Journal of Internet Services and Applications, 2025, 16:1, doi: 10.5753/jisa.2025.5155
This work is licensed under a Creative Commons Attribution 4.0 International License..

Urban Cultural Signature with Web Data: A Case Study with Google Places Venues

Fernanda R. Gubert [**Univ. Tecnológica Federal do Paraná** | ***fernandagubert@alunos.utfpr.edu.br***]
Gustavo H. Santos [**Univ. Tecnológica Federal do Paraná** | ***gustavohenriquesantos@alunos.utfpr.edu.br***]
Myriam Delgado [**Universidade Tecnológica Federal do Paraná** | ***myriamdelg@utfpr.edu.br***]
Daniel Silver [**University of Toronto** | ***dan.silver@utoronto.ca***]
Thiago H. Silva [**Universidade Tecnológica Federal do Paraná** | ***thiagoh@utfpr.edu.br***]

Departamento Acadêmico de Informática, Universidade Tecnológica Federal do Paraná, Av. Sete de Setembro, 3165 Rebouças 80230-901 Curitiba PR Brasil.

Received: 03 November 2024 • **Accepted:** 23 April 2025 • **Published:** 01 July 2025

Abstract
Providing knowledge about the characteristics of diverse cultural groups worldwide and identifying cultural similarities between their respective occupation regions can yield significant economic and social benefits. However, much of the existing research in this field relies on user behavior data, which may limit scalability and generalization due to the difficulty in obtaining such data. To address this, our work focuses on extracting venue data from Google Places and proposing a methodology based on the Scenes concept to enrich this dataset for generating cultural signatures of urban areas. This approach also considers the influence of different area sizes. Using Curitiba, Brazil, and Chicago, USA, as case studies, the results demonstrate that the proposed method can identify cultural similarities between regions while supporting an area-division strategy for analyzing cities across different countries. The findings show consistency, as evidenced by the segmentation of Curitiba and Chicago into culturally distinct clusters. This highlights the societal benefits of the proposal, such as location recommendations based on cultural criteria and real-time service validation.

Keywords: Cultural Signature, Multiscale Analysis, Geolocated Data, Cultural Similarities, Google Places

1 Introduction

According to the UNESCO (United Nations Educational, Scientific, and Cultural Organization) Report by Rivière *et al*. [2009], the world is marked by significant cultural diversity, and understanding the characteristics of these diverse cultures presents a considerable challenge. One of the difficulties lies in the dynamic nature of culture — society evolves, requiring continuous reassessment of cultural attributes. Traditional data collection methods, typically conducted through questionnaires and interviews, face limitations, primarily due to the high costs of gathering data from large populations. Besides the cost, these methods lack scalability, are challenging to execute quickly – such as World Values Survey (WVS[1]), which is updated on average every 5 years – and often do not maintain a level of standard and quality in the data, due to misinterpretations by respondents [Einola and Alvesson, 2021; Jaeger and Cardello, 2022]. To work around this limitation, many recent studies resort to data from web sources to address challenges across various fields [Ilieva and McPhearson, 2018; Zhang *et al*., 2018; Hu *et al*., 2020; Chen *et al*., 2024], producing meaningful results more efficiently. Consequently, our research also uses web data as raw material.

Beyond dynamic, the concept of culture is complex and lacks a single definition, making the task of finding data that satisfactorily describes it far from trivial. Culture can be understood as a set of aspects of a given group of people, including, for example, language, religion, cuisine, and arts [Spencer-Oatey and Franklin, 2012]. Some studies show that eating and drinking habits are elements capable of describing local culture [Silva *et al*., 2017; Sproesser *et al*., 2022; Heath, 1995; De Brito *et al*., 2018; Laufer *et al*., 2015]; however, data of this type – usually user check-ins – in addition to being difficult to obtain, also give analytical priority to users' tastes rather than the lifestyle evoked by the characteristics of a place. Another approach follows the discourse of Mehta and Mahato [2019], in which the availability of resources and services that meet the population's needs could provide a sense of identity to the place. What draws attention in this second approach is the possibility of considering various aspects of culture, since a city's resources, that is, its venues, can be associated with different categories, such as religion, cuisine, and arts, in addition to being a format that is still little explored.

Even so, relying solely on the types of venues in an urban area may be insufficient to create cultural signatures [Silva *et al*., 2017]. The concept of Scenes [Silver and Clark, 2016] transforms the everyday experiences of a population in a given location into elements of cultural significance, weighting these elements according to the types of venues present.

[1] https://www.worldvaluessurvey.org/wvs.jsp

Our study aims to present a strategy for obtaining venue data from a global-scale web source (Google Places) and proposes a methodology, grounded in the Scenes concept, to enrich these data for generating cultural signatures of urban areas. The results indicate that the proposed method can identify cultural similarities between areas while supporting an area-division strategy for analyzing cities across different countries.

Identifying cultural similarities and tracking changes more quickly (due to the large-scale automated process) can benefit the provision of services in near real-time, allowing a company, for example, to understand the preferences for its product or service in different markets and make decisions based on cultural information from different areas.

Our study can also help with problems related to local recommendations. A tourist who has visited a city may receive recommendations for similar cities based on cultural criteria, while people seeking a place to live could be offered options that align closely with their culture of origin or preference.

Furthermore, our proposal opens the door to developing new tools and frameworks that allow organizations to evaluate and interpret the cultural dynamics of various locations. By maintaining an up-to-date understanding of cultural landscapes, these tools could support diverse applications, such as monitoring the impact of public policies on local culture.

The contributions of the present study can be summarized as follows.

- We explore an approach for extracting relevant features of urban areas using Google Places data to derive their cultural profiles. Applied to neighborhoods of Curitiba, Brazil, and hexagonal grids, in both Curitiba and Chicago, USA, the proposed methodology shows promising results for the automated identification of cultural boundaries and distinctions;
- We assess the impact of varying urban area granularity (hexagonal sizes) on the generation of cultural signatures and their practical outcomes, highlighting implications for different levels of area detail;
- We outline a possible application of the proposed approach, identifying similar areas in different cities and showing the potential for new applications, such as new area recommendation systems based on cultural criteria.

The remainder of this paper is organized as follows. Section 2 presents related work, and Section 3 describes the foundations of the Scenes Theory and how its different dimensions are applied to obtain the cultural signature. Section 4 describes the methodology for extracting data from the Google Places API and expanding the dimensions that characterize the venues to include the cultural information, followed by the validation of the mapping process using data available in the literature for the city of Toronto, Canada. Section 5 presents the results for Curitiba and Chicago using different granularity levels in urban area partitioning. Finally, Section 6 concludes the paper and presents directions for future work.

2 Related Works

In this section, we first explore works discussing how cultural practices shape city travel behavior and mobility choices. Second, we review approaches for identifying and comparing cultural traits across regions and the Google Places dataset. In the sequence, we focus on the unique cultural attributes of locations and methods for generating urban signatures. Finally, we discuss gaps in the literature that justify our work.

2.1 Culture and Urban Mobility

Recent work is taking advantage of data from web sources to explore issues in various areas, including areas related to culture – [Laufer *et al*., 2015; De Brito *et al*., 2018; Silva and Silver, 2025]. Senefonte *et al*. [2020] evaluate how regional and cultural characteristics influence the mobility behavior of tourists and residents. For the study, Foursquare-Swarm data shared on Twitter have been used. In the proposed methodology, a mobility graph for residents and several mobility graphs for tourists were constructed for each country, depending on their respective countries of origin. This approach makes it possible to analyze how much the origin of users influences their choices, as well as the chosen destination. Transitions in the graph occur between categories of locations, and the matrix that represents the graph is transformed into a mobility vector, making it possible to calculate behavioral distances and explore the cultural characteristics of different nationalities and different destinations. The results show that the tourists' origin greatly influences their behavior, especially when there is a significant cultural distance.

Also, based on the study of user behavior, Candipan *et al*. [2021] state that racial segregation is not only linked to neighborhoods where people of different races reside but also to the places where these people move during their daily activities. For this study, the authors used Twitter data from 50 US cities and created a dynamic measure of racial segregation called the Segregated Mobility Index (SMI). This measure is based on a mobility graph, in which the nodes are the neighborhoods, and the edges indicate the existence of trips between these neighborhoods, showing the isolation of people who live in certain neighborhoods, even in their daily activities, which may come from a racist historical legacy. Within this context and using data from SafeGraph, De La Prada and Small [2024] examine the extent to which people's regular trips in US cities are to neighborhoods with a different racial composition than their own – and why. The authors found that, on average, the trip is to a neighborhood with less than half the racial difference of the neighborhood of origin, in addition to sustainable popular policies that encourage people to carry out most of their daily activities in venues 15 minutes from home, discourage integration into residentially segregated cities, as trips closer to home are less racially diverse. On the other hand, it was identified that some neighborhoods have POIs with characteristics different from the neighborhood standard, favoring the construction of diversity networks.

In parallel, other studies explore urban mobility through

telecom data analysis. For example, Furno *et al*. [2016] investigates mobile traffic patterns in ten cities, revealing how communication behaviors are linked to urban structures. Its methodology focuses on normalizing telecom data, employing hierarchical clustering and statistical techniques to discover patterns in residential, commercial, leisure, and transport zones. The results demonstrate significant variations between countries while identifying shared behavioral trends, showing that mobile traffic data are a great tool for understanding urban dynamics. Similarly, Tang *et al*. [2024] leverages aggregated and anonymized telecom traffic data to infer urban functions from urban land use. Their study was carried out in Shenzhen, China, and combines time series decomposition and urban texture analysis to map functions such as housing, work and recreation, even identifying areas with special functions such as urban villages and roadside shops. This research emphasizes the potential of high-frequency telecom data to address traditional limitations of urban planning.

2.2 Cultural Similarities Between Areas

The present work aims to conduct a comparative analysis of geographic areas to identify cultural similarities. The approach proposed by Le Falher *et al*. [2015] addresses the challenge of comparing neighborhoods across cities. Using geolocated data from Foursquare in cities across Europe and the USA, the authors develop a methodology to characterize neighborhoods based on the activities that take place within them. To achieve this, they represented each venue as a feature vector, capturing its characteristics and general activity. Since a neighborhood consists of a set of these vectors, the authors employ the Earth-Mover's Distance (EMD) to measure the similarity between neighborhoods by calculating the distance between their respective vectors.

Also using Foursquare data, Çelikten *et al*. [2016] develop a probabilistic model to characterize regions based on the activities that take place within them and to identify similar regions across cities. This model considers various factors, including location, user participation, and the time of day and day of the week when activities occur. A probabilistic model is constructed for 40 cities worldwide, capturing the geographic distribution of locations. One key finding is that user behavior in utilizing a city's resources plays a significant role in highlighting relevant regional characteristics.

To examine a city's current sociological trends regarding the identity of its neighborhoods, Olson *et al*. [2021] use data from Yelp reviews to characterize areas. To discover hidden trends, which would not be possible with direct analysis, the authors propose a deep autoencoder approach. For this purpose, a low-dimensional vector is created for each neighborhood using LSA (Latent semantic analysis); after the encoder stage, the embeddings are created, and finally, the decoder is performed for validation. Temporal analyses show changes in neighborhoods and by performing clustering with K-means, similar neighborhoods in Toronto, for example, are identified.

In our previous study [Gubert *et al*., 2024b], we propose methods to identify cultural similarities between urban areas using data from the Google Places API. Two approaches are tested: one based on the frequency of place categories and another on Scenes Theory, which associates categories with cultural dimensions. Data are collected from 14 global cities and every US state. The results indicate that the Scenes Theory-based approach captures cultural nuances more expressively, reflecting patterns identified in population research.

2.3 Studies with Google Places Data

The Google Places API has some benefits, such as its broad worldwide coverage, which facilitates scalability. Using such data, Sen and Quercia [2018] create a methodology to measure the spatial capital of a neighborhood in a cheap and standardized way, facilitating scalability. Spatial capital is related to the resources and daily lives of inhabitants, such as easy access to health facilities and less frequent use of cars, thus increasing environmental sustainability and making the neighborhood more "livable". As part of the data extraction strategy, areas of 200m$\times$200m are delimited, and a matrix is created for each of them to identify venues in 30 categories. This way, it is possible to assess whether the area offers different categories of venues within walking distance. Then, these areas are grouped, showing the different spatial capitals within the same city, in addition to making comparisons between cities. With this information, it is possible to determine urban interventions, such as identifying poor areas and recommending the introduction of new services and venues.

