A *Vibe* da Cidade: Extração e Estudo de Assinaturas de Categorias de Estabelecimentos Comerciais

Leonardo de Assis da Silva[1], **Thiago H. Silva**[1]

[1]Departamento Acadêmico de Informática
Universidade Tecnológica Federal do Paraná (UTFPR)
Avenida Sete de Setembro – 3165 – 80.230-901 – Curitiba – PR – Brasil

leosil@alunos.utfpr.edu.br, thiagoh@utfpr.edu.br

Abstract. *Urban computing is a field of study that among others objectives aims to help understand urban phenomenon envisioning to offer smarter urban services. Thus, an important aspect is the comprehension of functioning dynamics of businesses in the city. Performing this comprehension through time allows us, for instance, to use this information as a business descriptor that could be explored in new services. In this study, we collected and used a significant amount of data for business related to consumption of food and beverage in different cities in Brazil and the United States. Our main contributions are: (1) clustering and analysis of the collected time series representing the functioning dynamics of business in the city; (2) approach for identifying the signature that represents the behavior of certain categories of venues; (3) training and evaluation of an inference model for categories of establishments.*

Resumo. *A computação urbana é uma área de estudo que visa, dentre outros objetivos, auxiliar no entendimento de fenômenos urbanos visando oferecer serviços urbanos mais inteligentes. Nesse sentido, um importante aspecto é o entendimento da dinâmica de funcionamento de estabelecimentos comerciais na cidade. Realizar esse entendimento ao longo do tempo permite, por exemplo, empregar essa informação como um descritor de estabelecimentos, o que pode ser explorado em novos serviços. Neste estudo nós coletamos e exploramos uma quantidade significativa de dados para estabelecimentos relacionados com o consumo de comida e bebida em diferentes cidades no Brasil e nos Estados Unidos. Nossos principais resultados podem ser agrupados em: (1) agrupamento e análise das séries temporais representando a dinâmica de funcionamento de estabelecimentos comerciais na cidade; (2) abordagem para identificar a assinatura que representa a dinâmica de funcionamento para determinadas categorias de locais; (3) treinamento e avaliação de modelo de inferência de categoria.*

1. Introdução

Uma determinada categoria de estabelecimento possui horários de pico de popularidade que são ditados não apenas pelo serviço oferecido, mas também devido à diversas causas econômicas, sociais e culturais, como horário típico de entrada e saída de trabalhadores, período de sesta e *happy hour*. Além disso, estabelecimentos de mesma categoria também podem eventualmente exibir diferentes picos de popularidade de acordo com sua

localização ou características particulares do tipo de estabelecimento. Enquanto alguns tipos de restaurantes geralmente são mais populares durante a noite, restaurantes do tipo *fast-food* apresentam movimento mais uniformemente distribuído ao longo do dia.

Dessa forma, conhecer quais são os padrões popularidade de cada categoria de estabelecimento não é uma tarefa trivial, porém uma estratégia viável seria tentar identificá-los a partir da análise de uma base significativa de registros de visitas de uma amostra de estabelecimentos. Entretanto, a identificação de padrões confiáveis requerem a coleta de dados em larga escala. Como a principal variável necessária na identificação do comportamento de cada estabelecimento seria o número de visitas a cada hora, uma boa maneira de obter esta informação é através do rastreamento de aparelhos como *smartphones* e *notebooks* portados pelos visitantes no interior do estabelecimento. Esta abordagem, frequentemente chamada de sensoriamento social, consiste na coleta de informação através de pessoas, por intermédio de aparelhos conectados à rede celular e possui como vantagens, dentre outras, alcance geográfico, escalabilidade e tempo de coleta quando comparada a métodos tradicionais como censos e questionários [Silva et al. 2013a].

Recentemente, o Google lançou um serviço Web, chamado de Google Popular Times, que disponibiliza a dinâmica de funcionamento de estabelecimentos comerciais que atingem um limiar mínimo de visitas utilizando sensoriamento oportunístico. Este tipo de informação foi verificada consistente com dados provenientes de sensoriamento participativo, isto é, informação compartilhada voluntariamente de forma ativa através de *check-ins* no Foursquare em [Neves et al. 2016], que avaliou a possibilidade de reproduzir séries temporais geradas a partir do Google Popular Times utilizando dados extraídos do Foursquare para as cidades de Curitiba e Chicago. Ao melhor de nosso conhecimento, o presente trabalho é o primeiro a utilizar este tipo de informação para descobrir os padrões de comportamento emergidos nas cidades de um país. Para isso, aqui são analisadas as cidades de Curitiba, Rio de Janeiro, São Paulo, Chicago, Nova Iorque e San Francisco; considerando categorias relacionadas a hábitos de consumo de bebidas e alimentos para que seja possível explorar comportamentos emergidos em diferentes locais e investigar a relação entre países distintos.

Conhecer os diferentes comportamentos dependendo da localização pode auxiliar na melhoria da descrição do funcionamento de estabelecimentos e na compreensão da maneira como pessoas de diferentes locais interpretam a utilidade e a forma de uso de cada categoria de estabelecimento. Computacionalmente, essa informação pode ser empregada em conjunto com algoritmos de recomendação de estabelecimentos similares ao torná-los sensíveis ao contexto local e algoritmos de agendamentos ao identificar horários tipicamente de baixo movimento. Aplicações em outros domínios incluem o auxílio ao estudo de mercado para abertura de filiais de estabelecimentos em outros países e estudo de diferenças culturais como a interpretação de termos que, embora por tradução direta sejam equivalentes, podem carregar significados discrepantes para pessoas de origens distintas, isto é, enquanto um brasileiro pode entender que o uso principal de restaurante é o almoço, estado-unidenses podem assumir que o jantar e o fator que melhor define tal categoria.

