

Article

# Uncovering the Relationship Between Human Connectivity Dynamics and Land Use

Olivera Novović <sup>1,\*</sup>, Sanja Brdar <sup>1</sup>, Minuđer Mesaroš <sup>2</sup>, Vladimir Crnojević <sup>1</sup>  
and Apostolos N. Papadopoulos <sup>3</sup>

<sup>1</sup> BioSense Institute, University of Novi Sad, 21000 Novi Sad, Serbia; sanja.brdar@biosense.rs (S.B.); crnojevic@biosense.rs (V.C.)

<sup>2</sup> Department of Geography, Tourism and Hotel Management, Faculty of Sciences, University of Novi Sad, 21000 Novi Sad, Serbia; minuđer.mesaros@dgt.uns.ac.rs

<sup>3</sup> School of Informatics, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece; papadopo@csd.auth.gr

\* Correspondence: novovic@biosense.rs

Received: 31 December 2019; Accepted: 17 February 2020; Published: 26 February 2020



**Abstract:** CDR (Call Detail Record) data are one type of mobile phone data collected by operators each time a user initiates/receives a phone call or sends/receives an sms. CDR data are a rich geo-referenced source of user behaviour information. In this work, we perform an analysis of CDR data for the city of Milan that originate from Telecom Italia Big Data Challenge. A set of graphs is generated from aggregated CDR data, where each node represents a centroid of an RBS (Radio Base Station) polygon, and each edge represents aggregated telecom traffic between two RBSs. To explore the community structure, we apply a modularity-based algorithm. Community structure between days is highly dynamic, with variations in number, size and spatial distribution. One general rule observed is that communities formed over the urban core of the city are small in size and prone to dynamic change in spatial distribution, while communities formed in the suburban areas are larger in size and more consistent with respect to their spatial distribution. To evaluate the dynamics of change in community structure between days, we introduced different graph based and spatial community properties which contain latent footprint of human dynamics. We created land use profiles for each RBS polygon based on the Copernicus Land Monitoring Service Urban Atlas data set to quantify the correlation and predictiveness of human dynamics properties based on land use. The results reveal a strong correlation between some properties and land use which motivated us to further explore this topic. The proposed methodology has been implemented in the programming language Scala inside the Apache Spark engine to support the most computationally intensive tasks and in Python using the rich portfolio of data analytics and machine learning libraries for the less demanding tasks.

**Keywords:** network analysis; mobile phone networks; human dynamics; big data; knowledge discovery

## 1. Introduction

Network analysis refers to the tools applied on network-based data towards the discovery of useful knowledge. The research area of network analysis enjoys widespread use, mainly because there are numerous and significant diverse applications that require the manipulation and analysis of network-based data, such as social network analysis, searching and mining the Web, pattern mining in bioinformatics and neuroscience.

In this paper, we use information collected from CDRs to generate a network, representing the communication traffic between different parts of the mobile network. This network is represented by a graph  $G(V, E)$ , where  $V$  is the set of nodes (vertices) and  $E$  is the set of edges (links).

There is a wide spectrum of network science applications in diverse fields. For smart city policy making and urban planning, it is essential to know how people move around and how they use urban spaces. Mobile phone data can be of great value for urban policy making as they contain valuable information about users' mobility and activity. The spatial semantics of the location are determined by their land use, but even so, people are using urban spaces in many different ways. The activity detected through mobile phone data is changing over different day types but also over different day times [1].

Urban spaces are also affected by daily base migrations of people who are commuting to work which is reflected in telecom traffic. Telecom Radio Base Stations are forming a latent network over urban area consisted of links created by users' connectivity. A network, mathematically speaking could be represented as a graph structure and powerful Graph Theory methods can be applied. We can track the dynamics and changes in the network by analyzing local and global properties of the graph that represent the network.

Despite the unprecedented value of user generated networks that reflects human dynamics many obstacles are present when analyzing such data. User generated networks could be made of telecom connectivity data, location-based social media data, GPS data from devices, etc. Those networks are highly dynamic and very complex while at the same time they demand efficient analytics that could deliver the result as fast as possible. To address this issue, high-performance computing frameworks must be incorporated in analytical pipelines. To be able to exploit the value of user generated data, we must know the specific context of the location or the event. The importance of the place in urban zones is determined by its land use. Densely built up urban zones with many diverse amenities are likely to form different telecom traffic footprint than sparsely populated industrial zones. Knowing the importance of the place defined by land use and human interactions formed around the place, we could strengthen transport and telecom infrastructure where needed, plan events better, enhance services, etc. By knowing the correlation between land use and human dynamics, we could predict the impact of land use change to people's lives.

**Contributions.** This work investigates the potential of using innovative data sources, such as telecom CDR data together with land use data, to evaluate and analyze human dynamics. Human Dynamics is a very wide research area that often requires a multidisciplinary approach. Mobile phone data analysis is one way to perceive Human Dynamics that can be reflected through telecom activity, connectivity and users' mobility. In this work, we explore the aspect of Human Dynamics reflected through mobile communication network in the form of connectivity links. The interplay between Human Dynamics and Network Science has been demonstrated before in [2]. Our work is based on the rationale that if centrality measures can be predicted by using land use properties, then what-if scenarios can be applied in order to detect the factors that affect human connectivity dynamics through mobile networks. More specifically, the contributions of our work are summarized briefly as follows:

- Shaw et al. pointed out several challenges in handling big and mobile human dynamics data in a hybrid physical–virtual environment where activities and interactions occurring in virtual space are not independent from activities and interactions in physical geographical space [3]. One example of such physical–virtual cohesion of the events is experimental study conducted by Graells-Garrido et al., where the authors evaluated the impact of the Pokemon Go game to the change of people's movement and mobility patterns in physical geographical space [4]. In this work, we are determining the correlation between physical geographical space characterized by land use and human dynamics virtual space created by users telecom interactions.
- Based on CDR data and communication traffic between different areas, a weighted network is constructed to represent the linkage between different city areas. Community detection is applied in this network in a daily basis, in order to detect community evolution across time. Communities are clusters in the network where high inner connectivity is present. The evolution of clusters through time is triggered by users' dynamics and their activity variations in space and time.

- We further analyze the dynamics of community structure by calculating spatial as well as structural (i.e., network-based) properties of communities. It turns out that land-use information is strongly related to spatial and structural properties of communities, as it is shown by a thorough performance evaluation.

Earlier studies indicate that there is great potential in using mobile phone activity data to predict land use [5,6]. Mobile phone data could be used as proxy for human dynamics, to the best of our knowledge there is no prior work on how to predict human connectivity dynamics reflected through telecom data based on land use. Such results are valuable for urban planning that becomes more challenging as urban population grows with expectation that two thirds of the world population will be living in the cities by 2050. Changes in land use across cities and further expansions should be accompanied with assessment of changes in human connectivity.

Our methodology involves the use of Big Data Analytics platforms. More specifically, the most computationally intensive parts of the pipeline are implemented in the Apache Spark engine. Therefore, our technique is scalable as it can handle potentially large amounts of data.

**Roadmap.** The rest of the paper is organized as follows. Section 2 presents related work in the area. Fundamental concepts and definitions are given in Section 3, whereas the proposed methodology is detailed in Section 4. System architecture is presented in Section 5. Experimental results are presented in Section 6 and finally Section 7 concludes our work and presents briefly interesting future research directions.

## 2. Related Work

With an increasing number of devices and services that collect data about people activity, interactions and mobility in the last few years, human dynamics research became the key topic in computational social science [7]. Human dynamics research evolves with changing every day circumstances in which people live such as natural and urban environment, emerging new technologies, climate change and society. High presence of modern information and communication (ICT) technologies including location-aware devices, various sensors and mobile technology in every day life have great impact on shaping human activity and interaction patterns [8].

One of the richest data sources about human daily based activities is mobile phone data [9]. Many diverse applications with significant social impact are developed based on mobile phone data, such as urban sensing and planning [10,11], traffic engineering [12–14], predicting energy consumption [15], disaster management [16–18], epidemiology [19–21], deriving socio-economical indicators [22,23].

More specifically, Ratti et al. in their review [24] highlighted the potential of using mobile phone data for urban planning. Soto et al. [25] used *Call Detail Records* to extract the information to automatically identify land use behaviors in urban environments. They used fuzzy c-means to cluster the Radio Base Station signatures and detect the class representatives of the land use in urban environment. Grauwin et al. used mobile phone data to detect land use classes in three different cities, New York, London and Hong Kong [26]. Furno et al. conducted comparative analysis between ten different cities, in which they constructed specific mobile traffic signatures to determine dynamic patterns of human presence in urban areas [27]. Rios and Muñoz used a big mobile phone data set with 880 million records in a case study for Santiago, Chile for land use pattern detection. They used the latent variable clustering technique in detecting clusters of residential, office area, leisure-commerce and rush hour pattern areas [28]. Pei et al. used hourly relative pattern and the total call volume through semi supervised fuzzy c-means clustering approach in inferring land use types in Singapore, showing that the accuracy decreased with the increase in heterogeneity of land use and density of cell phone towers [6]. Furno et al. combined simultaneously Call Detail Records and vehicle GPS traces for revealing land use context in French and Italian cities [29].

Unveiling complex ties between land use and human dynamics properties derived from mobile phone data is an active area of research. The latest results demonstrate relations between dominant land

use for each Voronoi zone and corresponding human activity represented as aggregated CDRs [30], as well as land-use composition of city's neighborhoods and the time series of CDR intensities [31]. Both studies utilized clustering to group similar land uses on one side and human dynamics properties on the other side and finally estimated agreement between clustering results obtained from this two data sources. Another novel study quantified by regression models how urban land use influences the commuting flows [32].

Noyman et al. conducted the study that suggests a methodology of “reversed urbanism” to urban planning and decision making. The methodology considers human behaviour patterns extracted from mobile phone data as a key element of urban design and their association with the functionality of urban areas [33]. One recent study conducted by Cottineau et al. showed the relation of mobile phone data indicators such as number of calls, active days, duration of calls, entropy, etc. and socioeconomic organization of cities [34]. They showed how mobile phone data together with census and administrative data could be used for urban development.

User-generated mobile phone data are highly dynamic and usually very large in size. Analytical pipelines over such data sets could be computationally expensive, while at the same time delivery of the results needs to be efficient since many applications require almost real-time response. Brdar et al. [35] provided a broad overview of the entire workflow starting from raw data access, followed by demands for analytical performance and data fusion, to the final application. They pointed out the critical challenges in mobile phone data analysis that need to be addressed, in order to disclose the hidden potential of the data. To make large scale telecom data analytics more efficient the work in [36] suggested using the Apache Spark [37] platform for distributed data analytics.

Our methodology differs from previous work in many aspects. First of all, to analyze the mobile network we are combining graph mining techniques [38,39] such as community detection and centrality measures. Next, we provide a mechanism in order to be able to predict spatial as well as network-based properties from land-use data. Last but not least, and following our previous work in [36], we have implemented the most demanding tasks inside the Apache Spark engine.

