Criação de Assinatura Cultural de Áreas Urbanas com Estabelecimentos Geolocalizados na *Web*

Fernanda R. Gubert[1], Gustavo H. Santos[1], Myriam Delgado[1], Daniel Silver[2], Thiago H. Silva[1]

[1]Universidade Tecnológica Federal do Paraná (UTFPR)
Curitiba, Brasil

[2]University of Toronto
Toronto, Canada

fernandagubert,gustavohenriquesantos@alunos.utfpr.edu.br

Resumo. *O conhecimento a respeito das características dos diferentes grupos culturais que existem no mundo e a identificação de similaridades culturais entre suas respectivas áreas de ocupação podem trazer diversos benefícios econômicos e sociais, como a recomendação de locais sob critérios culturais. Pesquisas referentes ao estudo dessas diferentes culturas são realizadas, em grande parte, de maneira tradicional, as quais são caras e não escalam. Dessa forma, este trabalho consiste em obter características relevantes de áreas urbanas utilizando dados geolocalizados de fontes da web, e aplicar uma metodologia que enriquece esses dados obtidos para a geração de uma assinatura cultural de áreas urbanas. Em uma aplicação prática da proposta, o resultado se mostra muito coerente, separando os bairros de Curitiba em clusters com características culturais distintas.*

Abstract. *Knowledge about the characteristics of different cultural groups that exist in the world and the identification of cultural similarities between their respective areas of occupation can bring several economic and social benefits, such as recommending locations based on cultural criteria. However, much of the research in this domain relies on traditional methods, which are costly and lack scalability. Therefore, this work focuses on extracting pertinent features of urban areas using geolocated data from web sources. Subsequently, a methodology is applied to augment this collected data, generating thereby a cultural profile of urban areas. In practical terms, the outcomes exhibit consistency, as evidenced by the delineation of Curitiba's neighborhoods into culturally distinct clusters.*

1. Introdução

De acordo com o relatório mundial da UNESCO (Organização das Nações Unidas para a Educação, a Ciência e a Cultura) [Rivière et al. 2009] existe uma significativa diversidade cultural no mundo e ter o conhecimento das características dessas diferentes culturas é uma atividade desafiadora. Dentre seus desafios, enfrenta-se o fato da cultura ser uma entidade mutante, ou seja, a sociedade evolui culturalmente conforme o tempo passa e todas as suas características precisam ser revisitadas. As pesquisas relacionadas ao estudo do comportamento de usuários realizadas de formas tradicionais, que tipicamente acontecem através de questionários e entrevistas, possuem algumas limitações principalmente devido ao custo elevado para obter dados de milhões de pessoas, ou seja, tais pesquisas são caras, não escalam e dificilmente podem ser realizadas dentro de um curto espaço de tempo. Por conta disso, muitos estudos recentes utilizam dados provenientes de fontes da *web* para a exploração e resolução de problemas em diversas áreas [Ilieva and McPhearson 2018, Zhang et al. 2018, Hu et al. 2020], alcançando resultados relevantes e de forma mais rápida.

Por outro lado, o conceito de cultura é bastante complexo e não existe uma única definição, não sendo trivial a tarefa de encontrar dados que a descrevam de forma sa-

tisfatória. A cultura pode ser entendida como um conjunto de aspectos de um determinado grupo de pessoas, englobando idioma, religião, culinária e artes, por exemplo [Spencer-Oatey and Franklin 2012]. Alguns estudos mostram que hábitos alimentares e de bebida são elementos capazes de descrever a cultura local [Silva et al. 2017, de Brito et al. 2018, Laufer et al. 2015], porém, dados desse tipo – normalmente *check-ins* de usuários – além de serem difíceis de obter, também dão prioridade analítica aos gostos dos usuários ao invés do estilo de vida evocado pelas características de um local. Outra abordagem segue o discurso de [Mehta and Mahato 2019], no qual a disponibilidade de recursos e serviços, que atendem as necessidades da população, é uma forma de proporcionar um senso de identidade ao local. O que chama a atenção nesta segunda abordagem é a possibilidade de considerar vários aspectos da cultura, já que os recursos de uma cidade, ou seja, seus estabelecimentos, podem estar associados a diferentes categorias, como religião, culinária e artes, além de ser um formato ainda pouco explorado.

Mesmo assim, trabalhar apenas os tipos dos estabelecimentos de uma área urbana, pode não ser o suficiente para a criação de assinaturas culturais [Silva et al. 2017]. O conceito chamado *Scenes* [Silver and Clark 2016] transforma as "cenas" do dia-a-dia da população de um determinado local em elementos de significado cultural, ponderando esses elementos para cada tipo de estabelecimento presente. Dessa forma, o objetivo deste trabalho é mostrar uma estratégia para obter dados de estabelecimentos de uma fonte da *web* de escala global (*Google Places*) e apresentar uma metodologia baseada no conceito de *Scenes*, que demonstra a possibilidade de enriquecimento desses dados para a geração de uma assinatura cultural de áreas urbanas. Um experimento realizado evidencia que a estratégia utilizada tem potencial na identificação de similaridades culturais entre áreas.

A identificação de similaridades culturais e a possibilidade de acompanhar as mudanças mais rapidamente (devido ao processo automático de larga escala) podem beneficiar a habilitação de serviços quase em tempo real, possibilitando que uma empresa, por exemplo, entenda a preferência de seu produto ou serviço em diferentes mercados, e tome decisões de acordo com as informações culturais de diferentes áreas. Este tipo de estudo também pode auxiliar em problemas relacionados à recomendações de locais. Um turista que visitou uma determinada cidade pode receber recomendações para visitar cidades semelhantes sob esse critério cultural, ou ainda, um indivíduo que esteja procurando um local de moradia, pode receber como opção aqueles que mais se assemelham com sua cultura de origem ou de escolha.

