City Indicators for Mobility Data Mining

Mirco Nanni
ISTI-CNR, Pisa
Pisa, Italy
mirco.nanni@isti.cnr.it

Agnese Bonavita
Scuola Normale Superiore
Pisa, Italy
agnese.bonavita@sns.it

Riccardo Guidotti
University of Pisa
Pisa, Italy
riccardo.guidotti@di.unipi.it

ABSTRACT

Classifying cities and other geographical units is a classical task in urban geography, typically carried out through manual analysis of specific characteristics of the area. The primary objective of this paper is to contribute to this process through the definition of a wide set of city indicators that capture different aspects of the city, mainly based on human mobility and automatically computed from a set of data sources, including mobility traces and road networks. The secondary objective is to prove that such set of characteristics is indeed rich enough to support a simple task of *geographical transfer learning*, namely identifying which groups of geographical areas can share with each other a basic traffic prediction model. The experiments show that similarity in terms of our city indicators also means better transferability of predictive models, opening the way to the development of more sophisticated solutions that leverage city indicators.

1 INTRODUCTION

Classifying a geographical territory into semantic categories is one of the most common tasks in research areas such as urban geography, urban planning and mobility data analytics [7]. Characterizing human mobility is a key component of this process, and it is well known that mobility often does not work the same way across different regions. A movement pattern in a mountainous countryside may have other implications than the same pattern has in the suburbs of a large town. The movement trajectories in a planned city with rectangular streets and strict zoning laws might be completely different than the ones in a town that has grown organically without any clear structure. Therefore, any kind of property that was learned in a particular area, in general cannot simply be assumed to hold in another one.

This paper aims at making a first step towards the *characterization of a geographical area*. That is achieved through a range of quantitative measures that provide a multilayer description of urban regions and are a means for displaying differences between cities, municipalities, or other geographical units. Such a numerical description of urban areas can have a wide spectrum of applications. Among them, the measures presented in this work can be used as an input for *geographical transfer learning*, that is the transformation of knowledge gained in one geographical region in order to apply it to another region. This problem will be considered as a case study for the extracted indicators.

We consider two main approaches: *(i)* computing features that describe each area isolated from the others, that we call *local city indicators*; and *(ii)* computing features that describe its relation with the others, named *global city indicators*. The first group covers four different families of measures: spatial concentration indexes of human activities; network features of intra-city traffic flows; mobility characteristics of the individual mobility, obtained from networks that represent the places and movement of single users; last, characteristics of road networks and how traffic is distributed in them. The group of global city indicators, instead, looks at the mobility between cities as a graph, where each city is represented by a node, and extracts network features for each node. Both the complete network and the ego-network for each city are considered.

After describing all the city indicators we introduce a mobility prediction problem, and we use it to test how much predictive models are transferable across different regions. In particular, we study the relationship between *transferability* between two areas, i.e. the performances of a model built on one area and used to make predictions on the other one, and their *similarity* in terms of city indicators. The results confirm our hypothesis that cities with similar indicators are more likely to be transfer-compliant, this providing a first guide to understand which predictive models can be reused in other areas.

Finally, a key feature of this work is that all methods are implemented in a way that makes it possible to automatically calculate all characteristics for hundreds of different cities and entire regions. The resulting software (a Python library) enables the user to process an unlimited amount of data simply by passing a database with trajectories and a list containing the positions of the geographical areas of interest as an input.

The rest of this paper is organized as follows. Section 2 introduces some related works; Section 3 presents the dataset and geographical areas used as testbed in the paper; Sections 4 and 5 describe, respectively, the local and global city indicators; Section 6 presents our case study on evaluating the relations between geographical transferability of a simple predictive model between any pair of areas and their similarity based on our indicators; finally, Section 7 closes the paper with conclusive remarks.

2 RELATED WORK

Chracterizing urban spaces is a fundamental task of urban geography, which considers the spatial distribution of spaces and patterns of movement, focusing both on structural properties and how the different parts interact [7]. Historically, determining such characteristics was usually a domain expert-driven process, that required a huge amount of time, and also particular care to ensure that results are comparable across different places. Geographical Information Science introduced several innovations that helped also to automatize and extend the approach, including statistical methods for geography [33] and computational tools for managing large databases of information.

On another direction, city indicators have a important application in defining the sustainability characteristics of urban areas. Various attempts have been made to design indicators for monitoring sustainability at various levels, such as national [11] and city level [32]. As described in the review paper [17], the literature covers a wide range of aspects, including mobility-related ones (e.g. mobility space usage and functional diversity). However, very few attempts were made to systematically exploit big data sources to estimate them. One example was the Air Quality

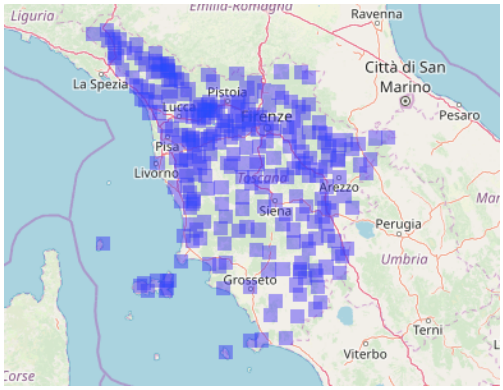

Figure 1: The areas of study: 10×10km squares centered on each municipality in Tuscany.

Now EU project [9], which used vehicular and public transport data to infer some measures. Yet, that is limited to direct and simple ones, such as traffic, speeds and exposure to pollution. The literature also considers mobility indicators and road network properties as potential measures to adopt, which is aligned with our approach [38].

Finally, exploiting big mobility data to understand the properties of geographical spaces is a very active area. It includes data mining methods to find mobility patterns and regularities [15], simulations to estimate various indicators, like the impact of alternative transportation means as car pooling [20], the visual exploration of patterns and contextual features [13], etc. However, to the best of our knowledge, no existing work tried to collect a wide set of complex indicators in a systematic and reproducible way, directly aimed to make cities comparable in a computational way.

