Urban Amenities Navigator

Joep Leeuwenburgh

Layout: typeset by the author using $\LaTeX$.
Cover illustration: Amenity plot of the neighbourhood Rozenknopje in Eindhoven

Urban Amenities Navigator

Characterising space through entropy-based measures

Joep Leeuwenburgh
13588044

Bachelor thesis
Credits: 18 EC

Bachelor *Kunstmatige Intelligentie*

University of Amsterdam
Faculty of Science
Science Park 900
1098 XH Amsterdam

Supervisor
dr. R. M. (Rafiazka) Hilman

Informatics Institute
Faculty of Science
University of Amsterdam
Science Park 900
1098 XH Amsterdam

Semester 2, 2023-2024

Abstract

This thesis proposes the implementation of entropy-based measures to analyse urban spaces with a focus on urban liveliness, using Open Street Map data and geodemographic data of the Netherlands from the Statistics Netherlands (*Centraal Bureau voor Statistiek*). The research aims to evaluate the behaviour and effectiveness of different entropy measures in characterising space based on the available urban amenities as well as turning them into measures of urban liveliness. Shannon, Leibovici and Altieri entropy are evaluated and used to capture the complexity and heterogeneity of urban areas, showing that higher entropy values are associated with a more diverse and well-distributed collection of amenities. Local Moran's I, a spatial autocorrelation technique, is used to identify hot spots and cold spots across the Netherlands. A significant outcome of the thesis is the development of an interactive dashboard that distils the findings of the research in a user-friendly format. Limitations include the inconsistent completeness of the Open Street Map Data and the computational constraints that prevent the implementation of customisable filters. To deal with this issue, a deep learning based remote sensing classification technique could be taken into consideration. Apart from that, suggested future research topics include enhancing the dataset used with a temporal component and the development of a predictive model to explore the dynamics between amenity distributions and socio-demographic features.

Contents

Chapter 1

Introduction

In the search for a new place to settle, we are confronted with many things to consider. Aside from taking into consideration the housing prices and square footage, we have to ask ourselves another question: "Where do I want to live?" While your work location might influence the answer to this question, you might have particular preferences regarding the living experience in a specific area. One might value living in a bustling neighbourhood, emphasising the direct availability of restaurants and cafes. At the same time, others might prefer a quieter area with schools for primary and secondary education in the direct vicinity. Answering this question can be challenging, especially for those unfamiliar with an area.

A website like Funda provides insight into the availability of nearby amenities (schools, daycare, public transport and supermarkets) and socio-demographic characteristics. However, it does not augment that information into a concise and comparable index nor present an option to compare different neighbourhoods. The presence of urban amenities can have a significant impact on our residential preferences as well as our experienced satisfaction while living there. Therefore, quantifying these aspects and conveying such information is indeed necessary.

Open Street Map (OSM) is an open-source project that creates a free editable world map. Combining OSM data with the geographies of different scales obtained from the *Centraal Bureau voor Statistiek* (CBS), an opportunity is presented to delve into the spatial configuration of various areas. Not only that, coupling OSM with CBS data also allows us to contextualise the socio-demographic configuration of living space.

This thesis explores the performance of different entropy measures and whether they can be used to construct a measure of urban liveliness. We define it as the ratio of spatial entropy to a logarithmic weighting of urbanity, capturing the balance between the complexity of human activities and the density of urban features within a given space. We also aim to investigate to what extent this entropy-based measure can be used together with spatial autocorrelation to extract spatial similarity between particular areas. Different entropy measures will be evaluated and then transformed into a measure of urban liveliness. Then, we will calculate spatial autocorrelation and use it to calculate similarities based on the distance between the values. It results in a set of neighbourhoods of comparable liveliness levels that could be used as alternative choices, such as in the housing search process. Lastly, a dashboard will be constructed to explore the different results for each measure and gain insight into specific areas and their surroundings. The dashboard will provide a seamless user experience and offer clear visual representations of the results to a wider audience.

Chapter 2

Literature review

This chapter will discuss the literature relevant to the thesis. First, we focus on the role of urban amenities in residential preferences/choices. Then, we survey the concept of spatial entropy to measure the variability of urban amenities in living spaces. Lastly, we discuss spatial autocorrelation, which serves to capture similarity across living spaces.

2.1 Urban amenities and urban liveliness

Urban amenities play a significant role in the choice of residential area and the experienced satisfaction therein (Kelly, 2006). Randall (2008) finds that most respondents to his survey are willing to live in more densely populated and built-up areas in exchange for more and better urban amenities. Saville-Smith (2010) similarly finds that location is an essential factor when deciding where to reside. Among others, significant factors are proximity to parks and recreational and educational amenities. Amenities affect the decision of where to live and neighbourhood satisfaction (McCrea, Stimson, & Marans, 2011).

The concept of urban vibrancy, referred to in this thesis as urban liveliness, captures the dynamic interplay of human presence, activities, and interactions within an urban space. This idea is rooted in the work of Jacobs (1961) and emphasises that lively, diverse, and attractive urban areas promote socio-economic activities, thus creating vibrancy. The built environment, namely points of interest such as urban amenities, plays a crucial role in supporting these activities and interactions (Yue et al., 2021). Moreover, a dense population is often associated with increased quality of life (Jacobs, 1961; Wu & Niu, 2019). Urban liveliness reflects the physical and social dynamics of a city and contributes to the overall quality of life, as highlighted by Lynch (1984).

2.2 Spatial entropy

Entropy can be used to distil information about the overall presence and location of urban amenities in a particular area and its surroundings. Shannon entropy can summarise a distribution into a continuous value, denoting the information needed to describe a distribution (Saraiva, 2023).

Leibovici (2009) builds upon Shannon's entropy by emphasising the use of co-occurrences of categories at multiple orders. It proposes a framework for entropy measurements. This approach makes it possible to account for multivariate data, which opens up possibilities for a wide range of spatial interaction models. It provides an introduction to moving from Shannon's entropy towards a system that accounts for spatial dependencies, broadly defined as spatial entropy.

In a later work, Altieri et al. (2019b) explore different methods for calculating spatial entropy and apply those methods to measure urban sprawl. Urban sprawl is defined as the uncontrolled development of cities into surrounding areas, adversely affecting the city's quality of life and the environment. To measure spatial entropy, they discuss three different expansions of Shannon entropy that do not account for spatial data. The methods covered are Batty's spatial entropy (Batty, 1974), Karlström and Ceccato's spatial entropy (Karlström & Ceccato, 2000) and Spatial mutual information and residual entropy (Altieri et al., 2019b). The authors test these methods on synthetic data before computing results based on the data from three different European cities. The paper provides insight into ways to measure spatial entropy and displays the strengths and weaknesses of the different methods.

Altieri et al. (2019a) further explore this topic by presenting two different entropy measures, allowing them to go further than Shannon's entropy: Spatial mutual information and spatial residual entropy. The first can be used to identify the role of space in a distribution. The second is a measure of heterogeneity depending on other sources than space. It can measure heterogeneity in data after space is accounted for. The measures they propose can go further than only looking at two realisations simultaneously, meaning they can find relations between more than two locations simultaneously. Furthermore, the authors proposed to discard the order within co-occurrences, improving the ability to detect spatial patterns. The article is relevant because the project aims to expose similarities within urban environments by measuring spatial entropy, and the authors propose a method for unveiling spatial patterns.

In a more recent article, Altieri et al. (2023) present a package for computing different measures of spatial entropy, including Batty, BattyLISA, Leibovici and Decomposable. The authors state that difficulties can arise in practical studies until the point of writing. This is due to the scarcity of computational tools for fast and easy implementation of the mentioned entropy indices. A new R toolbox called SpatEntropy is introduced. Using SpatEntropy, the authors apply the abovementioned indices to a dataset of gorilla nesting sites. The article demonstrates how SpatEntropy can quickly gather results from entropy-based measures on spatial data.

2.3 Spatial autocorrelation

Spatial autocorrelation is a concept that analyses how similar specific points in space are based on their location and neighbours' values. In this project, the concept is used to see if particular neighbourhoods or cities are similar to others. One way to measure spatial autocorrelation is by calculating Moran's I, a value between -1 and 1 that tells us if points in space are positively correlated or negatively - indicating the clustering or dispersion of variables. Anselin (1995) introduces a class of local indicators of spatial association (LISA). LISA allowed for the decomposition of global indicators, such as Moran's I, into local indicators. Anselin (1995) evaluates this local Moran's I by doing Monte Carlo simulations applied to a dataset of African countries. This work is of significant value to this project because it enables the evaluation of similarities in local areas.

This is later built upon by introducing local indicators of multivariate associations (Anselin, 2019). In Anselin's research, the local Geary's C statistic is conceptualised as a weighted distance in multivariate attribute space between observations and their neighbours. This method can provide an effective way of gaining insight into the similarities between different spatial distributions of amenities.