O restante deste artigo está organizado da seguinte maneira. A Seção 2 apresenta os trabalhos relacionados e na Seção 3 são descritos os fundamentos da Teoria *Scenes* e como suas diferentes dimensões são aplicadas para a obtenção da assinatura cultural. A Seção 4 descreve a metodologia referente à extração dos dados da API *Google Places* e à ampliação das dimensões que caracterizam os estabelecimentos, seguidas da validação do processo de mapeamento. A Seção 5 mostra uma aplicação prática na cidade de Curitiba. Finalmente, a Seção 6 apresenta a conclusão do artigo.

2. Trabalhos Relacionados

Trabalhos recentes estão se beneficiando de dados provenientes de fontes da *web* para explorar problemas em diversas áreas, inclusive em áreas relacionadas à cultura [de Brito et al. 2018, Silva et al. 2019, Silva and Silver 2024]. [Senefonte et al. 2020]

avaliam como as características regionais e culturais influenciam no comportamento de mobilidade de turistas e residentes. Para o estudo, foram utilizados dados do *Foursquare-Swarm* compartilhados no *Twitter*. Na metodologia proposta, foi construído para cada país um grafo de mobilidade para os residentes e diversos grafos de mobilidade para turistas, dependendo de seus respectivos países de origem. Esta abordagem possibilita analisar o quanto a origem dos usuários influencia em suas escolhas, assim como, o destino escolhido. As transições no grafo ocorrem entre categorias de locais e a matriz que representa o grafo é transformada em um vetor de mobilidade, sendo assim, é possível calcular as distâncias comportamentais, explorando as características culturais de diferentes nacionalidades e em diversos destinos. Os resultados mostram que a origem do turista tem grande influência em seu comportamento, principalmente quando existe uma distância cultural expressiva. Já [Bancilhon et al. 2021] verificaram que uma forma de quantificar a cultura de uma sociedade é através dos nomes das ruas das cidades, após descobrirem que estas refletem o sistema de valores da sociedade. Para isso, foram coletados dados de 4.932 ruas honoríficas (ruas dedicadas a figuras históricas) nas cidades de Paris, Viena, Londres e Nova York, provenientes de fontes públicas. Com este trabalho foi possível detectar a existência de preconceito de gênero, que tem sido mitigado com uma forte tendência recente de nomear novas ruas em homenagem a figuras femininas, quais profissões são consideradas de elite e o quanto uma cidade é influenciada pelo resto do mundo.

A API *Google Places* possui alguns benefícios, como sua ampla cobertura mundial, o que facilita a escalabilidade. Utilizando esses dados, [Sen and Quercia 2018] criaram uma metodologia para medir o capital espacial de um bairro de maneira barata e padronizada, enquanto [Hidalgo et al. 2020] estudaram os padrões de localização de estabelecimentos utilizando dados correspondentes a 47 cidades dos EUA, porém existem alguns desafios ao extrair dados dessa fonte. Promovendo o debate sobre a definição espacial dos limites da vizinhança, [Martí et al. 2021] realizaram um estudo utilizando dados da cidade Alicante, na Espanha, obtidos da *Google Places*. Um dos desafios foi a recategorização dos dados pois existiam muitas categorias semelhantes que dificultavam uma análise mais detalhada. Os autores criaram *clusters* funcionais em termos de atividade urbana, os quais foram na sequência contrapostos com os limites administrativos dos bairros. Como resultado, a pesquisa confirma a existência de uma desconexão entre partições administrativas tradicionais do bairro e a organização funcional da cidade, o que pode ser de grande valor no processo de planejamento urbano.

[Silva et al. 2017] representaram as preferências do usuário quanto aos hábitos alimentares e de bebida, utilizando *check-ins* do *Foursquare*. Através da metodologia proposta é possível identificar fronteiras culturais e semelhanças entre sociedades em diferentes escalas. De forma genérica, esta metodologia consiste em criar um vetor por usuário com valores binários, o qual indica suas preferências, a soma destes vetores caracteriza uma região e, calculando a similaridade do cosseno entre dois vetores de *features*, é possível comparar as regiões. Os resultados espaço-temporais obtidos mostram que têm potencial para explicar hábitos culturais dos usuários e, através de assinaturas culturais, medir a similaridade entre diferentes regiões. Já [Arribas-Bel and Fleischmann 2022] apresentam assinaturas espaciais como uma caracterização do espaço baseada na forma e função de um ambiente urbano, unindo características morfológicas e funcionais para a classificação do espaço.

Como abordado em alguns dos trabalhos relacionados, este estudo também se baseia na caracterização de áreas através dos recursos oferecidos pela cidade e o desenvolvimento de análises comparativas entre elas, mas com foco na criação de assinaturas digitais que permitam a identificação de similaridades culturais. Diferentemente dos trabalhos que levantaram características culturais de regiões utilizando hábitos alimentares e a mobilidade dos usuários, o objetivo deste estudo é obter tais características a partir dos tipos de estabelecimentos presentes em uma cidade, pois acredita-se que, desta forma, é possível abranger mais aspectos culturais, já que independe de ações de usuários, como *check-ins* e avaliações, as quais tendem a acontecer com mais frequência em locais específicos. Outra contribuição é a apresentação de uma metodologia que amplia as dimensões a partir das categorias dos estabelecimentos, enriquecendo a caracterização das regiões geográficas, além de criar as assinaturas culturais a partir de uma fonte de dados que oferece uma cobertura mundial maior, que é a *Google Places*.

3. Teoria *Scenes*

3.1. Fundamentos da teoria e as 15 dimensões

A Teoria *Scenes* busca equilibrar os significados, estilos e estética das características das experiências humanas com a precisão das ciências físicas. É baseada na ideia de combinar elementos culturais para formar as "cenas". Essas combinações podem acontecer de várias maneiras, criando cenas de diferentes momentos históricos e de diversos locais geográficos. Essa teoria é proposta por [Silver and Clark 2016] e é descrita a seguir.

O conceito de cena foca em descrever como, quando, onde e por que determinadas pessoas se reúnem em torno de conjuntos específicos de gostos e atividades culturais, o que vai além de "valores comuns" e "modos de vida" intrínsecos de cada cultura. Para determinar quais são os elementos que caracterizam as cenas, foi adotado um caminho entre a teoria sistemática e o empirismo, considerando uma série de fontes do mundo da cultura, como poesia, religião, jornalismo, pesquisas etnográficas e filosofia.

