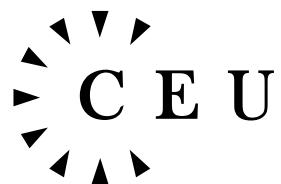


# **Networks and Cities**



**Luis Guillermo Natera Orozco**

Department of Network and Data Science  
Central European University

Supervisor: Federico Battiston  
External Supervisor: Michael Szell

A Dissertation Submitted in Partial Fulfillment of the Requirements  
for the Degree of Doctor of Philosophy in Network Science

2020

Luis Guillermo Natera Orozco: *Networks and Cities*, ©  
2020  
All rights reserved.

## RESEARCHER DECLARATION

---

I Luis Guillermo Natera Orozco certify that I am the author of the work Networks and Cities. I certify that this is solely my own original work, other than where I have clearly indicated, in this declaration and in the thesis, the contributions of others. The thesis contains no materials accepted for any other degrees in any other institutions. The copyright of this work rests with its author. Quotation from it is permitted, provided that full acknowledgement is made. This work may not be reproduced without my prior written consent.

### **Statement of inclusion of joint work**

Signature of PhD Candidate:

September, 2020

Signature of Dr. Federico Battiston, endorsing statement of joint work:

September, 2020

Signature of Dr. Michael Szell, endorsing statement of joint work:

September, 2020

Signature of Dr. Orsyola Vásárhelyi , endorsing statement of joint work:

September, 2020

## ABSTRACT

---



## ACKNOWLEDGEMENTS

---



# CONTENTS

---

<b>Contents</b>	<b>i</b>
<b>1 Extracting the multimodal fingerprint for transportation networks</b>	<b>1</b>
1.1 Prelude . . . . .	1
1.2 METHODS AND DATA . . . . .	2
1.3 FINDINGS . . . . .	4
<b>2 Life quality as walkability</b>	<b>7</b>
2.1 Prelude . . . . .	7
2.2 Data . . . . .	8
2.3 Quantifying Life Quality . . . . .	9
2.3.1 The services index: $Q^{services}$ . . . . .	11
2.3.2 Safety index: $Q^{safety}$ . . . . .	12
2.3.3 Environmental index $Q^{environment}$ . . . . .	13
2.4 Results . . . . .	14
2.4.1 Evaluation . . . . .	15
2.5 Discussion . . . . .	16



## LIST OF FIGURES

---

1.1 <b>(Map plot left)</b> Multiplex network representation of Manhattan with the four analyzed layers of transport infrastructure (pedestrian paths, bicycle paths, rail lines, and streets), with data from OpenStreetMap. <b>(Right)</b> Network information for each layer, number of nodes, links and average degree $\langle k \rangle$ . . . . .	3
1.2 <b>(a)</b> Schematic of multiplex layers in a city (left) and its transformation to the overlap census (right). In the overlap census, the vertical red line gives a visual separation of the left from the right half where nodes become active in the street layer. High spikes in the right half indicate car-centricity. <b>(b)</b> Clusters of cities based on similarity of their overlap census. We find six different clusters using a k-means algorithm (coloured areas), which explain more than 90% of the variance. <b>(c)</b> Overlap census for cities in each cluster. The first one corresponds to Amsterdam (the city with most active nodes in bicycle-only configurations). The Copenhagen-Manhattan-Barcelona-Portland city cluster has many active nodes in pedestrian-only and bicycle-only configurations, representing an active mobility city. The clusters of Los Angeles-Bogota and Mexico-Beihai-Jakarta are car-centric. . . . .	4
2.1 <b>(a)</b> Network representing Budapest pedestrian structure. The network was built following a primal approach, where the edges are sidewalks and pedestrian infrastructure, and nodes are intersections. <b>(b)</b> The graph-Voronoi tessellation of the Budapest network, generated using a subset of 15 parks as seeds. The color of the nodes represents the cell they belong to and the highlighted red dots are the seeds of each cell. The distance measure between two points is defined as the weighted shortest path on the graph, the weights being the average time required to cross a given edge. . . . .	10
2.2 Budapest neighborhoods, average life quality by categories and aggregated life quality index. . . . .	14

2.3 Districts of Budapest Life Quality Index (LQI) and its components correlated with real estate prices ( $m^2/EUR$ ) . . . . .	16
--	----

# CHAPTER 1

---

## EXTRACTING THE MULTIMODAL FINGERPRINT FOR TRANSPORTATION NETWORKS

---

Urban mobility increasingly relies on multimodality, combining the use of bicycle paths, streets, and rail networks. These different modes of transportation are well described by multiplex networks. Here we propose the overlap census method which extracts a multimodal profile from a city's multiplex transportation network. We apply this method to 15 cities, identify clusters of cities with similar profiles, and link this feature to the level of sustainable mobility of each cluster. Our work highlights the importance of evaluating all the transportation systems of a city together to adequately identify and compare its potential for sustainable, multimodal mobility.

