Uso de técnicas de visualização no apoio à análise de dados urbanos para a melhoria da transparência pública

Leandro Miranda[1], José Viterbo[1], Flavia Bernardini[2], Cristiano Maciel[3], Raissa Barcellos[1], Daniela Trevisan[1]

[1]Instituto de Computação
Universidade Federal Fluminense (UFF) – Niterói, RJ, Brasil

[2]Instituto de Ciência e Tecnologia
Universidade Federal Fluminense (UFF) – Rio das Ostras, RJ, Brasil

[3]Universidade Federal de Mato Grosso – Cuiabá/MT, Brasil

{lmiranda,viterbo}@ic.uff.br, fcbernardini@id.uff.br

cmaciel@ufmt.br, {rbarcellos,daniela}@ic.uff.br

Resumo. *Dados urbanos são disponibilizados gradativamente em grande volume. Diante disso, há a necessidade de ferramentas para exploração, análise e descoberta de novos conhecimentos sobre esses dados para os cidadãos de forma transparente. O objetivo deste trabalho é apresentar uma discussão sobre a importância de diferentes técnicas de visualizações de dados urbanos, que permitem uma análise univariada, bivariada e multivariada. Para guiar nossa análise, conduzimos uma análise exploratória de um conjunto de dados socioeconômicos da cidade de Chicago, EUA. Podemos observar que as diferentes formas de visualização complementam o processo de análise dos dados.*

1. Introdução

Houve, nos últimos anos, um aumento do número de novas tecnologias que permitem a coleta e armazenamento de grandes volumes de informações, distribuídos em grandes conjuntos de dados, descrevendo várias informações relacionadas a aspectos urbanos como dados geográficos, demografia local, entre outros [Veljković et al. 2011]. Dados governamentais abertos permitem que dados urbanos sejam disponibilizados sem restrições de acesso e possibilitam a redistribuição em diferentes formatos. A utilização desses dados, com apoio de tecnologias ubíquas, modelos analíticos e novos métodos de visualização, podem oferecer soluções para a melhoria do ambiente urbano e da qualidade de vida do cidadão [Zheng et al. 2014]. Chicago, EUA, oferece um portal Web para divulgação de dados abertos. O portal permite acesso a uma grande variedade de dados do governo municipal e autoriza a exploração e interpretação das informações disponíveis por tecnologia desenvolvida com a finalidade de extração de dados relacionados à cidade. Em Chicago e outras cidades, a oferta de dados abertos ainda não facilita um apoio ao cidadão na interpretação de conjuntos de dados urbanos. Interpretar esses dados não é uma tarefa trivial, devido à massiva quantidade de dados brutos, e a busca por técnicas e metodologias que permitam a interpretação de informações implícitas e dedução de novos conhecimentos com o auxílio de visualizações de dados se torna imprescindível [Schoffelen et al. 2015].

A maioria dos portais de dados abertos disponíveis mundialmente é baseada em duas plataformas, o Socrata (`https://socrata.com/`), que provê um conjunto de ferramentas de análise e visualização de dados na nuvem para dados abertos governamentais; e o CKAN (`ckan.org`), que é uma aplicação *web* que oferece uma catalogação de dados abertos, nos quais oferecem acesso a esses dados diretamente ou por APIs. Os usuários que acessam portais construídos com base nessas ferramentas em geral encontram mecanismos básicos de tabelas e visualização univariada, que favorecem apenas a análise do comportamento de variáveis isoladas. No entanto, recursos que permitam a visualização dos dados considerando a combinação das dimensões (variáveis) podem melhorar a interpretação de informações implícitas ou auxiliar na dedução de novos conhecimentos. Segundo [Gama and Lóscio 2014], plataformas de dados abertos devem oferecer alguns serviços básicos adicionais, para os quais eles propõem uma camada de Serviços de Análise e Visualização de Dados, provendo ferramentas de visualização uni ou bidimensionais, já comuns nos portais; e ferramentas de suporte à descoberta de informações úteis, incluindo técnicas de visualização adicionais (incluindo multidimensionais) e/ou ferramentas de mineração de dados e técnicas estatísticas.