### 3. Fundamental Concepts

In this section, we present some important background information that is essential for the subsequent discussion. In particular, we cover topics related to community detection in networks, centrality measures and distributed data analytics. These techniques are used as building blocks in our methodology which is presented in detail in Section 4.

#### 3.1. Community Detection

A core concept used in our proposal is *community detection*, which involves associating the nodes of a network into meaningful groups (also known as clusters or communities). Given a graph  $G(V, E)$ , where  $V$  is the set of nodes and  $E$  is the set of edges, the output of a community detection algorithm, in its simplest form, is a partitioning of  $V$  into  $c$  groups  $V_1, V_2, \dots, V_c$  where  $\forall i, j$  it holds that  $V_i \cap V_j = \emptyset$ . This definition corresponds to *non-overlapping communities* which are utilized in this work.

To discover community structure in the network, we perform community detection using modularity based algorithms, such as the one proposed in [40]. The concept of modularity [41] presented by Equation (1), is used as a goodness measure for the quality of partitions, where  $A_{ij}$  is the weight of the edge connecting the  $i$ -th and the  $j$ -th node of the graph,  $\sum_j A_{ij}$  is the sum of the weights of the edges attached to the  $i$ -th node,  $c_i$  is the community where the  $i$ -th node is assigned to,  $m = (1/2) \sum_{i,j} A_{ij}$ , and  $\delta(i, j)$  is zero if nodes  $x$  and  $y$  are assigned to the same community and 1 otherwise.

$$Q = \frac{1}{2m} \sum_{i,j} \left[ A_{ij} - \frac{\sum_j A_{ij} \cdot \sum_i A_{ji}}{2m} \right] \delta(c_i, c_j) \quad (1)$$

Unfortunately, computing communities based on the maximization of the modularity, is an  $\mathcal{NP}$ -hard problem. To provide an efficient solution, the algorithm proposed in [40] uses an iterative process that involves shrinking the graph, every time modularity converges. In each phase, each node is assigned to a neighboring community that maximizes the modularity of the graph. As long as nodes are moving around communities and modularity grows, we keep on executing this process. When there are no more changes, a shrinking process is applied. Upon shrinking the graph, each community produced during the previous phase, it is assigned to the same *super node* of the new graph. The same process is applied to the new graph. The algorithm terminates when the modularity detected in the new graph is less than the modularity detected in the previous one. The set of communities that maximize the modularity is returned as an answer. The outline of the technique is depicted in Algorithm 1. It is evident that this algorithm may reach a local maximum. However, in general it performs very well, it is efficient and the quality of the generated communities is high.

---

**Algorithm 1:** LOUVAIN ( $G(V, E)$ )
 

---

**Input:** the graph  $G$

**Result:** the communities of  $G$

```

1  $n \leftarrow |V|$  /* number of graph nodes */
2  $done \leftarrow false$ 
3 while not  $done$  do
4   assign each  $u \in V$  to a different community
5   while there is a change do
6     for every node  $u \in V$  do
7        $C \leftarrow$  a community that maximizes modularity /*  $C$  is a neighboring community or  $u$ 's community */
8   if  $newModularity > oldModularity$  then
9      $G \leftarrow$  shrink graph based on communities /* each community becomes a super node in the new graph */
10  else
11    return communities

```

---

### 3.2. Centrality Measures

Node centrality measures [42] quantify the importance of graph nodes. Among the most widely used node centrality measures are:

- *Degree Centrality* (DC): the number of edges incident to the node;
- *Betweenness Centrality* (BC): the number of shortest paths passing through the node;
- *PageRank* (PR): related to the probability that a random walker will visit the node; and
- *Core Number* (CN): a value associated with a node that quantifies how well the node is connected with respect to its neighborhood (we will provide more details on this).

BC and PR are the most computationally intensive, whereas DC and CN require linear time to compute. More specifically, the complexity of CN in undirected and unweighted graphs is  $\mathcal{O}(n + m)$ , where  $n$  is the number of nodes and  $m$  is the number of edges.

**Definition 1** (core number). *The core number of node  $u$ ,  $CN(u)$ , is the maximum value of  $k$  such that  $u$  belongs to the  $k$ -core of the graph  $G$ .*

**Definition 2** ( $k$ -core). *The  $k$ -core of a graph  $G$  is the maximal subgraph  $H$  such that every node  $u$  in  $H$  has at least  $k$  neighbors in  $H$ , i.e., the degree of all nodes in  $H$  is at least  $k$ .*

The core decomposition concept has many applications in diverse scientific areas. For a broad coverage of the topic the interested reader is referred to the work in [43]. Here we will present this concept briefly and highlight its most important characteristics.

Let  $G(V, E)$  be a graph. Also let  $N(v)$  denote the set of neighbors of node  $v$ . Assume that the minimum degree of all nodes of  $V$  is 1. This means that the whole  $G$  is also the 1-core, since for any node  $u$ , the degree of  $u$  is at least 1. This is illustrated in Figure 1a. Assume now that we remove from  $G$  the two nodes with a degree of one. Figure 1b presents the 2-core of  $G$ , since all nodes in this subgraph have a degree at least 2. Finally, the 3-core of  $G$  is shown in Figure 1c. This is also the maximum core, since there is no subgraph of  $G$  such that all nodes have at least four (4) neighbors. The pseudocode of the core decomposition process is given in Algorithm 2.

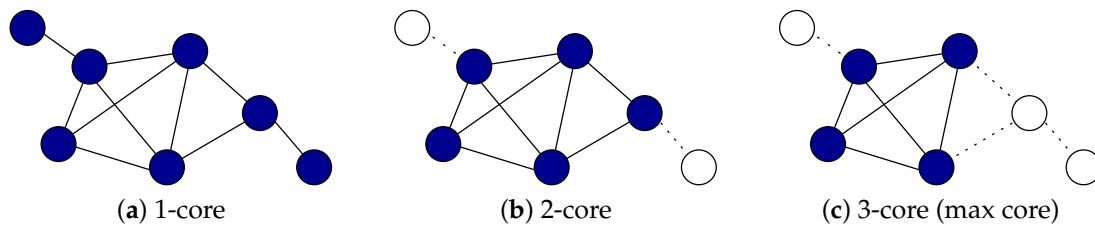


Figure 1. Core decomposition example.

---

**Algorithm 2:** COREDECOMPOSITION ( $G(V, E)$ )

---

**Input:** the graph  $G$

**Result:** the core numbers (array  $C$ )

```

1  $V \leftarrow$  set of vertices of  $G$ 
2 array  $D \leftarrow$  vertex degrees
3 sort array  $D$  in non-decreasing order
4 for each  $v \in V$  in the order do
5    $C[v] \leftarrow D[v]$ 
6   for each  $u \in N(v)$  do
7     if  $D[u] > D[v]$  then
8        $D[u] \leftarrow D[u] - 1$ 
9       reorder array  $D$  accordingly
10 return  $C$ 

```

---

The concept of  $k$ -core decomposition has been extended to other network types as well, such as directed, signed, weighted and many more. In the case of a weighted network, a straight-forward way to compute the core decomposition is to use the concept of the *weighted node degree*, which is defined as the sum of the weights of the edges incident to a node. However, the continuous space of weights is discretized, in order to achieve better runtime performance.

Despite its simplicity, the core number of a node  $u$  constitutes a very powerful centrality measure, which is easy to compute. It turns out that the core number can identify successfully important keywords in documents, influential nodes in social networks as well as dense subgraphs. In our case, nodes with high core number are most probably more important than others.

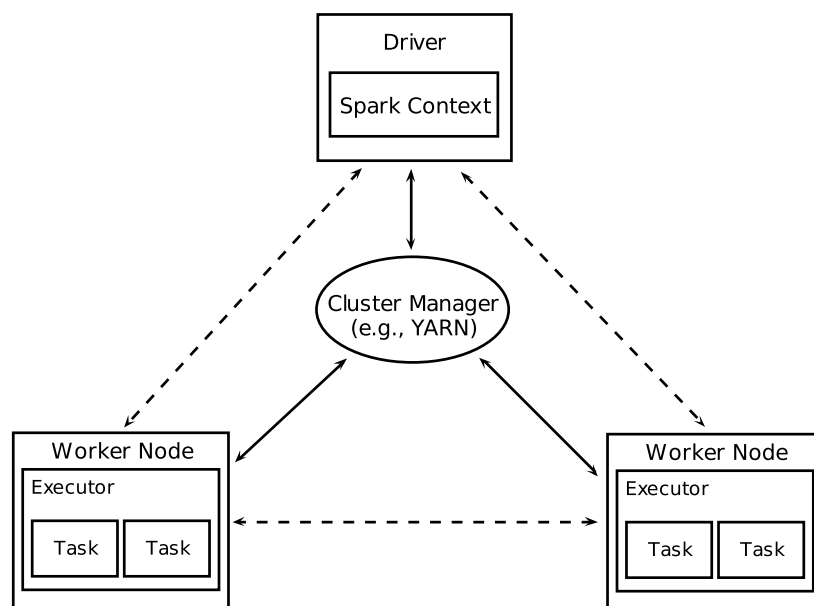
### 3.3. Distributed Data Analysis

One of the primary objectives of our proposal is to utilize technologies for big data analytics, in order to guarantee scalability. The most computationally intensive parts of our methodology are supported by Apache Spark. Apache Spark is a unified distributed engine with a rich and powerful API for Scala, Python, Java and R [37]. One of its main characteristics is that (in contrast to Hadoop



MapReduce) it exploits main memory as much as possible, being able to persist data across rounds to avoid unnecessary I/O operations. Spark applications are executed based on a master-slave model in cooperation with a cluster manager such as YARN (<http://hadoop.apache.org/>) or MESOS (<http://mesos.apache.org/>).

The basic Spark architecture is depicted in Figure 2. A Spark application consists of a *Driver* program that is responsible for executing the main function. The driver detects parts of the application that will be processed by *Worker Nodes* in parallel. A Spark application is composed of a number of Spark *Jobs*, and each job is defined as a sequence of *Transformations* ending by an *Action*.



**Figure 2.** Apache Spark architecture.

The Driver communicates with the *Resource Manager* (e.g., YARN or MESOS) in order to get access to cluster resources. The main execution component is the *Executor* which corresponds to a Java Virtual Machine (JVM) that runs on a Worker Node. Distributed processing is achieved by the communication between the Driver and the Executors. Spark Jobs are split into *Tasks* and each Spark Executor is responsible for the execution of one or more Tasks.

The main motivation for the use of Apache Spark is to provide efficient computation for large amounts of data. The methodology proposed in the paper is used as a proof-of-concept. However, the most important tasks are implemented inside the Apache Spark platform, meaning that the solution is highly scalable and supports larger data.

#### 4. Proposed Methodology

In this section, we present in detail our methodology. First we present the preprocessing of the raw data and then we proceed with the techniques applied towards knowledge discovery.

