Study of Google Popularity Times Series for Commercial Establishments of Curitiba and Chicago

Yuri B Neves*, Mozart C P Sindeaux*, William Souza*, Nádia P Kozievitch*,
Antonio A F Loureiro◇, Thiago H Silva*
*Universidade Tecnológica Federal do Paraná. Curitiba, Brasil
◇Universidade Federal de Minas Gerais. Belo Horizonte, Brasil.
yurineves,neto,william@alunos.utfpr.edu.br; loureiro@dcc.ufmg.br;
nadiap,thiagoh@utfpr.edu.br;

ABSTRACT

Urban computing is a recent area of study that helps us to understand the nature of urban phenomena. In this sense, an important aspect to study is the dynamics of commercial establishments popularity in the city. Recently, Google launched a new service that provides popularity time series of some commercial establishments in several cities. This is a valuable source of data that allow us to better understand the dynamics of establishments popularity, helping to change our perceived physical limits about the city, which can enable the development of new applications and urban services. The results of this study are: (1) characterization of Google popularity time series for bars and restaurants in the cities of Curitiba/Brazil and Chicago/USA. Among the results, we find cultural characteristics of these cities, as well as a favorable clustering of similar venues based on the temporal pattern of popularity; (2) evaluation of reproduction of Google popularity time series using Foursquare data. In this evaluation, we found evidence that Foursquare data might be used for this purpose. This means that for places where Google does not offer this service data from Foursquare, or other source, could be used. This enables the exploration of a greater number of establishments in, for example, a new venue recommendation engine.

Keywords

Computação Urbana; Google; Séries Temporais; Popularidade de Locais; Caracterização; Foursquare; Redes Sociais

1. INTRODUÇÃO

A computação urbana é uma área nova de estudo e pode ser definida como um processo de aquisição, integração e análise massiva de dados urbanos gerados por diversas fontes. Alguns dos principais objetivos dessa área são: oferecer serviços urbanos mais inteligentes e melhorar a qualidade de vida das pessoas que vivem em ambientes urbanos [17, 20]. Nessa direção, um dos aspectos importantes para estudo é a dinâmica de popularidade de estabelecimentos comerciais da cidade. Isso auxilia na mudança dos nossos limites físicos percebidos sobre a cidade, o que pode habilitar o desenvolvimento de novas aplicações e serviços urbanos.

Existem diversas fontes de dados que permitem pesquisas na área de computação urbana. Uma das mais valiosas são as redes sociais baseadas em localização, como o Foursquare, pois usuários podem ser considerados sensores sociais fornecendo dados sobre diversos aspectos da cidade de maneira espontânea [17]. Outro exemplo de fonte são serviços Web sobre áreas geográficas disponibilizados por empresas, como dados meteorológicos fornecidos pelo Clima Tempo[1] e séries temporais de popularidade de estabelecimentos comerciais fornecidos pelo Google. Esse último é um serviço recente oferecido pelo Google sobre a popularidade de estabelecimentos comerciais ao longo das horas do dia. Esse serviço está disponível para várias cidades ao redor do mundo para alguns estabelecimentos comerciais dessas cidades. Para essas cidades é possível saber, por exemplo, qual o horário mais popular e o menos popular de um estabelecimento que possui esse dado disponível. Esses dados permitem realizar um estudo sobre a dinâmica de popularidade de estabelecimentos comerciais de uma cidade.

Nesse trabalho utilizamos dados de *check-ins* do Foursquare e séries temporais de popularidade do Google para bares e restaurantes de Curitiba, Brasil, e Chicago, Estados Unidos. As contribuições deste trabalho podem ser divididas em dois grupos:

1. caracterização de séries temporais de popularidade do Google para bares e restaurantes de Curitiba e Chicago. No melhor do nosso conhecimento, este é o primeiro trabalho que estuda esses dados. Dentre os resultados, encontramos características culturais dessas cidades, bem como um elevado poder de agrupamento de locais similares com base no padrão temporal de popularidade;

2. avaliação da tentativa de reprodução das séries temporais de popularidade do Google usando dados do Foursquare. A verificação dessa possibilidade é interessante, pois as séries temporais de popularidade do Google não estão disponíveis para todos os estabelecimentos de Curitiba e Chicago. Nessa avaliação realizada, encontramos indícios de que dados do Foursquare apresentam um potencial de serem utilizados para essa finalidade. Isso significa que para locais onde o Google não oferece esse serviço dados do Foursquare poderiam ser utilizados. Isso habilita a exploração de um número maior de estabelecimentos em, por exemplo, novos mecanismos de recomendação de locais.

O restante deste trabalho é organizado como segue. A Seção 2 apresenta os trabalhos relacionados. A Seção 3 discute o conjunto

[1]http://www.climatempo.com.br.

WebMedia '16, November 08-11, 2016, Teresina, PI, Brazil

DOI: http://dx.doi.org/10.1145/2976796.2976862

de dados considerados neste trabalho, incluindo como obtê-los. A Seção 4 apresenta a caracterização das séries temporais de popularidade consideradas para Curitiba e Chicago. A Seção 5 apresenta a avaliação da tentativa de reprodução das séries temporais de popularidade do Google usando dados do Foursquare. Por fim, a Seção 6 apresenta as conclusões e trabalhos futuros.

2. TRABALHOS RELACIONADOS

Pesquisadores da área de Computação Urbana, tipicamente, adquirem novos conhecimentos através da procura de padrões, formulação de teorias e teste de hipóteses com base na observação de algum aspecto da cidade. A vasta disponibilidade na Web de dados espaciais e espaço-temporal com alta resolução oferece oportunidades para a obtenção de novos conhecimentos e melhor compreensão dos fenômenos geográficos complexos, tais como a dinâmica socioeconômica das cidades. A seguir discutimos alguns dos trabalhos relacionados nessa direção.