As principais contribuições deste trabalho são: (1) agrupamento e análise de séries temporais sobre a dinâmica de funcionamento de estabelecimentos comerciais. Encontramos padrões de comportamento que estão relacionados a fatores locais provenientes

das cidades do mesmo país; (2) abordagem para identificar a assinatura que representa a dinâmica de funcionamento para determinadas categorias de locais. Dessa forma, é possível identificar que, apesar de um estabelecimento estar originalmente rotulado com uma determinada categoria, o seu funcionamento pode ser mais similar a outra categoria. Isso é uma indicação que a nossa abordagem pode proporcionar uma melhor descrição de estabelecimentos comerciais que pode ajudar, por exemplo, a melhorar sistemas de recomendação de locais; (3) aplicação das assinaturas obtidas no treinamento e avaliação de um modelo de classificação para inferir a categoria de um novo estabelecimento dado sua série temporal de popularidade e a cidade em que está localizado, obtendo resultados satisfatórios.

O restante deste artigo está divido da seguinte forma. A Seção 2 apresenta os trabalhos relacionados. A Seção 3 mostra a descrição dos dados coletados e como estes foram processados. A Seção 4 apresenta como foram identificados os agrupamentos e o processo de geração de assinaturas de categorias. A Seção 5 discute os resultados, incluindo uma possível aplicação das assinaturas de categorias na tarefa de inferência de categoria a partir da série temporal do estabelecimento. A Seção 6 discute sobre a validação das meta-categorias propostas. Finalmente, a Seção 7 discorre sobre as considerações finais e trabalhos futuros.

2. Trabalhos relacionados

Dados extraídos da web tem sido explorado para várias finalidades diferentes com o intuito de entender dinâmicas do comportamento social urbano e do funcionamento de cidades. *Check ins* do Foursquare foram utilizados para identificar qual a função principal de sub-regiões urbanas em [Menezes et al. 2017], enquanto áreas de pontos de interesse como pontos turísticos e estabelecimentos populares foram identificados a partir de fotos compartilhadas no Instagram em [Silva et al. 2013b] e para compreensão do estado do trânsito em regiões urbanas através do estudo de *tweets* por [Pereira et al. 2017]. Assim, destacamos três grupos de estudo que são relevantes para o presente trabalho: distribuição de elementos populares, agrupamento de séries temporais e geração da série representante do agrupamento e extração semântica de dados geolocalizados. A seguir são elencados trabalhos de cada uma destas áreas.

Check-ins compartilhados no Foursquare foram usados para investigar as propriedades das Redes Sociais Baseadas em Localização (LBSN); sendo uma das descobertas a presença de uma distribuição *power law* da popularidade de estabelecimentos, isto é, um número reduzido de estabelecimentos recebem um alto número de visitas, enquanto a maioria dos estabelecimentos apresentam baixa popularidade. Consequentemente, a análise de uma amostra limitada a estabelecimentos populares ainda é capaz de revelar o comportamento de uma grande parcela de visitantes de uma categoria de estabelecimento. Essa constatação é aqui relevante pois o serviço Google Popular Times somente oferece informações para estabelecimentos acima de um determinado limiar de popularidade [Google 2017].

Agrupamento de séries temporais tem sido aplicado para encontrar padrões em diversos domínios através de abordagens distintas. A tarefa de identificar conjuntos de séries temporais necessita a adaptação de algoritmos de agrupamento convencionais ou a utilização de algoritmos convencionais em conjunto a uma medida de distância ade-

quada [Liao 2005]. O padrão de variação de menção de *memes* em mensagens do Twitter foi observado através da aplicação de um novo algoritmo chamado K-Spectral Centroid (K-SC), uma adaptação do algoritmo K-means, desenvolvido especialmente para lidar com dados temporais [Yang and Leskovec 2011]. Posteriormente, o algoritmo K-SC foi também aplicado no cenário de vídeos do Youtube para detectar os padrões de comportamento de aumento de visualizações [Figueiredo et al. 2014]. O critério utilizado para a definição do número de agrupamentos em ambos os trabalhos foram o método da Silhueta e o índice de Hartigan, medidas para validação de agrupamentos que avalia a similaridade entre elementos do mesmo grupo comparado a similaridade em relação a elementos dos demais grupos.

Em relação a aplicação de algoritmos convencionais e a necessidade de medidas de distância adequadas, a medida *Dynamic Time Warping* (DTW) tem obtido resultados satisfatórios em vários conjuntos de dados padrões [Petitjean et al. 2011], sendo uma técnica de programação dinâmica similar a distância de edição, ou distância de Levenshtein, que busca encontrar o alinhamento ótimo global entre duas séries temporais. Para amenizar a limitação desta técnica, sua complexidade temporal, Petitjean *et al.* propôs uma heurística para o cálculo de médias chamada *DTW Barycenter Averaging* (DBA).

A extração de conhecimento a partir de dados temporais pode ser realizada baseada em informações de diferentes domínios. Por exemplo, a tarefa de inferência da descrição de um local, como sua categoria, pode ser realizada examinando a distribuição do número de *check-ins* em estabelecimentos por hora e por dia da semana. Tais características foram utilizadas em conjunto com a localização espacial e informações de cada visitante, como idade e gênero, para atribuir uma etiqueta à locais através de uma máquina de vetor de suporte binária treinada com dados provenientes de LSBN em [Ye et al. 2011] e por *boosted decision trees* treinadas com dados coletados através de pesquisas em [Krumm and Rouhana 2013].

Este presente trabalho utiliza o algoritmo de particionamento K-means, ao invés de adaptar algoritmos convencionais como em [Yang and Leskovec 2011], porém empregando as mesmas medidas de validação de agrupamentos dos estudo citados acima para auxiliar na definição do número de agrupamentos. Diferente dos outros estudos de inferência de categoria, aqui os dados utilizados são obtidos do Google Popular Times, onde o envio de informação pelo aparelho do usuário ocorre por sensoriamento oportunístico, ou seja, esta fonte não é dependente da iniciativa das pessoas e, como se trata de um serviço anônimo, tem o potencial para ser menos afetada por fatores como desejo de omitir visitas à determinados estabelecimentos do que as LBSNs.

3. Coleta e limpeza de dados

No serviço Google Popular Times as distribuições de visitas por hora de cada dia da semana representam a média de visitas ao estabelecimento durante várias semanas e são geradas a partir de informações enviadas anonimamente pelo aparelho de pessoas que aceitaram participar do serviço Google History Location através do rastreamento automático da posição do aparelho ao longo do tempo através do GPS, WI-FI e rede de telefonia móvel [Google 2017].