2. Objetivos da Pesquisa

O objetivo deste trabalho é discutir a importância da disponibilização de diferentes formas de visualizações de dados abertos, considerando visualização univariada (ou unidimensional); bivariada, por meio de métricas de correlações entre variáveis; e multivariada, por meio de visualizações utilizando algoritmos de clusterização. Essas diferentes formas de visualização de dados podem auxiliar no processo de descoberta de conhecimento a partir de dados por parte da população, já que potencialmente facilitam a interpretabilidade de dados urbanos por parte dos usuários com transparência. Para este fim, realizamos uma análise exploratória a partir de dados disponíveis no portal de dados abertos da cidade de Chicago, disponível em `http://data.cityofchicago.org/`, contendo informações atualizadas referentes a um total de 27 indicadores socioeconômicos de cada região (área) comunitária da cidade.

3. Contribuições Esperadas

Vários trabalhos da academia apresentam tecnologias para aperfeiçoar e auxiliar a população e os governantes na compreensão das características de espaços públicos. Na cidade de Nova York, em [Barbosa et al. 2014] foi apresentada uma forma de padronizar e integrar, por meio de correlações, um grande volume de dados fornecido pela população local. Em [Piccolo et al. 2015], foram estudadas as características da vizinhança etnorracial das comunidades da cidade de Boston, nos Estados Unidos, com relação a diabetes tipo 2 (T2DM). Nesses trabalhos, verifica-se uma massiva coleta de dados referentes aos seus domínios de pesquisa, sendo a maioria deles fornecedor de análises de dados univariados e bivariados. O estudo de ferramentas de visualização envolvendo dados multivariados para dados abertos e apoio à transparência é carente em quantidade.

A contribuição deste trabalho está na capacidade de ressaltar a importância de análise de dados com o apoio de diferentes tipos de visualização de dados. Para a análise exploratória utilizando a base de dados de Chicago, foram utilizados, além de atributos de identificação das regiões — Código da comunidade e Nome da comunidade —,

os seguintes atributos:(1) Taxa de natalidade (*Birth Rate*); (2) Taxa de latrocínios (*Assault/Homicide*); (3) Taxa de pessoas abaixo do nível da pobreza (*Below Poverty Level*); (4) Taxa de pessoas com dependência financeira (*Dependency*); (5) Taxa de pessoas sem ensino superior (*No High School Diploma*); (6) Renda per capita (*Per Capita Income*); e (7) Taxa de desemprego (*Unemployment*). Para as visualizações de dados construídas neste trabalho, é indicado que as variáveis sejam normalizadas nas situações em que há atributos que apresentem intervalos de valores bastante distintos. Assim, todas as variáveis da base de dados foram normalizadas para o intervalo $[0, 1]$.

4. Resultados já Alcançados

Inicialmente, construímos visualizações univariadas utilizando a própria ferramenta do portal que disponibilizou os dados, apresentadas nas Figuras 1 (a), (b) e (c). Individualmente, representam a distribuição de atributos por comunidade de Chicago. Podemos inferir alta taxa de renda per capita nas regiões 7, 8, 32 e 33, de desemprego nas regiões 37 e 67, e de pessoas sem ensino superior nas regiões 30, 58 e 63; e há baixa taxa de renda per capita nas regiões 5, 6, 7 e 32, desemprego nas regiões 5, 6, 7 e 32, e de pessoas sem ensino superior nas regiões 6, 7, 8, 32, 41, 72 e 74. Avaliando-se as imagens em conjunto, segundo o comportamento dos gráficos, pode-se afirmar que aparentemente a taxa de desemprego está mais relacionada com a taxa de pessoas sem ensino superior do que com a renda per capita, porém não há garantia disso.