Em [1] os autores propõem uma forma de utilizar coordenadas geográficas e *check-ins* do Foursquare, para identificar regiões distintas da cidade, que refletem os padrões de atividades coletivas atuais. A compreensão destas informações pode ser utilizada para apresentar novos limites para os bairros. Também similar ao tema proposto neste artigo, podemos citar [12], onde os autores propuseram uma maneira de classificar áreas e usuários de uma cidade com base nas categorias de locais disponíveis no Foursquare. Isso pode ser utilizado para identificar grupos de pessoas que possuem interesses em locais de uma mesma categoria, o que seria útil para comparar áreas urbanas de cidades distintas ou para ser utilizado em um sistema de recomendação de locais. Em [6] os autores estudaram o problema da alocação ótima de lojas de varejo na cidade. Eles usaram dados do Foursquare para compreender como a popularidade de três redes de lojas de varejo em Nova Iorque é definida em termos de número de *check-ins*. Em [8] os autores apresentam um estudo sobre dados de ruídos produzidos em uma cidade. A partir desse trabalho inicial, a diminuição de ruído pode ser realizada e avaliada em locais estratégicos da cidade.

Mais relacionado com o estudo de cidades utilizando explorando a dimensão temporal, em [16] os autores propuseram uma nova metodologia para a identificação de fronteiras culturais e semelhanças entre sociedades, considerando hábitos alimentares e de bebida. Esse mesmo grupo também mostrou que utilizando dados do Foursquare e Instagram os locais parecem ter uma espécie de assinatura, ou seja, padrões que são característicos de um determinado tipo de estabelecimento [18].

O presente estudo diferencia de todos os outros, pois, no melhor do nosso conhecimento, esse é o primeiro estudo sobre as séries temporais de popularidade fornecidas pelo Google. Além disso, mostramos a utilidade desses dados no contexto de computação urbana, por exemplo, possibilitando o aprimoramento de estratégias de recomendação de locais como a apresentada em [2].

3. CONJUNTO DE DADOS

Esta seção apresenta dois conjunto de dados (*datasets*) utilizados neste trabalho. Primeiramente, a Seção 3.1 discute as séries temporais de popularidade do Google (também conhecida como Google *Popular Times*). Já a Seção 3.2 apresenta os dados do Foursquare, que é uma rede social baseada em localização.

3.1 Séries Temporais de Popularidade do Google

Séries temporais são séries estatísticas em que os dados são coletados a partir de observações durante um intervalo de tempo [3]. No escopo deste artigo, estudamos séries temporais fornecidas pelo Google. Os dados dessas séries temporais mostram a popularidade de um determinado local por hora do dia, para todos os dias de semana. Não se sabe ao certo como o Google gera essas informações, mas existem hipóteses de que é utilizado dados de localização obtidos com o auxílio do sistema operacional Android, que está presente em boa parte de dispositivos móveis que os usuários carregam.

Para visualizar as séries temporais de popularidade basta fazer uma pesquisa no Google por um determinado local e, eventualmente, um gráfico como o mostrado na Figura 1 será apresentado. Esses gráficos não estão disponíveis para todos os locais da cidade, mas sim para boa parte dos locais mais visitados. A Figura 1 é um exemplo de busca por um determinado local que possui uma série temporal de popularidade. No canto inferior direito da figura é possível observar essa série temporal em forma de um gráfico. Nesse gráfico, o eixo X representa as horas do dia, já o eixo Y representa a popularidade de cada horário. O dia da semana que esse gráfico se refere é especificado no *combobox* no canto superior direito do gráfico.

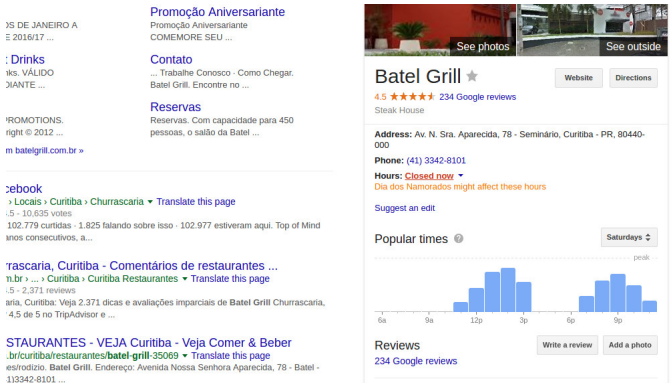


Figure 1: Exemplo de uma série temporal de popularidade do Google para um estabelecimento de Curitiba.

Para a realização deste trabalho, foram coletados as séries de popularidade do Google para diversos locais das cidades de Curitiba e Chicago. A coleta desses dados é feita por meio de coletores, também conhecidos como *web crawlers* [15, 17], que, neste trabalho, foram produzidos na linguagem *Python*. O coletor dos dados primeiramente necessita de uma lista com nomes de estabelecimentos. Para capturar os nomes dos estabelecimentos de interesse de Curitiba utilizamos dados do catálogo de endereços, fornecidos pelo website Apontador [2], que fornece dados dos estabelecimentos comerciais, incluindo a categoria dos mesmos. Para Chicago usamos a informação das licenças de estabelecimentos comerciais concedidas para essa cidade. Esses dados estão disponíveis publicamente[3] e, assim como os dados do catálogo, várias informações do estabelecimento estão disponíveis, incluindo a categoria do local.

