Finding and Predicting City regions via clustering.

Petr Rubin
December 19, 2020

1. Introduction

1.1 Background

City region is a term in use since about 1950 by urbanists, economists and urban planners to mean a metropolitan area and hinterland, often having a shared administration.
Administrative city divisions often do not reflect the actual condition of different areas of the city. Of course, the division of a city into regions can be carried out according to various criteria. Therefore, the division should be performed depending on the set administrative or business tasks. The actual boundaries of urban regions may change over time due to the intensity of activity in and around certain points. Assessing and predicting such points could play an important role in city and business planning.

1.2 Problem

Data that can contribute to the regional division of cities can include city venues. Such data can be accessed using the Foursquare API. The aim of this project is to cluster regions of Moscow using metro stations as reference points and on the basis of the results to predict the the regional division of other cities.

2. Data acquisition and cleaning

2.1 Data sources

The project used the Foursquare API. The Foursquare API provides location data and specifically venue data for various locations. Coordinates of Moscow metro stations can be found here, coordinates of St. Petersburg metro stations were scraped from here.

2.2 Data preparation

The city can be represented as a collection of venues. Obviously, the venues are not evenly distributed within the city, so the question arises: what to consider as a reference point. Daniel Preotiuc-Pietro et al. propose the following representation

options: city-centric, grid-centric and neighborhood-centric representation. The first option does not allow us to identify differences between the cities. To implement the second and third, there is a problem with determining the geographical coordinates of the grid and the availability of data on the coordinates of city neighborhoods.
In this context, it was decided to use subway stations as starting points. The data on the geographic coordinates of the metro stations were scraped from open sources using the BeautifulSoup and converted to pandas dataframe.

Using folium data about the location of Moscow metro stations are displayed on the map.

Using the Foursquare API, data were obtained for venues around each station within a radius of 1000 meters. The data in json format was filtered and converted into pandas dataframe.

	Station	Station Latitude	Station Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Академическая	55.68808	37.57501	Вкусвилл	55.686904	37.575196	Health Food Store
1	Академическая	55.68808	37.57501	Здоров.ру	55.687911	37.571558	Pharmacy
2	Академическая	55.68808	37.57501	Billy McDaniel	55.688104	37.571608	Pub
3	Академическая	55.68808	37.57501	Академический парк	55.691777	37.568886	Park
4	Академическая	55.68808	37.57501	HobbyGames	55.688509	37.570087	Toy / Game Store

Then one-hot encoding were applied. Rows were grouped by station and by taking the mean of the frequency of occurrence of each category. There are 412 uniques categories.

	Station	Accessories Store	Adult Boutique	Advertising Agency	American Restaurant	Amphitheater	Aquarium	Arcade	Arepa Restaurant	Argentinian Restaurant	...	Water Park	Waterfr
0	Савеловская	0.0	0.0	0.0	0.000000	0.0	0.0	0.010417	0.0	0.0	...	0.0	
1	Свиблово	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	...	0.0	
2	Севастопольская	0.0	0.0	0.0	0.013333	0.0	0.0	0.026667	0.0	0.0	...	0.0	
3	Семеновская	0.0	0.0	0.0	0.000000	0.0	0.0	0.020000	0.0	0.0	...	0.0	
4	Серпуховская	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	...	0.0	

