strategic planning more efficient. Determination of the sources (origins) and targets (destinations) of the socio-economic activities remains an unresolved point in the ODM estimation process. Here we present a novel method for activity spot (origins and destinations) identification from the anonymised records of telecommunications activity, and publicly available data on telecommunication network access points. Using the spatial tessellation algorithm, the origins and destinations can be effectively identified as the focal points of socio-economic activity, thus establishing the foundation for more efficient urban policy development and strategic planning.***

I. INTRODUCTION

The position-location duality in separated (physical and contextual) domains, introduced by one of us in [6], has gained the attraction among the researchers aiming to explore the integration of spatial and Information and Communication Technology (ICT) data for generation of completely new groups of systems and services.

Here we present a novel approach in advanced utilisation of telecommunications activity data (context) matched with spatial data (physical world) in advanced determination of areas of attractions (origins and destinations) for Origin – Destination (O-D) matrices estimation.

The manuscript is organised as follows.

Section II outlines the problem of determination of origins and destinations of an O-D matrix, and depicts the traditional approach to solve it. Section III describes the process of the O-D matrix estimation using telecommunications activity records. Section IV presents a novel approach in determination of origins and destinations using telecommunications activity records and a mathematical method called Voronoi tessellation. Section V presents results of the concept validation with the actual telecommunications activity record, and discusses advantages and shortcomings of the proposed approach. The manuscript concludes with the summary of research results and proposal for future research in Section VI.

II. PROBLEM DESCRIPTION

The origin-destination (O-D) matrix is a known indicator of individual and group mobility in transport science [10, 13], an invaluable input for socio-economic activity assessment of the community [11, 7, 5], and a tool for strategic planning and policy developments [7, 9]. An O-D matrix represents the socio-economic-driven intensity of local migration between the previously estimated areas of socio-economic attraction (residential areas, office areas, manufacturing areas, shopping malls, sport centres, concert halls etc.) [12, 5].

An O-D matrix is a scale-dependent tool. Its design begins with a proper determination of areas of attraction (origins and destinations) of local migrations [10]. The scale-dependency nature of an O-D matrix introduces constraints in definition of areas of attractions (origins and destinations). In traditional O-D matrix determination process, the origins and destinations are defined based on naïve estimations, interviews and examinations of cadastral data. Such an approach inevitably introduces uncertainty and noise, thus rendering the estimated O-D matrix less accurate and less fitted to the purpose.

Furthermore, a proper scaling of areas of interest requires skills and experience. Otherwise, it can produce a misleading evidence by either rendering areas of attraction too large (thus merging several actual areas of attraction into one), or too small (thus converting an O-D matrix into matrix of individual- or group-migration and loosing the original substance of an O-D matrix).

Therefore, a systematic and methodical approach in determination of origins and destinations of an O-D matrix, that will efficiently integrate the contextual knowledge with a proper scaling, is required.

III. O-D MATRIX DETERMINATION FROM TELECOMMUNICATIONS ACTIVITY DATA

The methodology of O-D matrix estimation comprises the following tasks: data preparation, determination of areas of activity (origins and destinations), migration identification (paring origins with destinations), O-D matrix estimation and graphical presentation [10, 5].

Let assume that we have n pre-defined disjoint areas, where n is a natural number. Let them count $1, 2, \ldots, n$. An O-D matrix is a square real matrix with non-negative elements. An (i, j)-element of an O-D matrix represents the number of migrations from i-th to j-th areas of attraction (origin-to-destination) in the observed time-span. Matrix rows represents origins, while columns represent destinations. It should be noted that, (i, j) element represents the number of migrations from the origin i to destination j, and (j, i) from origin j to destination i, with the two elements not necessarily equal (i.e. the O-D matrix is generally not symmetrical). An origin and a destination (comprising one migration O-D pair) are defined as the first and the last stop of the one migration event. In order to estimate the O-D matrix, the areas of observation and the number of migrations (migration connection) in predefined time-window between two areas from earlier defined set of areas of migration observations should be determined.

