CEFET/RJ
DIRETORIA DE PESQUISA E PÓS-GRADUAÇÃO

IDENTIFICAÇÃO DE ESTRANGULAMENTOS NA MOBILIDADE URBANA COM O USO DO GPS NO ÔNIBUS

Rafael Carreiro da Silva

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia de Produção e Sistemas, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, CEFET/RJ, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia de Produção e Sistemas.

Orientador

Diego Moreira de Araujo Carvalho, D.Sc.

Rio de Janeiro
Maio de 2016

IDENTIFICAÇÃO DE ESTRANGULAMENTOS NA MOBILIDADE URBANA COM O USO DO GPS NO ÔNIBUS

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia de Produção e Sistemas do Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia de Produção e Sistemas.

Rafael Carreiro da Silva

Aprovada por:

_____

Diego Moreira de Araujo Carvalho, D.Sc. (Orientador)

_____

Eduardo Soares Ogasawara, D.Sc.

_____

Felipe Maia Galvão França, Ph.D. (UFRJ)

Rio de Janeiro

Maio de 2016

Ficha catalográfica

Ficha catalográfica elaborada pela Biblioteca Central do CEFET/RJ

S586 Silva, Rafael Carreiro da
Identificação de estrangulamentos na mobilidade urbana com o uso do GPS no ônibus / Rafael Carreiro da Silva.—2016.
xv, 85f. : il. (algumas color.) , grafs. , tabs. ; enc.

Dissertação (Mestrado) Centro Federal de Educação Tecnológica Celso Suckow da Fonseca , 2016.
Bibliografia : f. 80-85
Orientador : Diego Moreira de Araujo Carvalho

1. Mobilidade residencial. 2. Sistemas inteligentes de veículos rodoviários. 3. Sistema de posicionamento global. 4. Planejamento urbano. I. Carvalho, Diego Moreira de Araujo (Orient.). II. Título.

CDD 304.8

Dedicatória

Dedico este trabalho a meus familiares que sempre me apoiaram e estiveram presentes durante os desafios enfrentados na minha vida, em especial meus pais que me incentivaram a seguir o caminho dos estudos e do gosto pelo conhecimento, minha esposa por ser minha companheira e cúmplice e meus filhos por serem minha fonte de inspiração.

Agradecimento

Gostaria de agradecer a todos que contribuíram para a elaboração deste trabalho, ao professor Diego Moreira de Araujo Carvalho que pacientemente me orientou ao longo desses dois anos, aos professores Eduardo Soares Ogasawara e Augusto da Cunha Reis que contribuíram com críticas construtivas durante minha qualificação e aos amigos Fabio Tadeu Cravo Carneiro e Tiago Seabra que sempre estiveram disponíveis a me ajudar com as dúvidas sobre linguagens de programação.

Epígrafe

“O maior bem que podemos fazer a um homem é levá-lo à Verdade. ”

São Tomás de Aquino

RESUMO

IDENTIFICAÇÃO DE ESTRANGULAMENTOS NA MOBILIDADE URBANA COM O USO DO GPS NO ÔNIBUS

Rafael Carreiro da Silva

Orientador: Diego Moreira de Araujo Carvalho, D.Sc.

Resumo da Dissertação de Mestrado submetida ao Programa de Pós-graduação em Engenharia de Produção e Sistemas do Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, CEFET/RJ, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia de Produção e Sistemas.

Com a mudança da matriz demográfica mundial e o crescimento acentuado da população urbana em todo mundo, as grandes cidades estão expostas a novos problemas de administração pública, principalmente desde a segunda metade do século XX. A questão da mobilidade urbana é um desses problemas que demanda por soluções inovadoras e que busquem a otimização dos sistemas de transporte público por meio da utilização de modernas tecnologias que despontam nas últimas décadas e que trazem ganhos de eficiência operacional. Este trabalho aponta soluções para o problema da identificação de gargalos no trânsito de grandes cidades que necessitem transportar uma parte considerável de sua população entre subúrbios e centros comerciais, utilizando meios de transporte rápidos, econômicos e confortáveis. É objetivo desta dissertação propor um método de análise de dados georreferenciados utilizando dispositivos GPS embarcados em ônibus, técnicas de mineração de dados e ciência da computação para identificar estrangulamentos no tráfego de veículos em ambiente urbano, contribuindo para um melhor entendimento da dinâmica do trânsito na cidade e para proposição de medidas que atenuem ou eliminem os transtornos causados pelos engarrafamentos.

Palavras-Chave:

Mobilidade urbana; Mineração de dados; Sistemas inteligentes de transporte

Rio de Janeiro

Maio de 2016

ABSTRACT

IDENTIFICATION OF BOTTLENECKS IN URBAN MOBILITY USING GPS ON BUS

Rafael Carreiro da Silva

Advisor: Diego Moreira de Araujo Carvalho, D.Sc.

Abstract of dissertation submitted to Programa de Pós-graduação em Engenharia de Produção e Sistemas - Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ as partial fulfillment of the requirements for the degree of Master.

The world demographic matrix has been changing and one realises a sharp increase of the urban population in worldwide. Because of that, large cities are exposed to new problems of public administration, especially during the second half of the twentieth century. Urban mobility is one of these problems that demand innovative solutions that seek to optimize public transport systems using modern technologies that have emerged in recent decades and bring operational efficiency gain. This work points out solutions to the problem of bottlenecks identification in road traffic of large cities that require carry a considerable part of its population between suburbs and downtown, using quick, cheap and comfortable transportation. This dissertation aims to propose a geospatial data analysis method, using GPS devices equipped on buses, data mining techniques and computer sciences to identify congestions in transit of urban environments, contributing to a better understanding of traffic dynamics in the city and to propose actions to mitigate or eliminate the inconvenience caused by traffic jams.

Keywords:

Urban mobility; Data mining; Intelligent transportation system.

Rio de Janeiro

Maio de 2016

SUMÁRIO

Capítulo I - Introdução 1
I.2 Justificativa para escolha do tema 2
I.3 Objetivo 4
I.4 Definição do escopo do trabalho 5
I.4.1 Etapa de Coleta e Tratamento dos Dados 5
I.4.2 Etapa de Mineração de dados e Análise dos Resultados 5
I.4.3 Estrutura do Trabalho 6
Capítulo II - Trabalhos Relacionados 7
II.1 Pesquisa Bibliográfica 7
II.2 Sistemas Inteligentes de Transporte 9
II.2.2 VANET 10
II.2.3 WSN 11
II.3 Cidades Inteligentes 11
II.4 Temas principais 12
II.4.1 Identificação de Padrões 13
II.4.2 Planejamento Urbano 13
II.4.3 Surveys 14
II.4.4 Roteirização e Predição de Trajetórias 14
II.4.5 Tempo de Viagem 15
II.4.6 Pontos de Interesse 16
II.4.7 Previsão de Fluxo 17
II.4.8 Estimativa de Velocidade 17
II.5 Geoposicionamento (Sistema de Coordenadas Geográficas) 17
II.6 Sistemas Globais de Navegação por Satélites 18
II.6.1 Funcionamento do GPS 19
II.6.2 Trilateração 20
II.6.3 Fórmula de *Haversine* 21
II.7 Sistemas de Informação Geográfica 22

II.8 Algoritmo *DBSCAN* 22
II.8.1 Vantagens e Desvantagens do DBSCAN 22
II.8.2 Parâmetros de Estimativa do DBSCAN (ε *e MinPts)* 23
II.8.3 Ponto Central, Ponto Periférico e Ruído 23
II.8.4 DBSCAN vs KMeans 24
II.9 Algoritmo *Grid Growing* 25
II.9.1 Parâmetros de Estimativa do *Grid Growing* 26
II.9.2 Vantagens e Desvantagens do *Grid Growing* 27
Capítulo III - Método Proposto 29
III.1 Ferramentas para análise dos dados 29
III.1.1 Bancos de Dados 29
III.1.2 Python 30
III.1.3 Google Earth 32
III.1.4 Google Fusion Tables 33
III.2 Conjunto de Dados 34
III.2.1 Estrutura dos Dados 34
III.3 Processo 37
III.4 Definição do período de análise 39
III.5 Etapa de pré-processamento 39
III.6 Garagens dos ônibus 40
III.7 Análise de Sensibilidade 41
Capítulo IV - Avaliação Experimental 48
IV.1 Caso 1: Estreia do Brasil na Copa do Mundo 48
IV.1.1 Caso 1: DBSCAN 48
IV.1.2 Caso 1: *Grid Growing* 53
IV.2 Caso 2: Greve do Rodoviários do Rio de Janeiro 58
IV.2.1 Caso 2: DBSCAN 59
IV.2.2 Caso 2: *Grid Growing* 63
IV.3 Caso 3: Acidente no Viaduto dos Marinheiros 67
IV.3.1 Caso 3: DBSCAN 68

IV.3.2 Caso 3: *Grid Growing* 71

Capítulo V - Conclusões 77

Referências Bibliográficas 80

LISTA DE FIGURAS

Figura I.1 População mundial urbana e rural – 1950 a 2050 1
Figura I.2 Viagens realizadas por modo principal (2012) 3
Figura I.3 Eixos de transporte e densidade populacional na RMRJ 4
Figura II.1 Quantidade de publicações nos últimos 10 anos 8
Figura II.2 Crescimento das cidades em todo o mundo 9
Figura II.3 Temas principais dos artigos pesquisados 12
Figura II.4 Sistema de Coordenadas – Latitude e Longitude 18
Figura II.5 Funcionamento do GPS 19
Figura II.6 Rota dos satélites 20
Figura II.7 Trilateração 20
Figura II.8 Deformações da crosta terrestre 21
Figura II.9 Ponto Central, Periférico e Ruído – Exemplo 1 24
Figura II.10 Ponto Central, Periférico e Ruído – Exemplo 2 24
Figura II.11 DBSCAN vs Variações KMeans 25
Figura II.12 Cálculo da malha de cada ponto 26
Figura II.13 Agrupamento utilizando método 4-vizinhos e 8-vizinhos 27
Figura II.14 Representação Cartesiana dos Vizinhos 27
Figura III.1 Representação Espacial do *Dataset* GPS ônibus – Linha 455 37
Figura III.2 Processo de identificação de estrangulamentos 38
Figura III.3 Matriz para Análise de Sensibilidade 42
Figura III.4 Agrupamento *Grid Growing* com Parâmetros 50x20 43
Figura III.5 Agrupamento *Grid Growing* com Parâmetros 160x140 43
Figura III.6 Análise de Sensibilidade por Quadrante 44
Figura III.7 Análise de Sensibilidade – Cálculo de Δ_C 46
Figura III.8 Análise de Sensibilidade – Cálculo de Δ_R 47
Figura IV.1 Análise de sensibilidade DBSCAN – 05/06/2014 (quinta-feira) 49
Figura IV.2 Análise de sensibilidade DBSCAN – 12/06/2014 (quinta-feira) 49
Figura IV.3 Cálculo de Δ_C para DBSCAN (05/06/14 e 12/06/14) 50