De posse de uma lista de nomes de estabelecimentos, simulamos uma pesquisa no Google com esses nomes, com isso é carregado

[2]http://www.apontador.com.br.
[3]https://data.cityofchicago.org.

uma página com resultados da pesquisa para cada local e o coletor salva estas páginas. Com estas páginas em mãos, por meio da biblioteca da linguagem Python BeautifulSoup [14], é feita uma pesquisa por classes específicas contidas no código HTML de cada página, a fim de extrair as séries temporais de popularidade, que são sete grupos de 24 valores (que representam as 24 horas de cada dia da semana). Em seguida, esses dados são normalizados pelo maior valor encontrado para cada cidade. Ao final do processo em questão foram coletados os dados de 634 locais na cidade de Curitiba e 1419 para a cidade de Chicago, os quais são usados e descritos com mais detalhes na seção 4.

3.2 Check-ins do Foursquare

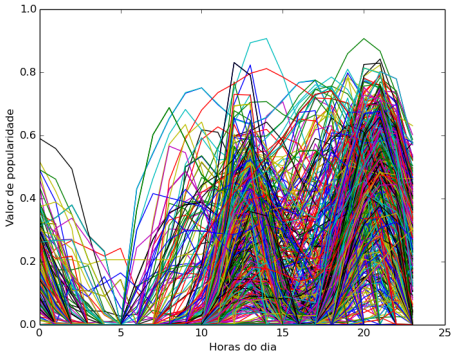
O Foursquare é uma rede social baseada em localização, que permite ao usuário informar o local onde o mesmo se encontra naquele horário. Além do tipo de dado fornecido pelo Foursquare ser bastante valioso, essa rede é bastante popular, com milhões de usuários[4], o que torna os *check-ins* uma fonte interessante de ser obtida e estudada para diversos propósitos [1].

Os dados do Foursquare foram coletados através do Twitter[5], que é um serviço de *microblogging*, ou seja, ele permite que os seus usuários enviem e recebam atualizações pessoais de outros contatos em textos de até 140 caracteres, conhecidos como *tweets*. Além de *tweets* de texto simples, os usuários também podem compartilhar *check-ins* a partir de uma integração com o Foursquare. Neste caso, *check-ins* do Foursquare anunciados no Twitter passam a ficar disponíveis publicamente, o que por padrão não acontece quando o *check-in* é publicado unicamente no sistema do Foursquare. Cada *check-in* é composto de coordenadas GPS (latitude e longitude), do horário do compartilhamento do dado, do ID do usuário compartilhador, categoria (por exemplo, comida) e um identificador do local. Mais informações sobre o *dataset* e como ele foi obtido podem ser encontradas em [19]. No total, esse *dataset* contém 4.672.841 de *check-ins* realizados por 1.929.237 de usuários em diferentes locais em abril de 2012 e nos meses de maio, junho e julho de 2014.

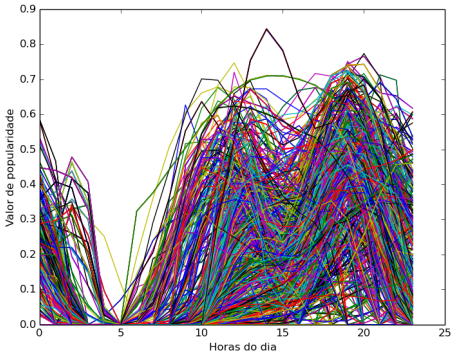
4. CARACTERIZAÇÃO DAS SÉRIES TEMPORAIS DE POPULARIDADE DO GOOGLE

Nessa seção caracterizamos as séries temporais de popularidade do Google para bares e restaurantes de Curitiba e Chicago. A Figura 2 mostra todas as séries temporais de popularidade para Curitiba (esquerda) e Chicago (direita). Cada linha do gráfico se refere a um estabelecimento, o eixo X mostra as horas do dia (24 horas) e o eixo y mostra a popularidade[6] do estabelecimento ao longo do dia. Como podemos observar, os resultados apresentados na Figura 2 não nos dizem muita coisa, e, aparentemente os resultados para as duas cidades estudadas são muito parecidos. No entanto, se analisarmos com mais profundamente esses resultados importantes diferenças são observadas.

Com o intuito de estudar mais detalhes dessas séries temporais, selecionamos locais que apresentaram no mínimo 10 *check-ins* observados na base de dados do Foursquare e que foram possíveis de serem identificados na base do Foursquare. Esse casamento de locais do Google com locais do Foursquare é descrito na Seção 5. A aplicação desse filtro nas séries temporais do Google é interessante para desconsiderar locais pouco populares entre usuários de redes



(a) Curitiba



(b) Chicago

Figure 2: Todas as séries temporais de popularidade para Curitiba (esquerda) e Chicago (direita).

sociais baseadas em localização, já que o público alvo para os potenciais resultados deste trabalho pertencem a essa classe. Neste estudo consideramos ainda apenas dias de semana (segunda a sexta). Após esse processo, obtemos 78 locais para Curitiba e 57 locais para Chicago.

Separamos essas séries temporais resultantes em grupos (*clusters*) utilizando um algoritmo de agrupamento de séries temporais baseado particionamento e em Dynamic Time Warping (DWT) [5][7]. O algoritmo constrói várias partições e as avalia usando algum critério, para isso é utilizada a distância *Dynamic Time Warping* (mais detalhes desta distância são fornecidos na Seção 5). As partições são criadas a partir da segmentação de um conjunto de dados em um conjunto de k *clusters*. Verificamos que as séries temporais de Curitiba são melhor separadas em três *clusters* e as séries de Chicago em dois *clusters*. As Figuras 3 e 4 mostram esses grupos para Curitiba e Chicago, respectivamente. Nessas figuras já podemos observar *clusters* bastante distintos.