Chapter 3

Dataset

This chapter will discuss the different datasets used in the project. To calculate the spatial entropy based on the amenities present in particular areas, first, the geometry of that area must be obtained. Then, the obtained geometry can be used to gather all amenities within that geometry. This is done using data from the CBS combined with data from OSM.

3.1 CBS

In this work, the "Wijk- en buurtkaart 2023" is used from CBS (2023) is used. This publicly available resource contains three datasets of the Netherlands in shapefiles for three scales: *Gemeenten* (municipalities), *wijken* (districts), and *buurten* (neighbourhoods). All three datasets contain the geometries of their respective elements and are supplemented with socio-demographic and geospatial demographic information. The CBS data accounts for a total of 342 *gemeenten*, $3,324$ *wijken* and $13,943$ *buurten*. In the future, the terms *gemeenten*, *wijken*, and *buurten* will be referred to using their English equivalents.

3.2 OSM

OSM is an open-source project that aims to create a free and editable map of the whole world (OpenStreetMap contributors, 2017). Volunteers do not only collect data by scraping online resources but also by surveying outside. OSM data is structured in nodes, ways and relations. Nodes represent specific points using latitude and longitude coordinates; ways consist of lists of nodes that define lines representing lines and polygons; relations are multi-purposed structures representing relations between different elements. Nodes, ways, and relations can have multiple tags attached that are represented as key-value pairs. These tags can then be combined to search for specific data. For example, by looking up nodes with the tags "amenity=restaurant" and "cuisine=Italian", it is possible to find all Italian restaurants in an area of interest. Data from OSM is obtained by using the Overpass-Turbo API. The data was retrieved in June 2024.

Chapter 4

Methods

Firstly, different entropy measures will be explored, after which those measures will be transformed to assess the liveliness of a particular area. Next, a method for measuring global and local autocorrelation is presented to characterise the similarity of spatial profiles within a particular area.

4.1 Entropy measurements

In information theory, entropy measures the degree of uncertainty of random events (Saraiva, 2023). This, in turn, is related to the amount of information that is needed to describe a distribution of categorical variables. The more information we need to describe a distribution, the higher the disorder of that distribution and the higher the entropy. When we need less information to describe a distribution, we will have fewer choices and, therefore, a smaller entropy value. In several studies, entropy is used to contextualise urban spatial configuration, such as in investigating the expansion of metropolitan areas in Italy (Altieri & Cocchi, 2022) and tracing the urban sprawl (Cabral, Augusto, Tewolde, & Araya, 2013). In the context of this project, when a particular area has a higher value of entropy, it is implied that it contains a more diverse landscape in terms of amenities and more amenities present. We will examine three entropy measurements: Shannon, Leibovici, and Altieri.

4.1.1 Shannon entropy

Shannon entropy (Shannon, 1948) has the following formula:

$$H(X) = -\sum_{i=1}^{n} p_i log_2(p_i) \tag{4.1}$$

In this formula, p_i denotes the probability of observing variable $X = x_i$. Using two as the base of the logarithm ensures that the outcome can be interpreted based on the number of bits of information we need to describe the distribution of X. Because this formula only considers the probability of observing x_i, this entropy measure does not account for the spatiality of the data. Therefore, Shannon entropy only provides information about how many different amenities we can expect in a specific area and how uniformly they are distributed.

4.1.2 Spatial entropy

New entropy measures must be constructed to account for the spatiality of the data points. This thesis will focus on the approaches introduced by Leibovici (2014) and Altieri (2019a).

Leibovici entropy

Leibovici (2014) introduces a new variable Z, defined as the co-occurrence of points. Entropy is then calculated as in Shannon's entropy but based on the probability that a co-occurrence is seen within a specified distance d. This results in the following formula:

$$H(Z|d) = \sum_{r=1}^{I^m} p(z_r|d) log(\frac{1}{p(z_r|d)}) \tag{4.2}$$

The sum considers every ordered combination of point types, resulting in I^m pairs to consider, where I denotes the number of types present and m denotes the combination size.

Altieri entropy

Altieri (2019a) expands on this by introducing another variable, W, which represents a set of sample windows across the data, and by splitting the entropy into spatial mutual information $MI(Z, W)$ and spatial residual entropy $H(Z)_W$.

$$H(Z) = \sum_{r=1}^{R} p(z_r) log(\frac{1}{p(z_r)}) = MI(Z, W) + H(Z)_W \tag{4.3}$$

The expression $MI(Z, W)$ stands for spatial mutual information, which measures how much the information from sample windows W reduces the uncertainty of Z within a specific window. Meanwhile, $H(Z)_W$ represents the remaining uncertainty within each sample window, averaged over all windows.

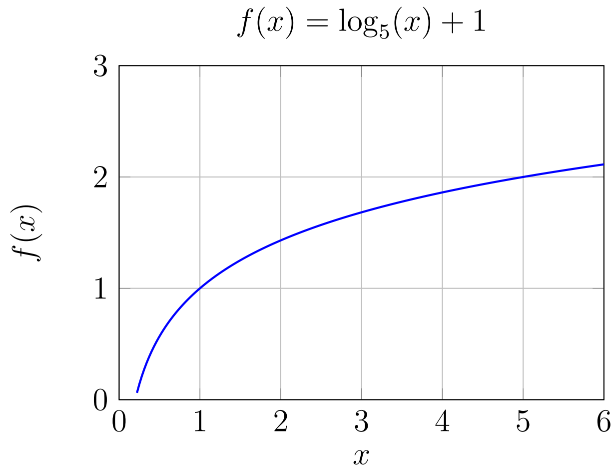


Figure 4.1: Behaviour of the function applied to the *stedelijkheid* (urbanity) variable.

4.2 Transforming entropy

A step beyond entropy is needed to measure the degree of liveliness in an area. Until now, the three different entropy measures only convey information about the distribution of amenities within a specific area. This means that when two specific areas have a similar distribution of amenities but vary greatly in terms of urbanity or size, they end up with similar entropy values. For this reason, the entropy values are divided by $\log_5(\text{sted}) + 1$, where sted is the variable *stedelijkheid* (urbanity) found in the CBS data. This value ranges from 1, denoting very urban, to 5, denoting not urban at all. Adding one to the logarithm ensures that the very urban areas are not weighed down while the least urban areas are weighed down by a maximum value of two. The function's behaviour applied to the *stedelijkheid* variable can be found in Figure 4.1.

4.3 Spatial autocorrelation

Spatial autocorrelation relates values of variables in specific areas with other values of the same variable in other areas (Rey, Arribas-Bel, & Wolf, 2023). It implies a relationship between what happens at two different points in space. Similar to non-spatial correlation, spatial autocorrelation can have positive, negative, and zero values. Positive values indicate that similar values are closer together while different values are further away. Negative spatial autocorrelation describes a situation where similar values are further away from one another. Otherwise, zero values indicate no correlation present.

4.3.1 Global spatial autocorrelation

Global spatial autocorrelation focuses on the overall trend within a specific area (Rey et al., 2023). When the global autocorrelation of a certain municipality (*gemeente*) or district is calculated, it conveys whether similar areas in terms of entropy or liveliness within those boundaries are close to one another or spread out. A very high value of global autocorrelation would imply that the area of interest has "hot spots". A low value would mean that lively and sleepy subareas are more evenly spread out. Global spatial autocorrelation is calculated by first obtaining the spatial weights matrix W, which provides equal weights that sum to one for all direct neighbours of a particular area. Next, we can calculate Moran's I by the formula:

$$I = \frac{n}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}\, z_i\, z_j}{\sum_i z_i^2} \tag{4.4}$$

Where n denotes the number of observations, z_i consists of the standardised value of our target variable at location i, and w_{ij} corresponds to the corresponding value in our weights matrix W.

4.3.2 Local spatial autocorrelation

Global spatial autocorrelation can provide valuable insights into whether the values in a specific area are clustered or dispersed. However, it does not offer information about the specific locations of these clusters (Rey et al., 2023). We can focus on the relationships between observations and their surroundings using local spatial autocorrelation. Similar to global spatial autocorrelation, we obtain the spatial weights matrix W from our area of interest. Then, we can use the local Moran's I statistic to classify high or low-value areas depending on them being surrounded by other high or low areas, resulting in classifications HH, LL, LH, and HL. The statistic is given by:

$$I_i = \frac{z_i}{m_2} \sum_j w_{ij} z_j \ ; \ m_2 = \frac{\sum_i z_i^2}{n} \tag{4.5}$$

Where m_2 denotes the variance of the target variable in the distribution, $z_i = y_i - \bar{y}_i$ and $w_{i,j}$ is the spatial weight. n is the total number of observations or, in our case, areas. We can extract the statistical significance of the classifications by simulating this process for several random distributions.

Chapter 5

Pipeline

This section will discuss the step-by-step process of moving from collecting amenities to calculating the liveliness of a specific area. For each area in the dataset of a certain scale (municipalities, districts, neighbourhoods), the amenities are collected, categorised, subcategorised and filtered using three different prespecified filters. After this, entropy and liveliness are calculated for the various types of entropy. Finally, global and local autocorrelation are calculated. A visualisation of the pipeline can be found in Figure 5.1.

