# APPEL: Uma extensão do Kepler para enriquecimento de dados geoespaciais

**Gabriel T. P. Coimbra**[1], **Cláudio Gustavo S. Capanema**[1], **Fabrício A. Silva**[1], **Thais R. M. Braga Silva**[1]

[1]Universidade Federal de Viçosa (UFV), Florestal, Brasil

{gabriel.coimbra,claudio.capanema,fabricio.asilva,thais.braga}@ufv.br

***Abstract.*** *The use of georeferenced data in the most different contexts has enabled promising studies, and aroused the interest of companies and research groups. This has translated into a demand for tools capable of manipulating spatial data. In this paper, we present the APPEL, a solution for data enrichment capable of performing reverse geocoding of large volumes of georeferenced data and providing hot-spots based on correlations between statistical data from each Brazilian municipality.*

***Resumo.*** *A utilização de dados georreferenciados nos mais diferentes contextos tem viabilizado estudos promissores, e despertado o interesse de empresas e grupos de pesquisa. Isso se traduziu em uma demanda por ferramentas capazes de manipular dados espaciais. Neste trabalho, apresentamos o APPEL, uma solução para enriquecimento de dados capaz de realizar geocodificação reversa de grandes volumes de dados georreferenciados e fornecer hot-spots com base em correlações entre dados estatísticos de cada município brasileiro.*

## 1. Introdução

A popularização dos dispositivos móveis trouxe uma crescente geração de dados georreferenciados provenientes do sensor de GPS e da rede de telefonia. Dessa forma, a possibilidade de se obter a localização de milhares de usuários tem sido um aspecto chave para o desenvolvimento em estudos envolvendo a mobilidade urbana [Bazzani et al. 2011, Zhao et al. 2016]. Além disso, segundo [Shahrour 2018] o conceito de *Smart Cities* está intimamente relacionado à utilização de dados georreferenciados, sendo importantes para a compreensão e a melhoria do ambiente urbano.

Além da informação espacial, outros aspectos relevantes para auxiliar nas tomadas de decisões envolvendo planejamento urbano em cidades inteligentes são as características demográficas, econômicas e sociais das cidades. Essas informações, que no Brasil são coletadas e disponibilizadas pelo IBGE (Instituto Brasileiro de Geografia e Estatística), são ricas fontes de dados. Porém, atualmente a correlação entre dados georreferenciados de diversas fontes e dados do IBGE não é possível sem um significativo esforço dos envolvidos.

Neste trabalho, é proposto o APPEL (*Augmented Point to Polygon Extension Layer*), uma extensão à ferramenta *Kepler* [1] para o enriquecimento de dados georreferenciados. Trata-se de uma solução apta a identificar, eficientemente, a região em que um

[1]https://kepler.gl/. Acesso em: 23/07/2019

dado ponto geográfico está contido. A partir desse processo, é possível correlacionar pontos com informações do IBGE das áreas correspondentes. Esses aspectos têm o objetivo de aprimorar a tarefa de análise de dados espaciais sem elevar a demanda por recursos computacionais e humanos.

## 2. Trabalhos relacionados

Nesta seção, são apresentadas as características dos sistemas com suporte a informação espacial mais comumente utilizados. As existentes são categorizadas de acordo com o seu foco: visualização, armazenamento e processamento de dados.

Dentre as principais ferramentas de visualização de dados georreferenciados se destacam as de código aberto QGIS [2], *Metabase* [3] e *Kepler*, além do ArcGIS [4], que é uma ferramenta proprietária. O QGIS, ArcGIS e *Metabase* dispõem de uma grande variedade de funcionalidades, sendo possível manipular informações de diferentes fontes (e.g. bancos de dados, arquivos locais e servidores *online*). A visualização é focada em dados geoespaciais (e.g. combinações de camadas de pontos, polígonos e contornos) para o ArcGIS, QGIS e *Kepler*. O *Metabase*, no entanto, fornece uma interface para bancos de dados SQL, logo possui ferramentas específicas para tabelas (e.g. filtros e agrupamentos). Apesar de possuir uma menor variedade de funcionalidades, o *Kepler* corresponde a um sistema *web* de implementação simples, o que o torna facilmente extensível. Além disso, sua interface intuitiva, ou seja, sem a sobrecarga de informações desnecessárias permite que usuários não técnicos possam explorar os seus dados.