Figura IV.4 Cálculo de Δ_R para DBSCAN (05/06/14 e 12/06/14) 51

Figura IV.5 DBSCAN *MinPts* = 90 e ε = 0.004 – 12/06/14 52

Figura IV.6 DBSCAN *MinPts* = 90 e ε = 0.004 – 05/06/14 53

Figura IV.7 Análise de sensibilidade GG – 05/06/2014 (quinta-feira) 54

Figura IV.8 Análise de sensibilidade GG – 12/06/2014 (quinta-feira) 54

Figura IV.9 Cálculo de Δ_C para GG (05/06/14 e 12/06/14)..... 55

Figura IV.10 Cálculo de Δ_R para GG (05/06/14 e 12/06/14)..... 56

Figura IV.11 GG *MinPts* = 60 e Malha = 140 – 12/06/14 56

Figura IV.12 GG *MinPts* = 60 e *Malha* = 140 – 05/06/14 57

Figura IV.13 GG *MinPts* = 60 e *Malha* = 140 – 05/06/14 (Zona Sul)..... 57

Figura IV.14 Análise de sensibilidade DBSCAN – 13/05/2014 (terça-feira) 59

Figura IV.15 Análise de sensibilidade DBSCAN – 20/05/2014 (terça-feira) 60

Figura IV.16 Cálculo de Δ_C para DBSCAN (13/05/14 e 20/05/14) 60

Figura IV.17 Cálculo de Δ_R para DBSCAN (13/05/14 e 20/05/14) 61

Figura IV.18 DBSCAN *MinPts* = 60 e ε = 0.003 – 13/05/14 61

Figura IV.19 DBSCAN *MinPts* = 60 e ε = 0.003 – 20/05/14 62

Figura IV.20 Análise de sensibilidade GG – 13/05/2014 (terça-feira) 63

Figura IV.21 Análise de sensibilidade GG – 20/05/2014 (terça-feira) 63

Figura IV.22 Cálculo de Δ_C para GG (13/05/14 e 20/05/14)..... 64

Figura IV.23 Cálculo de Δ_R para GG (13/05/14 e 20/05/14)..... 64

Figura IV.24 GG *MinPts* = 60 e *Malha* = 120 – 13/05/14 65

Figura IV.25 GG *MinPts* = 60 e *Malha* = 120 – 20/05/14 65

Figura IV.26 GG *MinPts* = 60 e *Malha* = 120 – 20/05/14 (Região Jacarepaguá) 66

Figura IV.27 Análise de sensibilidade DBSCAN – 16/09/2014 (terça-feira) 68

Figura IV.28 Análise de sensibilidade DBSCAN – 23/09/2014 (terça-feira) 68

Figura IV.29 Cálculo de Δ_C para DBSCAN (16/09/14 e 23/09/14) 69

Figura IV.30 Cálculo de Δ_R para DBSCAN (16/09/14 e 23/09/14) 69

Figura IV.31 DBSCAN *MinPts* = 170 e ε = 0.003 – 16/09/14..... 70

Figura IV.32 DBSCAN *MinPts* = 170 e ε = 0.003 – 23/09/14 70

Figura IV.33 Análise de sensibilidade GG – 16/09/2014 (terça-feira) 71

Figura IV.34 Análise de sensibilidade GG – 23/09/2014 (terça-feira) 72

Figura IV.35 Cálculo de Δ_C para GG (16/09/14 e 23/09/14) 72

Figura IV.36 Cálculo de Δ_R para GG (16/09/14 e 23/09/14) 73

Figura IV.37 GG *MinPts* = 170 e *Malha* = 110 – 16/09/14 73

Figura IV.38 GG *MinPts* = 170 e *Malha* = 110 – 23/09/14 74

Figura IV.39 GG *MinPts* = 170 e *Malha* = 110 – 16/09/14 (Região Centro-Tijuca) 74

Figura IV.40 GG *MinPts* = 170 e *Malha* = 110 – 23/09/14 (Região Centro-Tijuca) 75

Figura IV.41 GG *MinPts* = 110 e *Malha* = 150 – 16/09/14 (Região Centro) 75

Figura IV.42 GG *MinPts* = 110 e *Malha* = 150 – 23/09/14 (Região Centro) 76

LISTA DE TABELAS

Tabela III.1 Comparação entre *Google Earth Pro* e *Google Earth*..... 33

Tabela III.2 Descrição do *Dataset* – Conjunto GPS ônibus 35

Tabela III.3 Exemplo de registros – Conjunto GPS ônibus..... 36

Capítulo I - Introdução

Desde a Revolução Industrial no século XVIII observamos uma mudança na interação social e econômica que acarretou num fluxo migratório de pessoas de zonas rurais para centros urbanos. Esse contingente humano contribuiu para que as cidades crescessem em economia, oportunidades, mas também em problemas, uma vez que essa grande aglomeração de pessoas em espaços antes limitados acarretou no surgimento de novos problemas nas áreas de mobilidade urbana, segurança pública, saúde, educação, limpeza, conservação e muitos outras. Estima-se que 54% da população mundial viva em grandes centros urbanos e este número pode aumentar para 66% até 2050 como mostra a Figura I.1 (UNITED NATIONS; DEPARTMENT OF ECONOMIC AND SOCIAL AFFAIRS; POPULATION DIVISION, 2014). Os governos têm como grande desafio prover serviços de qualidade para milhões de pessoas, e certamente um dos meios mais eficientes para solucionarem este problema de escala é através da utilização de tecnologias que aumentem a produtividade e permitam uma melhor utilização dos recursos públicos (NEIROTTI et al., 2014).

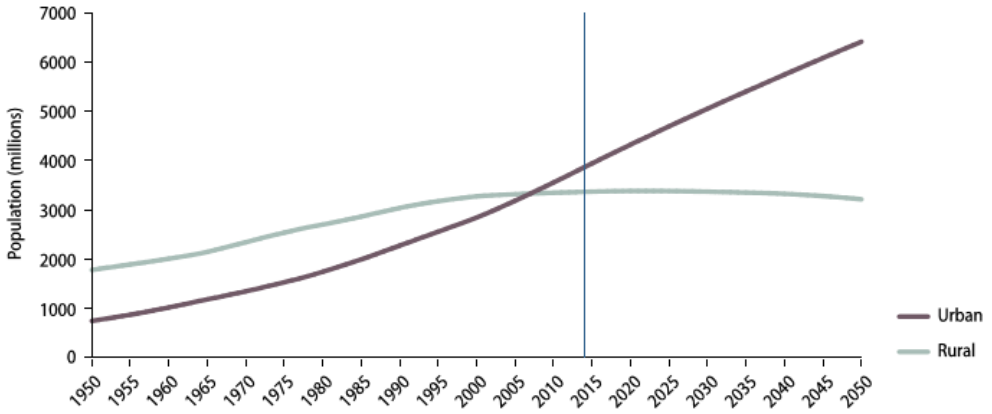


Figura I.1 População mundial urbana e rural – 1950 a 2050
Fonte: United Nations, Department of Economic and Social Affairs, Population Division 2014

BORGIA (2014) chama atenção para os avanços nos campos da microeletrônica, da telefonia celular, dos sistemas de comunicação sem fio e dos serviços ubíquos (que proveem conectividade ao usuário em qualquer lugar a qualquer momento), que se espalharam rapidamente na última década. Desde o final dos anos 1980 pesquisadores tem se voltado para adaptar a tecnologia às atividades cotidianas, passando de uma era dominada pelos computadores pessoais, para uma era onde qualquer dispositivo pode ter uma função na rede mundial de computadores (GUBBI et al. 2013). Sob este pano de fundo podemos considerar as cidades como organismos vivos (SALIM e HAQUE 2015) e não há soluções padronizadas para todos os problemas, cabendo uma análise que demande por soluções criativas e personalizadas. A mobilidade urbana é um tema importante dentro deste contexto, pois a

qualidade e eficiência dos meios de transporte públicos são pontos chave para um alto nível de satisfação da população em relação à administração pública.

Este trabalho tem como grande motivador contribuir para um melhor conhecimento dos problemas relacionados à mobilidade urbana nas metrópoles e apresentar alternativas criativas para solução desses entraves. Serão analisados dados do modal rodoviário de transporte público que nos permitem monitorar o trânsito com objetivo de identificarmos congestionamentos nas principais vias de acesso, contribuindo para uma melhor prestação de um dos serviços públicos mais importantes, a mobilidade urbana.

Para o experimento retratado neste trabalho foi escolhida a cidade do Rio de Janeiro como objeto de estudo por reunir as características necessárias à condução da pesquisa (elevada concentração de habitantes[1] e dependência do modal rodoviário na composição matriz de transporte[2]) e pela facilidade na obtenção dos dados. Serão utilizados os dados coletados em tempo real por dispositivos com tecnologia *Global Position System* (GPS) embarcados nos ônibus que atendem a Região Metropolitana do Rio de Janeiro (RMRJ) e disponibilizados pela Prefeitura Municipal do Rio de Janeiro (PMRJ). Estes dados serão tratados com objetivo de serem convertidos em informações úteis, por meio da aplicação de técnicas de Mineração de Dados, e auxiliarem na tomada de decisão por parte das autoridades da administração pública.

I.2 Justificativa para escolha do tema

A justificativa para escolha do tema deve-se ao contínuo aumento demográfico característicos dos grandes centros urbanos e por consequência, o aumento da complexidade do problema do transporte público, que deve ser capaz não só de atender uma quantidade maior de pessoas, mas também prestar um serviço de qualidade em uma malha com grande capilaridade. Outro ponto relevante é a limitação da capacidade de investimento dos Municípios, o que obriga as Prefeituras a priorizarem novos projetos que tragam maiores impactos em termos de benefício para a população.

A análise da Figura I.2 indica uma grande utilização do modal ônibus por parte da população da RMRJ totalizando mais de 8 milhões de viagens em 2012. Estendendo a análise para os demais modais de transporte, concluímos que apesar dos altos investimentos feitos nos últimos anos, a quantidade de pessoas atendidas por trens e metrô está bem abaixo da quantidade atendida pelos ônibus. A representatividade do transporte por ônibus é de 37,7%, e se desconsiderarmos o transporte a pé este valor sobe para 53,4%.

[1] 6.320.446 habitantes segundo censo IBGE 2010
[2] Ver Tabela 1



Figura I.2 Viagens realizadas por modo principal (2012)

Fonte: Plano Diretor de Transporte da Região Metropolitana do Rio de Janeiro – ago/14

Existem razões históricas que justificam uma maior capilaridade do transporte rodoviário em comparação aos demais modais na cidade do Rio de Janeiro, porém não é foco deste trabalho discutir tais fatos. Uma vez que os ônibus são responsáveis pelo deslocamento da maior parte da população, é razoável considerar que a melhorias no sistema rodoviário de transporte causem impactos positivos para os clientes deste modal. A Figura I.3 apresenta o eixo dos meios de transporte e a densidade populacional da RMRJ. O eixo das rodovias principais é mais extenso do que o de qualquer outro modal, além de ser a única opção para grande parte dos municípios e de cruzar as áreas com maior densidade populacional.

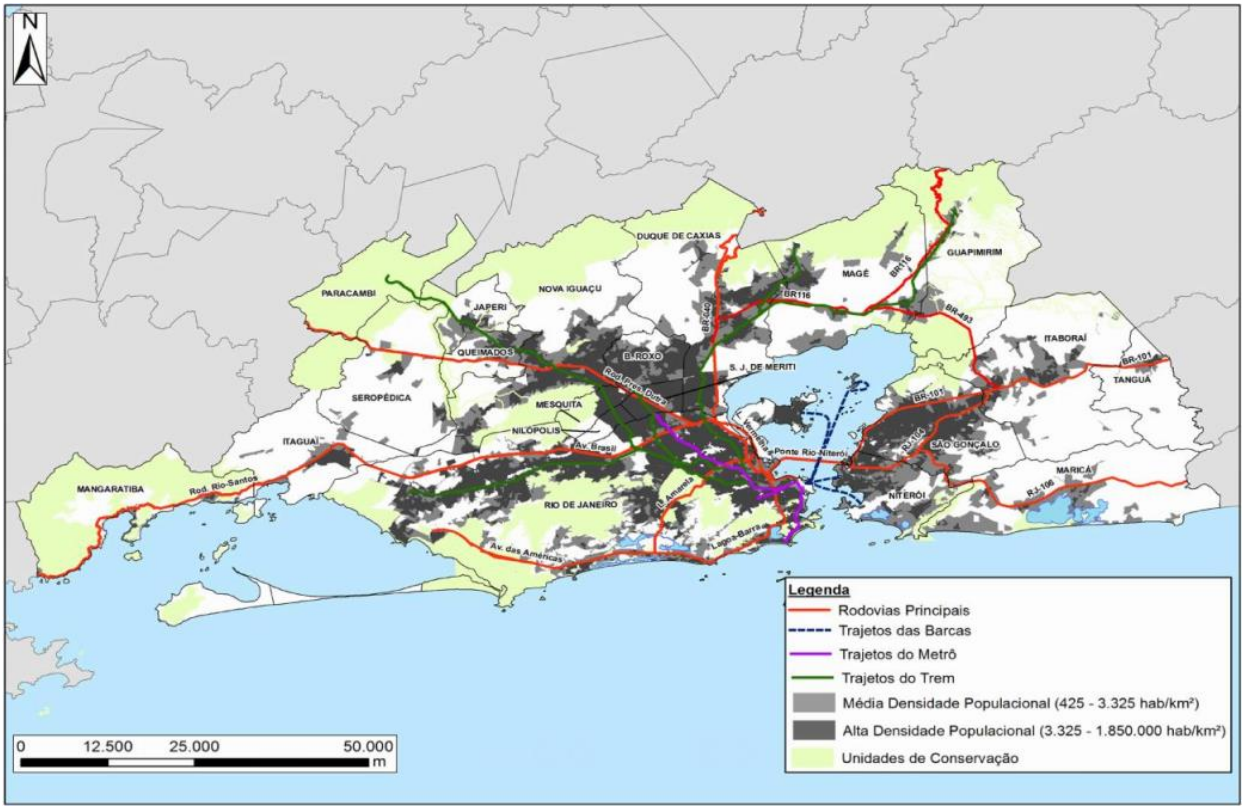


Figura I.3 Eixos de transporte e densidade populacional na RMRJ

Fonte: Plano Diretor de Transporte da Região Metropolitana do Rio de Janeiro – ago/14

Com base nos fatos acima, este trabalho tem como principal justificativa contribuir para uma melhor compreensão da dinâmica do transporte rodoviário, identificando regiões com maior densidade de veículos e como o sistema se adapta as constantes flutuações no fluxo de veículos ao longo do tempo. Em trabalhos futuros tais informações poderão ser utilizadas com intuito de prover um melhor atendimento aos usuários do transporte rodoviário, reduzindo tempo de espera e viagem.