### 1.1 Prelude

The infrastructure of different modes of transportation can be described as a mathematical object, the multiplex transport network [?, ?, ?, ?, ?, ?, ?, ?, ?]. A city's multiplex transport network contains the layer of streets and other coevolving network layers, such as the bicycle or the rail networks, which together constitute the multimodal transportation backbone of a city. Due to the car-centric development of most cities [?], streets form the most developed layers [?, ?] and define or strongly limit other layers: For example, sidewalks are by definition footpaths along the side of a street and make up a substantial part of a city's pedestrian space [?]. Similarly, most bicycle paths are part of a street or are built along the side. Yet, the different layers of a multimodal network typi-

cally serve as diverse channels to permeate a city. Here we consider the transport networks of 15 world cities and develop an urban fingerprinting technique based on multiplex network theory to characterize the various ways in which transport layers can be interconnected, identifying the potential for multimodal transport. Using clustering algorithms on the resulting urban fingerprints, we find distinct classes of cities, reflecting their transport priorities.

## 1.2 METHODS AND DATA

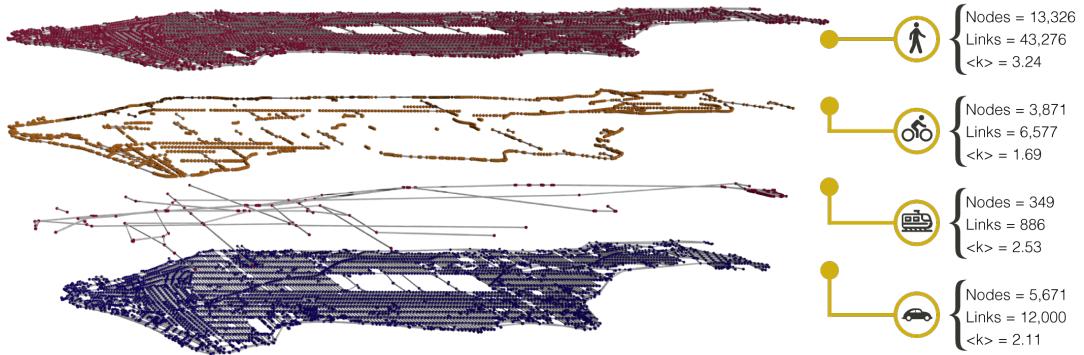
We acquired urban transportation networks from multiple cities around the world, defined by their administrative boundaries, using OSMnx [?]. These data sets are of high quality [?, ?] in terms of correspondence with municipal open data [?] and completeness [?]. The various analyzed urban areas and their properties are reported in Table 1.1. Figure 1.1 shows the different network layers for Manhattan, one of our analyzed cities.

	Pedestrian			Bicycle			Rail			Street			Population
	Nodes	Links	$\langle k \rangle$	Nodes	Links	$\langle k \rangle$	Nodes	Links	$\langle k \rangle$	Nodes	Links	$\langle k \rangle$	
Amsterdam	23,321	33,665	2.89	34,529	35,619	2.06	1,096	1,655	3.02	15,125	21,722	2.87	872,680
Barcelona	20,203	30,267	3.00	7,553	7,647	2.02	249	249	2.00	10,393	15,809	3.04	1,600,000
Beihai	2,026	2,978	2.94	0	0	0.00	59	62	2.10	2,192	3,209	2.93	1,539,300
Bogota	81,814	121,038	2.96	9,760	9,651	1.98	166	165	1.99	62,017	91,197	2.94	7,412,566
Budapest	73,172	106,167	2.90	10,494	10,318	1.97	1,588	1,964	2.47	37,012	52,361	2.83	1,752,286
Copenhagen	30,746	41,916	2.73	13,980	13,988	2.00	276	369	2.67	15,822	20,451	2.59	2,557,737
Detroit	47,828	78,391	3.28	3,663	3,626	1.98	20	21	2.10	28,462	45,979	3.23	672,662
Jakarta	140,042	191,268	2.73	248	231	1.86	58	54	1.86	138,388	188,637	2.73	10,075,310
LA	89,543	128,757	2.88	14,577	14,428	1.98	173	221	2.55	71,091	101,692	2.86	3,792,621
London	270,659	351,824	2.60	62,398	60,043	1.92	2,988	3,535	2.37	179,782	219,917	2.45	8,908,081
Manhattan	13,326	21,447	3.22	3,871	3,777	1.95	349	436	2.50	5,671	9,379	3.31	1,628,701
Mexico	108,033	158,425	2.93	5,218	5,278	2.02	370	364	1.97	95,375	140,684	2.95	8,918,653
Phoenix	111,363	157,075	2.82	35,631	35,979	2.02	105	138	2.63	73,688	102,139	2.77	1,445,632
Portland	50,878	72,958	2.87	24,252	24,325	2.01	230	340	2.96	35,025	49,062	2.80	583,776
Singapore	82,808	110,612	2.67	12,981	12,947	1.99	683	740	2.17	50,403	66,779	2.65	5,638,700

**Table 1.1.** Measures for the administrative area of analyzed cities. The number of nodes, links and average degree ( $\langle k \rangle$ ) for each layer in all cities of our dataset are highly diverse due to the varying developmental levels and focus of transport. The range of population in the analyzed cities goes from half million people to ten million people living in Jakarta, this allows to have a range of different sizes and cover different developmental stages.