Aiming to overcome the limitations of restricted availability of traditional socioeconomic data, Chen *et al*. [2024] propose an integrated framework for mapping large-scale urban building functions, combining geospatial data obtained from web platforms such as Google Maps and TripAdvisor. The methodology involves the automated collection of points of interest (POIs) and land use plots through web crawlers, in addition to the use of Microsoft building footprints. For building classification, an unsupervised machine learning algorithm (OneClassSVM) identifies residential structures based on landscape metrics, while the proportion of POI types and the area occupied by certain parcels are used to categorize non-residential functions such as hospitals, hotels, schools, stores, restaurants, and offices. The approach was validated in 50 cities in the United States, with detailed evaluations in Boston and Des Moines, demonstrating an average accuracy of 94%. The results indicate that the methodology is scalable and can be applied globally, offering a robust tool for urban planning, energy modeling, and socioeconomic studies in large urban areas.

Extending generative and parametric approaches in the context of urban design, such as ease of movement in a neighborhood and energy efficiency, Hidalgo *et al*. [2020] study the location patterns of venues using data corresponding to 47 US cities, coming from the Google Places API. The proposal consists of modeling the best combination of venues and identifying those that are over- or under-supplied in a neighborhood. A clustering algorithm is built to identify dense neighborhoods in venues to overcome the challenge of defining neighborhood boundaries. Once the neighborhoods are identified, the authors estimate the number of venues in each category, leveraging the kinship principle. In other words, the model predicts the number of venues expected to

be found in a neighborhood based on data on the other categories of venues already present. Next, a network is created connecting venues that are likely to be together, using Spearman's correlation and considering the number of times that the venue appears in the cluster. The final network shows the categories that tend to be together and the number of venues in each one. This network can be useful for predicting new venues given a set of inputs. Promoting the debate on the spatial definition of neighborhood limits, Martí P. and L. [2021] carry out a study using data from the city of Alicante, in Spain, also obtained from Google Places. One of the challenges is recategorizing the data, as many similar categories make a more detailed analysis difficult. The authors create functional clusters in terms of urban activity, which are then contrasted with the administrative limits of the neighborhoods. As a result, the research confirms the existence of a disconnection between traditional administrative partitions of the neighborhood and the functional organization of the city, which can be of great value in the urban planning process.

2.4 Cultural Signature of Areas

Aiming to stimulate the creation of cultural signatures for different areas, Silva *et al*. [2017] represent user preferences regarding eating and drinking habits using Foursquare check-ins. Their proposed methodology enables the identification of cultural boundaries and similarities between societies at various scales. The approach involves generating a binary-valued vector for each user to represent their preferences. The sum of these vectors characterizes a region, and comparing regions is achieved by calculating the cosine similarity between their corresponding feature vectors. The spatio-temporal results demonstrate the potential to explain users' cultural habits and, through cultural signatures, quantify the similarity between different regions.

To identify cultural similarities through beer preferences, De Brito *et al*. [2018] use data from Untappd, a location-based social network (LBSN) specializing in beer. First, the data are grouped according to a classification by ethnic characteristics; then, each city is represented by a vector that indicates users' preferences for each of the previously created categories, reflecting a kind of cultural signature. Using hierarchical clustering, similar areas were identified. As a result, the authors observed that the differences in preference for beer in cities in the same country were smaller when compared to the differences between cities in different countries, showing that this aspect can be significant in studying similarities between cultures.

Bancilhon *et al*. [2021] have found that one way of quantifying the culture of a society is through the names of city streets, after discovering that these reflect the society's value system. For this, data are collected from public sources from 4,932 honorific streets (streets dedicated to historical figures) in Paris, Vienna, London, and New York. Their findings revealed the presence of gender bias, though a recent trend shows an increasing number of streets being named in honor of female figures. The study also highlighted which professions are considered elite and how much external influences shape a city's identity.

The study of Gogishvili and Müller [2024] aims to analyze how iconic cultural buildings influence the cultural geography of cities over time. The methodology combines spatial data analysis with urban cultural theory, focusing on the geographic locations of significant cultural landmarks and their impact on the surrounding urban environment. By examining case studies from cities worldwide, the research traces the evolution of these buildings' locations and their role in shaping urban identities and cultural landscapes. The main findings suggest that the placement of such buildings has become a strategic tool in urban regeneration and cultural branding, with their locations shifting in response to economic and political changes. Additionally, the study highlights the role of these buildings in attracting tourists and reinforcing a city's cultural significance on a global scale, indicating a growing emphasis on cultural capital in urban development strategies.

Knowing that the spatial configuration of the different components of cities is relevant for codifying the aspects that created such an arrangement and also for being responsible for sustaining results, such as economic productivity and environmental sustainability, Arribas-Bel and Fleischmann [2022] present spatial signatures as a characterization of space based on the form and function of an urban environment. Firstly, a partition of the space is carried out, which is combined with a unifying approach to urban form and function called Enclosed Tessellation (ET) cells, uniting morphological and functional characteristics for the classification of the space. Then, the information from the ET cells is grouped using the K-means method, standardizing the data that reflect the form and function and, finally, generating the cities' spatial signatures.

Sparks *et al*. [2020] investigate the geosocial and temporal patterns of urban cultural behavior by analyzing the distribution of points of interest (POIs) across different cities. The authors aim to understand how the location and time-based activities associated with POIs reflect the cultural identities and behaviors of urban populations. The methodology involves using large-scale data from LBSNs like Foursquare and Yelp, which provide geotagged check-ins and user interactions at various POIs. By applying clustering and temporal analysis techniques, the authors identify distinct geosocial temporal signatures for each city, revealing how urban cultures differ regarding activity patterns, preferences, and social interactions. The main findings highlight significant differences in how cities globally structure their cultural and social behaviors, with certain cities exhibiting strong patterns of temporal clustering. In contrast, others show more diverse, less time-bound patterns. The study also demonstrates that these geosocial temporal signatures can be used to predict cultural trends and urban dynamics based on digital footprint data. Focusing on understanding the power of new machine learning methods based on graphs in urban area cultural signature prediction, Silva and Silver [2025] introduce a graph neural network method for predicting local culture signatures. They validate their method using Yelp data, showing that it could help predict local culture even when traditional local information, such as census data, is unavailable.

2.5 Discussion of Related Work

Studies such as Senefonte *et al*. [2020] and De La Prada and Small [2024] use mobility data from large-scale sources, such as LBSNs and SafeGraph, to evaluate how tourists and residents interact with urban space. Although they share the objective of mapping cultural patterns, these works focus on the movement of people and not on the structure of urban spaces themselves. On the other hand, our work proposes an independent model to express urban cultural characteristics, allowing a more structural characterization of urban areas.

Analyzing studies focused on city structure, Furno *et al*. [2016] and Tang *et al*. [2024] use telecom data to infer urban functions and traffic patterns. Although they also employ clustering and statistical analysis, their focus is on urban infrastructure rather than the culture of spaces. Likewise, works such as Arribas-Bel and Fleischmann [2022], Martí P. and L. [2021] and Chen *et al*. [2024] use spatial segmentation to define urban patterns, but their methods emphasize the form and function of cities, whereas our study explores urban culture through venues.

Our work is more closely related to studies presented by Silva *et al*. [2017], De Brito *et al*. [2018] and Sparks *et al*. [2020], which generate cultural signatures based on check-ins and user preferences. However, these studies rely on user behavior, which may limit their scalability and generalization, due to the difficulty of obtaining such data. The present work addresses these limitations by using location data directly, without requiring explicit user actions. Additionally, our study differs from other works, such as [Silva and Silver, 2025], that apply machine learning techniques to predict urban culture. Our study does not envision performing predictions.

In our previous study [Gubert *et al*., 2024a], we characterized urban areas based on city resources, developing comparative analyses focused on digital signatures that reveal cultural similarities. A key contribution of that research is the introduction of a methodology that expands dimensions based on venue categories, creating enriched cultural signatures using Google Places, a globally accessible data source. This methodology was compared with other less robust methods in another prior study [Gubert *et al*., 2024b], which examined cities and states at different granularities. The findings show that the methodology based on Scenes Theory offers a more nuanced understanding of cultural patterns. The present study builds on a previous work [Gubert *et al*., 2024a], with several key extensions: i) the inclusion of a second city, Chicago, USA, to broaden insights; ii) an assessment of the impact of varying urban area granularity (smaller subdivisions within cities); and iii) an indication of how the results can be applied to identify similar areas across cities.

3 Scenes Theory

This section describes Scenes Theory, including its 15 dimensions, the details of its scoring systems, and the possible cultural signature that results from it.

3.1 Fundamentals: The 15 Dimensions

The Scenes Theory aims to balance the meanings, styles, and aesthetics of human experience characteristics with the precision of physical sciences [Silver and Clark, 2016]. It combines cultural elements to form "scenes". These combinations can occur in various ways, creating scenes from different historical moments and geographic locations.

The concept of the scene aims to explain how, when, where, and why certain people come together around specific tastes and cultural activities, extending beyond the "common values" and inherent "ways of life" of each culture. To identify the elements that characterize scenes, a balanced approach is taken, integrating systematic theory with empirical analysis. This approach draws on diverse cultural sources, including poetry, religion, journalism, ethnographic research, and philosophy.

In Scenes Theory, three general types of meaning are addressed — theatricality, authenticity, and legitimacy — and such meanings exist in various traditions of thought, from Weber [1930] **on legitimacy** to Goffman [1974] **on theatricality** and Simmel [1971], **on authenticity**, among others. Authenticity evaluates how the scene points to something considered genuine rather than false, theatricality portrays how the scene describes the presentation, in its clothes, speech, manners, posture, bearing, and appearance, while legitimacy estimates what is believed to make actions right or wrong.

However, a need was identified to analyze scenes using more specific terms that convey their unique characteristics. These terms, referred to as **dimensions**, include 15 specific elements. The general types of meaning discussed above are interconnected and mutually reinforcing; the same applies to the dimensions. The 15 dimensions are organized below by general type of meaning, each followed by a brief description.

- **Theatricality**: performance, display.
  - Glamour: endowed with dazzling, sparkling aspects and mysterious and seductive characters.
  - Neighborliness: it's about friends and fellow comrades, coming together as a warm, caring community.
  - Transgression: breaks conventional appearance styles, opposing what is considered routine, whether concerning behavior, clothing, or good manners.
  - Formality: values highly ritualized and ceremonial dress patterns and aspects of speech and appearance in general.
  - Exhibitionism: the self becomes an object to be looked at, an exhibit to be admired.
- **Authenticity**: about the sources of your being, where the "real you" comes from, the dimensions expand from the particular to the generalized.
  - Locality: belonging to and rooted in this place and this place alone, not "contaminated" by foreign customs.
  - Ethnicity: these are ethnic customs, with deep, unchosen feelings, endowed with original practices.

  - State: extends characteristics, customs, ideas, and locations from the state to the national.
  - Corporateness: it is about the authenticity of big brands, which transcend states, regions and ethnicities, establishing themselves globally, being genuine with what they offer and claiming the loyalty of many.
  - Rationality: asserts that the true self is in the mind, the spontaneous exercise of reason is deeper than the arbitrary and external circumstances of location, ethnicity or nationality.

- **Legitimacy**: concerns the basis of moral judgments, the authority on which a verdict of right or wrong is founded, oriented by time (past, present and future) and space.
  - Tradition: the past is an enduring authority that extends into the present, it is the creation of a connection with the past that informs the reasons for acting in the here and now.
  - Charisma: it is an indescribable quality of great figures, such as artists and celebrities, leading others to follow them.
  - Utilitarian: is based on profit and productivity, evokes the importance of a cost and benefit analysis.
  - Egalitarian: consists of respect for human equality, all people deserve justice and equal treatment.
  - Self-Expression: is the expression of an individual's personality, with their unique vision, style and actions.

These 15 dimensions serve as tools to break down a scene into a series of distinct elements. Additional dimensions can be included, but these 15 already provide a strong foundation for capturing the scenes' cultural essence. When translating these dimensions into categories of venues, it can be observed that various venue types together form a scene, and this collection becomes a key indicator for measuring the scene. This approach creates a more holistic view, as the same venues can take on different meanings, demonstrating that no single venue alone creates a particular scene.

The selection of Scenes Theory for this research is grounded not only in its robust theoretical foundation but also in the fact that prior studies have initiated the application of this framework to analyze venue data in various regions [Silver and Clark, 2016; Gubert *et al*., 2024b; Silva and Silver, 2025], showing its usefulness in practice.

3.2 Dimension Scoring System

To translate seemingly non-cultural data into sources of information about cultural significance, in [Silver and Clark, 2016] the authors worked with a team of coders who helped assign weights to the dimensions of all types of venues present in their database, from NAICS (North American Industrial Classification System) and YP (Yellow Pages). NAICS is a government-maintained North American classification system that includes various indicators useful for compiling local "scenes," such as religious organizations, art galleries, environmental organizations, and more. These data are openly and publicly available in the U.S.[2], Canada[3] and Mexico[4]. YP makes online Canadian business, product and service data available on a platform called "Yellow Pages" [YP, 2022].