Para coletar os dados do Google Popular Times foi empregado um rastreador web alimentado com uma lista amostral de estabelecimentos contendo os atributos: nome,

categoria, cidade e país. Para elaboração dessa lista foi utilizada a plataforma Yelp Developers API [Yelp 2017] por oferecer informações sobre estabelecimentos em um formato padrão independente do país. Além de eventuais inconsistências de formato, essa preferência ocorre devido à disponibilidade de dados abertos ser dependente de políticas locais, de forma que a escalabilidade do número de cidades e países poderia ser prejudicada. Além disso, a plataforma Yelp permite a comunidade sugerir qual a categoria do estabelecimento, aumentando a fidelidade de que a descrição do estabelecimento corresponde ao comportamento percebido por seus usuários e amenizando possíveis erros cometidos pelos donos dos estabelecimentos durante o cadastro.

Para investigar como pessoas de diferentes locais interpretam a finalidade de cada tipo de estabelecimento, a identificação dos comportamentos frequentemente exibidos por estabelecimentos relacionados com o consumo de bebida e comida ocorreu através do estudo das seguintes categorias: padaria, bar, café, casa noturna e restaurante. Para que possamos comparar os comportamentos de lugares distintos, informações sobre estabelecimentos localizados no Brasil e nos Estados Unidos foram coletadas para tentar evidenciar possíveis diferenças de comportamento, enquanto a análise de três cidades de cada país foi usada para identificar se existem comportamentos exclusivos de cada cidade e se existem comportamentos típicos do cenário urbano destes países independente da cidade. O número total de estabelecimentos únicos coletados para cada cidade pode ser vista na Tabela 1.

Table 1. Número de estabelecimentos únicos coletados por cidade.

Cidades	Curitiba	Rio de Janeiro	São Paulo	Chicago	New York	San Francisco
Yelp	1755	2046	2964	3652	4280	3340
Google	1089	1324	2073	2672	3226	2320

Analisando a Tabela 1 é possível notar que nem todos os estabelecimentos presentes na lista previamente coletada na plataforma Yelp retornaram resultados do Google Popular Times durante a pesquisa no Google. As páginas HTML coletadas foram processadas para extrair os valores de popularidade de cada dia da semana. Esses valores foram então modelados como uma sequência discreta de valores v_k normalizados entre 0 e 1, onde k representa cada hora do dia, de forma que uma série temporal de popularidade S pode ser definida como: $S = (v_k | \forall k \in [0, 23], 0 \leq v_k \leq 1)$, onde para cada estabelecimento são geradas duas séries temporais através da técnica DBA, uma representando seu comportamento típico em dias úteis e outra o comportamento de finais de semana.

Como o objetivo é determinar qual o comportamento exibido pela maioria das séries temporais em cada agrupamento, membros anômalos foram removidos calculando a distância de cada série para o centroid do agrupamento e cortando os mais distantes. O limiar de corte aplicado seguiu a regra $Quartil_3 + Interquartil * 1.5$ da distribuição de distâncias ao centroid, que resultou na remoção de 5.95% em média do número de séries em um agrupamento.

4. Agrupamento e geração de assinaturas

4.1. Agrupamento de séries temporais

Identificar os comportamentos exibidos por cada categoria pode auxiliar a responder se uma categoria de estabelecimento tipicamente possui apenas um comportamento ho-

mogêneo ou se estabelecimentos exibem diferentes picos de popularidade mesmo pertencendo a mesma categoria. Para isso, estabelecimentos da mesma categoria com séries temporais similares foram detectados e separados em agrupamentos aplicando o algoritmo K-means com a distância DTW.

Escolher o número mais adequado de agrupamentos é um desafio comumente enfrentado ao realizar técnicas de aprendizagem não supervisionada em dados não classificados previamente [Davies and Bouldin 1979]. A heurística adotada adotada neste estudo foi o menor número entre a sugestão do método da Silhueta [Rousseeuw 1987] e o índice de Hartigan [Hartigan 1975]. Este critério é necessário devido a possibilidade de as duas métricas não convergirem para o mesmo valor, e pode ser justificado pela pouca variação encontrada ao aumentar o número de agrupamentos. Tal problema foi encontrado em [Yang and Leskovec 2011] e [Figueiredo et al. 2014] que também optaram por essas duas métricas.

Os experimentos de agrupamento de séries temporais foram divididos em três etapas: agrupamento de séries temporais de estabelecimentos da mesma categoria, agrupamento das assinaturas de diferentes categorias e agrupamento de assinaturas de diferentes cidades.

A etapa de agrupamento das assinaturas de diferentes categorias busca verificar se categorias podem exibir comportamentos semelhantes, isto é, se existem assinaturas de categorias que possam ser representadas através de uma única assinatura. O mesmo critério para a escolha do número de agrupamentos foi utilizado.

Finalmente, uma rodada para identificar as assinaturas de países associando as assinaturas de categorias mais semelhantes de cada cidade foi executada. Como o comportamento de uma categoria é considerada como representante do comportamento do país somente caso tal comportamento esteja presente em todas as cidades da amostra, o número de agrupamentos de assinaturas de países em cada categoria foi determinado como o número de assinaturas da cidade com menos agrupamentos na categoria.

4.2. Geração de assinatura

A geração de assinaturas para representar um conjunto de séries temporais pode ser realizadas de diferentes formas dependendo em como é dada a interpretação de similaridade entre séries no domínio analisado. No contexto de popularidade de estabelecimentos, assumimos que séries temporais são similares caso estas possuam curvas parecidas, ou seja, o comportamento é definido pelos horários de quedas e aumentos de visitas. Assim, uma distância como a Euclideana não seria adequada pois a comparação ocorre ponto a ponto, de forma que duas séries de comportamentos idênticos, porém separadas por um deslocamento de uma hora, seriam prejudicas no cálculo da distância.

Uma medida alternativa de distância que consegue contornar esse problema é a *Dynamic Time Warping* (DTW) que busca encontrar o melhor alinhamento global entre duas séries temporais tal que, por exemplo, uma série temporal com um pico estreito de início em 11h:00 e fim às 13h:00 é alinhada com outra série temporal de pico estreito com início às 21:00h e fim às 23h:00. Para evitar que o alinhamento global permita casos como do exemplo dado, onde curvas parecidas em horários totalmente distintos são alinhadas, uma janela de 2 horas foi aplicada.