Data and code to replicate the results are available in: (<https://doi.org/10.7910/DVN/GSOPCK>), and: (<https://github.com/naturaluis/Multimodal-Fingerprint>).

We characterize each city as a multiplex network [?, ?, ?] with  $M$  layers and  $N$  nodes that can be active in one or more layers in the system. Layers follow a primal approach [?] where nodes represent intersections (that may be present

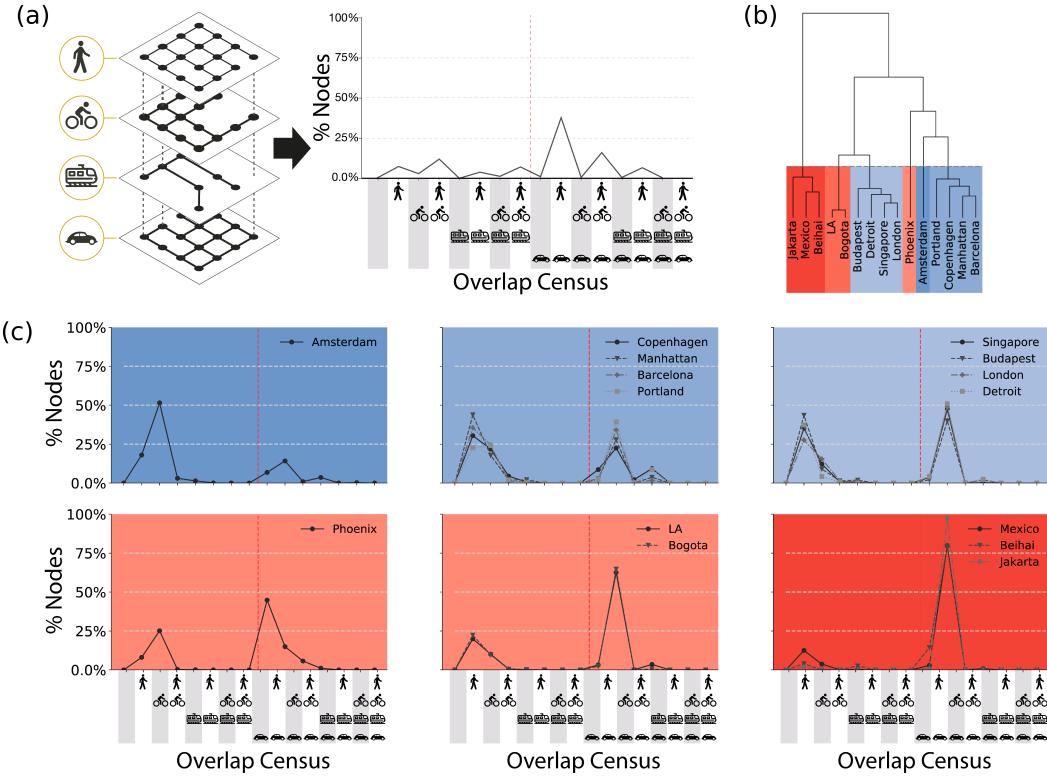


**Figure 1.1.** (Map plot left) Multiplex network representation of Manhattan with the four analyzed layers of transport infrastructure (pedestrian paths, bicycle paths, rail lines, and streets), with data from OpenStreetMap. (Right) Network information for each layer, number of nodes, links and average degree  $\langle k \rangle$ .

in one or more layers), and links represent streets (denoted by  $s$ ), bicycle paths and designated bicycle infrastructure ( $b$ ), subways, trams and rail infrastructure ( $r$ ), or pedestrian infrastructure ( $p$ ). Construction of these intersection nodes follows the topological simplification rules of OSMnx [?].

In a multimodal city, we expect to find many transport hubs that connect different layers, such as train stations with bicycle and street access, i.e. nodes that are active in different multiplex configurations. Here we propose a method to assess all such combinations of node activities in the system, helping us to learn how well connected different modes are. For each city, we build a profile based on the combinations of node activities, and refer to it as *overlap census* (Figure 1.2). The overlap census captures the percentage of nodes that are active in different multiplex configurations and provides an “urban fingerprint” of its multimodality [?]. To define the overlap census formally, given a multiplex transport network with  $M$  layers the overlap census is a vector of  $(2^M) - 1$  components, which accounts for the fractions of nodes that can be reached through at least one layer.

In Fig. 1.2(a) we show a schematic of how the overlap census is built: taking the multiplex network, and calculating the percentage of nodes that overlap in different configurations. The multiplex approach addresses the multimodality of a city: it not only counts how many nodes or links there are in each layer, but it shows how they are combined, revealing the possible multimodal mobility combinations in the city. Understanding the possibilities for interchange between mobility layers provides us with a better understanding of urban systems, showing us the complexity and interplay between layers.