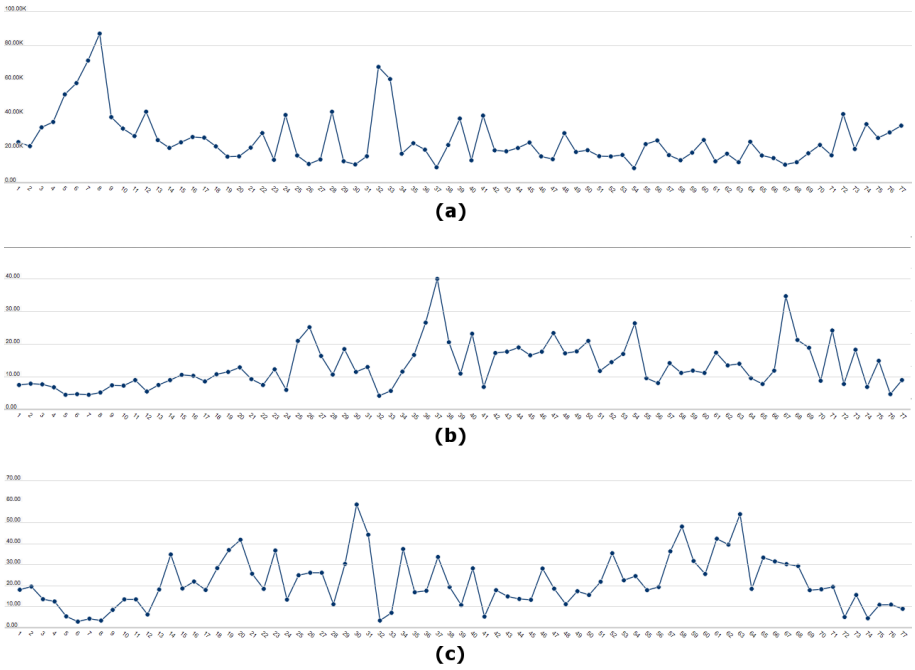

Figura 1. Gráficos que apresentam a distribuição da (a) renda per capita (em milhares de dólares), (b) taxa de desemprego (%) e (c) taxa de pessoas sem ensino superior (%).

Construímos outro tipo de visualização, também univariada, não disponível no portal de Chicago, apresentada nas Figuras 2 (a), (b) e (c). A visualização utiliza uma escala de cores sobre o mapa da área urbana de Chicago. Quanto mais escura a região, maior o valor da variável analisada. Esses gráficos permitem um entendimento mais claro

da distribuição geográfica das três variáveis pela área de Chicago. A análise individual de cada gráfico permite identificar facilmente as regiões onde a renda per capita, a taxa de desemprego e a taxa de pessoas sem ensino superior são mais altas ou mais baixas. A análise das imagens em conjunto permite identificar a área formada pelas regiões 7, 8, 32 e 33 e seu entorno como área próspera da cidade, com elevada renda per capita, baixa taxa de desemprego e baixa taxa de pessoas sem ensino superior, conhecimento esse que não pôde ser inferido na primeira visualização.

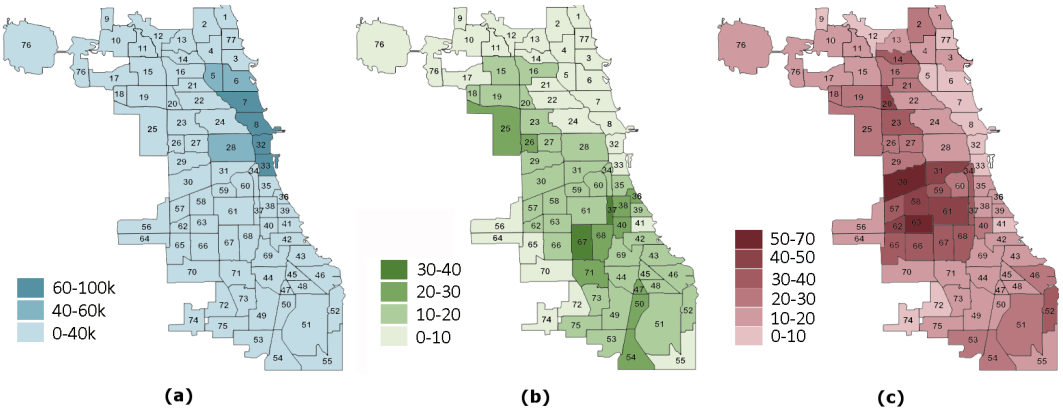

Figura 2. Mapas da cidade de Chicago que representam a distribuição da (a) renda per capita (em milhares de dólares), (b) taxa de desemprego (%) e (c) taxa de pessoas sem ensino superior (%).