According to Silver and Clark [2016], the coders received several instructions and immersed themselves in the project using a tutorial and a manual titled "The Coder's Handbook." This handbook includes a set of standardized questions for coding each dimension, highlights common pitfalls, and provides a series of examples with justifications. Coders focused on one dimension at a time, facilitating comparative analyses across different types of venues. The translation process lasted approximately one year and involved dozens of meetings, which led to repeated revisions and clarifications until consensus was reached in cases where scoring discrepancies occurred.

The scoring system was designed to ensure a clear and standardized procedure, guiding decision-making in each case. Each venue receives a score from 1 to 5 on each of the 15 dimensions. Scores of 4 or 5 indicate that the venue affirms the dimension, while scores of 1 or 2 suggest rejection. A score of 3 signifies a neutral stance toward the dimension. The most critical decision lies in assigning a positive (4 or 5) or negative (1 or 2) score. Coders reserve the extreme scores (5 and 1) for cases where a venue's label clearly and directly signals (or does not signal) a particular dimension as central to its meaning. Scores of 4 and 2 apply when a venue often or sometimes suggests a positive or negative orientation toward the dimension. Importantly, dimension scores are not classifications; rather, they serve as tools to identify the experience types that characterize each location, reflecting the overall experience promoted by all venues that comprise a scene.

The databases NAICS and YP, enriched with dimension scores called seed vectors, were analyzed by Silver and Clark [2016], alongside other important social domains. The study examines the scenes' contribution to economic growth and prosperity, their relationship with residential patterns, and the considerable variation in voting and other political activities according to local context. The findings reinforce the significant insights provided by scenes and affirm the effectiveness of translating venue types into the theory's dimensions. In this way, the scores assigned to various categories of venues serve as foundations for mapping additional datasets.

3.3 Cultural Signature

With the scoring system, each venue v ultimately receives one or more vectors, each encompassing the dimensions corresponding to its associated category(ies). Thus, its scenes model is represented as a matrix $S^v_{K\times D}$, where K denotes the number of categories associated with venue v, and D represents the total number of dimensions in the Scenes Theory. In this representation, each element $s^v_{k,d}$ at row k and column d corresponds to the d^{th} score in the k^{th} category assigned

[2]https://www.census.gov/naics/
[3]https://www.statcan.gc.ca/en/concepts/industry
[4]https://www.inegi.org.mx/SCIAN/

to venue v. The information stored in $S^v_{K\times D}$ can be averaged to derive a unique **scenes vector** $\mathbf{s}^v = \{s^v_1, s^v_2, \ldots, s^v_D\}$, $s^v_d = \frac{1}{K}\sum_{k=1}^{K} s^v_{k,d}$. This vector provides a compact representation of the venue's scene across all dimensions.

To obtain the cultural signature of a region, its scenes model is represented as a matrix $S_{V\times D}$, where V denotes the total number of venues in the region, and each row is given by the venue scenes vector $\mathbf{s}^v$. To measure the overall scene of a region, a vector $\mathbf{c} = \{c_1, c_2, \cdots, c_D\}$ is computed, where the d^{th} element of $\mathbf{c}$ is given by:

$$c_d = \frac{1}{V}\sum_{v=1}^{V} s^v_d. \tag{1}$$

Thus, each element in the vector $\mathbf{c}$ represents the average score across all V venues in the region. This result reflects the ***cultural signature*** of the region, represented by $\mathbf{c}$, also referred to as the performance score.

The *cultural signature* enables the cataloging and comparison of scenes without requiring physical visits. Additionally, automatically compiling all captured details and their context enhances results, as manual processing may introduce omissions and isolated interpretations. Specifically, manual analyses of only a small set of venues can create a misleading impression of the overall scene's meaning. Therefore, establishing a standardized measurement for the scene provides a more effective solution to this issue.

Aiming to expand the analysis to other urban areas and create new cultural signatures, this work proposes mapping the categories of venues in the dataset retrieved from Google Places to the $D = 15$ dimensions presented in the Scenes Theory. For this, we utilize existing mappings from the seeds of each category k in the Scenes Theory dataset $\{\mathbf{s}_k(\text{Scenes})\}$ and categories in the Yelp database $\{\text{categ(Yelp)}\}$ to provide the Google Places (GP) venue scenes matrix as

$$S^v_{K\times D}(\text{GP}) = f\left(\{\mathbf{s}^v_k(\text{Scenes})\}, \{\text{categ}^v(\text{Yelp})\}\right). \tag{2}$$

This mapping enables addressing areas as scenes and comparing their *cultural signatures* (Eq. 1), as it encompasses a diverse set of venues that provide different dimensions of meaning. For more details, refer to Subsection 4.2.

While our cultural signature model aims to capture a broad range of cultural dimensions, it does not cover all aspects, such as digital or informal cultural expressions, which may not be tied to physical venues. This limitation is also present in other studies exploring cultural differences, which often focus on specific aspects like eating and drinking habits, mobility patterns, and urban spatial configurations [De Brito *et al*., 2018; Senefonte *et al*., 2020; De La Prada and Small, 2024; Arribas-Bel and Fleischmann, 2022]. Despite this, our model mitigates part of those limitations by incorporating a broader variety of cultural categories per area and leveraging the cultural semantics derived from Scenes Theory.

4 Procedures for Cultural Extraction

This section outlines the procedures for cultural extraction used in the study, structured into three subsections. The first details the retrieval of venue data from the Google Places API for a given city (Toronto is used as an initial example, while Curitiba and Chicago are analyzed in the experiments). The second subsection describes the mapping of Google Places categories to the $D = 15$ dimensions of Scenes Theory, leveraging existing mappings from the Scenes dataset ($\{\mathbf{s}_k(\text{Scenes})\}$) and categories in Yelp ($\{\text{categ(Yelp)}\}$) to enhance semantic accuracy. The third one explains the validation method. It employs pre-existing datasets for Toronto, Canada, and evaluates results using Pearson and Spearman correlation analyses.

4.1 Extracting Data From Google Places

We chose to use Google Places as it provides one of the most comprehensive and reliable dataset for location information worldwide. It is important to clarify that we also considered open-source alternatives such as OpenStreetMap. However, we found that its level of detail and quality did not meet the requirements of our analysis, particularly for mapping the different categories of locations found in urban areas.

The Google Places API returns geolocated data on venues and points of interest. In addition to providing location coordinates as latitude and longitude pairs, each venue is associated with at least one category describing its type ($K \geq 1, \forall v$). There are 141 categories in total, i.e., $|\{\text{categ(GP)}\}| = 141$; however, these lack the specificity needed to create detailed cultural signatures. For example, the API provides a general "restaurant" category for venues that classify themselves as such, but it does not specify the type of cuisine, such as Italian or Japanese, which is essential for this work.

Aiming to address this issue, the optional keyword parameter was used in API calls. The Google Places service searches this parameter's text within the indexed content of venues, returning matches ordered by perceived relevance. Although this parameter is not specifically designed for venue-type searches, the API documentation ensures valid results when the entries include a location name, address, or venue category, making it a practical choice for our purposes. Categories from Yelp have been used as the keyword parameters due to their higher level of detail. The Yelp database consists of user venue reviews, with available categories organized into a four-level hierarchy. For this study, only the most specific categories (leaf nodes) are adopted, excluding some that are not relevant to our aims, resulting in 888 categories used as keyword parameters. Thus, for each venue v in GP, its keyword is given by kw(v,GP) = $\{\text{categ}^v(\text{leaf,Yelp})\}$, resulting in $700 \leq |\{\text{categ(GP)}\}| \leq 888$, which is almost 5 times larger than the original size (141).

For each API request, we must specify a pair of geographic coordinates, and to obtain them, we use the following strategy – illustrated in Figure 1 for Toronto. First, we must represent an area of interest in the city, and this is done by providing two coordinates representing the extreme northeast and southwest points to delimit a rectangle – see Figure 1 (part 1). Next, we create a grid with cells of sides of 9,000 meters, and the geographic coordinates of the central point of each cell are retrieved – see Figure 1 (part 2) – cells that cover areas outside the city of interest are disregarded after a manual inspection. Notice that they differ from the subdivisions

(neighborhoods or hexagons) used to generate the cultural signatures discussed in Sections 4 and 5.

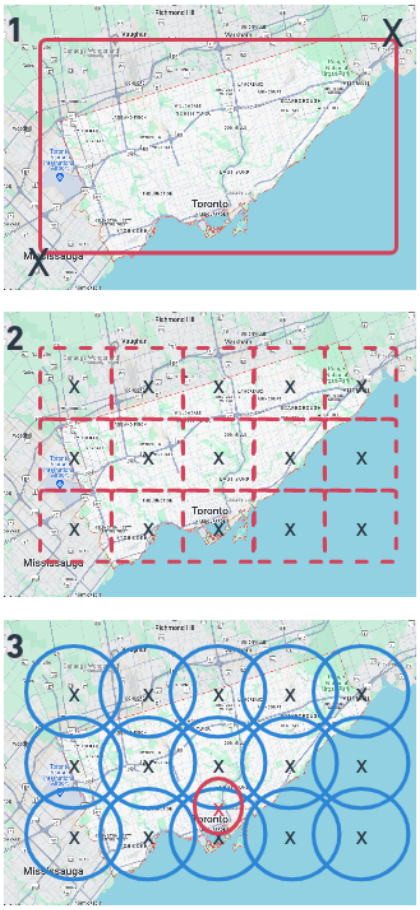

Figure 1. Rectangular area delimited for Toronto (part 1). Grid cells with sides of 9,000 meters (blue squares) (part 2). Automatically calculated circular areas (in blue) and highlighted city center (red circle) (part 3).

Considering the costs involved—each subarea requires one API call per category—and the goal of gathering substantial data, a radius of 6,000 meters is used for coordinates distributed throughout the city, covering the entire region – Figure 1 (part 3) blue circles. We standardize the radius size across cities rather than the number of coordinates, as cities vary widely in size. A fixed number of coordinates across cities would result in varying venue densities per subarea, potentially affecting the creation of cultural signatures and hindering fair comparison between the cities. To address density issues inside the same city, an exception to the radius size is made for a central coordinate with a radius of 3,000 meters – Figure 1 (part 3) red circle. The rationale for a smaller radius at the central coordinate is the API's limit on venues returned per request (20 per page, with a maximum of 60 using paging), which helps reduce venue loss in these denser areas.

For each coordinate and radius, we make 888 requests representing the enrichment categories. After data extraction, no missing data have been found, though duplicate records (around 30%) have been observed due to some overlap between small areas, as expected. These procedures facilitated the development of a tool to streamline this process[5] [Gubert and Silva, 2022]. The number of venues and unique categories found in each of the cities in this study are presented in Table 1, in addition to the number of geographic coordinates used in the requests to cover each area (the larger the area reported, the more coordinates).

[5]https://github.com/FerGubert/google_places_enricher

Table 1. Number of venues v, enriched unique categories $|\{\text{categ(GP)}\}|$, and coordinates used in each city to extract them.

City	Venues	Categories	Coordinates
Toronto	62.282	818	11
Curitiba	31.539	748	5
Chicago	55.063	839	9

As can be seen, thanks to our enrichment strategy, all cities in our final dataset have more than 700 categories, expressing a considerable diversity in terms of venues, much higher than the 141 basic categories provided by the GP API.

4.2 Mapping Categories Into Dimensions

The categories retrieved from Google Places must be mapped to the $D = 15$ dimensions of Scenes Theory. To accomplish this, both the existing category mapping from the Scenes dataset (referred to as "seeds") and an auxiliary Yelp category mapping are used as references [Silva and Silver, 2025]. Figure 2 provides an overview of the mapping process, which is detailed in what follows. For simplicity, the venue index v has been omitted.

The first part of the mapping process (top left in the figure) is based on the seeds of Scenes Theory, whose set $\{\mathbf{s}_k(\text{Scenes})\}$ encompasses score vectors of categories in NAICS and YP databases as shown in Eq. 3.

$$\{\mathbf{s}_k(\text{Scenes})\} = \{\mathbf{s}_k(\text{NAICS})\} \cup \{\mathbf{s}_k(\text{YP})\}, \qquad (3)$$

where $\mathbf{s}_k(.) = (s_{k1}, s_{k2}, ..., s_{kD})$, is the $D = 15$ dimensional score vector of a category k present in at least one of the databases. Each NAICS and YP score vector $\mathbf{s}_k$ is derived from the manual scoring described in Subsection 3.2.