**Figure 1.2.** (a) Schematic of multiplex layers in a city (left) and its transformation to the overlap census (right). In the overlap census, the vertical red line gives a visual separation of the left from the right half where nodes become active in the street layer. High spikes in the right half indicate car-centricity. (b) Clusters of cities based on similarity of their overlap census. We find six different clusters using a k-means algorithm (coloured areas), which explain more than 90% of the variance. (c) Overlap census for cities in each cluster. The first one corresponds to Amsterdam (the city with most active nodes in bicycle-only configurations). The Copenhagen-Manhattan-Barcelona-Portland city cluster has many active nodes in pedestrian-only and bicycle-only configurations, representing an active mobility city. The clusters of Los Angeles-Bogota and Mexico-Beihai-Jakarta are car-centric.

### 1.3 FINDINGS

Even in a multimodally “optimal” city there will be a high heterogeneity of node activities due to the different speeds and nature of transport modes, implying, for example, a much lower density of nodes necessary for a train network than for a bicycle network. Therefore, a good way to assess a city’s over-

lap census is by comparing it with the overlap census of other cities. We find similarities between cities via a k-means algorithm fed with fifteen vectors (one per city), where each vector contains the percentages of nodes active in each possible configuration. The algorithm separates the 15 analyzed cities into six different clusters [Fig. 1.2(b)].

On the left half of the overlap census, we show the configurations in which nodes are not active in the street layer, while the right half contains car-related configurations [Fig. 1.2(c)]. These clusters of cities are useful to explain similarities in infrastructure planning in different transport development paths [?, ?], with clusters of car-centric urbanization (like Mexico, Beihai, and Jakarta) opposed to clusters that show a more multimodal focus in their mobility infrastructure (like Copenhagen, Manhattan, Barcelona, and Portland). In the extreme cluster that contains only Amsterdam, close to 50% of nodes are active in the bicycle layer, whereas in the Mexico-Beihai-Jakarta cluster more than 50% of nodes are active in the street-pedestrian configuration. The concentration of nodes in just one configuration informs not only about the mobility character of the city, i.e. Amsterdam being a bicycle-friendly city, but unveils the importance of explicitly considering overlooked layers and their interconnections. For example, Singapore, Budapest, London, and Detroit have two main peaks indicating that most of their nodes are either active in the street-pedestrian or only in the pedestrian configuration. This is not the case in Los Angeles and Bogota, where the majority of nodes are active in the car-pedestrian combination, i.e. the pedestrians have to share most of the city with cars. Our multimodal fingerprint unravels how different transport modes are interlaced, helping identifying which layer (or set of layers) could be improved to promote multimodal, sustainable mobility.

To summarize, we propose the new “overlap census” method based on multiplex network theory allowing to rigorously identify and compare the multimodal potential of cities.



## CHAPTER 2

---

### LIFE QUALITY AS WALKABILITY

---

#### 2.1 Prelude

During the 20th century, most cities have evolved to accommodate a car-centric vision [?], allocating a privileged amount of urban space to motorized traffic [?, ?]. From a liveability perspective, this situation is suboptimal because the automobile infrastructure dominates and defines the walkable area, increasing car traffic, air pollution and deteriorating walkable conditions.

The concept of walkability is an important factor to consider in connection with liveability. Liveability refers to an environment from an individual perspective [?] which includes "a vibrant, attractive and secure environment for people to live, work and play and encompasses good governance, a competitive economy, high quality of living and environment sustainability" [?]. Thus in a liveable city, there must be an emphasis not only on sustainable transportation and built environment to reduce the harm on nature [?, ?] but also encouraging citizens to walk for supporting their physical and mental well-being [?]. However, improving walkability is more complex than we would think. Walking should be an available, safe and well-connected mode of transportation, but as Speck put it well, it should be interesting and comfortable as well, to have a feeling of the streets as 'outdoor living rooms' [?].

The pedestrian infrastructure that sustains walkability in a city can be described as a network [?]. This approach has been useful to identify street patterns [?, ?] and its evolution [?, ?], measure the morphology of cities [?], and how the streets connectivity impacts on pedestrian volume [?].

The various approaches to create a walkability index or so-called walk score consider mainly the following components: safety and security [?, ?]; convenience, attractiveness and public policy [?, ?], connectedness [?], but also reckon

with the land use mix and residential density of the certain area[?]. Another approximation rather accents the importance of its effect on air pollution, health problems, travel costs and even on the sense of community[?]. Thus measuring walkability not only captures the propensity to walk in a city but also includes the components a liveable city must have and support, under the umbrella of sustainability.

There are good examples of how sustainable city development initiatives tackle growing inequalities with data-driven approaches. Long Island used city data to analyze which amenities are needed to increase the quality of life in a newly built environment [?], other cities are investing in smart technologies to develop public transport, connecting spatially discriminated areas [?, ?].

Since the number of components which should be taken into consideration in creating a walkability index is high, the types of data are also mixed and thus difficult to integrate. While the information on connectedness, security, residential density, etc. is quantitative and in general easily available, gaining opinion about attractiveness, convenience, or even about the feeling of security is more complicated. Here we propose to use a data-driven approach as a proxy to quantify life quality, making it reproducible and easily expanded to include different data sources. We apply our methods to Budapest, but as having an emphasis on the online and easily available quantitative data, the methods can be generalized and applied to any city.