5.1 Obtaining area geometries and collecting amenities

To obtain the geometry of an area, a geodataframe is built from the CBS data using the Geopandas library (Jordahl, 2014). For each row in the dataframe, the

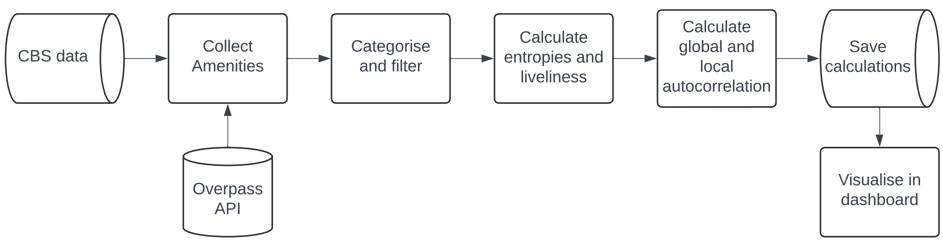


Figure 5.1: Schematic pipeline.

geometry bounding box is calculated and used to collect all amenities within that bounding box. This is done by sending a query through the Overpass Turbo API, which collects nodes, ways and relations of shops, amenities, railways, healthcare, sports, crafts, public transport and tourism. The API returns the query result in JSON format, which is then converted to a geodataframe.

The geometries of the ways and relations are converted to singular points by calculating the centroid of the geometry. This is done because the entropy calculations cannot be performed on lines and polygons.

Then, primary and secondary tags are extracted from the 'tags' column in the dataframe, which contains all categorical information about the nodes as a dictionary. The primary tag tells which type of node was returned by the API (e.g. an amenity, shops or sports node). The secondary tag is extracted by using the primary tag as a key for the dictionary.

5.2 Categorising amenities

After extracting the primary and secondary tags, the amenities can now be categorised into predefined categories. Each node gets a main category (L0) and a subcategory (L1). There are 13 L0 categories subdivided into a total of 53 L1 categories. A tree representation of how the categorisation is done can be found in Figure 5.2. The main categories and their respective amount of subcategories can be found in Table 5.1. The full distribution and distributions per category per scale of the L0 categories collected can be found in Appendix A (Figure A.1 and Figure A.2).

5.3 Applying filters

Because not every available amenity might interest every user, three different filters have been constructed to explore the impact of excluding specific amenities when doing entropy calculations. For instance, the category 'private transportation' consists mainly of car parking spaces. Not only is this category not relevant for a person who does not own a car, but it can also paint a distorted picture of the area as parking spaces tend to be counted individually, resulting in a neighbourhood having many private transportation amenities compared to the other amenities. The different constructed filters are in Appendix B.1.

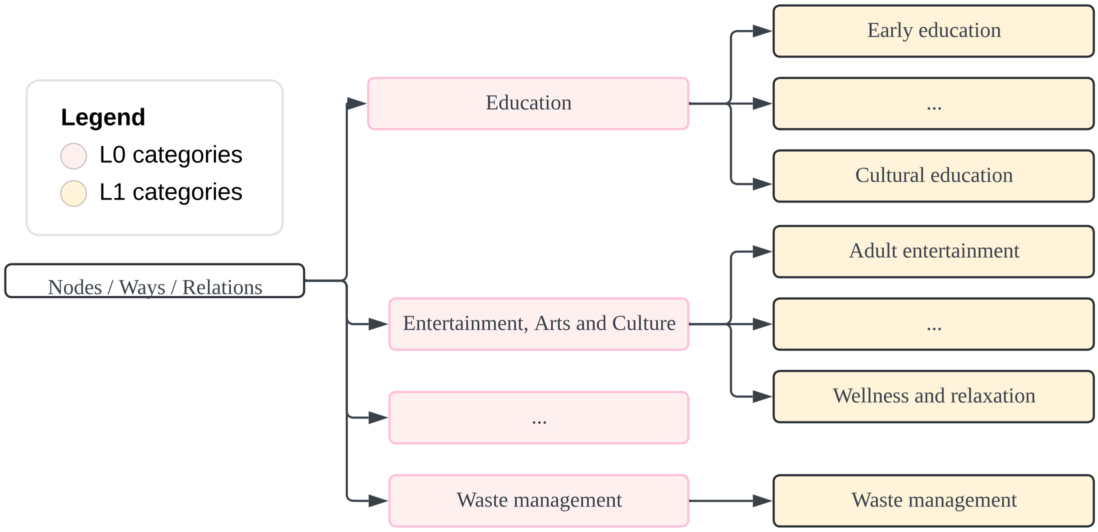


Figure 5.2: Tree giving an example of how the different amenities are categorised.

5.4 Calculating entropies and liveliness

After amenities have been collected, categorised and filtered, both Shannon and spatial entropy are calculated for the area. Shannon entropy is calculated by manually calculating the probabilities of the resulting distribution and applying the entropy function from the Scipy stats package (McKinney et al., 2010). Both Leibovici and Altieri entropy are calculated using the SpatialEntropy package (Zheng, 2021), which is a Python adaptation of the R package *SpatEntropy* constructed by Altieri, Cocchi and Roli (2023). Next, liveliness is calculated by directly applying the method found in Section 4.2.

5.5 Calculating global and local autocorrelation

Lastly, local and global autocorrelation are calculated by making use of the Pysal library (Rey & Anselin, 2007). Calculations are done on the different entropy measures and the measure of liveliness described in Section 4.2. The spatial weights matrix is calculated based on the Queen contiguity pattern (Getis & Aldstadt, 2004), meaning that every area is weighted against its direct neighbours. Using this pattern instead of a nearest neighbour approach ensures that a particular area's outskirts are not weighted against areas that might have multiple areas in between.

L0 category	**No. L1**
Education	5
Entertainment, Arts and Culture	6
Facilities	1
Financial	2
Healthcare	9
Places of worship	1
Private transportation	4
Public service	5
Public transportation	3
Shopping	12
Sports	4
Sustenance	4
Waste management	1

Table 5.1: Number of L1 categories per L0 category.

5.6 Calculating area similarity

To get insight into which areas are similar to one another, a function for measuring the Euclidian distance between multiple areas is set in place. This function is based on the transformed altieri entropy for the level 0 and 1 categories and Moran's I of those areas based on the same metrics. These features are chosen because they contain information about the spatial distributions of amenities, the area's urbanness, and its autocorrelation with its surrounding areas.

5.7 Dashboard architecture

A dashboard is constructed for users to explore the results of the different methods used. This dashboard is built using the dash framework, enabling easy integration of various HTML elements and graphs (Plotly Technologies Inc., 2015). The dashboard will feature a map of the Netherlands that users can interact with on municipalities, districts and neighbourhood scales. The aim is to give users an understanding of the different measures used to gather insights about the different spatial distributions of amenities in the Netherlands by enabling them to interact with different maps and change the applied metrics and filters. Further, we will add additional interactivity by providing insights into various areas. When particular areas are selected, the socio-demographic information of the respective area will be shown as a table showing the most similar places to the selected one.

Chapter 6

Results

This chapter provides an overview of the results gained from the application of the methods. Firstly, we will compare the performance of different entropy measures by applying them to synthetic data and by finding correlations within the actual data. Secondly, we will contextualise entropy by comparing results on three different areas in the data. Thirdly, the chapter will focus on the outcomes of the entropy transformations towards a notion of urban liveliness. Fourthly, an overview of the performance of spatial autocorrelation is given. Lastly, the elements of the constructed dashboard are highlighted and explained.

6.1 Comparing entropy measurements

This section will focus on building intuition about how the different entropy measures discussed in Section 4.1 behave in various scenarios. Firstly, the measures are tested on synthetic data to simulate comparative entropy measurements. Next, a view is given of how they correlate when applied to the actual data on the scales of municipalities, districts and neighbourhoods. Lastly, examples of different entropy scores for different districts from the data are given.

6.1.1 Behaviour of different measurements

To evaluate how the different entropy measures react to different point distributions. Two sets of 300 synthetic points representing synthetic geolocations have been constructed. Each point has been given one out of seven arbitrary labels representing different amenity categories; the labels are uniformly distributed for each set. For each set, the degree of clustering has been gradually increased. In the first set, the points gradually cluster into seven clusters of the same label. In the second set, the spatial clustering is the same, but the labels are shuffled, resulting in seven clusters of random labels. This has been done because, in the actual data, amenities are not always clustered by category, and we want to show the behaviour of the methods in both scenarios. The results can be seen in Figure 6.1.

