ERSI-RJ 2016

# W-SAGE

## Ferramenta Web para Análise de Dados Geoespaciais

Raul Sena Ferreira (UFRJ)
Carlos E. R. de Mello (UNIRIO)

ERSI-RJ 2016

# Introdução

A informação geográfica tem grande importância em diversas áreas como, marketing, agricultura, meio ambiente, saúde, planejamento urbano entre outros

Ajuda na tomada de decisões e estratégias

Agrega valor como um meio de representação visual mais expressiva do que uma representação discreta

ERSI-RJ 2016

# Introdução

Dados geográficos não possuem distribuição conhecida a priori mas podem revelar muito sobre um determinado domínio

Estimar densidades em cima de dados geográficos pode ajudar a determinar, de forma mais precisa, a probabilidade de ocorrência de um determinado fenômeno

ERSI-RJ 2016

# Objetivo

Construir uma ferramenta veloz com uma interface intuitiva para análise de dados geográficos

Utilizar um método estimador de densidade não-paramétrico para tornar possível inferir certas condições sobre os dados observados

Utilizar técnicas de paralelização em GPU com o intuito de aumentar a eficiência no processamento de algoritmos inferenciais estatísticos

Facilitar o reaproveitamento da ferramenta em outros sistemas mais complexos de visualização de dados

ERSI-RJ 2016

# Trabalhos relacionados

Visualização de dados geográficos voltado para auxiliar o combate ao tráfico humano fazendo uma análise sobre o problema de lavagem de dinheiro

- http://www.tandfonline.com/doi/full/10.1080/13658816.2015.1032294

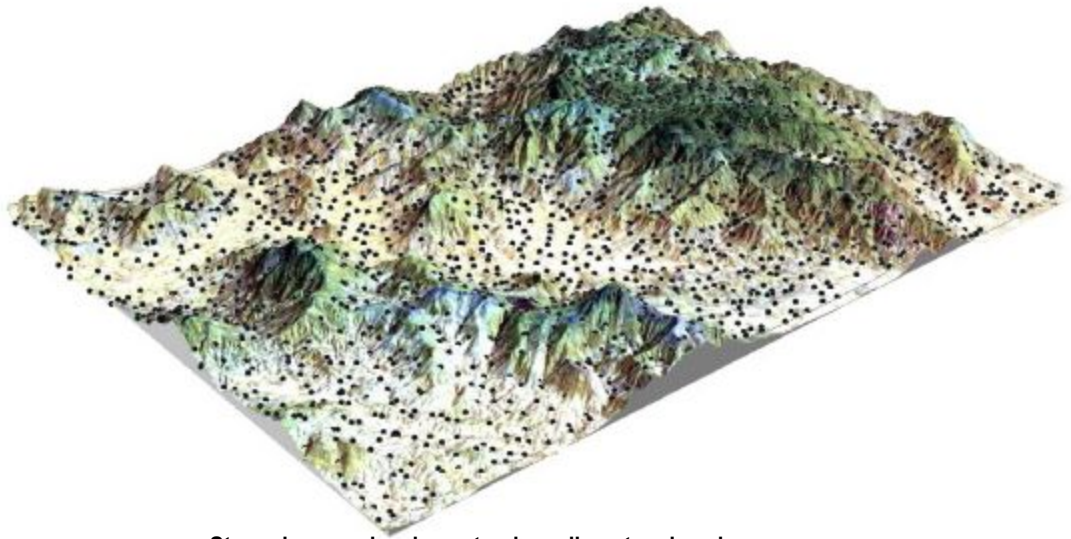
Workflow para tentar lidar com problemas geográficos que possuem a necessidade de computação intensa, usando métodos de paralelização tanto na parte de visualização quanto na parte de processamento

- http://gradworks.umi.com/10/11/10110968.html

ERSI-RJ 2016

# Trabalhos relacionados

Ferramentas de visualização para auxiliar o descobrimento de padrões geoquímicos na área de geologia, como anomalias e determinadas restrições dentro de uma região



**Streaming com locais contendo sedimentos de cobre**
Fonte: http://www.sciencedirect.com/science/article/pii/S0012825216300721

ERSI-RJ 2016

# Tecnologias







ERSI-RJ 2016

# Visualização geográfica

Kernel Density Estimation

- Uma das técnicas de estimativa de densidade mais comuns e bastante usada para normalizar e suavizar distribuições

