Socio-Economic Origin-Destination Matrix Derivation Through Contextualization of Material World

Ivana Stupar*, Petra Martinjak **, Vjera Turk***, Renato Filjar*
* Ericsson Nikola Tesla, Krapinska 45, Zagreb, Croatia
** Department of Mathematics, Faculty of Science, University of Zagreb, Croatia
*** Faculty of Engineering, University of Rijeka, Croatia
ivana.stupar@ericsson.com

***Abstract* - An origin-destination matrix contains data about individual and group mobility in a certain spatial area. As such, an origin-destination matrix represents an indicator of urban migrations and commuters' volume and can be used in strategic planning of the observed area. However, the origin-destination matrix does not contain information about the context of the migrations. In this paper we present an approach of expanding the origin-destination matrix data with the spatial data attributes containing information about the socio-economic context of activity's source and target derived from the material world. Origin-destination matrix enriched with such context can be used to gain a better understanding of socio-economic activities in an observed area, providing an insight to their nature and cause. Knowing the reason behind urban migrations can be useful for various socio-economic purposes. We describe developed information fusion algorithm, deploy the proposed concept, and present the graphical representation of the contextualized origin-destination matrix.**

Keywords – origin-destination matrix; ODM; socio-economic activities; contextualization; contextualized OD matrix

I. INTRODUCTION

An origin-destination (OD) matrix provides an objective and independent indication of mobility, related to both an individual and a group, in an observed area. It is often used in transport science as an information source for activity patterns assessment, and a tool for urban migration analysis and strategic planning. Traditionally, the OD matrix estimation is conducted by counting movement of individuals using human observers, video surveillance or other means. In recent developments, the utilization of various available datasets in the process of OD matrix estimation has been introduced. An example of such dataset is Charging Data Record (CDR) [1] used by telecommunication service providers. CDR is a collection of information about chargeable telecommunication events, e.g. performed phone calls, Short Messaging Services (SMS) and internet usage, used for billing purposes.

Here we propose a method for OD matrix estimation based on information extracted from CDRs and additional spatial data attributes extracted from OpenStreetMap [2], providing a novel framework for identification and estimation of the urban migrations due to socio-economic activities. OpenStreetMap (OSM) project is a collaborative initiative for creating a publicly available and editable map of the world. It provides information about spatial objects and relationships between them, following a topological data structure. Examples of object-related data obtainable from OSM include object's geographical position and type, i.e. object category. Object categories are related to the purpose of the object and include examples such as education (schools, universities), health (hospitals and other facilities providing healthcare services), work (company buildings, offices), and other religious, commercial, residential, and cultural facilities. Providing the information about the purpose of the origin and destination of a migration derives its socio-economic context. Hence, using this approach, it becomes possible to better understand daily and weekly urban migrations driven by socio-economic activity, which can be meaningful in identification of local regions of socio-economic attractions and urban strategic planning.

The remainder of this paper is organized as follows. Section 2 outlines the related work on the OD matrix estimation with the focus on estimations based on telecommunication activities datasets. Section 3 describes the methodology of our work. Results and the graphical representation of contextualized OD matrices can be found in Section 4, followed by conclusion and future work in Section 5.

II. RELATED WORK

The concept of enriching position (i.e. coordinates in a positioning system) with additional information, thus creating its context and transforming the position information into the location, has been described in [3]. One of recent commonly used approaches in creating information context is to utilize telecommunication activity data combined with other information sources for various purposes, such as emergency detection and management [4], estimation of human dynamics [5], regional delineation [6], or urban planning [7].