Em relação ao armazenamento de dados, o *PostgreSQL* por meio da sua extensão *PostGIS* é referência em bancos de dados espaciais, uma vez que permite que consultas envolvendo geometrias sejam executadas rapidamente.

Apesar de prover bons recursos para visualizações, *dCluster* [Capanema et al. 2017] se destaca pela variedade de opções de algoritmos para análise dos dados. O *dCluster* é um sistema *web* e gratuito que oferece medidas de estatística descritiva para cada atributo do conjunto de dados, associadas a diversas possibilidades de gráficos como barras, dispersão, mapas de calor dentre outros. Além disso, a biblioteca *Geopandas* [5] da linguagem *Python* tem ganhado destaque, já que permite a criação de estruturas de dados tabulares capazes de indexar geometrias e realizar operações espaciais.

## 3. O APPEL

Essa seção descreve os principais aspectos do trabalho, sendo que na seção 3.1 é apresentada uma visão geral do sistema. O algoritmo de geocodificação reversa é apresentado na seção 3.2. Por fim, na seção 3.3 são explicadas as possíveis visualizações dos dados.

### 3.1. Visão Geral

A figura 1 ilustra a arquitetura do sistema. Para utilizá-lo, primeiramente o usuário deve acessar o site através do navegador e enviar o seu conjunto de dados utilizando a interface do *Kepler*. Esses dados devem possuir os campos latitude e longitude, e opcionalmente, outros atributos. Em seguida, através da camada APPEL, o conjunto de dados é

[2]https://www.qgis.org/pt_BR/site/. Acesso em: 23/07/2019
[3]https://www.metabase.com/. Acesso em: 23/07/2019
[4]https://www.arcgis.com/index.html. Acesso em: 23/07/2019
[5]http://geopandas.org/. Acesso em: 25/07/2019.

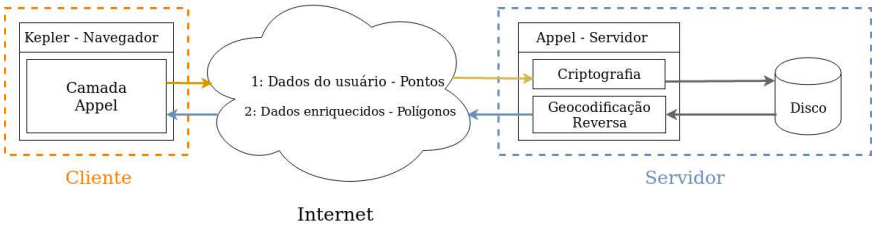


**Figura 1: Diagrama de arquitetura do sistema.**

enviado para o servidor, que é responsável pela criptografia, armazenamento e processamento. O cliente então requisita a transformação dos dados e o servidor responde com as informações processadas. Diferentemente do *Kepler*, parte do processamento do APPEL é realizado em um servidor. Logo, foi necessário criptografar tanto o conjunto de dados transmitidos pela rede, quanto os armazenados em memória persistente no servidor, considerando que a solução é para ser utilizada por múltiplos usuários simultaneamente.

### 3.2. Geocodificação Reversa

A geocodificação reversa é um termo comumente utilizado para indicar a transformação de uma coordenada em um endereço ou nome de um estabelecimento. Neste trabalho, no entanto, o termo geocodificação reversa será utilizado em referência à transformação de coordenadas em nomes de municípios brasileiros.