Na Teoria *Scenes* abordam-se 3 tipos gerais de significado — *theatricality, authenticity* e *legitimacy* — e tais significados existem em várias tradições de pensamento, de Max Weber sobre *legitimacy* [Weber 1930] a Erving Goffman sobre *theatricality* [Goffman 1974] e Georg Simmel sobre *authenticity* [Simmel 1971], entre outros. A *authenticity* avalia como a cena aponta para algo considerado genuíno em vez de falso, a *theatricality* retrata como a cena descreve a apresentação, em suas roupas, fala, maneiras, postura, porte e aparência, já a *legitimacy* estima em que se acredita para tornar as ações certas ou erradas.

Porém, foi percebida a necessidade de analisar a cena através de termos mais específicos que traduzam suas características, estes termos são chamados de dimensões e totalizam 15. Os tipos gerais de significado apresentados acima se relacionam e se complementam, consequentemente, as dimensões também. A seguir são listadas as 15 dimensões por tipo geral de significado, seguidas de uma breve descrição.

- ***Theatricality***: performance, exibição.
  - *Glamour*: dotada de aspectos deslumbrantes, cintilantes e com personagens misteriosos e sedutores.
  - *Neighborliness*: trata-se de amigos e companheiros de camaradagem, reunidos como uma comunidade calorosa e atenciosa.

  - *Transgression*: quebra os estilos convencionais de aparência, contrapondo o que é considerado rotineiro, seja em relação ao comportamento, vestimentas ou boas maneiras.
  - *Formality*: valoriza padrões de vestimenta altamente ritualizados e cerimoniais, além de aspectos da fala e aparência em geral.
  - *Exhibitionism*: o eu torna-se um objeto a ser olhado, uma exposição a ser admirada.

- ***Authenticity***: sobre as fontes do seu ser, de onde vem o "verdadeiro você", as dimensões se expandem do particular para o generalizado.
  - *Locality*: ser pertencente e enraizado neste lugar e somente neste lugar, não "contaminado" por costumes estrangeiros.
  - *Ethnicity*: trata-se de costumes étnicos, com sentimentos profundos, não escolhidos, dotados de práticas originais.
  - *State*: estende características, costumes, ideias e locais do estado para o nacional.
  - *Corporateness*: trata-se da autenticidade de grandes marcas, as quais transcendem estados, regiões e etnias, estabelecendo-se de forma global, sendo genuínas com o que oferecem e reivindicando a fidelidade de muitos.
  - *Rationality*: afirma que o verdadeiro eu está na mente, o exercício espontâneo da razão é mais profundo do que as circunstâncias arbitrárias e externas da localização, etnia ou nacionalidade.

- ***Legitimacy***: diz respeito à base dos juízos morais, à autoridade sobre que um veredicto de certo ou errado é fundamentado, orientada pelo tempo (passado, presente e futuro) e espaço.
  - *Tradition*: o passado é uma autoridade duradoura que se estende ao presente, é a criação de uma conexão com o passado que informa as razões para atuar no aqui e agora.
  - *Charisma*: pode ser traduzida também por carisma e trata-se de uma qualidade indescritível de grandes figuras, como artistas e celebridades, levando outros a segui-los.
  - *Utilitarian*: se baseia no lucro e na produtividade, evoca a importância de uma análise de custo e benefício.
  - *Egalitarian*: consiste no respeito pela igualdade humana, todas as pessoas merecem justiça e igualdade de tratamento.
  - *Self-Expression*: é a expressão da personalidade de um indivíduo, com sua visão única, estilo e ações próprios.

Essas 15 dimensões são como ferramentas para decompor uma cena em uma série de elementos. Outras dimensões podem ser adicionadas, mas estas já compõem um fundamento sólido para capturar o caráter cultural das cenas. Realizando uma tradução das dimensões de significado para categorias de estabelecimentos, pode-se dizer que é o conjunto de diferentes tipos de estabelecimentos que formam uma cena e esse conjunto passa a ser um indicador chave para medir a cena. Dessa forma, cria-se uma visão mais holística, pois um mesmo estabelecimento pode assumir significados diferentes, mostrando também que um estabelecimento não faz uma cena em particular.

3.2. Sistema de pontuação das dimensões

Com o objetivo de traduzir dados aparentemente não culturais em fontes de informação sobre o significado cultural, [Silver and Clark 2016] trabalharam com uma equipe de codificadores que auxiliou na atribuição dos pesos das dimensões a todos os tipos de estabelecimentos presentes em seu banco de dados, provenientes do NAICS (Sistema de Classificação Industrial da América do Norte) e YP (Páginas Amarelas). O NAICS é um sistema de classificação norte americano, mantido pelo governo, que inclui todos os tipos de indicadores úteis para compor as "cenas" locais, como organizações religiosas, galerias de arte, organizações ambientais e muito mais. Já o YP disponibiliza de forma *online* dados de negócios, produtos e serviços do Canadá, em uma plataforma chamada *Yellow Pages* [YP 2022].

De acordo com [Silver and Clark 2016], os codificadores receberam diversas instruções, realizando uma imersão no projeto e utilizando um tutorial, além do uso de um manual chamado "*The Coder's Handbook*" com uma série de perguntas padronizadas a serem feitas na codificação de cada dimensão e armadilhas comuns, bem como uma série de exemplos com justificativas. Os codificadores focavam sua atenção em uma dimensão de cada vez, tornando-se mais fácil realizar análises comparativas entre os diferentes tipos de estabelecimentos. Este processo de tradução durou cerca de um ano, com dezenas de reuniões que levaram a repetidas revisões e esclarecimentos, até encontrar um consenso para os casos em que havia divergência na pontuação.