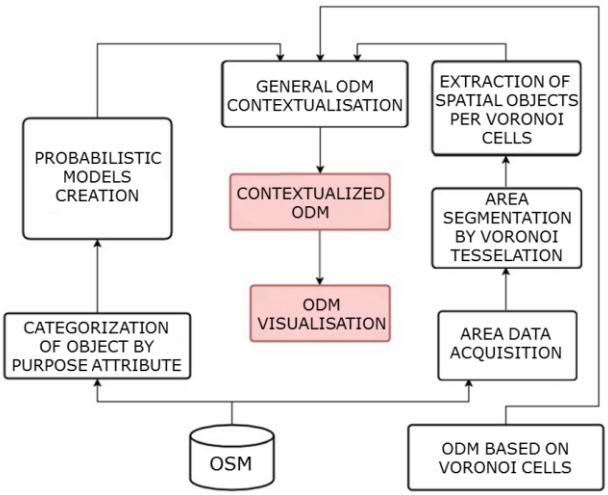

Fig. 1 ODM contextualization workflow

Several studies [4][8][9] introduce approaches of estimating OD matrices based on activities performed in a telecommunication network. Telecommunication network operators use data about chargeable user actions such as phone calls, SMS and internet usage, stored in a CDR. Based on CDR data, origins and destinations of an OD matrix can be determined such that areas of user activities are properly rendered [10] [11] [12].

As a base of our work we use OD matrix derived from telecommunications activity data as described in [13]. Authors of [13] use telecommunication activity data available in CDRs and match them with the spatial data to identify the origins and destinations in the process of OD matrix estimation, using the method of space partitioning based on Voronoi diagram called Voronoi tessellation. In our work, we expand the telecom-based OD matrix derivation described in [13] by providing the context of socio-economic activities derived from the OSM data, hence enabling better understanding of the socio-economic activity types in the observed region and the motivation behind them.

Fig. 2 OSM data and Voronoi tesellation -Shenzhen, China

III. Origin-Destination Matrix Contextualization

The proposed method of OD matrix contextualization described in this paper comprises the following tasks: acquisition of the area data and OD matrices derived from structured CDR subsets, area segmentation, spatial objects categorization and extraction, probabilistic model for migration flow contextualisation and graphical presentation, as outlined in Fig. 1.

A. Area Data Acquisition

In this work, we use an OD matrix estimated using structured CDR subset data collected in a mobile network in the Shenzhen metropolitan area [13]. The same CDR subset has been used for acquiring locations of base stations in the observed area. Every OD matrix is related to the predefined time-window. In this work we use OD matrices based on 3-hour time-window. The daily time span has been split into 8 time-windows: 0-3 h, 3-6 h, 6-9 h, 9-12 h, 12-15 h, 15-18 h, 18-21 h, 21-24 h thus estimating 8 OD matrices per day.

The area data acquisition necessarily precedes matrix contextualization algorithm development and deployment. The process consists of acquiring locations of base stations for area of interest and deriving spatial object data from the OSM database.

The set of base station locations used in this project originates from the CDR subset used for the OD matrix estimation which has been stripped off from the unnecessary data and stored in a Comma Separated Vale (CSV) format file. For each instance of a telecom activity in CDR data, an ordinal number of the activity was recorded, as well as latitude and longitude of the base station that detected activity. These records were also used for extraction of base station unique locations.

The OSM data describes spatial objects following a topological data structure, with four core elements defined as *nodes*, *ways*, *relations*, and *tags*. *Nodes* represent points with geographic position and are used to describe points of interest. *Tags* are key-value pairs which are used to store metadata about map objects (type, name, physical properties) and are always tied to a *node*, *way*, or *relation*. *Nodes*, *ways*, and *relations* are ranked hierarchically in the manner where *ways* consist of *nodes*, and *relations* consist of *ways* and *nodes*. In our work, *tags* assigned to *nodes* and *ways* were used to categorize objects described by *nodes* in purpose-based categories (described in detail in subsection *Purpose-Based Object Categorization*).

Due to the large area of interest (Shenzhen metropolitan area) and high number of objects in it, the spatial object data could not have been exported directly from the OSM database [14]. An algorithm was deployed with the purpose of extracting data related to smaller areas more convenient for data acquisition, eventually merging them into set of objects covering entire area of interest. A list of *nodes*, *ways* and *relations* was assigned to each area part, branched into specific purpose-based categories (e.g. home, education, health, etc.). *Nodes* were represented as a list of attributes (node ID, time of tagging, user, geolocation) and *tags*. Data from all area parts were

gathered and merged into single dataset representing all the OSM objects for the area of interest. In this step, we also acquired geolocations of base stations in the area of interest by retrieving them from the Open Cell ID database [15].