3 DATASET

The testbed considered in this paper is a dataset of GPS traces of private vehicles provided within the Track &Know project[1] moving in the Tuscany region, Italy. The experiments were performed on a sample of 18.9 million trajectories of 250,239 cars, which were collected during a period of seven weeks. The geographical unit adopted to model a *"city"* is the municipality.

All measures were calculated separately for each of the 276 municipalities of Tuscany. For the sake of simplicity and applicability to a wider range of situations, the areas to investigate were chosen to be a 10×10 km rectangle for each municipality, with the sides approximately parallel to the meridians and parallels, centered around the town or village center (see Figure 1).

It should be noted that, for the purpose of this work, only a partial subset of trajectories is considered, namely those starting and ending in Tuscany, and lasting less than 24 hours (indeed, longer trips are exceptional and not very representative).

4 LOCAL CITY INDICATORS

Here we introduce the *local* city indicators designed individually for each municipality. They are grouped in spatial concentration measures, flows measure, individual mobility and street network.

[1]https://trackandknowproject.eu/

4.1 Spatial Concentration

Spatial concentration is one of the most important aspects in the description of urban regions and answer the question *how the density of people and activities vary across the area?* This question was traditionally focused on people's residency and workplace, since that was the only available data, mostly coming from census or government records. More recent research is profiting from the availability of more detailed data from mobile phones, vehicle trackers and satellite imaging [2, 22, 39]. Spatial concentration is used in a vast range of different fields [16, 19, 20, 23, 36]. In this work, the concept of spatial concentration is focused on the overall amount of mobility, undifferentiated by types of activity. The question of interest is: *are the activities concentrated in cluster-like centers of high density or are they spread-out across the map?*

In the following, we present three approaches to answer this question: *spatial entropy*, *Moran's measure*, and the *average nearest neighbor distance*. The first two approaches can only be calculated after the geographical space has been partitioned into a set of disjoint areas. In this work, we do that adopting an equally-spaced grid, and divide the $100km^2$ region representing each area using different resolutions, including a grid of 10x10 (i.e. each cell is a square of side 1 km), 20x20 and 50x50 cells.

4.1.1 Entropy. It can be used to measures how equally activities are distributed across the grid. Let X be a discrete random variable modeling the positions of an individual ending up in n different fields [5]. The entropy is defined as [35]:

$$E(X) = -\sum_{i=1}^{n} P(x_i) \log P(x_i)$$

where $\{x_1, \ldots, x_n\}$ are the possible values of X and $P(x_i)$ is the probability of X being in state i. For maximum entropy ($log(n)$) there is an equal amount of activity in all fields; for minimum entropy (0) all the activity is amassed in a single field. In order to compare entropy scores of different-sized grids, the measure must be normalized by dividing it by the expected entropy of a uniform distribution, i.e., $log(n)$.

4.1.2 Moran's I. It overcomes the entropy weakness by considering how the fields are positioned in space: spatial autocorrelation [33] that represents the degree to which the fields' values are correlated to the value of neighboring fields. For spatial autocorrelation, the nearness between all pairs of fields must be defined with a so-called *weight matrix* w, where w_{ij} is the nearness between nodes i and j. A simple form of weight matrix is an adjacency matrix, with the value 1 if fields are adjacent, 0 otherwise. An important difference to the entropy is that spatial autocorrelation has two directions. A high autocorrelation indicates that values of the same magnitude are prone to be next to each other, while a low autocorrelation means that similar values are less likely to be near each other than under random positioning. Somewhere in between lies a value of autocorrelation in which the population of the fields is how one would expect it to be under a random distribution with no spatial autocorrelation. The most famous autocorrelation measures is *Moran's I* [26]:

$$I(X) = \frac{N}{W} \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

where N is the number of fields, x is the amount of activity or population, $\bar{x}$ is the average field value, and W is the sum of all the weights. The minimum and maximum values of *Moran's I* depend on the weight matrix. We highlight that the absence of

autocorrelation is given at *Moran's I* equals to $-1/(N-1)$, that tends to zero in grids with an high amounts of fields.

4.1.3 Nearest Neighbor Distance. The *Average Nearest Neighbor Distance* (ANND) is not dependent on a grid and its parameters. For every point, the distance to its nearest neighbor is calculated. The mean of those values is the *ANND* :

$$ANND = \frac{\sum_i min(d_i)}{N}$$

where d_i is a vector containing the distances of point i to all the other points, and N is the amount of points. The lower the *ANND* , the higher is the average spatial concentration in the areas surrounding the points. We highlight that this definition bears a similar weakness as the entropy. The expected *ANND* under assumption of a uniform distribution of points across the area is the *Mean Random Nearest Neighbor Distance* (*MRNND*) $MRNND = 0.5\sqrt{A/N}$, where A is the surface of the area and N the amount of points. By dividing the *ANND* by *MRNND* we obtain the *Nearest Neighbor Index* (*NNI*) which is comparable among samples with different sizes and areas. A *NNI* smaller than 1 indicates a higher spatial concentration than in a random case, whilst value above 1 shows that the points are spread out across the map more than one expects in a random scenario.

4.2 Flows in a Grid Network

In order to capture the information about flows in urban regions, the data can be transformed into a directed weighted graph that represents the flow of the people's trajectories:

- a set of nodes V representing places that are origins and destinations of trajectories,
- a set of edges E representing the directed connections between the nodes,
- a weight function $w : E \rightarrow \mathbb{R}$ that maps each edge to a weight, which indicates the amount of trajectories that occur along the edge.

The map is split into fields of a grid and all origins and destinations of the trajectories in the area are assigned to the field in which they lie. The network is created by assigning every node to a cell, and to each edge the weight the amount of flows occurring along the edge. The weight function w is equivalent to an origin destination matrix. The network allows us to gain knowledge about the structure of a region by looking at the properties of the resulting network described in the following.

4.2.1 Node Degrees. A basic property of the network is the distribution of its degrees. Degree is hereby defined as the total traffic (sum of in- and out-flow) of a grid field. This measure is sometimes also referred to as node-flux [34].