Clusterização e rasterização de pontos

- Rasterização é a conversão de uma imagem vetorial em uma imagem raster (pixels ou pontos). Dessa forma os pontos deixam de ser objetos e passam ser imagens
- Clusterização é usada para evitar que muitos pontos venham sobrecarregar a visualização dos dados no lado cliente

ERSI-RJ 2016

# Visualização geográfica

Sistema de cache

- É utilizado um banco de dados NoSQL, baseado em chave-valor, amplamente utilizado para problemas com alta latência de dados, conhecido como Redis

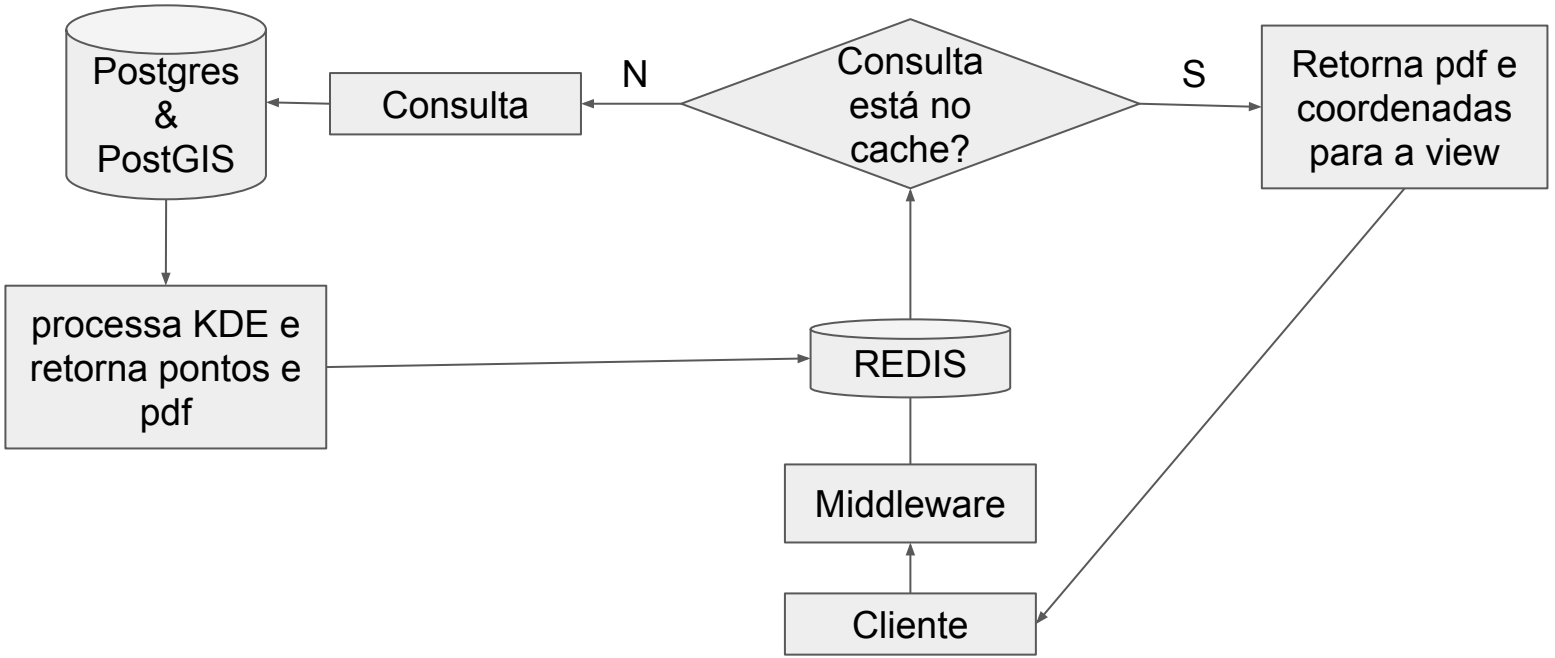
Consultas geográficas

- PostGIS e leaflet são usados em conjunto com o openstreemap

ERSI-RJ 2016

# W-SAGE

W-SAGE: **W**eb tool for **S**patial **A**nalysis of **GE**odata



ERSI-RJ 2016

# W-SAGE

Inferência estatística utilizando KDE multivariante $O(n^2k)$

- Dado n observações, calculamos curvas de densidade delas em relação à distância de um valor central, o núcleo, para cada um desses pontos e obtemos a estimativa de densidade final(PDF) somando-se esses valores
- Um kernel é uma função de ponderação padronizada, ou seja, o núcleo determina a suavização do PDF. Esta técnica é amplamente usada em vários algoritmos de aprendizado de máquina, principalmente em SVM (Support Vector Machines)