Analisando primeiro os *clusters* para as séries temporais de Curitiba observamos no primeiro *cluster* (figura mais à esquerda) que esse *cluster* representa lugares que possuem maior popularidade

[4]http://www.foursquare.com/about.
[5]http://www.twitter.com.
[6]Todos os valores de popularidade foram normalizados em ambas cidades estudadas.
[7]Utilizamos o pacote do *R DTWClust* [13].

à noite e de madrugada. O segundo *cluster* (figura do meio), representa locais populares no horário do almoço e jantar e que são impopulares durante a madrugada. Já o último *cluster* (figura mais à direita) representa locais que são bem mais populares por volta do horário do almoço do que por volta do horário do jantar (segundo horário de maior popularidade), bem como não são populares durante a madrugada.

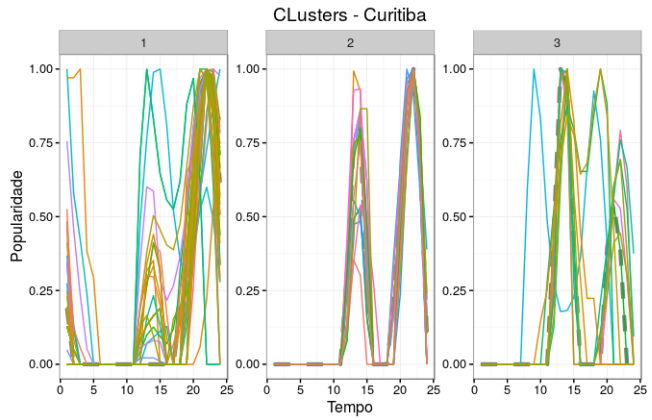


Figure 3: *Clusters* das séries temporais de popularidade do Google para Curitiba.

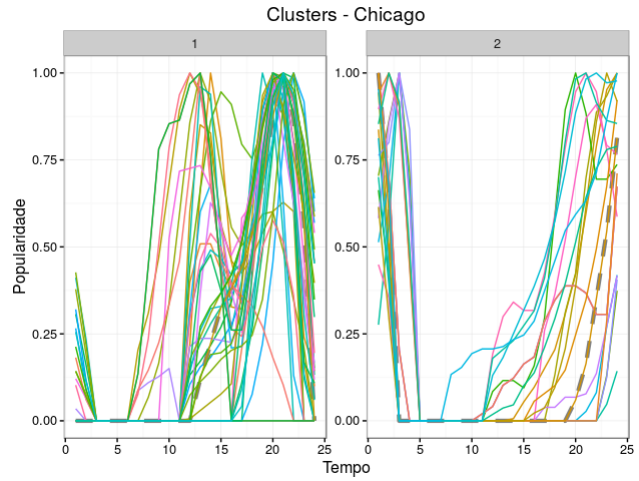


Figure 4: *Clusters* das séries temporais de popularidade do Google para Chicago.

Analisando agora os *clusters* para as séries de Chicago (Figura 4), o primeiro grupo (figura à esquerda) representa locais populares durante o almoço e jantar, com uma leve tendência para um volume maior de popularidade por volta do horário do jantar. Note que não observamos esse resultado com a mesma intensidade para os resultados de Curitiba. Isso é uma característica cultural que talvez possa ser explicada pela diferença das culturas presentes nas cidades analisadas. O segundo grupo para Chicago (figura à direita), representa locais que são mais populares à noite e durante a madrugada. Note que esses locais não são populares durante o horário do almoço, diferente do grupo similar encontrado para Curitiba. Isso também pode ser explicado por diferenças culturais entre essas cidades estudadas. Esse resultado vai de acordo com um estudo de diferenças culturais em séries temporais de popularidade utilizando dados do Foursquare desenvolvido pelos autores de [16].

Dado que alguns locais possuem séries temporais de popularidade parecidas, uma pergunta natural que surge é: as subcategorias dos locais explicam essas semelhanças nas séries temporais? Para tentar responder essa questão realizamos outro agrupamento, desta vez utilizando um agrupamento hierárquico com um critério de agrupamento por *complete linkage*[8] [4]. Para cada local consideramos a sua subcategoria (essas subcategorias são fornecidas pelo Foursquare, mais detalhes em [19]). O resultado é apresentado na forma de um dendograma [10], que pode ser observado nas Figuras 5 e 6[9] para Curitiba e Chicago, respectivamente. Como podemos observar nas Figuras 5 e 6 temos indícios de que locais semelhantes, em termos de mesma subcategoria de local, possuem séries temporais também similares. Se fizermos um corte no dendograma de Curitiba por volta da altura 0.04 nos podemos ver quatro *clusters* distintos. O primeiro, mais a esquerda, é composto majoritariamente por restaurantes, que parecem ser mais casuais, dado o elevado número de locais das subcategorias *fried chicken restaurant* e *BBQ joint*. O segundo *cluster*, logo após o primeiro, é composto por apenas dois locais que pertencem às subcategorias: *Café* e *vegetarian vegan restaurant*. O terceiro *cluster*, logo após o último mencionado, é composto também por restaurantes, e o que parece diferenciar do primeiro *cluster* é que esse grupo representa locais mais formais, dado que não foi encontrado muitas subcategorias de locais que sugerem locais casuais neste *cluster*. O último *cluster* é formado por restaurantes do tipo *fast food* e bares.

Analisando os resultados para Chicago, se fizermos um corte no dendograma para o valor da altura de 0.04 nós podemos ver quatro *clusters* distintos. O primeiro, mais a esquerda, é composto majoritariamente por bares. O segundo, logo após esse cluster mencionado, é composto por locais das subcategorias: *burger joint* e *brekfast spot*. Logo após, o terceiro *cluster* é composto por casas noturnas e bares. Talvez esses bares apresentem um padrão mais noturno, o que pode diferenciar dos bares do primeiro grupo. O último *cluster* é composto por bares e por restaurantes, que parecem ser mais sofisticados, já que, por exemplo, locais da categoria *sushi restaurant* e *mediterranean restaurant* constam nesse *cluster*.

