Caracterização topológica de redes viárias por meio da análise de vetores de características e técnicas de agrupamento

**Gabriel Spadon, Lucas C. Scabora, Marcos R. Nesso-Jr,
Caetano Traina-Jr, Jose F. Rodrigues-Jr**

Instituto de Ciências Matemáticas e de Computação (ICMC)
Universidade de São Paulo (USP)
São Carlos, SP – Brazil

{spadon, lucascsb, marcosnesso}@usp.br, {caetano, junio}@icmc.usp.br

Abstract. *Complex networks contribute to computational research by their ability to design systems modeled with vertices and edges. They provide means to describe urban structures through their street mesh, expressing predicates that refer to the flow and transportation in an urban zone. Towards the analysis of information from street networks, and by means of metrics from its elements, this paper aims at describing interactions between different cities using feature vectors. Our analysis is based on the use of digital maps; through them, we provide means for data modeling and feature extraction to support analytical activities. Our results are based on the analysis of 645 cities, which shape the Brazilian state of Sao Paulo. We show how the joint of features from complex-network metrics can describe urban indicators that are rooted in the network topology and how they can reveal differences among cities.*

Resumo. *As redes complexas contribuem para a pesquisa computacional por sua capacidade de projetar sistemas modelados por vértices e arestas. Eles fornecem meios para descrever estruturas urbanas por meio das malhas viárias, expressando predicados que se referem ao fluxo e ao transporte em zonas urbanas. Este trabalho tem o objetivo de descrever as interações entre diferentes cidades usando seus vetores de características pela análise de informações viárias e das métricas inerentes aos seus elementos. Propõe-se uma análise baseada no uso de mapas digitais, pois permitem abordagens para modelagem de dados e extração de características que suportam atividades analíticas. Os resultados deste trabalho são baseados na análise de 645 cidades, que formam o estado brasileiro de São Paulo; tais resultados demonstram como características extraídas por métricas de grafos descrevem indicadores urbanos que estão enraizados na topologia da rede, e como podem revelar diferenças entre cidades distintas.*

1. Introdução e Trabalhos Relacionados

As redes complexas são usadas para modelar sistemas reais e sintéticos, sendo exemplos disso as redes de interação proteica, as malhas viárias e as linhas metroviárias. Essas redes, como modelos matemáticos, se destacam devido às suas propriedades algébricas e potencial computacional, com aplicabilidade analítica para suportar processos cognitivos de tomada de decisão (Boccaletti et al. 2006). Por meio de métricas e métodos baseados

na topologia e/ou geometria das redes é possível identificar características de interesse que não são óbvias por inspeções humanas; porque as redes podem ser grandes (elevado número de vértices), intrincadas (elevado número de arestas), ou podem conter padrões e atributos não triviais, cuja observação depende da aplicação de técnicas algorítmicas.

No caso específico da representação de malhas viárias, as redes complexas descrevem fatores relacionados ao deslocamento de indivíduos, a localização e alocação de serviços, a melhoria de tarefas relacionadas ao transporte e até ao estudo de fatores advindos do comportamento coletivo. Neste contexto, foi observada a falta de estudos e/ou análises que caracterizam os grupos de cidades por meio da similaridade de suas características estritamente topológicas, que é o objetivo deste trabalho. Essa abordagem tem aplicações para a compreensão da morfologia urbana, bem como para identificar o porquê cidades partilham propriedades por estarem próximas ou distantes entre si.

A proposta deste trabalho se baseia na análise de 645 cidades do estado de São Paulo, visando fornecer compreensão sobre as peculiaridades existentes em diferentes cidades, interpretando suas características globais e usando métodos da área de aprendizado de máquina para a modelagem de dados, análise de agrupamentos e projeção multidimensional. Neste cenário, as seguintes premissas motivaram a presente pesquisa: **(A)** a topologia da rede é um poderoso conjunto de ferramentas que pode ser usado para identificar grupos de cidades com características semelhantes, potencialmente revelando disparidades (cidades que são muito grandes ou muito pequenas) sem utilizar dados demográficos; **(B)** embora cidades possam compartilhar fronteiras administrativas com outras, elas tendem a se agrupar com cidades das quais estão distantes; e, **(C)** pode haver correlação entre indicadores urbanos e/ou territoriais quando comparados com as características extraídas da topologia das redes viárias dentre o conjunto de 645 cidades.

Com o objetivo de resolver questões relacionadas ao cenário urbano, estudos foram realizados para descrever cidades considerando seu intenso fluxo de veículos (Masucci et al. 2013) e comportamento coletivo (Blumer 1971), enquanto outros analisaram a densidade de acidentes nas redes viárias (Anderson 2009) e as discrepâncias entre cidades por meio de seus indicadores urbanos (Grauwin et al. 2015). Alguns autores investigaram métodos métrico-analíticos aplicados em cidades (Crucitti et al. 2006, Costa et al. 2010), outros se concentraram no apoio ao desenho e ao planejamento urbano (Porta et al. 2009, Strano et al. 2012, Spadon et al. 2017), e há aqueles que avançaram com a análise e posicionamento de instalações (centros de serviço públicos e/ou privados) em cidades (Li and Parrott 2016). Entretanto, dentre estas aplicações, a análise de agrupamentos é ainda incipiente, mas é considerada um poderoso conjunto de ferramentas (Pan et al. 2013).

