Discovering Mobility Functional Areas: A Mobility Data Analysis Approach

Lorenzo Gabrielli[2], Daniele Fadda[2], Giulio Rossetti[1,2], Mirco Nanni[2], Leonardo Piccinini[3], Dino Pedreschi[1], Fosca Giannotti[2], and Patrizia Lattarulo[3]

[1] University of Pisa, Largo Bruno Pontecorvo, 2 Pisa, Italy
name.surname@di.unipi.it,
[2] KDD Lab. ISTI-CNR, via G. Moruzzi, 1 Pisa, Italy
name.surname@isti.cnr.it
[3] IRPET, Regione Toscana Via Pietro Dazzi, 1, Firenze, Italy
name.surname@irpet.it

Abstract. How do we measure the borders of urban areas and therefore decide which are the functional units of the territory? Nowadays, we typically do that just looking at census data, while in this work we aim to identify functional areas for mobility in a completely data-driven way. Our solution makes use of human mobility data (vehicle trajectories) and consists in a agglomerative process which gradually groups together those municipalities that maximize internal vehicular traffic while minimizing external one. The approach is tested against a dataset of trips involving individuals of an Italian Region, obtaining a novel territorial division which allows us to identify mobility attractors. Leveraging such partitioning and external knowledge, we show that our method is able to outperform the state-of-the-art algorithms. Indeed, the outcome of our approach is of great value to public administrations for creating synergies within the aggregations of the territories obtained.

Keywords: human mobility, community discovery, functional areas

1 Introduction

The traditional interpretation of the urban hierarchy simply refers to the size of the city, with its population and boundaries. From the theoretical point of view, a rather different perspective is given by the concept of polycentrism [1]: urban areas are often evolving from mono-centric agglomerations to more complex systems made of integrated urban centers (cores) and sub-centers. In other territories, a number of cities and towns are increasingly linking up, forming polycentric integrated areas.

The understanding of the spatial organization of homogeneous regions and of how places link among them can improve analytical approaches when facing governance challenges such as the economic development of nation wide complex systems. Indeed, policy makers are paying increasing attention to the role

of homogeneous economic agglomeration and to the capacity of local areas to contribute to social growth [2].

Moreover, the contraction of public expenditure has driven a process of service concentration towards denser urban areas.

Our work aims at contributing to this debate by providing a tool to researchers and policy makers to build a novel definition of regions seen as functional areas of similar behaviors [3].

The questions that drive our research are thus the following:

Q1: Can we identify mobility functional units just looking at human vehicular movement data?
Q2: Are such units mono-centric agglomerations or more complex polycentric integrated areas?
Q3: Which are the most relevant characteristics of such areas?

To answer such questions we developed a methodology whose identify a reasonable number of well-knit sub-regions that are significantly self-contained in terms of mobility fluxes, and therefore represent candidate functional areas w.r.t. mobility. The approach also tries to be not influenced by marginal municipalities that are substantially disconnected from the others and/or less interesting from the decision maker point of view, e.g. because of low traffic flows or small population. This latter characteristic is often neglected by traditional group discovery algorithms, whose final goal is to partition a generic set of linked elements disregarding any semantics attached to it.

As in [4], we model movements between municipalities as a network, and we compare our approach with competitors taken from network analysis studies (i.e., *community detection*) as well as from data mining ones (i.e., *clustering*).

2 Background

In this section we discuss some works in literature that are adopted – or might be adapted – to identify functional areas.

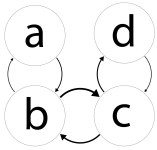
From a statistical and economical point of view, in [3] are illustrated different methodologies used to solve the problem of redefining urban areas. Among these, Dynamic Metropolitan Areas (DMA) are specifically designed to deal with the characteristics of policentricity. The first stage of the DMA algorithm has a top-down approach: it identifies first-order centers (seeds) which have at least 50,000 inhabitants and merges the surrounding municipalities that commute at least 15% of their inhabitants.

While (to the best of our knowledge) there are no works on data science tackling our specific problem, several group discovery methods might be applied to it, following a clustering of network-based perspective. Here we briefly mention some basic approaches in the field, while Section 1b will provide a detailed description of those we compared to.

