UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

FERNANDA REGINA GUBERT

DATA-DRIVEN DISCOVERY OF CULTURALLY SIMILAR URBAN AREAS USING GOOGLE PLACES

CURITIBA

2025

FERNANDA REGINA GUBERT

DATA-DRIVEN DISCOVERY OF CULTURALLY SIMILAR URBAN AREAS USING GOOGLE PLACES

Descoberta Orientada por Dados de Áreas Urbanas Culturalmente Similares usando Google Places

Dissertation presented in Program Postgraduate in Electrical Engineering and Industrial Informatics of the Federal Technological University of Paraná, as a partial requirement for obtaining the title of Master of Science.

Advisor: Prof. Dr. Thiago Henrique Silva

Co-Advisor: Profª. Drª. Myriam Regattieri De Biase da Silva Delgado

CURITIBA

2025

Ministério da Educação
Universidade Tecnológica Federal do Paraná
Campus Curitiba

FERNANDA REGINA GUBERT

DATA-DRIVEN DISCOVERY OF CULTURALLY SIMILAR URBAN AREAS USING GOOGLE PLACES

Trabalho de pesquisa de mestrado apresentado como requisito para obtenção do título de Mestre Em Ciências da Universidade Tecnológica Federal do Paraná (UTFPR). Área de concentração: Engenharia De Computação.

Data de aprovação: 11 de Abril de 2025

Dr. Thiago Henrique Silva, Doutorado - Universidade Tecnológica Federal do Paraná

Dra. Jussara Marques De Almeida Goncalves, Doutorado - Universidade Federal de Minas Gerais (Ufmg)

Dr. Ricardo Luders, Doutorado - Universidade Tecnológica Federal do Paraná

Documento gerado pelo Sistema Acadêmico da UTFPR a partir dos dados da Ata de Defesa em 11/04/2025.

ACKNOWLEDGEMENTS

I thank all the people who, in some way, were present and took part in this important phase of my life, including family, classmates, and professors. Special consideration goes to my advisor, Prof. Dr. Thiago Henrique Silva, for his wisdom and dedicated time, whose guidance was fundamental to my professional development and the completion of this stage. I also express my sincere gratitude to my co-advisor, Prof. Dr. Myriam Regattieri De Biase da Silva Delgado, for her support, valuable contributions, and dedication throughout this journey. I would also like to thank the professors on the examining board, Ricardo Luders and Jussara Marques De Almeida Goncalves, for agreeing to participate in this important moment in my academic career. Their valuable contributions, insightful observations and encouragement of critical thinking were fundamental to the improvement of this work.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 88887.954253/2024-00, by the SocialNet project of the São Paulo Research Foundation - FAPESP - Finance Code 2023/00148-0, by the National Council for Scientific and Technological Development - CNPq - Finance Codes 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 - Finance Code 405940/2022-0.

Culture is the system of living ideas that each era possesses. Or rather, the system of ideas through which time unfolds" (ORTEGA; GASSET, 1982).

RESUMO

Compreender as características de diversos grupos culturais em todo o mundo e identificar semelhanças culturais entre suas respectivas regiões pode gerar benefícios econômicos e sociais significativos. No entanto, grande parte da pesquisa existente nessa área depende de dados de comportamento do usuário, o que apresenta desafios de escalabilidade e generalização devido à dificuldade de obtenção desses dados. Para abordar essa limitação, nosso trabalho se concentra na extração de dados de estabelecimentos do Google Places (GP) e na introdução de uma metodologia baseada no conceito de Scenes para enriquecer o conjunto de dados do GP, permitindo a geração de assinaturas culturais de áreas urbanas. Propomos e avaliamos um método avançado que aprimora as categorias de locais usando Scenes Theory, o que nos ajuda a compreender o significado cultural da vida urbana cotidiana. Além disso, comparamos a abordagem baseada em Scenes com dois métodos mais simples que usam apenas os tipos de estabelecimentos e sua frequência nas áreas estudadas. Testamos todos os métodos propostos em 14 cidades ao redor do mundo e em todos os estados dos EUA. Nossos resultados indicam que uma abordagem direta baseada em frequências de categorias pode destacar grandes diferenças culturais. No entanto, o método baseado na Scenes Theory oferece uma melhor compreensão das nuances culturais, alinhando-se com os resultados apoiados por dados de pesquisa. Também exploramos o impacto da granularidade variável na geração de assinaturas culturais, analisando três níveis de granularidade com uma partição de grade hexagonal. As análises realizadas destacam os benefícios sociais da nossa abordagem, como recomendações de localização baseadas em critérios culturais e validação de serviços em tempo real.

Palavras-chave: assinatura cultural; análise multiescala; dados geolocalizados; similaridades culturais; google places.

ABSTRACT

Understanding the characteristics of diverse cultural groups worldwide and identifying cultural similarities between their respective regions can yield significant economic and social benefits. However, much of the existing research in this field relies on user behavior data, which poses challenges in scalability and generalization due to the difficulty of obtaining such data. To address this limitation, our work focuses on extracting venue data from Google Places (GP) and introducing a methodology based on the Scenes concept to enrich the GP dataset, enabling the generation of cultural signatures of urban areas. We propose and evaluate an advanced method that enhances venue categories using Scenes Theory, which helps us understand the cultural significance of everyday urban life. Moreover, we compare the Scenes-based approach with two simpler methods that use only venue types and their frequency within the studied areas. We tested all proposed methods in 14 cities worldwide and all US states. Our results indicate that a straightforward approach based on category frequencies can highlight major cultural differences. However, the Scenes Theory-based method offers a better understanding of cultural nuances, aligning with findings supported by survey data. We also explore the impact of varying granularity on the generation of cultural signatures by analyzing three levels of granularity with a hexagonal grid partition. The performed analyses underscore our approach's societal benefits, such as location recommendations based on cultural criteria and real-time service validation.

Keywords: cultural signature; multiscale analysis; geolocated data; cultural similarities; google places.

LIST OF FIGURES

Figure 1 – Representation of the meaning of possible score values in each dimension for each venue category. 18

Figure 2 – Rectangular area delimited for Toronto (part 1). Grid cells with sides of 9,000 meters (red squares) (part 2). Automatically calculated circular areas (in blue) and highlighted city center (red circle) (part 3). 26

Figure 3 – Process of mapping of the "seeds" Scenes to Yelp. 28

Figure 4 – Overview of mapping GP categories to the local cultural dimensions (from the Scenes Theory). 30

Figure 5 – 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. 33

Figure 6 – Hierarchical clustering dendrogram of cities represented by *Scenes*. . . 36

Figure 7 – Z-Score values of Scenes dimensions per cluster. Cluster numbers follow what is presented in Figure 6. 37

Figure 8 – Hierarchical clustering dendrogram of cities represented by *Frequency*. 38

Figure 9 – Z-Score values for the most distinct categories per cluster, considering the clustering of *Frequency*. 39

Figure 10 – Results of hierarchical clustering considering all states in the USA represented by *Scenes*. 41

Figure 11 – Results of hierarchical clustering considering all states in the USA represented by *Frequency*. 42

Figure 12 – Correlation ranges between AVS and *Scenes*. 44

Figure 13 – Comparing Euclidean distance of cultural signatures for all pairs of hexagons at each granularity level in Curitiba and Chicago. 46

Figure 14 – Number of different categories per hexagon, considering granularity levels 7 and 8, for Curitiba and Chicago. 47

Figure 15 – Calculation of metrics to evaluate hexagon signature clustering considering ganularity levels 7 and 8 for Curitiba. 48

Figure 16 – Dendrogram of the Agglomerative Clustering: neighborhood clusters for Curitiba. 49

Figure 17 – Clustering of Curitiba with images for the yellow cluster and its surroundings (blue cluster). **50**
Figure 18 – Z-Score values of Scenes dimensions per cluster of Curitiba. **51**
Figure 19 – Clustering of Chicago with images of some points in each cluster. . . . **52**
Figure 20 – Z-Score values of Scenes dimensions per cluster of Chicago. **53**
Figure 21 – Clustering Cultural Signatures of Curitiba and Chicago together. **54**
Figure 22 – Z-Score values of Scenes dimensions per cluster of Curitiba and Chicago together. **54**

LIST OF TABLES

Table 1 – Number of venues, unique categories, and coordinates used by each city. 27

Table 2 – Examples of two sentences Sent$_k^v$(GP) mapped to the dimensions of Scenes Theory s_k^v(GP) 32

Table 3 – Pearson correlation r (and its p-value) between the Euclidean distance of a particular state vs all others when describing them by AVS *versus Scenes* or *Frequency*. 43

LIST OF ABBREVIATIONS AND ACRONYMS

Pseudo-Acronyms

API	Application Programming Interface
AVS	American Value Survey
BERT	Bidirectional Encoder Representations from Transformers
CIC	Cidade Industrial de Curitiba
EMD	Earth-Mover's Distance
ET	Enclosed Tessellation
GP	Google Places
LBSN	Location-Based Social Network
LSA	Latent Semantic Analisys
PCA	Principal Component Analysis
POI	Point of Interest
SMI	Segregated Mobility Index
USA	United States of America
US	United States

CONTENTS

1 INTRODUCTION . . . 12

1.1 Problem description . . . 12

1.2 Objective . . . 13

1.3 Contributions . . . 13

1.4 Organization . . . 14

2 BACKGROUND AND RELATED WORKS . . . 15

2.1 Scenes Theory . . . 15

2.1.1 Fundamentals: The 15 Dimensions . . . 15

2.1.2 Dimension Scoring System . . . 17

2.2 Related Works . . . 19

2.2.1 Culture and Urban Mobility . . . 19

2.2.2 Cultural Similarities Between Areas . . . 20

2.2.3 Studies with Google Places Data . . . 21

2.2.4 Cultural Signature of Areas . . . 22

2.2.5 Discussion of Related Work . . . 23

3 GENERATING CULTURAL SIGNATURES FROM GOOGLE PLACES . . . 25

3.1 Extracting Data From Google Places . . . 25

3.2 Mapping Categories Into Dimensions . . . 28

3.3 Validation of the Mapping Process . . . 31

3.4 Proposed Cultural Signatures . . . 32

3.4.1 *Scenes-based approach* . . . 33

3.4.2 *Naive* and *Frequency-based approach* . . . 34

4 CULTURAL SIGNATURES TO IDENTIFY CULTURALLY SIMILAR AREAS . 36

4.1 Cities Worldwide . . . 36

4.1.1 *Scenes* for Dataset Cities . . . 36

4.1.2 *Naive* for Dataset Cities . . . 37

4.1.3 *Frequency* for Dataset Cities . . . 37

4.2 All States in the USA . . . 40

4.2.1 Evaluating Scenes-based approach for Dataset States . . . 40

4.2.2 Evaluating *Naive* and Frequency-based approaches for Dataset States . . . 40

4.3 Comparing with Survey Data 40
5 ANALYZING CULTURALLY SIMILAR AREAS AT DIFFERENT GRANULARITY LEVELS 45
5.1 Exploring the influence of granularity levels 45
5.2 Deep diving into grid level 7 with Curitiba and Chicago 48
5.3 Clustering cities together 53
6 CONCLUSION 55
REFERENCES 56

1 INTRODUCTION

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; ALVESSON, 2021; JAEGER; CARDELLO, 2022). To work around this limitation, many recent studies resort to data from web sources to address challenges across various fields (ILIEVA; MCPHEARSON, 2018; ZHANG *et al.*, 2018; HU; LI; YE, 2020; CHEN *et al.*, 2024), producing meaningful results more efficiently.

According to the report of UNESCO (United Nations Educational, Scientific, and Cultural Organization) produced 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.

Identifying cultural similarities and being able to track 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. This type of 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, this study opens the door to developing new tools 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.

1.1 Problem description

The concept of culture is complex and lacks a single definition, making the task of finding data that satisfactorily describe 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; 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; 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

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

of Mehta and Mahato (2019), in which the availability of resources and services that meet the population's needs is a way of providing 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.

1.2 Objective

The main goal of this study is to propose and evaluate three approaches to measuring local culture that rely only on basic information on urban venues, their location, and their categories and then use the best approach to support an area-division strategy for analyzing cities across different countries.

To accomplish this goal, this work:

- Considers Google Places (GP) as a potential data source. GP is interesting because it provides the demanded information and covers a vast portion of the globe[2].
- Evaluates different approaches for obtaining cultural signatures of urban areas with different levels of complexity.
- Studies the impact of the area size in generating cultural signatures.

1.3 Contributions

The main contributions of this work are:

- Proposition of three approaches for obtaining cultural signatures of urban areas using GP venue categories. The first approach only considers the presence/absence of each GP venue category available in the area (Naive-based approach). The second approach incorporates venue frequency (Frequency-based approach). The third approach enhances GP data using the Scenes concept (Scenes-based approach), transforming the everyday "scenes" of venues in a given urban area into elements of cultural significance. This method assigns weights to these elements based on venue type (category) and integrates category frequencies into the computation (SILVER; CLARK, 2016). The Scenes concept enables the generation of a more expressive cultural abstraction of any urban area where GP data is available.
- Evaluation of the proposed approaches was conducted using data from 14 cities across different continents and all U.S. states. The results indicate that a simple approach

[2] https://developers.google.com/maps/coverage.

(*Frequency*) can satisfactorily capture significant cultural differences. However, a more sophisticated approach (*Scenes*) enhances semantic expressiveness in representing cultural characteristics. This added expressiveness is evident from our comparison of outcomes with survey data, which suggests that *Scenes* better captures cultural nuances. These findings highlight promising alternative methods for automatically identifying culturally similar areas without relying on hard-to-obtain data.

- Analysis of the impact of variable granularity in urban areas on the generation of cultural signatures. The findings from this study highlight important implications for different levels of spatial detail and suggest the most appropriate granularity level for accurate cultural representation.

Based on these findings, we demonstrate a potential application of the proposed approach by identifying similar areas across different cities. This highlights the approach's potential for new applications, such as area recommendation systems based on cultural criteria.