I.3 Objetivo

O objetivo principal da dissertação é propor um método de análise de dados coletados nos ônibus da RMRJ para identificar estrangulamentos no tráfego de veículos, utilizando técnicas de mineração aplicadas a estes dados, sendo possível também a replicação em situações semelhantes identificadas em outras cidades. A partir dos resultados obtidos pelo algoritmo desenvolvido deverá ser possível extrair um diagnóstico objetivo dos eventos escolhidos para a experiência, servindo de base para uma validação do grau de aderência do método em relação a realidade observada.

I.4 Definição do escopo do trabalho

O escopo do trabalho se divide em duas fases: a fase de obtenção e tratamento dos dados e a de mineração de dados e análise dos resultados obtidos, gerando informações úteis para auxílio à tomada de decisão.

I.4.1 Etapa de Coleta e Tratamento dos Dados

Os dados utilizados na pesquisa são disponibilizados pela Prefeitura Municipal do Rio de Janeiro em seu portal de dados abertos[3] e são coletados pelos computadores do LAB-MOB (Laboratório de Mobilidade Urbana do CEFET-RJ), e devido suas características necessitam de um tratamento especial antes de sua aplicação nos modelos que serão propostos ao longo da dissertação. Podemos considerar que um determinado conjunto de dados possui qualidade se ele satisfaz os requisitos da utilização pretendida, principalmente quando avaliado à luz de fatores como exatidão, integridade, consistência, oportunidade, credibilidade e facilidade de interpretação (HAN et al., 2011). Em situações reais é muito comum que os dados utilizados apresentem algum grau de incompletude, imprecisão, ruído ou inconsistência causando confusão durante o procedimento de mineração e podendo resultar em saídas não confiáveis, portanto, se faz necessária a utilização de rotinas de limpeza de dados de modo a garantir sua qualidade (HAN et al., 2011).

I.4.2 Etapa de Mineração de dados e Análise dos Resultados

Na etapa seguinte o objetivo é dar significado aos dados coletados por meio de técnicas de mineração de dados. Devido à característica do *dataset* optamos pela utilização de dois algoritmos de agrupamento, um por densidade e outro baseado em grid, detalhados no Capítulo III - . Agrupamento refere-se ao processo de particionamento de um conjunto de objetos em subconjuntos que respeitem a similaridade dos objetos (HOPCROFT e KANNAN, 2011), no caso deste trabalho a similaridade almejada é a localização geográfica em um determinado período. Outro fator relevante para escolha destas técnicas é a farta quantidade de publicações com problemas de mobilidade urbana similares ao apresentado no presente trabalho. Ao final prevemos a análise dos resultados obtidos e as considerações sobre os casos analisados. Conforme estabelecido nos objetivos do trabalho, o problema a ser respondido é aonde ocorrem os congestionamentos no transporte rodoviário na área da RMRJ, e de quais locais carecem de uma melhor infraestrutura ou estratégia de transporte.

[3] data.rio

I.4.3 Estrutura do Trabalho

Este trabalho está dividido em cinco capítulos.

1. Capítulo I: São apresentadas as motivações, justificativas e objetivos do trabalho;
2. Capítulo II: São apresentados os critérios que basearam a pesquisa bibliográfica e os principais conceitos relacionados ao assunto do trabalho. Também é proposta de análise dos principais temas com objetivo de fazer um levantamento sobre o estado da arte no que se refere a como a temática da mobilidade urbana vem sendo tratada pela comunidade científica e com quais ferramentas e técnicas;
3. Capítulo III: Dedicado a apresentação do método desenvolvido para identificação de gargalos no transporte rodoviário, capítulo essencial para entendimento do *modus operandi* do experimento, com explicações detalhadas sobre cada uma das etapas do processo. Será apresentado método inovador para determinação dos parâmetros dos algoritmos de agrupamento;
4. Capítulo IV: Neste capítulo apresentaremos os resultados dos experimentos propostos e as análises que confirmarão o alcance do objetivo definido pela pesquisa. Aplicaremos o método exposto no capítulo III em três estudos de casos.
5. Capítulo V: Destinado às conclusões finais e sugestões de trabalhos futuros que possam gerar novas publicações dentro da mesma linha de pesquisa.

Capítulo II - Trabalhos Relacionados

O objetivo deste capítulo é discutir como o tema mobilidade urbana vem sendo abordado pela comunidade acadêmica, quais são as principais linhas de pesquisa em discussão até o momento, quais as ferramentas mais utilizadas nos trabalhos pesquisados e como o assunto vem atraindo o interesse dos pesquisadores nos últimos anos. Também serão abordados neste capítulo alguns conceitos chaves para contextualizarmos a relevância do tema para a sociedade, e que contribuirão para um entendimento mais amplo das questões relacionadas às mudanças tecnológicas ocorridas nas últimas décadas e suas aplicabilidades.

II.1 Pesquisa Bibliográfica

Como principal fonte de pesquisa de literatura existente sobre o tema e trabalhos correlatos, foi utilizada a base de dados da *Science Direct* a qual se mostrou bastante eficaz para obtenção de material de excelente qualidade e adequado ao propósito. A relação bibliográfica definitiva foi concluída em fevereiro de 2016 tendo como principais parâmetros de busca trabalhos publicados nos últimos 10 anos e palavras-chaves como "*public transportation", "intelligent transportation system", "urban traffic", "traffic jam", "mobility", "traffic monitoring", "gps"* e *"geo-spatial data".*

O quantitativo de artigos por ano encontrado na pesquisa está representado no gráfico de barras da Figura II.1. Nesta figura, podemos observar que existe uma elevada taxa de crescimento das publicações relacionadas ao assunto nos últimos 10 anos, o que indica um aumento no interesse da comunidade científica pelo tema. Importante destacar que os números referentes a 2016 estão incompletos uma vez que a consulta foi realizada em fev/2016 contemplando as publicações previstas para mar/2016. Se considerarmos a mesma quantidade de publicações dos 3 primeiros meses de 2016 para os demais trimestres, alcançaremos um valor total de 36 publicações, um aumento em torno de 57% em relação a 2015.

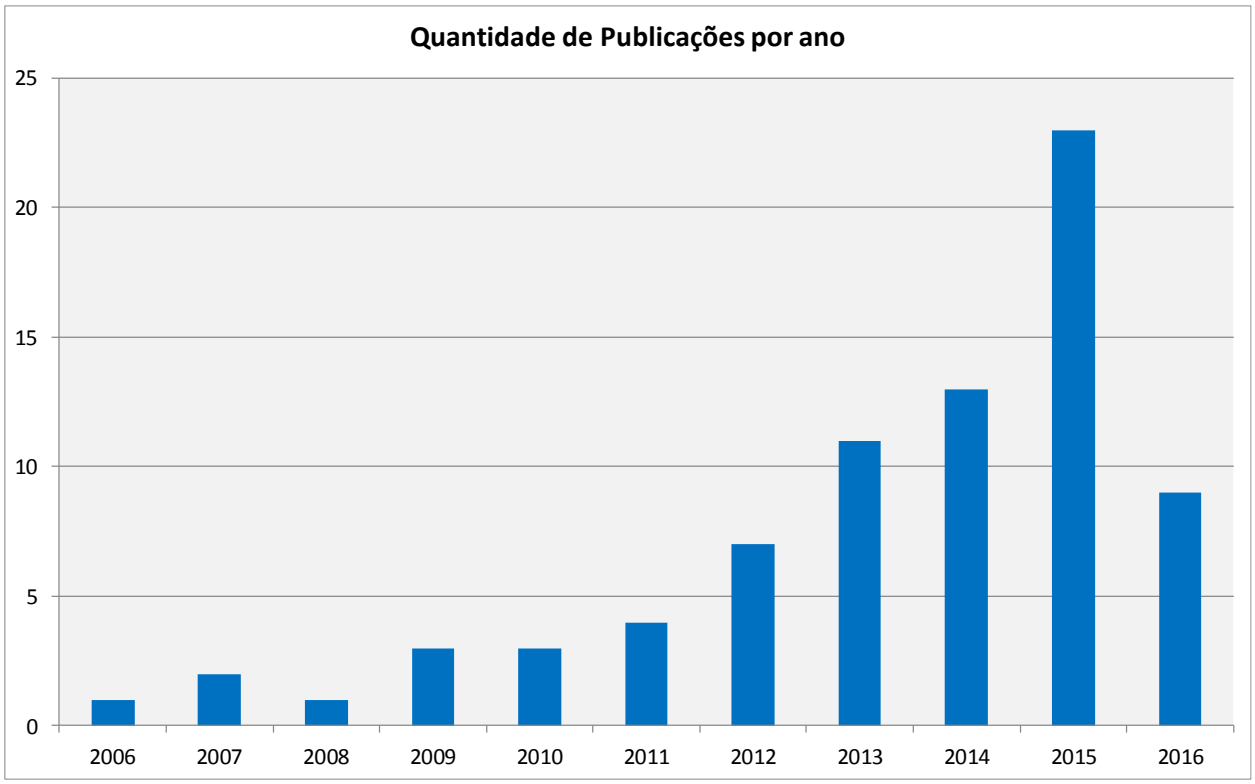


Figura II.1 Quantidade de publicações nos últimos 10 anos

Se compararmos este crescimento previsto para 2016 com as taxas dos últimos anos, observamos que os valores se encontram em linha com a tendência atual. Importante ressaltar que neste quadro apenas constam publicações indexadas a *Science Direct*, que apesar de ser uma excelente referência para trabalhos acadêmicos não esgota a quantidade de publicações sobre uso de tecnologias convergentes em problemas de mobilidade urbana. Em trabalhos futuros poderão ser avaliados o comportamento das taxas de crescimento de publicações em outras bases de dados de primeira linha, tais como, *Scopus ou Web of Science.*

Feita essa primeira avaliação quantitativa do material bibliográfico, se faz necessário nos debruçarmos mais atentamente sobre as questões qualitativas e sobre as especificidades abordadas nos diversos trabalhos pesquisados, a fim de respondermos os seguintes questionamentos: Por que cada vez mais o uso de novas tecnologias em questões de mobilidade urbana vem atraindo pesquisadores de diferentes partes do mundo? Esse interesse permanecerá nos próximos anos? Quais são os problemas atraem maior atenção e como eles são tratados?

II.2 Sistemas Inteligentes de Transporte

Uma resposta possível para o questionamento levantado anteriormente, passa pela mudança no comportamento demográfico observado em praticamente todo mundo. Desde meados do século XX, cada vez mais se intensifica o fluxo migratório de pessoas que deixam zonas rurais a procura de oportunidades e condições que propiciem uma melhor qualidade de vida (UNITED NATIONS; DEPARTMENT OF ECONOMIC AND SOCIAL AFFAIRS; POPULATION DIVISION, 2014). Um estudo do Departamento de Assuntos Econômicos e Sociais das Nações Unidas aponta que de 1950 até 2014 o percentual da população mundial urbana saltou de 30% para 56%, com expectativa de alcançar 66% em 2050. Esse aumento não é algo regionalizado ou restrito a países desenvolvidos, vem ocorrendo em praticamente todo o mundo.

Como consequência desse fenômeno, observamos o surgimento e crescimento de cidades que deverão suportar as necessidades dessas pessoas com serviços públicos e privados de qualidade, tais como saúde, educação, segurança pública e mobilidade urbana. É neste contexto que enxergamos a importância das tecnologias emergentes na solução dos problemas oriundos dessa transformação social que estamos vivendo. O mesmo estudo faz ainda uma análise do crescimento populacional das cidades ao longo dos últimos 25 anos e faz uma projeção para 2030, conforme Figura II.2.