Here, it is important to highlight that although the theory was initially developed with data sets from the USA and Canada, the transfer learning to a broader platform, as we are proposing in this work with Google Places, aims to extend the application of this theory to other regions of the world.

For the auxiliary mapping of Yelp categories (top center of Figure 2, representing the mapping Scenes → Yelp), Silva and Silver [2025] mapped the dimension scores for each category by semantically comparing each category k in Yelp (categ_k(Yelp)) with the every categories k' in Scenes Theory (categ'_k(Scenes)). Thus, each 15 dimension vector shown in the top center of the figure is given by Eq. 4.

$$\mathbf{s}_k(\text{Yelp}) = \mathbf{s}_{k^*}(\text{Scenes}), \qquad (4)$$

where

$$k^* = \text{argMax}_{k'}\ \text{SemMatch}\left[\text{categ}_k(\text{Yelp}), \text{categ}_{k'}(\text{Scenes})\right]$$

.

Although the Yelp database (Yelp) is also primarily focused on the Global North, in this work, it is used as an

auxiliary tool to support the direct mapping $Scenes \rightarrow$ GP through $Scenes \rightarrow$ Yelp $\rightarrow$ GP, which enables transferring knowledge from $Yelp$, without restricting generalization to other countries and regions.

To enhance semantic detail and mapping accuracy, descriptive sentences are created for each category in the Yelp database as shown in the top-right of Figure 2. The categories in Yelp are organized in a 4-level hierarchy, and the associated sentences incorporate all levels; i.e., for each category at the lowest level, the associated sentence includes all categories in the path up to the top level as shown in Eq. 5.

$$\text{Sent}_k(\text{Yelp}) = \text{concat}\left[\text{categ}_k(i, \text{Yelp})\right], i \leq 4. \quad (5)$$

For example, Active Life, at the top level ($i = 1$), is included to construct Yelp sentences associated with the categories Amusement Parks and Water Parks (both at i =leaf). Moreover, it is important to point out that this procedure does not affect the score vectors, which remain unchanged and will be transferred to the next mapping stage.

In the last mapping stage (Yelp $\rightarrow$ GP), as depicted at the bottom of Figure 2, the broader description of categories in Google Places can be turned more informative using the Yelp database. However, before this enrichment step, a cleaning procedure is necessary. When analyzing data from Google Places, we observed that two Google categories ("point of interest" and "establishment") accounted for 99% of the data, not offering therefore meaningful descriptions of venue types, so they were removed. For a more effective description of venues, in Google Places, the sentences encompass both the selected Yelp categories used in the requests and the broader categories provided by Google (categ(GP)). For example, if a venue v has the Yelp categories "Amusement Parks" and "Water Parks" along with the Google category "Tourist Attraction", the descriptive sentences are: "Amusement Parks Tourist Attraction" and "Water Parks Tourist Attraction". This first step in the last stage is described as:

$$\text{Sent}_k^v(\text{GP}) = concat\left[\text{categ}_k^v(\text{leaf,Yelp}), \text{categ}_k^v(GP)\right] \quad (6)$$

Given the existing mapping of "seeds" of Scenes Theory to the Yelp categories (Eq. 4) and the enrichment of Google data with some Yelp categories (Eq. 6), we chose to perform the second step in the last mapping stage directly with Yelp. This mapping process was carried out with SBERT, using the framework Sentence Transformers, in which several pretrained models with a large and diverse dataset of more than 1 billion training pairs are made available and can be used to calculate embeddings ($\mathcal{E}$) from sentences and texts for more than 100 languages [Reimers and Gurevych, 2019].

After selecting some suitable models for our purposes based on available documentation, we carried out experiments with sample data to evaluate the results. For each sentence associated with the venue v in category k in Google Places ($\text{Sent}_k^v(\text{GP})$), the generated embeddings $\mathcal{E}$ are compared using cosine similarity (cosS), retrieving the highest-scoring Yelp sentence as detailed in Eqs. 7 and 8.

$$\mathbf{s}_k^v(\text{GP}) = \mathbf{s}_{k^*}^v(\text{Yelp}), \quad (7)$$

$$k^* = \text{argMax}_{k'}\ \text{cosS}\left[\mathcal{E}(\text{Sent}_k^v(\text{GP})), \mathcal{E}(\text{Sent}_{k'}^v(\text{Yelp}))\right]. \quad (8)$$

Then, we select around 50 random sentences from Google Places and map them to the respective Yelp sentences. Through manual observation and judgment, it was decided to go with the "all-MiniLM-L6-v2" model, which generally shows more coherence and assertiveness. In reviewing sample results, we find that single-word Yelp categories, such as "German", lacked sufficient context for effective mapping, which led to unsatisfactory matches. These cases represented about 5% of the data and were subsequently excluded.

With this refined mapping, each venue v is associated with one or more vectors $\{\mathbf{s}_k^v\}$ reflecting the 15 dimensions of Scenes Theory, depending on the number of associated sentences $\text{Sent}_k^v(\text{GP})$ – during the transfer learning process, we confirmed that finding a matching was always possible. Notably, each vector carries equal weight in representing the venue, regardless of the specific categories forming the sentences. Resuming to Eq. 2, we can construct matrix $S_{K\times D}^v(\text{GP})$, where each element $s_{k,d}^v$ represents the d^{th} score of a venue v in category k and each row $\mathbf{s}_k^v$ is the score of venue v in category k given by Eq. 7. Therefore, the proposed mapping depicted in Figure 2 ($Scenes \rightarrow$ Yelp $\rightarrow$ GP) is based on knowledge transfer, through the score vector of venue v in category k ($\mathbf{s}_k^v$), for all v, as described by Eqs. 3 to 7.

To illustrate the final mapping results for Google Places data, Table 2 presents an example with sentences associated with two distinct venues retrieved from the GP API, and their associated scores $\mathbf{s}_k = (s_{k1}, ..., s_{kD})$.

Table 2. Examples of two sentences $\text{Sent}_k^v(\text{GP})$ mapped to the dimensions of Scenes Theory $\mathbf{s}_k^v(\text{GP})$

	Skin Care Store	*Hot Dogs Restaurant Food*
Theatricality		
Glamour	4	1
Neighborliness	4	1.8
Transgression	3	3
Formality	3	2.6
Exhibitionism	3	2.8
Authenticity		
Locality	3	1
Ethnicity	3	3
State	3	3
Corporateness	3	4.75
Rationality	2	3
Legitimacy		
Tradition	3	3
Charisma	4	2.6
Utilitarian	2	4.8
Egalitarian	3	3.4
Self-Expression	4	2.4

4.3 Validation of the Mapping Process

The validation of the mapping process described in the previous section uses data from Toronto. This city has been chosen because it is the only city in the literature with an existing Scenes mapping – based on data from other databases. Using Toronto allows us to validate our mapping process, which uses data from Google Places. The city is divided into regions known as Forward Sortation Areas (FSAs), geographic units defined by the first three characters of Canadian postal codes, totaling 99 regions. Each region is treated as a "scene"

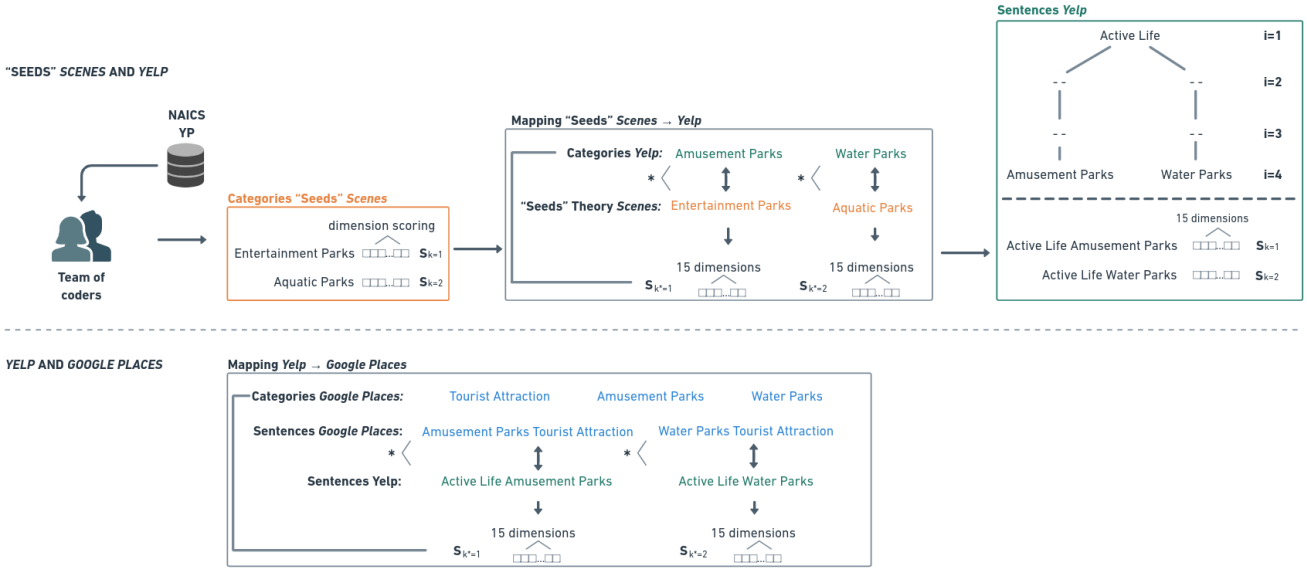

Figure 2. Process of mapping the categories to the Scenes Theory: 1) mapping of the "seeds" Scenes to Yelp, and 2) from the base Yelp to Google Places.

and is mapped to the 15 dimensions through the cultural signature created from the extracted data. Next, Pearson and Spearman correlation coefficients are calculated between the dimension values obtained in this study and those from a pre-existing mapping for these regions (NAICS and YP), as documented in the literature by Silver and Clark [2016]. Correlations are also calculated using Yelp data obtained by Silva and Silver [2025]. These sources provide reliable inputs for analyzing FSA regions and have been validated as trustworthy. Additionally, Silver and Clark [2016] work with these geographic units rather than larger entities like states or municipalities, as FSAs are sufficiently small and offer a high level of precision, with thousands of available categories for classification.

The Pearson correlation coefficient measures the linear relationship between two variables with a normal distribution, and a positive linear relationship is expected between the data from Google and the other databases. Additionally, analyzing a non-parametric classification statistic like Spearman, which evaluates the relationship between two variables described by an arbitrary monotonic function, is also relevant. This approach is justified since the dimensions can exhibit different behaviors, and the databases may not consistently present the same categories across each region. Therefore, using both correlation methods is appropriate [Hauke and Kossowski, 2011]. The results are shown in Figure 3.

The figure reveals that, except for "Tradition" and "Egalitarian," all other dimensions result in positive correlations across the three databases, particularly with YP, which shows overall strong results. Upon examining the mapped sentences to investigate the weaker and negative correlations, a few incoherent mappings related to the "Arts & Crafts" category could be identified. However, their limited number allowed for manual correction. Based on this analysis and the correlation results obtained with the YP database, we conclude that the mapping process is valid and allows the creation of reliable cultural signatures using Google Places.

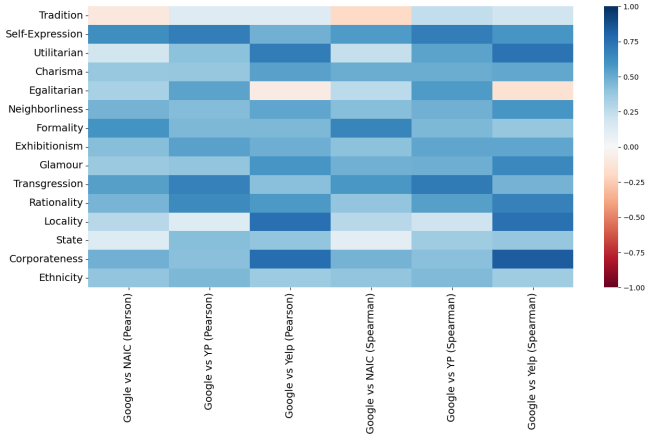

Figure 3. Validation in Toronto: results of Pearson (first three columns) and Spearman (last three columns) correlations, calculated between data from Google and the YP, NAICS, and Yelp databases.

5 Results

This section presents the findings from experiments on cultural signatures, considering different granularities in segmenting the city's area. First, a neighborhood-based approach is applied to Curitiba. Next, various granularity levels for hexagon grids in Curitiba and Chicago are analyzed. Finally, granularity level 7 is examined in depth regarding the characteristics of both cities, first separately and then by combining their datasets.