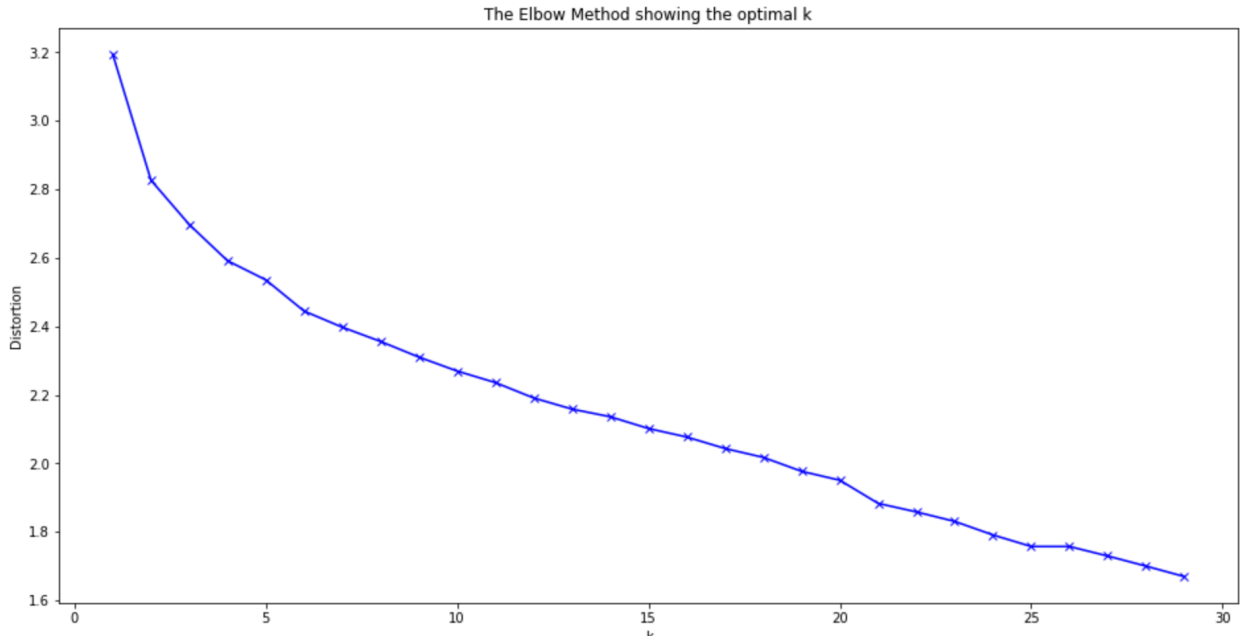
3. Methodology.

3.1 Clustering Moscow metro stations

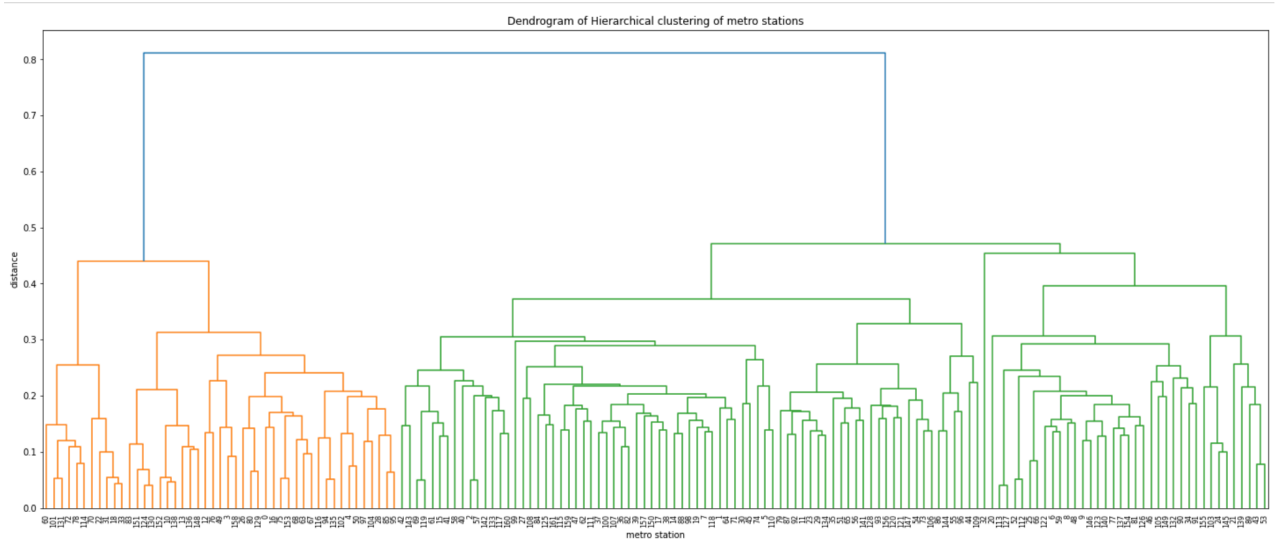
Let's see if the stations really differ in the frequency of venues around them. For each station, 10 most common venues were selected and sorted. As can be seen from the example of several stations there are significant differences, so in the area of some stations are dominated by catering facilities, somewhere there are parks, somewhere gyms.