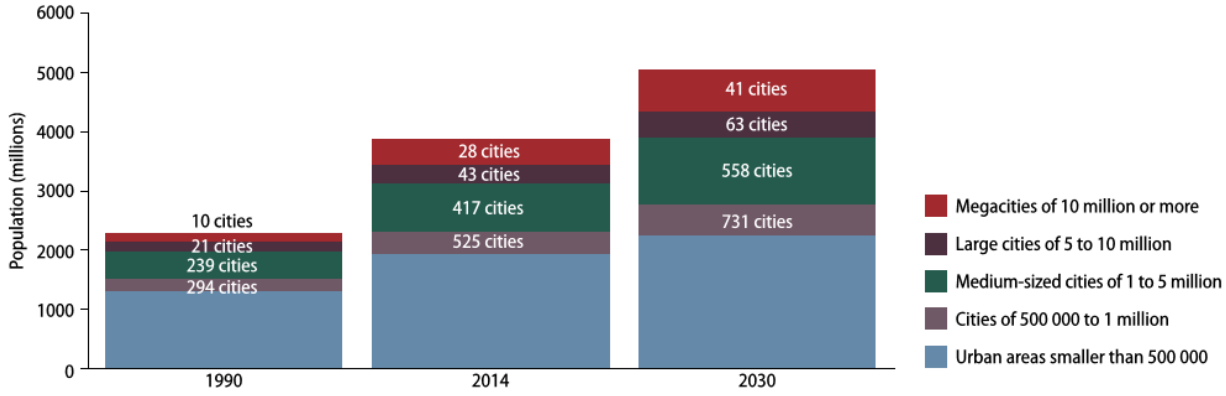


Figura II.2 Crescimento das cidades em todo o mundo
Fonte: United Nations, Department of Economic and Social Affairs, Population Division 2014

Como pudemos constatar, houve um aumento considerável na quantidade de cidades de todas as faixas de tamanho, entre 1990 e 2014 as megalópoles (maiores que 10 milhões de habitantes) quase que triplicaram e as cidades consideradas grandes (entre 5 e 10 milhões de habitantes) mais que dobraram.

Um cenário como este propiciou o desenvolvimento de um novo conceito que tivesse como característica principal o uso das diversas tecnologias disponíveis com objetivo de otimizar a utilização do transporte nas cidades. A este conceito foi dado nome de Sistemas Inteligentes de Transporte (ITS – *Intelligent Transportation Systems*) que pode ser definido como a aplicação e integração de tecnologias que permitem aos operadores e usuários um melhor gerenciamento e otimização dos sistemas de transporte, pelo uso de informações coletadas a partir vias de acesso, semáforos, veículos ou quaisquer outros entes que componham a rede de transporte em questão (PAN et al., 2014). Deste modo, os Sistemas Inteligentes de Transportam passam a ter uma maior importância junto aos órgãos das administrações públicas que necessitam gerenciar redes cada vez mais complexas com maior nível de eficiência, e que muito provavelmente explica o crescimento do interesse pelo assunto.

Dentro ainda do mundo dos ITS, abordaremos mais dois conceitos importantes que complementarão o entendimento que será explorado nos capítulos posteriores.

II.2.2 VANET

Elemento importante para a compreensão dos ITS, as Redes Veiculares *Ad Hoc* (VANET – *Vehicular Ad Hoc Networks)* são redes sem uma topologia previamente definida compostas por nós que se movem livremente, sendo a topologia modificada em tempo real e afetada principalmente por quatro aspectos (FONSECA e VAZÃO, 2013):

1. Cenário: são construídas sobre um mapa rígido, composto por vias de acesso, semáforos, cruzamentos, túneis, construções e outros obstáculos que limitam a livre movimentação dos nós e impactam na comunicação devido a interferência de sinal;
2. Padrão de mobilidade: os padrões de mobilidade não são aleatórios, mas seguem padrões limitados pelo cenário e, portanto, suscetíveis a conectividade irregular;
3. Propriedades da mobilidade: os nós movem-se rapidamente, podendo em um determinado momento estarem conectados e alguns segundos depois estarem situados fora da área de influência;
4. Propriedades dos nós: os nós são equipados com dispositivos que farão a comunicação entre os veículos (V2V) ou entre veículos e infraestrutura (V2I), passando informações coletados dos próprios nós, nos casos de telemetria, por exemplo, ou informações do meio, tais como, temperatura, humidade ou posicionamento. Além disso, não há preocupação com consumo de energia devido os veículos serem autônomos.

II.2.3 WSN

As Redes de Sensores Sem Fio (WSN – *Wireless Sensor Networks*) representam o principal conjunto de tecnologias utilizado para comunicação entre os veículos e a infraestrutura com objetivo de se alcançar uma gestão inteligente do transporte urbano. Segundo KAFI et al. (2012) WSN são tecnologias emergentes com potencial de superar dificuldades relacionadas a performance, custo, suporte e manutenção adicionando valor aos Sistemas Inteligentes de Transporte.

O avanço tecnológico das últimas décadas permitiu que sensores fossem desenvolvidos em tamanhos cada vez menores, tivessem sua autonomia de energia aumentada e preços reduzidos, viabilizando uma gama cada vez maior de aplicações. Outro ponto crucial é o desenvolvimento de novos protocolos de comunicação específicos para utilização neste tipo de rede que demandam baixas taxas de transmissão, geralmente a pequenas distâncias e que consumam a menor quantidade de energia possível. Além disso, há de se levar em conta medidas de redundância que considerem possíveis interferências causadas pelo ambiente, sejam elas problemas de sinal, danos aos equipamentos ou falhas na alimentação.

II.3 Cidades Inteligentes

Conforme visto anteriormente, as cidades já são responsáveis por abrigar a maior parte da população mundial, e as projeções apontam para continuidade desta tendência. Já dissertamos sobre a importância dos ITS para uma melhor gestão pública das cidades, porém este é apenas uma das preocupações acerca de um tema mais abrangente, o das Cidades Inteligentes (SC – *Smart Cities*).

O conceito de Cidades Inteligentes transpassa diversas áreas de conhecimento e com diversas interpretações na literatura atual, segundo STEENBRUGGEN et al. (2015) uma cidade pode ser considerada “inteligente” quando investimentos em capital social e humano, transporte, moderna infraestrutura de Tecnologia da Informação e Comunicação (TIC) alavancam um crescimento econômico sustentável e elevados índices de qualidade de vida por meio de um gerenciamento sensato dos recursos naturais e uma governança participativa. Já NEIROTTI et al. (2014) chamam a atenção para a questão do planejamento e controle de toda a cidade e da função central das TIC servindo como um sistema nervoso digital que obtém dados a partir das fontes mais heterogêneas (esgoto, estacionamentos, câmeras de segurança, semáforos, termostatos, etc.) e com objetivo de ganhos de produtividade por meio da automação de rotinas e fortalecendo o processo de tomada de decisão dos gestores públicos.

Podemos dividir a literatura de SC em domínios, cujo mais relevante para a iniciativa deste trabalho é Transporte, Mobilidade e Logística, com os seguintes subdomínios (NEIROTTI et al., 2014):

1. Logística da cidade: foca em melhorar os fluxos na cidade integrando efetivamente necessidades de negócio com as condições de tráfego, geografia e questões ambientais;
2. Informação de mobilidade: trata da distribuição e utilização de informações multimodais dinâmicas obtidas antes e durante a viagem, com objetivo de melhorar a eficiência do tráfego e transporte e garantir elevado grau de satisfação os usuários;
3. Mobilidade das pessoas: objetiva prover soluções de transporte inovadores e sustentáveis aos cidadãos, de tal forma que o desenvolvimento dos modais e veículos se baseie em combustíveis renováveis e limpos, suportados por novas tecnologias e um comportamento proativo da comunidade.

II.4 Temas principais

Por fim, conforme descrito na primeira seção deste capítulo, precisamos levantar quais são os problemas que demandam as maiores quantidades de publicação, seguindo as premissas e critérios estabelecidos anteriormente, e ainda, qual é o ferramental proposto pelos autores para responder a estes questionamentos. Após uma leitura minuciosa dos trabalhos levantados (Figura II.1), podemos subdividir os principais temas propostos conforme Figura II.3 a seguir:

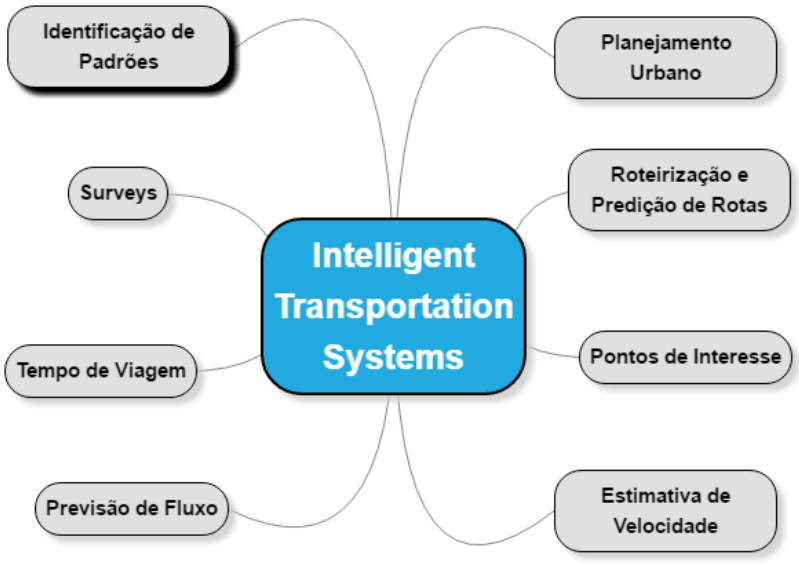


Figura II.3 Temas principais dos artigos pesquisados

II.4.1 Identificação de Padrões

A identificação de padrões é sem dúvida um dos principais temas mais retratados por artigos acadêmicos ao longo dos últimos anos. CALABRESE et al. (2013) utilizam dados coletados a partir de telefones móveis para obter padrões de mobilidade dos indivíduos em áreas metropolitanas, ANAGNOSTOPOULOS et al. (2015) propõem um sistema para otimização da validação de grupos de objetos móveis em Serviços Baseados em o Localização. Alguns trabalhos analisam os padrões de mobilidade em situações extremas causadas por fenômenos da natureza, como por exemplo, XU et al. (2013) quantificam o impacto de tempestades na operação de tráfego na malha rodoviária urbana e HARA e KUWAHARA (2015) analisam o padrão de evacuação da cidade de Ishinomaki (Japão) após a ocorrência de um terremoto com objetivo de criar planos de contingência que otimizem o escoamento de pessoas após desastres naturais. A identificação nos padrões de congestionamentos e suas implicações dão abordadas por KURZHANSKIY e VARAIYA (2015), YONG-CHUAN et al. (2011) e SUN et al. (2011). ASTARITA et al. (2014) apresentam um estudo para desenvolvimento de um sistema capaz de utilizar o GPS de dispositivos móveis para mapear o estilo de direção e estatísticas de direção dos motoristas, servindo de alerta para os operadores de trânsito sobre locais com maior ocorrência de incidentes e podendo recompensar os motoristas com estilo de direção mais seguros.

II.4.2 Planejamento Urbano

Muitos artigos selecionados tratam do tema do planejamento urbano auxiliado por tecnologias baseadas em Sistemas Inteligentes de Transporte. HAO et al. (2015) abordam a rápida expansão na utilização de *big data* em estudos de planejamento urbano na China a partir dos anos 2000, CARREL et al. (2015) introduzem um sistema para extração das rotas de viagens a partir do cruzamento de dados dos celulares dos participantes com os dados de localização dos veículos. VON FERBER et al. (2009) procuram identificar as propriedades distintivas das redes de transporte público de quatorze cidades quando interpretadas como redes de grafos complexas. ZHONG et al. (2014) propõem um método para inferir a utilização social de locais utilizando mineração de dados e PARK et al. (2014) também apresentam um método para mineração de dados, mas que utilizam *tablets* e telefones celular no desenvolvimento de uma interface integrada entre os diferentes bancos de dados de mobilidade da cidade de Nova Iorque.

LIU et al., 2015 fornecem, em seu estudo, uma visão sobre o uso dos dados no reconhecimento de padrões de viagem nas estruturas das cidades, que possam auxiliar no desenvolvimento e aplicação das políticas públicas de transporte, TOOLE et al. (2015), em seu trabalho, fazem uma proposta de implementação de um modelo de demanda baseado em um sistema que combina dados de chamada de telefonia celular, registros de censos,

levantamento de campo e dados geoespaciais para desenvolver um software flexível, modular, computacionalmente eficiente e que contemple o processo fim-a-fim, ou seja, desde a coleta do dado bruto até a transformação em estimativas de demanda.