A assinatura, isto é, uma série temporal criada para representar o agrupamento é gerada aplicando o método DBA que, em nosso caso, inicia com o centroid definido como a média de cada hora entre todas as séries temporais do agrupamento e passa para uma fase de cálculo iterativo de novos centroids através do alinhamento de uma série temporal do conjunto e o centroid temporário, de forma que ao final da execução o centroid representa uma série temporal construída que melhor alinha-se aos membros do conjunto.

5. Resultados

5.1. Agrupamento e assinatura

O número de agrupamentos de séries temporais de estabelecimentos de uma categoria sugerido pelo método utilizado, isto é, o número de comportamentos distintos existentes em uma mesma categoria, ficou geralmente entre 2 e 4 para dias úteis e finais de semana. Para verificar a real necessidade de mais de um agrupamento por categoria,uma assinatura única por categoria foi gerada inicialmente.

De maneira geral, ao gerar somente uma assinatura representativa de todas as séries temporais de estabelecimento em uma categoria, isto é, sem separá-las em agrupamentos, padrões de comportamento similares ainda emergem entre cidades de um mesmo país. Ao comparar as assinaturas representantes de cada categoria em dias úteis com as de finais de semana representadas nas Figuras 1 e 2 para o Brasil e Figuras 3 e 4 para os Estados Unidos, onde o eixo x representa as horas do dia e o eixo y o percentual de popularidade apresentado pelo estabelecimento a cada hora, é possível notar que a categoria com maior variação entre dia útil e final de semana são as casa noturnas, o que poderia ser esperado devido ao apelo noturno dos estabelecimentos desta categoria mais adequado aos finais de semana. Um comportamento interessante emergido nos Estados Unidos é a constante falta de movimento entre o horário de 04h:00 e 07h:00 para as categorias analisadas.

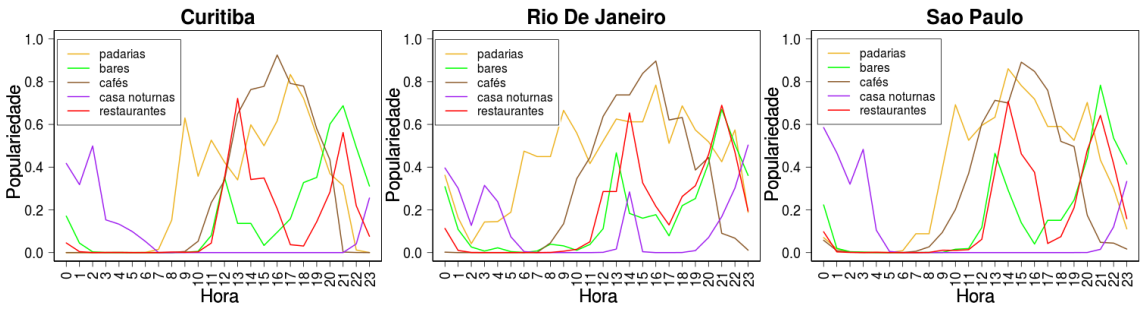

Figure 1. Assinaturas para o Brasil - dias úteis.

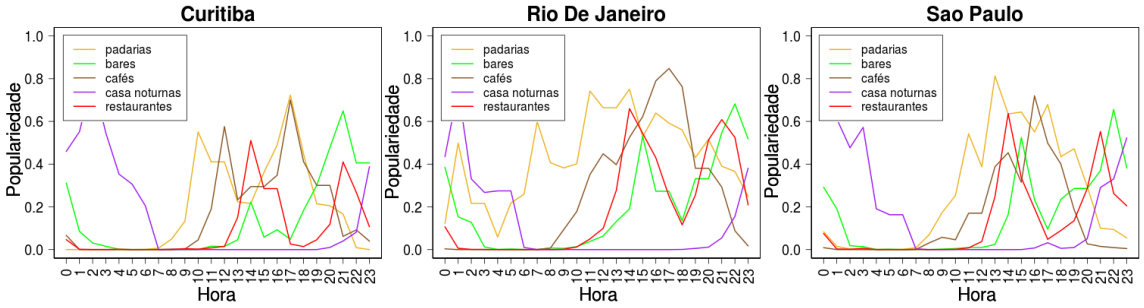

Figure 2. Assinaturas para o Brasil - finais de semana.

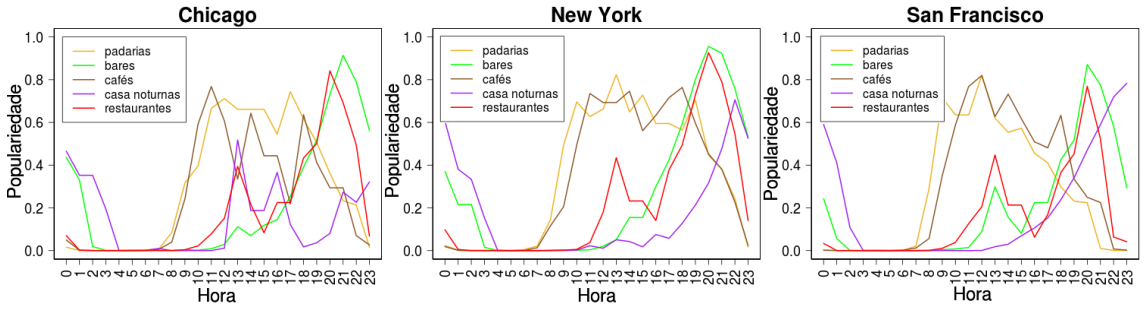

Figure 3. Assinaturas para os Estados Unidos - dias úteis.