4.2.2 Louvain Modularity. An interesting quality of networks is the degree to which nodes can be partitioned into groups, such that the connectivity is high within those groups, and low in between. In the context of urban regions, the corresponding question is: can the city be split into areas that are relatively autonomous and have only low interaction between them? In network science, *modularity* measures this property for a given partitioning: a graph partitioning separates the graph's nodes into non-overlapping communities. Modularity shows the difference between the relative amount of inner-community links and the expected relative amount under random linking in a non-directed weighted graph [4]. The modularity goes from -1 to $+1$, where 0 marks the value expected in a network where all possible edges have the same expected weight. We highlight that the direction of traffic flow is not important here. Thus, the grid networks in this work are transformed into non-directed networks before the modularity is calculated. Modularity does not describe a network on its own, but a network along with its partition. In order to quantify how well an urban region is separable into different subareas we adopt the *Louvain Algorithm* [6] that does not guarantee an optimal solution but it performs well empirically.

4.2.3 Interaction Models. The flow network allows us to test how well the empirical data aligns with two established models that describe human interaction in space. The *Gravitation Model* [1] idea is that the traffic flow from place i to place j depends on the origin population m_i and the destination population n_j. Highly populated places, attract flow towards them. The classic model predicts the traffic flow from i to j which have a distance of r as $G_{ij} = Am_i^{\alpha} n_j^{\beta} / r^{\gamma}$, where A is a normalization factor, and α, β, γ are the model's parameters. They can be optimized by multiple regression when fitting data to the model. In this work we adopt a simpler model [25] with $\alpha = \beta = 1$. The *Radiation Model* [37] updates G_{ij} by introducing s_{ij} that is the population within a circle around place i, with a radius of its distance to place j, minus m_i and n_j. The intuition is that outgoing trips are being attracted by nearby populations [25]. It predicts the flow T_{ij} as $\frac{T_i}{1-\frac{m_i}{M}} \frac{m_i n_j}{(m_i+s_{ij})(m_i+n_j+s_{ij})}$ where T_i is the sum of outflows from i, and $M = \sum_i m_i$ is the total sample population.

4.3 Individual Mobility

Here we consider the mobility at level of individual users. From this perspective, urban regions can be described by aggregated values of their inhabitants' mobility, therefore a set of statistics are calculated for each individual from their trajectories:

- Average distance and duration per trip
- Average driving distance and duration per day
- Average amount of trips per day

Also, following the methods described in [21, 31], individuals' mobility data can be transformed into *Individual Mobility Networks* (*IMN*), which is a representation of a person's travel behavior in the form of a weighted directed network, where the set of nodes V represents places that are visited once or repeatedly by the individual, and the edges E represent trajectories from one of those places to another. The edge's weights model the amount of times was followed a trajectory from one node to another. From an *IMN*, we can describe the individuals travel behavior with the following indicators:

- *Size of the network*: number of nodes and edges.
- *Temporal-uncorrelated entropy*: measure how equally the different places of the *IMN* are visited.
- *Radius of gyration* [28]: approximates the average distance of an individual from its center of mass [18].
- *Regularity of trajectories*: percentage of trips that are driven more often than a certain threshold per time [19, 21].
- *Modularity*: the *Louvain Algorithm* [6] applied to the *IMN*.

4.4 Roads and Traffic

4.4.1 Static Road Network. This section focuses on the road network modeled as a directed graph ($G = (E, V)$, where V is a set of nodes representing roads intersections, E is the set of directed edges which model the the road segments, and $l : E \rightarrow \mathbb{R}$

maps each edge to its length in meters. Some *basic statistics* of the road network can be calculated:

(1) amount of edges and nodes/node density
(2) amount of intersections/intersection density
(3) average node degree/average intersection degree
(4) total length of edges/mean edge length

In addition, since nodes in any network can be evaluated w.r.t. their centrality, we evaluate the *road network's closeness centrality* in terms of the length of the shortest path to any given node. The average of those path lengths is a node's average farness from other nodes. The reciprocal of this value is a node's closeness centrality $C(x) = \frac{1}{\sum_y d(y,x)}$, where x and y are nodes and the d returns the length of the shortest path between its arguments. As distance function we consider the length as the summed road lengths of the edges of the shortest path [30].

4.4.2 Traffic in the Road Network. To investigate how traffic is distributed in a road network one must *map match* the sequences of GPS locations that represent the trajectories to nodes and edges in the road network. There is a variety of algorithms that handle this problem, such as hidden Markov models [27]. In the case study of this work, a simpler algorithm was implemented due to the high reliability of the data. It independently maps every point of a trajectory to a node in the road network. The nodes are then connected and build a path that describes the individual's trajectory.

Given a map matching, it is possible to create a function that reveals the fraction of total traffic that flows through a given percentage of the most dense roads. For this purpose, all edges are sorted by their traffic flow in a non-ascending order. Cumulative traffic, measured as $\#cars \times meters$, is calculated for the end of every edge by multiplying the edge length with the amount of traffic flow and adding the result to the previous amount of cumulative traffic. The intermediary values within edges can be calculated by linear interpolation. For any given percentage of roads, the percentage of traffic in those roads is calculated by dividing the cumulative traffic until that point by the total amount of traffic.

5 GLOBAL CITY INDICATORS

In this section we introduce the *global* city indicators designed to compare two cities. To compare and cluster cities in groups, we need some quantitative features. Therefore, we have to define some metrics describing a city with respect to traffic. A possible approach is to exploit again a *network* structure where each city (in our case study, 276 municipalities in Tuscany) is a node, and edges are drawn based on the trajectories between them. Starting from the trajectories we infer descriptive attributes from two perspectives: (i) graph measures from the complete network of cities; (ii) graph measures from the ego-network of each city.