Com propósito semelhante ao deste trabalho, duas pesquisas do estado da arte (Strano et al. 2013, Domingues et al. 2017) usaram técnicas de agrupamento para analisar grupos de cidades. A primeira teve a intenção de medir graus de semelhança entre dez cidades europeias, enquanto a segunda realizou uma avaliação de agrupamentos considerando a proximidade e sobreposição de 1150 cidades, principalmente da América anglo-saxônica. Diversos algoritmos são capazes de agrupar dados, por exemplo, alguns fornecem melhores resultados para os dados que estão dispostos em polígonos convexos, outros não convexos; ainda, existem aqueles que se baseiam na hierarquia dos dados, enquanto outros não. Todavia, ambos os autores não discutiram a significância de seus

respectivos resultados, ou seja, a qualidade dos agrupamentos, comprovando a adequação dos métodos aos dados. Algumas das métricas com este objetivo são: *Silhouette*, *Dunn Index*, *Z-Score*, *Accuracy*, e *Precision-Recall* (Kremer et al. 2011); cada uma delas avalia diferentes perspectivas dos agrupamentos e sua combinação pode revelar melhores resultados ao descrever os dados.

Este trabalho contribui com técnicas que promovem a análise de sistemas urbanos por meio de grafos. Os resultados têm aplicações para a compreensão de semelhanças e diferenças entre cidades. Para apresentar as contribuições, este artigo está organizado como segue: a Seção 2 expõe a proposta e explica a validação dos resultados; a Seção 3 discute os resultados sobre a aplicabilidade dos métodos propostos; e, por fim, a Seção 4 apresenta as conclusões e considerações finais.

2. Proposta

Nesta seção é apresentada a proposta deste trabalho, cujo objetivo é promover a comparação entre cidades e o agrupamento delas. Essa seção está dividida do seguinte modo: na Seção 2.1 apresentam-se notações formais sobre as redes viárias e vetores de características; na Seção 2.2 detalha-se a modelagem e pré-processamento de dados; na Seção 2.3 discute-se a extração de características; e, por fim, na Seção 2.4 descrevem-se as técnicas utilizadas para a mineração e avaliação dos vetores, visando promover a projeção e o agrupamento dos vetores de características extraídos.

2.1. Conceitos Fundamentais

Grafos direcionados e ponderados são referidos neste texto como redes complexas e, apesar de diferentes, redes complexas e grafos são considerados equivalentes. Todo grafo $\mathbf{G} = \{V, E\}$ é composto de um conjunto de $|V|$ vértices (ou nós) e outro de $|E|$ arestas. Além disso, cada aresta $e \in E$ é conhecida por ser um par ordenado $\langle o, d \rangle$, em que $o \in V$ é o nó de *origem* e $d \in V$ é o de *destino*, $o \neq d$. Para cada aresta pode ser atribuído um peso numérico (d_{od}), o qual é referente à *distância dos grandes círculos* entre os vértices o e d na projeção esférica da superfície da Terra (Konstantopoulos 2012).

Um vetor de características $A = (a_1, \ldots, a_n) \in \mathbb{R}^n$ pode conter múltiplas métricas extraídas de uma rede complexa. A comparação entre dois vetores A e B é baseada em uma função de distância pré-definida $f(A, B) : A \times B \rightarrow [i, j],\ i \leq j$, no qual i indica os vetores mais próximos e j os mais distantes; por exemplo, a função de distância *Minkowski* é definida como $f(A, B) = \sqrt[p]{\sum_{i=1}^{n} |a_i - b_i|^p}$, onde diferentes valores de p indicam funções de distância distintas, por exemplo, a *Manhattan* (p_1) e a *Euclidiana* (p_2).

2.2. Aquisição e Preparação dos Dados

Para cada uma das 645 cidades do estado de São Paulo, obteve-se seus limites administrativos, indicadores territoriais e demográficos por meio do Instituto Brasileiro de Geografia e Estatística (IBGE)[1]. Os dados utilizados para modelar as redes complexas foram extraídos do OpenStreetMap (OSM)[2], uma rede social de mapeamento colaborativo de vias. Os limites territoriais foram utilizados para segmentar os dados geográficos do OSM em pequenas porções, cada qual representando uma cidade. Cada uma destas

[1] ⟨www.ibge.gov.br⟩
[2] ⟨www.openstreetmap.org⟩

porções descrevem o mundo real por meio de objetos georreferenciados. Estes objetos são descritos por relações, as quais se referem às vias e aos cruzamentos entre elas.

Existem duas possibilidades para construir uma rede complexa a partir de um arquivo do OSM. O primeiro é conhecido como *Grafo Primal* (Porta et al. 2006b), o qual considera as ruas como arestas e seus cruzamentos como vértices. O outro é o *Grafo Dual* (Porta et al. 2006a) em que as ruas são vértices e os cruzamentos são arestas. Levando em consideração que os atributos espaciais são essenciais para o domínio urbano e que não se pode calcular distância (em metros) por meio de dados não espaciais, foi escolhido o *Grafo Primal* ao invés do *Grafo Dual*. Consequentemente, usando um *Grafo Primal* assume-se que as redes são planares e que podem ser representadas em duas dimensões, no qual uma ou mais arestas se cruzam somente onde nós são definidos.