Clustering methods generally aim to group a set of objects putting together those that are similar to each other under some specific notion of similarity. The

three classical and most frequently adopted examples are: *k-means*, representing a family of partitioning methods that create compact clusters, trying to minimize the diversity within a cluster and to maximize it across different clusters; *hierarchical clustering*, producing several different partitioning at different levels of aggregation; *density-based clustering*, which puts together groups of objects that form locally dense areas, not enforcing any constraint on the size and shape of clusters.The solution we proposed belongs to the hierarchical methods, yet basing the aggregation of groups on complex self-containment considerations rather than on the standard maximization of mutual similarities within the cluster.

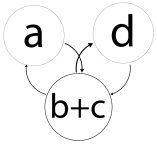
Network-based methods saerch for *communities*, i.e. groups of linked nodes that share common properties, defining them w.r.t. several objective functions[5]. In the context of territorial partitioning, community discovery has became an important tool for decision makers that need to study social complex systems, e.g. in grouping together municipalities showing similar characteristics [2]. Indeed, by adopting a community discovery approach we can obtain, in a bottom-up way, an unsupervised classification of territories. In this work we realize a dedicated method, that we called Mobility Functional Areas Discovery (MFAD), based on a context-specific combination of objective functions. As we will show in the experiment section, results prove the superiority of our solution.



(a) Iteration n

Local Function	Inner traffic
localQ(a,b)	50%
localQ(a,d)	0%
localQ(a,c)	0%
localQ(b,c)	**52%**
localQ(d,c)	48%

(b) $Evaluation(n)$



(c) Iteration $n+1$

Fig. 1: An iteration of *MFAD*: Considering a generic iteration n (a), we evaluate all possible combination of *localQ* (b). Once selected the best pair, i.e. (b,c), we proceed to the union of nodes and updating the data structures (c).

Algorithm 1 $MFAD(OD)$

Inputs: OD represents the mobility flows among municipality.
Output: The territorial tessellation $\tilde{T}$.
$G = CreateMobilityGraph(OD)$ ▷ loading the graph
$T = \emptyset$
while $|G.V| > 1$ **do**
  $C = ComputeConfigurations(G)$ ▷ computing all the possible fusions
  $bestPair = \arg\max_{(a,b)\in C} localQ(a, b, G)$ ▷ selecting the best fusion
  $G = update(G, bestPair)$ ▷ updating G
  $T = T \cup \{G\}$ ▷ Saving configuration
$\tilde{T} = \arg\max_{G\in T} globalQ(G)$ ▷ Evaluating configurations
return $\tilde{T}$

3 Mobility Functional Areas Discovery (*MFAD*)

Our final goal is to partition the territory considering mobility habits. We model the mobility between municipalities as a network graph $G = (V, E, F)$, and approach the task of defining a meaningful tessellation as the problem of identifying a community coverage of G. The municipalities define the set of the nodes V of G, while edges E represent the flows between nodes, their weights being denoted with $F(e)$ for each $e \in E$.

The general criterion we follow for obtaining an optimal solution is self containment of traffic flows: the traffic within a group of nodes should be much higher than that across different groups. The literature on community detection over networks provides a measure, called *modularity*, that seems to approximate this notion. However, directly maximizing modularity in our context leads to results that violate some basics expectations of the domain expert, e.g., it causes the appearance of geographically discontinuous groups (which is counter-intuitive) and the fact that densely populated areas tend to dominate the whole process (undesirable).

In order to overcome modularity limitations, adopt an agglomerative process, summarized in Algorithm 1, driven by a *local* measure that at each step of the process evaluates the benefits – in terms of self-containment of flows – of aggregating a pair of groups into a single larger one. Modularity is then used as global measure to decide when the aggregation process should stop. The process starts from a situation where each input node in the network G is kept separated from the others, meaning that each node forms a cluster by itself. Iteratively, two clusters are selected and merged together, thus reducing the number of clusters by one unit, and stop only when G contains just one cluster. In order to choose which clusters to merge, all possible pairs are taken into consideration, and for each of them our local measure *localQ* is computed, selecting the best one. Such measure, in particular, is computed as the fraction of the *local* traffic (i.e. the flows involving either node of the pair) that would be converted into internal traffic thanks to the merger:

$$localQ(a, b, G) = \frac{F(a, b) + F(b, a)}{\sum_{(x,y) \in E \wedge \{a,b\} \cap \{x,y\} \neq \emptyset} F(x, y)} \tag{1}$$