	Station	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Савеловская	Coffee Shop	Gym / Fitness Center	Dance Studio	Electronics Store	Photography Studio	Martial Arts School	Restaurant	Mobile Phone Shop	Caucasian Restaurant	Vietnamese Restaurant
1	Свиблово	Cosmetics Shop	Pizza Place	Food & Drink Shop	Gym	Gym / Fitness Center	Park	Sporting Goods Shop	Fast Food Restaurant	Convenience Store	Photography Studio
2	Севастопольская	Gym / Fitness Center	Clothing Store	Fast Food Restaurant	Health Food Store	Pizza Place	Supermarket	Salon / Barbershop	Bus Stop	Flower Shop	Pet Store
3	Семеновская	Coffee Shop	Photography Studio	Pet Store	Dance Studio	Hobby Shop	Vietnamese Restaurant	Gym / Fitness Center	Baby Store	Kids Store	Optical Shop
4	Серпуховская	Coffee Shop	Bakery	Yoga Studio	Cosmetics Shop	Spa	Gym / Fitness Center	Theater	Auto Workshop	Caucasian Restaurant	Clothing Store

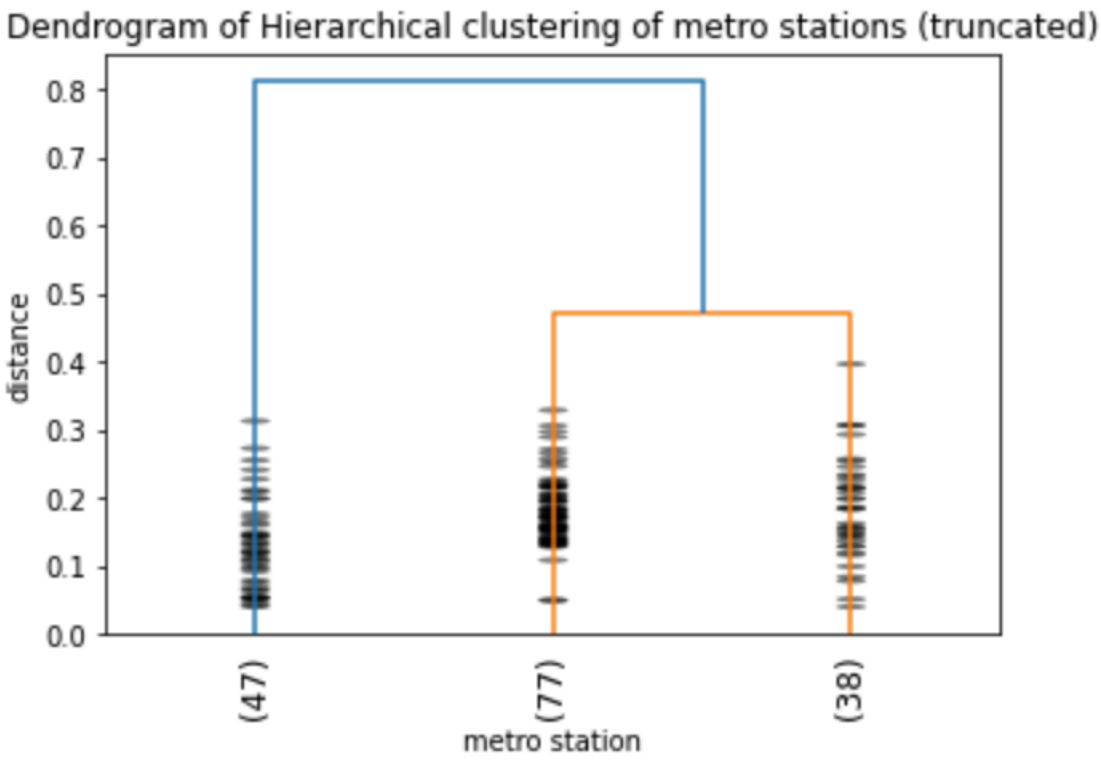
The K-means machine learning algorithm was used to cluster metro stations. The first thing to decide is how many clusters to use. One of the most popular techniques is the elbow method. The graph shows the results of clustering up to 30 clusters and the corresponding sum of squared distances from each point to its assigned center (distortions). The results do not allow us to determine the optimal number of clusters, because we did not get an obvious elbow.