Para entender melhor esses resultados apresentados no dendograma, a Figura 7 mostra as séries temporais de popularidade do Google de dois locais que constam no mesmo *cluster* (terceiro *cluster*, composto na maioria por casas noturnas) e que pertencem à mesma subcategoria (*bar*) para a cidade de Chicago. Como é possível observar as séries são bastante parecidas, resultado que, de certa forma, não é uma surpresa. Já a Figura 8 mostra séries de locais de subcategorias distintas, *bar* e *american restaurant*, que constam no mesmo *cluster* de Chicago (o primeiro *cluster*, mais à esquerda na figura). Como podemos constatar, locais podem ser de subcategorias distintas mas apresentar a mesma dinâmica, e isso pode ser um importante descritor sobre o tipo do local.

Em seguida escolhemos outro local, Square Bar & Grill, que também pertence à subcategoria *bar*, mas que pertence ao primeiro *cluster*, mais à esquerda, na mesma cidade. A Figura 9 mostra a série temporal de popularidade desse local juntamente com a série do local The Bar 10 Doors, estudado na Figura 7. Como pode-

[8]Julgamos que as características desse critério são interessantes para o problema estudado. No entanto outros critérios poderiam ser utilizados e essa avaliação está fora do escopo do presente estudo.

[9]Os números na frente do nome da subcategoria de local foi usado apenas para diferenciar o nome no algoritmo usado. Isso significa que Bar1 e Bar2, por exemplo, representa dois locais distintos da mesma subcategoria: Bar.

Figure 5: Dendograma para Curitiba

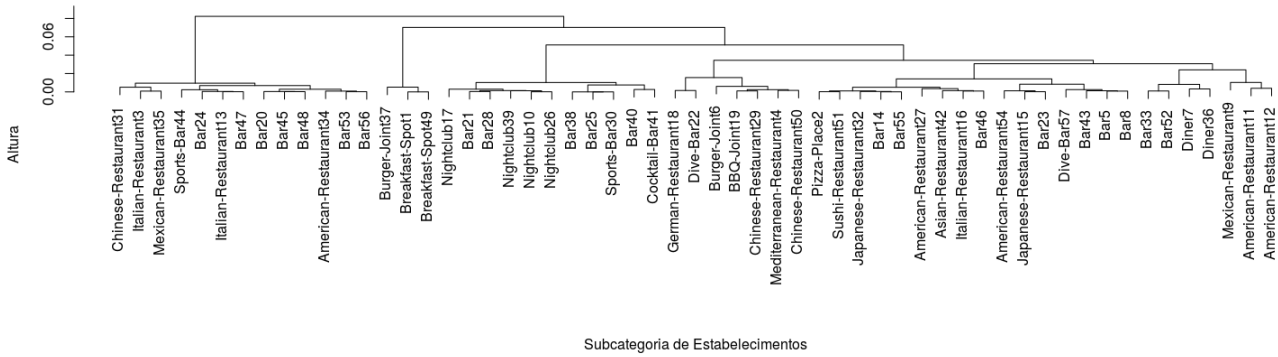


Figure 6: Dendograma para Chicago.

mos observar, existem lugares de mesma subcategoria que estão em *clusters* diferentes. Isso reforça o ponto de que a subcategoria não explica por si só o fato de dois locais estarem em um mesmo *cluster*, mas provavelmente a similaridade do comportamento dos locais explique. Note o potencial de recomendação que isso representa. No exemplo estudado, um usuário que visitou e gostou do local The Bar 10 Doors talvez goste do local Emporium Arcade Bar. É importante deixar claro que não estamos sugerindo que a recomendação deva se basear somente nesse critério, mas talvez esse critério também possa ser interessante de ser considerado em algoritmos mais sofisticados para recomendação de locais.

Isso pode significar que pessoas que frequentam determinados locais também poderiam gostar de frequentar outros locais que possuem um comportamento parecido, mostrando assim, que essa pode ser uma informação valiosa para, por exemplo, um algoritmo de recomendação de locais. Apesar da utilidade e aplicabilidade das séries temporais do Google, mostrada nesta seção, ainda existe uma limitação, ou seja, essas séries não estão disponíveis para todos os locais. Desta forma surge um novo questionamento: seria possível reproduzir esses dados com base em fontes alternativas? A próxima seção visa tentar responder essa pergunta.

5. REPRODUÇÃO DAS SÉRIES TEMPORAIS DO GOOGLE

Nesta seção descrevemos os procedimentos para tentar reproduzir as séries temporais de popularidade do Google a partir de uma fonte alternativa de dados, no caso *check-ins* do Foursquare.

5.1 Séries Temporais de Popularidade Usando Check-ins do Foursquare

Para exemplificar a metodologia proposta, nos utilizamos *check-ins* do Foursquare para as cidades de Curitiba e Chicago. No *dataset* do Foursquare (descrito na Seção 3.2), temos disponível o nome, id, subcategoria do local, latitude e longitude de cada estabelecimento. Para gerar as séries temporais de popularidade com *check-ins*, alocamos uma matriz de tamanho 7x24 (que repre-

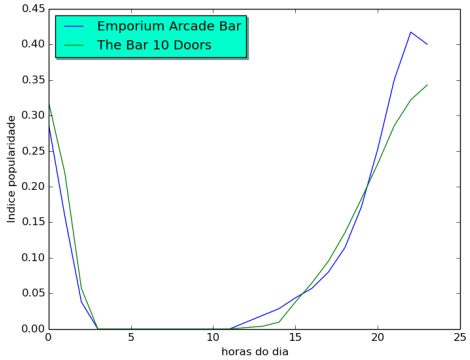


Figure 7: Séries temporais de popularidade do Google de dois locais de mesma subcategoria que constam no mesmo *cluster* de Chicago.