## 2.2 Data

We work with three different data sources: networks, points of interest and city attributes. The pedestrian network and points of interest were acquired using OSMnx [?], a python library to download and construct networks from OpenStreetMap (OSM). The data contained in OSM is of high quality [?, ?] in terms of correspondence with municipal open data [?] and completeness: More than 80% of the world is covered by OSM [?].

The majority of points of interest were downloaded from OpenStreetMap, from different classification keys (amenity, tourism, shop, office, leisure) using OSMnx [?]. We filtered the points of interest using the districts' demarcation [?], to get only the data within Budapest boundaries, having, as a result, more than 39,000 data points. We complement the data sets with secondary data sources as specialized directories of doctors and childcare facilities (see appendix).

We categorize the points of interest in six main categories: I) Family friendliness (Access to education and daycare, and family support services), II) Access to health care and sport facilities, III) Art and culture (e.g. museums, exhibitions), IV) Nightlife (e.g. bars, restaurants), V) Environment (air quality and ac-

cess to green areas), and VI) Public Safety. The points of interest and secondary data sources are available at [https://github.com/naturaluis/Budapest\\_LQI](https://github.com/naturaluis/Budapest_LQI)

The district-level data (population and crimes) were obtained from the Hungarian Police's public database, calculated based on the number of crimes committed in public places 100 thousand per capita in 2018 [?]. Population data is coming from the 2016 micro-census conducted by the Hungarian Statistical Bureau [?]. We took into account the air pollution, this data set coming from National Air Pollution Measurement Network [?], containing the geolocation of the air quality stations and different measures (annual median concentration of carbon monoxide, nitrogen dioxide, and PM10 dust).

Accuracy of the Life Quality Index (LQI) model highly depends on how comprehensive the distribution of listed services. We use OSM as our key data source, but to achieve a more comprehensive and country-specific database we collect publicly available data from various Hungarian websites for each category (See Appendix A for databases and sources)

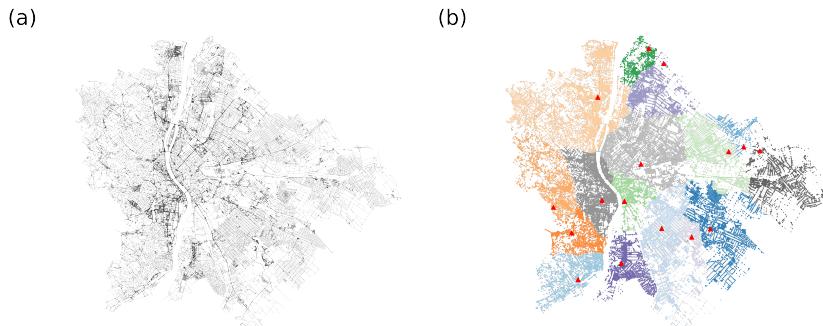
The network contains all the sidewalks and pedestrian designated infrastructure, it is conceptualized as undirected, nonplanar and primal network [?]. The pedestrian network is described as a weighted graph, with its adjacency matrix  $W = w_{ij}$  where the weight  $w_{ij}$  contains the length between  $i$  and  $j$  if connected, and 0 otherwise.

We assigned properties to the nodes of the network, matching the nodes with their corresponding districts, then assign nodes as attributes based on the district level data (population and crimes, see section 2.3.2). For the pollution data, we calculated the corresponding Voronoi cells, for the air quality stations, and matched the nodes with them, we divided the pollution by the number of nodes in each corresponding cell and assigned the value to the nodes (See section 2.3.3). For the edges, we encoded their length  $\ell_{ij}$  along with the traversal time  $Tt_{ij}$  between nodes  $i$  and  $j$  calculated as  $Tt_{ij} = \frac{\ell_{ij}}{ps}$  where  $ps$  is the pedestrian speed as a constant rate of 5km/h.

## 2.3 Quantifying Life Quality

The life quality of a person is largely subjective and hard to quantify. However, it is both intuitive and has been scientifically shown that the environment and personal well-being strongly correlate [?]. Thus using environmental factors as proxies, life quality and livability becomes quantifiable [?].

The main environmental factors we consider in our model are: the availability of services and amenities, the quality of the infrastructure, environmental factors and safety. The goal of our model is to quantitatively characterize the



**Figure 2.1.** (a) Network representing Budapest pedestrian structure. The network was built following a primal approach, where the edges are sidewalks and pedestrian infrastructure, and nodes are intersections. (b) The graph-Voronoi tessellation of the Budapest network, generated using a subset of 15 parks as seeds. The color of the nodes represents the cell they belong to and the highlighted red dots are the seeds of each cell. The distance measure between two points is defined as the weighted shortest path on the graph, the weights being the average time required to cross a given edge.

immediate environment of residents in the space of factors that affect life quality.