Para visualizar a relação entre duas variáveis, calculamos a matriz de correlação linear de Pearson entre pares de atributos do conjunto de dados, e construímos a visualização dessa matriz na forma de matriz de cores — Figura 3. Os tons vermelhos indicam uma correlação positiva e os tons azuis indicam uma correlação negativa. Pintamos a diagonal principal da matriz de vermelho pois $\rho(X_i, X_i) = 1$ para quaisquer $i \in [1, N]$. Quanto mais intensa a cor, maior o valor absoluto da correlação. Pode-se perceber nessa figura uma alta correlação entre os atributos X_1 **– taxa de natalidade** e X_5 **– taxa de pessoas sem ensino superior**; X_2 **– taxa de latrocínio** e X_3 **– taxa de pessoas abaixo do nível da pobreza**; X_2 **– taxa de latrocínio** e X_4 **– taxa de pessoas com dependência financeira**; X_2 **– taxa de latrocínio** e X_7 **– taxa de desemprego**; X_4 **– taxa de pessoas com dependência financeira** e X_7 **– taxa de desemprego**; e X_3 **– taxa de pessoas abaixo do nível da pobreza** e X_7 **– taxa de desemprego**.

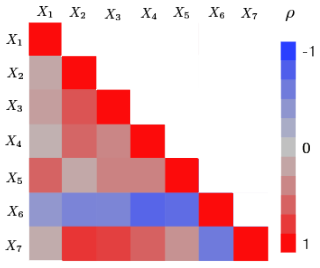

Figura 3. Mapa de cores com correlações entre as variáveis da base de dados de Chicago normalizada. ρ é a legenda da escala de cores da correlação.

A análise bivariada fornece novos conhecimentos, mas ainda assim carece de uma possibilidade de descoberta de conhecimento sobre toda a região de Chicago, que pode ser auxiliada pela visualização de dados multivariada. Pode-se utilizar técnicas para encontrar a similaridade entre os objetos utilizando algoritmos de clusterização. Utilizamos o algoritmo K-médias, algoritmo tradicional de clusterização, e como métrica de distância entre as comunidades, utilizamos a distância euclidiana. K, o número de clusters presente na partição construída pelo K-médias, é parâmetro do algoritmo. Para escolher o melhor valor para o parâmetro K utilizamos o critério CCC (*Cubic Clustering Criterion*) para estimar o melhor valor de K, que selecionou $K = 8$, utilizado para nossa análise. Na Figura 4 é apresentado o mapa da cidade de Chicago, onde cada cluster de região foi pintado de uma cor diferente. Observamos que o cluster 8, conjunto formado pelas regiões 6, 7, 8 e 32, identifica as regiões com bons índices sociais, e o cluster 4 agrega regiões onde os índices são muito ruins, revelando uma área que necessita de maior atenção do governo municipal. Visualizações uni e bivariadas não permitiram essas conclusões.

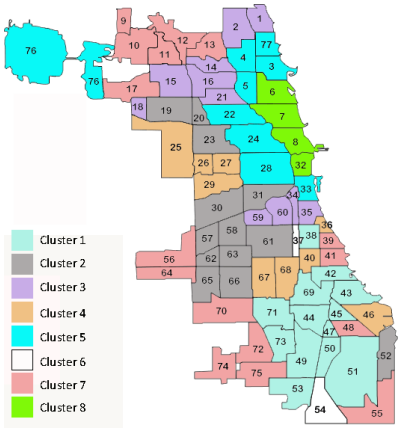

Figura 4. Representação de Chicago dividida em 8 conjuntos (clusters) de regiões que apresentam padrões de similaridade representadas em cores diferentes.