In a recently introduced approach [9], an O-D matrix can be estimated using the data sets describing telecommunications activity: voice calls, short-messages exchange, and data traffic over mobile internet. Summed up in internal Charging Data Records (CDRs) utilised by telecommunication network operators to charge the users for telecom services usage, those data sets can be used for distillation of the information necessary for the O-D matrices estimation procedure. Charging Data Record (CDR) is a collection of information about a chargeable telecommunication events (performed phone call, internet usage from mobile devices/units and Short Messaging Services (SMS), and ID of the base station involved in telecommunication activity) which are used by telecommunication network operators for user charging purposes. In a crucial development, a base station ID has been matched with the base station position in physical world, thus allowing for definition of transformation between physical and contextual (information) domain [9, 5, 4]. An O-D matrix based on telecommunications activity record provides more detailed and accurate insight into socio-economic-driven urban migrations that the classical O-D matrices based on road traffic data. Based on the CDR data set as the only source of information on migration of local population, we created the rules for determination of the areas of observation and mutual migration connection number [5].

The traditional approach in origins and destinations determination for telecom-based O-D matrix calls for the spatial partition founded on the estimation of the base station coverage. The base station radio coverage areas are not strictly separated in practice, allowing for a frequent overlapping of adjacent areas, thus creating a considerable ambiguity in origins and destination determination.

IV. DETERMINATION OF ORIGINS AND DESTINATIONS FROM CDR USING VORONOI DIAGRAM

A CDR is a massive data set generated during usage of a telecommunication network by a user. The set comprises numerous data that does not necessarily contribute to the ODM development, Additionally, the original CDR contains data from which the user's ID can be derived easily, pointing to his or her record of movement, among the other features. In due course, for the purpose of the ODM estimation the original CDR must be firstly anonymised in order to preserve the telecom users' privacy, followed by the data set reduction in the sense of extraction the ODM-relevant data only. The research presented here were conducted using the anonymised and ODM-related extract of the original CDR data set, a sub-set referred to as the CDR data in the remaining part of the manuscript. The CDR data sub-set comprised the following ODM-related data: (anonymised) user ID, time-stamp of telecom service initiation, user position approximated by position of the base station. A publicly available CDR data sub-set was used in this research [14].

A Voronoi diagram of a set S of n uniquely described points, called *seeds* or *sites*, in $\mathbf{R}^d$ is a partition of space into Voronoi cells, in a manner

that each cell consists of all the points closer to a particular seed than to any other. [1]. A Voronoi cell corresponding the seed with co-ordinates *(x, y)* (if the 2-dimensional Euclidean space is considered, or more generally: $(x_1, x_2, \ldots, x_n)$ for n-dimensional Euclidean space) bounds the area around *(x, y)* that is closest to the *(x, y)* relative to the other seeds of the same Voronoi diagram. The Euclidian metrics for distance determination is the most common one used, while the other metrics (such as: Manhattan and Mahalanobis distance metrics) may be used for particular case scenarios) [2, 8].
The Euclidian metrics, utilised in this research, determines the distance between the points A (x_A, y_A) and B (x_B, y_B) using the model depicted in (1).