O sistema de pontuação foi construído para garantir que o procedimento fosse realizado de forma clara e padronizada, direcionando a tomada de decisão em cada caso. Cada estabelecimento recebeu uma pontuação de 1 a 5 em cada uma das 15 dimensões. As pontuações 4 e 5 indicam que o estabelecimento afirma a dimensão. Pontuações 1 e 2 indicam que o estabelecimento rejeita a dimensão. Um ponto de 3 indica que as práticas do estabelecimento são neutras na dimensão. A decisão mais importante é entre uma pontuação positiva (4 ou 5) ou negativa (1 ou 2). Os codificadores reservaram pontuações extremas (5 e 1) para os casos em que o rótulo de um estabelecimento aparece de forma clara e indica (ou não) diretamente uma determinada dimensão como parte central de seu significado. As notas 4 e 2 foram para os casos em que o estabelecimento pode muitas vezes ou às vezes indicar uma postura positiva ou negativa em relação à dimensão. É importante ressaltar que as pontuações das dimensões não são classificações, mas sim ferramentas para discernir os tipos de experiências que se sobressaem em cada local, demostrando sua sensibilidade para capturar a experiência geral promovida por todos os estabelecimentos que compõem uma cena.

Essas bases de dados enriquecidas com as pontuações das dimensões foram analisadas com outros domínios sociais importantes no trabalho de [Silver and Clark 2016]. Foi investigada a contribuição das cenas para o crescimento econômico e a prosperidade, em como se relacionam com os padrões residenciais, além de mostrar como a votação e outras atividades políticas variam consideravelmente de acordo com o contexto local. Os resultados reforçam o conteúdo significativo carregado pelas cenas e a qualidade da tradução de tipos de estabelecimentos para as dimensões da teoria. Dessa forma, as pontuações atribuídas a inúmeras categorias de estabelecimentos tornaram-se sementes para o mapeamento de outros conjuntos de dados.

3.3. Assinatura cultural

Analisar um conjunto pequeno de estabelecimentos pode dar uma impressão enganosa do significado da cena geral e, se analisar um conjunto pequeno já é uma tarefa difícil de ser realizada manualmente, torna-se inalcançável quando se pensa em escalar. Ter uma medida da cena é uma resposta para esse problema.

Com o sistema de pontuação, cada estabelecimento recebe, por fim, um ou mais vetores com 15 dimensões, de acordo com a(s) categoria(s) associada(s). Dessa forma, o modelo padrão de cenas é a matriz na qual as linhas são os estabelecimentos e as colunas são as 15 dimensões, posicionando cada estabelecimento dentro de um complexo de significado cultural. Para medir a cena de uma região é calculada a média entre os valores de cada dimensão, considerando todos os vetores de estabelecimentos presentes.

Este resultado representado por um único vetor de 15 dimensões é chamado de pontuação de desempenho, e é o que representa a assinatura cultural proposta por este trabalho.

A assinatura cultural permite catalogar e comparar tantas cenas quanto possível, sem se limitar a recursos para visitar tais locais. Além de compor uma comparação que considera todos os detalhes capturados e seu contexto, o que feito de forma manual pode acarretar em esquecimentos e interpretações isoladas. Com o objetivo de incluir mais características às áreas urbanas para a criação da assinatura cultural, este trabalho propõe o mapeamento das categorias de estabelecimento existente no conjunto de dados recuperado da *Google Places* para as 15 dimensões apresentadas na Teoria *Scenes*. Este mapeamento vai propiciar que as áreas possam ser analisadas como cenas, já que englobarão um conjunto de diversos tipos de estabelecimentos que fornecem diferentes dimensões de significado.

4. Metodologia

4.1. Extração dos dados da *Google Places*

A API *Google Places* é um serviço que retorna dados geolocalizados de estabelecimentos e pontos de interesse. Além da localização em pares de latitude e longitude, os locais são associados pelo menos a uma categoria, com o objetivo de descrever o tipo do estabelecimento, ao todo existem 141 categorias. Porém, estas categorias não possuem o nível de especificidade necessário para viabilizar a criação das assinaturas culturais posteriormente. Por exemplo, a API fornece a categoria *restaurant* aos estabelecimentos que se classificam como tal, mas não fornece uma categoria mais específica sobre o tipo gastronômico, como italiano ou japonês, sendo insuficiente para a proposta deste trabalho.

Para contornar este problema foi utilizado o parâmetro opcional *keyword* nas chamadas. O serviço *Google Places* busca o texto deste parâmetro por todo o conteúdo indexado dos estabelecimentos, retornando as correspondências ordenadas com base na relevância percebida. Mesmo não sendo um parâmetro específico para a busca de tipos de estabelecimentos, a documentação da API garante retornar resultados válidos se as entradas forem um nome de local, endereço ou categoria de estabelecimentos, tornando-se assim uma opção conveniente para os fins desejados. Foram utilizadas as categorias da base *Yelp*, por fornecerem o nível de detalhe necessário. A base *Yelp* consiste em *reviews* de estabelecimentos feitos por usuários que visitaram tais locais e, assim como na *Google Places*, são disponibilizadas as categorias dos estabelecimentos. As categorias do *Yelp* possuem uma estrutura hierárquica de 4 níveis, fazendo sentido para este trabalho adotar apenas as do último nível. Algumas categorias ainda foram excluídas por não se apresentarem relevantes ao nosso propósito, resultando em 888.

Como cada requisição para essa API é realizada a partir de um par de coordenadas geográficas e um raio em metros, é necessária uma estratégia para definir estes pares em cada cidade do estudo, assim como, o nível de detalhes ideal para possibilitar a criação da assinatura cultural. Primeiramente, foram estabelecidos os extremos nordeste e sudoeste para delimitar o retângulo que engloba toda uma cidade, como mostra a Figura 1. Em seguida, foram criadas subáreas quadradas e recuperado o par de coordenadas geográficas central de cada uma. Após a definição automática das coordenadas, era necessário um ajuste manual, já que o retângulo pré-definido possui áreas fora da cidade de interesse.