Considering the example in Fig. 1, $localQ(b, c, G)$ would compute the ratio between the total traffic between b and c, i.e. $F(b, c) + F(c, b)$ (the same traffic that, if b and c are merged, will move from inter-group traffic to intra-group) and the total traffic from/to any of them, i.e. $F(a, b) + F(b, a) + F(b, c) + F(c, b) + F(c, d) + F(d, c)$.

When the best pair of nodes a, b is found they are merged thus replacing them in G by a single node $a\&b$. The edges to/from the new node is the union of those to/from either of the original nodes, and the flows associated to them are computed as the sum of the original flows, e.g. $F(a\&b, c) = F(a, c) + F(b, c)$.

When the iterative process comes to the end, T will contain the collection of all graphs obtained at each step, including the original graph G and the last

one where only two (big) nodes are left. In order to identify the most promising aggregation level, we adopt the *modularity* measure as global evaluation criterion, and find the graph $G \in T$ that maximizes it. That is computed by function *globalQ*:

$$globalQ(G) = \sum_{(i,j)\in E} F(i,j) - F(i \rightarrow) * \frac{F(\rightarrow j)}{K} \quad (2)$$

where $F(\rightarrow i)$ represents the total sum of outgoing flows from node i and $F(j \rightarrow)$ is the total of incoming flows to node j. Finally $K = \sum_{e\in E} F(e)$ represents the total flows in the network. Overall, the rightmost part of the formula provides the expected flow from i to j.

Computational costs: From a computational point of view, the Algorithm costs $o(n^3)$ where n is the cardinality of V. While high, cost is not a real issue in our application, since n is typically a low number; in our case study, covering the municipalities of a region, we have around 300 nodes, and running the whole process on a standard computer takes a few minutes. It is worth to notice that the expensive part of the algorithm is easily parallelizable.

4 Experiments

We apply our methodology on a dataset of trajectories capturing the mobility of individuals in a region. The dataset consists of 5 million trips produced by around 70 thousands cars within Tuscany (Italy) in a period of observation of 5 weeks[4]. Tuscany has about 4.8 million residents and 287 municipalities with a population density of 163 $residents/Km^2$.

MFAD is applied on the origin and destination matrix (OD Matrix) at municipality level. An OD Matrix is a network that describes the number of trajectories that start in a municipality and end in another one (not necessarily adjacent): the Tuscany dataset is composed by 287 nodes and 30 thousands of arcs. The average degree of the network is 119.74, the average clustering coefficient is 0.74 while the average shortest path is 1.64. The final output of *MFAD* is a set of 25 contiguous areas that, as will be discussed in deeper detail later, highlight some interesting structures in the region.

4.1 Competitors

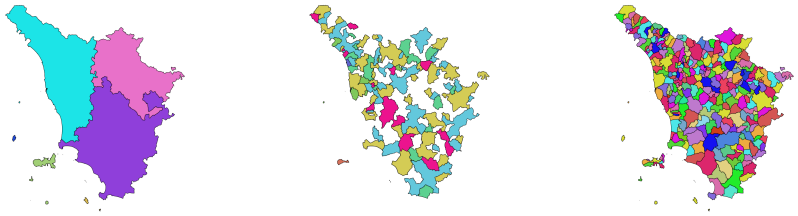
Here we introduce some of the main state-of-art methods for partitioning sets of elements into groups, which are then used as competitors against our proposed solution. Following the literature, we refer to partitioning methods based on networks (CD) or clustering. Finally, we introduce a simple random model as baseline solution.

[4] The analyzed trajectories are generated from raw GPS data using a tool called M-Atlas [6].