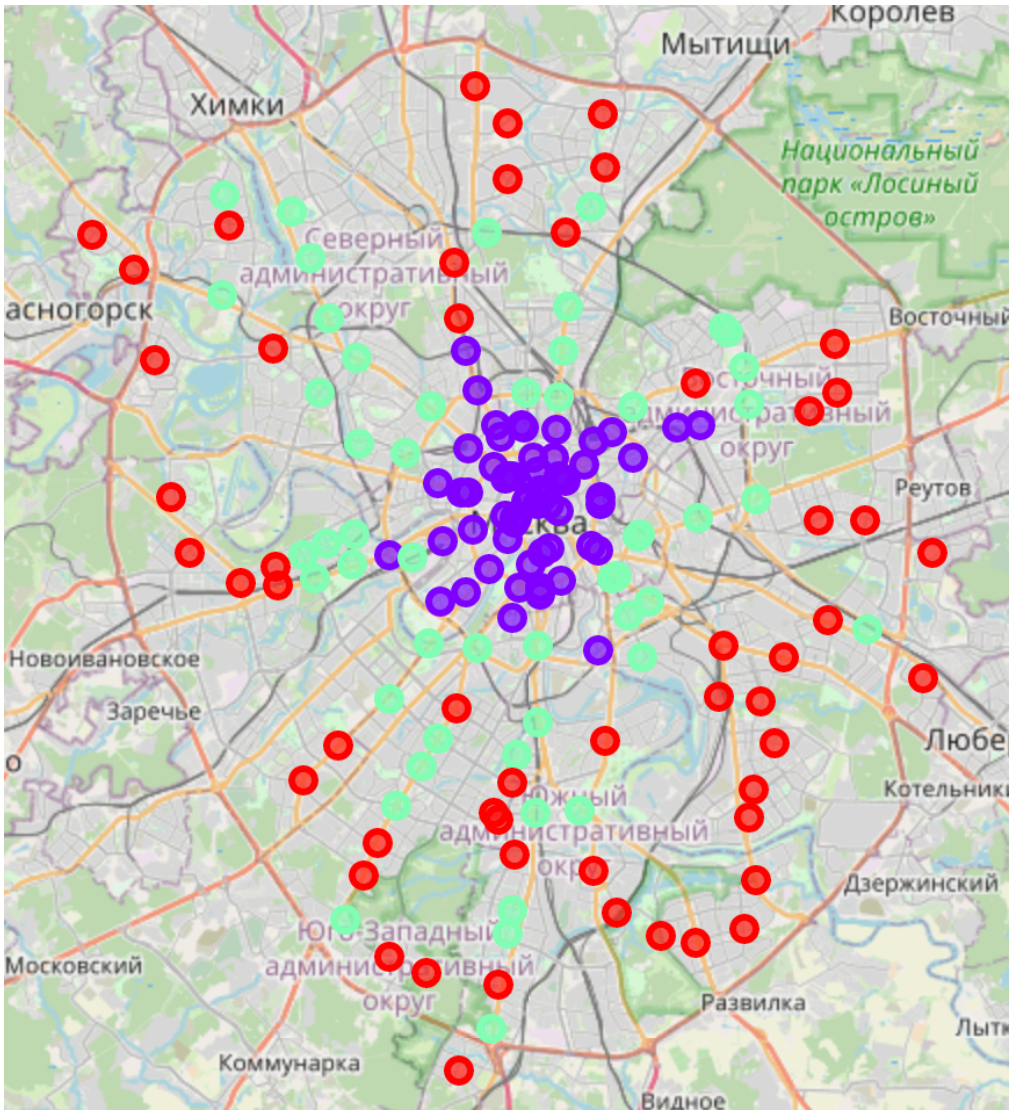


Due to the fact that the elbow method did not give satisfactory results, hierarchical clustering was applied, the results of which are shown in the following figure.