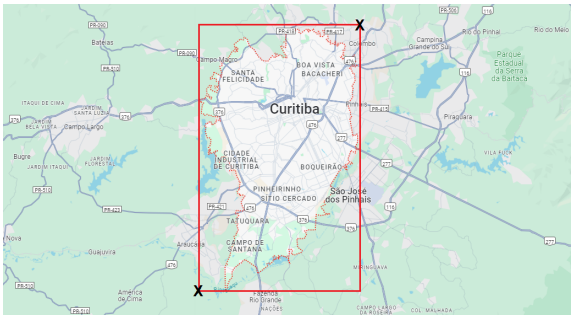

Figura 1. Área retangular delimitada para a cidade de Curitiba.

Considerando os custos envolvidos (lembrando que é necessária uma chamada por categoria em cada um dos pares de coordenadas) e a obtenção de uma quantidade significativa de dados, foi concluída a utilização de um raio de 6.000m em coordenadas distribuídas pela cidade, cobrindo toda a região, e uma coordenada na região central com raio de 3.000m. A justificativa para o uso da coordenada central com raio menor é a existência da limitação na quantidade de estabelecimentos retornados por requisição (20 estabelecimentos com possibilidade de paginação resultando em no máximo 60), dessa forma é possível minimizar a perda de estabelecimentos em regiões mais densas. Optou-se por utilizar tamanhos de raio padrão em todas as cidades ao invés de padronizar o número de coordenadas pois o tamanho das cidades pode variar bastante e, ao utilizar a mesma quantidade de coordenadas, obteríamos mais estabelecimentos por subárea de uma cidade em relação à outra, podendo influenciar na criação das assinaturas culturais e não possibilitar uma comparação justa entre as cidades. Após a extração, não foram detectados dados faltantes mas foi percebida a existência de registros duplicados, em torno de 30%, o que já era esperado pois algumas pequenas áreas acabam tendo sobreposição. Esses procedimentos resultaram em uma ferramenta para facilitar o processo[1] [Gubert and Silva 2022].

4.2. Mapeamento das categorias para as dimensões da Teoria *Scenes*

As categorias presentes nos dados recuperados da *Google Places* precisavam ser mapeadas para as 15 dimensões da Teoria *Scenes* e, para isso, foi utilizado como base e inspiração o mapeamento já realizado com as categorias presentes nos dados do *Yelp* [Silver and Silva 2021]. A Figura 2 resume todo o processo de mapeamento, o qual é descrito a seguir. Para as categorias do *Yelp*, o processo foi realizado a partir das "sementes" da Teoria *Scenes*, que tratam do resultado da pontuação manual descrita na Subseção 3.2, em que diversas categorias receberam pontuações para as 15 dimensões. Sendo assim, utilizando uma comparação semântica das categorias dos dados do *Yelp* com as "sementes" da Teoria *Scenes*, foi possível estabelecer os pesos para as dimensões de cada categoria presente na base [Silver and Silva 2021].

Ao analisar os dados obtidos da *Google Places*, constatou-se que para uma descrição mais eficaz dos estabelecimentos seria necessário utilizar tanto as categorias selecionadas do *Yelp*, que foram utilizadas nas requisições, quanto as categorias mais abrangentes disponibilizadas pela própria *Google*. Foi observado que duas das 141 categorias da *Google* estavam associadas a 99% dos dados e não forneciam descrição do tipo do estabelecimento, estas categorias são *point of interest* e *establishment*, portanto,

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

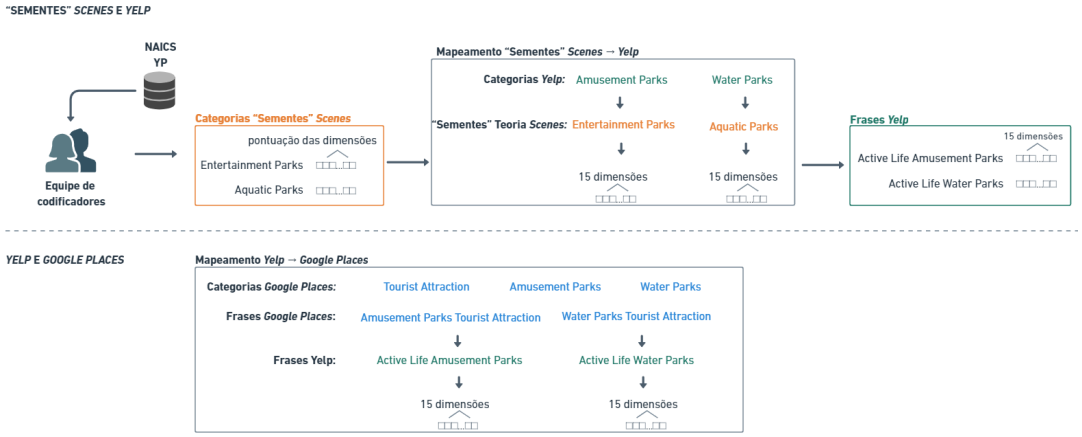

Figura 2. Processo de mapeamento das categorias para a Teoria *Scenes*, considerando o mapeamento das "sementes" *Scenes* para o *Yelp*, e da base *Yelp* para a *Google Places*.

foram removidas. Aproveitando-se da existência do mapeamento dos dados do *Yelp* com as "sementes" da Teoria *Scenes* e como os dados da *Google* já foram enriquecidos com algumas categorias do *Yelp*, optou-se por realizar o mapeamento diretamente com o *Yelp*. Para aumentar a capacidade semântica e consequentemente a precisão do mapeamento, foram criadas frases para cada estabelecimento. Estas frases foram compostas de uma categoria do *Yelp* com todas as categorias da *Google* existentes para aquele estabelecimento. Ou seja, se um estabelecimento possui as categorias *Amusement Parks* e *Water Parks* do *Yelp*, e a categoria *Tourist Attraction* da *Google*, as frases são:

- *Amusement Parks Tourist Attraction*
- *Water Parks Tourist Attraction*

Como dito anteriormente, as categorias do *Yelp* são dispostas em uma estrutura hierárquica de 4 níveis. Com a mesma intenção de aumentar a capacidade semântica, as frases do *Yelp* foram criadas utilizando todos os níveis, ou seja, para cada categoria do último nível, a frase associada consiste em todo o caminho a partir do primeiro nível. Como exemplo, foi adicionado *Active Life* para formar as frases do *Yelp* na Figura 2.