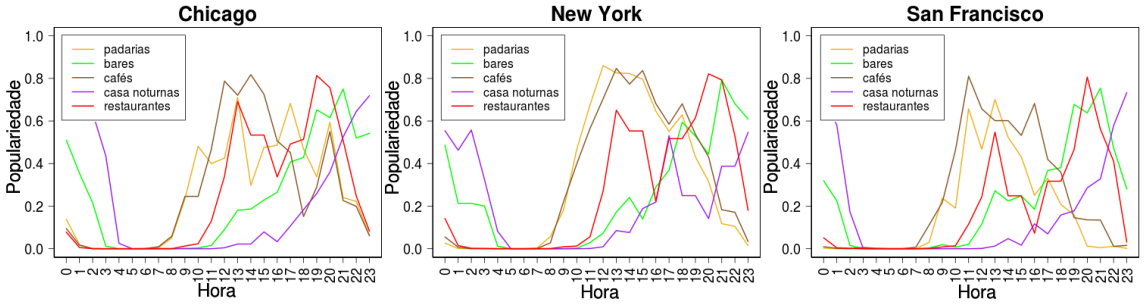

Figure 4. Assinaturas para os Estados Unidos - finais de semana.

O ponto negativo de gerar apenas um assinatura por categoria é a perda de eventuais comportamentos distintos exibidos por uma categoria, como pode ser notado para categoria bar em São Paulo ao comparar as assinaturas das Figuras 5 e 6 que representam dois agrupamentos de bares com comportamentos distintos à assinatura da Figura 1, onde bares de apenas um pico de visitação perderam representação.

As Figuras 5 e 6 mostram as várias séries temporais dos estabelecimentos em colorido e as assinaturas geradas em pontilhado para os dois agrupamentos distintos através do *K-means* para a categoria bar na cidade de São Paulo em dias úteis, conforme sugerido pelo método da Silhueta. É possível notar que tipicamente bares possuem dois grande picos em 12h:00 e 21h:00 ou apenas um pico às 21h:00 que se estende pela madrugada.

Analisando as diferentes assinaturas apresentadas por uma categoria, notamos que algumas destas são similares mesmo sendo de categorias diferentes. Isso poderia indicar a existência de estabelecimentos que, embora oficialmente declarados de uma categoria específica, apresentam características mais próximas a uma outra categoria. Para inves-

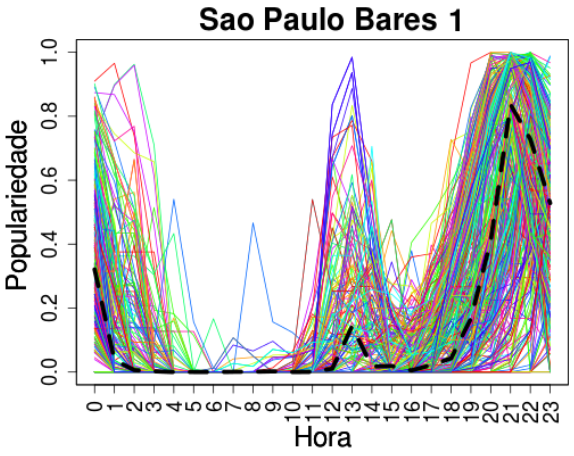

Figure 5. Bares 1 - dias úteis.

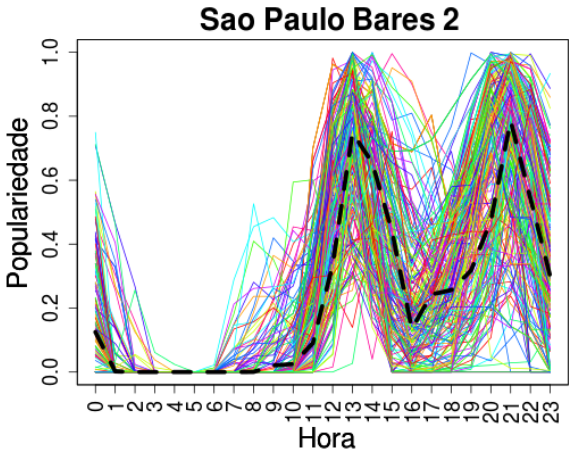

Figure 6. Bares 2 - dias úteis.

tigar este fenômeno agrupamos as assinaturas de diferentes categorias para descobrir se os agrupamentos seriam formados por assinaturas de uma mesma categoria ou resultadas da junção de diferentes categorias. Nas Figuras 7 e 8, por exemplo, as assinaturas das categorias bar e casa noturna ficaram no mesmo conjunto na cidade de Curitiba em dias úteis e em Chicago aos finais de semana.

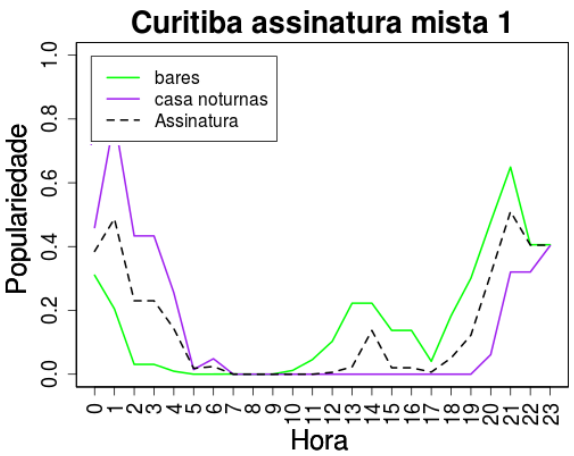

Figure 7. Assinatura bar-casa noturna em Curitiba.

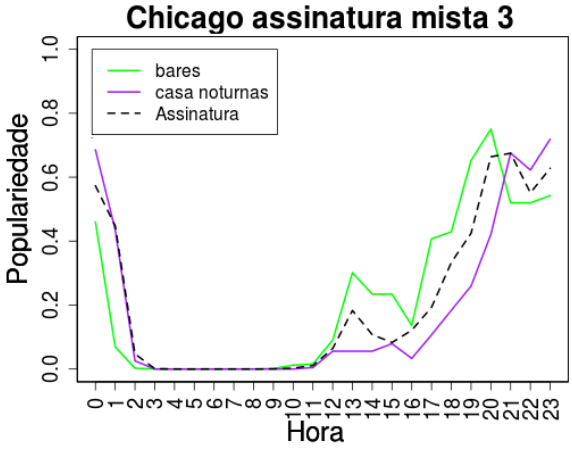

Figure 8. Assinatura bar-casa noturna em Chicago.