Em geral, determinar a região que contém um dado ponto geográfico é um processo computacionalmente caro, quando não são utilizados índices espaciais. Dessa forma, nesta seção é apresentada uma proposta para tornar mais eficiente o processo de geocodificação, tendo como base a utilização de atributos (e.g. população, área) e a estrutura geográfica das regiões. Para tanto, os polígonos a serem pesquisados (no caso repre-

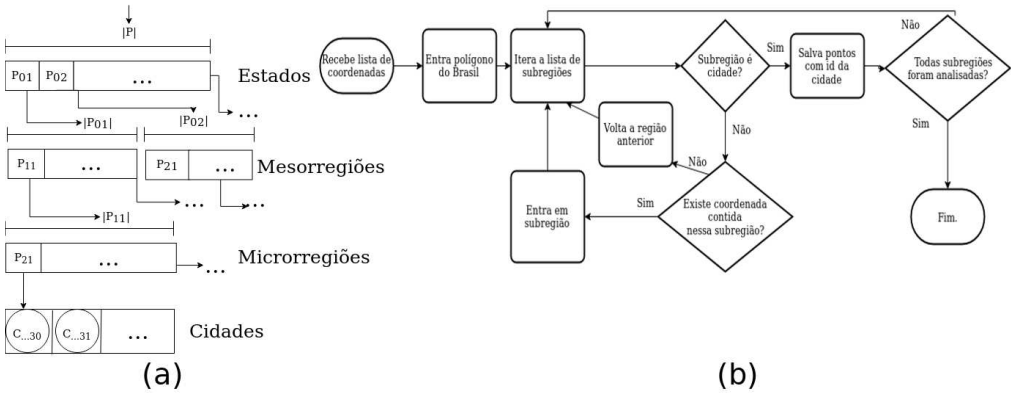


**Figura 2: (a) As letras $P$ e $C$ representam, respectivamente, os polígonos de regiões e cidades. Já $|P_{ij}|$ enumera a quantidade de subregiões contidas na região. (b) Fluxograma do algoritmo que percorre a estrutura em (a).**

sentando as cidades) são organizados através de uma árvore, semelhante à *R-Tree*. Porém, os seus níveis são pré-definidos com base em estados, mesorregiões, microrregiões e municípios (Figura 2(a)). Além disso, considera-se que é mais provável que os pontos geográficos fornecidos pelos usuários se localizem em regiões mais populosas devido às

aplicações desse sistema (e.g. infraestrutura, mobilidade). Portanto, essas áreas são as primeiras a serem pesquisadas em cada nível da árvore. Além disso, os polígonos das regiões e cidades foram simplificados utilizando um algoritmo da biblioteca GEOS que preserva a topologia [6]. Em suma, a simplificação diminuiu em 1% a acurácia em alguns casos. Mas, por outro lado, é medida uma redução de 98% da memória necessária para armazenar os polígonos.

Na figura 2 (b) é mostrado o fluxo para o processo de geocodificação. É importante notar que a operação que verifica se um ponto está contido em uma região utiliza operações de ponto em polígono, como o traçamento de raios [Haines 1994]. Essas operações são computacionalmente caras e o objetivo desse algoritmo é minimizar a sua utilização. Para isso, é empregada a estrutura de dados da figura 2(a). Este algoritmo procura primeiro a qual estado os pontos pertencem na lista de estados, depois a qual mesorregião estes pontos estão localizados dentro desse estado. A região de interesse diminui em área até ser encontrado o município ao qual o ponto pertence. Como as listas das regiões estão ordenadas por população, é esperado que as primeiras regiões já sejam suficientes para encontrar a maior parte dos pontos. Mesmo assim, na seção 4 é possível observar que o algoritmo tem um resultado satisfatório em casos onde a distribuição de pontos é aleatória, isso é, sem considerar a população daquela região. Para manter a eficiência para grandes volumes de dados, as coordenadas são divididas igualmente para serem tratadas em vários processos paralelos.

### 3.3. Correlações

O sistema APPEL utiliza os recursos de visualização de camadas da ferramenta *Kepler* para apresentar os dados enriquecidos, ou seja, correlacionados com as informações das cidades. A partir das informações processadas no servidor, é possível estabelecer correlações entre os demais dados fornecidos pelo usuário e os polígonos dos municípios do Brasil.