O processo do mapeamento foi realizado com o SBERT, utilizando o *framework Sentence Transformers*, no qual vários modelos pré-treinados com um grande e diversificado conjunto de dados de mais de 1 bilhão de pares de treinamento são disponibilizados e podem ser utilizados para calcular *embeddings* a partir de frases e textos para mais de 100 idiomas [Reimers and Gurevych 2019]. Após selecionar alguns modelos que se aplicavam aos nossos propósitos utilizando a documentação disponibilizada, foram realizadas experimentações e análises de amostras dos resultados para selecionar o modelo que apresentou maior eficácia nas comparações semânticas, dado o nosso contexto de frases. Para comparar os *embeddings* gerados foi calculada a similaridade do cosseno e, para cada frase relacionada aos estabelecimentos, foi recuperada a frase do *Yelp* com maior *score*. Analisando diferentes amostras deste resultado, foi verificado que as frases compostas por apenas uma categoria do *Yelp* constituída de uma única palavra, como por exemplo *German*, não obtiveram um bom mapeamento, justamente pela falta de significado. Estes casos correspondem em média a 5% dos dados e foram excluídos. Com este mapeamento, cada estabelecimento foi associado a um ou mais de um vetor contendo as 15 dimensões da Teoria *Scenes*, dependendo da quantidade de frases associada. Vale ressaltar que os

vetores entram com o mesmo peso para o estabelecimento, independente das categorias que compõem as frases. Para exemplificar o resultado final do mapeamento realizado para os dados da *Google Places*, a Tabela 1 mostra 3 frases associadas a estabelecimentos recuperados da API com seus respectivos pesos para cada dimensão.

Tabela 1. Exemplos de frases mapeadas para as dimensões da Teoria *Scenes*.

	Skin Care Store	*Hot Dogs Restaurant Food*	*Karaoke Bar*
Theatricality			
Glamour	4	1	3,25
Neighborliness	4	1,8	3,4
Transgression	3	3	3
Formality	3	2,6	3
Exhibitionism	3	2,8	4,2
Authenticity			
Locality	3	1	3
Ethnicity	3	3	3
State	3	3	3
Corporateness	3	4,75	3
Rationality	2	3	1,75
Legitimacy			
Tradition	3	3	3
Charisma	4	2,6	3,4
Utilitarian	2	4,8	1,6
Egalitarian	3	3,4	3
Self-Expression	4	2,4	4

4.3. Validação do processo de mapeamento

A validação do processo de mapeamento descrito na seção anterior foi realizada com os dados de Toronto. A cidade foi dividida nas regiões chamadas FSA (Área de Classificação Direta), que correspondem à unidades geográficas com base nos três primeiros caracteres de um código postal canadense, ao todo são 99 regiões. Cada uma dessas regiões foi tratada como uma "cena" e, portanto, mapeada para as 15 dimensões através da assinatura cultural criada a partir dos dados extraídos. Em seguida, foi calculada a correlação de *Pearson* e de *Spearman* entre os valores das dimensões obtidos neste trabalho e os valores das dimensões de um mapeamento pré-existente para essas regiões (NAICS e YP), disponíveis na literatura por [Silver and Clark 2016]. Assim como a correlação calculada com os dados do *Yelp*, estes obtidos em [Silver and Silva 2021]. Além dessas fontes já fornecerem subsídios para trabalhar com as regiões FSA e terem sido validadas como confiáveis, [Silver and Clark 2016] escolheram trabalhar com essas unidades geográficas, ao invés de estados ou municípios, pelo fato de serem pequenas o suficiente e com alto nível de precisão, com milhares de categorias disponíveis para classificação.

O coeficiente de correlação de *Pearson* é uma medida da relação linear entre duas variáveis com distribuição normal, e o mais esperado entre os dados da *Google* e as outras bases é justamente uma relação linear, neste caso, positiva. Em contrapartida, a análise de uma estatística de classificação não paramétrica, que avalia o relacionamento entre duas variáveis descritas por uma função monotônica arbitrária, também faz sentido, visto que as dimensões podem ter comportamentos distintos e as bases podem facilmente não apresentar as mesmas categorias em cada região, justificando o uso de ambas as correlações [Hauke and Kossowski 2011]. O resultado é mostrado na Figura 3.

Exceto *Tradition* e *Egalitarian*, todas as outras dimensões resultaram em

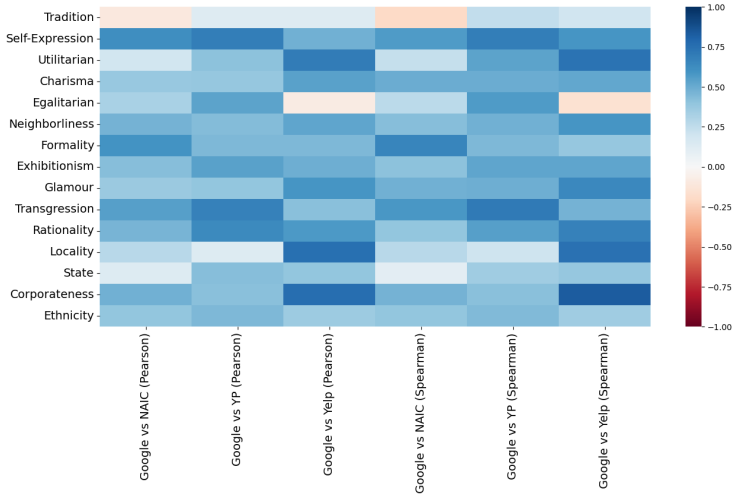

Figura 3. Resultado das correlações de *Pearson* (três primeiras colunas) e *Spearman* (três últimas colunas), calculadas entre os dados da *Google* e as bases YP, NAICS e *Yelp*.

correlações positivas para as 3 bases, em especial a YP, o que evidencia um bom resultado de forma geral. Observando as frases mapeadas, a fim de investigar as correlações mais fracas e as negativas, foram detectados alguns mapeamentos não muito coerentes relacionados à categoria *Arts & Crafts*, porém em uma quantidade pequena, o que possibilitou refazê-los manualmente. Dada essa análise e os resultados das correlações obtidos com a base YP, concluiu-se que o processo do mapeamento é válido para garantir a criação de assinaturas culturais confiáveis com os dados da *Google*.