Essa constatação pode auxiliar em diversas tarefas, por exemplo na detecção de estabelecimentos cadastrados erroneamente em categorias inadequadas e, após a confirmação da existência de outras características similares, na elaboração de meta-categorias que descrevem mais acuradamente o real comportamento de um tipo de estabelecimento, como a meta-categoria bar-casa noturna para estabelecimentos que embora sejam originalmente bares são vistos como casa noturna por seus visitantes.

Para confirmar se os comportamentos encontrados em cada cidade são consistentes para um país, isto é, se existem padrões de popularidade que tipicamente ocorrem em cenários urbanos de um país, as assinaturas de categoria de cada cidade foram associadas minimizando a distância DTW entre elas, o que resultou para cada categoria na

geração de duas assinaturas que são compartilhadas entre as cidades analisadas do país. Estudando as assinaturas geradas a partir deste processo nas Figuras 9 - 12, notamos que aparentemente existem comportamentos de utilização de estabelecimentos que pouco diferem no cenário urbano de um país, independente da cidade.

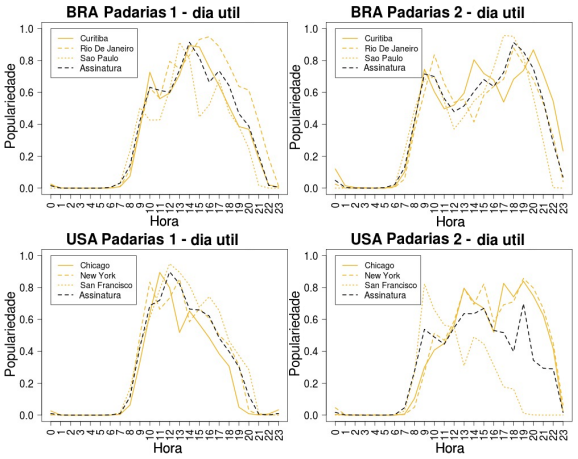

Figure 9. Assinaturas de padaria - dias úteis.

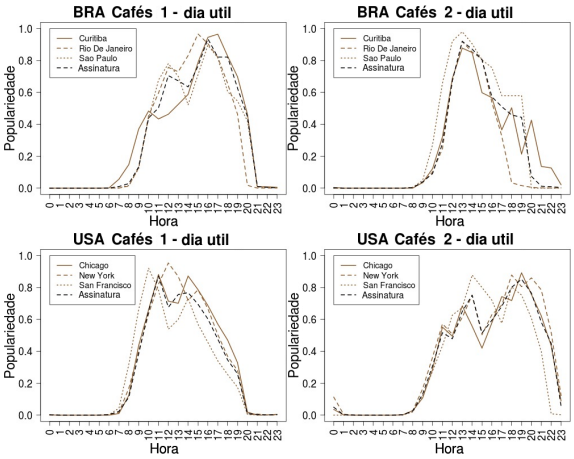

Figure 10. Assinaturas de café - dias úteis.

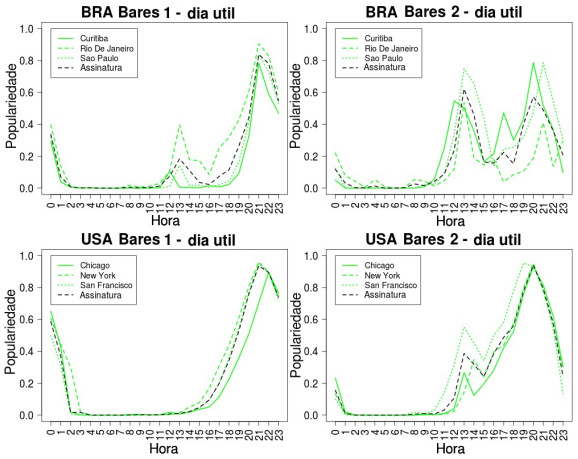

Figure 11. Assinaturas de bar - dias úteis.

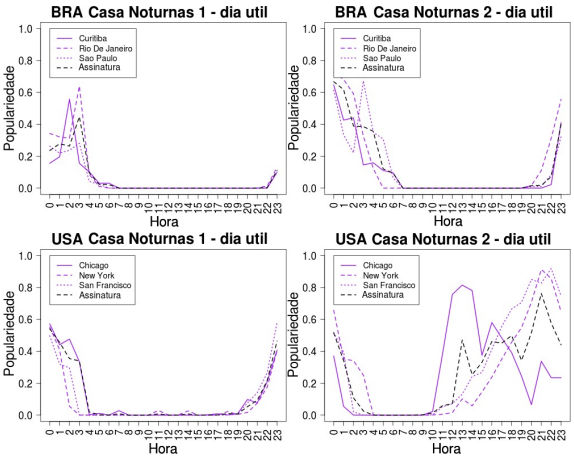

Figure 12. Assinaturas de casa noturna - dias úteis.

Enquanto as Figuras 9 e 10 denotam comportamentos similares entre Brasil e Estados Unidos para a categoria de padaria com duas assinaturas de três picos, uma mais afunilada com maior pico durante o horário de almoço e outra mais larga que se estende de 06h:00 às 23h:00, e café com uma assinatura de pico durante o almoço e outra ao final da tarde, as Figuras 11 e 12 exibem maiores diferenças entre os dois países, com o Brasil possuindo uma assinatura de bar com maior pico durante a tarde e os Estados Unidos apresentando uma assinatura de casa noturna popular no início da noite.

A categoria restaurante de maneira peculiar revelou um comportamento claramente distinto entre os dois países. Conforme as Figuras 13 e 14, enquanto o brasileiro

identifica alguns restaurantes como de uso majoritariamente para almoço, os americanos preferem frequentar este tipo de estabelecimento durante o jantar.

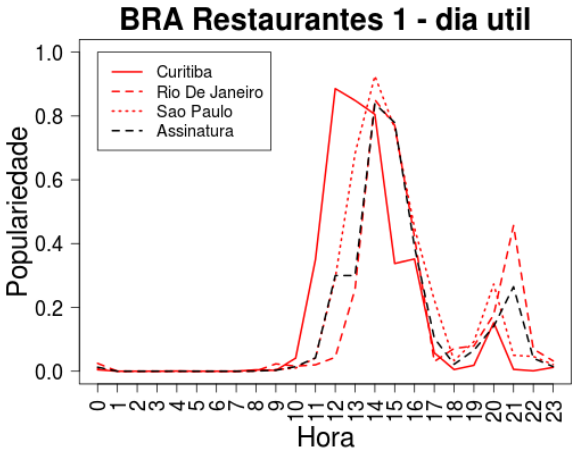

Figure 13. Assinatura de restaurante no Brasil - dias úteis.