5.1 Complete Network of Cities

We can derive a set of global indicators through a *network of cities* as described in the following. Given the trajectories on the territory, we can derive an *Origin-Destination Matrix* (OD), which measures the number of trips that starts from city A and ends in city B for each pair (A, B). Since connections established through very few trajectories might be not significant, a threshold is needed to establish if an edge should be drawn. In our case study, after empirical evaluation, we fixed this threshold to 110 trajectories by analyzing the results yielded by different values

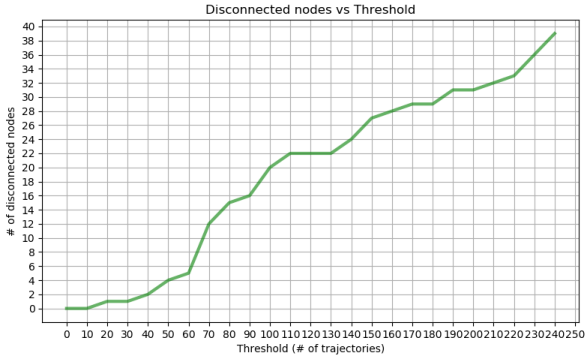

Figure 2: Disconnected nodes vs. flow threshold.

through Figure 2. The plot shows the number of disconnected nodes corresponding to a selected threshold. The fraction of "isolated" cities grows as the threshold increases, but there is a little *plateau* between 110 and 130, which led to our choice. With the selected threshold, the final graph consists of 276 nodes (corresponding to municipalities), 22 of which are disconnected from the *giant component*.

The properties related to each node of the network constitute the first set of attributes to be considered for clustering:

- ***Self-loops***: # trajectories starting and ending in that node.
- ***In/Out degree***: fraction of nodes its incoming/outgoing edges are connected to.
- ***Closeness***: the closeness centrality of a node u is the reciprocal of the average shortest path distance (see Section 4.4).
- ***Betweenness***: the betweenness of a node v is the sum of the fraction of all-pairs shortest paths that pass through v.
- ***Clustering coefficient***: the local clustering coefficient C_i for a vertex v_i is given by the proportion of links between the vertices within its neighborhood divided by the number of links that could possibly exist between them.
- ***Radius of Gyration***: the radius of gyration of a city c is defined as $r_g(c) = \sqrt{\frac{1}{N}\sum_i^N w_i(r_i - r_{cm})^2}$, where N is the total number of travels from c, w_i is the number of travels from c to i, r_i is the pair of coordinates of location i and r_{cm} is the center of mass (i.e., the average position) of the visited cities starting from c.
- ***Random Entropy***: the random entropy captures the degree of predictability of the destination starting from a city i if each location is visited with equal probability $S_{ran} = \log_2 M$, where M is the number of distinct cities visited starting from city i;
- ***Uncorrelated Entropy***: the temporal-uncorrelated entropy is the historical probability that a location j was visited starting from a city i, characterizing the heterogeneity its of visitation patterns $S_{unc} = -\sum_j^N p_j \log p_j$ where p_j is i's probability of visiting location j. We can also normalize the uncorrelated entropy by dividing it by $\log_2 N$.

5.2 Ego-Networks

In Social Network Analysis, it is usual to refer to *Ego Networks* as social networks made of an individual (called *ego*) along with all the social links he has with other users (called *alters*)[3, 10]. Several fundamental properties of social relationships can be characterized by studying them. Adapting the terms to the present

context, we can obtain an ego network for each city, where the ego is the city itself and the alters are its neighbors. The additional set of attributes obtained consists of:

- ***Number of nodes*** of the ego network.
- ***Number of edges*** of the ego network.
- ***Average clustering coefficient***: the clustering coefficient is the average $C = \frac{1}{n}\sum_{v \in G} c_v$, where n is the nbr. of nodes in G and c_v is the clustering coefficient of each node;.
- ***Diameter***: is the longest shortest path of the ego network.
- ***Assortativity***: is measured as the Pearson correlation coefficient of degree between pairs of linked nodes. It measures the preference for a network's nodes to attach to others that are similar in some way.

6 CASE STUDY: TRANSFER-COMPLIANT GEOGRAPHICAL LOCATIONS

The huge amount of urban data generated by smartphones, vehicles, and infrastructures (e.g., traffic cameras, air quality monitoring stations) opens up new opportunities to learn about city dynamics from a variety of perspectives and facilitates various smart city applications for traffic monitoring, public safety, urban planning, etc. – all contributing to what is called *urban computing*.

However, there are some questions that remains still almost unexplored: what if the administration of a city wanted to predict the impact of an event on the urban mobility without having historical data on it? Is it possible to infer some useful insights exploiting the experience gained by other municipalities? Can knowledge be transferred from any city or are there some constraints? How can you compare two cities, for example in terms of urban mobility? Lately there have been different attempts to overcome the data scarcity issue in "new" urban contexts. All these studies have in common the application of *Transfer Learning*, a very broad family of approaches which focuses on developing methods to transfer knowledge learned in one or more *"source tasks"*, and use it to improve learning in a related *"target task"*. This section studies these questions in the context of Machine Learning (ML) and big data analytics for mobility data. In particular, our goal it to verify the feasibility of a *model transfer*, i.e., a ML model is trained in the source domain and then transferred to the target domain, in the prediction of urban traffic, exploiting the city indicators developed in the previous sections.

The basic idea is that cities that are similar can be represented by the same model more easily than very different cities. For instance, a highly populated city with heavy traffic and users that frequently make long trips is expected to have mobility dynamics very different from small, country-side cities with low traffic. The approach proposed in this section is developed in three steps: first, using a similarity measure between cities based on the indicators presented in Sections 4 and 5, cities are clustered into similarity groups; next, for each city a traffic prediction task is defined, which is approached through a standard machine learning solution (XGBoost regression [8]); finally, the prediction model of a city is applied to make predictions in each of the others, aiming to test whether cities in the same cluster show a better transferability of their models.

6.1 City Clustering

In this step, the city indicators built in the previous sections are first preprocessed and filtered, and then used to cluster cities.