For each set, the three different entropy values are calculated and plotted in Figure 6.2. For both the regular and shuffled sets, the Shannon entropy stays the same for each degree of clustering. This is to be expected as Shannon's entropy cannot account for spatiality; the distribution of the different labels stays the same, and therefore, Shannon's entropy does not change. The Leibovici entropy shows little change regarding the shuffled set, while it does show change regarding the regular set. In Leibovici entropy, a new variable Z is introduced, which accounts for the co-occurrences of points in space. This results in a higher overall entropy and exhibits responsiveness when clustering increases between the same labels. When the labels are randomised, however, the measure becomes unresponsive as co-occurrence does not change. The Altieri entropy shows the same sensitivity for both the regular and shuffled sets, as can be seen by the lines in Figure 6.2 running almost directly on top of each other. This can be explained by Altieri's introduction of the variable W, which presents a series of sample windows for each pair of points on top of the variable Z. The decomposition into spatial residual entropy and spatial mutual information makes it more sensitive to clustering behaviour in general, which can also explain the inconsistency around cluster sizes 14 and 20.

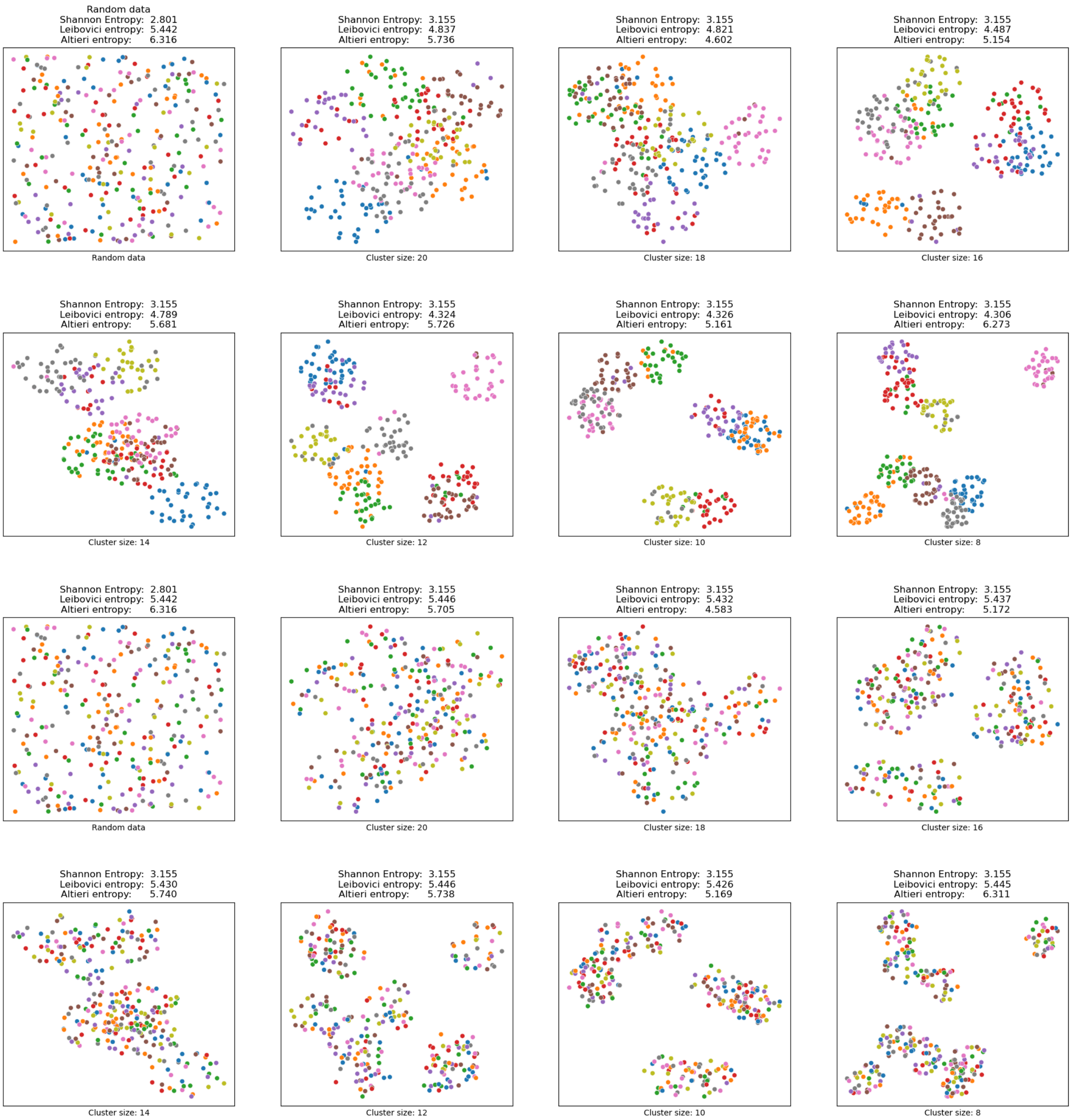


Figure 6.1: Scatters with entropy values for various degrees of clustering. Rows one and two show clusters per label, whereas rows three and four show clusters of random labels.

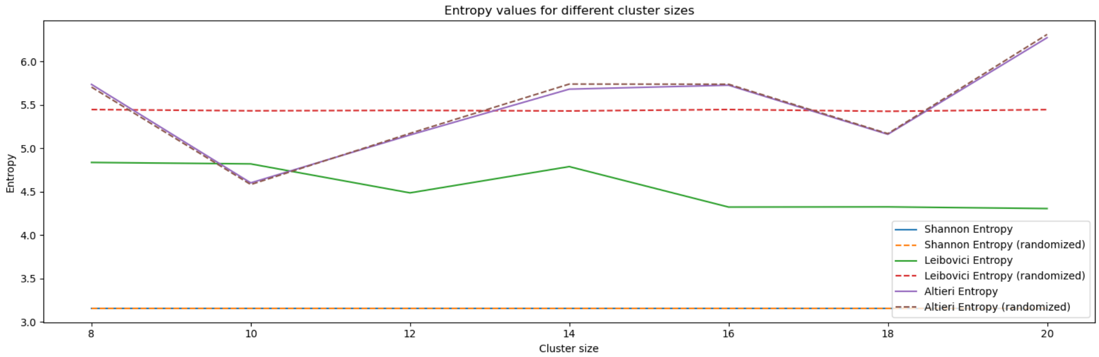
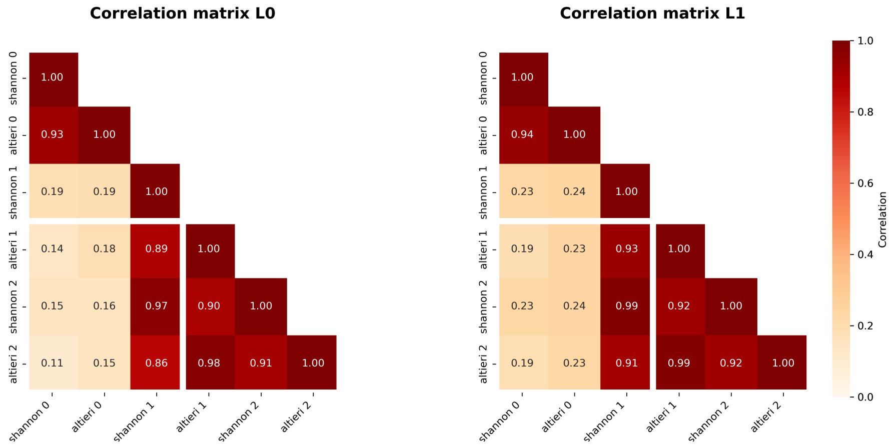


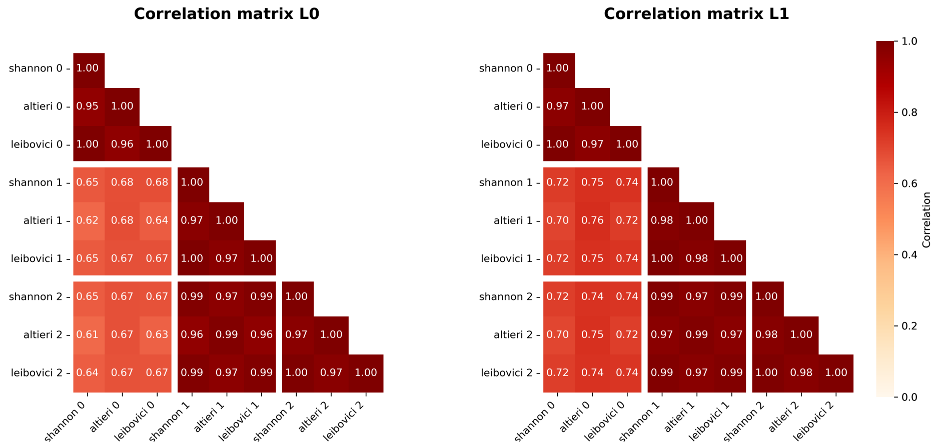
Figure 6.2: Lineplot showing how Shannon, Leibovici and Altieri entropy react to different degrees of clustering within the same distribution.