$$d(A,B) = \sqrt{((xB - xA)^2 + (xB - xA)^2)} \quad (1)$$

Each Voronoi cell is uniquely represented by its seed (the base station coordinates, in the case of telecommunication activity record utilisation) and the distance determination metrics (it determines the shape of a Voronoi cell). A procedure for the space partition using Voronoi diagram is often referred to as Voronoi tessellation.
Data preparation process comprises the extraction (identification) of all the base stations involved in telecommunications activity. Indirectly, the partition of the observed area (defined by CDR data set itself by observing the area around the base stations that CDR set comprises) onto Voronoi cells determines the areas of socio-economic activities exploiting the direct relationship between the position of base stations and the concentration of the mobile communication network users.
Each Voronoi cell corresponds to only one base station, or its seed. Following the change of base stations ID in the ordered set of CDRs which belongs to one user, a user migration can be determined. By determination of all migrations of all users referred to CDR remembering the first and the last stop of each migration, the number of migrations between specific areas can be determined. The spatial separation according to Voronoi cells represents the data arrangement in spatial domain.
Furthermore, the data preparation includes the extraction of information about the neighbouring Voronoi cells for each Voronoi cell in observation (set as an origin, for instance). A neighbouring cell of cell C is a cell which seed is from the C no more than D = 1km away. This way we define all neighbouring cells of cell C as all cells that can, theoretically, take over the voice call/Short Messaging Service (SMS)/data transfer from base station seed C due to its overload. The D value was set to 1 km in order to accept the common practice in telecom networks to disperse the base stations (in third-generation (3G) mobile communication networks) at 1 km – 3 km separations. [5]. The CDR set utilised in this project originates from urban area.

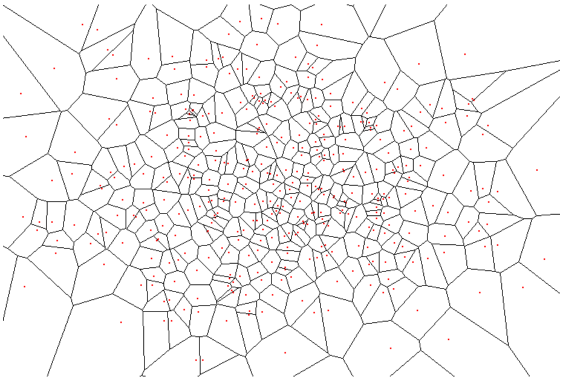

Fig 1 Voronoi tessellation

V. A CONCEPT VALIDATION AND DISCUSSION

The concept of determination of origins and destinations using telecommunication activity records and Voronoi tessellation was validated using openly available data set comprising anonymised CDR data sub-set collected in a 3G mobile network in Shenzhen, China [15]. The Voronoi tessellation algorithm based on Euclidian metrics was developed and implemented in the R programming environment for statistical computing [12, 3]. Spatial data provided through Google Maps API was used for identification of spatial objects and graphical presentation. The results of the Voronoi tessellation for the case presented are depicted in Fig 2.

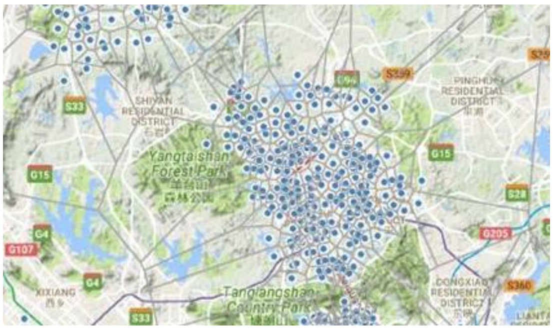

Fig 2 A Voronoi tessellation of Shenzhen, based on anonymised CDR

A provisional visual inspection of a digital map presentation revealed a good spatial separation of areas of socio-economic attraction by exploitation of the nature of radio network planning, that favours the situation of telecommunication network base station in centres of clusters of objects that attract the socio-economic activity. Still a more formal methodology for measurement of such a spatial fit remains to be developed in order to

provide an objective indicator for the quality of spatial partition based on socio-economic activity.
In comparison with the previous research accomplishments [7-10,12,13], we have identified the contextual value in matching the base station-centred Voronoi tessellation with the spatial data on socio-economic activities. This finding is to be explored further through the assessment of the other concepts in Voronoi tessellation and the advanced algorithms for identification of regions of attraction through spatial analytics of contextual data.
Additionally, the utilisation of the 3G-network base stations as the Voronoi cell seeds assures the proper scaling, while rendering a set of Voronoi-determined origins and destinations a proper foundation for an accurate determination of the O-D matrix. The proposed concept may be improved further with consideration of positions of the objects of interest (concert halls, shops, restaurants etc.) and tailoring the origins and destinations algorithm determination in order to evaluate the influence of small-sized attraction spatial objects.