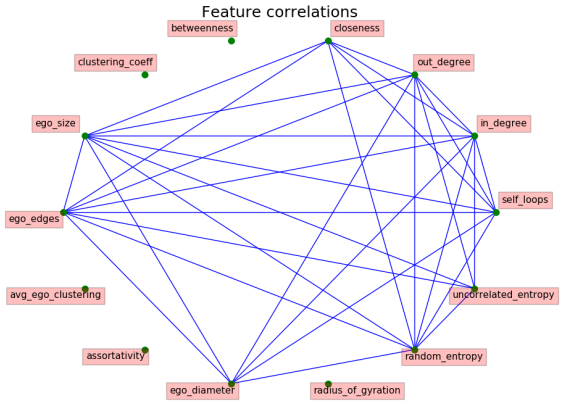

Figure 3: Network of correlations for the first set of attributes (total graph).

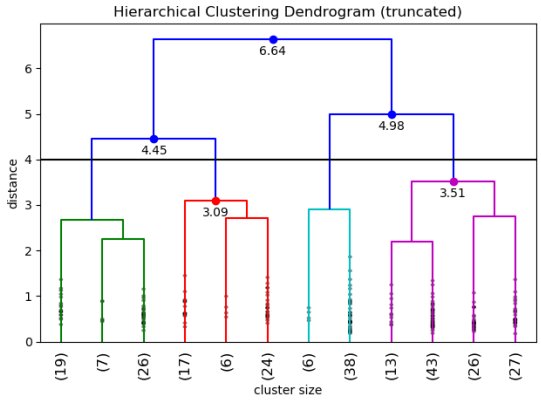

Figure 4: Dendogram and selected clusters.

Preprocessing and Feature Selection. Since the range of different indicators varies widely, we applied a form of normalization to make them homogeneous. We adopted the *min-max scaling*, where feature are re-scaled in the interval $[0, 1]$. Then, we performed a study of correlation on each set of features (local and global) to eliminate unnecessary ones. To efficiently filter them, we adopted a network-based correlations finder, where the features are interpreted as nodes of a graph, and a link is drawn between two features if they are highly correlated. As evaluation metrics, the standard *Pearson's Correlation Coefficient* is used [29].

Considering each couple of features (i, j), an edge is drawn if $\rho_{i,j} > 0.65$. The result obtained on the global features is shown as example in Figure 3. The removal of features is an iterative process that removes the node (feature) with the highest degree (thus is correlated to the highest number of non-filtered features) and repeats until the average degree of the network is 0. The remaining nodes are the features which are preserved. This preprocessing step is applied to global and local indicators separately, and then on the set of survived features of both categories. Applying the procedure to our case study, the initial set of indicators, composed of a total of 178 measures, was reduced to 21 features.

Hierarchical Clustering. The city clustering step has been realized through a Hierarchical agglomerative clustering schema, adopting Ward's linkage criterion, which at each step of aggregation aims to minimize the total within-cluster variance. In our case study, a small fraction of cities resulted to be disconnected

cluster id	# of cities	% of cities
0	22	8.0
1	53	19.2
2	47	17.0
3	110	39.9
4	44	15.9

Table 1: Cluster Population

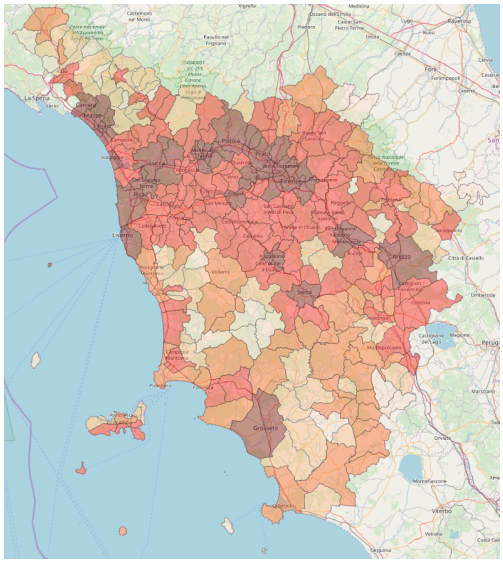

Figure 5: Map of clustered municipalities. Colors white, yellow, orange, red and dark red correspond resp. to clusters 0, 1, 2, 3 and 4.

from all the others in terms of flows, thus making them outliers w.r.t. the global features (e.g., the assortativity measure is null). Therefore, we decided to put them in a separate cluster, and apply the hierarchical clustering on the remaining ones. The results of applying the clustering to our dataset is shown in Figure 4 as a dendogram of the hierarchical clusters found (notice that the dendogram is truncated, in order to show only the last 12 aggregations. Based on the gaps between splits/merge points in the dendogram, the aggregation is stopped at distance 4.0, yielding four clusters. To these, we add another cluster (id 0) containing the isolated cities. A summary of clusters' size is in Table 1.

An analysis of the properties of each cluster reveals that they may be distinguished based on the kind of traffic flows they involve. Also, clusters are depicted on the map in Figure 5. Cluster 0 was named *Disconnected*, since it is composed by the nodes not connected in the inter-city flows network. These municipalities also have a low entropy and low Moran's I score, meaning a not significant pattern of traffic, and most of them are located at the boundary of Tuscany and in the country-side areas, where there is a lower concentration of roads. Cluster 1, named *Self Sufficient*, is characterized by high entropy, high modularity and high fraction of regular trips, yet a low radius of gyration and low diameter of the associated ego networks. Also, they are mostly far from the highways that cross the region. Cluster 2, called *Visited Sites*, have a very low entropy (almost as low as those in the disconnected group), low modularity and the lowest fraction of regular trips, and yet a relatively high betweenness. Cluster 3 was named *Drive Through*, as these cities are crossed by a great flow of traffic, which is however basically coming

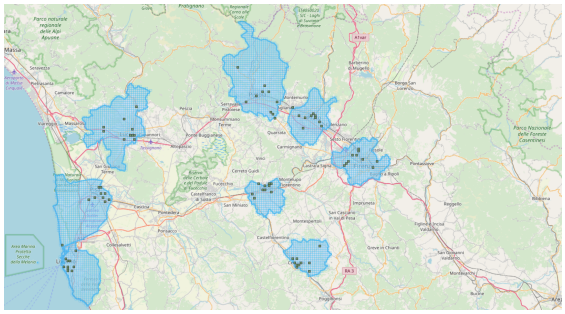

Figure 6: Selected cells for some municipalities.

from outside or going outside. Indeed, they have high values for entropy and low values for Moran's I, the highest number of nodes regularly visited from users and a large ego network radius. This cluster is the most populated, comprising almost 40% of the dataset. Finally, cluster 4 was called *Hubs*, since it comprises all the biggest cities, encompassing most of the busiest roads in Tuscany. Municipalities are pretty similar to those belonging to cluster 3, excepted that they have a large Moran's I, which reflects the presence of specific patterns within the city.