Louvain described in [7], is a fast and scalable algorithm based on a greedy modularity approach. It has been shown that modularity-based approaches suffer a resolution limit, and therefore, Louvain is unable to detect medium size communities [8]. This produces communities with high average density, due to the identification of a predominant set of very small communities and a few huge communities. **Demon**: introduced in [9], is an incremental and limited time complexity algorithm for community discovery. It extracts ego networks, i.e., the set of nodes connected to an ego node u, and identifies the real communities by adopting a democratic bottom-up merging approach of such structures. **Infohiermap**: is one of the most accurate and best performing hierarchical non-overlapping clustering algorithms for community discovery [10] studied to optimize community conductance. The graph structure is explored with a number of random walks of a given length and with a given probability of jumping into a random node.

K-medoids and **DBSCAN** are two of the existing methods for obtaining a homogeneous grouping of elements. We choose these algorithms because they can easily accommodate any distance function, which is a key feature for our problem, since standard measures (Euclidean, etc.) would not model it in a meaningful way. In particular, since the goal of our analysis is to identify groups that maximize the local score introduced in Section 3, we feed the algorithms with a distance which is the complement of $localQ$, i.e. $distance(a, b) = 1 - localQ(a, b, Q)$, which has the same range of values $[0, 1]$. **K-medoids** is a partitional clustering algorithm, that clusters the dataset into k clusters, where k is known a priori. It minimizes the overall distance (more specifically, the sum of squared distances) between the points of a cluster and its center. In contrast to the K-means algorithm, K-medoids chooses a real point as center and works with an arbitrary metrics of distances between data-points. We determine k using the standard *silhouette* score [11]. **DBSCAN** is a density-based clustering algorithm. It identifies each input point as *core point*, *border point* or *outliers*. A point p is a core point if at least $minPts$ other points are within distance ϵ from it. Border points are those that are not core but have a core point within distance ϵ. Finally, all remaining points labeled as outliers. The $minPts$ parameter is known to be not critical, and thus was set to the standard default value of 3. The (more critical) parameter ϵ was instead chosen through a grid search, selecting the one that optimizes the $globalQ$ function introduced in Eq. 2.

We compute a baseline method called **NM1** in order to test if there is a random configuration that generates a better territorial partitioning than *MFAD*. $NM1$ fixes a priori the number of territorial partitions (k), then randomly chooses k elements that represent the seed of clusters. The remaining municipalities are assigned sequentially, according to three criteria: the candidate municipality is assigned to an adjacent seed; if not possible, the municipality is assigned to an existing group that contains an adjacent municipalities; if all fails, the municipality is assigned randomly to a seed (even if not adjacent).



(a) Louvain (b) Demon (c) Infohiermap

Methods	Internal Edge Density	Conductance	Modularity	Communities	Contiguous
MFAD	0.27/0.75/0.49/0.20	0.014/0.97/0.88/0.19	-0.06	25	True
Louvain	0.15/0.32/0.21/0.07	0.014/0.58/0.38/0.27	0.16	7	True
Demon	0.12/0.50/0.28/0.18	0.37/0.90/0.50/0.17	-0.38	7	False
Infohiermap	0.09/0.50/0.18/0.10	0.90/0.98/0.95/0.24	0.006	29	False

Fig. 2: Communities identified. (a) Louvain produces very few and large communities; (b) Demon communities are slightly dispersed and show a significant overlapping (not visible from the figure); (c) Infohiermap produces several non-contiguous areas. We report the min., max., avg. and std. deviation of the measures. *MFAD* communities are denser on average, and have a good value of conductance even thought it was not explicitly among its optimization criteria.

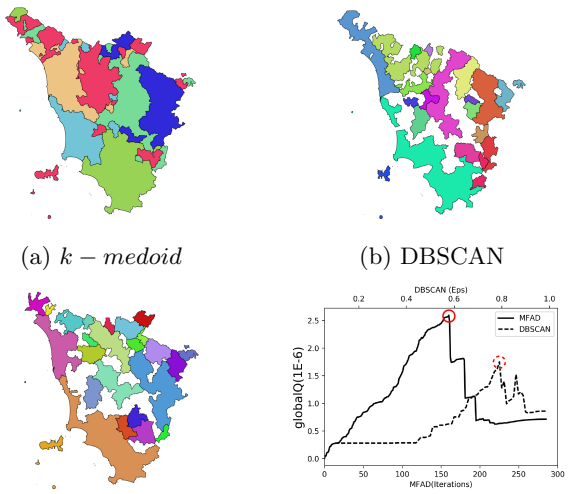
4.2 Results

Here we show the territorial partitions obtained with network-based methods and how they provide a good partitioning, yet not satisfying some requirements needed to answer our research questions. We will see the territorial partitions obtained with clustering methods and how DBSCAN proves to be the best competitor for our approach. Finally, we evaluate the territorial partitions obtained with a null model w.r.t. the final configuration of *MFAD*, proving that its random process fails to find better partitioning.