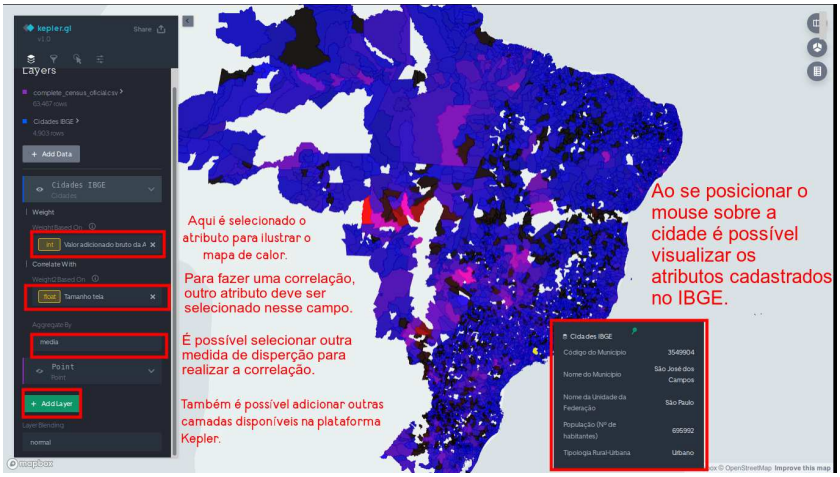


**Figura 3: Nessa figura é mostrada a correlação entre o atributo "Valor total adicionado bruto da Agropecuária" e um atributo do conjunto de dados do usuário. A figura também mostra as funcionalidades do APPEL.**

[6] http://geos.refractions.net/ro/doxygen_docs/html/classgeos_1_1simplify_1_1TopologyPreservingSimplifier.html

Através de um gráfico de *hot-spots* de calor (Figura 3), a cor de cada polígono varia de acordo com os valores de um atributo previamente selecionado para realizar a correlação, considerando os pontos contidos em cada região. E, como fonte de informações, o usuário tem a opção de usar atributos sobre cada cidade ou informações estatísticas sobre o conjunto de pontos que estão nesse município (e.g. máximo, mínimo, desvio padrão e média). Já para as visualizações, existem duas possibilidades: gerar o mapa de calor selecionando apenas um atributo ou escolher dois atributos das duas fontes de informações para ilustrar a correlação entre eles de forma geográfica.

Ao se fazer correlações entre dois atributos, na figura 3, o primeiro atributo da correlação corresponde a tonalidade de vermelho, enquanto que o segundo corresponde a intensidade do azul. Caso ambos atributos sejam significativos a cor se aproximará da cor-de-rosa. As cidades de cor branca não tiveram nenhum ponto detectado.

## 4. Testes

Para avaliar o desempenho da proposta deste trabalho mediante soluções bem conhecidas da literatura, foram executados diferentes testes, variando-se o volume e a distribuição dos dados. O banco de dados *PostgreSQL* através da extensão *PostGIS* e a biblioteca *Geopandas*, da linguagem *Python*, foram as soluções base selecionadas para a comparação.

O principal desafio envolvido no desenvolvimento do APPEL é o tempo de resposta da geocodificação reversa. Como o objetivo é que a camada criada seja utilizada posteriormente no *Kepler*, essa operação deve ser eficiente. Para a comparação com as outras abordagens ser justa, o APPEL foi executado sem paralelismo, uma vez que os outros sistemas não utilizam esse recurso no contexto apresentado.

O *PostGIS* realiza a tarefa de geocodificação através da função *ST_Contains* [7]. Além de se considerar o tempo de execução dessa função, também é calculado o tempo de inserção dos pontos enviados em uma tabela temporária, de modo a obter o desempenho do banco. Isso ocorre porque, analogamente, o sistema APPEL recebe novas coordenadas a cada nova requisição do usuário. Por outro lado, os polígonos de cada cidade do Brasil são armazenados de forma persistente em ambos sistemas, já que suas geometrias são fixas. Assim como o *PostGIS*, a biblioteca *GeoPandas* também é capaz de criar índices espaciais sobre geometrias, e dessa forma, é utilizada como referência no contexto de execução em memória principal.