The present work has also directly contributed to two publications:

- VIII Workshop em Computação Urbana (CoUrb 2024) (GUBERT *et al.*, 2024). An expanded version of this work was invited for submission to the Journal of Internet Services and Applications and is currently under revision (second round).
- 16th International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2024) (GUBERT *et al.*, 2024).

In addition to two other publications with indirect results:

- VIII Workshop em Computação Urbana (CoUrb 2024) (SANTOS *et al.*, 2024).
- Journal of Internet Services and Applications (JISA Vol.16 No.1 2025) (SANTOS *et al.*, 2025).

1.4 Organization

The rest of the study is organized as follows. Chapter 2 describes the foundations of the Scenes Theory and how its different dimensions are applied to obtain the cultural signature, then presents related works. Chapter 3 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. In addition to the description of the cultural signature approaches evaluated in this work. Chapter 4 presents and discusses the results regarding identifying culturally similar areas, in addition to presenting comparisons with survey data. Chapter 5 presents the results for Curitiba and Chicago using different granularity levels in urban area partitioning. Finally, Chapter 6 presents the conclusion and directions for future work.

2 BACKGROUND AND RELATED WORKS

This chapter describes Scenes Theory, including its 15 dimensions, the details of its scoring systems, and the possible cultural signature that results from it. Next we present the related works, first we 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 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.

2.1.1 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; 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 rein-

forcing; 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; CLARK, 2016; GUBERT *et al.*, 2024; SILVA; SILVER, 2025), showing its usefulness in practice.

2.1.2 Dimension Scoring System

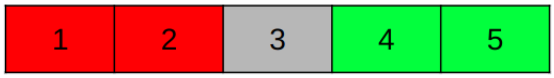
To translate seemingly non-cultural data into sources of information about cultural significance, in Silver and Clark (2016) the authors worked with a team of coders who helped assign weights to the dimensions of all types of venues present in their database, from NAICS (North American Industrial Classification 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.[1], Canada[2] and Mexico[3]. 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

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

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 category receives a score from 1 to 5 on each of the 15 dimensions, see Figure 1. Scores are not necessarily whole numbers. Scores of 4 to 5 indicate that the venue affirms the dimension, while scores of 1 to 2 suggest rejection. A score around 3 signifies a neutral stance toward the dimension. The most critical decision lies in assigning a positive (4 to 5) or negative (1 to 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 around 4 or 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.



Glamour	**Locality**	**Tradition**
Neighborliness	**Ethnicity**	**Charisma**
Transgression	**State**	**Utilitarian**
Formality	**Corporateness**	**Egalitarian**
Exhibitionism	**Rationality**	**Self-Expression**

Figure 1 – Representation of the meaning of possible score values in each dimension for each venue category.

The databases NAICS and YP, enriched with dimension scores called seeds 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.

2.2 Related Works

2.2.1 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 – (BRITO *et al.*, 2018; SILVA *et al.*, 2019; SILVA; SILVER, 2024). 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, 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) investigate 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) leverage aggregated

and anonymized telecom traffic data to infer urban functions from urban land use. Their study was conducted in Shenzhen, China, and combines time series decomposition and urban texture analysis to map functions such as housing, work and recreation, and even identifies 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.

2.2.2 Cultural Similarities Between Areas

The present work aims to conduct a comparative analysis of geographic areas to identify cultural similarities. The approach proposed by Falher, Gionis and Mathioudakis (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, Falher and Mathioudakis (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.*, 2024), 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.

2.2.3 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, Castañer and Sevtsuk (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 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 *et al.* (2021) have carried 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.

2.2.4 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, 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 (2024) 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.

2.2.5 Discussion of Related Work

Studies such as Senefonte *et al.* (2020) and 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 of user behavior, 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 *et al.* (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 by Silva *et al.* (2017), 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 works, such as that conducted by Silva and Silver (2024), which applies machine learning techniques to predict urban culture. Our study does not envision performing predictions.

In our previous study (GUBERT *et al.*, 2024), we characterize 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.*, 2024), 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 our previous work (GUBERT *et al.*, 2024), 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) a demonstration of how the results can be applied to identify similar areas across cities.

3 GENERATING CULTURAL SIGNATURES FROM GOOGLE PLACES

This chapter outlines the procedures for cultural extraction used in the study, structured into four sections. The first details the retrieval of venue data from the GP API for a given city (Toronto is used as an initial example). The second section describes the mapping of GP categories to the $D = 15$ dimensions of Scenes Theory, leveraging existing mappings from the Scenes dataset $\{\mathrm{s}_k(\text{Scenes})\}$ and categories in Yelp $\{\text{categ(Yelp)}\}$ to enhance semantic accuracy. The third section explains the validation method, which employs pre-existing datasets for Toronto, Canada, and evaluates results using Pearson and Spearman correlation analyses. Finally, the fourth section presents the approaches to creating a cultural signature.

3.1 Extracting Data From Google Places

We chose to use Google Places as it provides the most comprehensive and reliable dataset for location information worldwide. However, it is important to clarify that we also considered open-source alternatives such as OpenStreetMap. In this case, 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 GP API is a location-based social network that allows users to discover and share information about local venues, such as universities, cafes, and parks. Users can engage with the platform by posting reviews, uploading photos, and giving ratings. GP API provides geolocated venue data, resulting in one of the world's most accurate, up-to-date, and comprehensive venue models. In addition to latitude and longitude coordinates, venues are associated with at least one category designed to describe the venue type. In this study, we consider two datasets from GP, States and Cities, as described next.

The Dataset States has been provided by the authors Li, Shang and McAuley (2022) and Yan *et al.* (2023). It contains business metadata (geographic info, category information, and others) from GP up to Sep 2021 in all states of the United States. In this study, we focus on the geographic info and category information. It is composed of 4,963,111 unique venues and has 4,501 unique categories. We explore this dataset to study states. We have data for every state. The District of Columbia has the lowest number of distinct venues, totaling 11,003, while California has the highest count at 513,134 unique venues.

For Dataset Cities we have collected data from a set of cities. GP 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 GP 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 $600 \leq |\{\text{categ(GP)}\}| \leq 888$, which is almost 4 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 2 for Toronto.

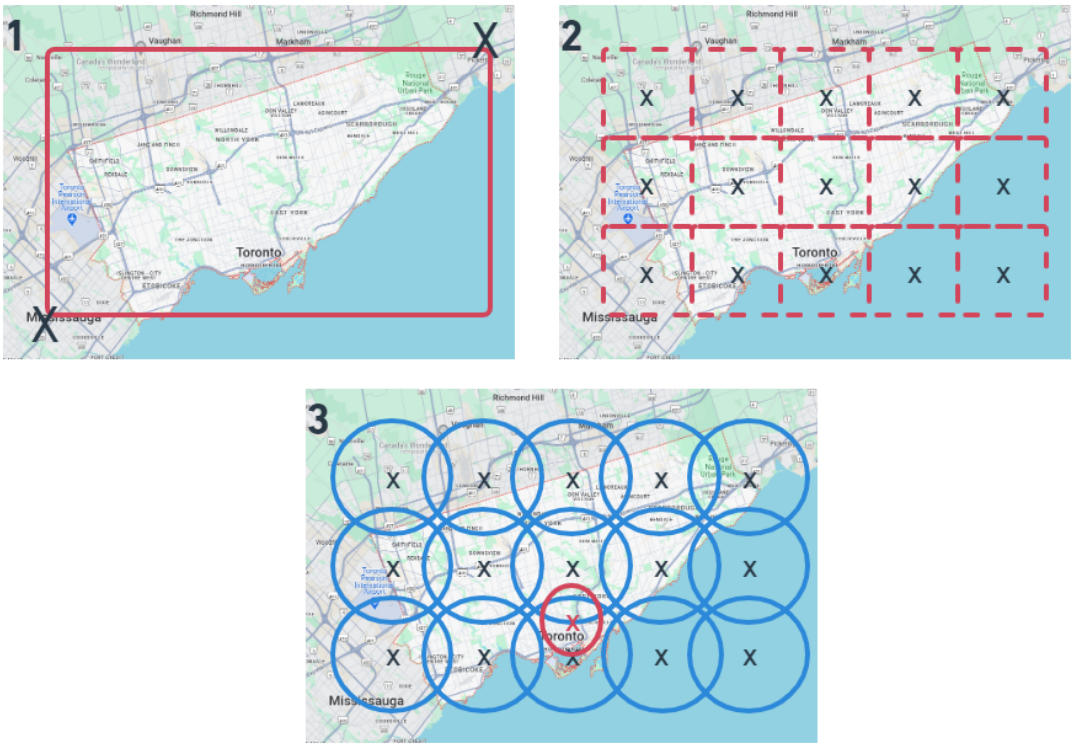


Figure 2 – Rectangular area delimited for Toronto (part 1). Grid cells with sides of 9,000 meters (red squares) (part 2). Automatically calculated circular areas (in blue) and highlighted city center (red circle) (part 3).

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 2 (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 2 (part 2) – cells that cover areas outside the city of interest are disregarded after a manual inspection. GP

API also demands a radius associated with each coordinate. For all cities, we choose a radius of 6,000 meters - Figure 2 (part 3) blue circles. For the city center, we perform an extra request where we choose a radius of 3,000 meters – Figure 2 (part 3) red circle. Doing this extra step for the city center is important because the number of venues returned per request is up to 60; thus, this extra request focused on a dense area minimizes potential losses of venues. 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[1] (GUBERT; SILVA, 2022).

Using the proposed strategy, we have collected data from 14 cities, namely: Curitiba and Rio de Janeiro in Brazil; Toronto and Vancouver in Canada; Chicago and Los Angeles in the USA; Berlin and Frankfurt in Germany; Paris and Lyon in France; Seoul and Busan in South Korea; and Nairobi and Mombasa in Kenya. These cities are important ones in the countries they are located in and cover regions with different cultural characteristics. The numbers of venues and unique categories found in each of the cities are presented in Table 1, in addition to the number of geographic coordinates used in the requests to cover each of the areas (the bigger the area informed, the more coordinates).

Table 1 – Number of venues, unique categories, and coordinates used by each city.

City	Venues	Categories	Coordinates
Curitiba	31.539	748	5
Rio de Janeiro	83.819	773	17
Toronto	62.282	818	11
Vancouver	20.536	796	2
Chicago	55.063	839	9
Los Angeles	115.761	834	20
Berlin	72.338	825	18
Frankfurt	25.739	735	7
Paris	36.380	817	4
Lyon	16.796	690	1
Seoul	50.721	716	8
Busan	45.910	622	19
Nairobi	35.131	746	4
Mombasa	9.799	612	1

As can be seen, thanks to our strategy, all cities in our final dataset have more than 600 categories, expressing a considerable diversity in terms of venues, much higher than the original number of basic categories provided by the GP API (141 categories).

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

3.2 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; SILVER, 2025). Figure 3 and 4 provide an overview of the entire mapping process, which is detailed below. The venue index v has been omitted for simplicity.

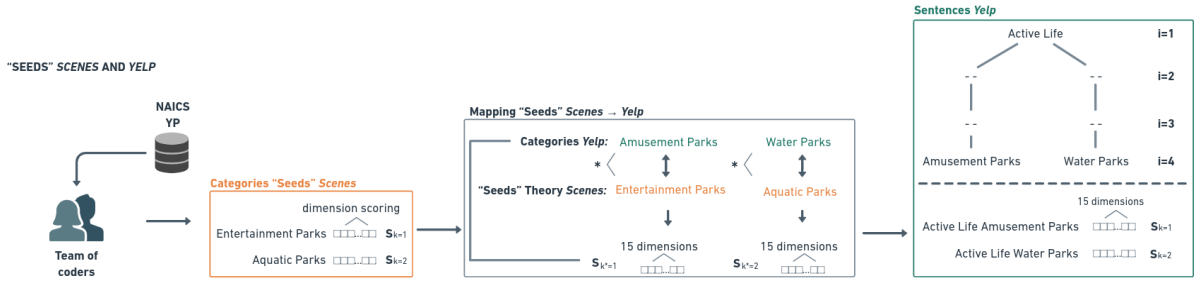


Figure 3 – Process of mapping of the "seeds" Scenes to Yelp.

The first part of the mapping process (left in the Figure 3) 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. 1.

$$\{\mathbf{s}_k(\text{Scenes})\} = \{\mathbf{s}_k(\text{NAICS})\} \cup \{\mathbf{s}_k(\text{YP})\}, \tag{1}$$

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 2.1.2.

Here, it is important to highlight that although the theory was initially developed with data sets from the USA and Canada, the knowledge transfer 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 to Yelp categories (center of Figure 3, representing the mapping $Scenes \rightarrow Yelp$), 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 ($\text{categ}_{k'}$(Scenes)). Thus, each 15 dimension vector shown in the center of the figure is given by Eq. 2.

$$\mathbf{s}_k(\text{Yelp}) = \mathbf{s}_{k^*}(\text{Scenes}), \tag{2}$$

where

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

Although the Yelp database is also primarily focused on the Global North, in this work, it is used as an auxiliary step $Scenes \rightarrow$ Yelp $\rightarrow$ GP to support the direct mapping $Scenes \rightarrow$ GP. This intermediate step 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 right of Figure 3. 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. 3.

$$\text{Sent}_k(\text{Yelp}) = \text{concat}\left[\text{categ}_k(i,\text{Yelp})\right], i \leq 4. \tag{3}$$

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 in Figure 4, the broader description of GP categories can be turned more informative using the Yelp database. It illustrates an example for two different venues, each one provided by a different dataset, venue A from Dataset States and venue B from Dataset Cities.

To provide a richer description of venues, both the selected Yelp categories used in the requests and the broader categories available from Google Places (GP) are incorporated. To enhance semantic capacity and improve mapping accuracy, descriptive sentences are generated for each venue v following the procedures specific to each dataset.