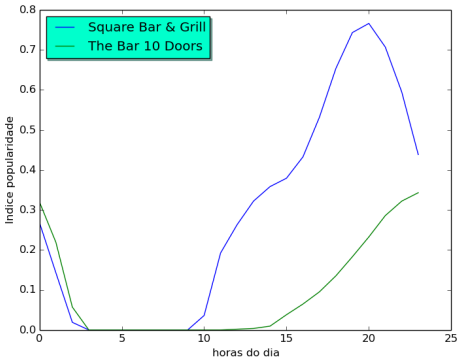


Figure 9: Séries temporais de popularidade do Google de dois locais de mesma subcategoria que constam em *clusters* distintos de Chicago.

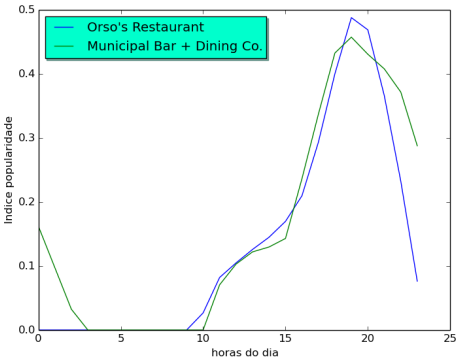


Figure 8: Séries temporais de popularidade do Google de dois locais de subcategorias distintas que constam no mesmo *cluster* de Chicago.

senta os dias e horas do dia, respectivamente). Então, buscamos por *check-ins*, utilizando o id dos locais e pra cada *check-in* encontrado extraimos a data e hora do mesmo. Com estes dados em mãos, usando um método da biblioteca *Date*, determina-se o dia da semana correspondente àquela data e a hora, usando estas informações como índice para incrementar a matriz descrita acima. Em seguida normalizamos os valores pelo maior encontrado, após isso a série temporal do local está finalizada.

Nas séries temporais do Google cada local também possui um nome, mas não necessariamente os nomes são idênticos aos nomes da base do Foursquare. Com isso, precisamos realizar uma casamento de nomes (*string matching*), para vincular um local do Foursquare, com o seu respectivo local no *dataset* do Google, com base em seus nomes. Para realizar esse casamento, utilizamos o algoritmo de Levenshtein[10] [9].

Utilizamos uma biblioteca da linguagem Python chamada *Distance*[11] que fornece o algoritmo de Levenshtein já implementado que, ao passar como parâmetro um par de nomes, é retornado um índice de diferença entre os nomes normalizado, entre 0 e 1. Quanto menor o valor retornado mais próximos são os nomes. Foi necessário definir um limite na resposta para os casamentos de $0,18$ (valor definido empiricamente que gerou 100% de acerto nos casamentos). Esse limite representa uma linha de corte para determinar o que é, ou não, um par de nomes iguais no nosso contexto. O resultado desse casamento gerou 127 registros de locais disponíveis no *dataset* do Google vinculado a um id no Foursquare para Curitiba e 103 para Chicago.

5.2 Distância Entre as Séries Temporais

A comparação entre as séries do Google e as séries geradas utilizando os *check-ins* é feita por meio da distância DTW [5]. O DTW, resumidamente, recebe duas séries S e Q, de tamanho n e m respectivamente, e as compara. Para realizar a comparação, cria-se uma matriz D (para armazenar as distâncias), de tamanho n por m, onde D_{ij} possui o resultado da seguinte expressão: $(s_i - q_j)^2$, sendo i um valor entre 0 e n, e j entre 0 e m.

Para encontrar o emparelhamento e calcular a distância entre as séries, deve-se definir um caminho que minimiza a distância acumulativa entre elas. Isto pode ser feito através do calculo da fórmula 1.

$$DTW(S,Q) = min\sqrt{\sum_{k=1}^{K} wk}, \quad (1)$$

onde wk representa o elemento da matriz D_{ij}, que pertence ao k-ésimo elemento do caminho W.

Para realizar os cálculos de distâncias entre as séries, foi utilizado uma biblioteca da linguagem *R*, chamada *TSdist*[12], que já possui diversos métodos de cálculo de similaridade de séries, dentre estes está o algoritmo de DTW. Nesta implementação passa-se como parâmetro dois vetores com valores numéricos e o método nos retorna o valor de similaridade entre esses vetores. Dessa forma, utilizamos dois vetores (um do Google e um do Foursquare) de 24 valores, que representam os índices de popularidade nas horas do dia.

Para avaliarmos a estratégia acima criamos dois *heatmaps*, mostrados nas Figuras 10 (para Curitiba) e 11 (para Chicago), onde cada ponto do eixo X diz respeito a um estabelecimento, e cada ponto no eixo Y, refere-se a cada dia da semana. Nessa avaliação, além dos locais com no mínimo 10 *check-ins*, consideramos tam-

[10](Mesmo sendo antigo, esse método continua sendo amplamente utilizado [11]).

[11]http://pypi.python.org/pypi/Distance.

[12]https://cran.r-project.org/web/packages/TSdist/index.html.

bém locais com no mínimo 30 e 50 *check-ins*. Optamos por discutir os resultados para locais com no mínimo 50 *check-ins*, pois eles refletem a mensagem principal de maneira mais compacta. Vale ressaltar que, com mais *check-ins* observamos uma ligeira tendência de reproduzir melhor os resultados, o que não é uma surpresa.