ZHOU et al. (2016) apresentam um sistema de predição de passageiros para usuários de *smartphones* composto por um servidor que processa e analisa o fluxo de dados e um *app* que monitora continuamente o número de passageiros em cada estação. CUI et al. (2016) em seu trabalho desenham um processo para identificação de regiões com problema de mobilidade, utilizando um conjunto de indicadores que medem a performance do transporte rodoviário a partir de dados coletados de GPS embarcados em taxis da cidade de Harbin (China).

II.4.3 Surveys

Como seria de se esperar, o maior interesse pela produção de trabalhos sobre Sistemas Inteligentes de Transporte, também acarretou em aumento no número de publicações de artigos que tem como finalidade servirem de *surveys* e *reviews* de temas derivados. QURESHI e ABDULLAH (2013) se preocupam em citar uma grande variedade de aplicações e tecnologias nas diversas áreas das ITS, BALI et al. (2014) discutem desafios das VANET, propõe soluções e apresentam uma taxonomia de diferentes técnicas de agrupamento de redes veiculares ad hoc, NEIROTTI et al. (2014) elaboram uma proposta de taxonomia para Cidades Inteligentes com ramificação na área de transporte, logística e mobilidade, STEENBRUGGEN et al. (2015) também abordam o tema das Cidades Inteligentes sobre a ótica da melhora no desenvolvimento urbano utilizando aplicações inovadoras que utilizem dados georreferenciados e BOUKERCHE et al. (2008) pesquisam diversas técnicas utilizadas para estimativa de posicionamento de veículos elencando as vantagens e desvantagens de cada uma quando aplicadas a VANET.

Questões de sinalização são tratadas por FLORIN e OLARIU (2015) e KAFI et al. (2012), de protocolos de roteamento de veículos por BILAL et al. (2013) , FONSECA e VAZÃO (2013) e SHAREF et al. (2014). RAO et al. (2012) fazem uma revisão literária sobre diferentes métodos de representação, análise, mineração e visualização de dados espaço temporais, e WHAIDUZZAMAN et al. (2014) apresentam uma extensiva revisão literária sobre o estado da arte em *Vehicular Cloud Computing*, uma nova tecnologia de computação nas nuvens que disponibiliza serviços de informação a um baixo custo para os motoristas.

II.4.4 Roteirização e Predição de Trajetórias

Uma das questões mais citadas em trabalhos acadêmicos acerca de mobilidade urbana é a escolha do “melhor trajeto” para a viagem, ou roteirização. Muitas variáveis podem ser

consideradas na definição de “melhor trajeto”, tais como, distância, tempo, custo, restrições de vias não asfaltadas, faixas exclusivas para transporte público etc. ROUHIEH e ALECSANDRU (2012) propõem um sistema dinâmico de escolha de rotas para linhas de ônibus de Quebec utilizando teoria Markoviana, ZHOU e HIRASAWA (2014) utilizam técnicas de programação de redes genéticas *(GNP – Genetic Network Programming)* combinadas com um simulador de tráfego em tempo real em seu experimento. MEI et al. (2015) avaliam a utilização de três diferentes métodos de agrupamento na classificação das trajetórias de veículos em corredores de tráfego misto: *Constrained-K-Means (CKM), Seeded-K-Means (SKM) e Semi-Supervised Fuzzy c-Means (SFCM).* CHEN et al. (2011) apresentam um sistema de predição de rotas que roda em celulares com funcionalidade GPS, utiliza arquitetura cliente/servidor economizando processamento nas pontas e que utiliza o algoritmo inovador *Continuous Route Pattern Mining (CRPM).*

LIU e KARIMI (2006) abordam dois modelos de predição de trajetória denominadas *Probability-based Model e Learning-based Model* (adotam algoritmos de *machine learning*) com objetivo de fornecer informações georreferenciadas como parâmetros essenciais para sistemas que necessitem de tal informação para prestação do serviço *(location-aware computing)*. JABBARPOUR et al. (2015) destacam a utilização de algoritmos de otimização da colônia de formigas *(ACO – Ant Colony Optization)* em sistemas de roteamento de tráfego “verdes” *(Green VTRS – Vehicle Traffic Routing Systems)* com intuito de reduzir o consumo de combustível e a emissão de CO2 para atmosfera. ŞTEFĂNESCU et al. (2014) analisam diversos planejadores de viagem *(trip planners), softwares* que orientam os passageiros com relação às informações sobre transporte público (horário de partida, rotas, preços, distâncias, pontos de interesse, conexões com outros meios de transporte etc.), discutem a importância de tais programas para os operadores e usuários de transporte público e apresentam um planejador de viagem desenvolvido para cidade de Timisoara na Romênia.

II.4.5 Tempo de Viagem

Outro problema clássico de mobilidade urbana abordado frequentemente é o cálculo do tempo de viagem dado os pontos de origem e destino. FENG et al. (2014) utilizam distribuições de probabilidade para calcular tempo de viagem e apresentam dois estudos de caso, um em Mineápolis e outro em Atlanta. MAZLOUMI et al. (2009) atentam para questão da confiabilidade na qualidade do serviço prestado analisando dados coletados de Melbourne e também utilizam distribuição de probabilidade para compreensão do fenômeno da variabilidade do tempo de viagem (*TTV – Time Travel Variability*). JIMÉNEZ-MEZA et al. (2013) também utilizam conceitos de nível de serviço para segmentar as vias públicas e propõem um *framework* com tal finalidade utilizando, como exemplo, dados coletados de taxis da cidade de Beijing.

Muitos autores utilizam técnicas de mineração de dados para cálculo do tempo de trajeto, HAGE et al. (2012) desenvolvem uma metodologia de estimativa em tempo real utilizando como ferramenta de predição uma variante do filtro de Kalman*, UKF (Unscented Kalman Filter).* Filtro de Kalman também é utilizado por CHEN et al. (2012) para ajustar os valores obtidos a partir de um modelo baseado em *SVM (Support Vector Machine)* e testado com dados coletados do *BRT (Bus Rapid Transit)* de Shangai. XINGHAO et al. (2013) propõem um modelo de predição dos tempos de viagem utilizando dados de ônibus de Shangai, mas também sugere a utilização de sistemas *RFID (Radio Frequency Identification)* para reduzir o impacto das interferências que afetam sistemas baseados apenas em GPS. Por fim o modelo de predição de XIN e CHEN (2016) utiliza técnicas de baseadas em KNN *(K-Nearest Neighbour)* com a finalidade de determinar o tempo de permanência dos ônibus nas estações e utilizando como base dados coletados de Changzhou na China.

II.4.6 Pontos de Interesse

Questões acerca de levantamento de pontos de interesse (*PoI – Points of Interest*) tem chamado atenção, principalmente nos últimos cinco anos, e podem ter seu número de publicações aumentado nos próximos anos. CAO et al. (2010) apresentam uma técnica para extração de semântica das localizações a partir de dados de GPS e gerando grafos que representem as relações entre locais e entre locais e usuários. PAN et al. (2013) discutem a classificação de uso do espaço urbano (*land-use classification*) utilizando rotas de taxi da cidade de *Hangzhou* na China e técnicas de mineração de dados (*DBSCAN*). ZHAO et al. (2015) também utilizam dados coletados de GPS em taxi e técnicas clássicas de mineração de dados (*K-means, Spectral Clustering, DBSCAN* dentre outros) para comprovar a eficácia e eficiência de seu método proposto, *Grid Growing Clustering* (discutiremos este método com mais detalhes no próximo capítulo).

JOSSE et al. (2015) conciliam dois objetivos, minimizar o tempo de viagem e guiar usuários pelos pontos de interesse mais populares, para tal utiliza técnica de caminho ótimo de Pareto, e MOREIRA-MATIAS et al. (2016) também utilizam dados de taxis portugueses para criar uma Matrix O-D (origem e destino) e utilizando técnicas de mineração em fluxos de dados contínuos encontrar regiões de interesse. KERAMAT JAHROMI et al., 2016 propõem um modelo que simule a movimentação da população pelos PoI com objetivo de descrever o comportamento social dos indivíduos nos espaços urbanos e extrair as informações de como os recursos e serviços das cidades são utilizados e POUKE et al. (2016) apresentam um método computacional que exploram os pontos de interesse para gerar uma simulação de fluxo de multidão utilizando a rede *wi-fi* gratuita da cidade de Oulu (Finlândia).

II.4.7 Previsão de Fluxo

A previsão do tráfego de curto prazo tem sido parte fundamental para ITS e muitas áreas de pesquisa sobre transporte desde o início dos anos 1980 (VLAHOGIANNI et al., 2014) e muitos autores têm utilizado múltiplos métodos combinados para melhorar o resultado das estimativas (MANNINI et al., 2015). KONG et al. (2015) abordam o uso de sensores móveis em veículos (*floating cars*) para investigar grandes quantidades de tráfego urbano em tempo real, e utiliza no algoritmo de predição método do enxame de partículas (PSO – *Particle Swarm Otimization*) e SVM. Já MANNINI et al. (2015) exploram o trânsito de Roma e utilizam como ferramenta de predição modelos de fluxo de tráfego macroscópico (*Macroscopic Traffic Flow Model*) corrigido por um filtro de Kalman extendido (*Extended Kalman Filter*). XIA et al. (2016) propõem um modelo espaço-temporal ponderado KNN (STW-KNN) comparando seus resultados com outros modelos de previsão, tais como, *KNN* Convencionais, Redes Neurais Artificiais, *Random Forest*, *Naive Bayes* e C4.5, e por fim VLAHOGIANNI et al. (2014) atualizam sua revisão literária proposta em 2004, com as principais publicações sobre previsão de tráfego nos últimos dez anos.

II.4.8 Estimativa de Velocidade

A formulação de modelos que façam predições de velocidade do trânsito ainda é um tópico pouco explorado, apenas dois trabalhos levantados se debruçam sobre este assunto. MA et al. (2015) abordam a utilização de uma variante de rede neurais (*LSTM NN*) na predição da velocidade de tráfego, com um estudo de caso em Beijing e comparando resultado com outras estruturas de redes neurais, SVM, ARIMA (*Autoregressive Integrated Moving Average*) e *Kalman Filter*. BACHMANN et al. (2013) comparam diversas técnicas de fusão de dados multissensoriais como estimadores de velocidade de tráfego utilizando dados coletados a partir de dispositivos *bluetooth* em uma estrada de Toronto, Canada. Por fusão de dados multissensorial entenda-se a combinação de dados coletados a partir de múltiplas fontes referentes a um mesmo fenômeno tornando sua representação mais consistente, precisa e útil (KHALEGHI et al., 2013).

II.5 Geoposicionamento (Sistema de Coordenadas Geográficas)

Os *datasets* utilizados durante a avaliação experimental possuem uma característica espaço-temporal, ou seja, as informações relevantes para as análises dependem basicamente de dados espaciais (coordenadas geográficas) em um determinado tempo. O posicionamento geográfico é apresentado por meio de coordenadas geográficas, latitude e longitude, que definem o posicionamento de um determinado ponto aferido na projeção do globo terrestre. Latitude de um ponto é o ângulo entre o plano do equador e uma reta que passe pelo centro da

terra e pelo próprio ponto, com isso todas as medições de latitude são paralelas à linha do equador e os valores possíveis variam entre 90º Norte e 90º Sul. Já a longitude de ponto é medida pelo ângulo entre o plano do meridiano de referência (que passa pelo observatório de Greenwich, Inglaterra) e o meridiano do ponto em questão, variando entre 180º Leste e 180º Oeste. A Figura II.4 apresenta uma visualização do sistema de coordenadas geográficas.