5.1 Neighborhood-level Experiment

In a practical application of our methodology, Google Places data are collected for Curitiba. This city has been chosen because it is the one the authors are most familiar with, allowing for a more thorough validation of the results. The resulting dataset includes 31,539 venues and 748 unique categories, with data retrieved using five geographic coordinates (subareas) to cover the area of the city effectively.

An analysis based on signatures clustering is performed for Curitiba at the neighborhood level, excluding 11 of the 75 neighborhoods with fewer than 100 venues, as their lack of

information and possibly low diversity could impact the calculation of cultural signatures. This left 64 neighborhoods, whose cultural signatures $\mathbf{c}$ are calculated based on Eq. 1, with $V \geq 100$ and clustered through the hierarchical agglomerative clustering, to be analyzed using $D = 15$ dimensions of the Scenes Theory as features. This results in the dendrogram shown in Figure 4.

Ward's method with Euclidean distance as the chosen metric merges the clusters. The number of clusters is defined by cutting at the second-largest distance, as using only two clusters (the largest distance) would be less meaningful in this context. This approach results in four clusters, which are analyzed and interpreted in what follows.

Cluster 1 (purple in Figure 4) is the largest, encompassing 31 neighborhoods. The most prominent among them are Bairro Novo, Boqueirão, Pinheirinho, and Tatuquara; it is further away from the Center and more concentrated in the city's south. Most of the neighborhoods in these regions are characterized by museums, parks, squares, and tree-lined streets, as well as nightlife attractions, such as bars and clubs. Cluster 2 (blue in Figure 4) has 19 neighborhoods with a predominance of Matriz, characterized by being the commercial center, with areas that lead the city's economic indexes. Its greatest representation is in the retail and service sectors, such as food, beverage, office, and administrative support. In turn, cluster 3 (red in Figure 4) has 11 neighborhoods, which are located on the outskirts of the Center. These are areas with good commercial and leisure infrastructure, in addition to parks with extensive green areas. The study of Viezzer *et al*. [2022] describes the green areas of Curitiba, reinforcing its existence in the southern portion and in greater quantity, but smaller size, in the central and northern portions, aligning with some of the characteristics that we consider to be relevant in Cluster 1 and 3. Finally, cluster 4 (green in Figure 4) has only 3 neighborhoods, namely: Abranches, Alto da Glória and São Francisco. Regarding geographic location, Alto da Glória and São Francisco are close to the Center, while Abranches is a little further away, but in the city's northern region as well. São Francisco has peculiar characteristics, known for being the "coolest" neighborhood in Curitiba, full of bars, casual pubs with rock shows, hamburgers, Arabic restaurants and a market Sunday called Feira do Largo da Ordem, with stalls selling street food and handicrafts. Alto da Glória, located nearby, may share some characteristics with São Francisco and is home to Couto Pereira Stadium. Abranches features Ópera de Arame, known for music and theater events, along with Pedreira Paulo Leminski, which hosts performances by prominent national and international artists.

Upon examining the 15 dimensions of the Scenes Theory across clusters, several parallels can be observed with the previous descriptive analysis, as depicted in Figure 5.

Briefly, cluster 2 exhibits high values in the Tradition and Corporateness dimensions. Cluster 1, by contrast, is prominent in Utilitarian, Transgression, Rationality, and Corporateness but has one of the lowest values in Formality. Cluster 4 stands out with the highest scores in Tradition, Self-Expression, Charisma, Neighborliness, Formality, Glamour, Locality, and Ethnicity, while it scores lowest in Utilitarian, Egalitarian, Transgression, Rationality, and Corporateness. Cluster 3 is notable for high values in Egalitarian and Exhibitionism. Figure 6 summarizes the cluster characteristics and key dimensions.

5.2 Analysis of Granularity Levels Influence

Working with pre-established city divisions, such as neighborhoods, can limit comparative analyses between cities in different countries. For example, expanding analyses from Curitiba to a U.S. city may encounter issues, as not all cities in that country are divided into neighborhoods; instead, some use Census Tracts or ZIP codes, which vary widely in area size. Using a standardized size for city divisions allows for fairer comparisons. It can yield more valuable insights by maintaining independence from existing divisions that may not account for cultural aspects.

To achieve a flexible division for comparative analyses, we apply a hexagonal grid system with three granularity levels—6, 7, and 8—where higher values yield finer resolutions and smaller hexagon sizes. The tool H3-Cities[6] aids in specifying city boundaries and desired granularity. For example, in Curitiba, level 6 typically covers areas larger than neighborhoods, while levels 7 and 8 segment the city into smaller areas. This multi-level approach is applied to Curitiba and Chicago to examine its effectiveness, with Chicago's dataset comprising 55,063 venues, 839 unique categories, and 9 geographic coordinates for comprehensive area coverage. These values reflect nearly double the number of venues and geographic coordinates compared to those in Curitiba, highlighting a significantly broader dataset and bigger area for Chicago.

Cultural signatures are calculated for each hexagon in a particular grid level, providing distinct cultural profiles at the various granularity levels. The Euclidean distance between all pairs of hexagons in the same grid level helps assess the (dis)similarity of cultural characteristics within each city.

The method enables analysis across different granularities, with the cumulative distribution function (CDF) results presented in Figure 7. These findings illustrate how cultural characteristics vary within cities and granularity levels, providing comparative insights into the cultural landscapes of Curitiba and Chicago. Overall, the results are similar for both cities. It is noteworthy that at granularity level 6, there is an excessive lack of diversity, which is confirmed by the values close to zero of the distances between the hexagons. At granularity level 8, the distances between the hexagons increase significantly, which may indicate little information in the creation of their cultural signatures. Finally, level 7 shows a good compromise in this regard.

Based on the previous analysis and given the extensive area covered by granularity level 6, which sometimes leads to minimal division within a city, the analysis proceeds solely with finer levels (7 and 8). To gain insights into the richness of captured information, the number of unique categories per hexagon at each level is calculated, including the enriched categories added to Google Places from the Yelp database. Figure 8 presents these results, showing category diversity across different granularity levels for both Curitiba

[6]https://h3-cities.streamlit.app/

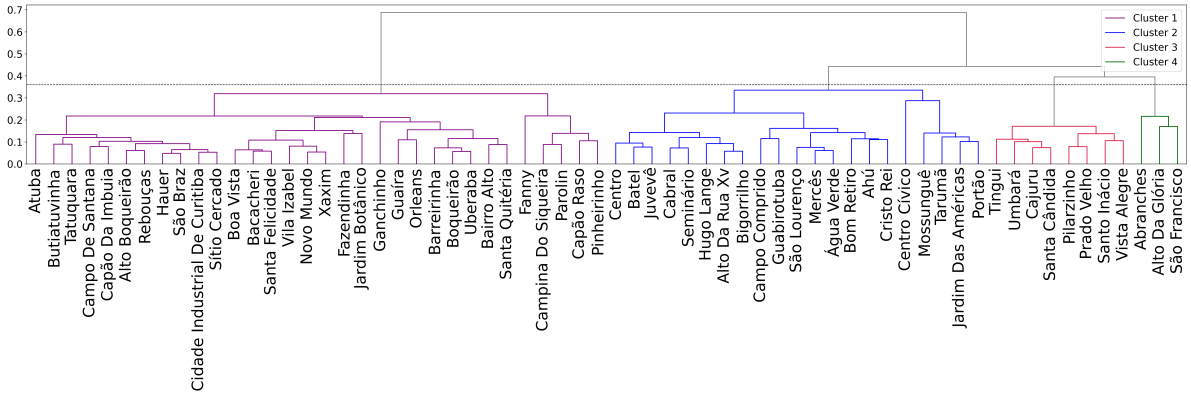

Figure 4. Dendrogram of the Agglomerative Clustering: neighborhood clusters for Curitiba.

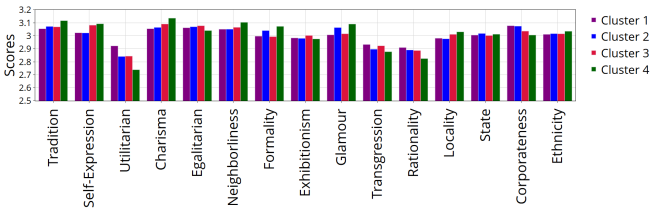

Figure 5. Dimension values per neighborhood signature cluster of the city of Curitiba.

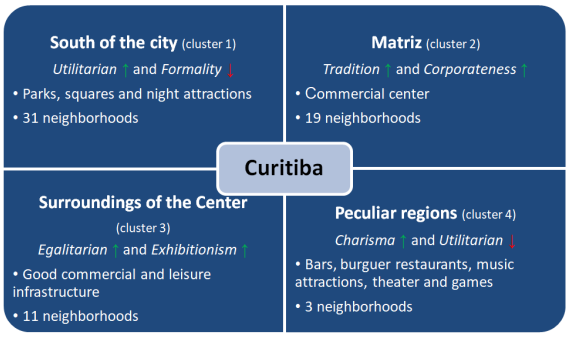

Figure 6. Summary of cluster characteristics in the city of Curitiba.

and Chicago.

According to Figure 8, the CDF of granularity level 7 in both cities shows a slower growth, indicating a higher number of hexagons with distinct categories. For instance, at level 8 in Curitiba, nearly 90% of hexagons contain up to 100 distinct categories (around 80% in Chicago), while at level 7, this proportion drops to about 30% for Curitiba (and 15% for Chicago). This indicates that level 7 captures a broader category diversity, enhancing the representation of an area's cultural profile.

Overall, the results for Curitiba and Chicago display similar patterns for both analyses (signature diversity and number of distinct categories among hexagons), so a focused analysis can be performed to assess clustering quality in Curitiba

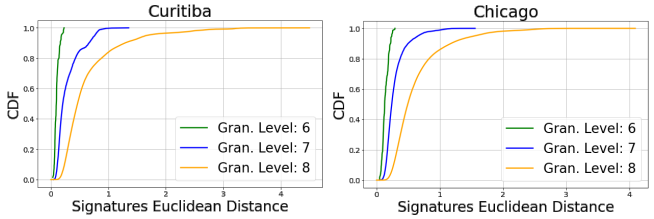

Figure 7. Comparing Euclidean distance of cultural signatures for all pairs of hexagons at each granularity level in Curitiba and Chicago.

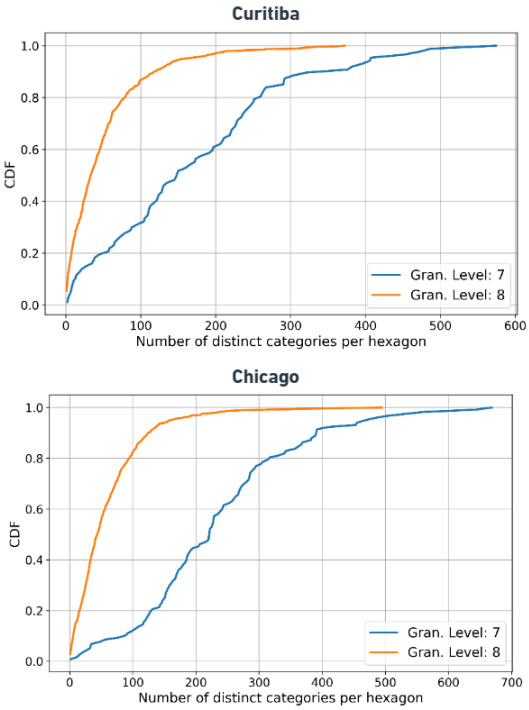

Figure 8. Number of different categories per hexagon, considering granularity levels 7 and 8, for Curitiba and Chicago.

for granularity levels 7 and 8, as shown in Figure 9. Given the importance of having a minimum number of venues per area to accurately capture cultural characteristics ($V \geq \xi$, in Eq. 1), yet lacking a strict threshold (ξ), three experiments were conducted: the first use all hexagons ($\xi = 0$), the second consider ($\xi = 25$) included only hexagons with more than 25 venues, and the third ($\xi = 50$), use those with at least 50. For each experiment, clustering quality is evaluated with the Calinski-Harabasz, Davies-Bouldin, and Silhouette Index metrics, considering 2 to 15 clusters. Hierarchical Agglomerative Clustering is applied using Ward's method and Euclidean distance, with the 15 Scenes Theory dimensions as features.