ERSI-RJ 2016

# KDE

| | | |
|---|---|---|
| Epanechnikov | $\frac{3}{4}(1-\frac{1}{5}t^2)/\sqrt{5}$ for $\lvert t\rvert < \sqrt{5}$, $0$ otherwise | $1$ |
| Biweight | $\frac{15}{16}(1-t^2)^2$ for $\lvert t\rvert < 1$, $0$ otherwise | $(\frac{3087}{3125})^{1/2} \approx 0.9939$ |
| Triangular | $1-\lvert t\rvert$ for $\lvert t\rvert < 1$, $0$ otherwise | $(\frac{243}{250})^{1/2} \approx 0.9859$ |
| Gaussian | $\frac{1}{\sqrt{2\pi}}\, e^{-(1/2)t^2}$ | $(\frac{36\pi}{125})^{1/2} \approx 0.9512$ |
| Rectangular | $\frac{1}{2}$ for $\lvert t\rvert < 1$, $0$ otherwise | $(\frac{108}{125})^{1/2} \approx 0.9295$ |

Tipos de núcleos(kernels)

```
for i ← 0 to n do
    soma_kernel ← 0.0
    for j ← 0 to n do
        prod_kernel ← 1.0
        for k ← 0 to xLen do
            prod_kernel *K((x[i][k] − x[j][k])/h)/h
        end
        soma_kernel ← soma_kernel + prod_kernel
    end
    pdf[i] ← soma_kernel / n
end
```

Pseudocódigo do KDE

ERSI-RJ 2016

# Experimentos

Foram realizados dois estudos de caso:

- Localização das 570 agências do IBGE pelo país
- Base anonimizada referente à alguns dados de matrícula de 14027 alunos da UFRRJ entre 2000 e 2013

Foi montado o mapa de calor baseado no resultado do KDE tanto em matlab quanto pelo W-SAGE

Também foi levado em consideração o tempo de processamento do KDE

ERSI-RJ 2016

# Resultados

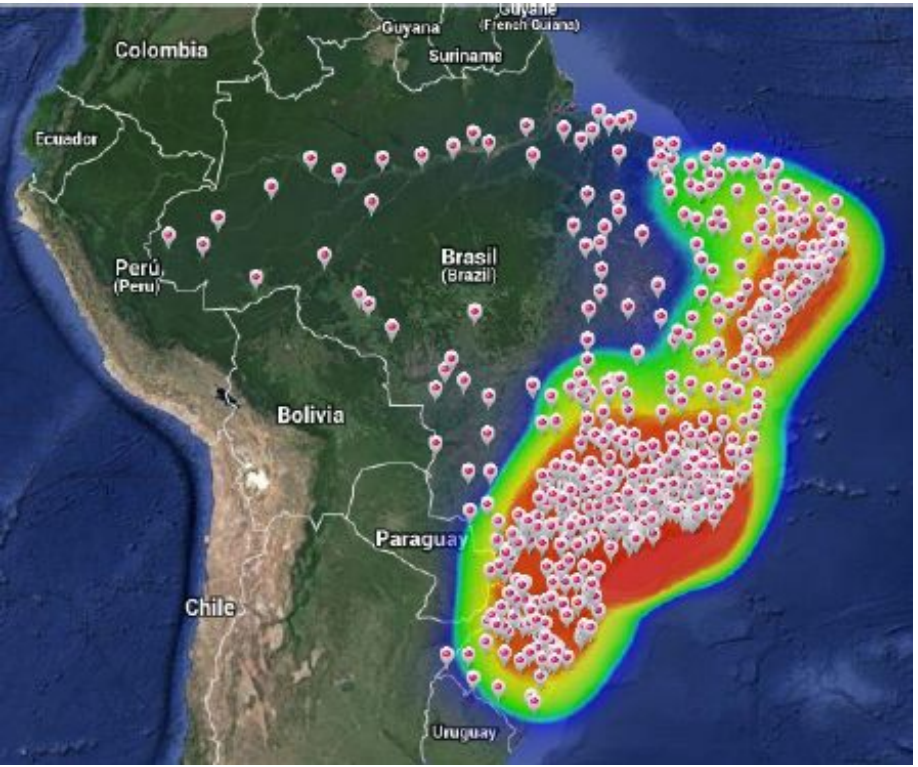
Técnica de paralelização de trabalho anterior para processamento do KDE mostrou-se promissor para ser usado em ambiente web com muitos registros