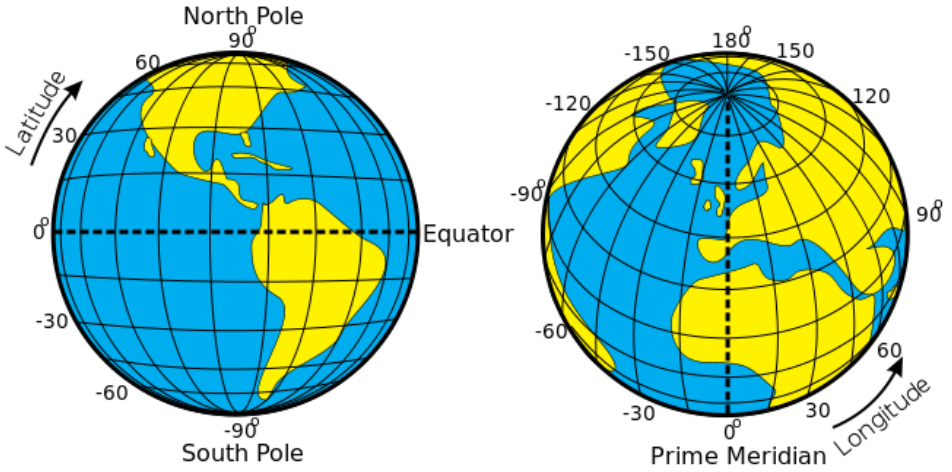


Figura II.4 Sistema de Coordenadas – Latitude e Longitude

Fonte: https://commons.wikimedia.org/wiki/File:Latitude_and_Longitude_of_the_Earth.svg#filelinks

II.6 Sistemas Globais de Navegação por Satélites

Por meio da utilização de Sistemas Globais de Navegação por Satélite *(GNSS - Global Navigation Satellite Systems)* é possível obtermos de maneira rápida e precisa a identificação da latitude e longitude de um ponto. O primeiro sistema do tipo foi desenvolvido pelo Departamento de Defesa dos Estados Unidos, com nome de projeto Navstar/GPS[4]. Primeiramente esse sistema tinha como finalidade apenas utilização militar, mas em 1983 foi aberto também para utilização civil com alguma limitação de precisão (em torno de 100m), porém em 2000 essa limitação foi suspensa permitindo uma precisão de ±15m nos dispositivos comerciais mais simples.

Outros projetos GNSS também estão sendo desenvolvidos por outros países, por exemplo, GLONASS[5] (Rússia), BeiDou[6] (China) e Galileo[7] (União Europeia). Alguns dispositivos mais recentes já se aproveitam não só das informações enviadas pela constelação Navstar/GPS, mas também dos demais sistemas de navegação por satélite, garantindo maior disponibilidade de serviço e precisão na utilização. É comum por uma questão de hábito generalizar por GPS os diferentes sistemas de navegação por satélite, apesar de termos a

[4] http://www.gps.gov/
[5] https://www.glonass-iac.ru/en/
[6] http://en.beidou.gov.cn/
[7] http://www.gsa.europa.eu/galileo/why-galileo

consciência de que grande parte das aplicações existente faz uso conjunto dos GNSS para uma melhor precisão no posicionamento. Outro modo de se aprimorar a precisão é por meio da utilização de dispositivos compatíveis com A-GPS (*Assisted* GPS) que permitem o recebimento de dados das operadoras de telefonia móvel, via suas redes GPRS, 3G e 4G, o que pode melhorar acurácia da informação em 5 a 10 metros, tornar a sincronia mais rápida, melhorar a performance em locais fechados e reduzir o consumo de energia (SINGHAL e SHUKLA, 2012).

II.6.1 Funcionamento do GPS

O GPS é composto por receptores móveis de pequenas dimensões situados nos pontos a serem identificados, e que recebem informações dos satélites que orbitam a Terra constantemente (de 24 a 32 simultaneamente já considerando sobressalentes). Os satélites estão situados a uma altitude de aproximadamente 20 km da Terra e percorrem suas rotas a uma velocidade estimada de 14.000 km/h, emitindo sinais na velocidade da luz com informações sobre seu posicionamento e hora com excelente grau de confiabilidade uma vez que todos eles são equipados com relógios atômicos sincronizados entre si. Com base nestas informações os receptores conseguem calcular as coordenadas de seus pontos por meio do processo de trilateração (detalhado na próxima seção). Para a correta definição de um ponto no globo são necessárias informações de ao menos quatro satélites, a Figura II.5 ilustra o funcionamento do GPS.



Figura II.5 Funcionamento do GPS
Fonte: https://commons.wikimedia.org/wiki/File:Good_gdop.png

A Figura II.6 apresenta um exemplo de trajetória de um dos satélites da constelação Navstar/GPS (trajetória verde) e de um satélite da constelação GLONASS (trajetória vermelha).

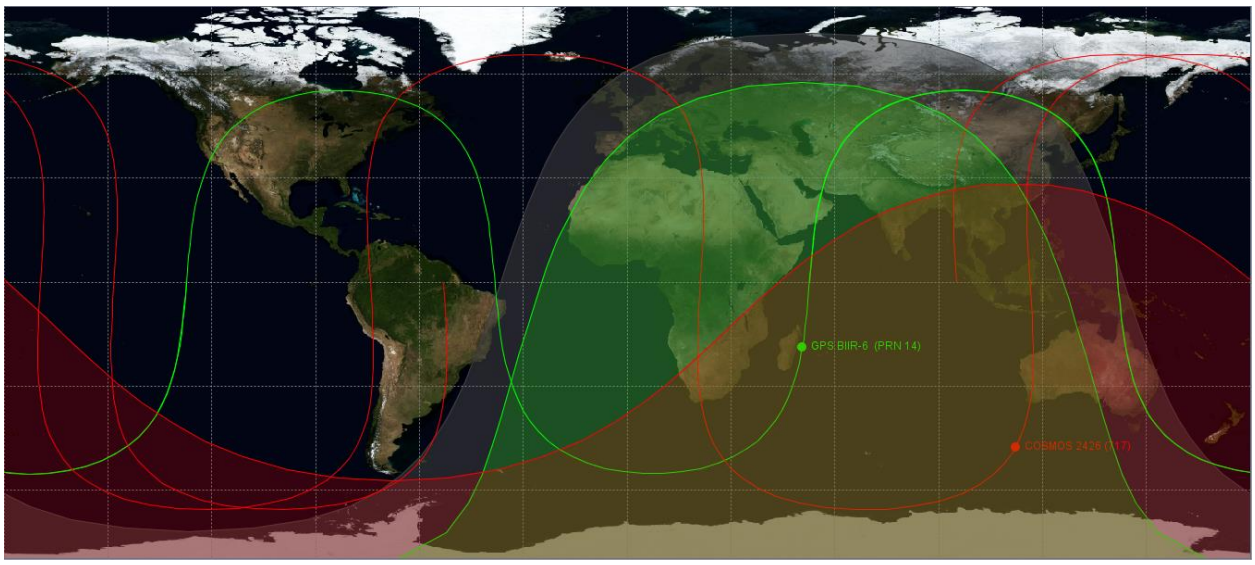


Figura II.6 Rota dos satélites
Fonte: Aplicativo JsatTrak

II.6.2 Trilateração

O cálculo para determinação de um ponto utilizando a informação de quatro satélites não é complexo, o método usado é o da trilateração. Conhecida as distâncias de pelo menos três satélites para um ponto na Terra, calculando a interseção entre estas distâncias define-se a projeção aproximada em que o ponto observado se encontra (NICULESCU e NATH, 2003). O quarto satélite é utilizado na definição da altitude e para refinamento da posição inicialmente determinada pelos demais satélites. A Figura II.7 ilustra a o processo de trilateração.

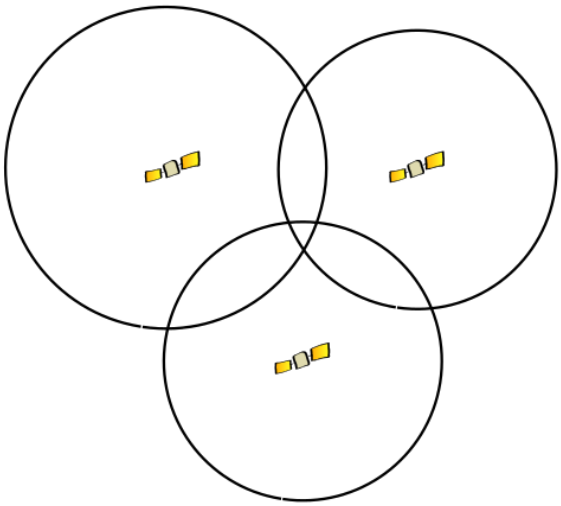


Figura II.7 Trilateração

Fonte: https://commons.wikimedia.org/wiki/File:Trilateration-with-3-satellites.svg

II.6.3 Fórmula de *Haversine*

O globo terrestre não é uma esfera perfeita e não possui uma superfície regular, devido às deformações encontradas nas diferentes placas tectônicas (Figura II.8), se aproximando de um formato elipsoide. Este fato torna muito complexo qualquer modelagem matemática que busque utilizar o sistema de coordenadas para cálculo exato de distâncias em projeções do globo. Neste trabalho utilizaremos a fórmula de *Haversine* para cálculo das distâncias entre dois pontos, este método é amplamente conhecido e utilizado pela comunidade acadêmica que promove trabalhos relacionados à geoposicionamento e pode ser aplicado sem prejuízos aparente devido as distâncias calculadas durante o experimento serem pequenas em relação ao tamanho da Terra (ALVES, 1986). Na aplicação da fórmula de *Haversine* que serão realizadas nas seções seguintes utilizaremos como parâmetros o raio da Terra aproximado de 6371 km, que é o raio de uma esfera com a mesma área de superfície que o elipsoide terrestre e está de acordo com a norma cartográfica WGS84 - *World Geodetic System* definida em 1984 (ŠEDŠNKA e GASTI, 2014).

$$D = R * \cos^{-1}\big(\cos\big(\pi * \big(90 - latB\big)/180\big) * \cos\big((90 - latA) * \pi/180\big) + \sin\big((90 - latB) * \pi/180\big) * \sin\big((90 - latA) * \pi/180\big) * \cos\big((longA - longB) * \pi/180\big)\big)$$

$D = Distância\ entre\ pontos$

$R = Raio\ da\ Terra\ (6.371km)$

$lat\ e\ long\ em\ graus$

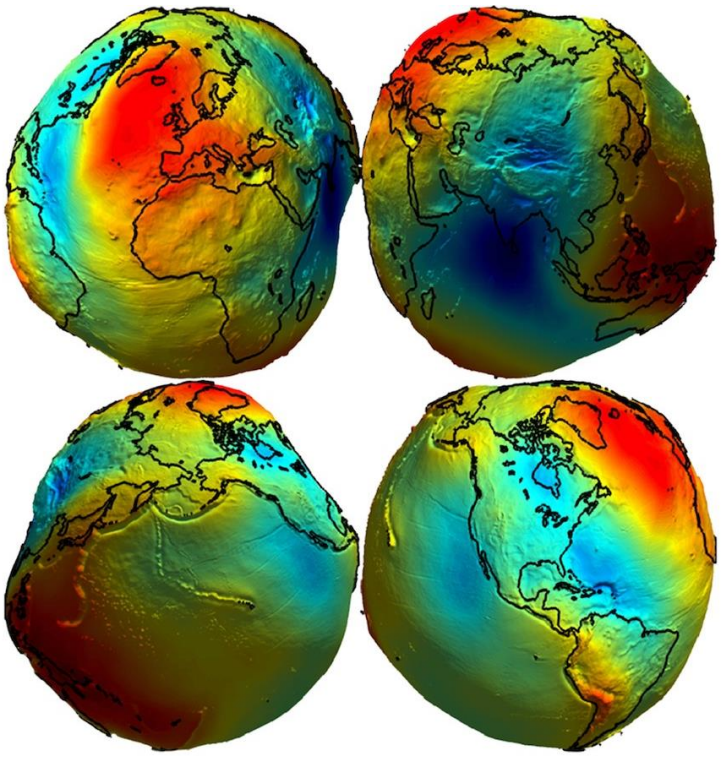


Figura II.8 Deformações da crosta terrestre
Fonte: https://timeandnavigation.si.edu/multimedia-asset/geoid

II.7 Sistemas de Informação Geográfica

Cada vez mais SIG ou Sistemas de Informação Geográfica *(GIS – Geographic Information Systems)* vem sendo utilizados em questões relacionadas a transportes (DANTAS et al. (1996). SIG são definidos por SANTOS et al. (2010) como sistemas informatizados que tem a finalidade de adquirir, armazenar e analisar dados geográficos, para MITCHELL (1999) os Sistemas de Informação Geográfica são subutilizados e quase que exclusivamente relegados a elaboração de mapas, porém existe um potencial bem maior em tais ferramentas, como por exemplo, descoberta de relações entre diversos fenômenos geoespaciais. DUNCAN et al. (2009) classificam os SIG como plataformas que mapeiam e manipulam dados com objetivo de examinar as relações e padrões de informação geograficamente referenciada, e ressalta possíveis áreas de utilização, como, saúde, transporte e serviços.