In evaluating clustering quality, we use three metrics, each emphasizing different aspects of cluster structure: The **Calinski-Harabasz index** prioritizes well-separated clusters, with higher values indicating better-defined clusters. The **Davies-Bouldin index** favors compact and distinct clusters, where lower values correspond to improved quality. The **Silhouette Index** (ranging from -1 to 1) assesses how

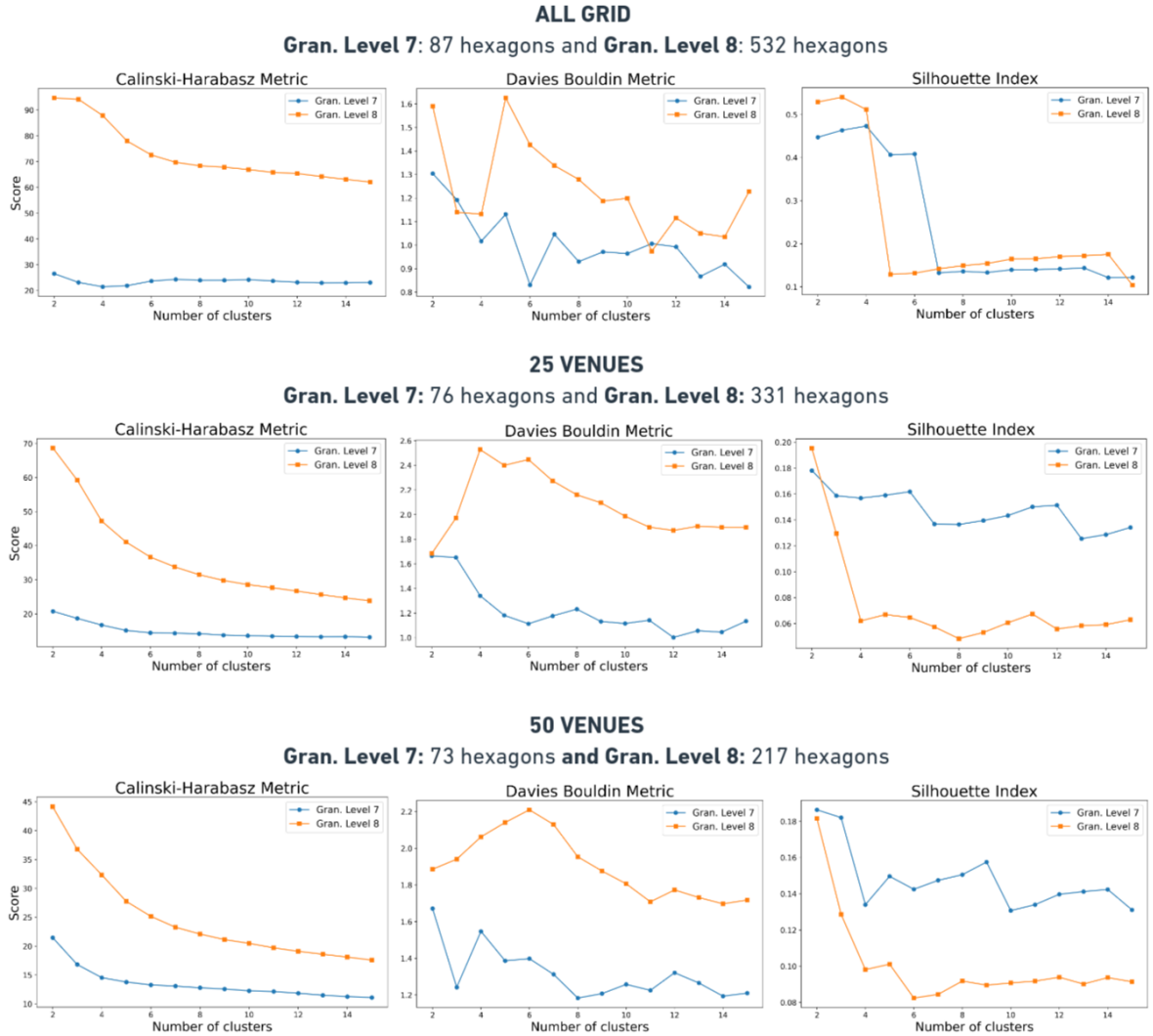

Figure 9. Calculation of metrics to evaluate hexagon signature clustering considering ganularity levels 7 and 8 for Curitiba.

appropriately individual points are grouped: values near -1 suggest poorly defined clusters, values around 0 indicate points near cluster boundaries, and values close to 1 imply well-separated and cohesive clusters. Initial results from the experiment including all hexagons ($\xi = 0$) were inconclusive, primarily due to overlapping outcomes in the Davies-Bouldin and Silhouette metrics. This likely stems from hexagons with limited venue data, which can introduce noise. Consequently, we focus on experiments with hexagons meeting minimum venue thresholds ($\xi = 25$ and $\xi = 50$). The metrics exhibited some apparent contradictions. The Calinski-Harabasz suggests that granularity level 8 provided better cluster separation. However, the Davies-Bouldin and Silhouette metrics favor level 7, emphasizing compactness and point-level cohesion. Given that the latter two metrics focus on intra-cluster compactness (Silhouette) and inter-cluster distinctness (Davies-Bouldin), we conclude that level 7 offers a more robust and balanced assessment of clustering quality. This granularity level was therefore selected for further analysis.

5.3 Deep Diving Into Hexagon Grid Level 7

We conducted analyses in Curitiba and Chicago using granularity level 7 and filtering for hexagons with $\xi = 50$. The choice of a 50-venue threshold reduces the number of hexagons relative to a 25-venue threshold. Still, it preserves broad coverage of key areas while focusing on hexagons that offer rich, informative data about each city's cultural landscape.

With the cultural signatures $\mathbf{c}$ established for each hexagon, Hierarchical Clustering groups those signatures for both cities, using the Ward linkage method and Euclidean distance. Differently from the analysis performed in Figure 4 which is based on neighborhood signature clusters, here, the number of hexagon signature clusters is selected based on the metrics from the prior analysis (Figure 9 with $\xi = 50$ for Curitiba and replicating the same calculation and analysis for Chicago): by choosing 5 clusters for Curitiba and 4 for Chicago we balance the Davies-Bouldin and Silhouette Index metrics as an attempt to achieve a meaningful clustering structure for each city.

The result of clustering signatures using hexagons in Curitiba is illustrated in Figure 10. Notice that three large clusters are apparent, similar to the neighborhood-approach find-

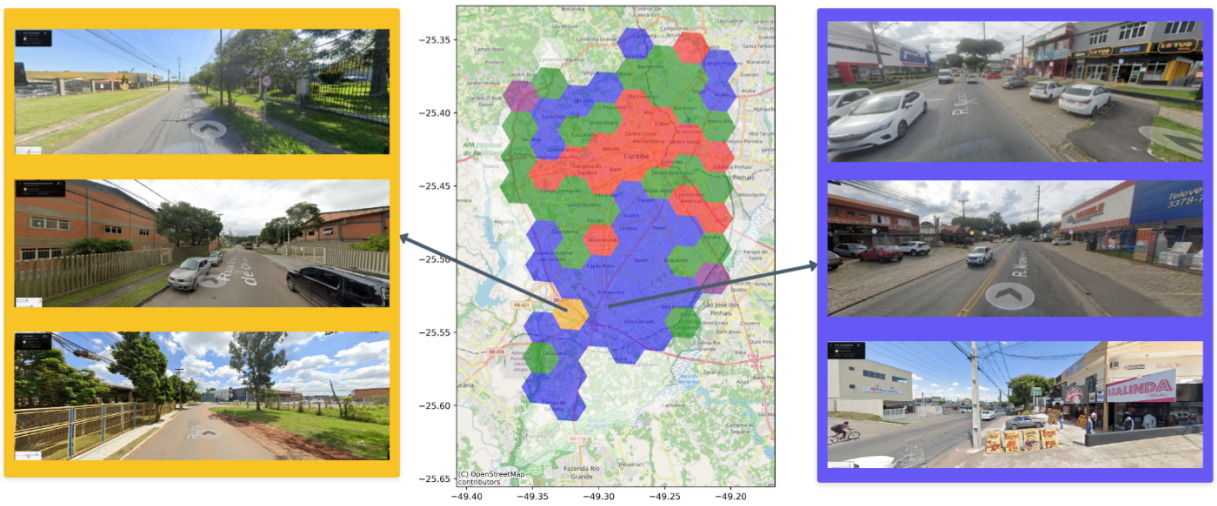

Figure 10. Clustering of Curitiba with images for the yellow cluster and its surroundings (blue cluster).

ings: the city center (red), the southern area (blue), and the area surrounding the center (green). The purple cluster, although joining distant hexagons, reflects similar characteristics shaped by European immigration and is represented by landmarks such as Italiano, São José, and Náutico parks. Italiano Park commemorates the contributions of Italian immigrants who arrived in Paraná at the end of the 19th century, while São José Park is linked to Polish immigration, which played a significant role in the rural development of the area. Although Náutico Park is not directly associated with a single immigrant group, it reflects the broader cultural diversity of the region, with German, Polish, Ukrainian, and Italian influences. European influence in these areas helped shape the development around these parks, impacting public spaces, urban planning, and recreational facilities. The study of Mascarenhas Rocha [2023] helps substantiate this argument as it also shows how the urban design of Curitiba is influenced by immigration, in this case, Polish and Germanic, leaving cultural traces in different parts of the city.

To gain deeper insights into these results, the Z-Score is calculated for the values of the Scenes Theory dimensions, as illustrated in Figure 11. The Z-Score represents the number of standard deviations from the citywide average, where the city average is defined as the centroid of all cluster centroids. This approach aids in comparing clusters by highlighting characteristics that are distinct within each cluster relative to the citywide overview. For instance, Cluster 5 (the purple cluster) displays notably high values in the Tradition, Egalitarian and Ethnicity dimensions, underscoring unique cultural attributes in this area. Cluster 3 (yellow) is represented by a single hexagon. According to the Z-score analysis (Figure 11), this cluster stands out in 5 of the 15 dimensions compared to the other clusters: Utilitarianism, Formality, Rationality, State, and Corporateness. Additionally, it exhibits other peculiarities, such as extremely low values for Self-Expression, Charisma, Neighborliness, Exhibitionism, and Locality—dimensions where other clusters are closer to the average. The cultural signature of this cluster aligns with the area it represents: the southern part of the Cidade Industrial (CIC) neighborhood in Curitiba.

To further understand this region and its differences from surrounding areas, Street View provides detailed images from the area, as shown in Figure 10. The yellow hexagon region has a significant concentration of industries and logistics-focused companies, owing to its strategic location near important highways. This makes it primarily dedicated to production and distribution, contrasting with other parts of the city that are more residential or commercial. Also, the infrastructure in the southern part of CIC is more oriented toward the industrial sector, with fewer leisure options and public spaces. This indicates the proposed approach's potential to identify cultural signatures and provide a comprehensive overview of geographic areas by extracting their key dimensions.

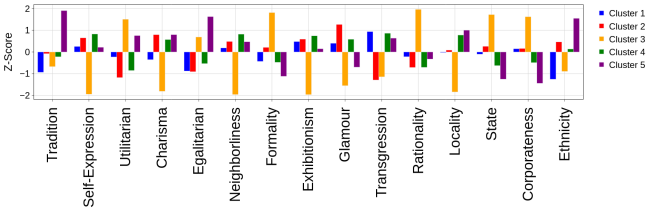

Figure 11. Z-Score values of Scenes dimensions per cluster of Curitiba.

The divergent clusters observed in the Neighborhood- *vs.* Hexagon-based approach analyses for the city of Curitiba can be attributed to the greater granularity and precision of hexagons at granularity level 7 associated with a threshold $\xi = 50$, compared to neighborhoods. Specifically, analyzing smaller areas containing a minimum of venue information allows identifying cultural variations within the same neighborhood that may not be apparent when treating the entire area as a single unit.

In Chicago, the result on the map in Figure 12 shows the general pattern of the city. The red cluster spans most of the lakefront on the North side of the city, as well as areas in the downtown Loop, and some of the South Loop. Overall, these are higher-income areas, often with newer condominium developments, thriving restaurant scenes, and rich nightlife. The Austin neighborhood has one of the highest incomes in the city, represented here by the two red hexagons on the west

side. The south side and west loop, mostly denoted by the yellow cluster, are home to Chicago's African-American communities. The blue cluster (considering the south and west part of the city) is quite diverse, with working-class communities that mix African-Americans, Latinos, East Asians, mixing middle-class neighborhoods and areas with high rates of poverty. The north of the city shows more internal variation, perhaps due to the diverse ethnic immigrant populations there, including Indians, Pakistanis and Vietnamese, compared to the rest of the north region, which is predominantly white, with upper middle-class neighborhoods. Images of some selected points assist in this general interpretation.