The fundamental framework of our model and our calculations is the network representation of Budapest's pedestrian infrastructure. The nodes of the network represent intersections, while links are sidewalks and pedestrian infrastructure. The output of our model is an index, that characterizes every node of the Budapest network, giving a high-resolution quality-landscape of the city. The index is ultimately a number aggregated from multiple sub-categories, and its main value is highlighting inequalities and relative deficiencies within the city.

The final value of the index is a weighted sum, characterizing every node (intersection) in the network:

$$Q_i = w^{services} \tilde{Q}_i^{services} + w^{safety} \tilde{Q}_i^{safety} + w^{environment} \tilde{Q}_i^{environment} \quad (2.1)$$

In the equation  $i$  represents an individual node in the network. The “tilde” above the  $Q$  terms means that the values of the different category indices are normalized within the category. The weights  $w$  assigned to every term are arbitrary and are highly context-dependent. We include the weights used for producing the results of this paper in Appendix B. All terms of the equation are discussed in the following sections.

### 2.3.1 The services index: $Q_i^{services}$

The number quantifying each node in terms of how well it is connected with amenities and services is a weighted sum of sub-categories as well.

$$Q_i^{services} = \sum_c w^c Q_i^c \quad (2.2)$$

where,  $c$  denotes categories (family, culture, health, sport, and nightlife), and  $w^c$  the importance (weight, see Appendix B) of category  $c$ . Some categories, like family, have further subcategories. Even though we have also had data and made a separate analysis on tourism, its effects on life quality of the residents are ambiguous, so we decided to omit it from the index.

What sets categories apart is that they incorporate different sets of amenities, with a few overlaps. The details of the categorization of amenities are included in the Appendix.

For every service/amenity class we have a given set of points of interest (POI) along with where the amenities of that class are available, with exact geolocation. We assign every POI of a given amenity class (e.g supermarket, pharmacy, school, etc.) to the nearest node on the infrastructure network. Each set of POIs organically generates a spatial partitioning of the city with one partition per POI. The partition of a POI is the set of all the nodes from which that particular POI can be reached faster than any other POI of the same class.

Mathematically these partitions are called graph-Voronoi cells [?, ?], where every node of a cell is assigned to its closest seed (POI). Distance, in this case, is not euclidean or geometric distance, but the distance on the network, where we use the weighted shortest path between two nodes as the distance measure. The weight of links is a temporal parameter encoding the average time required to cross the represented street from one end to the other, thus the weight is a simple product of average speed and length of the street. This is in principle very similar to the way navigation systems find routes between points. For an example of a graph-Voronoi partitioning see Figure 1 (b).

To assess how well connected a node is to amenities we consider the following factors:

- How important is an amenity - weight ( $w_a$ )
- How long does it take to reach the amenity - time to reach ( $t_{ia}$ )
- Relatively how many nodes (or people) does the amenity share with - exclusivity ( $P_a$ )

From the three factors, the latter two are calculated using the city infrastructure network. The index for an amenity class, from the perspective of node  $i$ , is proportional with its importance (weight) and it is inversely proportional with the time to reach the closest POI from  $i$  and with the degree of exclusivity.

$$q_i^a = \frac{w_a}{(P_a + 1)(t_{ia} + 1)} \quad (2.3)$$

There can be certain singular cases when a Voronoi cell is empty ( $P_a = 0$ , i.e. no residents in the area) or the node  $i$  in question is right at the POI ( $t_{ia} = 0$ ). To avoid anomalies in the index we added 1 to both parameters.

The index of one category is proportional to the sum of its amenity-indices (calculated in (2.3)). To treat this number on the right scale (in practice we can get very large and very small numbers) we take the natural logarithm of the sum across amenities.

$$Q_i^c = \log\left(\sum_a q_i^a\right) : \quad (2.4)$$

As we have mentioned earlier the final services index is the weighted sum of the indices of the sub-categories.

$$Q_i^{services} = \sum_c w^c Q_i^c$$

Finally we normalize the values of  $Q^{services}$  so its values are comparable to the other values of the final  $Q$  equation (2.1):

$$\tilde{Q}_i^{services} = \frac{Q_i^{services} + |\min(Q^{services})|}{\max(Q^{services}) + |\min(Q^{services})|} \quad (2.5)$$

### 2.3.2 Safety index: $Q^{safety}$

The safety index is calculated across districts based on the number of crimes committed per one hundred thousand residents. Since the highest resolution data available to us was on the district level, every node  $i$  in the same district will have the same safety index value. The crime index:

$$Q_i^{crime} = \frac{N_{crime}^{district}}{n_i^{district}}$$

Where  $N_{crime}^{district}$  is the number of crimes committed in a district in a year, and  $n_i^{district}$  is the number of nodes in the district. The safety index is one minus the normalized crime index.

$$\tilde{Q}_i^{safety} = 1 - \frac{Q_i^{crime}}{\max(Q^{crime})} \quad (2.6)$$

### 2.3.3 Environmental index $Q_{environment}$

The environmental index is made up of two components: air pollution ratio and ratio of natural areas.