Cada célula no *heatmap* representa o resultado da distância DTW calculada, que é representado por uma cor de acordo com a escala mostrada juntamente com a figura. Vale ressaltar que no eixo Y tem uma linha além dos dias da semana, que é a da média. A referida linha foi calculada considerando todas as séries temporais da seguinte maneira: para cada local consideramos as sete séries temporais de popularidade (segunda a domingo). Para cada horário, somamos seu correspondente nas sete séries e o dividimos por sete, gerando desta forma uma nova série com 24 valores (representando as horas do dia), sendo cada um deles a média dos horários correspondentes. Executamos esses passos para as duas séries temporais consideradas (Google e Foursquare), e então seguimos o mesmo processo para determinar a distância DTW.

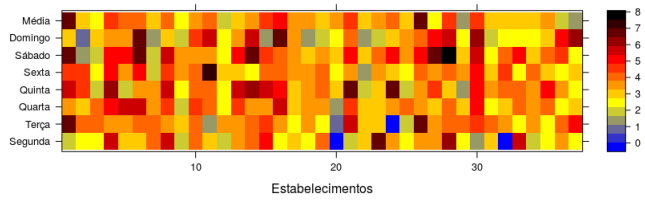


Figure 10: *Heatmap* de similaridade entre as séries do Google e as séries geradas usando o Foursquare para a cidade de Curitiba, considerando dias de semana.

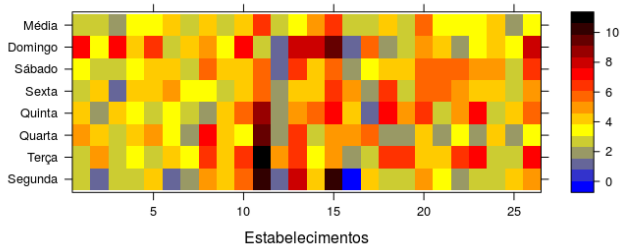
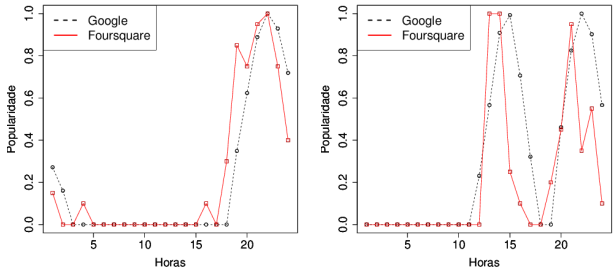


Figure 11: *Heatmap* de similaridade entre as séries do Google e as séries geradas usando o Foursquare para a cidade de Chicago, considerando dias de semana.

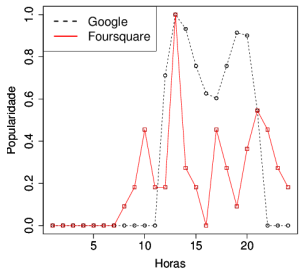
Estudando os resultados apresentados nas Figuras 10 e 11, podemos observar que para todos os locais alguns dias da semana apresentam maior similaridade entre as duas séries analisadas, ou seja, menor distância entre as séries do Google e Foursquare. Um dos motivos que pode explicar esse resultado é que o número de *check-ins* pode variar muito, isso significa que em um determinado dia um local pode ter mais *check-ins* do que outro. Por essa razão calculamos a média de todos os dias de semana, com isso acreditamos que a série temporal gerada representa melhor o comportamento dos dias de semana que, como já foi estudado em [19], tende a ser muito similar entre todos os dias de semana. Observe que os resultados para a média dos valores não é pior do que o pior resultado encontrado. Isso significa que considerar a média dos dados é mais interessante do que considerar algum dia de semana específico.

Para entendermos melhor o que os valores das distâncias significam, as Figuras 12 e 13 mostram as séries com o menor, o maior e um caso médio para as distâncias DTW encontradas para Curitiba e Chicago, respectivamente. Podemos observar que no melhor caso, com distância menor do que 2 para ambos os casos, a reprodução da série temporal de popularidade do Google com a série temporal de popularidade do Foursquare tende a ser muito boa. Percebemos também que vários casos médios, assim como os ilustrados nas Figuras 12b e 13b, capturam corretamente a essência da popularidade do local, como o horário de maior popularidade. A maioria dos resultados observados estão até o caso médio e confirmamos que a maioria capturou corretamente informações importantes sobre o horário de popularidade do local.

Resultados com valores de distância acima do caso médio podem não ser tão informativos. Ilustramos esse caso mostrando a maior distância encontrada para a série que representa a média das horas do dia de ambas as cidades, Figuras 12c e 13c. O resultado para Curitiba ainda consegue capturar o pico máximo de popularidade, no entanto para Chicago o resultado pouco útil. Observamos que o número de *check-ins* recebidos pelos locais pode estar influenciando nesses resultados, no entanto esse ponto ainda pede uma maior investigação. Acreditamos que para esses casos realizar um processo de redução de dimensionalidade por, por exemplo, *Piecewise Aggregate Aproximation* [7] pode ser interessante.



(a) Menor distância = 1,48 (b) Caso médio = 3,11

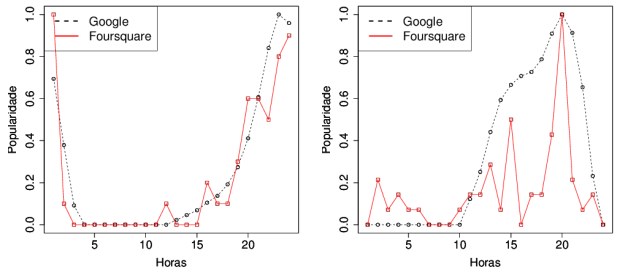


(c) Maior Ditância = 6,67

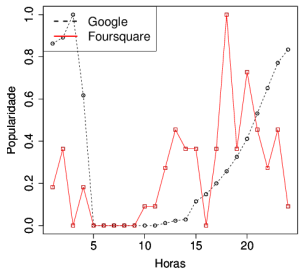
Figure 12: Séries temporais de exemplo para Curitiba.