B. Area Segmentation

The OD matrices derived from CDR use space partition based on Voronoi cells that define base station areas of coverage [13]. This approach allows for accurate mapping of the contextual information from telecommunication network into spatial domain of socio-economic activities in the physical world. The purpose of this step is to determine which *nodes* lay in each of the cells, since nodes are the elements of the OSM data related to a single geolocation.

A data frame of spatial points was generated from the node geolocations, consisting of latitudes, longitudes, and nodes IDs. Such representation of *nodes* allowed separation of spatial data into smaller sets. We implemented spatial sub-setting using R [16] [17]. For each subset of spatial points, the corresponding *nodes* were retrieved from the initial set of data based on matching node IDs. This action results in the initial OSM-extracted dataset being split into a list of datasets matching the Voronoi cells. Results of the area segmentation for the city of Shenzhen are depicted in Fig. 2, containing both OSM objects and the Voronoi cells based on the tessellation method.

C. Purpose-Based Object Categorization

In *Area Data Acquisition* subsection, we presented the structure of data collected from OSM. The *tags* assigned to *nodes* and *ways* are used for the object categorization, along with geolocation of nodes. Objects are categorized by their purpose into six initial categories: *home*, *education*, *leisure*, *health*, *work* and *other*. The categories listed are also referred to as attributes. The category *other* comprises the objects whose purpose does not fall into any of the preceding categories but are still meaningful for the estimation of socio-economic activities (e.g. supermarkets, shops, amenities). Based on observation of tagging system used by OSM, each category is given a tag value that is used for filtering and categorization. Some categories (work, other) are given tag keys to categorize objects more specifically.

An additional method of object detection was deployed for *home* category due to an inconsistency of the OSM data tagging system. During the study of OSM data we observed that the tag *residential* is commonly used with *ways* that describe streets in residential suburbs. Since *ways* are used for representing linear features but do not have geographic position, they are represented with a list of *nodes* whose coordinates describe polyline of a *way*. The *nodes* parented by *ways* tagged *residential* were not tagged themselves, thus making them undetectable with method based on tags attached exclusively to *nodes*. For this category, a method of detecting nodes based on *tags* attached to their parent *ways* was deployed so that untagged *nodes* related to *ways* tagged *residential* were categorized as *home* objects.

D. Spatial Object Filtration Algorithm

As described in *Area Segmentation* subsection, the area of interest and OSM data are segmented per Voronoi cells, thus allowing for further categorization of objects by purpose. The following methods for filtering desirable nodes are applied on each Voronoi cell data: detection based on *tags* attached to *nodes* and detection of *nodes* based on *tag* values attached to their parent *way*.

The set of OSM data for each Voronoi cell is scanned for *nodes* whose *tags* correspond to specified category of purpose, as a part of detection based on *tags* attached to *nodes*. These *nodes* are counted and stored separately into list representing *nodes* distributed per purpose attributes for each Voronoi cell. Once counted in a specific category, nodes are removed from set of OSM data for the area to avoid the assignment of a single *node* to multiple categories, which was important for the process of probabilistic model creation.

Detection of *nodes* based on *tag* values attached to their parent *way* is only used for the *home* category, due to previously described specifics of tagging residential areas. The set of the OSM data for the current Voronoi cell is processed for ways matching tag *residential*. Given the method of OSM data distribution per Voronoi cell, there is a possibility of a single *way* appearing in two or more cells. Hence children *nodes* of residential *ways* are compared with *nodes* in OSM dataset for the current area and only *nodes* existing in the current Voronoi cell were taken as *nodes* belonging to *home* category.