6.1.2 Correlations from the dataset

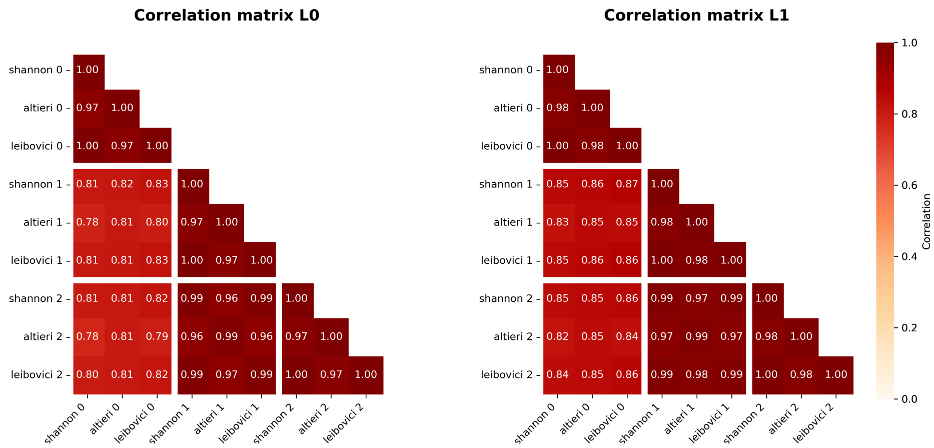
After applying the different entropy measures to municipalities, districts, and neighbourhoods, the values have been normalised to inspect possible correlations. For each area scale, a correlation matrix has been calculated for all filter levels on both the L0 and L1 categories; these can be found in Figure 6.3. When inspecting these matrices, a recurring pattern can be spotted. Each measure within each filter level correlates strongly with the other measures within that filter level. Next to that, filter levels 1 and 2 correlate strongly with each other, while filter level 0 tends to be more weakly correlated with the other filter levels. This is most clear at the scale of municipalities as the values at level 0 tend only weakly to correlate with the rest. Therefore, the organization of living spaces can be better understood by examining similarities at higher detail levels, such as across neighbourhoods and districts rather than municipalities. These patterns are consistent across different categories.



(a) Correlation matrices for municipalities



(b) Correlation matrices for districts



(c) Correlation matrices for neighbourhoods

Figure 6.3: Correlations between the different measures of entropy on different filter levels.

6.1.3 Examples of different districts

Three different districts have been sampled from the municipality (*gemeente*) of Amsterdam, with a low, medium and high Altieri entropy value for the L0 category and filter 1. Both a plot of the district and a bar plot of the amenities can be found in Figure 6.4. The first district, Tuindorp Buiksloot, has very few overall amenities and little variation in the amenities available, resulting in a low entropy score of 2.114. The second district, Rijnbuurt, has many amenities, but the distribution is dominated by the Entertainment, arts, and culture categories, resulting in a medium entropy value. Lastly, Oostelijke eilanden/Kadijken, does not have the most amenities out of the three but has a more varied and more evenly spread distribution, resulting in a higher entropy value. This implies that spatial entropy measures enable us to contextualise space not only by the number of amenities present and their placement but also by the distribution of different categories.

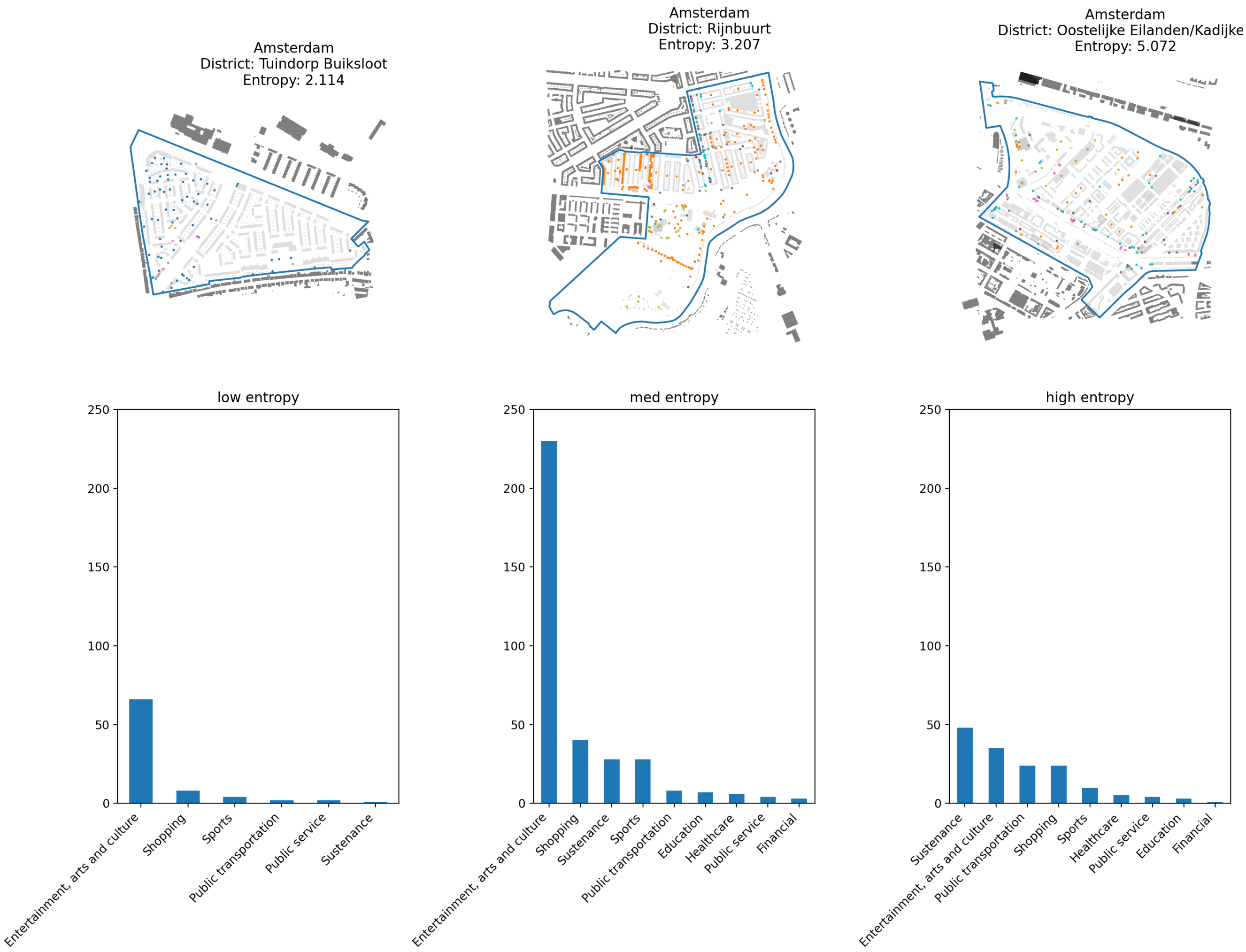
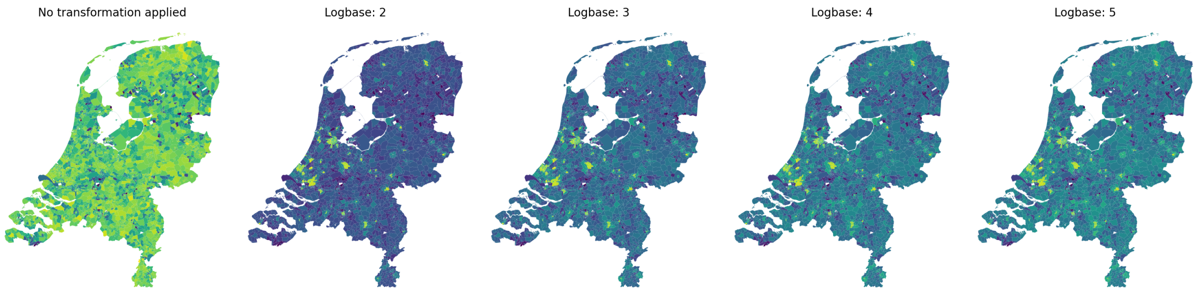


Figure 6.4: Amenity plots of three sampled districts in Amsterdam.