##### 4.1. Data Description and Preparation

Telecom data is very rich in user behaviour and therefore could reveal significant information about the personal profile of the user. Telecom operators follow rigorous procedures for data anonymization to preserve the privacy of users, before sharing the data to third parties. Call Detail Records (CDR) can be anonymized by performing temporal and/or spatial aggregation. CDRs can be temporally aggregated in predefined time slots in a way that all communication that occurred between two base stations is aggregated as one weighted link. Spatial aggregation is performed in order to hide the exact location of Radio Base Station where telecom traffic is distributed onto cells in regular grid

covering predefined spatial areas. Another data anonymization approach is to add noise to original data up to the level not affecting the statistics significantly while preserving the users' privacy [35].

In this study, we have used CDR data provided by the Semantics and Knowledge Innovation Lab (SKIL) of Telecom Italia [44]. The data refers to the area of Milan city and the surroundings and contains CDR data for a time period of two months, November and December 2013. The data is provided in the form of text files where each line represent aggregated telecom traffic that occurred between two square IDs for a given time interval. The Radio Base Stations coverage network is provided in the raster form, which we vectorized and approximated with adjacent Voronoi polygon network. First step in our processing pipeline is to import data into Apache Spark and transform it to DataFrame structure [45]. We decided to use Sparks' DataFrame structure to be able to exploit powerful SQL semantics while keeping the performance at high levels. The next step in our processing pipeline is to perform additional space/time data aggregation to obtain daily based snapshots of connectivity network across the observed area. Connectivity network is in the form of graph and methods for statistical filtering could be applied to remove the noise from the data. We used Disparity filter [46] as a method for extraction of the relevant connections in weighted networks, for distinguishing between strong and weak links in the graph. We have used Equation (2) which calculates the *link significance*,  $\alpha_{ij}$ , where  $\alpha$  denotes the significance *threshold*,  $p_{ij}$  is the probability of having a link between nodes  $i$  and  $j$ , and  $n$  is the number of nodes in the network.

$$\alpha_{ij} = 1 - (n - 1) \int_0^{p_{ij}} (1 - x)^{n-2} dx < \alpha \quad (2)$$

We note that smaller values of  $\alpha_{ij}$  denote more significant edges. Therefore, filtering is applied by keeping all edges where  $\alpha_{ij} \leq \alpha$  and thus removing all edges where  $\alpha_{ij} > \alpha$ . In our experiments, we applied filtering with probability 95% as it represent the less strict filtering, allowing us to keep more edges. The  $\alpha$  value used is 0.05.

For community detection we applied the LOUVAIN [40] modularity-based algorithm implemented in the Scala programming language, and run it in the Apache Spark engine. This algorithm is very efficient for large networks and also it does not require the number of communities as an input. The number of communities is determined by the algorithm. The LOUVAIN algorithm is applied for every single day, and the generated communities are used together with graph-based properties and lang-use profiles for knowledge discovery.

For spatial visualization of inspected communities we used QGIS software. We used the results of community detection to inspect spatial community properties. Spatial community properties that we calculated are mean intersection area, standard deviation of intersection area, average coverage area and diameter. Other properties used to evaluate human dynamics that we have explored in this study are graph based properties such as: betweenness centrality, weighted degree, PageRank and core number. Together with telecom data used to evaluate the community structure from connectivity network we used spatial open data that describe the spatial semantics of locations. One of the open spatial data sources is Copernicus Land Monitoring Service from which we have used Urban Atlas 2012 data set [47]. Urban Atlas provides detailed land use and land cover data for 800 Functional Urban Areas across Europe for the year 2012. For the spatial area of Milan city with the surrounding suburbs, there are 21 different land use classes present in the data, as shown in Figure 3. For each Voronoi polygon from the coverage area network we have created the profile vector of its land use classes.



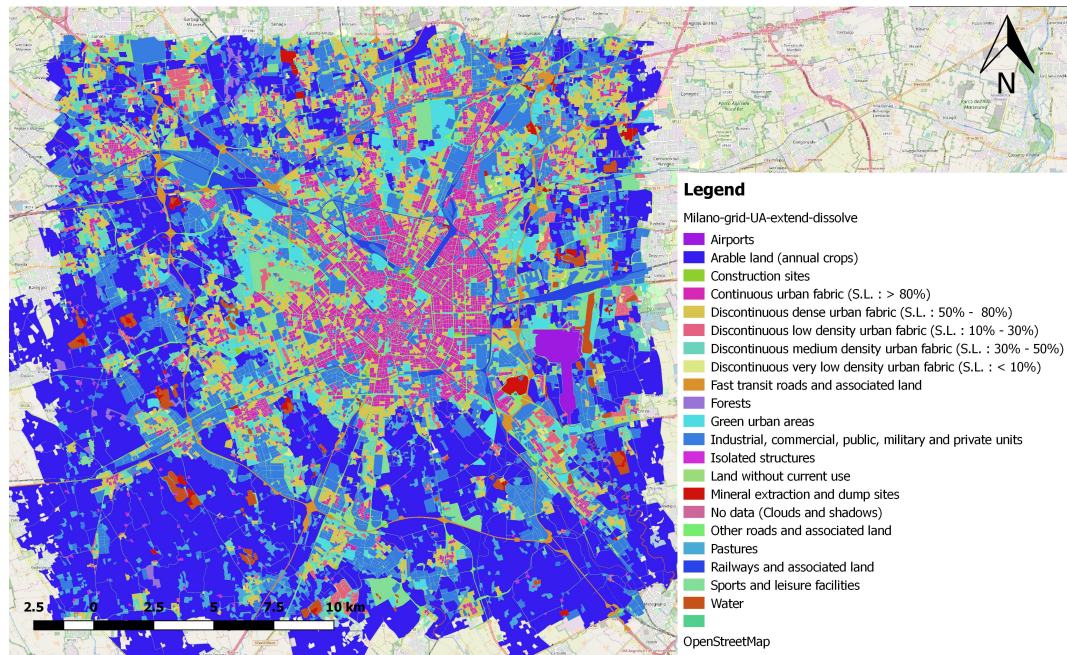


Figure 3. Land use classes over the extended spatial area of Milan.

We further investigated the relationship between land use and graph based and spatial community properties, which are related to human dynamics. We have used programming language Python and its rich portfolio of data analytics libraries to compute graph based and spatial community properties, which we further used to apply machine learning algorithms. The pipeline used is illustrated in Figure 4.

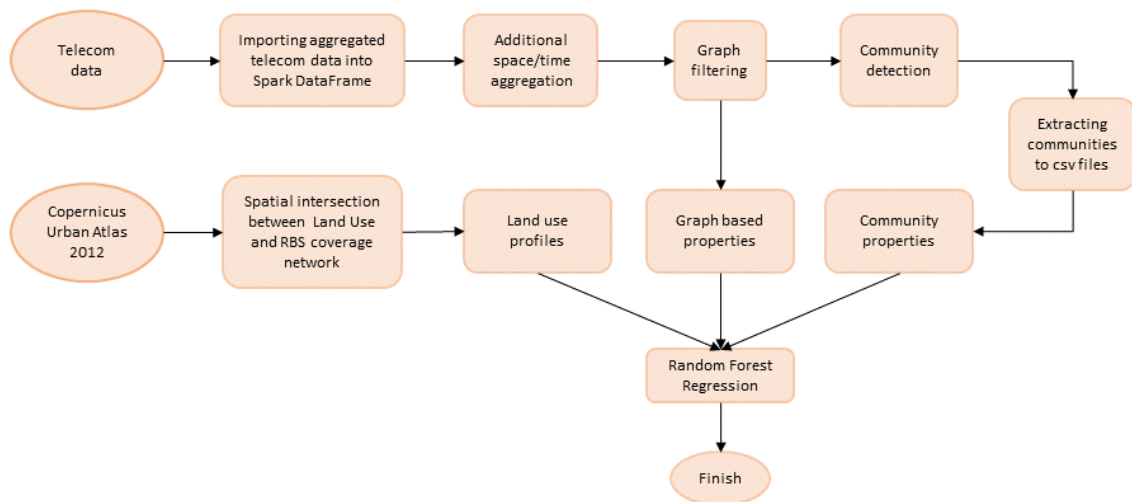


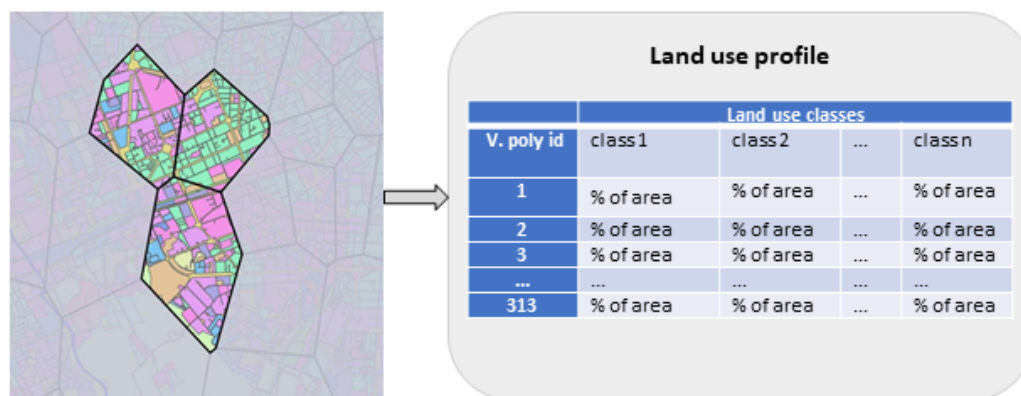
Figure 4. The pipeline of the proposed methodology.

#### 4.2. Feature and Properties Description

To investigate the dependence between land use data and human dynamics we calculated land use features and special properties related to human dynamics and run a predictive model over it. In this study, the spatial units are the Voronoi polygons generated from the RBS coverage network. For each Voronoi polygon we extracted the land use profile and we calculated local graph and spatial properties related to community structure.

### 4.3. Land Use Profiles

To generate a profile vector with land use types for each Voronoi polygon in the network, we first performed spatial intersection between RBS coverage area and Urban Atlas 2012 data layer [47]. The high resolution land use/land cover (LULC) Urban Atlas dataset was produced using Earth Observation data, road network datasets and topographic maps. For each LULC polygon feature resident population data was derived from various available sources, including GEOSTAT grid 2011, based on the 2011 census. The minimum mapping unit for all artificial surfaces is 0.25 hectares and 1 hectare for all the remaining LULC classes [48]. We extracted land use classes present in each polygon and calculated the percentage of area covered by the specific class. We further used those vector profiles as polygon based land use features. An example of intersection between three neighbouring Voronoi polygons and land use layers, and extraction of land use profile vectors is illustrated in Figure 5.



**Figure 5.** Example of three Voronoi polygon intersection with land use layer and Land use profile extraction.

### 4.4. Graph-Based Properties and Communities

Along with community detection we have performed a deeper analysis of connectivity graphs and quantified local graph properties. Local graph properties uncover localized patterns in the graphs, focusing on adjacent node neighbourhood [1]. We have evaluated different centrality measures, as described in Section 3, i.e., *betweenness centrality*, *node degree*, *PageRank* and *core number*.