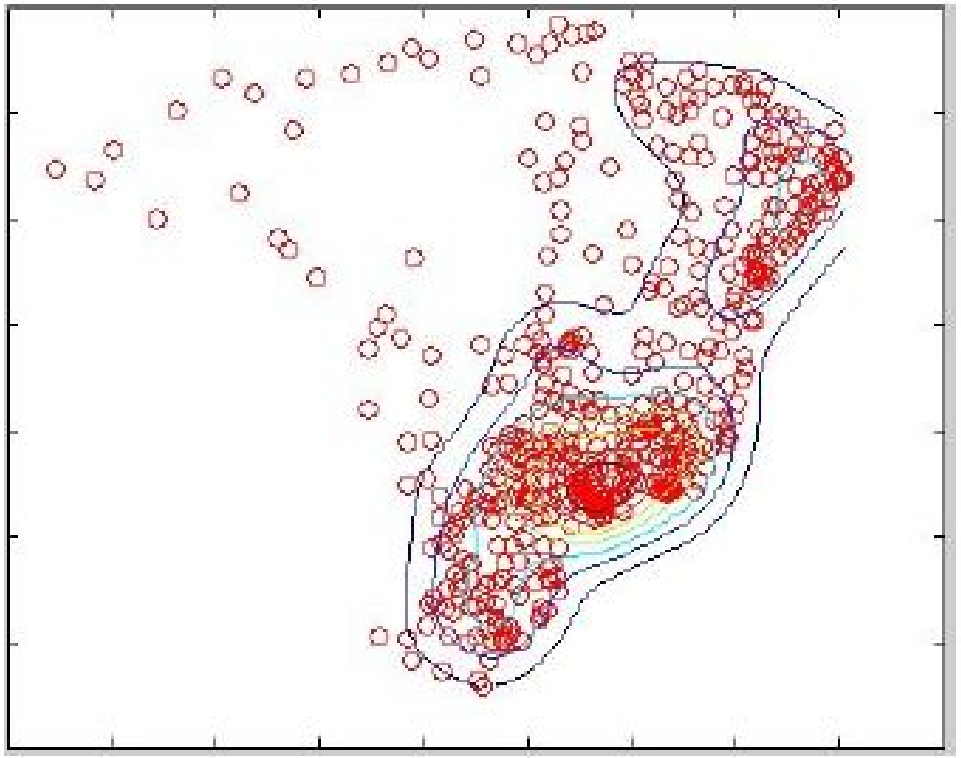
**Tabela 1.** Tempos medidos.

| KDE Sequencial | KDE Paralelizado Matlab | KDE c/ CUDA | KDE c/ CUDA otimizado | Speed-Up (GPU x Serial) | Speed-Up (GPU x Matlab) |
|---|---|---|---|---|---|
| 31.088s | 6.355s | 1.680s | 1.028s | 30,241 | 6,181 |