Since **Louvain** optimises the partition modularity, its modularity score is higher than *MFAD*, yet the latter provides communities with higher average densities and higher conductance. The main drawback of Louvain is the reduced number of communities it produces, which comes from the tendency of modularity-based approaches to build up few very large communities along with small sized ones (Fig. 2a). **Demon** is a good method because it manages to handle noise and overlapping communities. Yet, *MFAD* improves both density and conductance (Fig. 2b). In our context, moreover, offering crisp and non-overlapping partitions is a plus, since it simplifies the interpretation of results. Demon's overall very good results are therefore not very appealing for our goal. **Infohiermap** creates communities that are on average less dense than *MFAD*. Also, it produces a comparable conductance (*MFAD*=0.88, Infohiermap=0.95), while not optimizing this measure explicitly. Finally, Infohiermap groups are

consistently non-contiguous, which is a counterintuitive result from the the application viewpoint (Fig. 2c). Overall, *MFAD* produces results that outperform CD approaches, since the latter tend to find either too few and big communities, or non-contiguous ones.

Now we show the territorial partitions obtained with K-medoids and DBSCAN cluster methods. For **K-medoids** we selected the k value that optimizes the *silhouette* coefficient, which results to be $k = 6$. As shows in Fig. 3a), the algorithm produces non-contiguous areas and a fragmented spatial partitioning, which also happens for any other value of k. **DBSCAN** was performed for several values of ϵ in the interval $[0, 1]$, computing the *globalQ* score for each result, as reported in Fig. 3d). The best score is obtained for $\epsilon = 0.79$. Fig. 3b) depicts the corresponding territorial partitioning showing that DBSCAN basically satisfies the requirements we mentioned before, producing a reasonable number of contiguous communities and isolating/removing uninteresting (noisy) municipalities. In terms of *globalQ* score, however, we can see that the best DBSCAN can reach is still largely inferior to *MFAD* (around 35% smaller). Also in this case, despite DBSCAN is so far the best competitors, *MFAD* remains, however, the best option, since it reaches much higher values of our *globalQ* quality function.



(a) $k - medoid$ (b) DBSCAN

(c) *MFAD* (d) performance comparison

Fig. 3: Territorial partitioning. (a) K-medoids generates a fragmented partitioning useless for our purpose. (b) DBSCAN provides 21 contiguous communities with $\epsilon = 0.79$ (the optimal value). Visually, the result is good and comparable to (c) *MFAD* . As shown in (d), however, the optimal value of *globalQ* for DBSCAN is much lower than *MFAD*.

Finally, the random heuristics called **NM1** has been applied with all possible values of k between 10 and 30, running the method 100 times for each k. Varying the number of areas produced by the model $NM1$, which assigns each municipality randomly to one of the k groups, we obtain lower values of *globalQ* w.r.t. *MFAD*. We can clearly see that the random approach consistently behaves much worse than our solution, regardless of the number of groups it seeks.

5 Evaluation

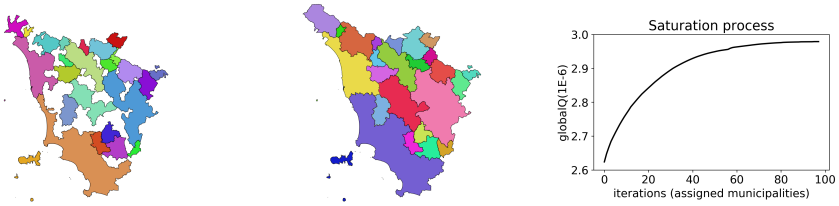
In this section we evaluate the functional areas obtained by *MFAD* with the aid of domain experts (co-authors of this paper) working for a public agency on topics related to territorial policies. For this kind of problems the expert needs a complete tessellation of the territory, therefore in Sec. 5.1 we show an assignment criterion for municipalities not grouped by *MFAD*. In Sec. 5.2 we show the internal structure of the main areas identified and, finally, we report some domain expert's comments in terms of how the obtained results can be used (Sec. 5.3).