For Dataset States one sentence is created per venue v, combining all associated categories. For example, if the venue has the categories “Italian”, “Restaurant” and “Food”, the sentence is: “Italian Restaurant Food”.

Recall that Dataset Cities by default does not have categories with the necessary level of specificity. Therefore, sentences consist of a Yelp category used in the requested data and all the GP categories associated with that venue. 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:

$$\text{Sent}_k^v(\text{GP}) = concat\left[\text{categ}_k^v(\text{leaf,Yelp}), \text{categ}_k^v(GP)\right] \tag{4}$$

Given the existing mapping of "seeds" of Scenes Theory to the Yelp categories (Eq. 2) and the enrichment of Google data with some Yelp categories (Eq. 4), we chose to perform the second step in the last mapping stage (bottom of Figure 4) directly with Yelp. This map-

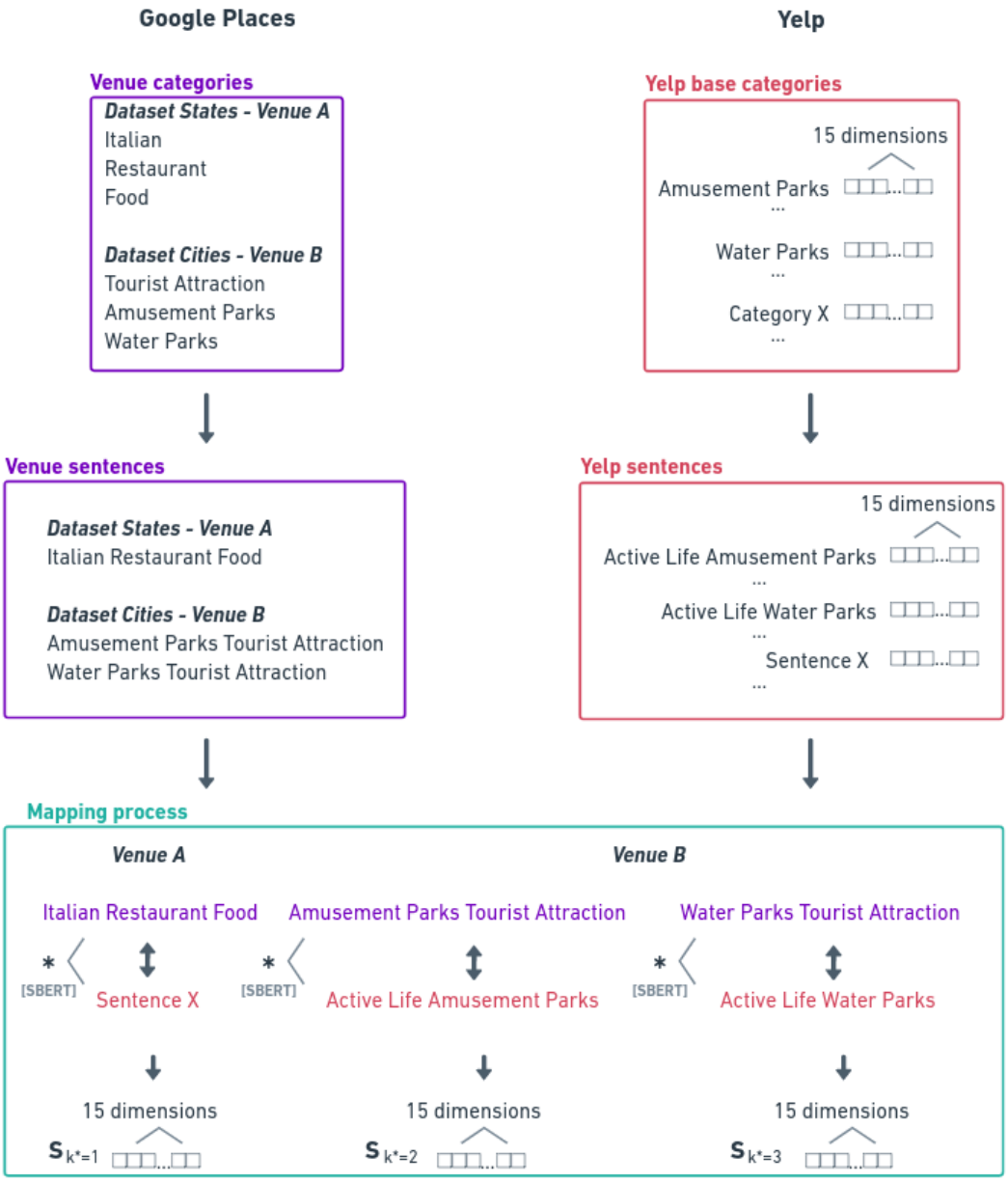


Figure 4 – Overview of mapping GP categories to the local cultural dimensions (from the Scenes Theory).

ping process was carried out with SBERT, using the framework Sentence Transformers, in which several pre-trained 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; 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 GP category k ($\text{Sent}_k^v(\text{GP})$), the generated embeddings $\mathcal{E}$ are compared using cosine similarity (cosS), retrieving the highest-scoring Yelp sentence as detailed in Eqs. 5 and 6.

$$\mathbf{s}_k^v(\text{GP}) = \mathbf{s}_{k*}^v(\text{Yelp}), \tag{5}$$

$$k^* = \text{argMax}_{k'} \cos\text{S}\left[\mathcal{E}(\text{Sent}_k^v(\text{GP})), \mathcal{E}(\text{Sent}_{k'}^v(\text{Yelp}))\right]. \quad (6)$$

Then, we selected around 50 random sentences from Google Places and mapped them to respective Yelp sentences. This sample was analyzed with different models and after observation and manual judgment, the “all-MiniLM-L6-v2" model was chosen, which generally showed more coherence and assertiveness given our context. In reviewing sample results, we found 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 knowledge transfer 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. 10, 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. 5. Therefore, the proposed mapping depicted in Figure 3 and 4 ($Scenes \rightarrow$ Yelp $\rightarrow$ GP) is based on knowledge transfer, through the score vector ($\mathbf{s}_k^v$) of venue v in category k, for all v, as described by Eqs. 1 to 5.

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

3.3 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. This selection 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. 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. Each region is treated as a "scene" and is mapped to the 15 dimensions through the cultural signature (Eq. 9) 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; CLARK, 2016). Correlations are also calculated using Yelp data obtained by Silva and Silver (2025).

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

	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

These sources provide reliable inputs for analyzing FSA regions and have been validated as trustworthy.

The Pearson correlation coefficient measures the linear relationship between two variables, 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; KOSSOWSKI, 2011). The results are shown in Figure 5.

The figure reveals that except for the "Tradition" and "Egalitarian," all other dimensions resulted 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 effectively ensures the creation of reliable cultural signatures using Google Places.

3.4 Proposed Cultural Signatures

We propose three approaches to creating cultural signatures, which are described below.

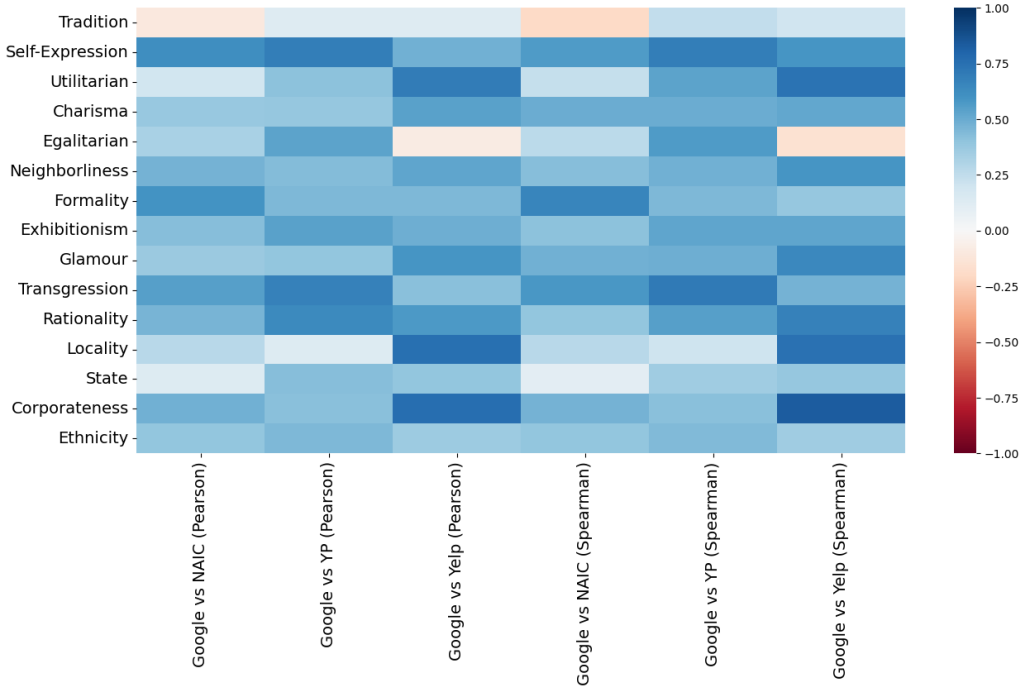


Figure 5 – Validation in Toronto: results of Pearson (first three columns) and Spearman (last three columns) correlations, calculated between data from Google and the YP, NAICS, and Yelp databases.

3.4.1 *Scenes-based approach*

The first cultural signature proposal is called Scenes-based approach. With the scoring system, detailed in the subsection 2.1.2, 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 $S^v_{K^v \times D}$, where K^v 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 to venue v. The information stored in $S^v_{K^v \times D}$ can be averaged to derive a unique **scenes vector**

$$\mathbf{s}^v = \{s^v_1, s^v_2, \ldots, s^v_D\}, \tag{7}$$

where s^v_d is computed as

$$s^v_d = \frac{1}{K} \sum_{k=1}^{K^v} s^v_{k,d}. \tag{8}$$

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:

$$c_d = \frac{1}{V} \sum_{v=1}^{V} s_d^v. \tag{9}$$

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

$$S^v_{K\times D}(\text{GP}) = f\left(\{\mathbf{s}_k^v(\text{Scenes})\}, \{\text{categ}^v(\text{Yelp})\}\right). \tag{10}$$

This mapping enables addressing areas as scenes and comparing their *cultural signatures* (Eq. 9), as it encompasses a diverse set of venues that provide different dimensions of meaning. For more details, refer to Subsection 3.2.

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 (BRITO *et al.*, 2018; SENEFONTE *et al.*, 2020; PRADA; SMALL, 2024; ARRIBAS-BEL; 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.

3.4.2 *Naive* and *Frequency-based approach*

We also also consider two other alternative approaches to creating cultural signatures. They disregard the "Scenes" information, using only the venue categories:

- *Naive-based approach* - This approach considers only the existence or not of the category in the area, for a particular urban area (region r), we have a vector describing it by all unique categories found in that area. For example, an area could be described by the categories [University, Restaurant, Coffee Shop, American Restaurant] and another by [Italian Restaurant, Wine Shop]. This strategy disregards the frequency of categories.

$$\mathbf{n}^r = \{n_1^r, n_2^r, \ldots, n_K^r\}, \tag{11}$$

where $K = |\{\text{categ}(DS)\}|$ is the total categories addressed by a particular dataset DS and n_k^r is computed as

$$n_k^r = \begin{cases} 1 & \text{if } k \in \{\text{categ}(DS)\}_r \\ 0 & \text{otherwise} \end{cases}$$

where $\{\text{categ}(DS)\}_r$ is the set of categories addressed by DS in a region r.

- *Frequency-based approach* - As in *Naive*, it also considers all unique categories in an urban area, but their frequency comes into play here.

$$\mathbf{f}^r = \{f_1^r, f_2^r, \ldots, f_K^r\}, \tag{12}$$

The frequency values f_k^r are normalized per category as:

$$f_k^r = \text{norm}\left(\sum_{v \in r} \chi_{k,v}^r\right) \rightarrow [0,\text{max}_k],$$

where

$$\chi_{k,v}^r = \begin{cases} 1 & \text{if } k,v \in \{\text{categ}(DS)\}_r \\ 0 & \text{otherwise} \end{cases}$$

Naive helps answer the question: Is the existence of certain types of venues in two different urban areas enough to explain their cultural differences? *Frequency* helps answer a complementary question: Is the quantity of categories helpful in this task?

4 CULTURAL SIGNATURES TO IDENTIFY CULTURALLY SIMILAR AREAS

In this chapter, the results for Dataset Cities and Dataset States are presented, followed by evidence through comparison with Survey Data.

4.1 Cities Worldwide

Using Dataset Cities we applied the knowledge transfer methodology and created cultural signatures ($\mathbf{c}$), naive ($\mathbf{n}$) and frequency vector ($\mathbf{f}$) for all 14 cities – as described by Eqs. 9, 11, and 12, respectively.

4.1.1 *Scenes* for Dataset Cities

First, we evaluate the results of the cultural signatures ($\mathbf{c}$, Eq. 9) generated by the Scenes-based approach. We perform hierarchical clustering using Ward's linkage method and Euclidean distance, with the 15 dimensions of Scenes Theory as features. The results are represented in the dendrogram depicted in Figure 6, where a division into six clusters is identified.

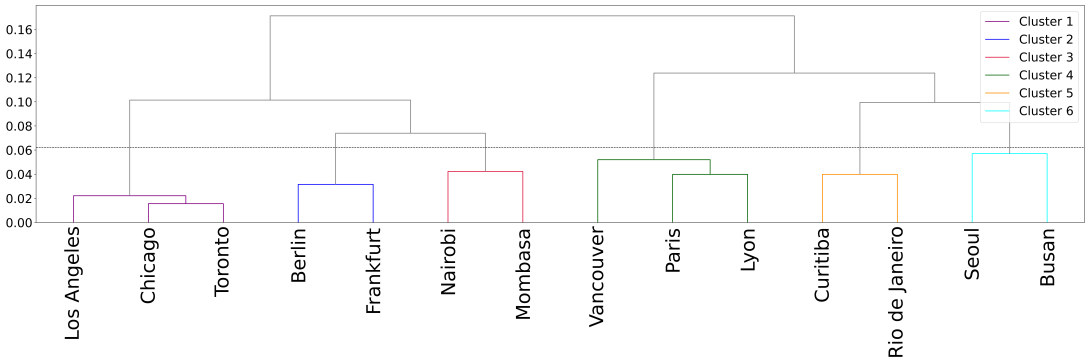


Figure 6 – Hierarchical clustering dendrogram of cities represented by *Scenes*.