For weighted networks *k*-core decomposition should consider both edge weight and node degree [49]. The *core number* of the node in our case represents the weighted degree of node of its maximal core. The average of all values for each node is used as a property. Among other centrality measures we calculated the *PageRank* of each node and used it as a property. The degree of a node is another important property that represents the number of incoming and/or outgoing edges connected to that node. The *weighted degree* of a node is calculated considering the edge weights connected to that node. All calculated properties are used as graph based properties associated to Voronoi polygons in the RBS coverage network.

For a time period of two months, from November to December 2013, we performed community detection over daily based connectivity graphs. Community formation is a highly dynamic process that depends on the connectivity network structure, and their spatial location, center, geographical spreading fluctuate between days. To quantify the spatial dynamics of community structure we introduced measures that reflect some aspects of the dynamics. Communities over inspected area are formed of single, very often neighbouring Voronoi polygons, and they are overlapping between days to some extent. For each polygon we detected and communities to which that polygon belongs to, we calculated the geographical intersection between sequential communities. *Mean intersection area* is the property that reflects the extent to which sequential communities are spatially overlapping. Another property related to spatial intersection is *standard deviation of intersection area*. This property

is introduced to evaluate the dispersion of the sequential intersection area values. Daily based communities vary in size and spatial distribution. We have also introduced *coverage area* as a measure of overall geographical reach of communities. We calculated the average of all coverage areas, and used it as a property. For each Voronoi polygon, we have computed the distances between the polygon center and the center of communities. The *center of a community* is defined as the geometrical center of the polygon that is created by joining all Voronoi polygons that are part of the community. The maximum of all distances, *the diameter*, is used as a measure of community center displacement from its starting Voronoi polygon center.

#### 4.5. Regression Models

We evaluated predictive power of land use profiles to produce predictions on graph based and spatial community properties based on Voronoi polygons with Random Forest [50] and Ridge regression [51] methods. Random Forest algorithm is an ensemble method based on many decision trees which predictions are averaged to form final output. From random forest model it is possible to derive feature importance [52]. Feature importance values sum to 1 and the higher values indicate higher importance. Alternative approach, Ridge regression was used due to its property that it overcomes the problem of multicollinearity amongst regression predictor variables.

### 5. System Architecture

In this section, we provide an overview of the HPC (High-performance computing) system architecture used to perform analytics.

To infer correlation between human dynamics and land use we used two input data sets. The first data set is about telecom connectivity that reflects human dynamics related to their activity and communication. The second data source consists of spatial data that contains land use classification of inspected area. Telecom connectivity data is provided in form of textual files, where each file represent daily base telecom traffic data aggregated in time slots of 10 min. The raw textual files contain many records which affects file size. For aggregation, graph filtering and community detection we decided to use Apache Spark platform to be able to process large files. We used powerful SQL semantic supported in Spark to transform the data and Apache Hive database to store the results. Community detection algorithm implementation is modified to read from Hive database and to iteratively update the clustering result until local maximum of modularity function is reached. To visualize the communities in geographical space we extracted the results from Hive database into csv format and loaded it into QGIS software. Land use data and RBS coverage network based on Voronoi polygons is provided in spatial vector format, and loaded into QGIS software to perform geographical intersection between layers. Result of intersection is spatial vector layer that contains land use profile for each Voronoi polygon. Land use profiles, communities extracted from Hive database and telecom connectivity graphs are further used for feature extraction in the pipeline part implemented in Python. We used Python's rich data analytics libraries to calculate features related to land use, graph and spatial properties. Human dynamics is reflected in the graph based and spatial properties, while land use profiles contain fine grained classification of land use types for each Voronoi polygon. A predictive model based on Random Forest and Ridge regression is used to evaluate the correlation and predictiveness of graph and spatial properties based on land use. The final result of the processing is detecting the dependence between land use and human dynamics reflected through graph based and spatial properties and evaluating the importance of specific land use class in the predictive model. The system architecture is illustrated in Figure 6.

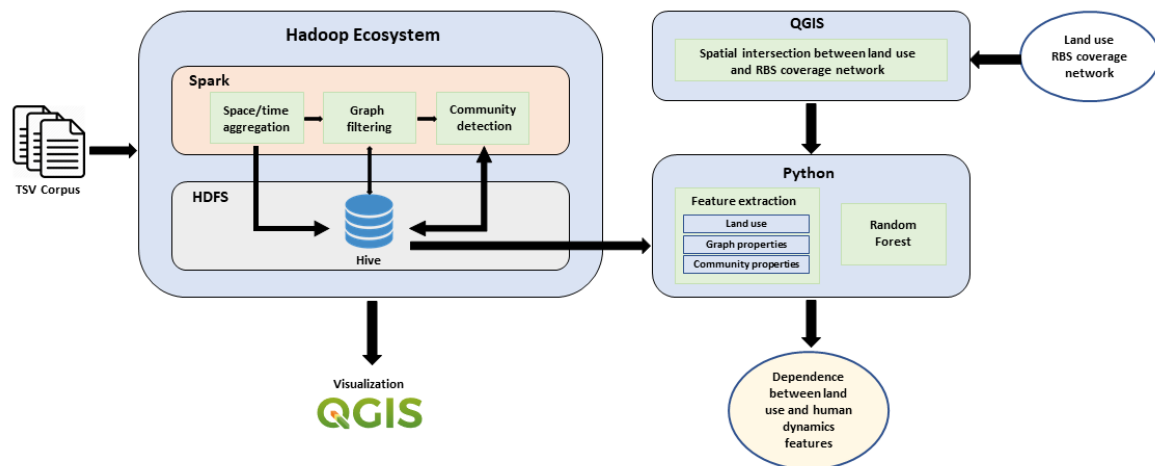


Figure 6. System architecture.

## 6. Results

In this section, we present some representative experimental results of our study.

The Telecommunication Radio Base Station (RBS) network consists of carefully designed and spatially distributed equipment used to provide the signal and to transfer telecom traffic. Each RBS is characterized with defined signal coverage, and in urban areas signal coverage is ubiquitous. In telecommunications, signal coverage is often determined with Voronoi polygons with antenna in the polygon center. In our connectivity network, telecom traffic is aggregated and distributed from one RBS to another. The connectivity network is modelled as graph, where base stations represent nodes and weighted links represent edges in the graph. We further applied graph filtering to keep only significant edges, and community detection to obtain community structure over inspected area.

For graph filtering, we applied the *Disparity filter*, which is a method for statistical filtering of weighted networks [46]. To select the appropriate  $\alpha$  value, we compared the results of community detection for the same graph when different filtering levels are applied. The results are presented in Table 1. From this table we observe a significant drop in the number of edges when filtering is applied, which is expected when using the Disparity filter. The number of detected communities remains stable when filtering with  $\alpha = 0.05$  is applied, while with smaller  $\alpha$  values a higher number of communities is observed. To evaluate how similar are the results when different filtering levels are applied, we computed the *Adjusted Random Index* (ARI) which is a common measure for evaluating similarity between clusters [53]. Higher values of ARI indicate more similar clusters. We considered the results of community detection performed over the unfiltered graph as the *ground truth*. The ARI between the clustering of the ground truth graph and the graph filtered with  $\alpha = 0.05$  is 0.83 which indicates very high similarity. When filtering is applied with  $\alpha = 0.01$  threshold, ARI drops slightly to 0.81 which also indicates a high similarity between clustering, while the number of detected communities increases 20% which is significant compared to the increase in communities when filtering with  $\alpha = 0.05$  is applied. With more strict filtering, when the  $\alpha$  threshold is set to 0.001, ARI drops to 0.71 and the number of detected communities increases for 38% compared to the ground truth. The aim of graph filtering before running community detection is to eliminate a large number of weak edges, in order to make the graph structure less dense and therefore to increase performance, while preserving network properties. Filtering with  $\alpha = 0.05$  threshold showed the best results, eliminating a significant number of edges, while the high value of ARI and the change in number of detected communities of only 4% indicate a very high similarity between clustering.

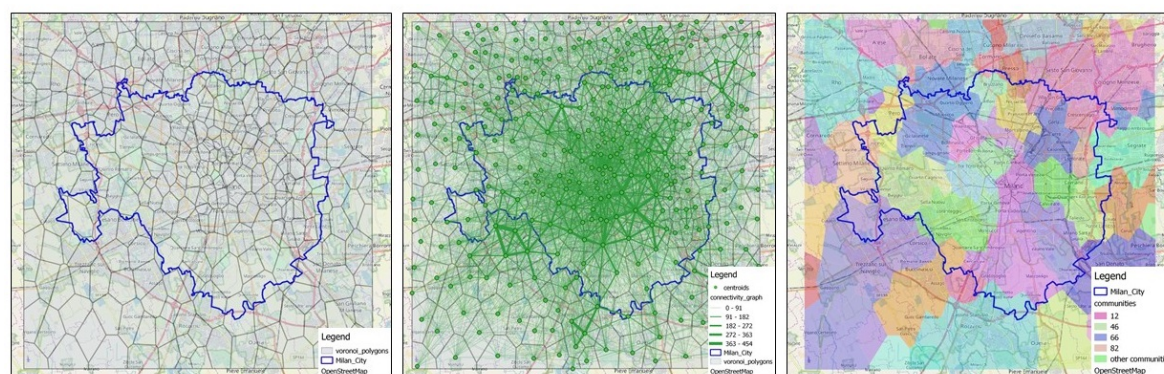


**Table 1.** Number of edges, communities and ARI when different filtering level is applied.

$\alpha$	Edges	Communities	ARI
1	95 860	53	-
0.05	6 322	55	0.83
0.01	3 809	64	0.81
0.001	2 323	73	0.71

We performed community detection over graphs filtered with  $\alpha = 0.05$  using Louvain algorithm to obtain snapshots of daily based community structure. Louvain algorithm starts with assigning each node to its own community and proceeds with moving nodes across communities to maximize the modularity function. Although the initial assignment of nodes might depend on data partitioning and sorting, Louvain algorithm showed very stable results between different iterations for the same graph. We run Louvain algorithm for community detection 10 times for each day and calculated ARI between different iterations. ARI between different iterations for the same input graph is always 1.0 indicating that the algorithm is very stable. The change in community structure between days is caused by change in input graph structure which reflect the dynamics of human connectivity. To evaluate the change in community structure between sequential days we calculated average ARI, which is 0.73.

The path from RBS coverage area network to communities detected based on connectivity network graphs is presented in Figure 7. On the first map, far left in Figure 7, the network coverage consisting of Voronoi polygons over the inspected area of Milan city with the surroundings is presented. On the middle map in Figure 7, the overlaid connectivity graph which is the input for community detection is presented. On the last map from Figure 7, the result of community detection for the input graph is presented.