2.3. Extração e Seleção de Características

Métricas de grafos podem ser divididas entre locais e globais (Scripps et al. 2010); as métricas locais descrevem as propriedades para cada um dos elementos que formam a rede, enquanto as métricas globais caracterizam toda a rede por um valor que descreve todos seus elementos em conjunto. Note que este trabalho faz uso das métricas globais pois permitem a comparação direta de cidades distintas, enquanto as métricas locais não.

Para obter os resultados destas métricas, foi desenvolvido um extrator de características que produz um vetor de características para qualquer rede complexa. Em um primeiro momento, foram selecionadas várias métricas de grafos como potenciais candidatos para prover características das cidades. Destas, 29 foram mantidas por sua capacidade de prover informações sobre redes viárias. Todas as métricas selecionadas têm base na análise da topologia da rede, uma vez que a topologia descreve a malha viária das cidades, que é a base para o desenvolvimento desta pesquisa.

As métricas que compõem o conjunto de testes são: **(1)** número de auto conexões; **(2)** número de nós; **(3)** número de arestas; **(4)** número de vias unidirecionais; **(5)** número de vias bidirecionais; **(6)** média do grau de entrada; **(7)** média do grau de saída; **(8)** grau médio; **(9)** grau ponderado de entrada; **(10)** grau ponderado de saída; **(11)** entropia da distribuição de grau; **(12)** coeficiente de correlação do grau dos nós; **(13)** média ponderada de distâncias da rede; **(14)** média das distâncias das vias; **(15)** raio da rede; **(17)** diâmetro da rede; **(17)** entropia da distribuição de distâncias; **(18)** entropia da distribuição de caminhos mínimos; **(19)** média dos caminhos mínimos; **(20)** densidade; **(21)** densidade de rede planar; **(22)** transitividade; **(23)** índice de agrupamento médio; **(24)** índice de agrupamento global; **(25)** dominância do ponto central; e, coeficiente de assortatividade de **(26)** entrada$\times$entrada, **(27)** entrada$\times$saída, **(28)** saída$\times$entrada, e **(29)** saída$\times$saída.

Após coletar todas as métricas, as não relevantes foram removidas por meio da análise de correlação. Foi calculado o coeficiente de correlação de Pearson (Chiang 2003) para cada par de métricas. Esse coeficiente é definido no intervalo $[-1, 1]$, no qual os valores extremos indicam, respectivamente, a correlação máxima negativa e positiva, enquanto 0 indica nenhuma correlação linear. Na sequência, foram definidos dois valores-limite $\left[-\frac{1}{2}, \frac{1}{2}\right]$ dentro do intervalo original, permitindo remover todos as características com forte correlação mútua. Nos casos em que duas características estão fora do intervalo de corte, uma das métricas foi descartada aleatoriamente. O processo de seleção garante que apenas métricas não relacionadas umas com as outras serão usadas para

descrever as cidades. Como resultado, cada vetor de características é definido como $F = (\mathcal{H}, \mathscr{L}, \mathcal{R}, \mathcal{E}, \mathcal{D}, \mathcal{P}, \mathcal{B}, \mathcal{G}_c)$, contendo apenas 8 das 29 métricas avaliadas. As 8 métricas escolhidas são definidas de acordo com Costa et al. 2007 como se segue:

Entropia da Distribuição de Grau **($\mathcal{H}$).** A distribuição de grau de uma rede descreve seus vértices por meio de probabilidades de acordo com a quantidade dos vértices que possuem o mesmo grau. Considerando que, a entropia representa a quantidade de incerteza e aleatoriedade em uma determinada informação, ao usar a entropia em uma distribuição de grau de uma cidade, pode-se medir a incerteza entre as conexões de suas vias. A métrica é descrita na Equação 1, onde P_k é a proporção de vértices com grau k.

Média dos Caminhos Mínimos **($\mathscr{L}$).** Quantifica a média de todos os caminhos mínimos (d_{ij}^S) que conectam todos os pares de vértices em um grafo (ver Equação 2). É usada para quantificar a capacidade de locomoção por meio dos caminhos mais curtos de uma cidade.

Coeficiente de Assortatividade **($\mathcal{R}$).** Refere-se ao grau de correlação entre pares de nós. Valores positivos indicam que os nós com grau similar tendem a se conectar uns aos outros, enquanto que valores negativos indicam o mesmo, mas em relação a nós com graus diferentes. Pode ser entendida como a probabilidade de passar de uma rua sem importância para uma rua importante, com base apenas no número de ruas adjacentes a ambas. A métrica (ver Equação 3) usa e_{xy} para indicar à fração de arestas que conectam nós com grau x e y, a_x e b_y para a fração de arestas que começam e terminam em vértices com grau x e y; e σ_a e σ_b para o desvio padrão das distribuições de a_x e b_y.

$$\mathcal{H} = -\sum_{k=0}^{\infty} P_k \times \log(P_k) \quad (1) \qquad \mathscr{L} = \frac{\sum_{i=1}^{|V|}\sum_{j=1}^{|V|} d_{ij}^S}{|V|(|V|-1)} \quad (2) \qquad \mathcal{R} = \frac{\sum_{xy} xy(e_{xy} - a_x b_y)}{\sigma_a \sigma_b} \quad (3)$$

Excentricidade **($\mathcal{E}$).** Esta métrica é local e mede para um conjunto de vértices o maior caminho mínimo entre todos os outros vértices do grafo (ver Equação 4). Em uma perspectiva global, a maior excentricidade de um grafo é conhecida como o *diâmetro*, enquanto a menor é denominada de *raio*. O diâmetro e o raio podem indicar cidades que sofrem com problemas de locomoção urbana, este é o caso de redes geograficamente esparsas, que são redes viárias que possuem o raio muito pequeno em relação ao diâmetro.