The result aligns with what is expected concerning the cultural characteristics of the studied areas. Most of the clusters grouped cities from the same country, which is coherent because, in general, countries have distinct cultural characteristics; the exceptions in this sense are clusters 1 and 4. In cluster 1, Toronto was grouped with Chicago and Los Angeles; note also that Los Angeles is the most dissimilar city in the grouping. The result of Chicago and Toronto being together and more similar makes sense, in that they are often considered to be culturally similar to one another, even compared to Los Angeles. Regarding cluster 4, Vancouver was grouped with Paris and Lyon. We found significant similarities between the most recurrent categories of French cities and Vancouver, such as "Art galleries," which could help explain this result. Although German cities (Berlin and Frankfurt) and French cities (Paris and Lyon) are on

the same continent, they are quite distinct culturally, and so their location in separate clusters seems reasonable.

To facilitate a comparative analysis by contrasting the values of each cluster dimension with its corresponding overall average, we calculate the Z-Score, as shown in Figure 7. The Z-Score is the number of standard deviations concerning the average of what is being observed. This facilitates comparing clusters by extracting the characteristics that stand out in each compared with a general overview, i.e., the centroid of clusters' centroids. For example, cluster 3, representing Kenya, has one of the lowest values for Tradition. In contrast, for cluster 4 with the cities Vancouver, Paris, and Lyon, this dimension represents one of the most important characteristics. Looking at cluster 1, composed of Chicago, Los Angeles and Toronto, we see that Tradition is not as predominant as in cluster 4. This reinforces the potential for characterizing cultural signatures and enabling an overview of geographic areas by simply extracting their most evident dimensions.

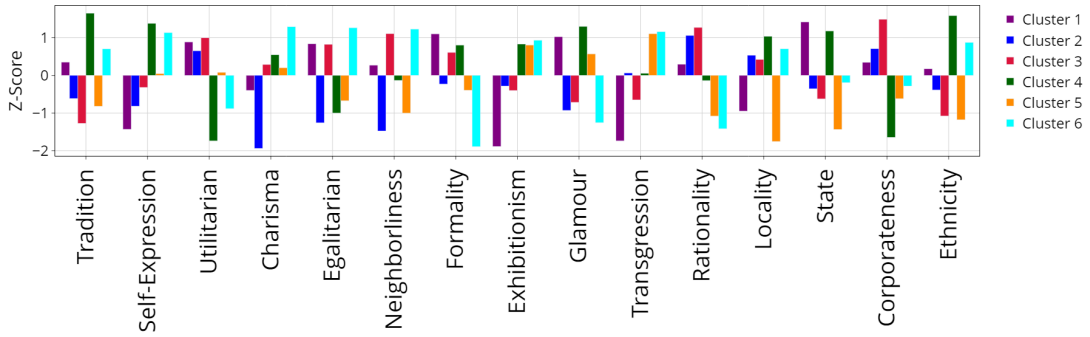


Figure 7 – Z-Score values of Scenes dimensions per cluster. Cluster numbers follow what is presented in Figure 6.

4.1.2 *Naive* for Dataset Cities

For *Naive*, we perform hierarchical clustering using Ward's linkage criteria and Euclidean distance – best combination tested – on naive vectors $\mathbf{n}$ (Eq. 11). The results are very far from those achieved by the Scenes-based approach. Variations in dendrogram structure are insignificant, even with other combinations of clustering parameters to verify possible changes and improvements in the results.

4.1.3 *Frequency* for Dataset Cities

For *Frequency*, we perform hierarchical clustering using the Complete linkage criteria and Cosine distance – the best combination tested. As depicted in Figure 8, the results for *Frequency*, as with *Scenes*, align with what is expected regarding grouping cities of the same

country. However, using *Frequency* differently, Chicago is more similar to Los Angeles, and Vancouver is more related to Toronto than to the French cities.

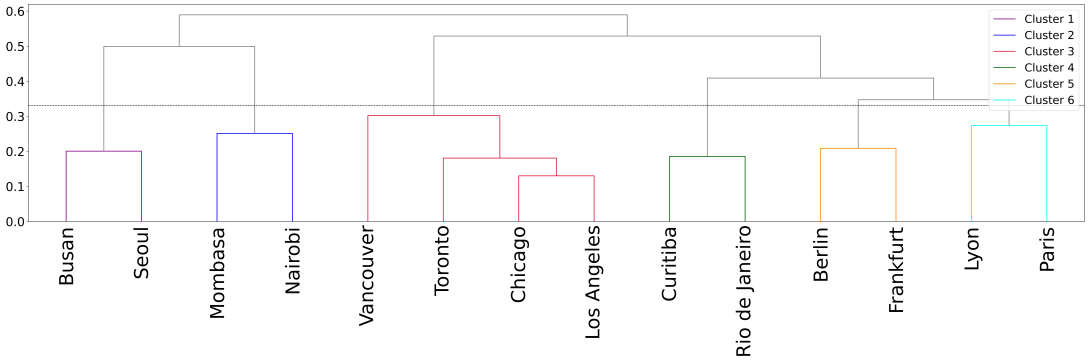


Figure 8 – Hierarchical clustering dendrogram of cities represented by *Frequency*.

The results obtained demand reflection because although Toronto and Vancouver are in the same country, they are not necessarily similar in terms of immigration patterns, governance, geography, ecology, and cultural style. Toronto and Chicago, on the other hand, have much in common: they are both Great Lakes cities, with strong industrial heritages and are now in the midst of a post-industrial transformation. Hence, they are often compared as similar cases (ROBSON *et al.*, 2019; KOLPAK; WANG, 2017). In fact, a hallmark of the scenes approach is that lower-level units (neighborhoods or cities) can sometimes be more similar to one another than higher-level units (states, countries), which are often missed by other approaches. Therefore, using *Scenes* might allow these sorts of similarities to be identified. Nevertheless, validating such representations for urban settings lacks a definitive ground truth.

We can reveal specific characteristics of each cluster by extracting the five most distinct categories for each of them – we do that by calculating the distance of the category from its cluster centroid. After that, we calculate the Z-Score for the selected categories against the overall average. The result of this process is illustrated in Figure 9.

Certain categories in some clusters stand out so notably that they not only significantly deviate from their overall average, but also emerge as the sole positive value compared to others. For example, in French cities, “municipality”, in Brazilian cities, “hang gliding”, and in Korean cities, “face painting” exhibits this distinct characteristic. Making a comparison with the Z-Score values illustrated in Figure 7, we can relate these specific findings depicted in Figure 9 to the aspects highlighted in Tradition for cluster 4 (predominantly French), Transgression for cluster 5 (Brazil) and Self-Expression and Charisma for cluster 6 (South Korea).

On the other hand, to analyze the clusters that differ from *Scenes* to *Frequency* approaches, we look for the most evident characteristics in each. For clusters 1 and 4 of *Scenes*, we select the three dimensions that stand out most in each one and retrieve the most important sentences for each of them. While for *Frequency*, we look for the 50 most frequent categories for cluster 3. In *Scenes*, the main characteristics become more evident as the sentences provide more depth of meaning within the scope of each dimension. For example, Los Angeles, Chicago,

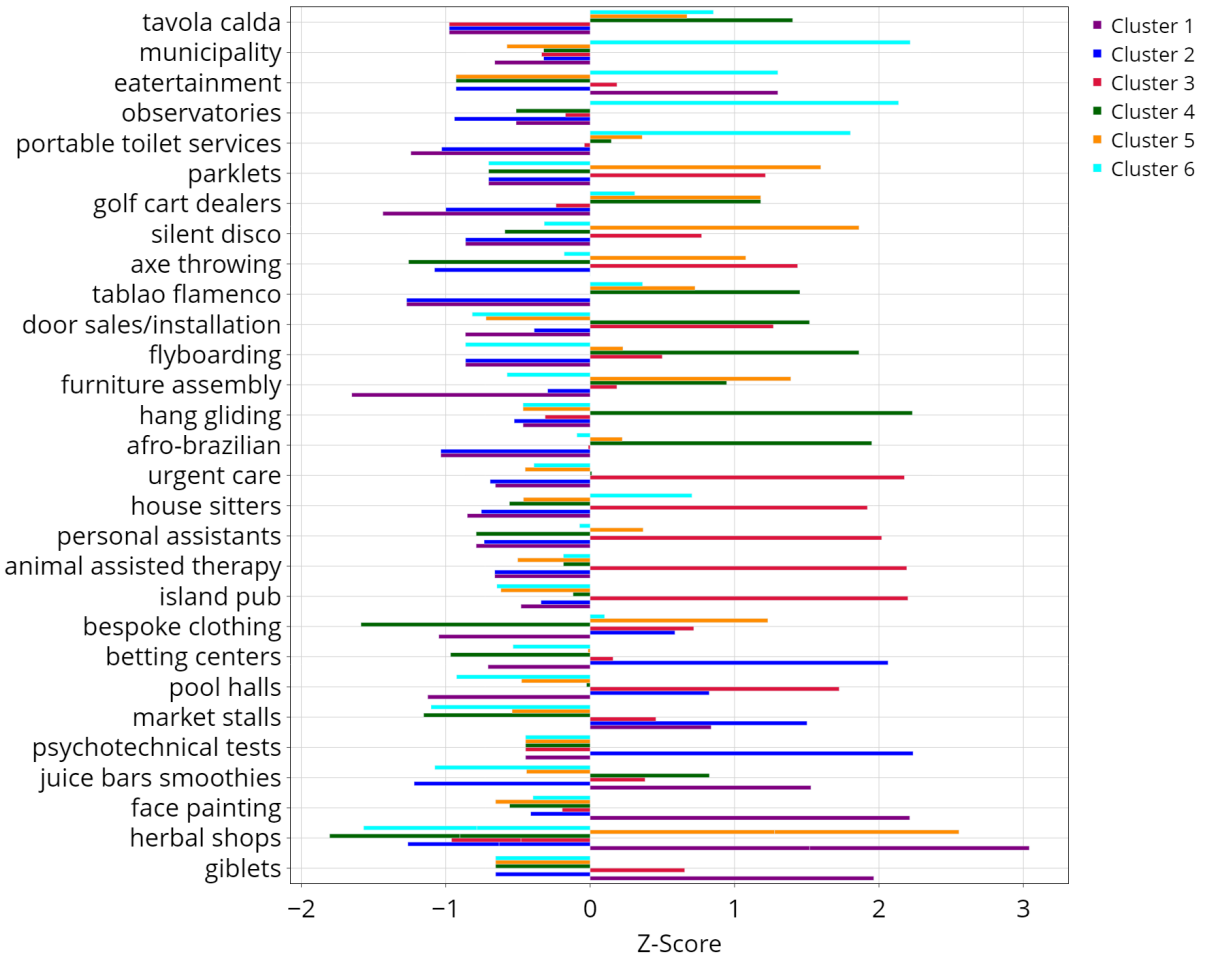


Figure 9 – Z-Score values for the most distinct categories per cluster, considering the clustering of *Frequency*.

and Toronto have “Business Consulting”, “Libraries” and “Gastropubs” in common, whereas Vancouver, Paris, and Lyon are marked by “Antiques Book Store”, “Art Gallery”, “Comedy and Night Club” and gastronomic diversity, such as “Portuguese Bakery”, “Spanish Meal Delivery”, “Sushi Bars” and “Tapas Bars”. In *Frequency*, many categories can be found that summarize these characteristics, such as “Gastropubs”, “Art Installation”, “Imported Food”, “Meal Takeaway” and “Souvenir Shops”. What this result suggests is that differently from *Frequency*, *Scenes* appears to discern slight differences among categories that, typically, convey similar meanings – this is done by exploring human knowledge in the *Scenes*’ dimensions. In fact, when clustering on PCA components that explains 100% of the variability of *Frequency* features, we have a result more similar to *Scenes*, i.e., Toronto is closer to LA and Chicago, and Vancouver is closer to Paris and Lyon, despite the clustering quality in general not being as good as with *Scenes*. This highlights the influence of feature space size on the clustering performance.

4.2 All States in the USA

Using Dataset States, we applied the knowledge transfer methodology and created cultural signatures ($\mathbf{c}$), naive ($\mathbf{n}$) and frequency vector ($\mathbf{f}$) for all US states – as described by Eqs. 9, 11, and 12, respectively.

4.2.1 Evaluating Scenes-based approach for Dataset States

To analyze cultural signatures ($\mathbf{c}$) by every US state using *Scenes*, we also perform hierarchical clustering considering the 15 dimensions of the Scenes Theory as features, maintaining the Ward linkage criteria and Euclidean distance.

By inspecting the dendrogram, Figure 10(a), we observe a tendency to group by geographic region. Mapping one of the clearest cuts in the dendrogram, we obtain Figure 10(b), which makes it easier to see this information. Note also that culturally similar regions, e.g., the US South, are all grouped. These results reinforce the usefulness of the proposed method in identifying culturally similar regions.