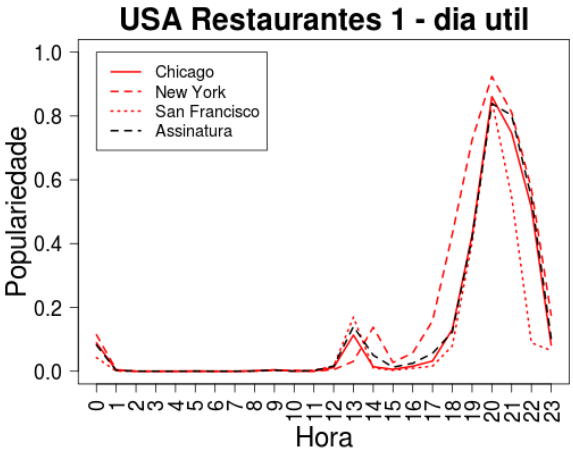

Figure 14. Assinatura de restaurante nos Estados Unidos - dias úteis.

Essas semelhanças e diferenças possuem importantes implicações, a maneira como pessoas de países distintos utilizam certos estabelecimentos pode ser afetado por diversos fatores econômicos e sociais, de forma que a distância entre assinaturas de países pode ser explorada ao estudarmos a cultura de cada país, podendo ser utilizado como uma índice de similaridade entre países. Na próxima Seção a tarefa de inferência de categoria do estabelecimento a partir de sua séries temporal é apresentada como uma possível aplicação.

5.2. Inferência de categoria

A tarefa de inferência de categoria pode auxiliar a entender se os fatores temporais como distribuição de visitas e dia da semana e geográficos como a cidade do estabelecimento são suficientes para realizar essa classificação. Para avaliar quão bem as assinaturas conseguem generalizar os comportamentos das categorias são comparados dois modelos de classificação, baseando-se em todas as séries temporais e nas assinaturas.

Um modelo $M1$ de classificador por árvore de inferência condicional foi usado para inferir a categoria de um estabelecimento dado como atributos sua série temporal, a informação se esta representa o dia útil ou final de semana e a cidade de onde ela foi obtida. O método de validação cruzada k-fold, com k = 10, foi empregado no conjunto de todas as 12704 séries temporais. Um segundo modelo de inferência, chamado de $M2$, foi desenvolvido utilizando as 87 assinaturas de categorias e meta-categorias, onde um estabelecimento é categorizado de acordo com a assinatura mais próxima à sua série temporal na distância DTW, porém com pequena tolerância na classificação como meta-categoria.

Conforme os valores da Tabela 2, notamos que o modelo $M1$ apresentou melhor desempenho na acurácia e f1-score, enquanto o modelo $M2$ obteve melhor precisão em dias úteis e acurácia nos finais de semana. Esse caso pode ser explicado pela pequena perda de representatividade produzida pelo processo de atribuir uma assinatura a

um grande grupo de séries temporais. É ainda possível notar que o desempenho em ambos modelos são levemente piores para finais de semana, o que pode indicar a existência de uma menor consistência no comportamento exercido pelos visitantes comparado a dias úteis.

Table 2. Comparação de desempenho entre os modelos M1 e M2.

	Dias úteis		Finais de semana	
	M1	M2	M1	M2
Acurácia	0.70	0.67	0.59	0.65
F1-score	0.69	0.63	0.66	0.60
Precisão	0.64	0.66	0.67	0.62
Revocação	0.76	0.60	0.66	0.58

Como o desempenho apresentado por ambos modelos são similares, o uso de assinaturas realmente permite representar os diversos comportamentos da dinâmica de estabelecimentos comerciais utilizando um número reduzido de séries temporais com pouca perda de informação.

6. Validação de meta-categorias

Um experimento foi conduzido para validar as assinaturas de meta-categorias com dez voluntários de idade variando entre 18 e 30 anos instruídos a responder um questionário online de múltipla escolha no qual eles deveriam selecionar as categorias que melhor descreveriam o estabelecimento dentre as cinco categorias listadas. Uma opção de texto livre também foi disponibilizada para que os usuários fossem capazes de apontar qualquer informação complementar que julgassem necessária. Os voluntários tiveram liberdade para interpretar a definição de cada categoria. Para cada país, dez estabelecimentos que haviam sido classificados por meta-categoria durante a tarefa de inferência foram selecionados aleatoriamente, e para cada estabelecimento foram apontados links para que os voluntários fizessem uma breve busca sobre o estabelecimentos no Foursquare, ou na página do Facebook ou nos resultados de pesquisa no Google.

Considerando apenas a resposta com maior número de votos para os 20 estabelecimentos analisados, 17 corresponderam a categoria considerada como nosso *ground-truth*, isto é, a categoria do estabelecimento na plataforma Yelp, enquanto para os 3 estabelecimentos restantes a categoria verdadeira recebeu o segundo maior número de votos. Portanto, segundo a amostra a categorização principal realmente corresponde às expectativas. No entanto, ao analisar todas as opções de respostas que mais de três pessoas concordaram notamos que 14 dos estabelecimentos foram classificados com mais de uma categoria e correspondente à nossa meta-categoria, ou seja, a categoria principal não é capaz de descrever completamente o estabelecimento. Isto pode ser visualizado na Tabela 3 que mostra a avaliação dos voluntários sobre um estabelecimento tipo padaria que acabou recebendo 6 votos para a categoria café.

Os 6 casos nos quais as categorias atribuídas pelos voluntários não corresponderam totalmente com a meta-categoria parecem ser devido a confusões na interpretação da categoria café, na qual a opção de texto livre incluiu comentários como ”casa de chá”, ”doceria” e ”sorveteria”, além de perguntas sobre o que consistia a classe café. Esse

Table 3. Classificação de uma padaria por voluntários

Categoria	Padaria	Bar	Café	Casa noturna	Restaurante
Votos	9	1	6	0	1

fenômeno ocorre também na classificação utilizada em websites, por exemplo, enquanto o estabelecimento Momofuku Milk Bar de Nova Iorque é classificado como doceria pelo Google, a plataforma Foursquare o cataloga como café e padaria.