Densidade da Rede Planar **($\mathcal{D}$).** A densidade (ver Equação 5) de um grafo planar é definida como a relação entre o número de arestas $|E|$ e o número de todas as arestas possíveis em um grafo com $|N|$ nós, no qual as arestas não se cruzam a não ser nos nós da rede; é capaz de revelar o quão densa é a malha viária de uma cidade ou de um bairro.

Dominância do Ponto Central **($\mathcal{P}$).** A métrica avalia a centralidade global de uma rede por meio do desvio padrão entre os valores de *Betweenness* de seus vértices, que é uma métrica de centralidade baseada em distância. Para valores próximos de 0, sabe-se que

existem muitas rotas eficientes que são semelhantes às mais curtas; enquanto que, para valores próximos de 1, a métrica indica que a rede é vulnerável sem o nó central pois o mesmo é usado para conectar diferentes componentes, servindo como ponto de acesso (*e.g.* pontes, viadutos e túneis). Na Equação 6, usa-se $\bar{v}$ como o vértice com o maior *Betweenness* e $\mathcal{B}(v)$ como o *Betweenness* normalizado do vértice v, definido entre $[0, 1]$.

$$\mathcal{E}_i = \frac{1}{max\{d_{ij}^S | \forall j \in V\}} \quad (4) \qquad \mathcal{D} = \frac{|E| - |N| + 1}{2|N| - 5} \quad (5) \qquad \mathcal{P} = \frac{\sum_v^{|V|} \mathcal{B}_{\bar{v}} - \mathcal{B}_v}{|V|(|V| - 1)} \quad (6)$$

***Vias Bidirecionais* ($\mathcal{B}$).** Consiste no número de arestas bidirecionais de um grafo; elas representam vias que fornecem rotas em dois sentidos entre o mesmo par de vértices.

***Agrupamento Global* ($\mathcal{G}_c$).** Esta métrica consiste na fração do número de triângulos $\mathbb{N}_\triangle$ e triplas $\mathbb{N}_3$, que é dado por $\mathcal{G}_c = (3 \times \mathbb{N}_\triangle) \div \mathbb{N}_3$. Descreve como as ruas tendem a se agrupar nos cruzamentos de uma determinada cidade, de modo que quanto maior o valor, maiores as possibilidades de locomoção em menos etapas entre pares de vértices distintos.

2.4. Análise de Vetores de Características

Esta etapa concentrou-se na aplicação de dois métodos da literatura de mineração de dados: o primeiro de projeção multidimensional e o segundo de análise de agrupamentos. A projeção multidimensional permite a visualização de dados, reduzindo seu espaço dimensional, revelando particularidades e comportamentos a serem explorados por meio de análise de suas relações. A análise de agrupamentos, por sua vez, se concentra no estudo das interações entre os dados, inferindo que dois elementos são semelhantes porque estão no mesmo grupo ou são dissimilares porque estão em grupos distintos. Deste modo, a combinação destes dois métodos contribui para a avaliação das cidades pelo seu elevado potencial de revelar características e padrões intrínsecos ao conjunto de dados.

Em relação à projeção multidimensional, foram aplicadas duas técnicas de redução de dimensionalidade (Spiwok et al. 2015); a primeira é chamada Isomap e a segunda é conhecida como Análise de Componentes Principais (PCA). Isomap é uma técnica de redução de dimensionalidade não linear, que fornece uma projeção em uma dimensão inferior, mantendo a distância geodésica entre os dados. PCA é uma técnica linear que usa conversões ortogonais para transformar um conjunto de variáveis em valores linearmente não correlacionados com a maior variância mútua. Na análise de agrupamentos foi usado KMeans (MacQueen et al. 1967), que divide os dados em grupos de igual variância, minimizando a distância da soma de quadrados entre eles.

Para escolher as duas técnicas de projeção, usou-se conhecimento sobre o domínio; manteve-se registro de algumas cidades discrepantes já conhecidas, buscando abordagens que as diferenciasse nitidamente. Já a técnica de agrupamento foi escolhida por ser um algoritmo amplamente utilizado na literatura relacionada que é conhecido por ser escalável para um grande número de amostras e que tem sido usado para uma variedade considerável de propósitos em muitos domínios de aplicação.

Para a validação dos resultados, consideram-se métricas de avaliação de qualidade de agrupamentos (Kremer et al. 2011). O foco de tais métricas é analisar a semelhança

entre elementos que foram atribuídos ao mesmo grupo. As que atendem a este objetivo são: *Silhouette* e *Dunn Index*; ambos conhecidos como métricas de qualidade interna, as quais não demandam dados rotulados. A medida de *Silhouette* é definida no intervalo $[-1, 1]$; os valores são atribuídos a cada grupo inerente aos dados e quanto mais próximo de 1, melhor. *Dunn Index* foi usado para evitar casos em que a medida de *Silhouette* falha, uma vez que valores de *Dunn Index* maiores que 1 indicam resultados confiáveis.