4.2.2 Evaluating *Naive* and Frequency-based approaches for Dataset States

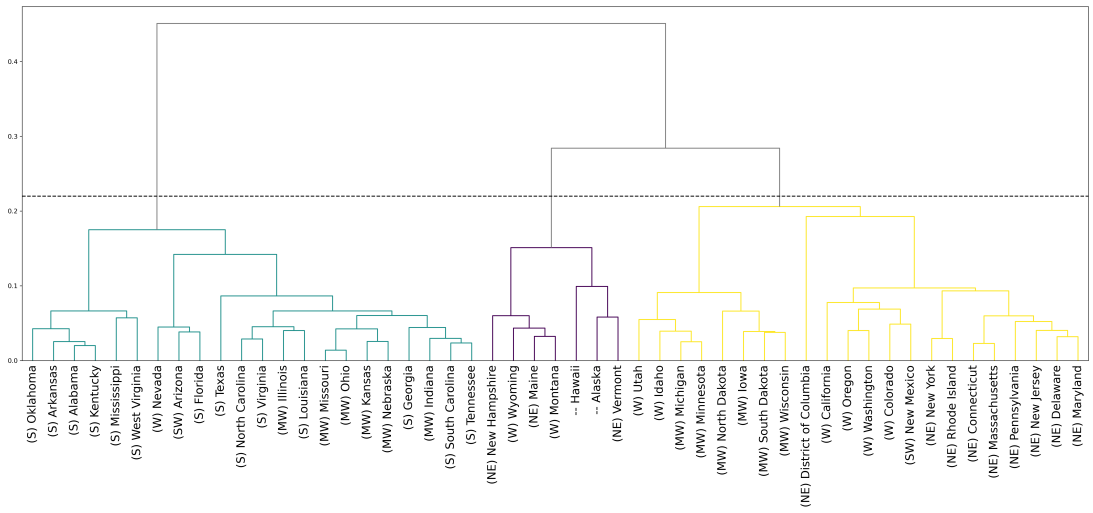
Both, *Naive* using $\mathbf{n}$ and *Frequency* using $\mathbf{f}$, are also applied to data from US states. For these cases, we perform hierarchical clustering using Ward linkage criteria and Euclidean distance – other combinations were experimented with, yet none proved superior. Now, we observe a more significant difference between the two approaches and the results obtained with *Scenes*.

Figure 11 illustrates the results provided by the Frequency-based approach, the best between the two simpler approaches. Figure 11(a) shows the resulting dendrogram, and Figure 11(b) illustrates the three mapped clusters.

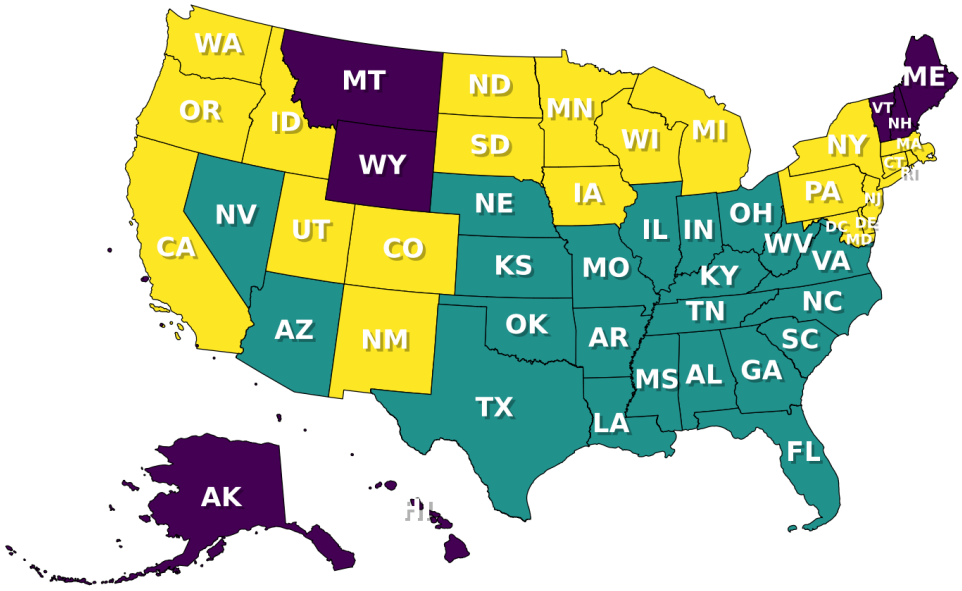
Looking carefully at the results, it is not possible to detect clear patterns in these results, at least as clear as identified by *Scenes*, regardless of the number of clusters adopted. Surprisingly, Alaska and Maine are positioned within clusters larger than with *Scenes*. Alaska is situated among states such as Washington, Oregon, North Dakota, Minnesota, and Michigan. Maine is part of the largest cluster, which includes most of the remaining states. Thus, *Scenes* provides extra semantic expressiveness in smaller dimensions.

4.3 Comparing with Survey Data

There is no clear way to access the ground truth of our results. However, we explore in this work a source where we expect some correlation: the American Value Survey (AVS, access https://www.prri.org). The survey was conducted among a representative sample of 5,031



(a) Dendrogram



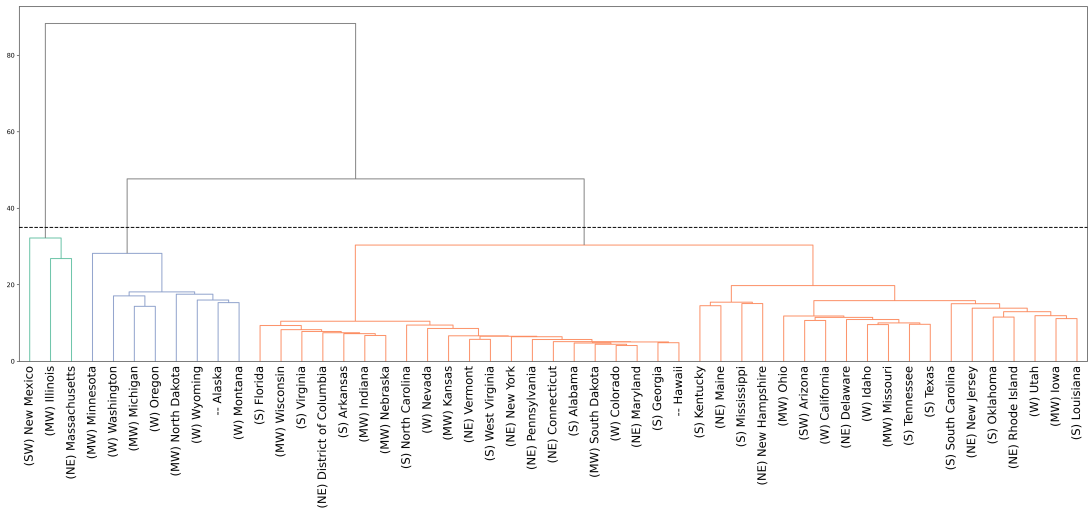
(b) Clusters mapped

Figure 10 – Results of hierarchical clustering considering all states in the USA represented by *Scenes*.

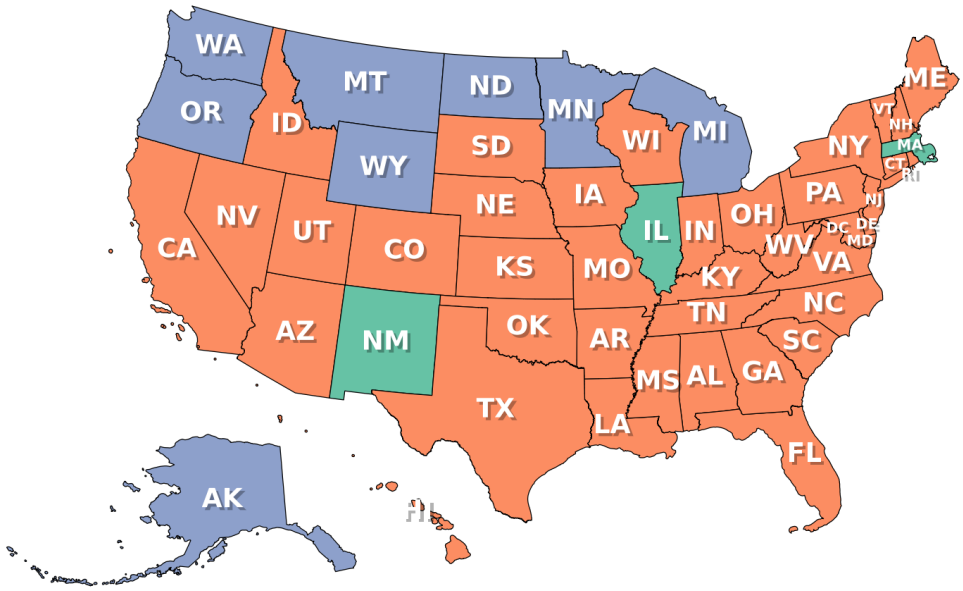
adults (age 18 and up) living in all 50 states in the United States, having a statistically valid representation of the USA population, including many minorities or hard-to-reach populations. Interviews were conducted online between September 16-29, 2021 and September 1-11, 2022. Wyoming had only three respondents in both years, so we removed it from our comparison. Additional details about the methodology can be found on the Ipsos website[1].

The questions in the survey consider political aspects and basic beliefs. We represent the questions as features to describe states, where the values are the mean answers regarding

[1] https://www.ipsos.com/en-us/solutions/public-affairs/knowledgepanel.



(a) Dendrogram



(b) Clusters mapped

Figure 11 – Results of hierarchical clustering considering all states in the USA represented by *Frequency*.

all participants for a particular state. We disregard all political questions and keep only basic beliefs[2].

To assess the relationship between the results of the AVS and two of our proposals (*Scenes* and *Frequency*), we used the Pearson correlation for the Euclidean distance between all pairs of states when describing them by AVS and our approaches. In other words, we calculate the distance between all vectors (states) in the context of *Scenes*, *Frequency* and AVS and then

[2] The complete list of questions used can be found at: https://sites.google.com/view/neighbourhood-change.

calculate correlation of these distances between AVSx*Scenes* and AVSx*Frequency*. By doing that, we got a moderate correlation of 0.51 ($p < 10^{-4}$) for *Scenes*. Using *Frequency*, on the other hand, resulted in a Pearson correlation of -0.06 ($p < 10^{-1}$) for the Euclidean distance between all pairs of states. To better understand the correlation results individually across the entire USA, we divided the analysis by state. This involved calculating the Euclidean distance of each state in relation to all others, considering its descriptions using AVS and each of our proposals. Then, we calculate the Pearson correlation of these values. Table 3 shows the results for both approaches, *Scenes* and *Frequency*.

Table 3 – Pearson correlation r (and its p-value) between the Euclidean distance of a particular state vs all others when describing them by AVS *versus Scenes* or *Frequency*.

State	r (p-value)		State	r (p-value)	
	Scenes	*Frequency*		*Scenes*	*Frequency*
Alaska	-0.221 (0.13)	-0.257 (0.07)	Virginia	0.516 (0.00)	-0.128 (0.38)
Hawaii	-0.046 (0.76)	-0.106 (0.47)	Massachusetts	0.517 (0.00)	-0.008 (0.96)
Maine	-0.038 (0.80)	-0.146 (0.32)	California	0.521 (0.00)	-0.099 (0.50)
Montana	0.013 (0.93)	-0.096 (0.51)	Kansas	0.536 (0.00)	-0.023 (0.88)
Vermont	0.019 (0.90)	-0.175 (0.23)	North Carolina	0.560 (0.00)	-0.069 (0.64)
Wisconsin	0.070 (0.63)	-0.120 (0.41)	Illinois	0.562 (0.00)	-0.034 (0.82)
New York	0.122 (0.41)	-0.096 (0.51)	Connecticut	0.567 (0.00)	-0.068 (0.64)
New Hampshire	0.131 (0.37)	-0.091 (0.54)	Arkansas	0.569 (0.00)	-0.017 (0.91)
Rhode Island	0.133 (0.36)	-0.058 (0.69)	Ohio	0.595 (0.00)	-0.073 (0.62)
Pennsylvania	0.277 (0.05)	-0.012 (0.93)	Missouri	0.597 (0.00)	-0.028 (0.85)
Oregon	0.294 (0.04)	-0.219 (0.13)	Colorado	0.597 (0.00)	-0.112 (0.44)
North Dakota	0.298 (0.04)	-0.048 (0.75)	Nevada	0.602 (0.00)	-0.131 (0.37)
South Dakota	0.349 (0.01)	0.050 (0.73)	Texas	0.626 (0.00)	-0.016 (0.92)
Idaho	0.394 (0.01)	-0.142 (0.33)	Nebraska	0.639 (0.00)	0.071 (0.63)
Iowa	0.396 (0.00)	-0.007 (0.96)	Mississippi	0.640 (0.00)	0.036 (0.81)
Delaware	0.409 (0.00)	0.009 (0.95)	Kentucky	0.643 (0.00)	-0.062 (0.67)
Minnesota	0.418 (0.00)	-0.216 (0.14)	West Virginia	0.649 (0.00)	0.149 (0.31)
New Jersey	0.443 (0.00)	-0.042 (0.77)	Tennessee	0.652 (0.00)	0.043 (0.77)
Washington	0.446 (0.00)	-0.196 (0.18)	Florida	0.652 (0.00)	0.002 (0.99)
Dist. Columbia	0.458 (0.00)	-0.154 (0.29)	Arizona	0.659 (0.00)	-0.061 (0.68)
Alabama	0.462 (0.00)	0.035 (0.81)	South Carolina	0.663 (0.00)	0.006 (0.97)
New Mexico	0.472 (0.00)	0.061 (0.68)	Georgia	0.667 (0.00)	0.037 (0.80)
Maryland	0.501 (0.00)	-0.212 (0.14)	Oklahoma	0.678 (0.00)	0.034 (0.82)
Utah	0.504 (0.00)	-0.168 (0.25)	Indiana	0.694 (0.00)	0.001 (0.99)
Michigan	0.509 (0.00)	-0.185 (0.20)	Louisiana	0.709 (0.00)	0.091 (0.53)

Note that for *Scenes*, $r \in [-0.221, 0.709]$ and approximately 75% of all states exhibit either a moderate or high correlation. Additionally, it is worth mentioning that Alaska is the only state with a negative correlation. Figure 12 maps these correlations, where we can see geographical patterns, e.g., the tendency of lowest correlations on top border states. By looking at

the results for *Frequency*, with $r \in [-0.257, 0.149]$, it is clear that it shows a worse association with another source (AVS) regarding cultural beliefs. In all cases presented here we also calculated Spearman correlation, however Pearson presented a better result, proving the linear relationship, for this reason we omitted the results with Spearman.