#### Air pollution ratio

We use the data provided by Budapest's air pollution measuring stations for the year 2018. For this study, we used the yearly median value of three polluters: carbon monoxide, nitrogen dioxide, and PM10 dust-pollution. As an approximation, we project the geometric Voronoi cells of the measuring stations onto the city map and each node will receive the pollution metrics of the geometrically closest station. We divide these values with the yearly upper health limit for the given polluter to assess to what degree do these values approximate the health limit. Thus the air pollution index of one node is formalized as follows:

$$C_i = \frac{c_i^{CO}}{c_{limit}^{CO}} + \frac{c_i^{NO_2}}{c_{limit}^{NO_2}} + \frac{c_i^{Pm10}}{c_{limit}^{Pm10}}$$

Where  $c_{limit}^{CO} = 3000g/m^3$ ,  $c_{limit}^{NO_2} = 40g/m^3$ ,  $c_{limit}^{Pm10} = 40g/m^3$  are the yearly upper limits based on [?].

#### Ratio of natural areas

For this index, we have data on the neighborhood level, which is a more granular level of administrative partitioning the city than the districts are. In this case, we project the same index onto every node in the same neighborhood. We consider as natural areas forests, parks and water surfaces (ponds, rivers, etc).

The index:

$$T_i = \frac{R_{water}^{nh(i)} + R_{forest}^{nh(i)} + R_{park}^{nh(i)}}{\max(T)}$$

Where  $R_x^{nh(i)}$  is the relative surface area of natural area  $x$  within the neighborhood that  $i$  belongs to ( $nh(i)$ ). In other words, the surface area of a natural area is divided by the number of nodes in the neighborhood and the surface area of the neighborhood. Thus  $R_x^{nh(i)} = \frac{T(x)}{T(nh(i))n_{nh}}$ , where  $T(x)$  is the surface area of  $x$  natural area,  $T(nh)$  is the surface area of  $nh$  neighborhood and  $n_{nh}$  is the number of nodes in neighborhood  $nh$ . The final environmental index:

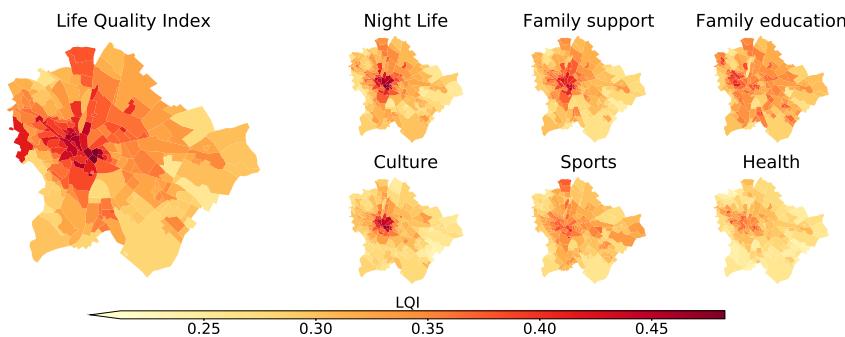
$$Q_i^{environment} = \frac{1 + T_i}{1 + C_i}$$

, That after a normalization is:

$$\tilde{Q}_i^{environment} = \frac{Q_i^{environment}}{\max(Q^{environment})}$$

## 2.4 Results

We quantify life quality in terms of each category (family support, education healthcare, sport, culture, nightlife, environment), and an overall measurement which contains all 6 categories and crime rate normalized by the population for the city of Budapest. Our method allows us to measure life quality for each intersection of the city, which helps to capture within neighborhood inequalities too. Analysis on the category level is beneficial for targeted policy interventions for better service allocation.



**Figure 2.2.** Budapest neighborhoods, average life quality by categories and aggregated life quality index.

Figure 2.2 shows our overall life quality index (LQI) and by categories. Heatmaps reveal important features of Budapest. Similarly to most European cities life quality is much better in the inner districts [?, ?], especially in the case of Night Life and Culture.

Budapest is divided by the Danube river into two main parts: Buda and Pest. The river does not only serve as a geographical border but due to historical reasons, it also divides the citizens by social status. Hilly Buda, on the West side of the river, used to be the capital of the country, with the residence of the former Hungarian king. On the other side, the mainly flat Pest used to be the agricultural supporter of the aristocrats in Buda [?]. Even though the city has changed dramatically since the Monarchy, the division of Buda and Pest persists, and our life quality index captures it well. However certain services

are legally guaranteed to be evenly distributed in the city, such as education and healthcare, for precise modeling one should take into account private care too, which highlights inequalities. So, the traditional division of Buda and Pest is even visible in categories where there should not be that much of a difference (Education, Family Support, Healthcare).

Results also highlight that category LQI-s are highly correlated, less liveable neighborhoods are constant regardless of the amenity category, and well-performing neighborhoods do not change either. It is caused by two main factors: the lack of amenities and the relatively high walking distances in the suburbs.