VI. CONCLUSION AND FUTURE RESEARCH

A novel approach in determination of origins and destinations for an O-D matrix, based on the utilisation of telecommunications activity records and the Voronoi spatial tessellation is presented in this manuscript. Validated with the real 3G telecommunications activity record data using the R-based software developed by our team, the proposed concept delivered improved quality of origins and destinations determination (in a sense of ability to identify the areas of socio-economic activity) over the traditional approach.
Future research will focus on the proposed concept improvement through exploration of the additional contextual and spatial relations between the Voronoi cell seeds (locations of base stations) and points of socio-economic activity (contextual data on shops, concert halls, residential areas etc.).

REFERENCE

[1] Arya, S, Malamatos, T, and Mount, D M. (2002). Space-Efficient Approximate Voronoi Diagrams. Proc of the 34th annual ACM symposium on Theory of computing, 721 – 730. Montreal, Canada.

[2] Aurenhammer, F, and Klein, R. (2000). Voronoi Diagrams. In Handbook of Computational Geometry, Chapter V, (J. Sack and G. Urrutia, editors), 201-290. Elsevier Science Publishing. Amsterdam, The Netherlands.

[3] Crawley, M J. (2013). The R Book (2nd ed). John Wiley & Sons. Chichester, UK.

[4] Filić, M, Filjar, R, and Vidovic, K. (2016). Graphical presentation of Origin-Destination matrix in R statistical environment. *Proc of KoREMA Automation in Transport 2016 Conference*, 26-30. Krapina, Croatia.

[5] Filjar, R, Filic, M, Lucic, A, Vidovic, K, and Saric, D. (2016). Anatomy of Origin-Destination Matrix Derived from GNSS Alternatives. *Coordinates*, **12**(10), 8-10.

[6] Filjar, R, Jezic, G, and Matijasevic, M. (2008). Location-Based Services: A Road Towards Situation Awareness. *J of Navigation,* **61**, 573 – 589.

[7] Florez, M et al. (2016). Measuring the impact of economic well being in commuting networks – A case study of Bogota, Colombia. Available at: http://bit.ly/2cVpihm, accessed on 20 March, 2017.

[8] Groff, E R, and McEwen, T. (2006). Visualization of Spatial Relationship in Mobility Research: A Primer (Final Report). Institute for Law and Justice. Alexandria, VA. Available at: http://bit.ly/2cdC9Li, accessed on 19 March, 2017.

[9] Gonzalez, M C, Hidalgo, C A, and Barabasi, A-L. (2008). Understanding individual human mobility patterns. Nature, **453**, 779-782. doi:10.1038/nature06958

[10] Peterson, A. (2007). The Origin-Destination Matrix Estimation Problem – Analysis and Computations (MSc thesis). Linkoeping University. Norrkoeping, Sweden.

[11] R Development Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3--900051--07--0. Available at http://www.r-project.org, accessed on 5 March, 2017.

[12] Toole JL, Herrera-Yaque C, Schneider C M, Gonzalez, M C. (2015). Coupling human mobility and social ties. *J. R. Soc. Interface* **12**: 20141128. http://dx.doi.org/10.1098/rsif.2014.1128. Available at: http://bit.ly/2d0GWBp, accessed on 10 March, 2017.

[13] Wang, P et al. (2012). Understanding Road Usage Patterns in Urban Areas. *Scientific Reports*, **2**, e01001. doi: 10.1038/srep01001.

[14] Zhang, D et al. (2015). UrbanCPS: A Cyber-Physical System based on Multi-source Big Infrastructure Data for Heterogeneous Model Integration. 6th ACM/IEEE International Conference on Cyber-Physical Systems (ICCPS'15), 2015. Available at: http://bit.ly/2bTEkWk, accessed on 1 March, 2017.