6.2 Traffic Forecasting in City Grids

Urban traffic prediction is a discipline that aims to exploit ML models to capture hidden traffic characteristics from substantial historical mobility data, making then use of trained models to predict traffic conditions in the future [24]. However, there is a main problem to face: is it possible to extract specific traffic patterns that reflect the peculiarities of a city structure?

A Grid to Split the City. Following one of the most used approach in traffic prediction problems [24], we divide every geographical area corresponding to municipalities in adjacent squared cells having side of 0.5 km, and our predictive objective is to forecast the traffic flow that crosses a given cell. In our case study we select a subset of representative cells and, in order to avoid the possible issues emerging when a random or top-frequency subset is selected, we adopt a mixed approach, randomly selecting 5 cells among those having a traffic volume above the 90^{th} percentile over the municipality, and other 5 cells among those having a traffic volume between the 80^{th} and the 90^{th} percentiles.

Time Series Preprocessing. Based on the trajectories that cross the representative cells identified above, we compute a time series for each cell with a 1-hour sampling rate, by counting the number of vehicles that crossed the cell within each hour of each day. A first operation performed was to compute a *moving-average smoothing* of the time series, since a preliminary test with the *Augmented Dickey-Fuller* test (ADF) [14] reveals that they are not stationary, i.e. it could not be rejected the null hypothesis that a *unit root* is present in the time series sample (ADF=-2.38 against a critical 90% threshold at -2.57). On the contrary, after smoothing, the null hypothesis is rejected with a very large confidence (ADF=-5.57 against a 99% threshold at -3.43, p-value = $2 \cdot 10^{-5}$).

Predictive Features. Similarly to what done by several time series forecasting solutions [12], we base our predictions for the next value of the time series on more recent observations of the same time series. In particular, we adopt as basic features the 24 most recent lagged values, i.e., the observations of the last 24 hours. We remark that in this simplified approach we do not

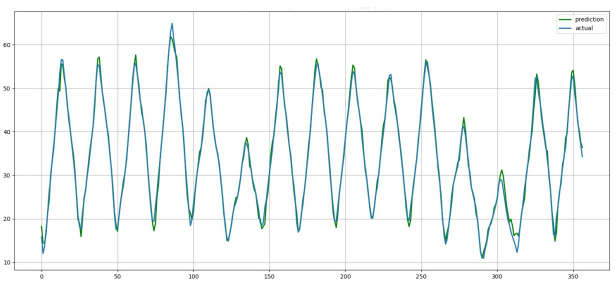

Figure 7: XGBoost traffic forecasting on Florence (green) against real values (blue). The two curves differ in very few points.

include features about other time series in the same municipality, as done in more complex solutions that exploit the spatial autocorrelation of this kind of phenomena.

Another important property that can be encoded is related to the weekday; at this regard, we introduce the Boolean feature *is_weekend* that is true if the weekday is Saturday or Sunday and false otherwise, since we expect to see different behaviors in the weekends. Finally, we can encode information about a weekday by inserting the average traffic volume at that day.

Having a total of 26 new features, we can now try to forecast the smoothed time series.

Predictive Model. As regressive model, we selected the popular and effective algorithm *XGBoost* [8]. XGBoost has proved to be highly reliable in regression tasks, providing in general a good accuracy of predictions and remarkable speed of execution, yielding good results in term of robustness with its default settings, which simplifies our task. XGBoost adopts a Boosting procedure, i.e., is a ML ensemble meta-algorithm for primarily reducing bias and variance in supervised learning, where a set of weak learners is turned into a single strong learner.

In Figure 7 we can see an example of XGBoost predictions exploiting the features previously introduced over the municipality of Florence, which shows results very close to the real values. The model performance is evaluated through the standard *Normalized Root Mean Squared Error*, defined as RMSE $= \sqrt{\frac{\sum_{t=1}^{T}(\hat{y}_t - y_t)^2}{T}}$, having predicted values $\hat{y}_t$ for times t of a regression's dependent variable y_t, with variables observed over T times. RMSE is always non-negative, the lower is the value the better are the predictions. Since RMSE is scale-dependent, we adopt the Normalized RMSE (NRMSE), computed as: NRMSE $= \frac{\text{RMSE}}{\sigma}$, where σ is the standard deviation of the observed values.

Empirical evaluation shows that the most important feature is the value of traffic 1 hour before, as expected, while the previous hours have all a comparable influence. Instead, it is apparently almost irrelevant to know if a day is a week-end day or not.

6.3 Testing Model Transferability

In this section we study the transferability of the predictive models built above, and its relation with the similarity groups found through clustering. The hypothesis we want to test is that the similarity based on our city indicators is indeed useful to identify groups of areas such that any model built from an area in the cluster is usable in other areas within the same cluster.