Still difficult to determine the optimal number of clusters, after several experiments it was decided to choose 3 clusters. The truncated dendrogram and the corresponding number of stations in each cluster are shown in the following figure.



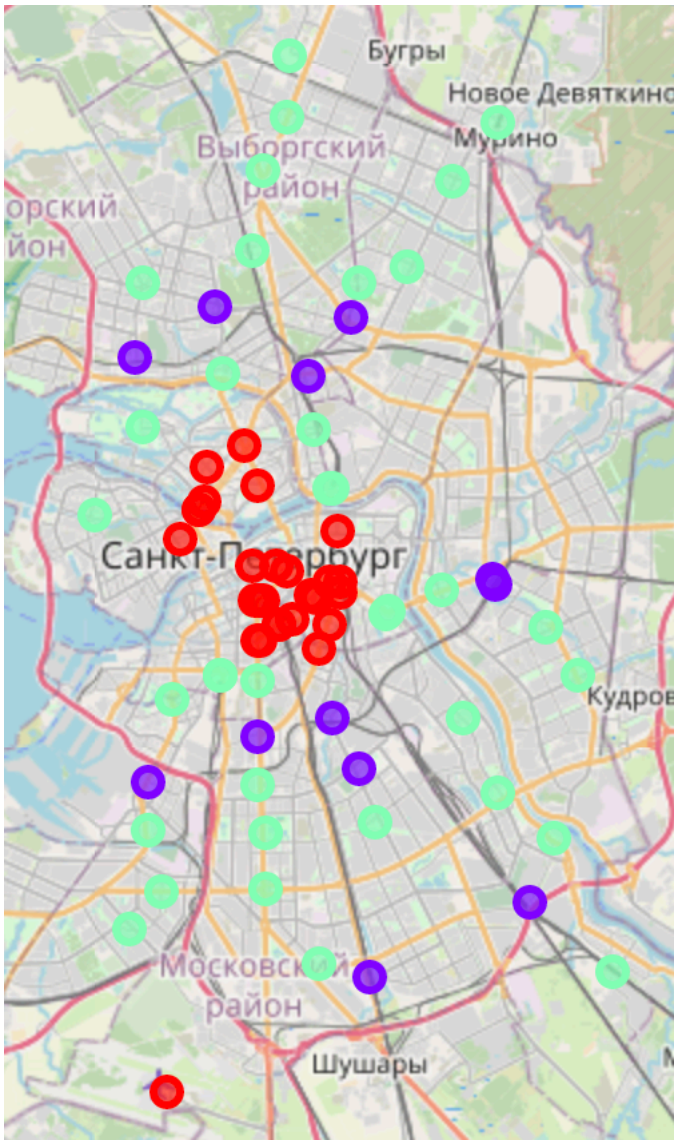


Let's show the distribution of metro stations by clusters on the map.

As you can see, we got a certain pattern: the clusters spread from the center of the city to the periphery. Let's look at the sample venues that represent each cluster.

	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	0	Supermarket	Park	Gym / Fitness Center	Health Food Store	Pizza Place	Cosmetics Shop	Pharmacy	Café	Coffee Shop	Sushi Restaurant
1	1	Coffee Shop	Theater	Gym / Fitness Center	Hotel	Dance Studio	Café	Yoga Studio	Caucasian Restaurant	Plaza	Bakery
2	2	Coffee Shop	Café	Auto Workshop	Gym / Fitness Center	Electronics Store	Clothing Store	Park	Supermarket	Cosmetics Shop	Sporting Goods Shop