**Figure 7.** From coverage area to telecom data graphs and communities.

The next step in our processing pipeline is to deeper investigate the time and space evolution of community structure. First we extracted communities from Hive database layer to common csv files and we visually examined the spatial distribution of communities using QGIS software. By visual inspection we noticed that although the form and number of communities differ between days some patterns are repeating. The granularity, size and spatial distribution of communities formed in densely populated, built up areas differ much compared to those of communities formed in peripheral parts of urban zone. High dynamics of community structure formed over urban core of the city can be explained by high dynamics of connectivity network over same area, reflected in input graph. The presence of high dynamics in community structure over urban core of the city can be explained by high dynamics of connectivity network over same area. In our previous work [1], we observed that core of the connectivity network, where the strongest links occur, is located over urban core of the city. Communities formed in the urban core of the city tend to be small in size and dynamically distributed

in space between days, while communities formed in less urbanized peripheral zones tend to form large, more stable structures. To quantify the observed phenomena, we introduce spatial *community properties* described in Section 4.

In our previous work [1], we detected the social event presence through analysing local and global graph properties of telecom connectivity networks. Those findings motivated us to compute the graph properties described in Section 4, to evaluate their correlation with land use and communities. To calculate community and graph properties, we used Python and its rich portfolio of libraries such as Pandas, NetworkX, NumPy, SciPy, Matplotlib, GeoPy, GeoPandas, etc. To evaluate the predictiveness of graph and community properties which contain latent footprint of human dynamics, based on land use we applied regression. Regression is applied sequentially over graph and community properties.

The results of regression are presented for the graph properties in Table 2 and for spatial community properties in Table 3. For all properties, Random Forest regression showed better results than Ridge regression. From Table 2, we observe that the predicted values of the property avgCN, i.e., *average core number* show the best correlation with real values, followed by property avgWND, i.e., *average weighted node degree* which is expected since both measures are considering weights over nodes edges which is important structural property of the network. This result is significant since it emphasize that the variability in core number and weighted node degree can be explained by the variability in the land use. Another property that shows strong correlation between predicted and real values considering Spearman coefficient is avgPR, i.e., *average PageRank*, while the other metrics, Pearson and R2, indicate a weaker correlation. The property that is the least associated with land use is avgBC, i.e., *average betweenness centrality*. From Table 3 we observe that the predicted values of the spatial community property avgCA, i.e., *average coverage area* show the best correlation with real values. Predicted values of spatial community property avgIA, i.e., *average intersection area* show fair correlation with real values considering Spearman coefficient, but show less correlation considering Pearson and R2 metric. Community properties that seem the least associated to land use are st.dev IA, i.e., *standard deviation intersection area* and avgDiam, i.e., *average diameter*. Overall Spearman correlation test shows the best results for all properties except for avgCN, avgDiam and st.dev IA Ridge regression where the Pearson correlation was higher. Spearman correlation test is expected to show robust results in the most of the cases since it does not carry any assumptions about the distribution of the data. All correlation results are significant at 0.01 (all  $p$ -values < 0.006).

**Table 2.** Prediction results for structural properties.

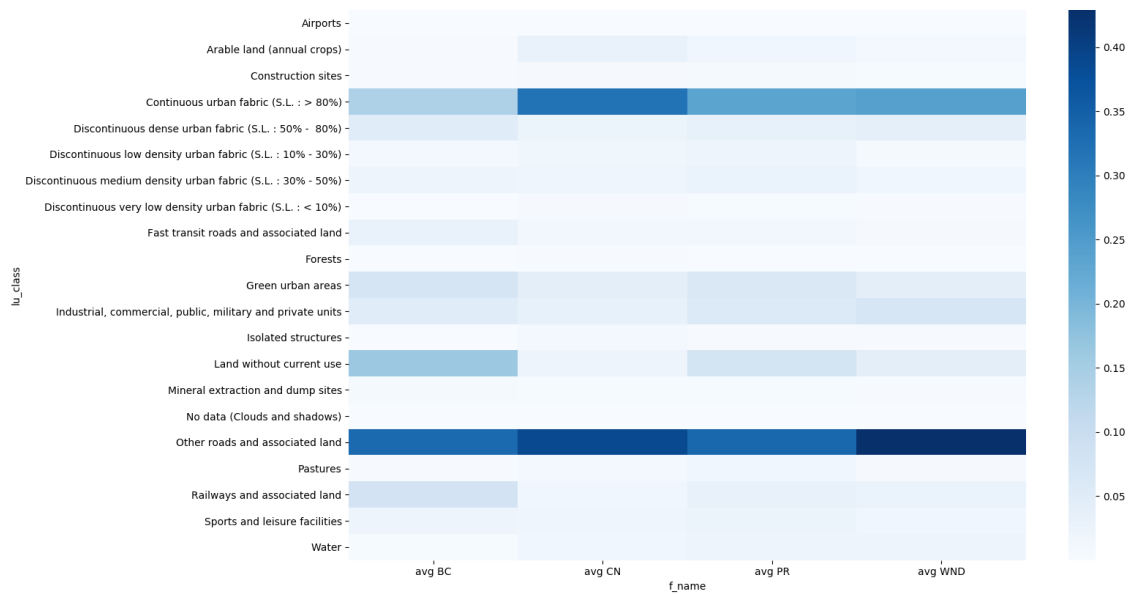
Metric	avg BC		avg WND		avg PR		avg CN	
	RF	Ridge	RF	Ridge	RF	Ridge	RF	Ridge
R2	0.111	−0.0192	0.422	0.424	0.310	0.321	0.628	0.623
Pearson	0.357	0.163	0.659	0.653	0.581	0.574	0.793	0.790
Spearman	0.269	0.289	0.776	0.707	0.715	0.644	0.778	0.753

**Table 3.** Prediction results for spatial community properties.

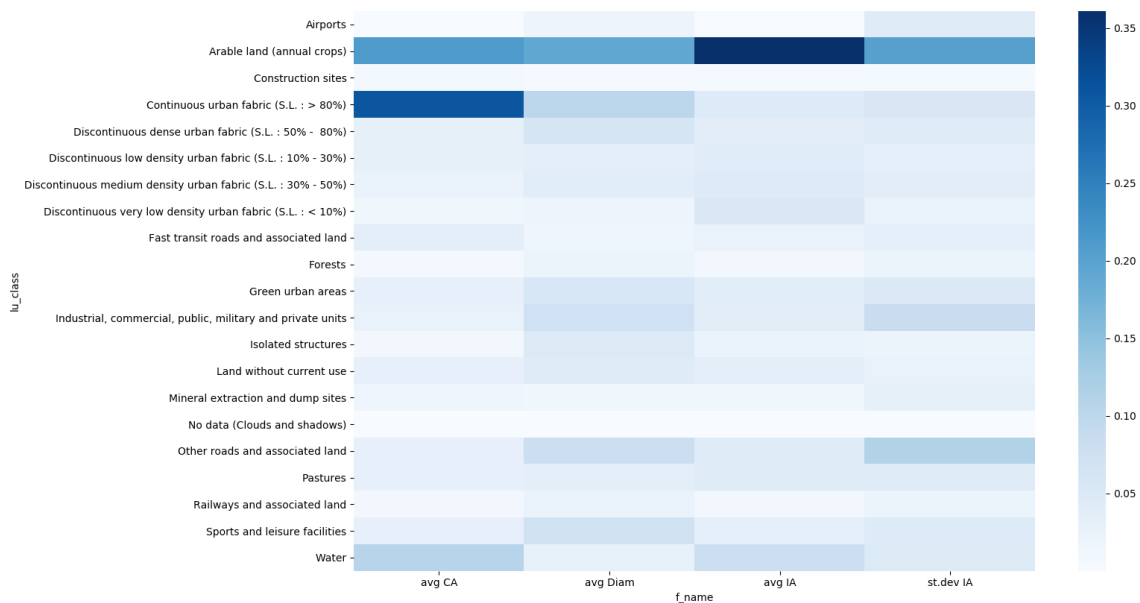
Metric	avg IA		st.dev IA		avg CA		avg Diam	
	RF	Ridge	RF	Ridge	RF	Ridge	RF	Ridge
R2	0.353	0.196	0.236	0.109	0.521	0.410	0.211	0.140
Pearson	0.596	0.485	0.488	0.385	0.722	0.660	0.474	0.403
Spearman	0.630	0.579	0.490	0.382	0.727	0.692	0.414	0.335

To explore how each specific land use class is important for prediction, we calculated feature importances. The importance of land use classes for predicting each graph based and community property is presented in heatmaps in Figures 8 and 9.





**Figure 8.** Importance of land use classes for predicting graph based properties.



**Figure 9.** Importance of land use classes for predicting community properties.

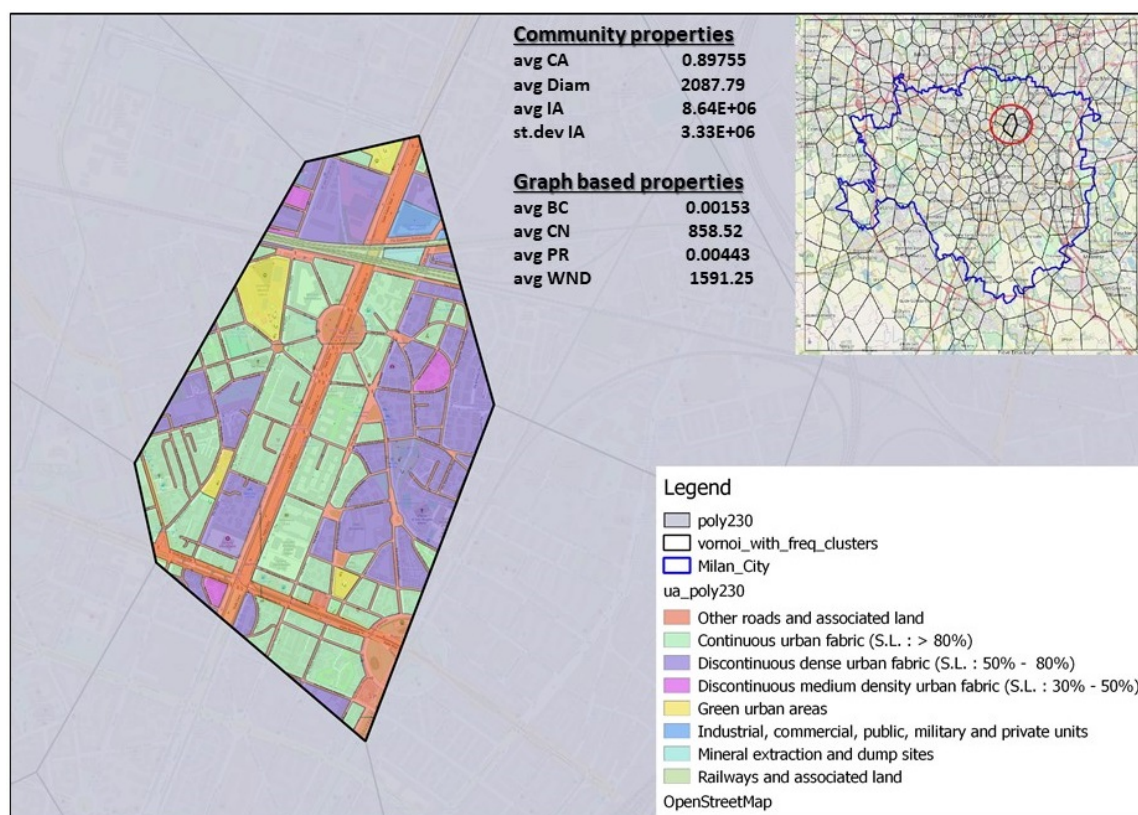
For predicting the graph based properties based on land use, the most important classes are *Other roads and associated land*, *Continuous urban fabric (S.L.: > 80%)*, as shown in Figure 8. The highest predictive power has land use class *Other roads and associated land* for the property *average weighted node degree* (avgWND). For predicting the property *average core number* (avgCN) the most important land use classes are *Other roads and associated land* and *Continuous urban fabric (S.L.: > 80%)*, which also have the highest impact on predicting the property *average PageRank* (avgPR). Land use has the least predictive power for the property *average betweenness centrality*, which is shown in Table 2 with small values of correlation metrics. From the heatmap in Figure 8 we can notice that only land use class *Other roads and associated land* has impact on predicting the property *average betweenness centrality*, but due to small values of correlation metrics we can conclude that impact is not significant.