By analyzing the cultural signatures of the clusters using the Z-Score values calculated based on the dimensions of Scenes Theory and illustrated in Figure 13, it is possible to observe how the red cluster stands out in dimensions that reflect the previously described characteristics, such as Tradition, Charisma, Exhibitionism, Glamour, and Transgression. This mix of attributes aligns with much recent research charting the distinctive rise of Chicago's "entertainment machine" especially in the downtown loop and northside neighborhoods along the lakefront, which increasingly stress nightlife and entertainment, while, compared to other major cities, remaining tied more closely to heritage and neighborhood traditions (see Clark and Silver [2012]).

The behavior of the green cluster is similar to that of the yellow cluster in Curitiba, as both are the most distinct from the others. The green cluster also shows signs of divergence across several common dimensions, particularly in Self-Expression, Utilitarianism, Formality, and Rationality. Chicago's yellow cluster features Neighborliness, Egalitarianism, Localism, Charisma, and Self-Expression. This mix too reflects what ethnographers have long reported, notably in studies of African-American communities (such as Patillo Pattillo and Lareau [2013]) where church and community life are central, mixed with distinctive fashion and entertainment in which charismatic performers are often prized. The blue cluster stands out for low values in State and Formality, which have often been featured in sociological studies of low-income areas of Chicago as contributing to the emergence of informal economies and street culture, also featuring to some extent in the relatively higher values in Transgression and Exhibition (though not as high as the dense amenity-rich red areas [Venkatesh, 2008; Stuart, 2020].

Chicago is perhaps the most studied city in the history of urban sociology, and it was the laboratory where the Chicago School of Sociology developed in the early 20th century the major perspectives that have defined the field ever since. In one model, social groups occupy areas according to competition processes for access to the central city, resulting in five concentric zones. In this model, Zone I is the main business, commerce, and transportation area, located in the Chicago Loop. Zone II is the area adjacent to Zone I. Zone III comprises working-class residences, neighborhoods such as Bridgeport and Back of the Yards. Zone IV encompasses middle-class residences, including neighborhoods such as Hyde Park and parts of Lincoln Park. Finally, Zone V is the residential suburbs with residents who commute to the city center, such as Evanston and Oak Park [Burgess, 1925, 2015]. While this model never perfectly fits Chicago and its applicability to other cities has often been questioned, it remains a useful starting point [Quinn, 1940; Haggerty, 1971; Beveridge, 2011; Florida, 2013]. Our findings, in line with others highlighting cultural and symbolic areas, do not neatly map onto the concentric model. Even so, some loose aligning with zonal characteristics can be found in our clustering, such as Zone III being represented by the blue and yellow cluster (southwest and south of the city), Zone IV by the red cluster (close to the city center), and Zone V by the red and yellow cluster (north and west of the city). The Chicago Loop and its adjacent area have greater diversity in the divisions of our results.

Chicago School sociologists complemented the concentric zone model with parallel Community Area studies. These assume that urban cultures are segmented by symbolic practices, moral codes, and consumption patterns, which would persist despite competitive pressures [Firey, 1945; Hunter, 1974; Anderson, 2000; Zukin, 2009; Merriman, 2015; Bennett, 2019; Stuart *et al*., 2024]. Our results also reveal this aspect of Chicago. For example, our findings align with research showing that the south of Chicago, in addition to having a strong ethnic identity (Latino, African-American, Polish), has an intense religious life and community values (yellow cluster in the south of the city). By contrast, the north of the city is characterized by upper-middle-class families with a focus on meritocracy, mass consumption, and intense schooling, which can be seen in some ways in the images of the red and yellow cluster (in the north) in Figure 12. This is, of course, a very cursory overview, and for a more precise analysis, including explanations of overlapping clusters in the same geographic area, it would be necessary to consider these and other aspects in more detail.

5.4 Clustering cities together

Finally, to explore the comparison of areas in different cities applying the methodology proposed in this work, Curitiba and Chicago are clustered together, using the same parameter criteria and filters used in the previous section. The result in Figure 14 is consistent with segregating most hexagons according to the city they belong to – yellow clusters in Curitiba and blue and red in Chicago. About the hexagons of different cities that are in the same cluster, four cases are worth further attention. One of them includes the central region of Curitiba, called Matriz, in the blue cluster, which is predominantly formed by Chicago hexagons, demonstrating that the Curitiba Matriz is the region most similar to Chicago, in general. The second case is the yellow cluster, which can be interpreted as the areas in Chicago that most resemble the city of Curitiba. In both cities, these are dispersed and more residential areas, with local commerce and socioeconomic diversity, denoting the population's daily life.

Another case is the red cluster that maintained the same hexagons as the previous cluster in Chicago, but in this cluster, it found 4 hexagons in Curitiba that have similar characteristics. The hexagons of this cluster in Curitiba are mainly characterized by being high-income residential areas, with good infrastructure and quality of life, such as the Jardim Social and São Lourenço neighborhoods [Viezzer *et al*., 2022]. Furthermore, they have tree-lined streets and a range of val-

Figure 12. Clustering of Chicago with images of some points in each cluster.

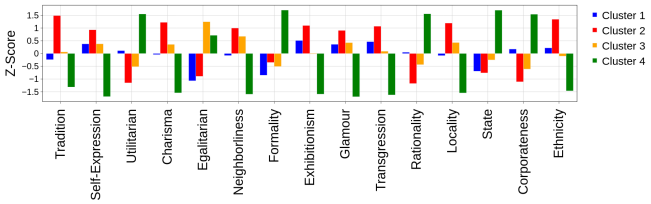

Figure 13. Z-Score values of Scenes dimensions per cluster of Chicago.

ued properties, with a greater predominance of houses and few buildings, unlike the blue neighboring areas, which are denser. An analysis of the Z-Score values of cluster red in Figure 15 reveals a predominance of the Tradition, Self-Expression, Charisma, Neighborliness, Glamour, Locality, and Ethnicity dimensions. This pattern may help explain the similarity with the hexagons in the red cluster in Chicago.

In contrast, the green cluster in Curitiba—previously composed of a single hexagon—expanded to include four hexagons in the current analysis, now incorporating areas from Chicago. Two of these hexagons, already distinct from others, correspond to non-residential and warehouse zones, characterized by highways and limited commercial activity. These features resemble the southern region of CIC in Curitiba and reinforce the same dimensions previously observed there—Utilitarianism, Formality, Rationality, and Corporateness—as shown in Figure 15. This result illustrates the potential to identify latent patterns between areas of different cities that may not be apparent through conventional analyses.

6 Conclusion

Extracting cultural characteristics on a large scale involves several challenges. To address this, the present work proposed a methodology designed to capture key attributes of geographic areas, supporting the generation of cultural signatures and the identification of similarities among them. To ensure scalability, the experiments relied on geolocated web data sources that do not require explicit user actions, thereby addressing limitations commonly associated with user behavior data, which can be difficult to obtain.

The proposed methodology involved merging sentences collected from two different databases (Yelp and Google Places) for each venue and mapping them to a 15-dimensional space using Scenes Theory. A cultural signature was then derived for each region—specifically, divisions in Curitiba and Chicago—by averaging the score vectors of the venues within that region. The size of each region depended on the chosen approach: neighborhoods for Curitiba and a hexagon-based division for both Curitiba and Chicago, with granularity levels of 6, 7, or 8 in the hexagonal approach. In addition to validating the proposed mapping based on previous analyses for Toronto, our work presented experiments that explore various area divisions and their signature clustering to assess the effectiveness of cultural characteristics in Curitiba and Chicago. We also performed a clustering analysis combining both cities to evaluate which areas are similar.

The results yielded interpretations consistent with most of the cultural characteristics of the analyzed regions. The cultural signatures may offer a basis for identifying similarities between areas and could support applications such as place recommendation, validation of service delivery based on cultural criteria, and monitoring the cultural impact of public policies. In particular, cities like Curitiba have been the subject of limited research in the literature regarding their cultural characteristics, and this study contributes to addressing that gap. Although this study focused on two cities to ensure reliable validation of results and methodological soundness—cities with which we are deeply familiar—we acknowledge this as a limitation.

Future work can build upon the methodology presented and extend its application to other cities using Google Places data, further enhancing the validity and generalizability of the findings. Also, mechanisms and experts can be used for cultural analyses, especially in cities in the Global South and the East, to verify whether there is any interference due to only using databases from the Global North (Canada and the USA) in the initial mapping of the Scenes theory. The methodology could also be replicated using data from alternative sources, enabling a more diversified approach to data collection tailored to specific needs. This could also help identify and mitigate potential location-specific biases if they exist.

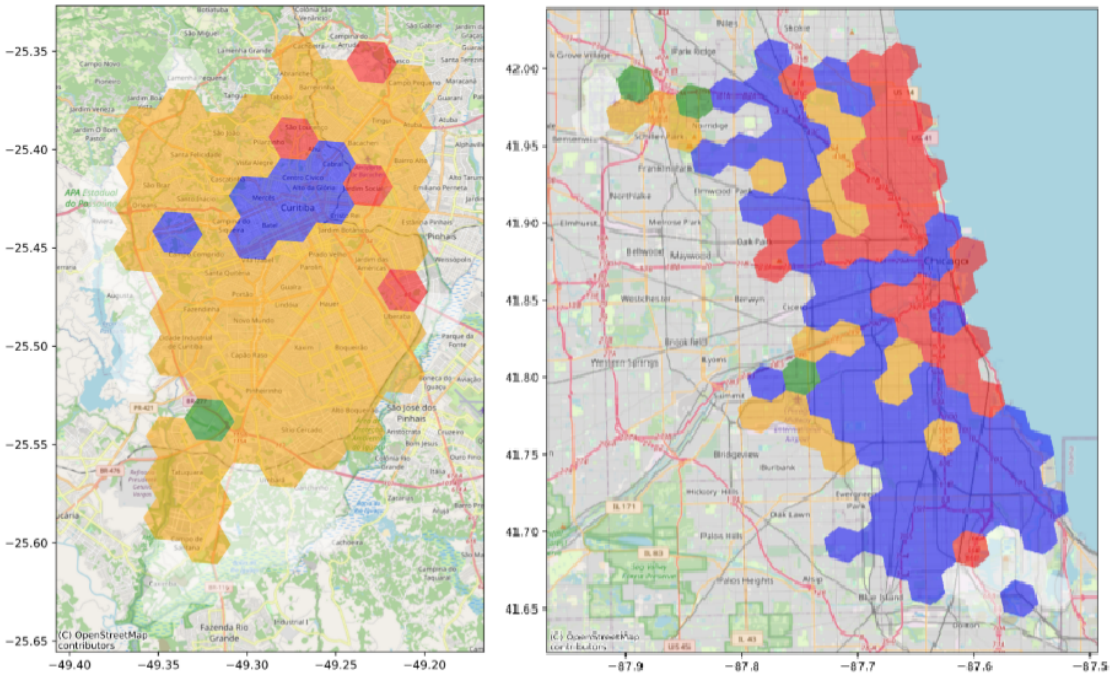

Figure 14. Clustering Cultural Signatures of Curitiba and Chicago together.

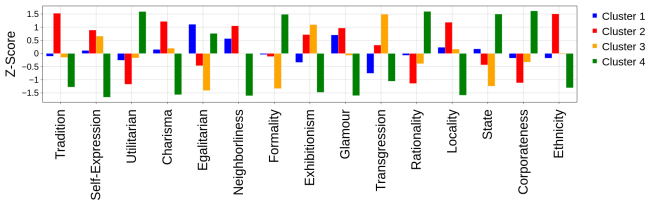

Figure 15. Z-Score values of Scenes dimensions per cluster of Curitiba and Chicago together.

Acknowledgements

This research was partially supported by the SocialNet project (process 2023/00148-0 of the São Paulo Research Foundation - FAPESP), by the National Council for Scientific and Technological Development - CNPq (processes 313122/2023-7, 314603/2023-9, 441444/2023-7, 409669/2024-5, and 444724/2024-9). This research is also part of the INCT ICoNIoT funded by CNPq (proc. 405940/2022-0) and CAPES Finance Code 88887.954253/2024-00.

Competing interests

The authors declare that they have no competing interests.

Authors' Contributions

FG performed the experiments. FG, GS, MD, DS and TS helped in the conceptualization of the study and writing of the manuscript. FG is the main contributor and writer of this manuscript. All authors read and approved the final manuscript.

Availability of data and materials

The tool h3-cities used in the study is available at: *https* : *//h3 − cities.streamlit.app/*. And the tool for data extraction is available at: *https* : *//github.com/FerGubert/google_places_enricher*

References

Anderson, E. (2000). *Code of the street: Decency, violence, and the moral life of the inner city*. WW Norton & Company. Book.

Arribas-Bel, D. and Fleischmann, M. (2022). Spatial signatures-understanding (urban) spaces through form and function. *Habitat International*, 128:102641. DOI: 10.1016/j.habitatint.2022.102641.