A junção dos processos de mineração, juntamente com as métricas de validação, permite interpretar as características da rede (Seção 2.3) em relação à sua semântica no domínio do problema; os resultados relacionados serão discutidos na próxima seção.

3. Resultados

Esta seção apresenta e discute os resultados deste trabalho. Ela está dividida em duas partes; na Seção 3.1 são descritos os resultados da técnica de projeção multidimensional e na Seção 3.2 discute-se os que se referem a análise de agrupamentos.

3.1. Projeção Multidimensional

Em relação à população[3], a maioria das cidades do conjunto de dados é considerada como de pequeno porte, mas ainda apresenta um conjunto substancial de cidades de médio porte e um pequeno conjunto de cidades de grande porte, no qual São Paulo — a maior cidade brasileira — está localizada. Algumas análises preliminares podem ser feitas observando a Figura 1, onde as cidades (pontos) foram dimensionadas pela sua quantidade de vértices.

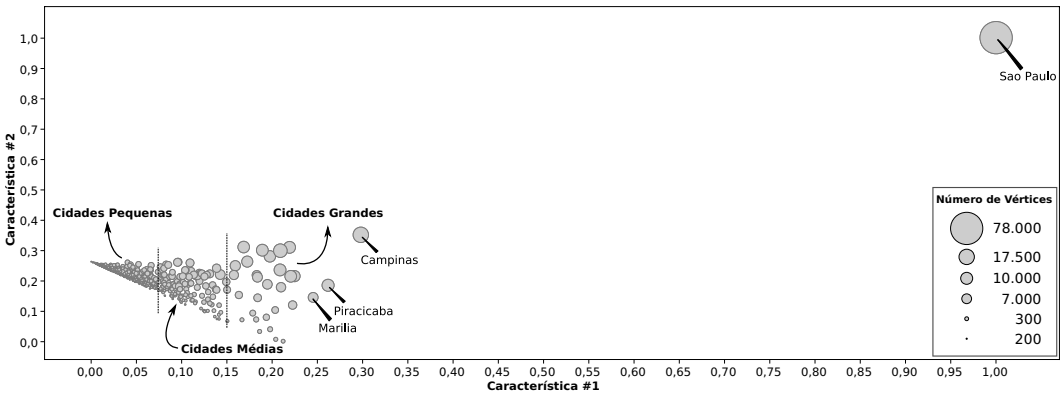


Figura 1: Representação dos vetores de características projetados em duas dimensões usando PCA; cidades são pontos dimensionados pelo número de vértices em suas redes.

Um indício de que as características topológicas selecionadas podem descrever conhecimento relevante sobre cidades é que a cidade de São Paulo está isolada das demais. Uma reação semelhante pode ser observada, em pequena escala, considerando as cidades de Campinas, Marília e Piracicaba, que estão separadas do grupo principal de cidades — localizado na parte esquerda da imagem. Acredita-se que esse comportamento está relacionado com a demografia das cidades; de modo que, em larga escala, as características topológicas podem inferir sobre a população das cidades, enquanto que, em pequena escala, podem apontar para bairros densamente ou escassamente povoados.

[3]Os detalhes sobre as categorias de tamanho são descritos nas Figuras 6a e 6b.

Na etapa subsequente, a cidade de São Paulo foi removida do conjunto de dados e o resultado desse processo foi ilustrado na Figura 2, considerando as técnicas de projeção PCA e Isomap, respectivamente; note que na imagem os valores foram normalizados no intervalo $[0, 1]$ a fim de facilitar a interpretação visual dos resultados.

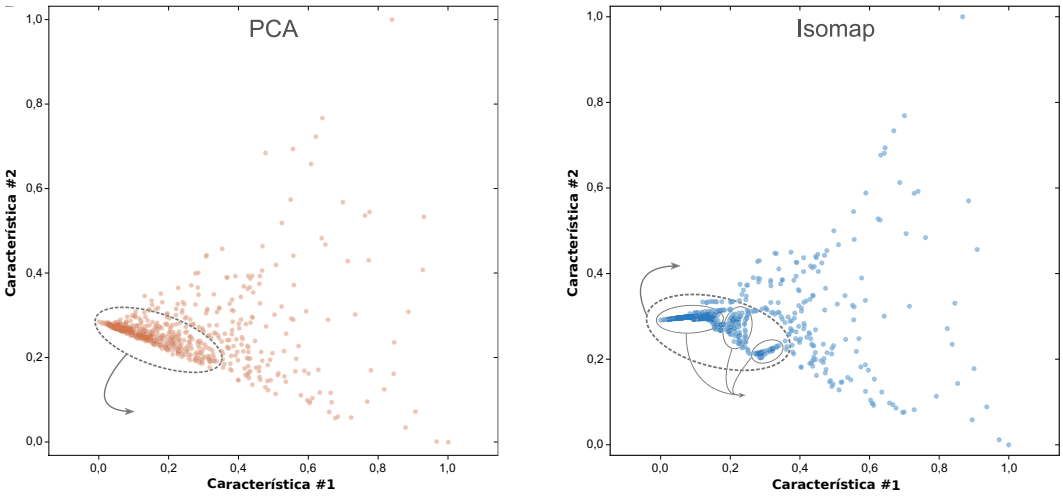


Figura 2: Projeção dos vetores de características das cidades usando PCA e Isomap.