For predicting the community properties based on land use, the most important classes are *Arable land*, *Continuous urban fabric (S.L.: > 80%)*, as shown in Figure 9. Land use class *Arable land* has the

most impact on predicting the community property *average intersection area* (avg IA), which can be observed from the heatmap in Figure 9. The most predictive community property based on land use is *average coverage area* (avg CA), where the most impact have land use classes *Continuous urban fabric* (S.L.: > 80%) and *Arable land*. From the heatmap in Figure 9 we can notice that only land use class *Arable land* has the impact on predicting the community properties *average diameter* (avg Diam) and *standard deviation intersection area* (st.dev IA), but from the Table 3 we can observe small values of correlation metrics which indicate that the impact is not significant.

We investigated deeper the urban profiles to detect polygons from network coverage area where the classes of high importance for predicting properties are dominant in area.

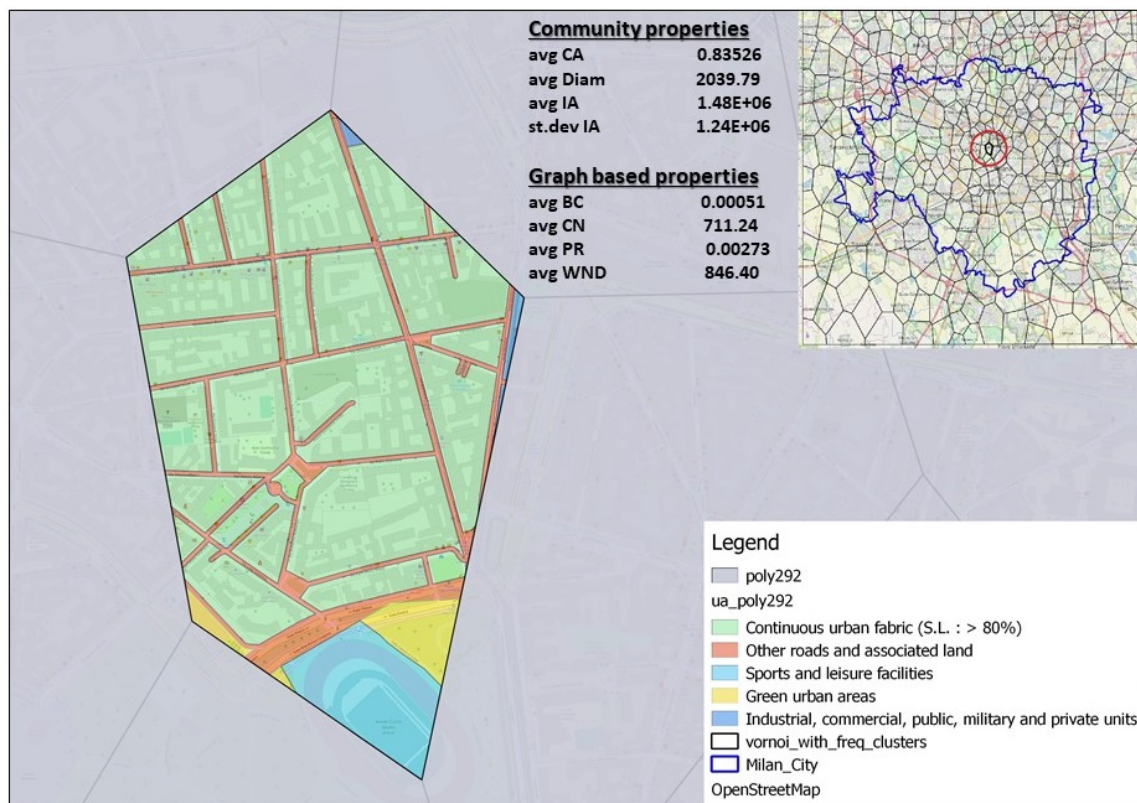
In Figure 10 is presented network coverage polygon where the dominant land use class is *Other roads and associated land* together with its graph and community spatial properties and its geographical position in the network.



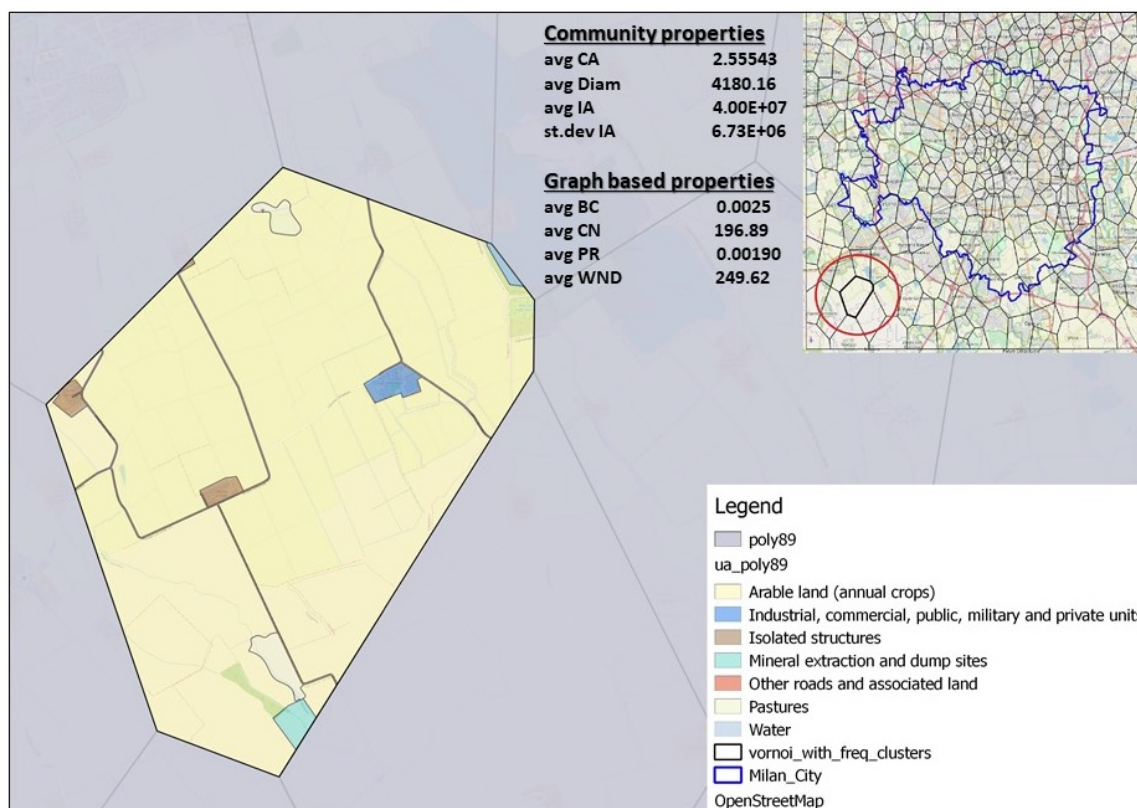
**Figure 10.** Land use profile of polygon where the dominant class is *Other roads and associated land*.

In Figure 11 is presented network coverage polygon where the dominant land use class is *Continuous urban fabric* (S.L.: > 80%) together with its graph and community spatial properties and its geographical position in the network.

In Figure 12 is presented network coverage polygon where the dominant land use class is *Arable land* together with its graph and community spatial properties and its geographical position in the network.



**Figure 11.** Land use profile of polygon where the dominant class is *Continuous urban fabric* (S.L.: > 80%).



**Figure 12.** Land use profile of polygon where the dominant class is *Arable land*.

By observing those three special cases presented in Figures 10–12 with specific dominant land use class we can notice some trends related to spatial community and graph properties. Polygon in the



highly urban area where the dominant class is *Continuous urban fabric* (S.L.: > 80%) is characterized with lower *mean intersection area* than polygons where the most dominant classes are *Other roads and associated land* and *Arable land*. This property indicates that the observed polygon forms highly dynamic communities, which are distributed in space in diverse manner and therefore the spatial intersection between communities is small. On contrary, high value of *mean intersection area* indicates the presence of more stable community structure that is not prone to dynamical change. The property *st.dev. intersection area* shows similar behaviour and indicates the same characteristics. Community properties *average coverage area* and *diameter* indicates the maximal spatial reach of all communities formed including observed polygon. High value of *average coverage area* indicates that formed communities are large in size, covering significant spatial area. This behaviour can occur due to initial size and/or granularity of Voronoi polygons contained in the communities, where we can have fewer very large polygons in the community or many smaller ones. Larger *average coverage area* of the community indicates wider reach of information spread inside community. High value of *diameter* indicate larger and wider reach of the communities, but only as a linear measure and therefore shows not significant correlation with land use which is area feature.

Graph properties of polygons show diverse behaviour depending on the dominant land use class. The property *average betweenness centrality* gains the highest value for polygons with dominant land use class *Arable land*. This can be explained with the geographical position of the polygon in the network and network dynamics. Polygons that are located far from urban core of the city may act as bridges between distinct parts of the network and therefore gain higher value of *betweenness centrality*. Also, if the network around specific polygon is changing fast day by day, it is more likely that the observed polygon will have less impact to network connectivity between days. The property *average PageRank* has the highest value for the polygon where the dominant land use class is *Other roads and associated land*. It is interesting to notice how physical transitivity of a node, i.e., polygon in the network characterized by land use class *Other roads and associated land* is reflected to its communication transitivity defined by *average PageRank* property.