5. Aplicação prática

Para demonstrar uma aplicação prática da metodologia proposta neste trabalho, foram coletados dados da cidade de Curitiba, que foi escolhida por ser a de maior conhecimento dos autores e possibilitar uma validação mais criteriosa dos resultados, totalizando 31.539 estabelecimentos e 748 categorias únicas, além de utilizar 5 coordenadas geográficas nas requisições para cobrir a área.

As análises foram realizadas em nível de bairro. Um total de 11 bairros foi desconsiderado visto que estes possuíam menos de 100 estabelecimentos - foi observado que nestes casos não havia muita diversidade, por se tratarem de zonas residenciais ou rurais, e isto poderia prejudicar o cálculo da assinatura cultural e as análises subsequentes. Em seguida, foi calculada a assinatura cultural para os demais, totalizando 64 bairros. Ao realizar o Agrupamento Hierárquico Aglomerativo utilizando como *features* as 15 dimensões da Teoria *Scenes*, obteve-se o dendrograma da Figura 4. A estratégia utilizada para mesclar os pares de *clusters* foi *ward*, tendo como métrica a distância Euclidiana para calcular essa ligação. O corte para determinar o número de *clusters* foi realizado no agrupamento com a segunda maior distância, já que considerar apenas dois *clusters* (maior distância) não faz muito sentido para o contexto analisado. Dessa forma, foram obtidos 4 *clusters*, os quais foram analisados e interpretados a seguir.

O *cluster* 1 é o maior de todos, com 31 bairros. As regionais predominantes são Bairro Novo, Boqueirão, Pinheirinho e Tatuquara, concentrando-se em regiões mais afastadas do Centro e ao sul da cidade. Em geral, os locais são caracterizados pela presença de

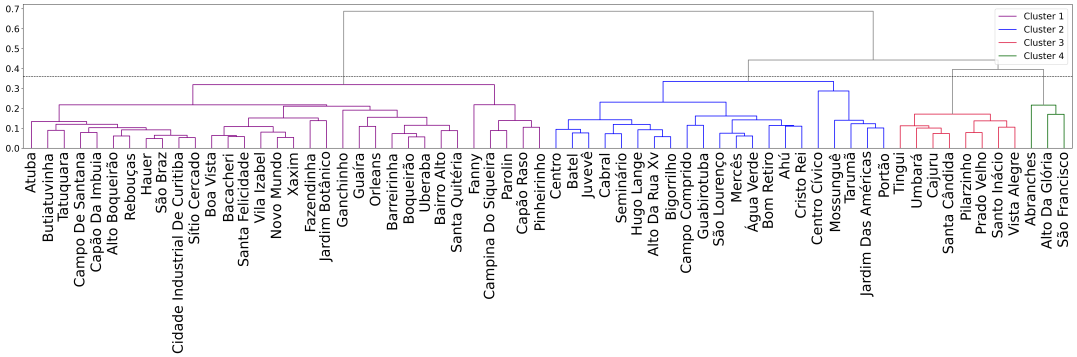

Figura 4. Dendrograma representando o resultado do Agrupamento Aglomerativo de Curitiba.

museus, parques, praças, ruas arborizadas, além de atrações noturnas, como bares e clubes. O *cluster* 2 possui 19 bairros com predominância da regional Matriz, caracterizada por ser o centro comercial, com regiões que lideram os índices econômicos da cidade. A maior representatividade está nos setores de comércio varejista e de serviços, como alimentação, bebida, escritório e apoio administrativo. Por sua vez, o *cluster* 3 apresenta 11 bairros, que se encontram nos arredores do Centro. São regiões com boa infraestrutura de comércio e lazer, além de possuir parques com área verde extensa. Por último, o *cluster* 4 possui apenas 3 bairros, sendo eles: Abranches, Alto da Glória e São Francisco. Em relação à localização geográfica, Alto da Glória e São Francisco ficam próximos ao Centro, enquanto Abranches está um pouco mais afastado porém na região norte da cidade também. São Francisco tem características peculiares, conhecido por ser o bairro mais *"cool"* de Curitiba, repleto de bares, *pubs* casuais com shows de *rock*, hamburguerias, restaurantes árabes e um mercado dominical chamado Feira do Largo da Ordem, com barracas comercializando comida de rua e artesanato. Alto da Glória por ser geograficamente próximo pode ter sido influenciado por algumas das características de São Francisco, é também onde encontra-se o Estádio Couto Pereira. Em Abranches estão o Parque das Pedreiras e a Ópera de Arame, com atrações de música e teatro. Além da Pedreira Paulo Leminski, com apresentações de grandes artistas nacionais e internacionais.

Ao verificar as 15 dimensões da Teoria *Scenes* por *cluster*, pôde-se identificar várias semelhanças com a análise descritiva realizada anteriormente. Na Figura 5 é possível comparar os valores das dimensões. Em suma, o *cluster* 2 possui valores mais altos nas dimensões *Tradition* e *Corporateness*. Em comparação com os outros, o *cluster* 1 se sobressai em *Utilitarian*, *Transgression*, *Rationality* e *Corporateness*, sendo um dos *clusters* com menor valor em *Formality*. Também de forma comparativa, o *cluster* 4 possui o valor mais alto em *Tradition*, *Self-Expression*, *Charisma*, *Neighborliness*, *Formality*, *Glamour*, *Locality* e *Ethnicity*. E possui o valor mais baixo em *Utilitarian*, *Egalitarian*, *Transgression*, *Rationality* e *Corporateness*. Já o *cluster* 3 se sobressai em *Egalitarian* e *Exhibitionism*. O resumo das características dos *clusters*, assim como, as dimensões de destaque, podem ser encontrados na Figura 6.