5.1 Saturation

MFAD produces clusters which do not include the totality of the municipalities. For some applications the domain expert requires the assignment of all the municipalities in a cluster. This may be the case, for example, if we use the partition to redefine the perimeters of universal public services (health care, education, transport). In this scenario, we must assign every municipality to a cluster, since we cannot have a territory where the service is not provided. For this reason, we applied a *saturation* process, that iteratively (i) selects the unassigned municipality m and the area a such that they are adjacent and their merger maximizes the *globalQ* function; (ii) assigns m to a; (iii) reiterates the process until all areas have been assigned. Geographically isolated municipalities, if any, form singleton areas.

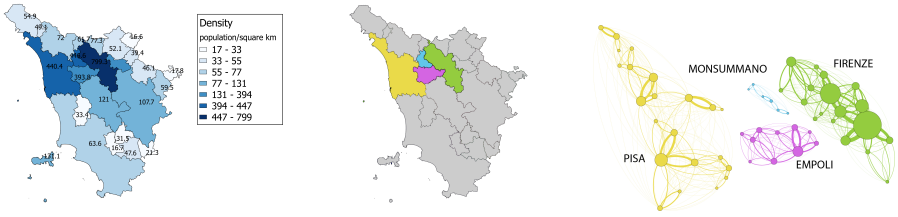
5.2 The policentric structure of the urban areas

As requested by the domain expert we analyzed the structure of the communities identified, with particular reference to the highly populated areas. In Fig. 5a) we note that rural areas (mainly situated in the South) are defined by larger aggregations, while the central areas are comprised of smaller ones. The Northern border, which is mainly mountainous, shows more fragmentation than the population density would suggest. This could be due to a combination of two factors: insulated communities and a border effect (since our data are trimmed at the administrative regional border). After selecting some interesting areas for the domain expert (Fig. 5b), we can observe the internal structure of the communities with the highest population density. Fig. 5c) shows the structure of 4 communities, depicting the flow between municipalities and the in/out flows of



(a) *MFAD* (b) *MFAD* with Saturation (c) Global Improvement

Fig. 4: Figures (a) and (b) show the result provided by *MFAD*, before and after the saturation process applied to include also the unclustered municipalities to the detected areas. In (c) we show the growth of *globalQ* value for each reallocated municipality.



(a) Population Density (b) Most populated areas (c) Network structure

Fig. 5: Here we evaluate the characteristics of the areas obtained by *MFAD*. (a) The size and density of communities depends on the socio-economic characteristics of the territories, observing a very low density in rural areas (mainly in the South of Tuscany and mountainous locations) and very high one in the most urbanized zones. (b) Among the 25 areas we select those belonging the more urbanized part of the region. (c) observing the internal structure of the network generated by vehicular flows, we can observe several polycentric sub-areas.

each single municipality as the size of lines and points. The densest area has a main hub in Firenze, which is accompanied by a second, slightly smaller one at North-West, and together they keep all the area tightly connected to them. The area on the West is centered on Pisa, and has a different and more diffuse structure. Indeed, there are several poles of comparable size (Pisa being slightly larger), each capturing a part of the area, and in most cases they are only weakly connected to other poles. The area around Empoli is quite similar to Pisa, at a smaller scale. Finally, the small area around Monsummano is very homogeneous and made of municipalities of approax. the same weight.

5.3 Exploitation potential

The proposed bottom-up approach defines a partition of the selected region that can be be interpreted in different ways. The first application could be an

analytical approach: mobility patterns tell us a story about territorial integration that goes beyond the administrative borders. If we want to analyze the socio-economic dynamics and the determinants of local development, the algorithm can suggest us which might be the boundaries of our analysis.