Resulting from the application of the two methods presented, an attribute matrix is assembled describing objects distribution per Voronoi cell and object purpose. For n (where n is a natural number) Voronoi cells, the matrix contains n rows. The attributes, i.e. purpose categories, are represented by columns. The (i, j) element of attribute matrix defined in such manner represents number of objects in i^{th} Voronoi cell whose purpose is from j^{th} category. The attribute matrix servers as an input for the next step, probabilistic model creation.

E. Probabilistic Model Creation

In this section, the development of the probabilistic model for describing targeted objects within a Voronoi cell is presented, as well as its algorithmic representation.

Probability of migration flow segmentation on objects based on their purpose is determined by a number of objects for each category within a cell and time-window during which migrations occur. Time-windows have been chosen accordingly to OD matrices and following common practice of splitting daily span onto 8 time-windows of a 3 hours-duration. Considering each time-window, the attributes have been assigned the appropriate weight values, depending on the influence that the categories have on a socio-economic motivation of migrations. Weight values in a form of natural numbers on scale 1 to 10 have been assigned in naive manner, based on the identification of a nature of socio-economic movement thorough various time slots (e.g. at 6-9 h time-window people are more likely to visit health institutions than at 18-21 h time-slot, while the 15-18 h time-slot holds an increased likelihood of home objects being

destinations rather than workplaces). Using weight values and number of objects per cell for each category allows for deployment of weighted distribution of migration flow in a specific time-slot.

Let O be the origin cell and D destination cell. Furthermore, let $a_1,\ldots,a_6$ be the purpose attributes defined in previous subsection, and w_{ai} weight value assigned to attribute a_i.

For every attribute a_i the probability for a single migration being directed to object belonging to a_i is calculated and multiplied with weight value assigned to attribute a_i as expressed in (1):

$$f_{a_i} = \frac{n_{a_i}}{n} w_{a_i} \tag{1}$$

where n is the number of all objects in cell and n_{ai} is the number of objects in category a_i.

Weighted distribution of a migration flow from origin area O to destination area D is given as in (2):

$$P \sim \begin{pmatrix} a_1 & a_2 & a_3 & a_4 & a_5 & a_6 \\ p_{a_1} & p_{a_2} & p_{a_3} & p_{a_4} & p_{a_5} & p_{a_6} \end{pmatrix} \tag{2}$$

where p_{a1} is described with (3).

$$p_{a_1} = \frac{f_{a_i}}{\sum_{k=1}^{6} f_{a_k}} \tag{3}$$

The contextualization has been conducted for each original OD matrix separately. Every migration flow (stored in OD matrix cell at i^{th} row and j^{th} column) was distributed according to above mentioned distribution and stored as vector. Novel contextualized origin-destination matrix (COD matrix) is assembled as 3-dimensional array, where the first and the second dimensions are origin and destination from the original OD matrix respectively, and the third dimension is an attribute to which the migration flow is segmented. The algorithm for probabilistic model development described in this section was deployed as algorithm in an R-based script.

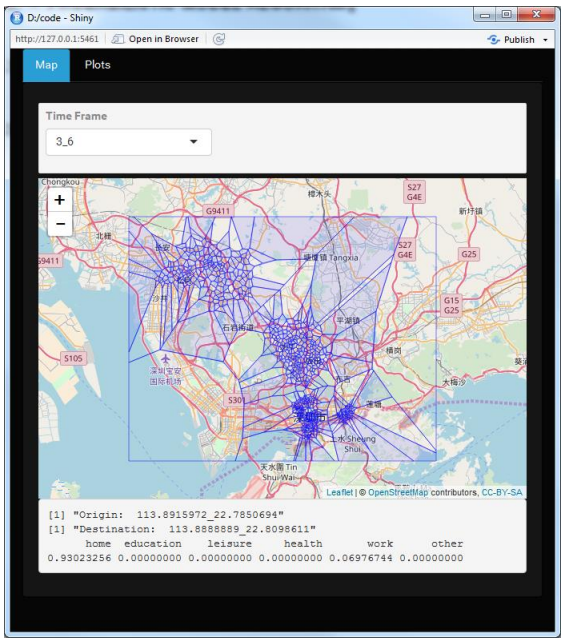

Fig. 3 Interactive map

IV. Graphical Representation of Results

As a graphical representation of OD matrix contextualization, we use several plots and diagrams. An interactive map (Fig. 3) was designed to show Voronoi cells on top of the OSM foundation layer. The map allows selection of the time-slot and the origin cell of interest, and displays possible destination cells based on the selection. After destination cell has been chosen, the spatial analysis results are displayed bellow map showing distribution of the initial migration flow.