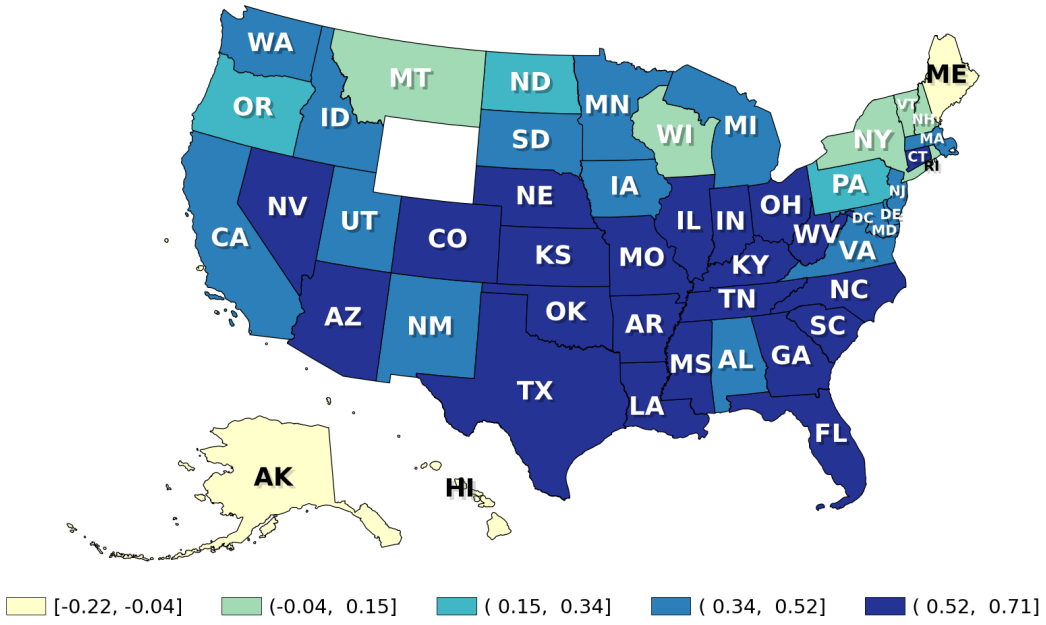


Figure 12 – Correlation ranges between AVS and *Scenes*.

5 ANALYZING CULTURALLY SIMILAR AREAS AT DIFFERENT GRANULARITY LEVELS

This chapter explores different levels of granularity to recommend the one capable of delivering more reliable cultural signatures. It then uses this result in a practical analysis with the cities of Curitiba and Chicago.

5.1 Exploring the influence of granularity levels

In more granular analyses, 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 by 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 consistent and 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[1] 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 results presented in Figure 13. 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 level 6, the areas do not differ much compared to the other granularities, which is confirmed by the values close to zero of the distances between the hexagons, not seeming like a good option in our analysis. At granularity level 8, the distances between the hexagons increase significantly, which may indicate high sensitivity in distance variation and little information in the creation of their cultural signatures. Finally, level 7 shows a good compromise in this regard.

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

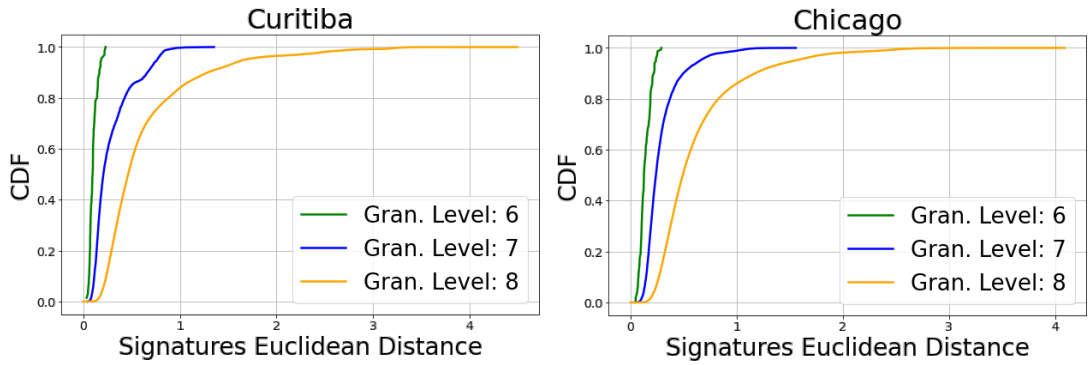


Figure 13 – Comparing Euclidean distance of cultural signatures for all pairs of hexagons at each granularity level in Curitiba and Chicago.

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 14 presents these results, showing category diversity across different granularity levels for both Curitiba and Chicago.

According to Figure 14, 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 for granularity levels 7 and 8, as shown in Figure 15. Given the importance of having a minimum number of venues per area to accurately capture cultural characteristics ($V \geq \xi$, in Eq. 9), but lacking a strict threshold (ξ), three experiments were conducted: the first used all hexagons ($\xi = 0$), the second ($\xi = 25$) included only hexagons with more than 25 venues, and the third ($\xi = 50$), 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, the Calinski-Harabasz metric prioritizes well-separated clusters, where higher values are associated with better-defined clusters. Conversely, the Davies-Bouldin metric favors clusters that are compact and distinct from one another, with lower values indicating improved quality. The Silhouette Index, ranging from -1 to 1, assesses how appropriately points are grouped; values near -1 suggest bad-defined clusters, around 0 indicate

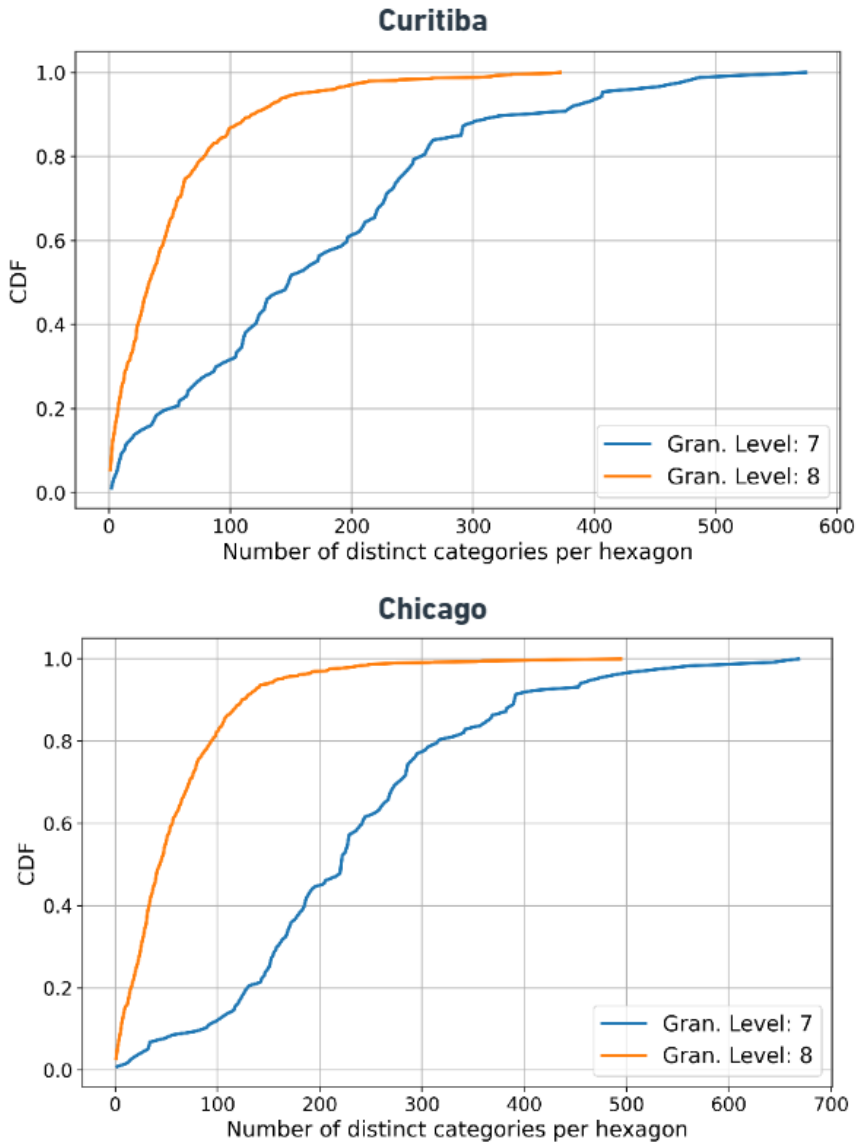


Figure 14 – Number of different categories per hexagon, considering granularity levels 7 and 8, for Curitiba and Chicago.

boundary points between clusters, and close to 1 imply well-defined, well-separated clusters. Results from the experiment that includes all hexagons ($\xi = 0$) seem inconclusive mainly because of overlap in metrics Davies-Bouldin and Silhouette Index, likely due to hexagons with limited information. Consequently, the analysis focuses on the experiments with ($\xi = 25$ and $\xi = 50$), i.e., hexagons that meet the 25- and 50-venue minimums. The differences in emphasis among the three metrics explain some apparent contradictions: Calinski-Harabasz, which rewards well-separated clusters, indicates that granularity level 8 offers better separation. However, level 7 performs better in the Davies-Bouldin and Silhouette metrics, prioritizing compact clusters. As these two metrics offer an assessment more focused on the perspective of each point (Silhouette) and the relative separation between clusters (Davies-Bouldin), we consider that level 7 presents a more robust clustering quality, being the most appropriate according to a joint and balanced assessment of the metrics.

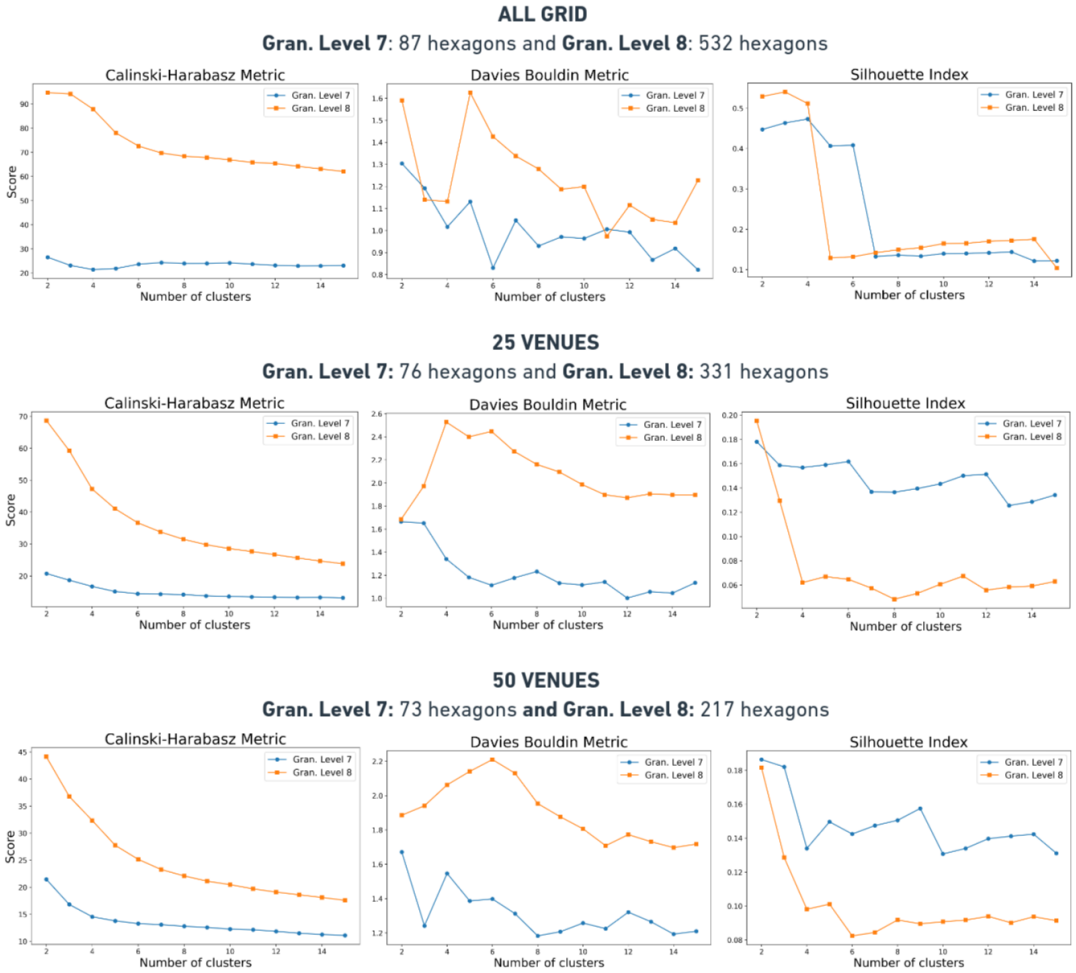


Figure 15 – Calculation of metrics to evaluate hexagon signature clustering considering ganularity levels 7 and 8 for Curitiba.

5.2 Deep diving into grid level 7 with Curitiba and Chicago

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 but preserves broad coverage of key areas while focusing on hexagons that offer rich, informative data about each city's cultural landscape.

For comparison and validation purposes of results with granularity level 7, a previous analysis based on signatures clustering is performed for Curitiba at the neighborhood level, maintaining the Scenes approach. It was decided to exclude 11 of the 75 neighborhoods with fewer than 100 venues, because their lack of information and possibly diversity could impact the calculation of cultural signatures. It is important to mention that as the analyses here are at the neighborhood level, we consider a greater number of venues as the minimum limit, compared to granularity levels 7 and 8 (25 and 50 venues). This left 64 neighborhoods, whose cultural signatures are calculated and clustered through the Hierarchical Agglomerative Clustering using Ward's method and Euclidean distance, with the 15 dimensions of the Scenes Theory as

features. This results in the dendrogram shown in Figure 16. The number of clusters is defined by cutting at the second-largest distance, as using only two clusters (largest distance) would be less meaningful in this context. This approach results in four clusters.