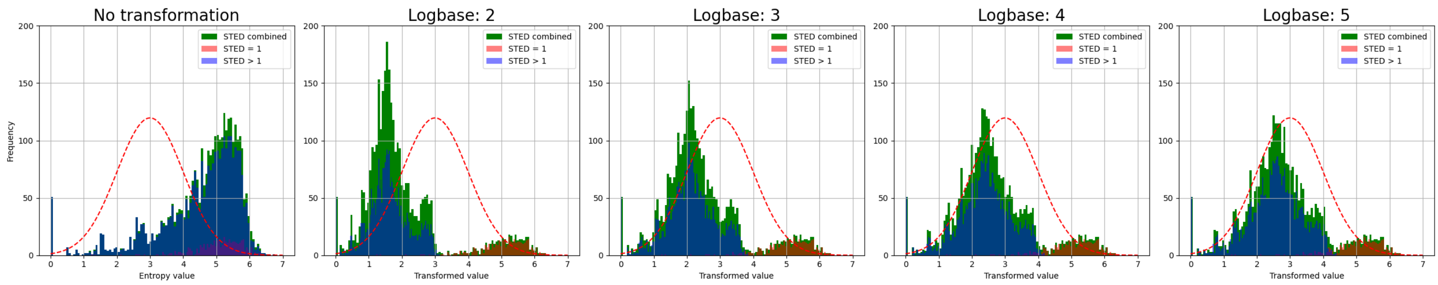
6.2 Entropy transformations

To find the right log base for transforming the entropy values into a measure of liveliness, the function described in Section 4.2 is applied with multiple log bases. The results are plotted as a map and as a histogram and can be found in Figure 6.5. In Figure 6.5b, it can be seen that when increasing the log base, the distribution of transformed values increasingly matches a bimodal distribution, with the more significant part matching a normal distribution. The divide in this bimodal distribution reflects the shift from rural to urban areas, as can be seen in

Figure 6.5b. Figure 6.5a shows that after the transformations are applied, large cities in the Netherlands are left with higher values than rural areas, indicating that urban areas tend to be more lively than rural ones.



(a) Map of districts with entropy values denoting colours



(b) Value distributions of untransformed and transformed entropy values

Figure 6.5: Entropy transformation function applied to L0 altieri values with filter 1 at the scale of districts.

6.3 Spatial autocorrelation

Local spatial autocorrelation is calculated on districts using the transformed altieri values on the L0 category with filter 1. 5 is used as the log base for the transformation. For each district, local Moran's I is calculated, and the values are divided into five quantiles, which are then plotted under local statistics in Figure 6.6. Next, each district is given a label denoting whether it is a high or low value and if it is surrounded by high or low values. This is plotted under the name Scatterplot Quadrant. Lastly, the statistical significance of each district is used as a mask to indicate which district has a p-value of less than 0.5. It would be consistently observed when performing multiple random simulations. This procedure leaves

36.83% of the district as significant. What stands out is that the bigger cities in the Netherlands tend to be hot spots, while a substantial part of the north of the Netherlands tends to have low values.

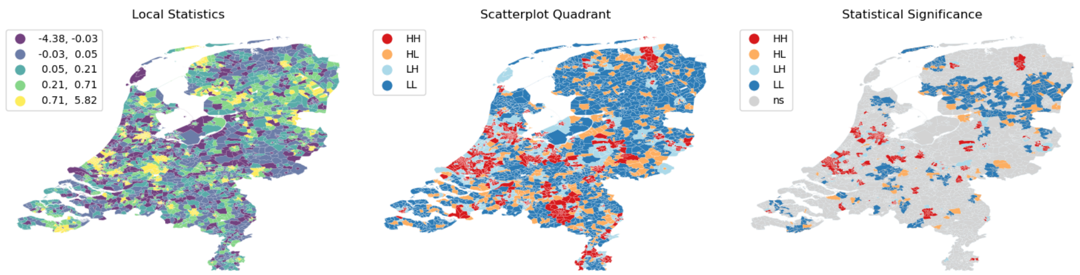


Figure 6.6: Local autocorrelation for district using the transformed L0 altieri values with filter 1.

6.4 Dashboard

The dashboard that has been created features two aspects: The main view, where the values of different measures can be explored and a side panel, where specific areas can be inspected for further insights. Both can be observed in Figure 6.7.

6.4.1 Main view

The main view of the dashboard features a map that the user can interact with by clicking and dragging. municipalities, districts, and neighbourhood scales can be selected at the top right. Beneath the map, different controls are placed to tweak different parameters. The user can select which entropy measurement, category level, and filter level are to be represented between the normalised and transformed entropy values.

6.4.2 Side panel

The side panel reveals itself when a specific area on the map is selected. It aims to provide the user with further insights into the area by giving a map with the amenities present and a barplot of the amenities per L0 category. Besides this, a second barplot reveals information about the socio-demographic composition of the area. This plot tells the user about the descent, ages and gender proportions of the inhabitants. Lastly, a table is included that displays the five most similar areas based on the function described in Section 5.6.

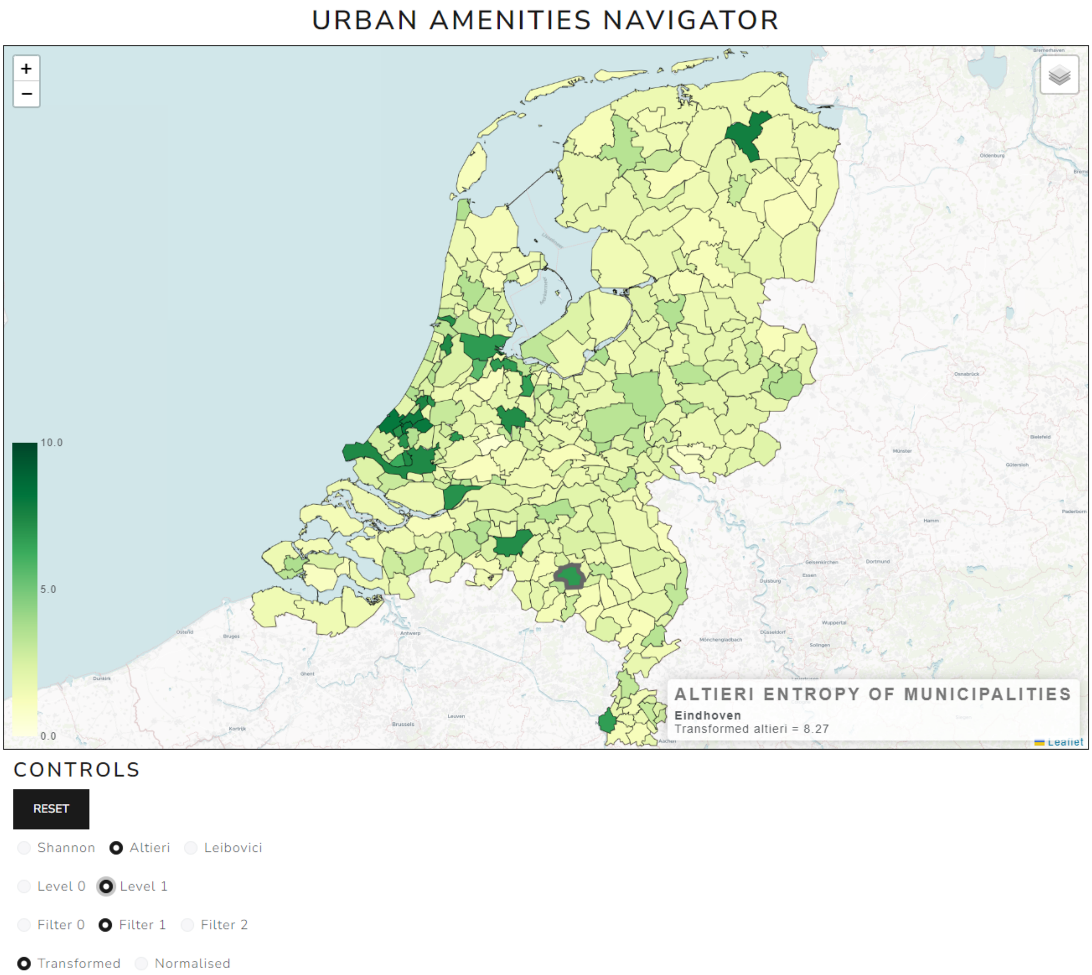
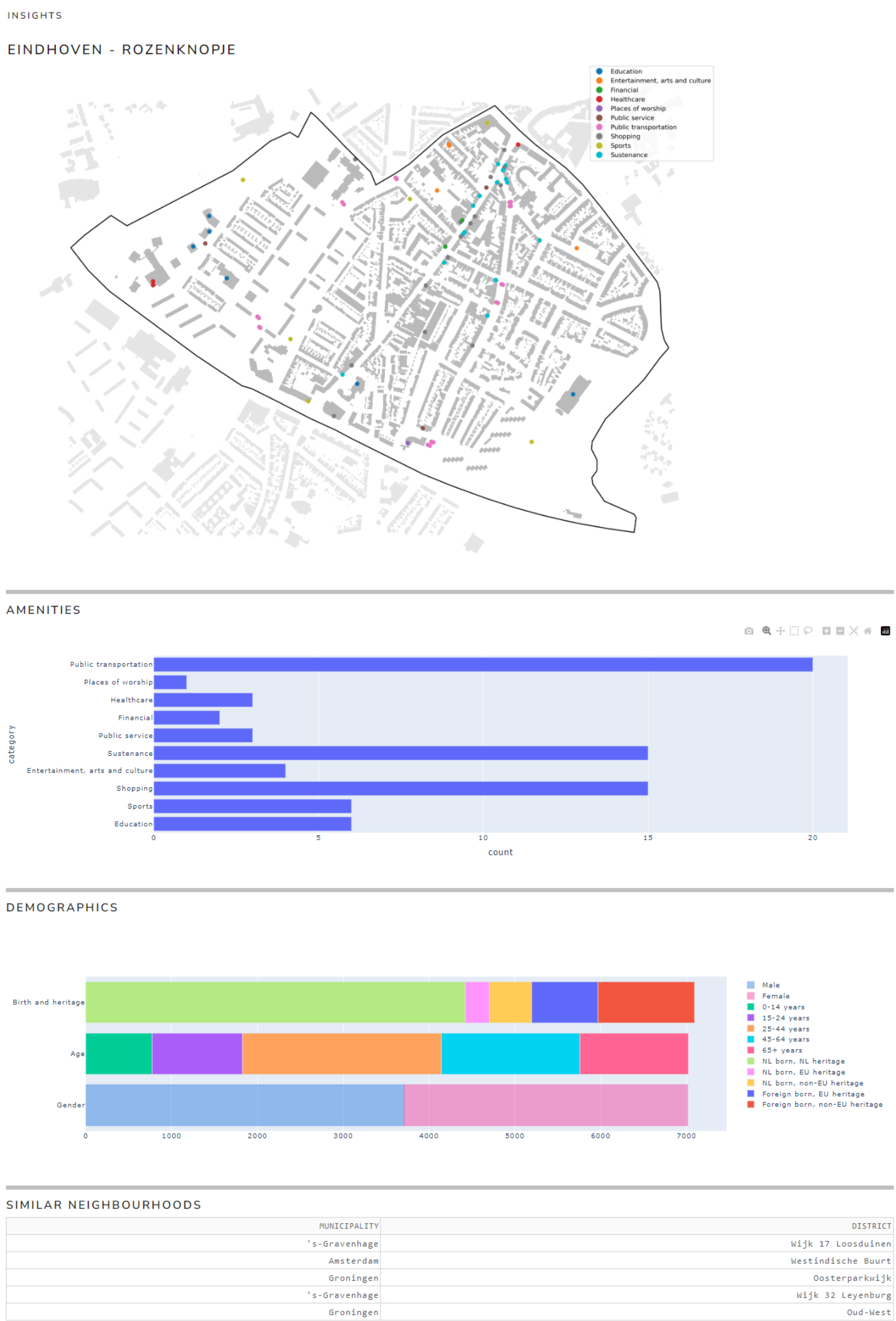