3.2 Clustering St. Petersburg metro stations

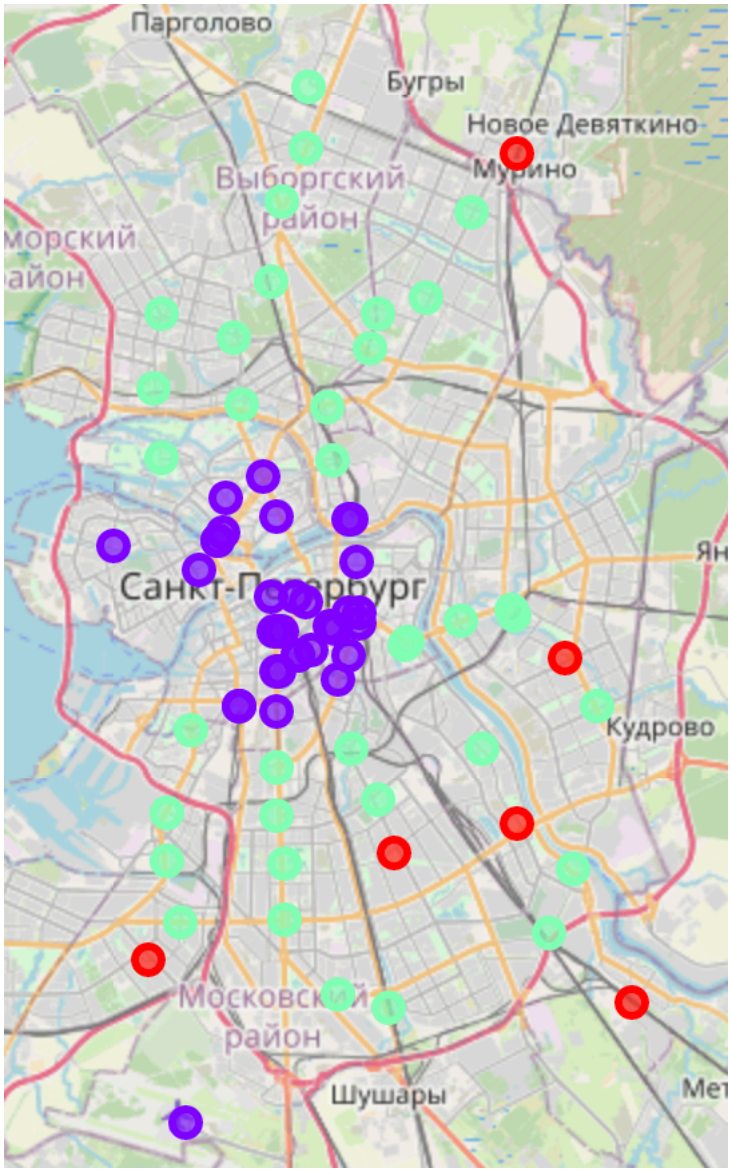


The above procedure was carried out to cluster St. Petersburg metro stations. As can be seen in the figure, the same pattern persists: the clusters are distributed from the center of the city to the periphery. There are 354 uniques venue categories.

4. Results.

First of all it is necessary to solve the problem of the different number of unique venues (features) in Moscow and St. Petersburg, respectively 412 and 354. For further analysis we will select only matching categories. There are 314 such categories.
Re-perform the clustering of Moscow and St. Petersburg metro stations into 3 clusters, using 314 categories.

I used the K-means model derived from Moscow metro stations data to cluster St. Petersburg metro stations. The results are shown on the map.



Prediction error was calculated using clustering data from St. Petersburg stations and predicted clustering using a model derived from Moscow metro station data. Clustering prediction accuracy is 0.87.

5. Conclusion

In this work was created a clustering model, based on Foursquare data about the venues located in the area of Moscow metro stations. This model was used to cluster St. Petersburg metro stations. The resulting model can be used to determine the category of a particular point in the city.
What problems were encountered and what further ways of developing the work? First, it is not an easy task to determine the number of clusters. As hyperparameters, you can use the radius of the area in which the venues are considered (in this work 1000 meters), specific points of urban space, the selection of venues as features. Perhaps accuracy can be improved by using the venues as a feature for other machine learning methods or ensemble models. It is also possible to collect and add other features, such as population density, traffic data, property values etc.
Thus, the data from Foursquare offer great potential for zoning urban space using machine learning models.

References.

1. https://en.wikipedia.org/wiki/City_region
2. Daniel Preo¸tiuc-Pietro, Justin Cranshaw, Tae Yano. Exploring venue-based city-to-city similarity measures. UrbComp '13: Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing. August 2013 Article No.: 16 Pages 1–4. https://www.sas.upenn.edu/~danielpr/files/cities13urbcomp.pdf