Probability distribution of the incoming migration flow to a destination Voronoi cells was graphically represented in a bar plot (Fig. 4) depicting probabilities of a selected category being migrated to destination Voronoi cells. The bars represent Voronoi cells with every color corresponding to a specific attribute (i.e., purpose category). Probabilities of a migration flow being appointed to an object of certain category are shown on y-axis.

The OD matrix graphical presentation method we used in this work was inherited from [18]. The authors describe methodology for OD matrix graphical presentation, in which they represent the intensity of migration through the color gradient rather than numerically. In addition to the two-dimensional heat-map we introduce a surface plot (Fig. 5) where bar height represents the number of migrations. The intensity of color red reflects the number of migrations from i^{th} to j^{th} area of observation.

Both two-dimensional heat-maps and surface plots were generated for selected time window through the following procedure. We loaded the OD matrix and probabilities matrix for the selected time window and performed OD matrix contextualization according to the method described in Section 3, *Probabilistic Model Creation*. As the contextualized OD matrix contains category-related probability distributions, the two-dimensional heat-maps and surface plots were generated for each category of the contextualized OD matrix.

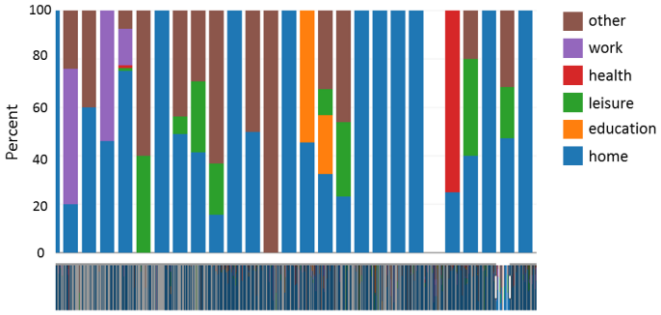

Fig. 4 Bar plot for probabilities in 9 to 12h time-window

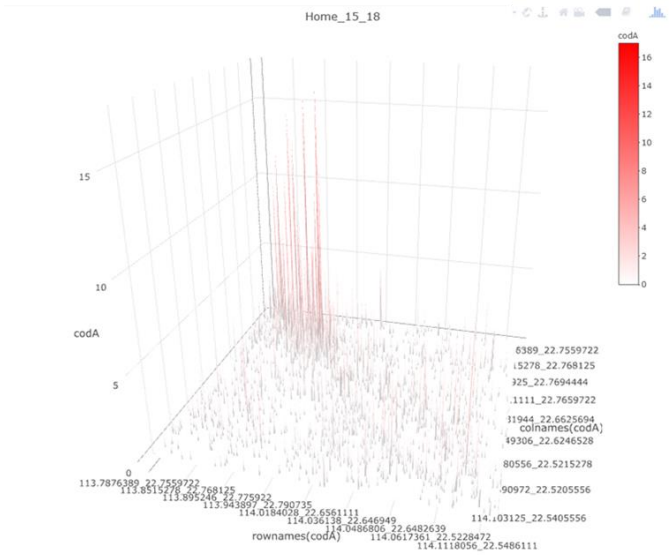

Fig. 5 Surface plot of *Home* OD matrix for 15 to 18h time-window

V. Conclusion and Future Work

In this paper we present an approach of providing socio-economic context to traditionally created OD matrices. We expanded CRD data used for OD matrix estimation with the spatial data attributes available from OpenStreetMap initiative, resulting in gaining more insight on the socio-economic motivation of the migration flows. Using this approach, it becomes possible to better understand reasons behind daily and weekly urban migrations driven by socio-economic activity, which can be meaningful in identification of local socio-economic attractions and urban strategic planning.