Figure 6.7: The main view of the dashboard where different settings can be applied.



MUNICIPALITY	DISTRICT
's-Gravenhage	Wijk 17 Loosduinen
Amsterdam	Westindische Buurt
Groningen	Oosterparkwijk
's-Gravenhage	Wijk 32 Leyenburg
Groningen	Oud-West

Figure 6.8: The inspection panel that appears when selecting a specific district.

Chapter 7

Conclusion and discussion

This section contains final concluding remarks summarising this thesis's achievements. Finally, limitations and suggestions for further research are provided.

7.1 Conclusion

This thesis explored the performance of Shannon entropy (Shannon, 1948), Leibovici entropy (Leibovici et al., 2014) and Altieri entropy (Altieri et al., 2019a) when applied to spatial distributions, and the possibility of transforming those values into a measure of urban liveliness (Jacobs, 1961; Yue et al., 2021; Wu & Niu, 2019) and using those measures for computing spatial autocorrelation (Anselin, 2019). Through the application of these methods, several insights were obtained, ranging from capturing the spatial context of urban living space to establishing a similarity profile among areas.

We have found that entropy measures can effectively capture the complexity and heterogeneity of urban areas. High entropy values are typically associated with areas that not only have many amenities present but whose amenities are also distributed well over multiple categories, as can be seen in Section 6.1.3. Combining this with the measure of urbanity by applying a log-based weighting shown in Section 4.2 proved an effective way of capturing urban liveliness and resulted in a refined delineation between urban and rural characteristics shown in Section 6.2

The use of spatial autocorrelation methods displayed in Section 6.3, particularly local Moran's I, enabled us to find hot and cold spots within the Netherlands's urban landscape. This allowed for better insight into the relationships between different areas with various spatial resolutions, providing us with the information that even though the Netherlands might seem very uniform in its urban landscape, that is not always the case.

All the methods used and their results have been condensed into a dashboard. This dashboard, displayed in Section 6.4, serves as an interactive platform where users can explore different areas in the Netherlands at the municipal, district, and neighbourhood levels. It allows users to not only see which areas have high or low urban liveliness but also provides information about the socio-demographic characteristics of a specific area and identifies areas similar to the one being inspected.

7.2 Discussion

Despite the thesis's promising achievements, there is no guarantee that the data supporting these achievements is complete. OSM data is predominantly collected by volunteer contributors, resulting in varying levels of documentation across different cities in the Netherlands and differences between urban and rural areas in general. It would be intriguing to apply the same methods in the future when the OSM dataset is even more comprehensive. An alternative to consider is enforcing deep learning techniques for satellite image processing that could classify remote sensing data.

While the dashboard is designed to offer a user-friendly interface for exploring various aspects of the Netherlands, the current results are restricted to only three predetermined filter levels. To better meet the specific needs of individual users, it would be beneficial to implement customizable filters. However, this was not achieved in this thesis as it was considered to be too computationally expensive. In an ideal solution, entropy calculations could be done on a remote server with the hardware capable of providing results in a quick manner.

In the future, it would be beneficial to incorporate a temporal aspect into the project by including OSM data from different years. This could lead to intriguing findings related to liveliness measures and spatial autocorrelation by shedding light on the behaviour of various areas. Such an approach could offer valuable insights into clustering behaviour and the dynamics between different areas. It could, for instance, show whether development policies have made an impact on the urban landscape or shed light on the agglomeration or separation of broader areas in terms of urban growth. Another suggestion for future research is to explore the possibility of building a model to predict the socio-demographic aspects of a particular area based on the distribution of amenities or vice versa. A model like this could reveal interesting patterns within our landscape and could guide policymakers to better design an urban built environment that could enhance social cohesion and urban dynamics.

References

Altieri, L., & Cocchi, D. (2022). Understanding the expansion of italian metropolitan areas: A study based on entropy measures. *Environment and Planning B: Urban Analytics and City Science*, *49*(2), 447–463.

Altieri, L., Cocchi, D., & Roli, G. (2019a). Advances in spatial entropy measures. *Stochastic environmental research and risk assessment*, *33*, 1223–1240.

Altieri, L., Cocchi, D., & Roli, G. (2019b). Measuring heterogeneity in urban expansion via spatial entropy. *Environmetrics*, *30*(2), e2548.

Altieri, L., Cocchi, D., & Roli, G. (2023). Efficient computation of spatial entropy measures. *Entropy*, *25*(12), 1634.

Anselin, L. (1995). Local indicators of spatial association—lisa. *Geographical analysis*, *27*(2), 93–115.

Anselin, L. (2019). A local indicator of multivariate spatial association: extending geary's c. *Geographical Analysis*, *51*(2), 133–150.

Batty, M. (1974). Spatial entropy. *Geographical analysis*, *6*(1), 1–31.

Cabral, P., Augusto, G., Tewolde, M., & Araya, Y. (2013). Entropy in urban systems. *Entropy*, *15*(12), 5223–5236.

Centraal Bureau voor Statistiek. (2023, 11). *Wijk- en buurtkaart 2023.* Retrieved from `https://www.cbs.nl/nl-nl/dossier/nederland-regionaal/geografische-data/wijk-en-buurtkaart-2023`

Getis, A., & Aldstadt, J. (2004). Constructing the spatial weights matrix using a local statistic. *Geographical analysis*, *36*(2), 90–104.

Jacobs, J. (1961). *The death and life of american cities, the failure of town planning.* Pelican Book, Londres.

Jordahl, K. (2014). Geopandas: Python tools for geographic data. *URL: https://github. com/geopandas/geopandas*.

Karlström, A., & Ceccato, V. (2000). A new information theoretical measure of global and local spatial association.

Kelly, A. H. (2006). Securing urban amenity: does it coincide with biodiversity conservation at the local government level? *Australasian Journal of Environmental Management*, *13*(4), 243–253.

Leibovici, D. G. (2009). Defining spatial entropy from multivariate distributions of co-occurrences. In *Spatial information theory: 9th international conference, cosit 2009 aber wrac'h, france, september 21-25, 2009 proceedings 9* (pp. 392–404).

Leibovici, D. G., Claramunt, C., Le Guyader, D., & Brosset, D. (2014). Local and global spatio-temporal entropy indices based on distance-ratios and co-occurrences distributions. *International Journal of Geographical Information Science*, *28*(5), 1061–1084.

Lynch, K. (1984). *Good city form.* MIT press.

McCrea, R., Stimson, R., & Marans, R. W. (2011). The evolution of integrative approaches to the analysis of quality of urban life. In *Investigating quality of urban life: theory, methods, and empirical research* (pp. 77–104). Springer.

McKinney, W., et al. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th python in science conference* (Vol. 445, pp. 51–56).