As duas técnicas aplicadas mostram que os dados estão concentrados em uma pequena região de cada imagem, com poucos pontos esparsos ao longo dos eixos. A principal diferença entre as duas técnicas é que a Isomap implica na existência de múltiplas áreas com alta densidade, enquanto a PCA tem uma única área densa e muitos pontos esparsos. Isso é evidência de que cidades de pequeno porte tendem a se agrupar isolando cidades de médio e grande porte. Isomap, por outro lado, mostra que, apesar das cidades de pequeno porte serem semelhantes, elas têm particularidades que as divide em pequenos grupos dentro de um maior. Além disso, pode-se inferir que, por serem espalhadas, as cidades de tamanho médio e grande não possuem um padrão claro, mas ainda assim, elas podem compartilhar características comuns para serem exploradas pela análise de agrupamentos. Todavia, pode-se provar, usando correlação, que a topologia dos grafos pode inferir sobre a demografia das cidades do estado de São Paulo (ver Figure 3).

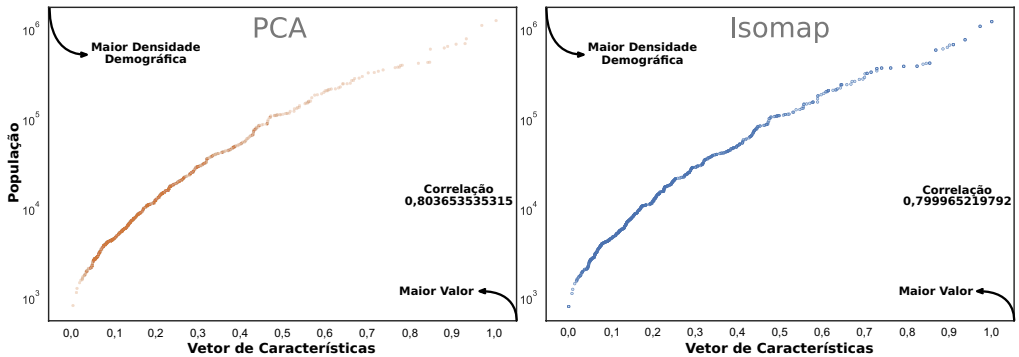


Figura 3: Teste de correlação entre o número de habitantes e as características das cidades projetadas em uma única dimensão usando ambas as técnicas PCA e Isomap.

Para avaliar a dependência entre os indicadores mediu-se sua correlação. Para esse fim, a dimensionalidade dos vetores de características foi reduzida para uma única dimensão. Sobre os valores resultantes, foram correlacionados os dados demográficos com as características unidimensionais de cada cidade. Como resultado, foi obtido 0,803 e 0,799 de correlação para PCA e Isomap, respectivamente. Ambos os valores indicam forte correlação entre os dados, permitindo afirmar que, no caso do estado brasileiro de São Paulo, as características topológicas e demográficas estão fortemente correlacionadas.

3.2. Análise de Agrupamentos

Na análise de agrupamentos utilizou-se KMeans variando o parâmetro da quantidade de grupos de 2 a 645 unidades, buscando pela configuração que proporciona a maior pontuação média de *Silhouette* (AVG) nos casos em que o *Dunn Index* (DNN) é maior do que 1. O resultado deste processo revelou que os dados são melhor agrupados em dois grupos, de modo que a cidade de São Paulo se separa das demais (ver Figuras 1 e 6c); este cenário possui os maiores valores de AVG e DNN, com AVG igual a 0,94 e DNN igual a 3,75 (ver Figura 4).

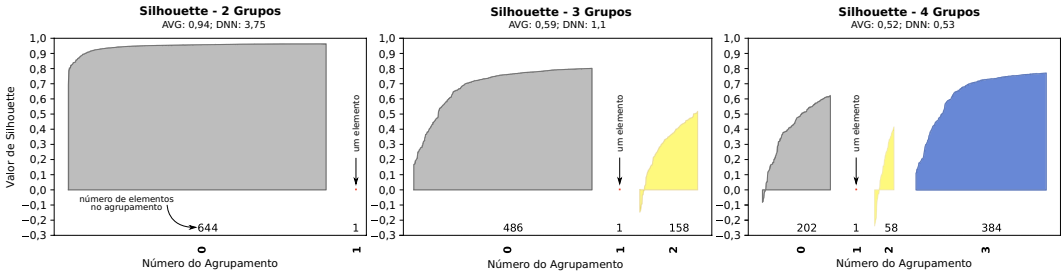


Figura 4: Avaliação da qualidade do agrupamento para todo o conjunto de dados.

Nota-se que, geralmente, a cidade de São Paulo está em um agrupamento isolado, enquanto que as outras cidades tendem a se agrupar, mesmo sendo geograficamente dispersas e notoriamente dissimilares umas às outras (ver Figuras 1 e 2). Ao remover São Paulo, é melhor dividir os dados em dois grupos (ver Figura 5), com os maiores valores de AVG e DNN sendo iguais a 0,59 e 1,10 respectivamente. Neste cenário, os agrupamentos aparentam um melhor equilíbrio quanto a quantidade de elementos por grupo, apontando para a existência de um padrão.