The value of *average PageRank* decreases for the polygons where transit infrastructure defined with land use class *Other roads and associated land* is less present, as shown in Figures 11 and 12. The property that is the most predictive based on land use is *average core number*, and it has the highest value for the polygon where the dominant land use classes are *Other roads and associated land* and *Continuous urban fabric* (S.L. : > 80%), as shown in Figure 10. Polygons located in the urban core of the city where the dominant land use classes are *Continuous urban fabric* (S.L. : > 80%) and *Other roads and associated land* also show high value for the property *average core number*. In contrast, polygons where the most dominant land use class is *Arable land* and classes *Continuous urban fabric* (S.L. : > 80%) and *Other roads and associated land* are not significantly present show very small values of *average core number*. We can conclude that both classes *Continuous urban fabric* (S.L. : > 80%) and *Other roads and associated land* have a significant impact on the value of the property *average core number*, which is also confirmed by the algorithm when calculating feature importance, as shown in Figure 8.

The property *average weighted degree* shows similar trends as *average PageRank* and *average core number*, which is expected since those properties are correlated. Even though those values are highly correlated, they represent different very important graph properties related to connectivity and transitivity and we considered them independently in relation to land use profiles. Correlation between centrality measures is commonly observed in many different networks, but as they represent different property they might have diverse impact on information flow through the network [54].

Polygons that have the highest value of *average weighted degree* are the ones where the dominant land use class is *Other roads and associated land*. Polygons that have the dominant class *Continuous urban fabric* (S.L. : > 80%) also have a high value of *average weighted degree* but significantly less than the polygons with dominant class *Other roads and associated land*, as shown in Figure 11. Polygons with the dominant class *Arable land*, Figure 12, have very small *average weighted degree*. We can conclude that the land use classes that have the most impact on *average weighted degree* property are *Other roads and*

*associated land* and *Continuous urban fabric* (S.L. : > 80%) which is also detected by the algorithm when calculating feature importance, as shown in Figure 8.

## 7. Conclusions and Future Research

Human dynamics in geography and computational social sciences is related to human movement and mobility, but there are other aspects of human dynamics that need to be considered. Virtual activity reflected through social media, location-based services, telecommunications, internet usage, even some online games can have severe impact on shaping the patterns of human dynamics. The relationship between physical and virtual space is indisputable which adds even more complexity to human dynamics research.

In this study, we have analyzed human dynamics reflected by telecom traffic network through connectivity links. User generated data is usually very large in size, dynamic and it has complex structure. To analyze such data we have designed and developed a processing pipeline based on Apache Spark Big Data platform using programming language Scala. We have introduced graph filtering as a method for performance enhancement by keeping only statistically significant links in the graph.

To investigate community structure of the network we applied modularity-based algorithm for community detection, to investigate the space-time network evolution. We analyzed the community structure in a daily basis and observed some persistent patterns over time although the number, structure and spatial distribution of communities differs between days. We also noticed that communities formed over the urban core of the city are smaller in size but highly dynamic while communities formed over peripheral parts of the city and in sub-urban zones are larger in size and more stable. This observation motivated us to deeper investigate the correlation between land use and community structure. We have created land use profiles for each polygon of the RBS coverage network and quantified graph based and spatial community properties, which contain latent information about human dynamics. We further used the properties to learn the predictive model and to evaluate the correlation and predictiveness of properties based on land use.

Our results have shown strong correlation between land use and the properties *average core number*, *average coverage area*, fair correlation for the properties *average weighted degree*, *average page rank* and *average intersection area*, weak correlation for properties *average diameter*, *standard deviation of intersection area* and almost no correlation for property *average betweenness centrality*. To deeper investigate the impact of specific land use classes on properties predictiveness we have calculated feature importance.

Based on the results, for predicting all properties, the land use classes *Other roads and associated land*, *Continuous urban fabric* (S.L. : > 80%) and *Arable land* have the highest impact. The land use class *Other roads and associated land* has the highest impact on predicting all graph based properties. For predicting the property *average core number*, which shows the highest correlation to land use, the land use classes *Other roads and associated land* and *Continuous urban fabric* (S.L.: > 80%) have the highest impact. This can be explained by considering the concept of importance, both in physical space and in networks. In physical geographical space, the important urban zone is the one with many amenities, densely built up zone with developed transit infrastructure and good connections. Such zones would be characterized with high presence of land use classes *Continuous urban fabric* (S.L. > 80%) and *Other roads and associated land*.

In network science, graph centrality measures are used to quantify the importance of a node. Therefore a node with high importance would participate in the high degree core, which means its core number would be high. The results of feature importance calculation indicate that highly important zones in physical geographical space would also have an important role in the virtual telecom traffic network. Similar behaviour could refer to the property *average weighted degree*, its expected that highly urban zones would reflect to network nodes with high weighted degree. For predicting the property *average page rank* the most important land use class is *Other roads and associated land*. This can be explained by considering the concept of transitivity. Transitivity of the urban zone is reflected in

the prevalence of important roads, public transport and other transport infrastructure, while the transitivity of a node in the network is associated with its page rank property. Our results indicate that transitivity in physical space is linked with transitivity in the telecom connectivity network. The results of this study lead us to the conclusion that physical geographical space and virtual communication space need to be considered together as one entity.

To evaluate space–time dynamics of communities we have introduced spatial community properties. Spatial community property that shows the highest correlation with land use is *average coverage area*. To predicting the property *average coverage area* the most impact has the land use class *Continuous urban fabric* (S.L. : > 80%) and *Arable land*. Based on the feature importance result both classes have significant impact, although they predict opposite values. Polygons where dominant land use class is *Arable land* tend to have very high coverage area, while those where the dominant land use class is *Continuous urban fabric* (S.L. : > 80%) form smaller coverage area. This can be explained by observing the size and granularity of the neighbouring Voronoi polygons in the highly urban and non urban area. In highly urban areas, communities are formed by many small polygons, while in non urban zones communities are formed by few very large polygons, affecting the area of the community. The property that is the most related to spatial dynamic of community is *mean intersection area*. If a polygon is forming a stable community structure that does not differ much between days, its *mean intersection area* would be high comparing to polygons which form dynamic community structure with little or non spatial overlapping between days. The land use class *Arable land* has the most impact on predicting *mean intersection area*, but it is interesting to notice that the land use classes associated with highly urban zones such as *Continuous urban fabric* (S.L. : > 80%) and *Other roads and associated land* do not show significant predictive power. Based on the results, the land use class *Arable land* is very predictive for the property *mean intersection area* and it indicates high values of the property, but the opposite pattern is not observed. Land use classes that are associated with highly urban zones, *Continuous urban fabric* (S.L. : > 80%) and *Other roads and associated land* do not seem to have any significant impact on predicting the value of the property *mean intersection area*, neither high or low. This imply that even in the highly urban zones it is possible to form stable community structure that would persist in size and spatial distribution over time.

**Future work.** In this study, we evaluated the correlation and predictive power of land use features for predicting graph based and spatial properties of communities, derived from the telecom connectivity network. Graph-based and spatial properties reflect human dynamics related to their communication patterns. There are several interesting directions for future work, such as: (i) adding day type dimension to the analytics by distinguishing between working days, weekends and holidays, (ii) a deeper investigation of communication patterns associated with specific land use classes, and (iii) developing all the steps of analytical pipeline inside Apache Spark to have unified HPC framework, without the need to of additional processing tools such as Python.

The results of this study could be beneficial for urban planning and city policy making since it emphasize the correlation between land use and human dynamics reflected through properties evaluated from telecom connectivity network. Although human dynamics is often considered through the lenses of movement and mobility patterns, communication as one of the main aspects of human activity shouldn't be neglected when analysing human behaviour patterns. Telco providers are facing many challenges related to network infrastructure management, demands of new services which needs to be supported by the infrastructure, etc. With knowing the importance of the place, its transitivity and dynamics telco providers could enhance and optimize the network infrastructure and services they are providing.

The topic of human dynamics has been studied by many researchers from diverse disciplines over years. In the era where technology is present in our every day life more then ever, when data records are generated by almost every action we take, it is a huge challenge for the research community to design and develop methods to explore human dynamics which will help us answer essential questions about urban development and society in general.



**Author Contributions:** Conceptualization, Olivera Novović, Sanja Brdar and Apostolos N. Papadopoulos; methodology, all authors; software, Olivera Novović, Sanja Brdar and Apostolos N. Papadopoulos; validation, Olivera Novović and Minučer Mesaroš; formal analysis, Minučer Mesaroš; writing—original draft preparation, all authors; visualization, Olivera Novović; supervision, Vladimir Crnojević and Apostolos N. Papadopoulos; project administration, Sanja Brdar; funding acquisition, Vladimir Crnojević. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by The Ministry of Education, Science and Technological Development of the Republic of Serbia, Grant No. III 44006.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

- Novović, O.; Brdar, S.; Crnojević, V. Evolving connectivity graphs in mobile phone data. In Proceedings of the Main Conference on the Scientific Analysis of Mobile Phone Datasets (NetMob 2015), Boston, MA, USA, 8–10 April 2015; pp. 73–75.
- Song, C.; Wang, D.; Barabasi, A.L. Connections between Human Dynamics and Network Science. *arXiv* **2012**, arXiv:1209.1411.
- Shaw, S.L.; Tsou, M.H.; Ye, X. Editorial: Human dynamics in the mobile and big data era. *Int. J. Geogr. Inf. Sci.* **2016**, *30*, 1687–1693.
- Graells-Garrido, E.; Ferres, L.; Caro, D.; Bravo, L. The effect of Pokémon Go on the pulse of the city: A natural experiment. *EPJ Data Sci.* **2017**, *6*, 23.
- Toole, J.L.; Ulm, M.; González, M.C.; Bauer, D. Inferring land use from mobile phone activity. In Proceedings of the ACM SIGKDD International Workshop on Urban Computing, Beijing, China, 12–16 August 2012; pp. 1–8.
- Pei, T.; Sobolevsky, S.; Ratti, C.; Shaw, S.L.; Li, T.; Zhou, C. A new insight into land use classification based on aggregated mobile phone data. *Int. J. Geogr. Inf. Sci.* **2014**, *28*, 1988–2007.
- Mann, A. Core Concepts: Computational social science. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 468–470.
- Shaw, S.L.; Sui, D. *Human Dynamics Research in Smart and Connected Communities*; Springer: 2018.
- Blondel, V.D.; Decuyper, A.; Krings, G. A survey of results on mobile phone datasets analysis. *EPJ Data Sci.* **2015**, *4*, 10, doi:10.1140/epjds/s13688-015-0046-0.
- Becker, R.A.; Caceres, R.; Hanson, K.; Loh, J.M.; Urbanek, S.; Varshavsky, A.; Volinsky, C. A tale of one city: Using cellular network data for urban planning. *IEEE Pervasive Comput.* **2011**, *10*, 18–26, doi:10.1109/MPRV.2011.44.
- Calabrese, F.; Ferrari, L.; Blondel, V.D. Urban Sensing Using Mobile Phone Network Data: A Survey of Research. *ACM Comput. Surv.* **2014**, *47*, 25:1–25:20, doi:10.1145/2655691.
- Alexander, L.; Jiang, S.; Murga, M.; González, M.C. Origin–destination trips by purpose and time of day inferred from mobile phone data. *Transp. Res. Part C Emerg. Technol.* **2015**, *58*, 240–250, doi:10.1016/j.trc.2015.02.018.
- Järv, O.; Ahas, R.; Saluveer, E.; Derudder, B.; Witlox, F. Mobile phones in a traffic flow: A geographical perspective to evening rush hour traffic analysis using call detail records. *PLoS ONE* **2012**, *7*, e49171, doi:10.1371/journal.pone.0049171.
- Calabrese, F.; Colonna, M.; Lovisolo, P.; Parata, D.; Ratti, C. Real-time urban monitoring using cell phones: A case study in Rome. *IEEE Trans. Intell. Transp. Syst.* **2010**, *12*, 141–151.
- Bogomolov, A.; Lepri, B.; Larcher, R.; Antonelli, F.; Pianesi, F.; Pentland, A. Energy consumption prediction using people dynamics derived from cellular network data. *EPJ Data Sci.* **2016**, *5*, doi:10.1140/epjds/s13688-016-0075-3.
- Lu, X.; Wrathall, D.J.; Sundsøy, P.R.; Nadiruzzaman, M.; Wetter, E.; Iqbal, A.; Qureshi, T.; Tatem, A.J.; Canright, G.S.; Engø-Monsen, K.; et al. Detecting climate adaptation with mobile network data in Bangladesh: anomalies in communication, mobility and consumption patterns during cyclone Mahasen. *Clim. Chang.* **2016**, *138*, 505–519.
- Pastor-Escuredo, D.; Morales-Guzmán, A.; Torres-Fernández, Y.; Bauer, J.M.; Wadhwa, A.; Castro-Correa, C.; Romanoff, L.; Lee, J.G.; Rutherford, A.; Frias-Martinez, V.; et al. Flooding through the lens of mobile phone activity. In Proceedings of the Global Humanitarian Technology Conference (GHTC), San Jose, CA, USA, 10–13 October 2014; pp. 279–286, doi:10.1109/GHTC.2014.6970293.