6. CONCLUSÃO

Neste trabalho realizamos, no melhor do nosso conhecimento, o primeiro estudo sobre as séries temporais de popularidade do Google. Consideramos estabelecimentos comerciais para cidades



(a) Menor distância = 1,98 (b) Caso médio = 3,8



(c) Maior Ditância = 6,67

Figure 13: Séries temporais de exemplo para Chicago.

em dois países distintos, uma no Brasil (Curitiba) e outra nos Estados Unidos (Chicago). Mostramos que as séries temporais de popularidade do Google são bastante valiosas para um melhor entendimento da dinâmica de estabelecimentos comerciais de uma cidade. Mostramos também o potencial de recomendação de locais em que esses dados poderiam ser explorados. Apesar de inúmeras vantagens, as séries temporais de popularidade do Google não estão disponíveis para todos os estabelecimentos, o que limita certos tipos de estudo. Com isso, avaliamos a reprodutibilidade das séries temporais de popularidade do Google usando uma fonte alternativa de dados: *check-ins* do Foursquare. Nessa avaliação, encontramos indícios de que dados do Foursquare podem ser utilizados para a reprodução das séries temporais do Google. Isso abre um leque de novos trabalhos futuros, por exemplo a integração dos resultados que podem ser obtidos com esse estudo a um novo algoritmo de recomendação de locais.

Agradecimentos

Este trabalho foi parcialmente financiado com recursos da FAPEMIG, Fundação Araucária, CAPES e CNPq.

7. REFERENCES

[1] J. Cranshaw, R. Schwartz, J. I. Hong, and N. Sadeh. The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City. In *Proc. of ICWSM*, 2012.

[2] M. A. Domingues, T. E. Santos, R. Hanada, B. C. Cunha, S. O. Rezende, and M. d. G. C. Pimentel. A platform for the recommendation of points of interest in brazilian cities: Architecture and case study. In *Proc. of WebMedia*, pages 229–236, New York, NY, USA, 2015. ACM.

[3] R. S. Ehlers. Análise de séries temporais. *Universidade Federal do Paraná*, 2007.

[4] E. B. Fowlkes and C. L. Mallows. A method for comparing two hierarchical clusterings. *Journal of the American statistical association*, 78(383):553–569, 1983.

[5] T.-c. Fu. A review on time series data mining. *Engineering Applications of Artificial Intelligence*, 24(1):164–181, 2011.

[6] D. Karamshuk, A. Noulas, S. Scellato, V. Nicosia, and C. Mascolo. Geo-spotting: Mining online location-based services for optimal retail store placement. In *Proc. of KDD '13*, pages 793–801, Chicago, Illinois, USA, 2013. ACM.

[7] E. J. Keogh and M. J. Pazzani. A simple dimensionality reduction technique for fast similarity search in large time series databases. In *Knowledge Discovery and Data Mining.*, pages 122–133. Springer, 2000.

[8] N. P. Kozievitch, L. C. Gomes-Jr, T. M. C. Gadda, K. V. O. Fonseca, , and M. Akbar. Analyzing the Acoustic Urban Environment: A Geofencing-Centered Approach in the Curitiba Metropolitan Region, Brazil. In *Proc. of SMARTGREENS*, 2016.

[9] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710, 1966.

[10] O. Maimon and L. Rokach. *Data mining and knowledge discovery handbook*, volume 2. Springer, 2005.

[11] G. Navarro. A guided tour to approximate string matching. *ACM computing surveys (CSUR)*, 33(1):31–88, 2001.

[12] A. Noulas, S. Scellato, C. Mascolo, and M. Pontil. An Empirical Study of Geographic User Activity Patterns in Foursquare. In *Proc. of ICWSM'11*, 2011.

[13] T. Oates, L. Firoiu, and P. R. Cohen. Clustering time series with hidden markov models and dynamic time warping. In *Proc. of IJCAI*, pages 17–21. Citeseer, 1999.

[14] L. Richardson. Beautiful soup. *Crummy: The Site*, 2013.

[15] M. A. Russell. *Mining the Social Web*. O'Reilly Media, 2013.

[16] T. Silva, P. Vaz de Melo, J. Almeida, M. Musolesi, and A. Loureiro. You are what you eat (and drink): Identifying cultural boundaries by analyzing food e drink habits in foursquare. In *Proc. of ICWSM*, Ann Arbor, USA, 2014.

[17] T. H. Silva and A. A. Loureiro. Computação urbana: Técnicas para o estudo de sociedades com redes de sensoriamento participativo. In *Anais da XXXIV JAI*, volume 8329, pages 68–122. SBC, 2015.

[18] T. H. Silva, P. O. S. Vaz de Melo, J. M. Almeida, J. Salles, and A. A. F. Loureiro. A picture of Instagram is worth more than a thousand words: Workload characterization and application. In *Proc. of DCOSS'13*, Cambridge, MA, USA, May 2013.

[19] T. H. Silva, P. O. S. Vaz de Melo, J. M. Almeida, J. Salles, and A. A. F. Loureiro. Revealing the city that we cannot see. *ACM Trans. Internet Technol.*, 14(4):26:1–26:23, Dec. 2014.

[20] Y. Zheng, L. Capra, O. Wolfson, and H. Yang. Urban computing: concepts, methodologies, and applications. *ACM TIST*, 5(3):38, 2014.