Bancilhon, M., Constantinides, M., Bogucka, E. P., Aiello, L. M., and Quercia, D. (2021). Streetonomics: Quantifying culture using street names. *Plos one*, 16(6):e0252869. DOI: 10.1371/journal.pone.0252869.

Bennett, L. (2019). *The third city: Chicago and American urbanism*. University of Chicago Press. Book.

Beveridge, A. A. (2011). Commonalities and contrasts in the development of major united states urban areas: A spatial and temporal analysis from 1910 to 2000. In *Navigating time and space in population studies*, pages 185–216. Springer. DOI: 10.1007/978-94-007-0068-0_8.

Burgess, E. W. (1925). The growth of the city: an introduction. *Chicago, USA: Sociedad Sociológica Mexicana*. Available at: `https://files.eportfolios.macaulay.cuny.edu/wp-content/uploads/sites/6914/2020/01/16233305/Burgess-growth-of-a-city.pdf`.

Burgess, E. W. (2015). The growth of the city: an introduction to a research project. In *The city reader*, pages 212–220. Routledge. Available at: `https://langurbansociology.wordpress.com/wp-`

content/uploads/2013/01/burgess-the-growth-of-the-city.pdf.

Candipan, J., Phillips, N. E., Sampson, R. J., and Small, M. (2021). From residence to movement: The nature of racial segregation in everyday urban mobility. *Urban Studies*, page 0042098020978965. DOI: 10.1177/0042098020978965.

Chen, W., Zhou, Y., Stokes, E. C., and Zhang, X. (2024). Large-scale urban building function mapping by integrating multi-source web-based geospatial data. *Geo-spatial Information Science*, 27(6):1785–1799. DOI: 10.1080/10095020.2023.2264342.

Clark, T. N. and Silver, D. (2012). Chicago from the political machine to the entertainment machine. In *The Politics of Urban Cultural Policy*, pages 28–41. Routledge. Chapter 05. DOI: 10.4324/9780203088777.

De Brito, S. A., Baldykowski, A. L., Miczevski, S. A., and Silva, T. H. (2018). Cheers to untappd! preferences for beer reflect cultural differences around the world. In *Proc. of AMCIS'18*, New Orleans, USA. Available at: https://aisel.aisnet.org/amcis2018/SocialComputing/Presentations/13/.

De La Prada, A. G. and Small, M. L. (2024). How people are exposed to neighborhoods racially different from their own. *Proceedings of the National Academy of Sciences*, 121(28):e2401661121. DOI: 10.1073/pnas.2401661121.

Einola, K. and Alvesson, M. (2021). Behind the numbers: Questioning questionnaires. *Journal of Management Inquiry*, 30(1):102–114. DOI: 10.1177/1056492620938139.

Firey, W. (1945). Sentiment and symbolism as ecological variables. *American Sociological Review*, 10(2):140–148. DOI: 10.2307/2085629.

Florida, R. (2013). The most famous models for how cities grow are wrong. Available at: https://www.bloomberg.com/news/articles/2013-08-09/the-most-famous-models-for-how-cities-grow-are-wrong.

Furno, A., Fiore, M., Stanica, R., Ziemlicki, C., and Smoreda, Z. (2016). A tale of ten cities: Characterizing signatures of mobile traffic in urban areas. *IEEE Transactions on Mobile Computing*, 16(10):2682–2696. DOI: 10.1109/tmc.2016.2637901.

Goffman, E. (1974). *Frame analysis: An essay on the organization of experience.* Harvard University Press. Book.

Gogishvili, D. and Müller, M. (2024). Culture goes east: Mapping the shifting geographies of urban cultural capital through major cultural buildings. *Urban Studies*, page 00420980241289846. DOI: 10.1177/00420980241289846.

Gubert, F., Santos, G., Delgado, M., Silver, D., and Silva, T. (2024a). Criação de assinatura cultural de Áreas urbanas com estabelecimentos geolocalizados na web. pages 127–140. DOI: 10.5753/courb.2024.3260.

Gubert, F. and Silva, T. (2022). Google places enricher: A tool that makes it easy to get and enrich google places api data. In *Anais Estendidos do XXVIII Simpósio Brasileiro de Sistemas Multimídia e Web*, pages 91–94, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/webmedia$_e stendido$.2022.227245.

Gubert, F. R., Santos, G. H., Delgado, M., Silver, D., and Silva, T. H. (2024b). Culture fingerprint: Identification of culturally similar urban areas using google places data. In *International Conference on Advances in Social Networks Analysis and Mining*, pages 286–297. Springer. DOI: 10.1007/978-3-031-78538-2$_2$5.

Haggerty, L. J. (1971). Another look at the burgess hypothesis: Time as an important variable. *American Journal of Sociology*, 76(6):1084–1093. DOI: 10.1086/225034.

Hauke, J. and Kossowski, T. (2011). Comparison of values of pearson's and spearman's correlation coefficients on the same sets of data. *Quaestiones geographicae*, 30(2):87. DOI: 10.2478/v10117-011-0021-1.

Heath, D. B. (1995). *International handbook on alcohol and culture*. Bloomsbury Publishing USA. Book.

Hidalgo, C. A., Castañer, E., and Sevtsuk, A. (2020). The amenity mix of urban neighborhoods. *Habitat International*, 106:102205. DOI: 10.1016/j.habitatint.2020.102205.

Hu, L., Li, Z., and Ye, X. (2020). Delineating and modeling activity space using geotagged social media data. *Cartography and Geographic Information Science*, 47(3):277–288. DOI: 10.1080/15230406.2019.1705187.

Hunter, A. D. (1974). *Symbolic communities: The persistence and change of Chicago's local communities*. University of Chicago Press. Book.

Ilieva, R. T. and McPhearson, T. (2018). Social-media data for urban sustainability. *Nature Sustainability*, 1(10):553–565. DOI: 10.1038/s41893-018-0153-6.

Jaeger, S. R. and Cardello, A. V. (2022). Factors affecting data quality of online questionnaires: Issues and metrics for sensory and consumer research. *Food Quality and Preference*, 102:104676. DOI: 10.1016/j.foodqual.2022.104676.

Laufer, P., Wagner, C., Flöck, F., and Strohmaier, M. (2015). Mining cross-cultural relations from wikipedia: a study of 31 european food cultures. pages 1–10. DOI: 10.48550/arXiv.1411.4484.

Le Falher, G., Gionis, A., and Mathioudakis, M. (2015). Where is the soho of rome? measures and algorithms for finding similar neighborhoods in cities. DOI: 10.1609/icwsm.v9i1.14602.

Martí P., Serrano-Estrada L., N.-C. A. and L., B. J. (2021). Revisiting the spatial definition of neighborhood boundaries: Functional clusters versus administrative neighborhoods. *Journal of Urban Technology*, pages 1–22. DOI: 10.1080/10630732.2021.1930837.

Mascarenhas Rocha, R. (2023). A contribuição das imigrações polonesa e germânica para a formação da cidade de curitiba (pr): bairros e endereços que trazem marcas da imigração. *Idéias*, 14. Available at: https://periodicos.sbu.unicamp.br/ojs/index.php/ideias/article/view/8671164.

Mehta, V. and Mahato, B. (2019). Measuring the robustness of neighbourhood business districts. *Journal of Urban Design*, 24(1):99–118. DOI: 10.1080/13574809.2018.1500137.

Merriman, B. (2015). Three conceptions of spatial locality in chicago school sociology (and their significance

today). *The American Sociologist*, 46:269–287. DOI: 10.31235/osf.io/2khse.

Olson, A. W., Calderón-Figueroa, F., Bidian, O., Silver, D., and Sanner, S. (2021). Reading the city through its neighbourhoods: Deep text embeddings of yelp reviews as a basis for determining similarity and change. *Cities*, 110:103045. DOI: 10.31235/osf.io/8jbvg.

Pattillo, M. and Lareau, A. (2013). *Black Picket Fences: Privilege & Peril among the Black Middle Class*. University of Chicago Press. Book.

Quinn, J. A. (1940). The burgess zonal hypothesis and its critics. *American Sociological Review*, 5(2):210–218. DOI: 10.2307/2083636.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. DOI: 10.18653/v1/d19-1410.

Rivière, F. *et al*. (2009). *Investing in cultural diversity and intercultural dialogue*, volume 2. Unesco. Available at:`https://unesdoc.unesco.org/ark:/48223/pf0000184755`.

Sen, R. and Quercia, D. (2018). World wide spatial capital. *PloS one*, 13(2):e0190346. DOI: 10.1371/journal.pone.0190346.

Senefonte, H., Frizzo, G., Delgado, M., Lüders, R., Silver, D., and Silva, T. (2020). Regional influences on tourists mobility through the lens of social sensing. In *International Conference on Social Informatics*, pages 312–319. Springer. DOI: 10.1007/978-3-030-60975-$7_2$3.

Silva, T. H., de Melo, P. O. V., Almeida, J. M., Musolesi, M., and Loureiro, A. A. (2017). A large-scale study of cultural differences using urban data about eating and drinking preferences. *Information Systems*, 72:95–116. DOI: 10.1016/j.is.2017.10.002.

Silva, T. H. and Silver, D. (2025). Using graph neural networks to predict local culture. *Environment and Planning B: Urban Analytics and City Science*, 52(2):355–376. DOI: 10.1177/23998083241262053.

Silver, D. A. and Clark, T. N. (2016). *Scenescapes: How qualities of place shape social life*. The University of Chicago. Book.

Simmel, G. (1971). On individuality and social forms: Selected writings, ed. *Donald N. Levine. Chicago: UP of Chicago*. Book.

Sparks, K., Thakur, G., Pasarkar, A., and Urban, M. (2020). A global analysis of cities' geosocial temporal signatures for points of interest hours of operation. *International Journal of Geographical Information Science*, 34(4):759–776. DOI: 10.1080/13658816.2019.1615069.

Spencer-Oatey, H. and Franklin, P. (2012). What is culture. *A compilation of quotations. GlobalPAD Core Concepts*, pages 1–22. Available at: `https://warwick.ac.uk/fac/soc/al/globalpad-rip/openhouse/interculturalskills_old/core_concept_compilations/global_pad_-_what_is_culture.pdf`.

Sproesser, G., Ruby, M. B., Arbit, N., Akotia, C. S., dos Santos Alvarenga, M., Bhangaokar, R., Furumitsu, I., Hu, X., Imada, S., Kaptan, G., *et al*. (2022). Similar or different? comparing food cultures with regard to traditional and modern eating across ten countries. *Food Research International*, 157:111106. DOI: 10.1016/j.foodres.2022.111106.

Stuart, F. (2020). Ballad of the bullet: Gangs, drill music, and the power of online infamy. Book.

Stuart, F., Collins, C. R., Wade, B., Gleit, R. D., and Louis Moore, C. (2024). Where do neighbourhood reputations come from? analysing chicago community areas using a systematic neighbourhood reputation score, 1985–2020. *Urban Studies*, page 00420980241297088. DOI: 10.1177/00420980241297088.

Tang, J., Cheng, X., Liu, A., Huang, Q., Zhou, Y., Huang, Z., Liu, Y., and Xu, L. (2024). Inferring "high-frequent" mixed urban functions from telecom traffic. *Environment and Planning B: Urban Analytics and City Science*, 51(8):1775–1793. DOI: 10.1177/23998083231221867.

Venkatesh, S. (2008). *Gang leader for a day: A rogue sociologist takes to the streets*. Penguin. Book.

Viezzer, J., de Moraes, E. N., Biondi, D., Martini, A., and Scarano, F. R. (2022). Áreas verdes, populaÇÃo e renda em curitiba, pr, brasil. *Revista da Sociedade Brasileira de Arborização Urbana*, 17(2):37–49. DOI: 10.5380/revsbau.v17i2.85848.

Weber, M. (1930). *The Protestant Ethic and the Spirit of Capitalism*. New York: Routledge Classics. Available at: `https://gpde.direito.ufmg.br/wp-content/uploads/2019/03/MAX-WEBER.pdf`.

YP (2022). Yellow pages. Available at: `https://www.yellowpages.ca/`.

Zhang, Z., He, Q., Gao, J., and Ni, M. (2018). A deep learning approach for detecting traffic accidents from social media data. *Transportation research part C: emerging technologies*, 86:580–596. DOI: 10.1016/j.trc.2017.11.027.

Zukin, S. (2009). *Naked city: The death and life of authentic urban places*. Oxford University Press. DOI: 10.1093/oso/9780195382853.001.0001.

Çelikten, E., Le Falher, G., and Mathioudakis, M. (2016). Modeling urban behavior by mining geotagged social data. *IEEE Transactions on Big Data*, 3(2):220–233. DOI: 10.1109/tbdata.2016.2628398.