Further research will focus on relating additional data to the process of OD matrix derivation, as well as optimization of presented contextualized migration flow probabilistic model.

References

[1] 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects. (March 2018) Telecommunication management; Charging management; Charging Data Record (CDR) parameter description (Release 15)" 3GPP TS 32.298 V15.2.0.

[2] OpenStreetMap project, https://www.openstreetmap.org, accesed on 7 February, 2018

[3] Filjar, R, Jezic, G, Matijasevic, M. (2008). Location-Based Services: A Road Towards Situation Awareness. The Journal of Navigation, 61(4), 573-589.

[4] Pastor-Escuredo, D. et al. (2014, October). Flooding through the lens of mobile phone activity. In Global Humanitarian Technology Conference (GHTC), 2014 IEEE (pp. 279-286). IEEE.

[5] Šćepanović, S. et al. (2015). Mobile Phone Call Data as a Regional Socio-Economic Proxy Indicator. PLoS One, 10(4), e0124160. doi:10.1371/journal.pone.0124160. Available at: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0124160, accessed on 7 February, 2018.

[6] Ratti, C. et al. (2010). Redrawing the Map of Great Britain from a Network of Human Interactions. PLoS ONE 5(12): e14248., https://doi:10.1371/journal.pone.0014248. Available at: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0014248, accessed on 7 February, 2018.

[7] Alhazzani, M., et al. (2016). Urban Attractors: Discovering Patterns in Regions of Attraction in Cities. arXiv preprint arXiv:1701.08696.

[8] Gonzalez, M. C., Hidalgo, C. A., Barabasi, A. L. (2008). Understanding individual human mobility patterns. Nature, 453(7196), 779-782.

[9] Dong, Y. et al. (2015, September). Inferring unusual crowd events from mobile phone call detail records. In Joint European conference on machine learning and knowledge discovery in databases (pp. 474-492). Springer, Cham.

[10] Vidović, K., Lučić, A., Šarić, D. (2016). Methodology of O-D matrix determination from telecommunication charging data records. Proc of KoREMA Automation in Transport Conference, 26-30. Krapina, Hrvatska, and Ljubljana and Maribor, Slovenija, 2016.

[11] Filjar, R. et al. (2016). Anatomy of Origin- Destination Matrix derived from GNSS alternatives. Coordinates, 12(10), 8-10. Available at: http://mycoordinates.org/anatomy-of-origin-destination-matrix-derived-from-gnss-alternatives/, accessed on 7 February, 2018.

[12] Iqbal, M. S. et al. (2014). Development of origin–destination matrices using mobile phone call data. Transportation Research Part C: Emerging Technologies, 40, 63-74.

[13] Dešić, S., Filić, M., Filjar, R. (2017, May). Determination of origins and destinations for an OD matrix based on telecommunication activity records. In Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2017 40th International Convention on (pp. 424-427). IEEE.

[14] OpenStreetMap Data Extracts. Avialable at: http://download.geofabrik.de, accessed on 5 September, 2017.

[15] OpenCellid, Open Database of Cell Towers, https://opencellid.org/, acessed on 7 February, 2018

[16] Lansley, G., Cheshire, J. (2016). An Introduction to Spatial Data Analysis and Visualisation in R. University College of London (UCL)/Consumer Data Research Centre. London, UK. Available at: http://bit.ly/2rlCuEi, accessed on 15 September, 2017.

[17] Lovelace, R. et al. (2017). Introduction to visualising spatial data in R. Available at: http://bit.ly/1kFLrjz, accessed on 5 September, 2017.

[18] Filić, M, Filjar, R, Vidović, K. (2016). Graphical presentation of origin-destination matrix in R statistical environment. Proc of KoREMA Automation in Transport Conference, 26-30. Krapina, Hrvatska, and Ljubljana and Maribor, Slovenija, 2016.