Com base no exposto acima, concluímos que SIG são fundamentais para análise dos dados georreferenciados, e sua utilização facilitará a visualização dos dados e suas diversas interações. O método proposto prevê a utilização da biblioteca *Basemap* (integrante do pacote *matplotlib* do *Python*), *Google Earth e Google Fusion Tables* como *softwares* a serem utilizados na representação dos dados coletados. Todas as ferramentas que serão empregadas durante o capítulo de avaliação experimental serão apresentadas com maior riqueza de detalhes na seção seguinte.

II.8 Algoritmo *DBSCAN*

A principal etapa do trabalho consiste na utilização de técnicas de mineração de dados que sejam capazes de encontrar pontos de estrangulamento na malha rodoviária da RMRJ. Devido às características do problema a ser endereçado neste trabalho, e com base na literatura pesquisada, optamos por utilizar dois algoritmos de agrupamento, um por densidade e outro baseado em *grid*. Nesta seção apresentamos o algoritmo DBSCAN *(Density-Based Spatial Clustering of Applications with Noise)* que é um dos algoritmos de agrupamento por densidade mais conhecidos e utilizados pela comunidade acadêmica (ESTER et al., 1996). O DBSCAN possui boa *performance* em *datasets* com grande quantidade de dados e permite a identificação de grupos de diferentes formas e tamanhos (TANG et al., 2015).

II.8.1 Vantagens e Desvantagens do DBSCAN

Podemos considerar como principais vantagens do DBSCAN:

- Não há necessidade de especificar previamente a quantidade de *clusters*;
- Possibilidade de identificar também os *outliers* durante a fase de agrupamento;
- Capaz de agrupar arbitrariamente cluster de diferentes tamanhos e formas;
- Funcional quando aplicado a *datasets* com grande quantidade de registros;

- Fácil configuração devida necessidade de apenas dois parâmetros (*MinPts* e ε).

Também existem limitações que devemos considerar, porém devido a natureza dos conjuntos de dados selecionados, essas limitações não impactam de modo relevante nos resultados do experimento.

- Limitado nos casos de clusters com densidades variáveis;
- Limitado nos casos que utilizem multidimensionais;
- Caso os dados e a escala não estejam bem compreendidos, a escolha da distância e limites pode ser de difícil análise;

II.8.2 Parâmetros de Estimativa do DBSCAN (ε *e MinPts)*

O princípio básico do DBSCAN é bem simples, baseia-se na definição dos conceitos de vizinhança e quantidade mínima de pontos para formação do *cluster*. O parâmetro ε define o raio que delimita a área de vizinhança de um ponto, enquanto *MinPts* representa a quantidade mínima de pontos que devem existir na vizinhança do ponto para que seja formado um cluster, as variações de *MinPts* são descritas a seguir:

II.8.3 Ponto Central, Ponto Periférico e Ruído

Uma vez definidos os conceitos básicos de vizinhança e quantidade mínima de pontos, podemos aprofundar o entendimento do DBSCAN com as definições de Ponto Central, Ponto Periférico e Ruídos.

- Ponto Central: Um ponto *p* é considerado ponto central se houver ao menos a quantidade mínima de pontos (*MinPts*) em sua vizinhança *(ε).* No exemplo da Figura II.9 consideramos *MinPts* igual a seis. O ponto $p1$ é considerado central por conter ao menos outros seis pontos dentro de sua vizinhança definida por ε;
- Ponto Periférico: Um ponto p é considerado ponto periférico se não houver a quantidade mínima de pontos em sua vizinhança, mas pertencer à vizinhança de um ponto central. No mesmo exemplo da Figura II.9, o ponto $p2$ é um ponto periférico ou de borda, pois na área de sua vizinhança definida por ε existem apenas cinco pontos, mas $p2$ é um ponto que pertence à vizinhança do ponto central $p1$;
- Ruído: São os pontos que não possuem em sua vizinhança a quantidade mínima estabelecida e não pertencem a vizinhança de qualquer ponto central.

Na Figura II.9 o ponto $p3$ é considerado ruído, pois em sua vizinhança não possui qualquer outro ponto, seja ele central ou periférico.

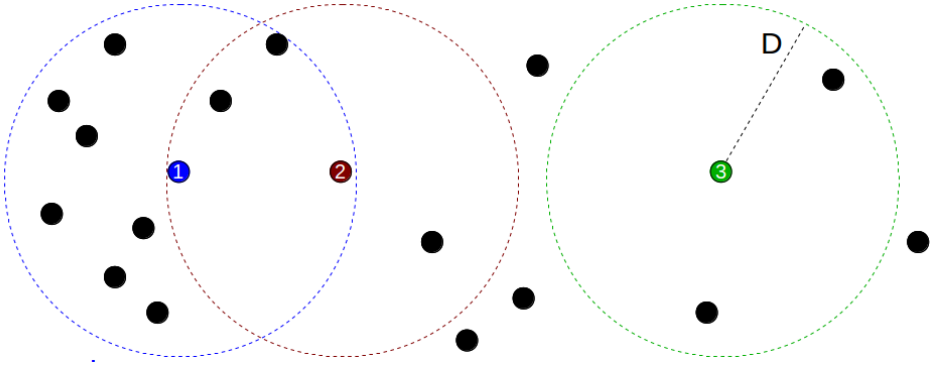


Figura II.9 Ponto Central, Periférico e Ruído – Exemplo 1
https://commons.wikimedia.org/wiki/File:Classificacao.png

A Figura II.10 apresenta mais um exemplo de agrupamento utilizando DBSCAN considerando *MinPts* igual a três. Os pontos vermelhos são centrais, os amarelos são periféricos e o azul ruído.

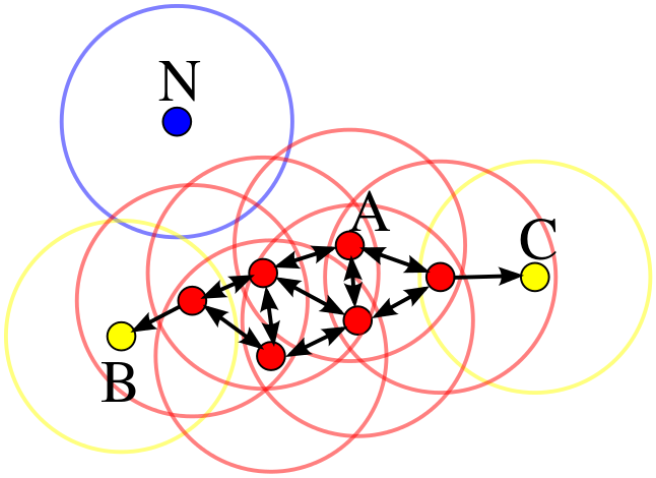


Figura II.10 Ponto Central, Periférico e Ruído – Exemplo 2
Fonte: https://commons.wikimedia.org/wiki/File%3ADBSCAN-Illustration.svg

II.8.4 DBSCAN vs KMeans

A Figura II.11 apresenta uma comparação entre o DBSCAN e outros algoritmos baseados em *KMeans* para quatro *datasets* de diferentes tamanhos e formas. A natureza espacial dos dados e o tipo de agrupamento que desejamos identificar são muito próximos aos resultados apresentados pelo DBSCAN. No primeiro *dataset* se ambos os círculos representassem vias de acesso engarrafadas, seria exatamente essa formação de *clusters* regida pela densidade das diversas partes que estaríamos buscando, assim como ocorre nos demais casos da figura abaixo.

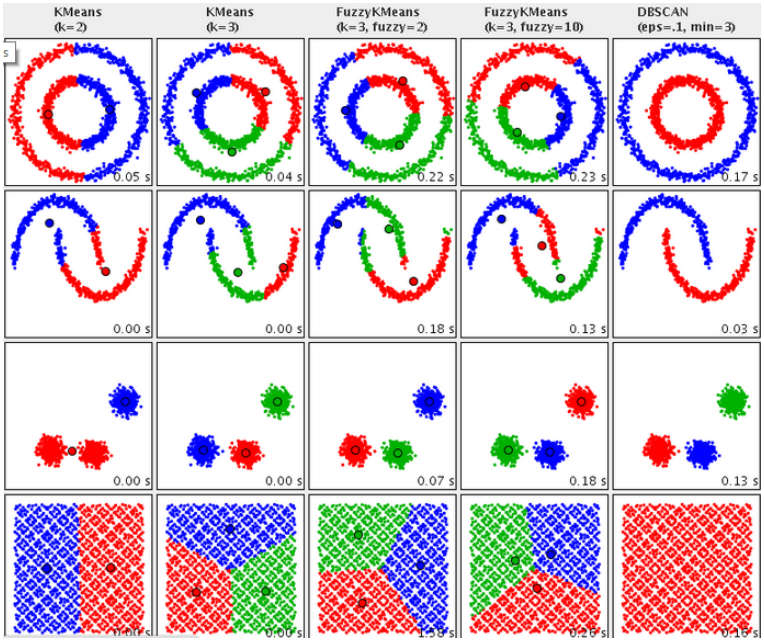


Figura II.11 DBSCAN vs Variações KMeans
Fonte: http://commons.apache.org/proper/commons-math/userguide/ml.html

II.9 Algoritmo *Grid Growing*

O outro algoritmo escolhido para fazer parte do experimento é o *Grid Growing Clustering* (ZHAO et al., 2015). Este algoritmo utiliza uma estrutura de grid e uma técnica de agrupamento que tem como grande benefício a eficiência computacional. No artigo original o método foi utilizado para identificar pontos de interesse com base em embarques e desembarques de taxis. Uma das contribuições do presente trabalho é estender a aplicação inicialmente proposta pelos autores, adaptando o algoritmo utilizado no experimento para auxiliar na identificação de estrangulamentos no trânsito.

Nos últimos anos vêm surgindo alguns trabalhos que correlacionam técnicas baseadas em *grid* com algoritmos de agrupamento tendo como objetivo uma melhor representação dos fenômenos de mobilidade urbana pelo mundo. RORIZ JUNIOR et al. (2016) propõe um método de agrupamento combinando DBSCAN e processos de fluxo de dados baseados em *Complex Event Processing* (CEP) com objetivo de detecção contínua de *clusters* em tempo real. MOREIRA-MATIAS et al. (2016) utilizam técnica de decomposição da cidade em *grid* com agrupamento dos pontos de embarque e desembarque de taxis de Portugal com a finalidade de identificar pontos de interesse. CAO et al. (2009) desenvolvem um algoritmo agrupamento híbrido mesclando métodos hierárquicos e de grid com a finalidade de descobrir padrões em dados especiais.

II.9.1 Parâmetros de Estimativa do *Grid Growing*

O primeiro passo para utilização do método é a definição da região que se pretende analisar, essa região é delimitada pelas coordenadas de uma das diagonais do quadrilátero que delimita a área desejada, longitude ($Xmin$ e $Xmax$) e latitude ($Ymin$ e $Ymax$). A partir desta premissa definiremos a quantidade de malhas que a área selecionada deverá conter (Nx no eixo da longitude e Ny no eixo da latitude), e por consequência o tamanho das mesmas. As observações contidas no *dataset* representadas pelos pontos $D(x,y)$ são plotadas no mapa e são atribuídas a cada uma das malhas do *grid* utilizando a fórmula (Figura II.12).

$$t = \left\lceil \frac{x_{\max} - x}{x_{\max} - x_{\min}} \cdot n_x \right\rceil$$
$$s = \left\lceil \frac{y_{\max} - y}{y_{\max} - y_{\min}} \cdot n_y \right\rceil$$

Figura II.12 Cálculo da malha de cada ponto

O próximo passo é definir a quantidade de pontos m para servir de *seed* do processo de agrupamento. *Seeds* são os pontos que utilizaremos para determinar as regiões por onde iniciaremos a formação dos *clusters*. Durante os experimentos realizados utilizaremos como critério para escolha de m as trezentas malhas mais densas da região determinada. Esse valor foi definido durante a fase de testes, após verificar que não havia aumento considerável na quantidade de *clusters* formados quando m > 300. Uma vez definido m, submeteremos cada uma das malhas aos critérios previamente selecionados, como quantidade mínima de pontos por malha e se o algoritmo deverá varrer apenas os quatro vizinhos situados nos eixos horizontais e verticais ou varrer todos os oito vizinhos possíveis, incluindo os localizados nas diagonais. A Figura II.13 demonstra os agrupamentos de um mesmo *dataset* utilizando os métodos de 4-vizinhos (b) e 8-vizinhos (c), onde cada cor representa um *cluster*.