A simples análise da visualização não permite uma compreensão clara de quais atributos mais influenciam na formação de cada cluster. Os valores dos centróides de cada cluster, ou seja, o ponto médio de cada cluster, pode auxiliar nessa análise. Esses são apresentados na Figura 5, onde as colunas correspondem aos atributos $X_i, i = 1, ..., 7$; cada linha corresponde a um centróide de um cluster $c_k, k = 1, ..., 8$; e a última coluna apresenta o número de regiões por cluster — $|c_k|$. Os valores apresentados na Figura 5 apresentam grande importância na representação de cada cluster. Quanto mais escuro o tom de azul, mais próximo o valor do centróide naquele atributo se aproxima de 1; e quanto mais claro, mais próximo de 0. Analisando conjuntamente as Figuras 4 e 5, observamos que o cluster 1 une regiões com índices de latrocínio consideráveis e alta dependência financeira; o cluster 2 apresenta comunidades com altos índices demográficos e taxa de pessoas com alta dependência financeira e com pouca escolaridade; o cluster 3 apresenta características semelhantes ao cluster 2, porém com índices demográficos e taxas menores que as regiões do cluster 2; o cluster 4 une regiões mais perigosas para morar, com altos índices de latrocínio, pobreza, dependência financeira, baixa escolaridade e desemprego; o cluster 5 une regiões mais populosas e com a segunda maior renda per capita entre as todas as regiões de Chicago; o cluster 6 une regiões mais pobres por

apresentar a mais alta proporção de pessoas vivendo no nível abaixo da linha da pobreza, consequência de altas taxas de latrocínio, baixa escolaridade e desemprego; o cluster 7 une regiões com alta dependência financeira; e o cluster 8 une as regiões mais ricas de Chicago, concentrando alta renda per capita e baixíssima taxa de desemprego.

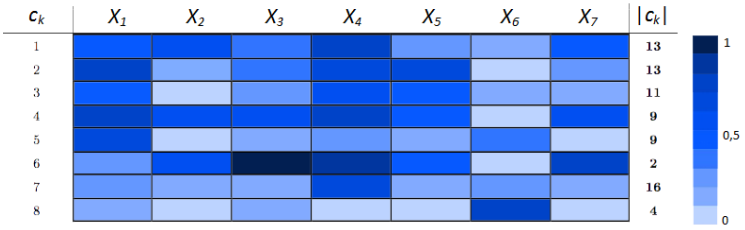

Figura 5. Centróides por cluster c_k construído para $K = 8$ utilizando a base normalizada. X_i são os atributos da base de dados com $i \in [1..7]$ e $|c_k|$ é o número total de regiões pertencentes a cada cluster. Ao lado da figura, é exibida a escala de cores, variando conforme o valor do centróide em um dado atributo.

5. Conclusões

Analisamos neste trabalho técnicas de visualização de dados univariada, bivariada e multivariada, utilizando um conjunto de dados socioeconômicos da cidade de Chicago, EUA, para apresentar a potencialidade de uso dessas diversas maneiras de visualização em dados abertos. As diferentes técnicas de visualizações permitiram inferir diferentes conhecimento complementarmente. Em trabalhos futuros, pretendemos realizar avaliação das visualizações com usuários possuidores de pouco ou nenhum conhecimento sobre análise de dados, para verificar se as ferramentas oferecem melhora no processo de interpretação de dados urbanos. Ainda, técnicas para clusterização considerando a regionalização dos dados também estão sendo pesquisadas, com o objetivo de facilitar a interpretação dos dados.

Referências

Barbosa, L., Pham, K., Silva, C., Vieira, M. R., and Freire, J. (2014). Structured open urban data: Understanding the landscape. *Big Data*, 2(3):144–154.

Gama, K. and Lóscio, B. F. (2014). Towards ecosystems based on open data as a service. In *Proceedings of the 16th International Conference on Enterprise Information Systems*, pages 659–664.

Piccolo, R. S., Duncan, D. T., Pearce, N., and McKinlay, J. B. (2015). The role of neighborhood characteristics in racial/ethnic disparities in type 2 diabetes: Results from the boston area community health (bach) survey. *Social Science & Medicine*, 130(0):79 – 90.

Schoffelen, J., Claes, S., Huybrechts, L., Martens, S., Chua, A., and Moere, A. V. (2015). Visualising things. perspectives on how to make things public through visualisation. *CoDesign*, 11(3-4):179–192.

Veljković, N., Bogdanović-Dinić, S., and Stoimenov, L. (2011). Municipal open data catalogues. In *Conference for E-Democracy and Open Government*, page 195.

Zheng, Y., Capra, L., Wolfson, O., and Yang, H. (2014). Urban computing: Concepts, methodologies, and applications. *ACM Trans. Intell. Syst. Technol.*, 5(3):38:1–38:55.