Durante a validação constatamos que o conceito do que consiste cada categoria é algo difícil de definir. Além disso, conforme pode ser notado pelas assinaturas de cada país exibidas na Seção 4, a interpretação do significado de cada categoria parece ser afetada pelo contexto cultural de cada país, de forma que apenas traduzir nomes de categorias pode não ser capaz de transmitir completamente o significado desejado. Portanto, o uso de abordagens como a apresentada neste trabalho pode ser usada para enriquecer a descrição de locais por incluir conhecimento sobre o comportamento de seus usuários.

7. Conclusão

Neste estudo foram empregadas informações do serviço Google Popular Times sobre distribuições de visitas de estabelecimentos comerciais de diferentes cidades do Brasil e Estados Unidos para encontrar padrões de assinaturas de popularidade que representam um grande número de estabelecimentos nas cidades amostradas. Através das assinaturas de categorias foi possível estabelecer uma associação entre diferentes categorias para a criação de meta-categorias, como bar-danceteria, que foram posteriormente validadas através de um questionário com 10 voluntários. Ao comparar as 10 assinaturas representantes do Brasil e Estados Unidos é possível notar similaridade entre 6 pares de assinaturas de dias úteis e 4 de finais de semana. Esse tipo de informação poderia ser utilizada para calcular a distância entre países de acordo com a forma que pessoas tendem a frequentar estabelecimentos comerciais.

Como aplicação mostramos que é possível inferir a categoria de um estabelecimento a partir de sua série temporal e cidade de localização. Comparamos ainda dois modelos de inferência: um utilizando as séries temporais de todos os estabelecimentos e outro apenas as assinaturas de categorias e de meta-categorias. Embora ao utilizar apenas as assinaturas não houveram grandes ganhos de desempenho, utilizá-las diminui o número de séries temporais necessárias para realizar a classificação.

Investigar quais são os fatores que definem se um cenário urbano seguirá ou não as assinaturas do país pode auxiliar na melhor compreensão da dinâmica destas cidades. Como trabalho futuro propomos a aplicação da abordagem para descoberta de assinaturas em um número maior de países, de forma a identificar possíveis similaridades entre países e compará-las à índices padrões. Outra oportunidade consiste em adotar uma perspectiva na escolha das categorias de interesse não focada nos picos de popularidade, mas sim na identificação de horários de baixo movimento, o que pode ser explorado para facilitar o uso de vários serviços como serviços bancários, médicos e governamentais.

Agradecimentos

Bolsista da UTFPR/Brasil. Este trabalho foi parcialmente apoiado pelo projeto CNPq-URBCOMP (processo 403260/2016-7).

References

[Davies and Bouldin 1979] Davies, D. L. and Bouldin, D. W. (1979). A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227.

[Figueiredo et al. 2014] Figueiredo, F., Almeida, J. M., Gonçalves, M. A., and Benevenuto, F. (2014). On the dynamics of social media popularity: A youtube case study. *ACM Transactions on Internet Technology (TOIT)*, 14(4):24.

[Google 2017] Google (2017). Google popular times. `https://support.google.com/business/answer/6263531`. Acessed em: 2017-09-10.

[Hartigan 1975] Hartigan, J. A. (1975). *Clustering algorithms*, volume 209. Wiley New York.

[Krumm and Rouhana 2013] Krumm, J. and Rouhana, D. (2013). Placer: semantic place labels from diary data. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing, Zurich, Switzerland*, pages 163–172. ACM.

[Liao 2005] Liao, T. W. (2005). Clustering of time series data—a survey. *Pattern recognition*, 38(11):1857–1874.

[Menezes et al. 2017] Menezes, A. A., Santos, J. W., Souza, B. A., Almeida, T. G., Nakamura, F. G., Nakamura, E. F., and Figueiredo, C. M. (2017). Um método de detecção de regiões funcionais utilizando dados de redes sociais. In *Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC)*. Belém, Pará.

[Neves et al. 2016] Neves, Y. C., Sindeaux, M. P., Souza, W., Kozievitch, N. P., Loureiro, A. A., and Silva, T. H. (2016). Study of google popularity times series for commercial establishments of curitiba and chicago. In *Proceedings of the 22nd Brazilian Symposium on Multimedia and the Web, Teresina, Piauí, Brazil*, pages 303–310. ACM.

[Pereira et al. 2017] Pereira, B., Rettore, P., Ramos, H., Vieira, L., and Loureiro, A. (2017). T-maps: Modelo de descrição do cenário de trânsito baseado no twitter. In *Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC)*. Belém, Pará.

[Petitjean et al. 2011] Petitjean, F., Ketterlin, A., and Gançarski, P. (2011). A global averaging method for dynamic time warping, with applications to clustering. *Pattern Recognition*, 44(3):678–693.

[Rousseeuw 1987] Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

[Silva et al. 2013a] Silva, T. H., de Melo, P. O. S. V., Almeida, J. M., and Loureiro, A. A. (2013a). Social media as a source of sensing to study city dynamics and urban social behavior: Approaches, models, and opportunities. In *Ubiquitous Social Media Analysis*, pages 63–87. Springer.

[Silva et al. 2013b] Silva, T. H., de Melo, P. O. V., Almeida, J. M., and Loureiro, A. A. (2013b). Uma fotografia do instagram: Caracterização e aplicação. *Revista Brasileira de Redes de Computadores e Sistemas Distribuídos. Brasília, Distrito Federal.*

[Yang and Leskovec 2011] Yang, J. and Leskovec, J. (2011). Patterns of temporal variation in online media. In *Proceedings of the fourth ACM international conference on Web search and data mining, Kowloon, Hong Kong*, pages 177–186. ACM. Kowloon, Hong Kong.

[Ye et al. 2011] Ye, M., Shou, D., Lee, W.-C., Yin, P., and Janowicz, K. (2011). On the semantic annotation of places in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 520–528. ACM. San Diego, California.

[Yelp 2017] Yelp (2017). Yelp developers. `https://www.yelp.com/developers/documentation/v3`. Acessed: 2017-09-10.