Para os testes, foram utilizados duas categorias de conjuntos de dados. Os testes Proporcionais e Aleatórios correspondem a dados gerados artificialmente, sendo que no primeiro a quantidade de pontos em cada cidade é proporcional às suas respectivas populações e no segundo o número de pontos é totalmente aleatória, com distribuição uniforme ao longo de todo o território.

A tabela 1 apresenta os tempos de execução, em segundos, de cada método utilizando os conjuntos de testes. A acurácia de todos os métodos se manteve próxima de 99% em todos os casos. Há uma pequena perda de acurácia no APPEL, devido a não apenas os polígonos das cidades serem simplificados como nos outros métodos, os polígonos dos estados, micro e mesorregiões, também são simplificados.

[7] https://postgis.net/docs/ST_Contains.html. Acessado em 25/07/2019.

| Teste | Pontos | APPEL (s) | *PostGIS* (s) | *GeoPandas* (s) |
|---|---|---|---|---|
| Proporcionais | 10000 | $0,57 \pm 0,01$ | $1,43 \pm 0,03$ | $4,46 \pm 0,28$ |
| | 100000 | $2,18 \pm 0,12$ | $45,79 \pm 0,85$ | $11,46 \pm 0,36$ |
| | 1000000 | $14,25 \pm 0,01$ | $435,78 \pm 2,22$ | $67,88 \pm 0,48$ |
| Aleatórios | 10000 | $0,82 \pm 0,01$ | $4,3 \pm 0,24$ | $4,52 \pm 0,02$ |
| | 100000 | $2,8 \pm 0,01$ | $44,21 \pm 0,92$ | $7,59 \pm 0,06$ |
| | 1000000 | $20,92 \pm 0,02$ | $439,77 \pm 10,73$ | $72,47 \pm 0,57$ |

**Tabela 1: Comparação de tempo médio de pesquisa com *PostGIS* e *GeoPandas* com desvio padrão.**

O desempenho do APPEL se manteve melhor do que os outros métodos, especialmente quando se consideram os dados Proporcionais. É importante notar que as pesquisas no *GeoPandas* e no *PostGIS* são rápidas quando há indexação espacial dos pontos, porém, é demandado um certo tempo e memória para a construção da estrutura *R-tree*. Isso aumenta o tempo de processamento em um contexto no qual é necessário inserir novos dados a cada requisição.

## 5. Conclusão e trabalhos futuros

Nesse trabalho foram apresentados os resultados parciais de uma solução capaz de gerar visualizações de correlações entre dados georreferenciados diversos com informações das cidades fornecidas pelo IBGE. Para tanto, foi desenvolvido um método de geocodificação reversa, que se mostrou eficiente diante de outras abordagens. Como trabalhos futuros, pretende-se diminuir a granularidade da geocodificação reversa para procura dos pontos em setores censitários dentro das cidades, assim como aumentar o alcance do sistema para outros países além do Brasil.

## Referências

Bazzani, A., Giorgini, B., Gallotti, R., Giovannini, L., Marchioni, M., and Rambaldi, S. (2011). Towards congestion detection in transportation networks using gps data. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust*, pages 1455–1459.

Capanema, C. G. S., Silva, F. A. S., and Silva, T. R. M. B. (2017). DCluster: Um sistema para análise exploratória de grandes volumes de dados georreferenciados. In *Satellite Events of the 32nd Brazilian Symposium on Databases (SBBD)*.

Haines, E. (1994). Point in polygon strategies. *Graphics gems IV*, 994:24–26.

Shahrour, I. (2018). Use of gis in smart city projects - 04/10/2018.

Zhao, K., Tarkoma, S., Liu, S., and Vo, H. (2016). Urban human mobility data mining: An overview. In *2016 IEEE International Conference on Big Data (Big Data)*, pages 1911–1920. IEEE.