OpenStreetMap contributors. (2017). *Planet dump retrieved from https://planet.osm.org .* `https://www.openstreetmap.org`.

Plotly Technologies Inc. (2015). *Collaborative data science.* Montreal, QC: Plotly Technologies Inc. Retrieved from `https://plot.ly`

Randall, T. A. (2008). Preferences of suburban residents in thunder bay, ontario towards neighbourhood intensification and rediversification. *Canadian Journal of Urban Research*, *17*(2), 28–56.

Rey, & Anselin. (2007). PySAL: A Python Library of Spatial Analytical Methods. *The Review of Regional Studies*, *37*(1), 5-27.

Rey, Arribas-Bel, D., & Wolf, L. J. (2023). *Geographic data science with python.* Chapman and Hall/CRC.

Saraiva, P. (2023). On shannon entropy and its applications. *Kuwait Journal of Science*, *50*(3), 194–199.

Saville-Smith, K. (2010). Neighbourhoods and intensification: Measuring sustainability impacts of higher density and mixed use. *SB10 New Zealand*.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, *27*(3), 379–423.

Wu, W., & Niu, X. (2019). Influence of built environment on urban vitality: Case study of shanghai using mobile phone location data. *Journal of Urban Planning and Development*, *145*(3), 04019007.

Yue, W., Chen, Y., Thy, P. T. M., Fan, P., Liu, Y., & Zhang, W. (2021). Identifying urban vitality in metropolitan areas of developing countries from a comparative perspective: Ho chi minh city versus shanghai. *Sustainable Cities and Society*, *65*, 102609.

Zheng, Y. (2021). *Spatialentropy.* `https://github.com/Mr-Milk/SpatialEntropy`. GitHub.

Appendix A

Dataset size

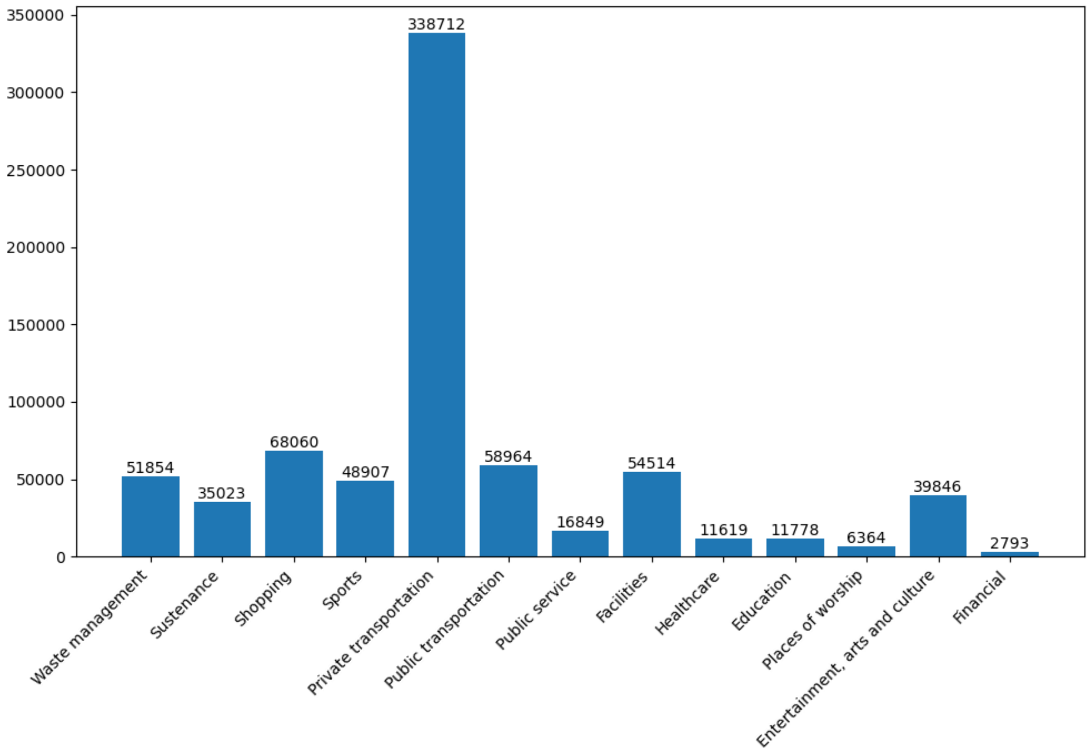


Figure A.1: Barplot of the amount of amenities for each category.

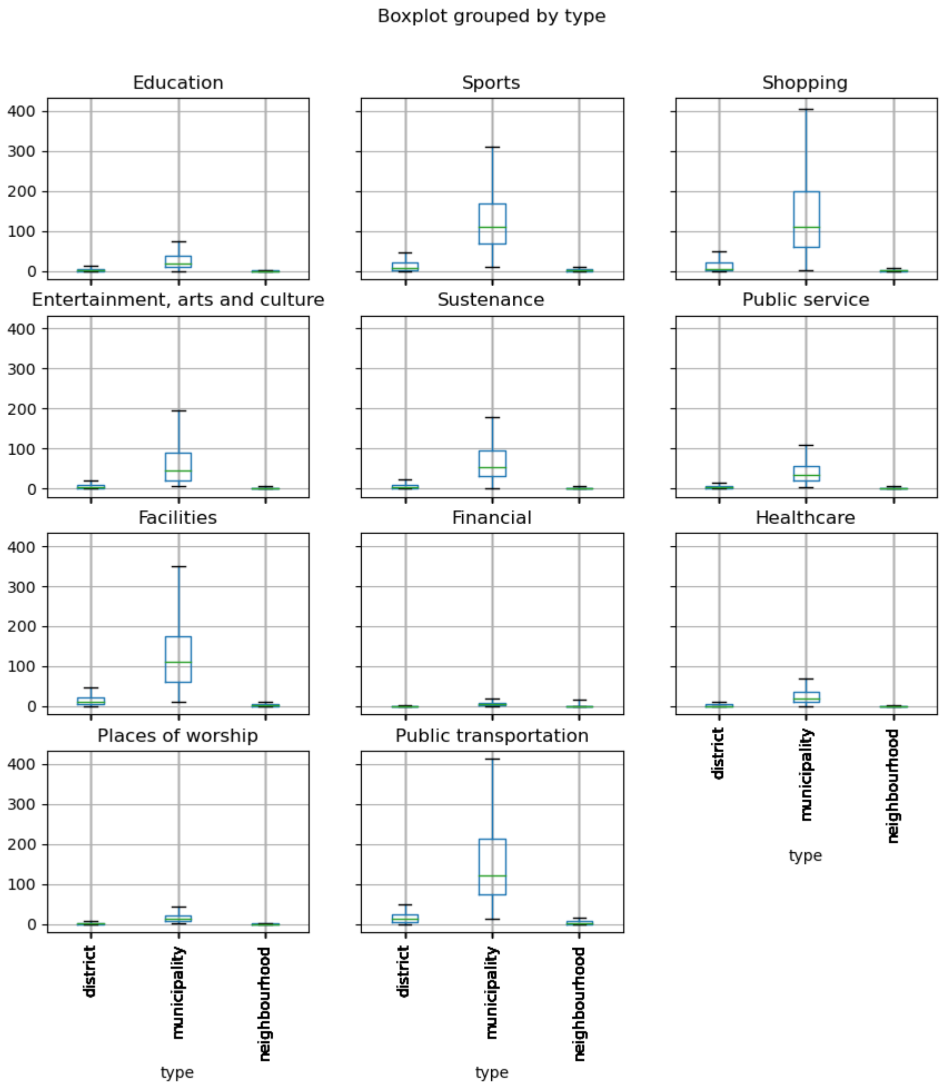


Figure A.2: Boxplots describing the distribution of amenities for each L0 category for the scales of municipality, district and neighbourhood.

Appendix B

Collecting and filtering amenities

B.1 Overpass API query

```
[out:json];
(
    nwr[shop][!leisure]({{bbox}});
    nwr[amenity][!leisure]({{bbox}});
    nwr[leisure][!amenity]({{bbox}});
    nwr[railway~"station"]({{bbox}});
    nwr[sport][!shop][!amenity]({{bbox}});
    nwr[healthcare][!amenity]({{bbox}});
    nwr[craft][!amenity]({{bbox}});
    node[public_transport][!railway]({bbox}});
    nwr[tourism~"gallery|theme_park|zoo|museum|aquarium"]
        ({{bbox}});
);
out geom;
```

B.2 Filters

Filter level	**L0 category**	**L1 category**
0	Uncategorised	Uncategorised
1	Private transportation	**X**
	Facilities	**X**
	Waste management	**X**
	Healthcare	Other
	Shopping	Other
	Public Service	Other
2	Financial	**X**
	Entertainment, arts and culture	Other
	Shopping	Clothing and accessories, Crafts House and interior Media appliances and hardware Mobility

Table B.1: Table portraying the different filters, **X** means the whole category is filtered. A higher filter level always includes the elements of the previous filter levels.