The first step is to split the traffic time series of each city in training and test sets. In this way it is possible to obtain a matrix

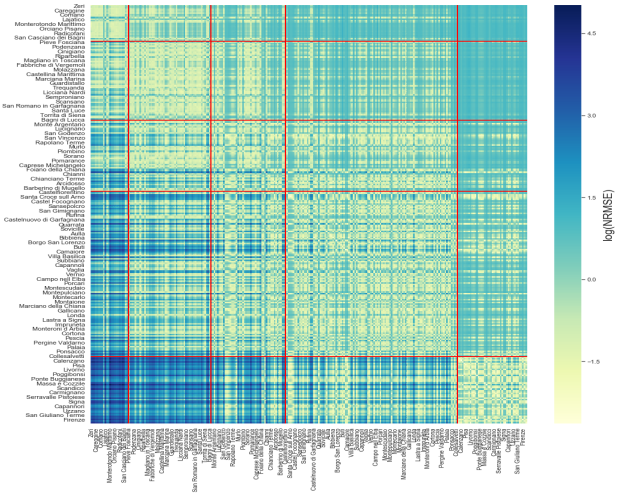

Figure 8: Transfer scores matrix with cluster separation (red lines). Each row/column represents a municipality.

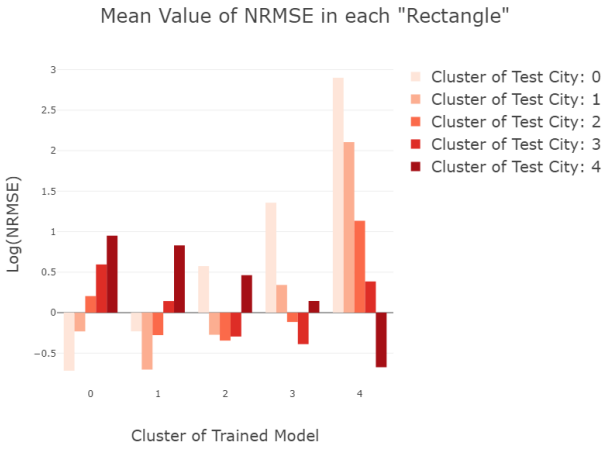

Figure 9: NMRSE mean values for all train-test pairs.

of prediction scores where on the rows there are the cities in which the model is trained and in the columns those where the model is tested. The algorithm implemented iteratively trains a model on each city, tests it against all the cities and fills the score matrix with the corresponding NRMSE score obtained. To enable a more meaningful comparison, NRMSE scores are *log*-transformed to reduce the skewness.

The final result is visually shown Figure 8 which shows the transfer scores by sorting the cities based on their cluster belonging. Keeping in mind that the squares around the diagonal represent training and testing on cities of the same cluster, while the other rectangles depict training and testing on different clusters, we can observe:

(1) the transfer is far better between cities of the same cluster (the NRMSE values obtained making predictions on a municipality using a model built on a different one is lower if the two belong to the same cluster);
(2) it is worth noting that also cluster 0, that we built up artificially behave exactly as the others;
(3) the matrix is not symmetric: training on city A and testing on B is different from training on B and testing on A.

The trend noticed in Figure 8 can be better identified by computing the average error among the clusters, i.e., considering all the possible *source* areas in each cluster (where the models are built) and all the possible *target* areas in each other cluster (where the model is tested), including the case source = target. This is shown in Figure 9, where each bar corresponds to one of the *rectangles* outlined in red in Figure 8. We observe that the lowest mean values are always those corresponding to central squares, where the source and the target cities are from the same cluster. Overall, these results confirm our hypothesis, namely that the similarity of cities based on our city indicators is a good proxy of model transferability, at least for the simple predictive task we adopted.

7 CONCLUSIONS

In this work we have defined a large array of local and global city indicators, we have calculated them on a real case study, and we have proved that they can be successfully exploited in a task of mobility transfer learning. In particular, we have clustered municipalities based on the mobility behavior described by the city indicators. Then, we have assessed the transferability of a machine learning model for traffic forecasting. Experimental results show that models trained on a municipality perform markedly better when tested on other municipalities belonging to the same cluster, and thus more similar (according to the city indicators) to the first one.

As future work, it would be interesting to extend the set of features used to describe a city, for example including census and cartographic data or some indicators related to economy, industry level and information about the most florid commercial activities in each area. All these extra properties would also help to interpret the results of clustering, to identify patterns of similarity and eventually to supervise with some kind of feedback the allocation of a city to a determinate group. More models should be analyzed and compared to evaluate which is the most effective. Finally, the approach presented here works on a city-to-city transfer, namely the model of a single city is used to make prediction on the destination city. That assumes that there exists at least one origin city that is similar enough to perform the transfer. Alternatively, all the data and known city models can be exploited to achieve better prediction on the target city.

ACKNOWLEDGMENTS

This work is partially supported by the European Community H2020 programme under the funding scheme *Track &Know* (Big Data for Mobility Tracking Knowledge Extraction in Urban Areas), G.A. 780754, https://trackandknowproject.eu/ and *SoBigData++*, G.A. 871042, http://www.sobigdata.eu. We thank Christoph Pfaltz and Matteo Centonze for their contribution to preliminary materials of this work.