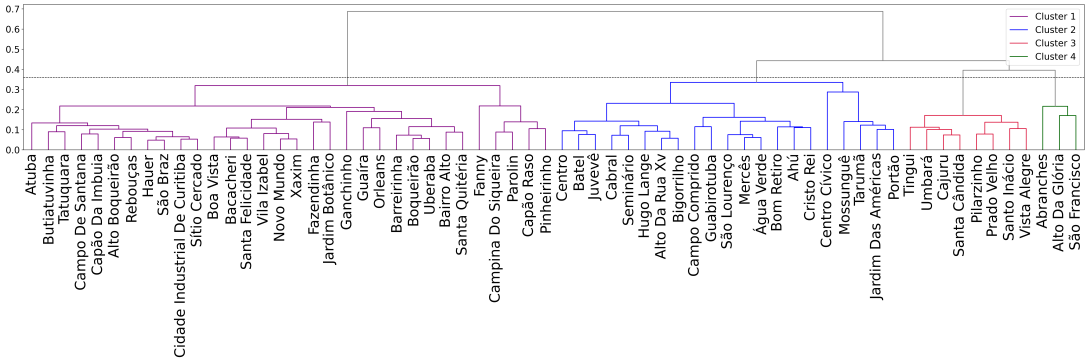


Figure 16 – Dendrogram of the Agglomerative Clustering: neighborhood clusters for Curitiba.

Cluster 1 (purple color in Figure 16) is the largest one, with 31 neighborhoods. The predominant regions are Bairro Novo, Boqueirão, Pinheirinho, and Tatuquara; it is further away from the Center and more concentrated on the south of the city. Most of the neighborhoods in these regions are characterized by the presence of museums, parks, squares, and tree-lined streets, as well as nightlife attractions, such as bars and clubs. Cluster 2 (blue in Figure 16) has 19 neighborhoods with a predominance of the Matriz region, characterized by being the commercial center, with regions 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 16) has 11 neighborhoods, which are located on the outskirts of the Center. These are regions with good commercial and leisure infrastructure, in addition to parks with extensive green areas. Finally, cluster 4 (green in Figure 16) 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.

Returning to the experiment with granularity level 7 and with the cultural signatures established for each hexagon $\mathbf{c}$, Hierarchical Clustering groups those signatures for both cities, using the Ward linkage method and Euclidean distance too. Differently from the analysis performed in Figure 16 which was based on neighborhood signature clusters, here, the number of hexagon signature clusters is selected based on the metrics from the prior analysis (Figure 15 with $\xi = 50$ for Curitiba and replicating the same calculation and analysis for Chicago): by choosing 5 clus-

ters for Curitiba and 4 for Chicago we balance the Davies-Bouldin and Silhouette Index metrics to achieve a meaningful clustering structure for each city.

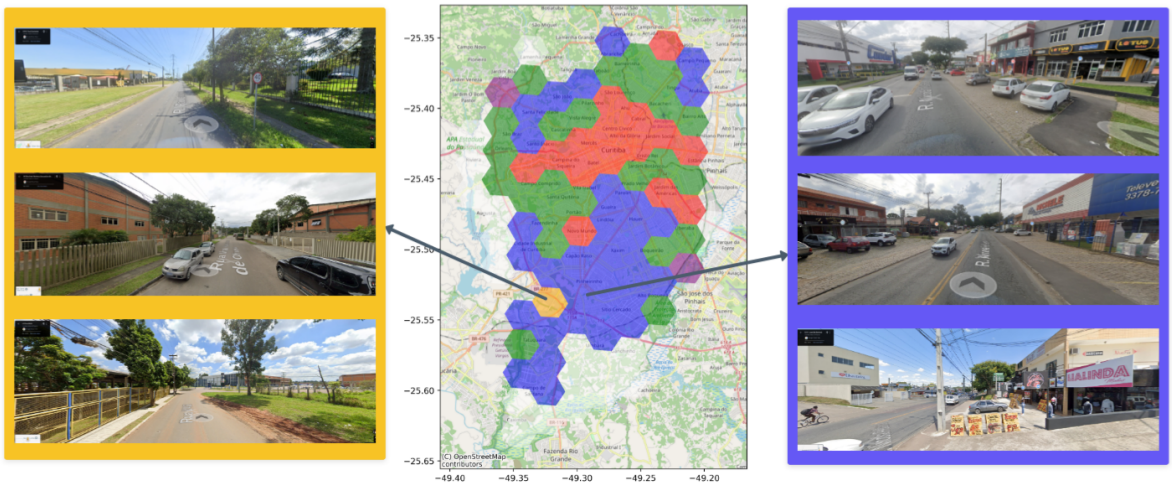


Figure 17 – Clustering of Curitiba with images for the yellow cluster and its surroundings (blue cluster).

The result of clustering signatures using hexagons in Curitiba is illustrated in Figure 17. Notice that three large clusters are apparent, similar to the neighborhood-approach findings: 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 Rocha (2023) helps to 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 18. 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 18), 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 17. This region (the yellow hexagon) 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. Furthermore, the infrastructure in the southern part of CIC is more oriented toward the industrial sector, with fewer leisure options and public spaces. This underscores the potential to identify cultural signatures and provide a comprehensive overview of geographic areas by extracting their key dimensions.

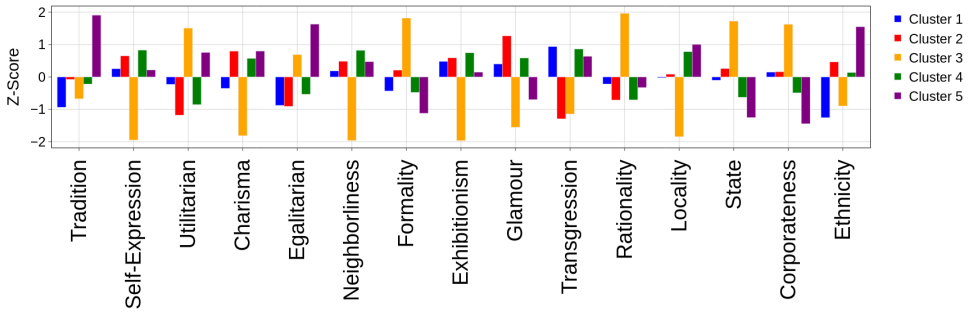


Figure 18 – Z-Score values of Scenes dimensions per cluster of Curitiba.

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 compared to neighborhoods. Specifically, analyzing smaller areas allows for the identification of 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 19 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 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, is 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

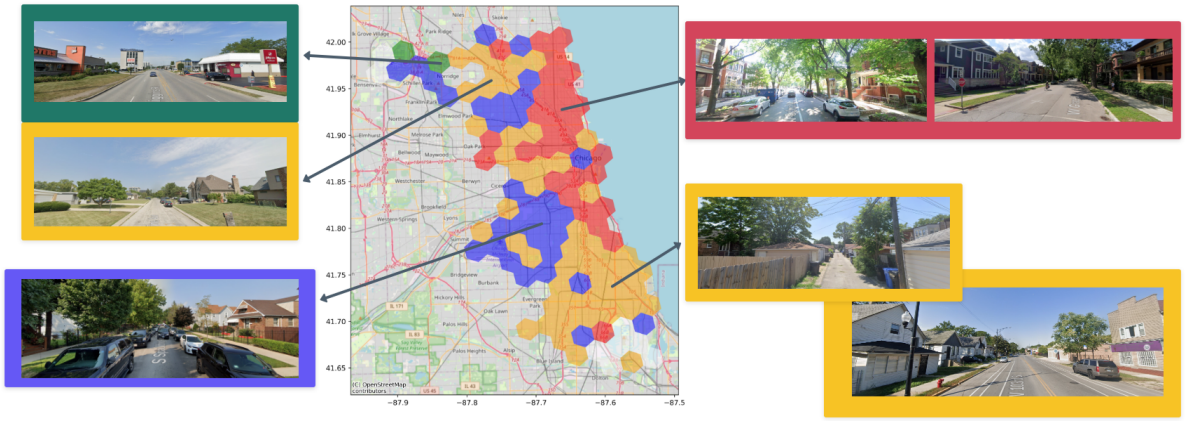


Figure 19 – Clustering of Chicago with images of some points in each cluster.

region which is predominantly white, with upper middle-class neighborhoods. Images of some selected points assist in this general interpretation.

Analyzing the cultural signature of the clusters through the Z-Score values calculated according to the dimensions of the Scenes Theory and illustrated in Figure 20, it is possible to verify how the red cluster stands out in dimensions that highlight the characteristics described above, 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 in terms of being the most peculiar and different from the others, in addition to showing evidence or distancing in many common dimensions, highlighting Self-Expression, Utilitarian, Formality and Rationality. The yellow clusters feature neighborliness and egalitarianism, along with some localism, charisma, and self-expression. This mix too reflects what ethnographers have long reported, notably in studies of African-American communities (such as 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 clusters stand 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). 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 other aspects besides social groups and ethnic communities.

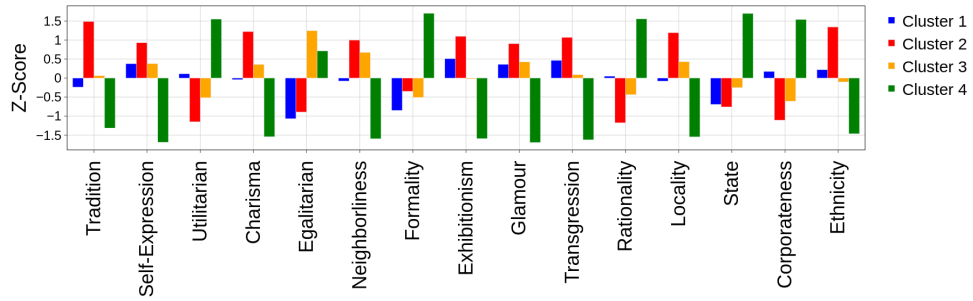


Figure 20 – Z-Score values of Scenes dimensions per cluster of Chicago.

5.3 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 21 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 daily life of the population.

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 valued properties, with a greater predominance of houses and few buildings, unlike the blue neighboring areas which are denser. And analyzing the Z-Score values in Figure 22, there is a predominance of the Tradition, Self-Expression, Charisma, Neighborliness, Glamour, Locality and Ethnicity dimensions. These evidences may justify the similarity with the hexagons in the red cluster in Chicago.

In contrast, the Curitiba cluster with a single hexagon in the previous analysis expanded to 4, with Chicago hexagons, of which 2 of them were already different from the rest of the city of Chicago. Here represented by the color green, Chicago's hexagons cover non-residential and warehouse areas, with highways and few businesses, resembling the characteristics of the south of CIC in Curitiba and reinforcing the same dimensions previously seen in the CIC, such as Utilitarian, Formality, Rationality and Corporateness, which can be found in Figure 22. This

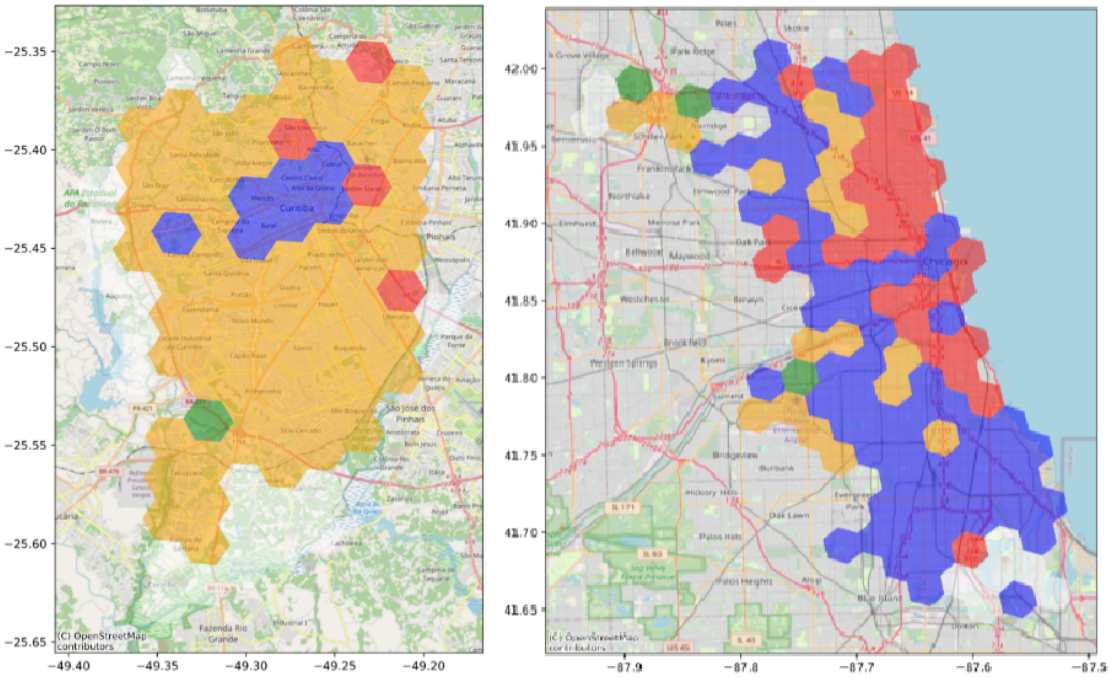


Figure 21 – Clustering Cultural Signatures of Curitiba and Chicago together.