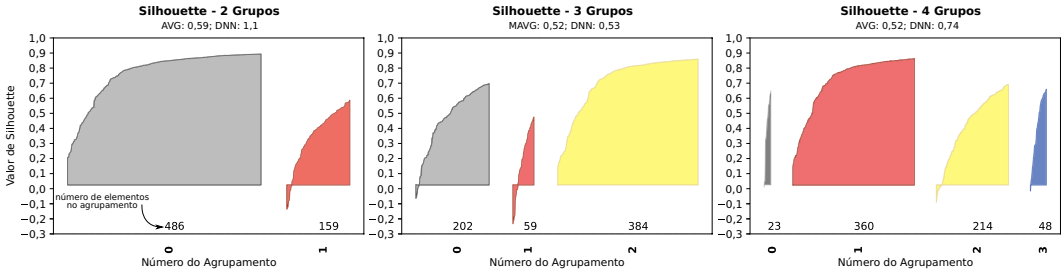
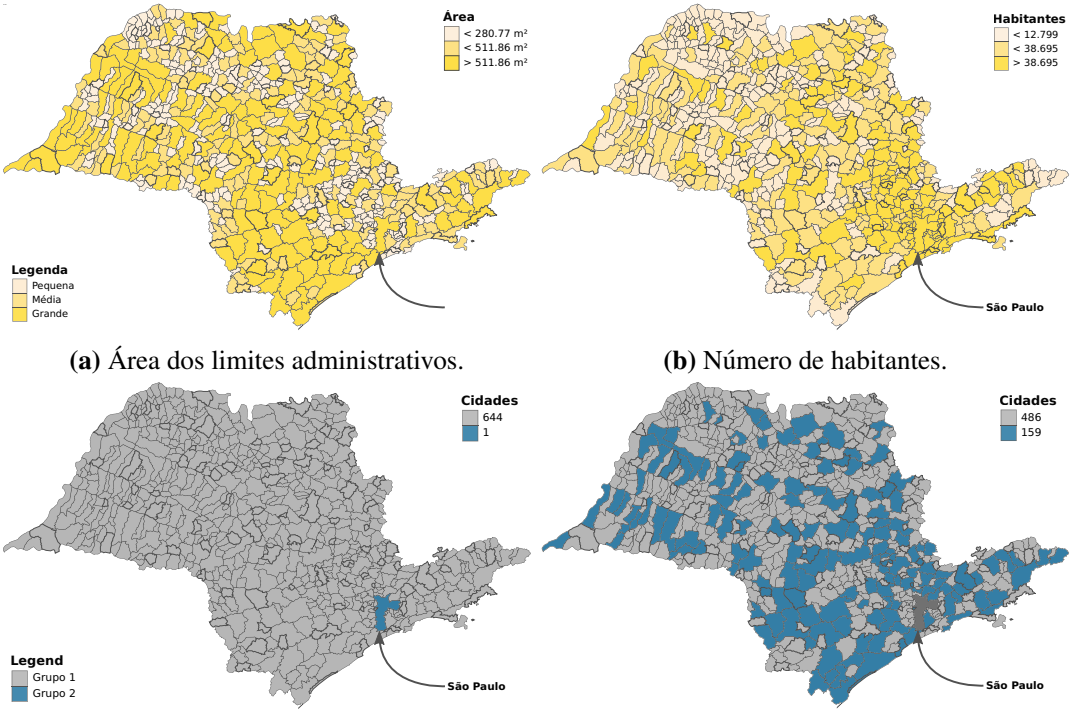


Figura 5: Avaliação da qualidade do agrupamento dos dados sem a cidade São Paulo.

Quanto ao último agrupamento, a relação que o favorece não está ligada ao número de habitantes, como a hipótese da seção anterior, mas sim à extensão territorial

(área em metros quadrados) de cada cidade. Observe que as características extraídas das redes estão conectadas a sua topologia a qual, por sua vez, está relacionada ao tamanho e geometria das malhas viárias. Além disso, 61,20% da população do estado está no primeiro grupo e 38,80% está no segundo (veja Figuras 5 e 6d). Não obstante, o segundo grupo parece ser povoado principalmente por cidades que são consideradas de extensão territorial média ou grande e apenas por algumas cidades pequenas (veja a Figura 6a).



(a) Área dos limites administrativos. **(b)** Número de habitantes.

(c) Agrupamento *com* São Paulo. **(d)** Agrupamento *sem* São Paulo.

Figura 6: Análise das cidades por meio de agrupamentos e indicadores urbanos relacionados à área dentro de seus limites administrativos e ao seu número de habitantes.

Esse padrão pode ser entendido como a forma com que as cidades se organizam em seu espaço disponível. De fato, em relação à extensão territorial, 61,54% das cidades do primeiro grupo são de tamanho pequeno, 25,78% são de tamanho médio e 12,58% são de tamanho grande; enquanto que 13,91% das do segundo grupo são de tamanho pequeno, 22,78% são de tamanho médio e 63,29% são de tamanho grande. Portanto, conclui-se que em relação a quantidade populacional as cidades do primeiro grupo podem ser consideradas pequenas e densamente povoadas, enquanto que as do segundo grupo podem ser consideradas grandes e escassamente povoadas.