18. Wilson, R.; zu Erbach-Schoenberg, E.; Albert, M.; Power, D.; Tudge, S.; Gonzalez, M.; Guthrie, S.; Chamberlain, H.; Brooks, C.; Hughes, C.; et al. Rapid and near real-time assessments of population displacement using mobile phone data following disasters: the 2015 Nepal Earthquake. *PLoS Curr.* **2016**, *8*, doi:10.1371/currents.dis.d073fbeece328e4c39087bc086d694b5c.
19. Brdar, S.; Gavrić, K.; Čulibrk, D.; Crnojević, V. Unveiling spatial epidemiology of HIV with mobile phone data. *Sci. Rep.* **2016**, *6*, doi:10.1038/srep19342.
20. Lima, A.; De Domenico, M.; Pejovic, V.; Musolesi, M. Disease containment strategies based on mobility and information dissemination. *Sci. Rep.* **2015**, *5*, doi:10.1038/srep10650.
21. Wesolowski, A.; Qureshi, T.; Boni, M.F.; Sundsøy, P.R.; Johansson, M.A.; Rasheed, S.B.; Engø-Monsen, K.; Buckee, C.O. Impact of human mobility on the emergence of dengue epidemics in Pakistan. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 11887–11892, doi:10.1073/pnas.1504964112.
22. Pappalardo, L.; Pedreschi, D.; Smoreda, Z.; Giannotti, F. Using big data to study the link between human mobility and socio-economic development. In Proceedings of the 2015 IEEE International Conference on Big Data (Big Data), Santa Clara, CA, USA, 29 October–1 November 2015; pp. 871–878. doi:10.1109/BigData.2015.7363835.
23. Steele, J.E.; Sundsøy, P.R.; Pezzulo, C.; Alegana, V.A.; Bird, T.J.; Blumenstock, J.; Bjelland, J.; Engø-Monsen, K.; de Montjoye, Y.A.; Iqbal, A.M.; et al. Mapping poverty using mobile phone and satellite data. *J. R. Soc. Interface* **2017**, *14*, doi:10.1098/rsif.2016.0690.
24. Ratti, C.; Frenchman, D.; Pulselli, R.M.; Williams, S. Mobile landscapes: Using location data from cell phones for urban analysis. *Environ. Plan. B Plan. Des.* **2006**, *33*, 727–748.
25. Soto, V.; Frias-Martinez, E. Robust land use characterization of urban landscapes using cell phone data. In Proceedings of the 1st Workshop on Pervasive Urban Applications, in Conjunction with 9th International Conference on Pervasive Computing, San Francisco, CA, USA, 12–15 June 2011; Volume 9.
26. Grauwin, S.; Sobolevsky, S.; Moritz, S.; Gódor, I.; Ratti, C. Towards a comparative science of cities: Using mobile traffic records in New York, London, and Hong Kong. In *Computational Approaches for Urban Environments*; Springer: 2015; pp. 363–387.
27. Furno, A.; Fiore, M.; Stanica, R.; Ziemlicki, C.; Smoreda, Z. A tale of ten cities: Characterizing signatures of mobile traffic in urban areas. *IEEE Trans. Mob. Comput.* **2016**, *16*, 2682–2696.
28. Ríos, S.A.; Muñoz, R. Land Use detection with cell phone data using topic models: Case Santiago, Chile. *Comput. Environ. Urban Syst.* **2017**, *61*, 39–48.
29. Furno, A.; El Faouzi, N.E.; Fiore, M.; Stanica, R. Fusing GPS probe and mobile phone data for enhanced land-use detection. In Proceedings of the 2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS), Naples, Italy, 26–28 June 2017; pp. 693–698.
30. Aung, T.; Lwin, K.K.; Sekimoto, Y. Identification and Classification of Land Use Types in Yangon City by Using Mobile Call Detail Records (CDRs) Data. *J. East. Asia Soc. Transp. Stud.* **2019**, *13*, 1114–1133.
31. Bernini, A.; Toure, A.L.; Casagrandi, R. The time varying network of urban space uses in Milan. *Appl. Netw. Sci.* **2019**, *4*, 1–16.
32. Liu, Y.; Fang, F.; Jing, Y. How urban land use influences commuting flows in Wuhan, Central China: A mobile phone signaling data perspective. *Sustain. Cities Soc.* **2020**, *53*, 101914.
33. Noyman, A.; Doorley, R.; Xiong, Z.; Alonso, L.; Grignard, A.; Larson, K. Reversed urbanism: Inferring urban performance through behavioral patterns in temporal telecom data. *Environ. Plan. B Urban Anal. City Sci.* **2019**, *46*, 1480–1498.
34. Cottineau, C.; Vanhoof, M. Mobile phone indicators and their relation to the socioeconomic organisation of cities. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 19.
35. Brdar, S.; Novović, O.; Grujić, N.; González-Vélez, H.; Truić, C.O.; Benkner, S.; Bajrovic, E.; Papadopoulos, A. Big Data Processing, Analysis and Applications in Mobile Cellular Networks. In *High-Performance Modelling and Simulation for Big Data Applications: Selected Results of the COST Action IC1406 cHiPSet*; Kołodziej, J., González-Vélez, H., Eds.; Springer International Publishing: Cham, Switzerland, 2019; pp. 163–185. doi:10.1007/978-3-030-16272-6\_6.
36. Truić, C.O.; Novović, O.; Brdar, S.; Papadopoulos, A.N. Community detection in who-calls-whom social networks. In Proceedings of the International Conference on Big Data Analytics and Knowledge Discovery, Regensburg, Germany, 3–6 September 2018; pp. 19–33.

37. Karau, H.; Konwinski, A.; Wendell, P.; Zaharia, M. *Learning Spark: Lightning-Fast Big Data Analytics*, 1st ed.; O'Reilly Media, Inc.: 2015.
38. Aggarwal, C.C.; Wang, H. *Managing and Mining Graph Data*; Springer: 2010. doi:10.1007/978-1-4419-6045-0.
39. Cook, D.J.; Holder, L.B. *Mining Graph Data*; John Wiley & Sons: 2006, New Jersey, USA. doi:10.1002/0470073047.
40. Blondel, V.D.; Guillaume, J.L.; Lambiotte, R.; Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, 2008, P10008, doi:10.1088/1742-5468/2008/10/P10008.
41. Newman, M.E.J.; Girvan, M. Finding and evaluating community structure in networks. *Phys. Rev. E* **2004**, 69, 026113, doi:10.1103/PhysRevE.69.026113.
42. Oldham, S.; Fulcher, B.; Parkes, L.; Arnatkeviciute, A.; Suo, C.; Fornito, A. Consistency and differences between centrality measures across distinct classes of networks. *arXiv* **2018**, arXiv:1805.02375.
43. Malliaros, F.D.; Giatsidis, C.; Papadopoulos, A.N.; Vazirgiannis, M. The core decomposition of networks: Theory, algorithms and applications. *VLDB J.* **2019**, 1–32.
44. Barlacchi, G.; De Nadai, M.; Larcher, R.; Casella, A.; Chitic, C.; Torrisi, G.; Antonelli, F.; Vespignani, A.; Pentland, A.; Lepri, B. A multi-source dataset of urban life in the city of Milan and the Province of Trentino. *Sci. Data* **2015**, 2, 150055, doi:10.1038/sdata.2015.55.
45. Armbrust, M.; Xin, R.S.; Lian, C.; Huai, Y.; Liu, D.; Bradley, J.K.; Meng, X.; Kaftan, T.; Franklin, M.J.; Ghodsi, A.; et al. Spark sql: Relational data processing in spark. In Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, Melbourne, VC, Australia, 31 May–4 June 2015; pp. 1383–1394.
46. Serrano, M.Á.; Boguná, M.; Vespignani, A. Extracting the multiscale backbone of complex weighted networks. *Proc. Natl. Acad. Sci. USA* **2009**, 106, 6483–6488, doi:10.1073/pnas.0808904106.
47. Land Copernicus. Copernicus Land Monitoring Service Urban Atlas. 2012. Available online: <https://land.copernicus.eu/> (accessed on 13 November 2019).
48. e Silva, F.B.; Poelman, H. *Mapping population density in Functional Urban Areas*; Technical Report, 2016.
49. Garas, A.; Schweitzer, F.; Havlin, S. Ak-shell decomposition method for weighted networks. *New J. Phys.* **2012**, 14, 083030, doi:10.1088/1367-2630/14/8/083030.
50. Breiman, L. Random forests. *Mach. Learn.* **2001**, 45, 5–32.
51. Hoerl, A.E.; Kennard, R.W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **2000**, 42, 80–86.
52. Louppe, G.; Wehenkel, L.; Suter, A.; Geurts, P. Understanding variable importances in forests of randomized trees. In *Advances in Neural Information Processing Systems*; 2013; pp. 431–439.
53. Wagner, S.; Wagner, D. *Comparing clusterings: An overview*. Technical Report; 2007.
54. Valente, T.W.; Coronges, K.; Lakon, C.; Costenbader, E. How correlated are network centrality measures? *Connections* **2008**, 28, 16.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).