From the public administration perspective, this method of clustering territories could be helpful in the policy design phase. Since we are looking at highly integrated areas, we might want to tailor the intervention on the characteristics of the aggregated partition, since we expect that the outcome at the municipality level can have a spillover effect on the surrounding territories, and, vice-versa, that the socio-economic conditions of surrounding territories affect the potential outcome of the single municipality. Therefore, an integrated and coordinated policy implementation approach can maximize the desired outcome and prevent potential drawbacks.

Moreover, as we mentioned in the introduction, public service provision can be more cost-effective when implemented at an aggregated level. This is especially true in the case of Italian municipalities, where excessive fragmentation of administrative units has been recognized as one of the sources of inefficient public expenditure.

6 Conclusion

The evolution of the economy and society affects the way metropolitan areas change over space and time. It is therefore necessary an accurate boundary delimitation of services in order to increase the efficiency of public administrations without marginalizing the surrounding territories. We propose to use Big Data, to be precise GPS data produced by vehicles, to overcome the limitations of traditional sources in the measurement of the real boundaries of the city. The position of our work is to contribute to provide a tool to policy makers for building a novel definition of regions considered as mobility functional areas [3].

The results highlighted in the paper show on a real dataset that *MFAD* outperforms state-of-the-art methods optimizing an objective function defined with domain experts. We have shown that 25 communities emerge from data, observing only the private vehicle mobility (ref. question Q1 in the introduction). The identified communities show a polycentric structure, with centers apparently corresponding to highest population density and presence of transport infrastructures that facilitate connections to/from other municipalities (ref. Q2). Finally, the areas found have very diverse population density and size, the tighter connections corresponding to highest populated areas (ref. Q3, and see Fig. 5a)).

Planned developments of the work include the exploration further aspects with domain experts, in particular how communities change by applying our method only to the systematic vs. occasional traffic, and by including public transportation. Also, it would be very helpful including social and productive aspects in the global objective function. Comparison with the Local Labour Communities defined by the Italian Statistical Institution (ISTAT) shows that the

border effect (i.e. influence of neighbouring municipalities outside our dataset) is actually relevant in those areas. This suggests that further developments should include cross-border data. Finally, we plan modify the initialization phase of our method – now consisting in putting each municipality in a separated group – by following the approach in [3], which might provide better initial seeds for the computation, injected through local domain knowledge.

Acknowledgment

This work is partially supported by the European Community's H2020 Program under the scheme 'INFRAIA-1-2014-2015: Research Infrastructures', grant agreement #654024 'SoBigData: Social Mining & Big Data Ecosystem'[5].

References

1. Brezzi, M.: Redefining" urban": A New Way to Measure Metropolitan Areas. OECD (2012)
2. ISTAT: Local labour system
3. Boix, R., Veneri, P., Almenar, V.: Polycentric metropolitan areas in europe: towards a unified proposal of delimitation. In: Defining the Spatial Scale in Modern Regional Analysis. Springer (2012) 45–70
4. Rinzivillo, S., Mainardi, S., Pezzoni, F., Coscia, M., Pedreschi, D., Giannotti, F.: Discovering the geographical borders of human mobility. KI - Künstliche Intelligenz **26**(3) (2012) 253–260
5. Fortunato, S.: Community detection in graphs. Physics Reports **486**(3-5) (2010) 75 – 174
6. Trasarti, R., Rinzivillo, S., Pinelli, F., Nanni, M., Monreale, A., Renso, C., Pedreschi, D., Giannotti, F.: Exploring real mobility data with m-atlas. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer Berlin Heidelberg (2010) 624–627
7. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment **2008**(10) (2008) P10008
8. Fortunato, S., Barthélemy, M.: Resolution limit in community detection. PNAS **104**(1) (January 2007) 36–41
9. Coscia, M., Rossetti, G., Giannotti, F., Pedreschi, D.: Demon: a local-first discovery method for overlapping communities. In 0001, Q.Y., Agarwal, D., Pei, J., eds.: KDD, ACM (2012) 615–623
10. Rosvall, M., Bergstrom, C.T.: Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Sciences **105**(4) (2008) 1118–1123
11. Rousseeuw, P.J.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics **20** (1987) 53 – 65

[5] http://www.sobigdata.eu