REFERENCES

[1] W Alonso. 1976. A Theory of Movements: Introduction. *Working Paper 266* (1976).
[2] Gennady Andrienko et al. 2020. (So) Big Data and the transformation of the city. *International Journal of Data Science and Analytics* (2020).
[3] Valerio Arnaboldi, Marco Conti, Andrea Passarella, and Robin IM Dunbar. 2017. Online social networks and information diffusion: The role of ego networks. *Online Social Networks and Media* 1 (2017), 44–55.
[4] Albert-László Barabási and Márton Pósfai. 2016. *Network science.* Cambridge University Press, Cambridge.
[5] Michael Batty. [n.d.]. Spatial Entropy. *Geographical Analysis* 6, 1 ([n. d.]), 1–31. https://doi.org/10.1111/j.1538-4632.1974.tb01014.x
[6] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008, 10 (2008), P10008.
[7] Harold Carter. 1995. *The Study of Urban Geography.* E. Arnold publications.
[8] Tianqi Chen et al. 2016. XGBoost: A Scalable Tree Boosting System.
[9] CITEAIR consortium. 2007. Air Quality in Europe web site. http://www.airqualitynow.eu/ [Online; accessed 21-December-2020].
[10] Michele Coscia, Giulio Rossetti, et al. 2012. Demon: a local-first discovery method for overlapping communities. In *ACM SIGKDD*. 615–623.
[11] H.G De Sherbinin, A.; Bittar. London, UK, 2003. The Role of Sustainability Indicators as a Tool for Assessing Territorial. *Environmental Competitiveness; International Forum for Rural Development* (London, UK, 2003).
[12] George E. P. Box et al. 2015. *Time Series Analysis: Forecasting and Control.* John Wiley and Sons.
[13] Liu F. et al. 2020. Citywide Traffic Analysis Based on the Combination of Visual and Analytic Approaches. *J geovis spat anal* 4, 15 (2020).
[14] W. A. Fuller. 1976. *Introduction to Statistical Time Series.* John Wiley and Sons.
[15] Fosca Giannotti, Mirco Nanni, Dino Pedreschi, et al. 2011. Unveiling the complexity of human mobility by querying and mining massive trajectory data. *The VLDB Journal* 20, 5 (Oct. 2011), 695–719.
[16] Fosca Giannotti, Mirco Nanni, Fabio Pinelli, and Dino Pedreschi. 2007. Trajectory pattern mining. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. 330–339.
[17] D. Gillis, I. Semanjski, and D. Lauwers. 2015. How to Monitor Sustainable Mobility in Cities? Literature Review in the Frame of Creating a Set of Sustainable Mobility Indicators. *Sustainability* 8 (2015), 29.
[18] Marta C. Gonzalez, Cesar A. Hidalgo, et al. 2008. Understanding individual human mobility patterns. *Nature* 453, 7196 (June 2008), 779–782.
[19] Riccardo Guidotti and Mirco Nanni. 2020. Crash Prediction and Risk Assessment with Individual Mobility Networks. In *2020 21st IEEE International Conference on Mobile Data Management (MDM)*. IEEE, 89–98.
[20] Riccardo Guidotti, Mirco Nanni, Salvatore Rinzivillo, Dino Pedreschi, and Fosca Giannotti. 2017. Never drive alone: Boosting carpooling with network analysis. *Information Systems* 64 (2017), 237–257.
[21] Riccardo Guidotti, Roberto Trasarti, Mirco Nanni, Fosca Giannotti, and Dino Pedreschi. 2017. There's a path for everyone: A data-driven personal model reproducing mobility agendas. In *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 303–312.
[22] M. K. Jat, P. K. Garg, and D. Khare. 2008. Modelling of urban growth using spatial analysis techniques: a case study of Ajmer city (India). *International Journal of Remote Sensing* 29, 2 (2008), 543–567. https://doi.org/10.1080/01431160701280983 arXiv:https://doi.org/10.1080/01431160701280983
[23] Gabriel Lang, Eric Marcon, and Florence Puech. 2016. Distance-based Measures of Spatial Concentration: Introducing a Relative Density Function. (Sept. 2016).
[24] Z. Liu, Z. Li, K. Wu, and M. Li. 2018. Urban Traffic Prediction from Mobility Data Using Deep Learning. *IEEE Network* 32, 4 (2018), 40–46. https://doi.org/10.1109/MNET.2018.1700411
[25] A Paolo Masucci, Joan Serras, Anders Johansson, and Michael Batty. 2013. Gravity versus radiation models: On the importance of scale and heterogeneity in commuting flows. *Physical Review E* 88, 2 (2013), 022812.
[26] P. A. P. Moran. 1950. Notes on Continuous Stochastic Phenomena. *Biometrika* 37, 1/2 (1950), 17–23.
[27] Paul Newson and John Krumm. 2009. Hidden markov map matching through noise and sparseness. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, 336–343.
[28] Luca Pappalardo, Salvatore Rinzivillo, Zehui Qu, Dino Pedreschi, and Fosca Giannotti. 2013. Understanding the patterns of car travel. *The European Physical Journal Special Topics* 215, 1 (01 Jan 2013), 61–73.
[29] Karl Pearson. 1895. Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London* 58 (1895), 240–242.
[30] S. Porta, P. Crucitti, and V. Latora. 2006. Centrality measures in spatial networks of urban streets. *Physical Review E* 73, 3, part 2 (24 3 2006), 036125–1.
[31] Salvatore Rinzivillo, Lorenzo Gabrielli, Mirco Nanni, Luca Pappalardo, Dino Pedreschi, and Fosca Giannotti. 2014. The purpose of motion: Learning activities from individual mobility networks. In *2014 International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 312–318.
[32] Antônio Nélson Rodrigues da Silva et al. 2015. A comparative evaluation of mobility conditions in selected cities of the five Brazilian regions. *Transport Policy* 37 (2015), 147 – 156.
[33] P.A. Rogerson. 2010. *Statistical Methods for Geography: A Student's Guide.* SAGE Publications. https://books.google.ch/books?id=Zz69Ab8i0QsC
[34] Meead Saberi, Hani S. Mahmassani, Dirk Brockmann, and Amir Hosseini. 2017. A complex network perspective for characterizing urban travel demand patterns: graph theoretical analysis of large-scale origin–destination demand networks. *Transportation* 44, 6 (November 2017), 1383–1402.
[35] Claude Elwood Shannon. 1948. A Mathematical Theory of Communication. *The Bell System Technical Journal* 27, 3 (7 1948), 379–423.
[36] Sulochana Shekhar. 2004. Urban sprawl assessment Entropy approach. *GIS Development* 2004, Vol 8 issue 5, Page ., 6 Pages (05 2004), 43 – 48.
[37] Filippo Simini, Marta C. Gonzalez, Amos Maritan, et al. 2012. A universal model for mobility and migration patterns. *Nature* 484, 7392 (2012), 96–100.
[38] Pavlos Tafidis et al. 2017. Sustainable urban mobility indicators: policy versus practice in the case of Greek cities. *Transportation Research Procedia* 24 (2017), 304 – 312. 3rd Conference on Sustainable Urban Mobility.
[39] Roberto Trasarti, Riccardo Guidotti, Anna Monreale, and Fosca Giannotti. 2017. Myway: Location prediction via mobility profiling. *Information Systems* 64 (2017), 350–367.