ERSI-RJ 2016

# Resultados



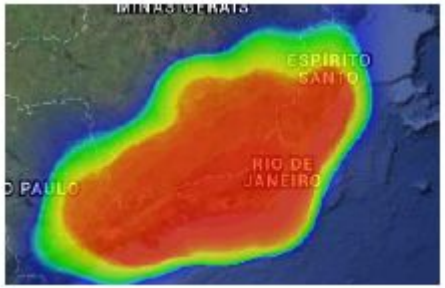



ERSI-RJ 2016

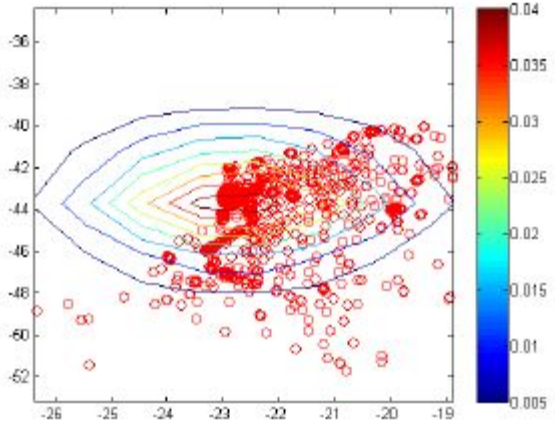
# Resultados

Origem dos alunos matriculados na UFRRJ

W-SAGE



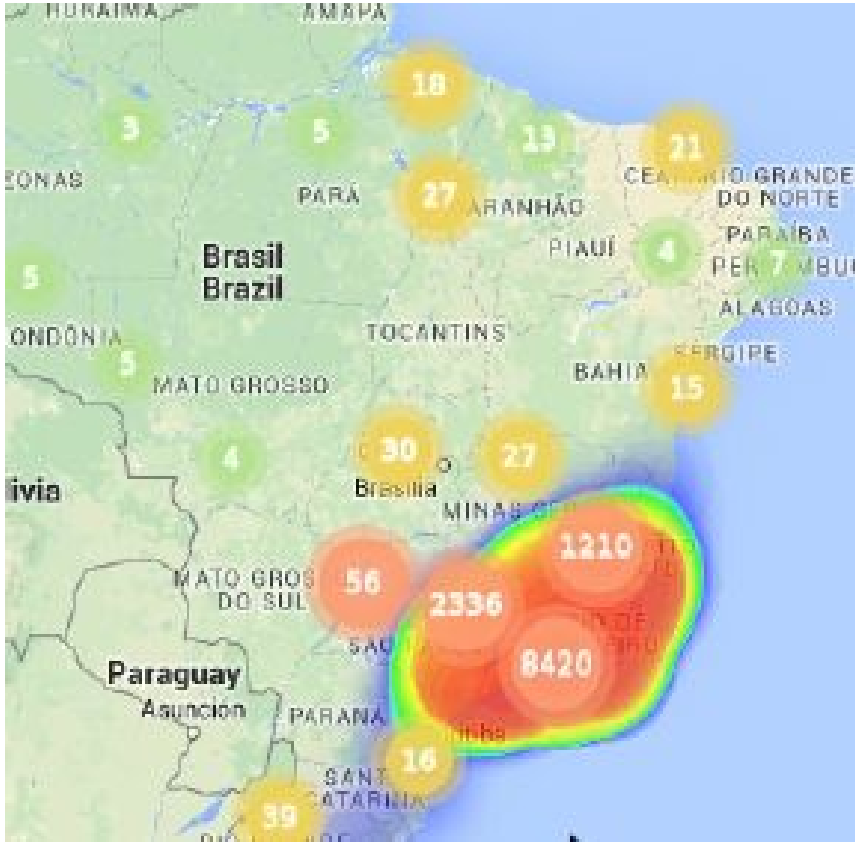

Matlab



ERSI-RJ 2016

# Resultados

W-SAGE mostrando pontos clusterizados e rasterizados juntamente com o mapa de calor baseado nos pesos dados pelos PDFs gerados pelo KDE



ERSI-RJ 2016

# Resultados

Ferramenta está sendo usada em um sistema de informação dentro do Instituto de Pesquisa Econômica e Aplicada (IPEA)

- Versão em uso no IPEA utilizará GPU mas o paralelismo por enquanto é feito utilizando a biblioteca joblib do python

ERSI-RJ 2016

# Resultados

Código utilizado para construir o W-SAGE

- https://github.com/raulsenaferreira/W-SAGE

Código utilizado para paralelizar o KDE

- https://github.com/raulsenaferreira/Computer-Science-UFRRJ/tree/master/TEPC/Paper-%20II%20RAIC

ERSI-RJ 2016

# Conclusão

Conseguimos integrar um método estimador de densidades, com resultados parecidos com o de softwares reconhecidos pelo mercado, à uma aplicação web

A ferramenta se mostra escalável ao paralelizar uma parte do algoritmo usando GPU e ao utilizar técnicas de clusterização de pontos e rasterização de imagens

Sistema de cache evita reprocessamento de consultas e do KDE

Ferramenta pode ser facilmente integrada à outros sistemas web

ERSI-RJ 2016

# Trabalhos futuros

Utilizar bancos de dados distribuídos e uma arquitetura assíncrona para receber muitas requisições (Node.js) e processar milhões de pontos

Substituição do KDE tradicional por um modelo mais rápido do algoritmo, proposto este ano na literatura onde a complexidade, que é quadrática, é reduzida para linear


# Contato

Raul Sena Ferreira

- raulsf@cos.ufrj.br
- www.raulferreira.com.br

Carlos E. R. de Mello

- mello@uniriotec.br
- http://www.uniriotec.br/~mello