The compact city concept focuses on building more sustainable and livable cities while designing practical neighborhoods where citizens can maintain everyday life without a car [?]. Since, the walkability of a neighborhood highly correlates with its liveability [?] and the suburbs in Budapest do not show any compact city design features, both long distances and the lack of amenities effects suburban habitats lives negatively.

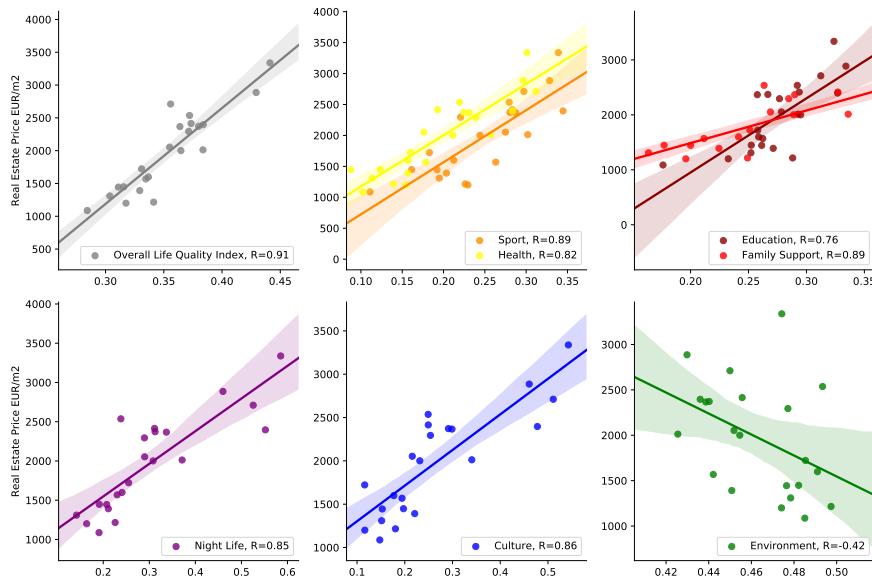
### 2.4.1 Evaluation

Multiple methods have been developed to evaluate the accuracy of quality of life metrics: Scholars used expert validation with geographic visualization [?, ?], correlations with socioeconomic characteristics [?] and surveying citizens' perceptions of the conditions of life [?].

Our evaluation is based on the micro-economical *hedonic approach* of estimating the values of public goods. In a capitalist market, real estate prices reflect the recognition of a neighborhood's characteristics: Prices are formed based on demand, more desirable places are more expensive, due to the underlying assumption of providing a higher quality of life. [?] Estimating neighborhoods life-quality with real estate prices has a long tradition in urban literature [?, ?, ?], therefore we adopt this method to evaluate our model.

We collected the average  $m^2/EUR$  price for all 23 districts of Budapest in January 2019 [?] and correlated each LQI category averaged by district with it. Figure 2.3 shows that our overall LQI correlates the most ( $R=0.91$ ) with the real-estate prices. Most of its components have a positive correlation with real-estate prices, except the environment which is calculated based on air pollution and green surface proximity. The life quality (LQI) in Budapest is much higher in densely populated downtown districts, which are lack of green surface and suffers from high air pollution due to heavy traffic.

**Summary** Locals of Budapest, like in most European cities, traditionally values downtown areas. The relative closeness to CBD, good access to public



**Figure 2.3.** Districts of Budapest Life Quality Index (LQI) and its components correlated with real estate prices ( $m^2 / EUR$ )

transport, and vital city life kept it as a desirable area for living [?]. However, in recent years, the city is facing new challenges: due to gentrification [?] and over-tourism (eg.:Airbnb) real estate prices are sky-rocketing in downtown areas. In contrast with the early 2000-s when (upper) middle-class moved to the suburbs, nowadays, lower-income families and young professionals are leaving the downtown behind in hope for more affordable living.

As our findings show, Budapest is quite centralized and the quality of life highly correlates with real estate prices, which possibly lead to even more inequalities in the future. This spatial discrimination with longer traveling time, less fulfilling environment, and potential segregation reduces the chances of upward mobility and the quality of life of individuals [?].

## 2.5 Discussion

We have proposed a methodology to quantify life quality as a function of walkability on urban networks. We have used open data to capture inequalities between neighborhoods and districts in the city. We have shown that the real estate market reflects the life quality that our methods found.

A data-driven approach for quantifying life quality at such a granular level like our proposed method can help decision-makers to tackle social and envi-

ronmental challenges better. Designing compact, liveable neighborhoods, considering also the upcoming environmental crisis is the number one priority of many cities worldwide.

The use of open data sources and algorithmic approaches adds up towards a systematic framework for understanding urban liveability. Our current approach is not the last word in this development since it does not yet account for multiple other variables, such as the quality of services and infrastructure, and other qualitative variables. To capture the more specific indicator of liveability in different cities it would be necessary to work with more granular and city-dependant data.

We anticipate a future stream of research focused on the use of worldwide open data sets to quantify urban liveability, including longitudinal studies in multiple cities, along with algorithmic modeling, simulations, and machine learning approaches, to first quantify the liveability, propose changes and test them with the ground truth data.