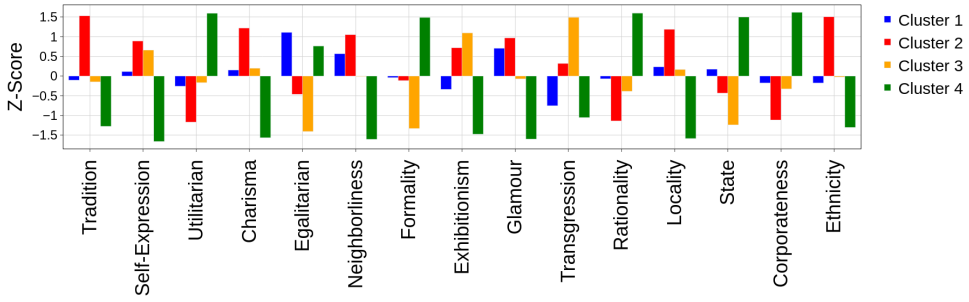


Figure 22 – Z-Score values of Scenes dimensions per cluster of Curitiba and Chicago together.

result shows how it is possible to find latent patterns between areas of different cities that are not immediately visible through conventional analyses.

6 CONCLUSION

Obtaining cultural characteristics on a large scale presents significant challenges. Recognizing this, we examined data from Google Places (GP) and developed methods to establish cultural signatures of urban areas. The proposals were then assessed for their effectiveness in cities worldwide and all states in the United States. We obtained evidence that the proposed approaches, even a simple one based on *Frequency*, could capture the cultural character of geographic areas. We gathered evidence based on a comparison with survey data that one of the approaches, based on the Scenes Theory, could capture better cultural nuances.

Unlike other approaches that require proxy data on user preferences, such as user check-ins, our approach only requires simple data—namely, venue categories—- which are easily obtainable from GP for almost any urban area. In addition to evidence at the city and state levels, we explored area divisions based on hexagons with varying granularity levels (6, 7, and 8). Beyond validating the proposed mapping using previous analyses of Toronto, this work presents experiments that examine different area divisions and their signature clustering to assess the effectiveness of cultural characteristics in Curitiba and Chicago. We also conducted a clustering analysis combining both cities to identify similar areas.

The results provided interpretations that aligned with the cultural characteristics of the regions. Thus, these signatures hold significant potential for identifying cultural similarities between locations and can be applied in various ways to benefit society, such as recommending locations, validating service delivery in near real-time based on cultural criteria, and monitoring the impact of public policies on local culture. Additionally, some cities, such as Curitiba, are rarely studied, making it difficult to find literature that explores their cultural aspects—a gap that this study helps address. We recognize that the application of the methodology in certain areas, mainly in East Asia, can be seen as a limitation of this work, due to the foundations of Scenes Theory in Canadian and North American bases and evaluators, however, possible biases in Scenes scores are outside the scope of this work.

Future research can build upon the methodology presented here and extend its application to other cities using Google Places data, further enhancing the validity and generalizability of the findings. Additionally, the methodology could 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.

REFERENCES

ARRIBAS-BEL, D.; FLEISCHMANN, M. Spatial signatures-understanding (urban) spaces through form and function. **Habitat International**, Elsevier, v. 128, p. 102641, 2022.

BANCILHON, M. *et al.* Streetonomics: Quantifying culture using street names. **Plos one**, Public Library of Science, v. 16, n. 6, p. e0252869, 2021.

BRITO, S. A. de *et al.* Cheers to untappd! preferences for beer reflect cultural differences around the world. *In*: **Proc. of AMCIS'18**. New Orleans, USA: [*s.n.*], 2018.

CANDIPAN, J. *et al.* From residence to movement: The nature of racial segregation in everyday urban mobility. **Urban Studies**, SAGE Publications Sage UK: London, England, p. 0042098020978965, 2021.

ÇELIKTEN, E.; FALHER, G. L.; MATHIOUDAKIS, M. Modeling urban behavior by mining geotagged social data. **IEEE Transactions on Big Data**, IEEE, v. 3, n. 2, p. 220–233, 2016.

CHEN, W. *et al.* Large-scale urban building function mapping by integrating multi-source web-based geospatial data. **Geo-spatial Information Science**, v. 27, n. 6, p. 1785–1799, 2024.

CLARK, T. N.; SILVER, D. Chicago from the political machine to the entertainment machine. *In*: **The Politics of Urban Cultural Policy**. [*S.l.*]: Routledge, 2012. p. 28–41.

EINOLA, K.; ALVESSON, M. Behind the numbers: Questioning questionnaires. **Journal of Management Inquiry**, Sage publications Sage CA: Los Angeles, CA, v. 30, n. 1, p. 102–114, 2021.

FALHER, G. L.; GIONIS, A.; MATHIOUDAKIS, M. Where is the soho of rome? measures and algorithms for finding similar neighborhoods in cities. *In*: **Proc. of the ICWSM'15**. Oxford, UK: [*s.n.*], 2015.

FURNO, A. *et al.* A tale of ten cities: Characterizing signatures of mobile traffic in urban areas. **IEEE Transactions on Mobile Computing**, IEEE, v. 16, n. 10, p. 2682–2696, 2016.

GOFFMAN, E. **Frame analysis: An essay on the organization of experience.** [*S.l.*]: Harvard University Press, 1974.

GOGISHVILI, D.; MÜLLER, M. Culture goes east: Mapping the shifting geographies of urban cultural capital through major cultural buildings. **Urban Studies**, SAGE Publications Sage UK: London, England, p. 00420980241289846, 2024.

GUBERT, F. *et al.* Criação de assinatura cultural de Áreas urbanas com estabelecimentos geolocalizados na web. *In*: **Anais do VIII Workshop de Computação Urbana**. Porto Alegre, RS, Brasil: SBC, 2024. p. 127–140. ISSN 2595-2706. Disponível em: https://sol.sbc.org.br/index.php/courb/article/view/29995.

GUBERT, F.; SILVA, T. Google places enricher: A tool that makes it easy to get and enrich google places api data. *In*: **Proc. of WebMedia'22, Extended Proceedings**. Curitiba, PR, Brasil: SBC, 2022. p. 91–94. ISSN 2596-1683.

GUBERT, F. R. *et al.* Culture fingerprint: Identification of culturally similar urban areas using google places data. *In*: SPRINGER. **International Conference on Advances in Social Networks Analysis and Mining**. [*S.l.*], 2024. p. 286–297.

HAUKE, J.; KOSSOWSKI, T. Comparison of values of pearson's and spearman's correlation coefficients on the same sets of data. **Quaestiones geographicae**, De Gruyter Poland, v. 30, n. 2, p. 87, 2011.

HEATH, D. B. **International handbook on alcohol and culture**. [*S.l.*]: Bloomsbury Publishing USA, 1995.

HIDALGO, C. A.; CASTAÑER, E.; SEVTSUK, A. The amenity mix of urban neighborhoods. **Habitat International**, Elsevier, v. 106, p. 102205, 2020.

HU, L.; LI, Z.; YE, X. Delineating and modeling activity space using geotagged social media data. **Cartography and Geographic Information Science**, Taylor & Francis, v. 47, n. 3, p. 277–288, 2020.

ILIEVA, R. T.; MCPHEARSON, T. Social-media data for urban sustainability. **Nature Sustainability**, Nature Publishing Group, v. 1, n. 10, p. 553–565, 2018.

JAEGER, S. R.; CARDELLO, A. V. Factors affecting data quality of online questionnaires: Issues and metrics for sensory and consumer research. **Food Quality and Preference**, Elsevier, v. 102, p. 104676, 2022.

KOLPAK, P.; WANG, L. Exploring the social and neighbourhood predictors of diabetes: a comparison between toronto and chicago. **Prim. health care resear. & devel.**, Cambridge University Press, v. 18, n. 3, p. 291–299, 2017.

LAUFER, P. *et al.* Mining cross-cultural relations from wikipedia: a study of 31 european food cultures. *In*: **Proc. of the ACM WebSci'15**. Oxford, UK: [*s.n.*], 2015. p. 1–10.

LI, J.; SHANG, J.; MCAULEY, J. UCTopic: Unsupervised Contrastive Learning for Phrase Representations and Topic Mining. *In*: **Proc. of the ACL'22**. Dublin, Ireland: ACL, 2022. p. 6159–6169.

MARTÍ, P. *et al.* Revisiting the spatial definition of neighborhood boundaries: Functional clusters versus administrative neighborhoods. **Journal of Urban Technology**, Taylor & Francis, p. 1–22, 2021.

MEHTA, V.; MAHATO, B. Measuring the robustness of neighbourhood business districts. **J. of Urban Design**, Taylor & Francis, v. 24, n. 1, p. 99–118, 2019.

OLSON, A. W. *et al.* Reading the city through its neighbourhoods: Deep text embeddings of yelp reviews as a basis for determining similarity and change. **Cities**, Elsevier, v. 110, p. 103045, 2021.

ORTEGA, Y.; GASSET, J. Apontamentos para uma educação para o futuro. *In*: MADRID: ALIANZA EDITORIAL. **Mission de la Universidad**. [*S.l.*], 1982. p. 225–238.

PATTILLO, M.; LAREAU, A. **Black Picket Fences: Privilege & Peril among the Black Middle Class**. University of Chicago Press, 2013. ISBN 9780226021225. Disponível em: https://books.google.ca/books?id=-R4CAAAAQBAJ.

PRADA, À. G. de la; SMALL, M. L. How people are exposed to neighborhoods racially different from their own. **Proceedings of the National Academy of Sciences**, National Academy of Sciences, v. 121, n. 28, p. e2401661121, 2024.

REIMERS, N.; GUREVYCH, I. Sentence-bert: Sentence embeddings using siamese bert-networks. *In*: **Proc. of the EMNLP'19**. Hong Kong, China: ACL, 2019.

RIVIÈRE, F. *et al.* **Investing in cultural diversity and intercultural dialogue**. [*S.l.*]: Unesco, 2009. v. 2.

ROBSON, K. *et al.* A comparison of factors determining the transition to postsecondary education in toronto and chicago. **Res. in Comp. Inter. Educ.**, v. 14, p. 338–356, 2019.

ROCHA, R. M. A contribuição das imigrações polonesa e germânica para a formação da cidade de curitiba (pr): bairros e endereços que trazem marcas da imigração. **Idéias**, v. 14, 2023.

SANTOS, G. H. *et al.* Redes de interesse: comparando o google places e foursquare na captura da escolha de usuários por áreas urbanas. *In*: SBC. **Workshop de Computação Urbana (CoUrb)**. [*S.l.*], 2024. p. 99–112.

SANTOS, G. H. *et al.* Modeling interest networks in urban areas: A comparative study of google places and foursquare across countries. **Journal of Internet Services and Applications**, v. 16, n. 1, p. 25–42, Mar. 2025. Disponível em: https://journals-sol.sbc.org.br/index.php/jisa/article/view/5152.

SEN, R.; QUERCIA, D. World wide spatial capital. **PloS one**, Public Library of Science San Francisco, CA USA, v. 13, n. 2, p. e0190346, 2018.

SENEFONTE, H. *et al.* Regional influences on tourists mobility through the lens of social sensing. *In*: SPRINGER. **International Conference on Social Informatics**. [*S.l.*], 2020. p. 312–319.

SILVA, T. H. *et al.* A large-scale study of cultural differences using urban data about eating and drinking preferences. **Information Systems**, Elsevier, v. 72, p. 95–116, 2017.

SILVA, T. H.; SILVER, D. Using graph neural networks to predict local culture. **arXiv**, 2024.

SILVA, T. H.; SILVER, D. Using graph neural networks to predict local culture. **Environment and Planning B: Urban Analytics and City Science**, v. 52, n. 2, p. 355–376, 2025. Disponível em: https://doi.org/10.1177/23998083241262053.

SILVA, T. H. *et al.* Urban computing leveraging location-based social network data: A survey. **ACM Comput. Surv.**, ACM, v. 52, n. 1, p. 17:1–17:39, fev. 2019. ISSN 0360-0300.

SILVER, D. A.; CLARK, T. N. **Scenescapes: How qualities of place shape social life**. [*S.l.*]: The University of Chicago, 2016.

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

SPARKS, K. *et al.* A global analysis of cities' geosocial temporal signatures for points of interest hours of operation. **International Journal of Geographical Information Science**, Taylor & Francis, v. 34, n. 4, p. 759–776, 2020.

SPENCER-OATEY, H.; FRANKLIN, P. What is culture. **A compilation of quotations. GlobalPAD Core Concepts**, p. 1–22, 2012.

SPROESSER, G. *et al.* Similar or different? comparing food cultures with regard to traditional and modern eating across ten countries. **Food Research International**, Elsevier, v. 157, p. 111106, 2022.

STUART, F. Ballad of the bullet: Gangs, drill music, and the power of online infamy. Princeton University Press, 2020.

TANG, J. *et al.* Inferring “high-frequent” mixed urban functions from telecom traffic. **Environment and Planning B: Urban Analytics and City Science**, SAGE Publications Sage UK: London, England, v. 51, n. 8, p. 1775–1793, 2024.

VENKATESH, S. **Gang leader for a day: A rogue sociologist takes to the streets**. [*S.l.*]: Penguin, 2008.

VIEZZER, J. *et al.* Áreas verdes, população e renda em curitiba, pr, brasil. **Revista da Sociedade Brasileira de Arborização Urbana**, v. 17, n. 2, p. 37–49, 2022.

WEBER, M. **The Protestant Ethic and the Spirit of Capitalism**. [*S.l.*]: New York: Routledge Classics, 1930.

YAN, A. *et al.* Personalized showcases: Generating multi-modal explanations for recommendations. *In*: **Proc. of the SIGIR'23**. Taipei, Taiwan,: Association for Computing Machinery, 2023. (SIGIR '23), p. 2251–2255. ISBN 9781450394086.

YP. **Yellow Pages**. 2022. https://www.yellowpages.ca/.

ZHANG, Z. *et al.* A deep learning approach for detecting traffic accidents from social media data. **Transportation research part C: emerging technologies**, Elsevier, v. 86, p. 580–596, 2018.