4. Conclusão

Neste trabalho, foram analisadas características extraídas de 645 cidades que formam o estado de São Paulo. A metodologia aplicada baseou-se em processos de mineração de dados, com foco em projeção multidimensional e análise de agrupamentos, dividindo-se em processos de (i) Aquisição e Preparação de Dados, (ii) Extração e Seleção de Características, e (iii) Análise de Vetores de Características. Os resultados descrevem as

relações entre redes viárias, sua demografia e extensão territorial, explicando associações entre a topologia e indicadores urbanos das cidades. Mais precisamente, as contribuições deste trabalho estão na descrição de como a topologia da rede é capaz de revelar grupos de cidades com características semelhantes, na análise de correlação entre a quantidade de população das cidades e suas características, e no estudo de porquê as cidades se agrupam com outras distantes e não com aquelas as quais fazem fronteira.

Agradecimentos

Os autores são gratos ao CNPq (167967/2017-7), a FAPESP (2016/17078-0, 2016/17330-1 e 2017/08376-0) e a CAPES (10095541/M) pelo apoio financeiro a este trabalho.

Referências

Anderson, T. K. (2009). Kernel density estimation and k-means clustering to profile road accident hotspots. *Accident Analysis & Prevention*, 41(3):359–364.

Blumer, H. (1971). Social problems as collective behavior. *Social problems*, 18(3):298–306.

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., and Hwang, D. (2006). Complex networks: Structure and dynamics. *Physics Reports*, 424(4-5):175–308.

Chiang, C. (2003). *Statistical Methods of Analysis*. World Scientific.

Costa, L. F., Rodrigues, F. A., Travieso, G., and Boas, P. R. V. (2007). Characterization of complex networks: A survey of measurements. *Advances in Physics*, 56(1):167–242.

Costa, L. F., Travençolo, B. A. N., Viana, M. P., and Strano, E. (2010). On the efficiency of transportation systems in large cities. *EPL (Europhysics Letters)*, 91(1).

Crucitti, P., Latora, V., and Porta, S. (2006). Centrality measures in spatial networks of urban streets. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 73(3).

Domingues, G. S., Silva, F. N., Comin, C. H., and Costa, L. F. (2017). Topological characterization of world cities. *arXiv preprint arXiv:1709.08244*.

Grauwin, S., Sobolevsky, S., Moritz, S., Gódor, I., and Ratti, C. (2015). Towards a comparative science of cities: Using mobile traffic records in new york, london, and hong kong. In *Computational approaches for urban environments*, pages 363–387. Springer International Publishing.

Konstantopoulos, T. (2012). *Introduction to projective geometry*. Number September. Dover Publications.

Kremer, H., Kranen, P., Jansen, T., Seidl, T., Bifet, A., Holmes, G., and Pfahringer, B. (2011). An effective evaluation measure for clustering on evolving data streams. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 868–876, New York, NY, USA. ACM.

Li, X. and Parrott, L. (2016). An improved genetic algorithm for spatial optimization of multi-objective and multi-site land use allocation. *Computers, Environment and Urban Systems*, 59:184–194.

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297.

Masucci, A. P., Stanilov, K., and Batty, M. (2013). Limited Urban Growth: London's Street Network Dynamics since the 18th Century. *PLoS ONE*, 8(8).

Pan, G., Qi, G., Zhang, W., Li, S., Wu, Z., and Yang, L. T. (2013). Trace analysis and mining for smart cities: issues, methods, and applications. *IEEE Communications Magazine*, 51(6):120–126.

Porta, S., Crucitti, P., and Latora, V. (2006a). The network analysis of urban streets: A dual approach. *Physica A: Statistical Mechanics and its Applications*, 369(2):853–866.

Porta, S., Crucitti, P., and Latora, V. (2006b). The Network Analysis of Urban Streets: A Primal Approach. *Environment and Planning B: Planning and Design*, 33(5):705–725.

Porta, S., Latora, V., Wang, F., Strano, E., Cardillo, A., Scellato, S., Iacoviello, V., and Messora, R. (2009). Street Centrality and Densities of Retail and Services in Bologna, Italy. *Environment and Planning B: Planning and Design*, 36(3):450–465.

Scripps, J., Nussbaum, R., Tan, P.-N., and Esfahanian, A.-H. (2010). Link-Based Network Mining. In *Structural Analysis of Complex Networks*, pages 403–419. Springer Nature.

Spadon, G., Gimenes, G., and Rodrigues-Jr, J. F. (2017). Identifying Urban Inconsistencies via Street Networks. volume 108, pages 18 – 27. Elsevier BV. International Conference on Computational Science, ICCS 2017, 12-14 June 2017, Zurich, Switzerland.

Spiwok, V., Oborsky, P., Pazurikova, J., Keenek, A., and Kralova, B. (2015). Nonlinear vs. linear biasing in trp-cage folding simulations. *The Journal of Chemical Physics*, 142(11).

Strano, E., Nicosia, V., Latora, V., Porta, S., and Barthélemy, M. (2012). Elementary processes governing the evolution of road networks. *Scientific Reports*, 2.

Strano, E., Viana, M., Costa, L. F., Cardillo, A., Porta, S., and Latora, V. (2013). Urban Street Networks, a Comparative Analysis of Ten European Cities. *Environment and Planning B: Planning and Design*, 40(6):1071–1086.