6. Conclusão

A obtenção de características culturais em larga escala é uma atividade desafiadora. Sabendo disso, este trabalho explorou uma metodologia capaz de obter características

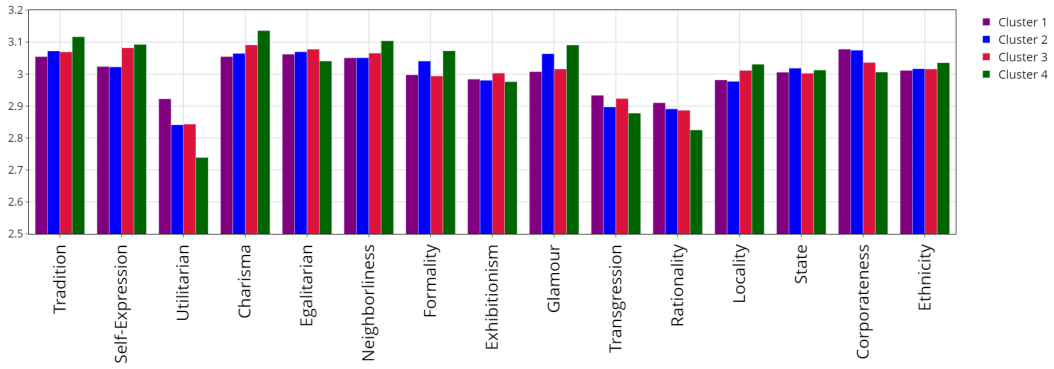

Figura 5. Valores das dimensões por *cluster* da cidade de Curitiba.

Figura 6. Resumo dos *clusters* da cidade de Curitiba.

relevantes de áreas geográficas que permitissem a criação de assinaturas digitais e a identificação de similaridades culturais entre essas áreas. Para os experimentos, foram utilizadas fontes da *web* para justamente ultrapassar as limitações inerentes ao processo clássico baseado em questionários e entrevistas.

A metodologia proposta consistiu em mesclar frases obtidas de duas bases de dados diferentes (*Yelp* e *Google*) para cada estabelecimento considerado e depois mapear (através da Teoria *Scenes)* essas frases para um espaço de dimensão 15 que representa a respectiva cena. Para cada região geográfica (foram consideradas regiões nas cidades de Toronto e Curitiba), um único vetor de cena foi obtido a partir da média dos vetores dos estabelecimentos localizados na região. Além de validar o mapeamento proposto para a cidade de Toronto tendo por base trabalhos prévios da literatura, foi apresentada neste artigo, uma aplicação prática com o objetivo de estudar a efetividade das assinaturas culturais na cidade de Curitiba.

O resultado trouxe interpretações que se mostraram condizentes com as características culturais das regiões. Com isso, pode-se dizer que existe um grande potencial de utilizar essas assinaturas para identificar similaridades culturais entre locais e aplicá-las de forma a trazer diversos benefícios para a sociedade, como a recomendação de locais e a validação de habilitação de serviços em quase tempo real sob critérios culturais. Trabalhos futuros podem facilmente aplicar a metodologia aqui apresentada e expandir o

estudo para outras cidades com os dados da *Google Places*, o que poderia fortalecer as conclusões. Outra possibilidade é replicar a metodologia para dados de outras fontes, criando uma abordagem mais diversificada na coleta de dados, de acordo com a necessidade.

Agradecimentos

Este trabalho foi parcialmente apoiado pelo projeto SocialNet (processo 2023/00148-0 da Fundação de Amparo à Pesquisa do Estado de São Paulo - FAPESP) e CNPq (processo 314603/2023-9 e 441444/2023-7).

Referências

Arribas-Bel, D. and Fleischmann, M. (2022). Spatial signatures-understanding (urban) spaces through form and function. Habitat International, 128:102641.

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

de Brito, S. A., Baldykowski, A. L., Miczevski, S. A., and Silva, T. H. (2018). Cheers to untappd! preferences for beer reflect cultural differences around the world. In Proc. of AMCIS'18, New Orleans, USA.

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

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

Hauke, J. and Kossowski, T. (2011). Comparison of values of pearson's and spearman's correlation coefficients on the same sets of data. Quaestiones geographicae, 30(2):87.

Hidalgo, C. A., Castañer, E., and Sevtsuk, A. (2020). The amenity mix of urban neighborhoods. Habitat International, 106:102205.

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

Ilieva, R. T. and McPhearson, T. (2018). Social-media data for urban sustainability. Nature Sustainability, 1(10):553–565.

Laufer, P., Wagner, C., Flöck, F., and Strohmaier, M. (2015). Mining cross-cultural relations from wikipedia: a study of 31 european food cultures. In Proceedings of the ACM Web Science Conference, pages 1–10.

Martí, P., Serrano-Estrada, L., Nolasco-Cirugeda, A., and Baeza, J. L. (2021). Revisiting the spatial definition of neighborhood boundaries: Functional clusters versus administrative neighborhoods. Journal of Urban Technology, pages 1–22.

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

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

Rivière, F. et al. (2009). Investing in cultural diversity and intercultural dialogue, volume 2. Unesco.

Sen, R. and Quercia, D. (2018). World wide spatial capital. PloS one, 13(2):e0190346.

Senefonte, H., Frizzo, G., Delgado, M., Lüders, R., Silver, D., and Silva, T. (2020). Regional influences on tourists mobility through the lens of social sensing. In International Conference on Social Informatics, pages 312–319. Springer.

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

Silva, T. H. and Silver, D. (2024). Using graph neural networks to predict local culture. arXiv.

Silva, T. H., Viana, A. C., Benevenuto, F., Villas, L., Salles, J., Loureiro, A., and Quercia, D. (2019). Urban computing leveraging location-based social network data: A survey. ACM Comput. Surv., 52(1):17:1–17:39.

Silver, D. and Silva, T. H. (2021). Complex causal structures of neighbourhood change: Evidence from a functionalist model and yelp data.

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

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

Spencer-Oatey, H. and Franklin, P. (2012). What is culture. A compilation of quotations. GlobalPAD Core Concepts, pages 1–22.

Weber, M. (1930). The Protestant Ethic and the Spirit of Capitalism. New York: Routledge Classics.

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

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