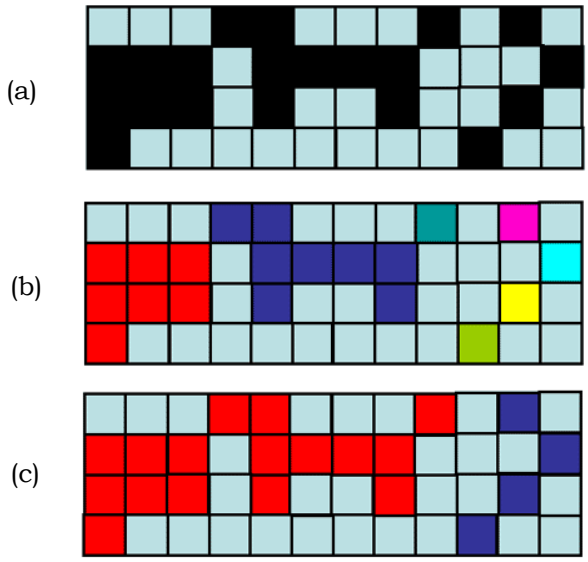


Figura II.13 Agrupamento utilizando método 4-vizinhos e 8-vizinhos
Fonte: https://www.cs.auckland.ac.nz/courses/compsci773s1c/lectures/ImageProcessing-html/topic3.htm

A Figura II.14 apresenta o cálculo utilizado no algoritmo para varrer os vizinhos de cada malha e aplicar os critérios mínimos de definição de *cluster*.

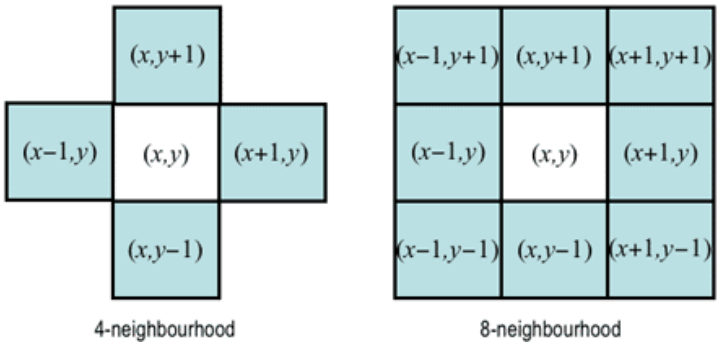


Figura II.14 Representação Cartesiana dos Vizinhos
Fonte: https://www.cs.auckland.ac.nz/courses/compsci773s1c/lectures/ImageProcessing-html/topic3.htm

Como já seria de se esperar a utilização de 4-vizinhos permite uma maior segregação de *clusters* adjacentes e consome menos recurso computacional. Porém para o tipo de problema proposto neste trabalho, o método de 8-vizinhos é mais adequado, uma vez que certamente existirão vias de acesso que cruzarão as malhas diagonalmente. Deste modo evitamos interrupção de *clusters* que se encontram nestas condições. Esse processo ocorre de forma iterativa e para todos os pontos m e ao final os pontos que não forem agrupados serão caracterizados como *outliers*.

II.9.2 Vantagens e Desvantagens do *Grid Growing*

Alguns autores compararam a eficiência entre métodos de agrupamento por densidade e baseados em *grid*, por exemplo, MONTOLIU e GATICA-PEREZ (2010) reforçam a

possibilidade do método proposto por eles limitar o tamanho dos *clusters* formados. ZHAO et al. (2015) destacam a eficiência de seu método calculando a complexidade de seu algoritmo como *O(N log N)*. Essa talvez seja a principal vantagem do *Grid Growing* sobre os demais métodos, pois com fluxo de dados contínuos em ininterruptos, o custo computacional é um fator bastante relevante alcançar os resultados esperados com a rapidez necessária. A técnica de agrupamento por *grid* utilizada neste trabalho também compartilha de pontos fortes contidos no DBSCAN:

- A quantidade de *clusters* é definida *a posteriori*;
- *Outliers* identificados durante o agrupamento;
- Forma *cluster* de diferentes formatos e dimensões;
- Aplicável a grandes *datasets*.

Porém, diferente do DBSCAN, a configuração dos parâmetros demanda maior atenção e cuidado, pois é sensível a aspectos que dependem das premissas assumidas preliminarmente, como por exemplo, o tamanho da malha, quantidade mínima de observações por malha, *seeds* e número de vizinhos considerados.

Capítulo III - Método Proposto

Neste capítulo abordaremos os detalhes do método que foi desenhado para esta pesquisa e que terá seus resultados avaliados no próximo capítulo. Serão apresentados o conjunto de dados e as ferramentas utilizados durante o experimento, as etapas e técnicas de mineração de dados selecionadas e os passos a serem seguidos para correta implementação do procedimento descrito.

Importante neste momento recordarmos o objetivo principal desta dissertação, que é a proposição de um método que possibilite a identificação de estrangulamentos no trânsito de veículos a partir do uso de dados coletados por meio de dispositivos GPS. Tais estrangulamentos serão identificados a partir da aglomeração de uma quantidade determinada de registros em uma área limitada durante certo período de tempo, e podem ter como causa motivos diversos, tais como, fechamentos total ou parcial de vias de acesso, grande volume de veículos em horários de pico ou redução da velocidade média por conta de fenômenos climáticos. A pesquisa também possui um caráter generalista, ou seja, apesar de utilizar dados referentes ao sistema de ônibus da RMRJ, sua aplicação deverá ser possível para qualquer outra cidade, desde que, estejam disponíveis as informações necessárias para utilização do método.

III.1 Ferramentas para análise dos dados

Nesta seção serão apresentadas as aplicações que fazem parte do método definido para identificação dos estrangulamentos no trânsito de grandes cidades. Também serão sugeridos alguns *softwares* alternativos aos utilizados na pesquisa, caso seja relevante para futuras aplicações deste método por parte de terceiros.

III.1.1 Bancos de Dados

Conforme veremos mais a frente, a quantidade de dados necessária para condução do experimento pode alcançar patamares bastante significativos e, portanto, é fundamental que nos preocupemos como os dados serão capturados, armazenados, apresentados, processados e disponibilizados ao usuário final (RIGAUX et al. 2002) de modo que sejam reduzidas as influência de ruídos nos resultados obtidos. Os dados são disponibilizados pela PMRJ em formatos de arquivos, JSON *(JavaScript Object Notation)* ou CSV *(Comma Separeted Value)* e a leitura pode ser realizada diretamente pelo *script* de programação simplificando a implementação. Porém, do ponto de vista de organização e gerenciamento dos dados, é uma boa prática garantir que estes dados estejam em uma única base gerida por um RDBMS *(Relational Database Management System)* (RIVEST et al., 2005).

Como os dados possuem uma característica espacial, é importante, mas não obrigatório, que sejam instaladas as extensões que permitam ao RDBMS escolhido a utilização de funcionalidades que disponibilizem consultas e operações utilizando os atributos georreferenciados. Atualmente, a maioria dos RDBMS encontrados no mercado dispõe de tal recurso, tornando a escolha pela aplicação uma questão mais subjetiva e dependente das preferências do pesquisador que conduzirá a experiência. No caso do presente trabalho são utilizadas as ferramentas *SQLite* e *PostgreSQL* com *PostGIS* devido serem *softwares* de licença livre e que não demandam custos adicionais para aquisição de licença, mas existem também outras opções comerciais ou não que ser consideradas, como, *Microsoft SQL*, *MySQL* ou *Oracle Database.*

III.1.2 Python

Python é uma linguagem de programação interpretativa de alto nível desenvolvida sob licença *open source* e administrada pela *Python Software Foundation*[8]. O fato de ter licença de código aberta permite que a comunidade execute o software para qualquer propósito, que modifique livremente seu código e distribuição para os demais usuários e a gratuidade na utilização da ferramenta. Existe hoje uma comunidade de usuários *Python* bem ativa e numerosa que promove diversas conferências e encontros por todo mundo, desenvolve novos pacotes com finalidades diversas e elabora documentação de apoio e tutoriais para aqueles que têm o interesse em aprender a linguagem.

Apesar da versão 3 estar disponível desde 2008, durante o experimento foi utilizada a versão 2.7, pois ainda existem incompatibilidades no *Python 3* que precisam ser sanadas, e uma grande quantidade de usuários ainda opta pela versão 2.7 (LUTZ, 2009). BEAZLEY e JONES (2013) enumeram algumas vantagens do Python:

- Qualidade do *software*: *Python* foca na facilidade de leitura, coerência e na manutenção do código, muito mais que outras linguagens interpretadas. A uniformidade traz facilidade na leitura dos scripts mesmo para os que não participaram da escrita do código, além de ter suporte à programação orientada a objetos;
- Produtividade: Estima-se que o tempo de elaboração de um código em *Python* seja um terço a um quinto do tempo de elaboração do código equivalente em linguagens compiladas (C, C++, Java e outras);
- Portabilidade: É multiplataforma permitindo que códigos gerados em um determinado sistema operacional sejam facilmente interpretados em outros *(Windows, Linux, Mac OS* etc);

[8] https://www.python.org/about/

- Bibliotecas: Como apontado anteriormente, uma extensa lista de bibliotecas complementares, que tornam a ferramenta extremamente poderosa e adaptável a praticamente qualquer tipo de necessidade;
- Integração: Fácil integração com outras linguagens, sendo facilmente evocada por C, C++, Java ou .NET;
- Facilidade: Devido à facilidade no aprendizado e simplicidade de uso *Python* oferece uma experiência mais agradável e menos trabalhosa do que outras linguagens.

O *Python* é uma ferramenta simples, porém muito poderosa e bastante versátil, e com apoio da comunidade desenvolvedora, inúmeros pacotes estão disponíveis com a implementação de um sem números de algoritmos, fazendo desta ferramenta uma das preferidas entre pesquisadores e estudantes para manipulação de dados (estatística, *machine learning*, *big data*, *data mining* etc) tanto para uso industrial quanto científico. A seguir uma relação dos principais pacotes utilizados no experimento.

- *matplotlib*[9]: biblioteca para plotagem de gráficos 2D (histogramas, espectros de potência, gráficos de barra, gráficos de dispersão, gráficos de pizza e muitos outros) que produz figuras em alta qualidade de publicação e uma extensa variedade de formatos em um ambiente interativo.
- *pandas*[10]: biblioteca especializada na manipulação de estruturas de dados e em ferramentas de análise;
- *numpy*[11]: pacote fundamental para projetos científicos, possibilitando o uso de *arrays* n-dimensionais, álgebra linear, transformada de *Fourier*, códigos em C/C++ e *Fortran* e muitas outras funções. Além de suas aplicações científicas, *numpy* pode ser utilizado como repositório de dados multidimensionais.
- *seaborn*[12]: biblioteca de visualização de dados estatísticos baseada no *matplotlib* que permite a elaboração de gráficos profissionais, facilitando a exploração e interpretação dos dados, devendo ser utilizada como complemento ao *matplotlib* e não em substituição.
- *math*[13]: permite acesso às funções matemáticas da biblioteca C padrão;
- *sys*[14]: este modulo prove acesso a variáveis e funções utilizadas e mantidas pelo interpretador;

[9] http://matplotlib.org/
[10] http://pandas.pydata.org/
[11] http://www.numpy.org/
[12] https://stanford.edu/~mwaskom/software/seaborn/
[13] https://docs.python.org/2/library/math.html
[14] https://docs.python.org/2/library/sys.html

- *json*[15]: este pacote permite a utilização de dados contidos em arquivos do formato JSON;
- *sqlite3*[16]: biblioteca C que permite acesso a bancos de dados no formato .db (*SQLite*) possibilitando consultas, manipulações de dados e criação e manutenção de bancos de dados por meio de comandos SQL inseridos no script do *Python*. Aplicações podem utilizar *SQLite* como armazenamento provisório dos dados e posteriormente migrar os dados para bancos mais sofisticados, como *PostgreSQL*, *MySQL* ou Oracle;
- *psycopg2*[17]: é o pacote mais popular para utilização do *PostgreSQL* integrado ao *Python*, possibilitando acesso diretamente a bancos de dados deste RDBMS.
- *datetime*[18]: módulo que possibilita a manipulação de datas e horas em diversos formatos.

A ferramenta também possibilita a utilização em diferentes ambientes de desenvolvimento integrado *(IDE – Integrated Development Enviroments)*, *ipython*, *Python Notebook*, S*pyder*, *PyCharm* e tantos outros que permitem aos usuários diferentes formas para se adaptar a linguagem, geralmente disponibilizando um editor de texto, um *debbuger* e *prompt* de comando (GOODRICH et al., 2013). Outras linguagens poderiam ser utilizadas nesta pesquisa com resultados similares, por exemplo, R que é uma ferramenta muito utilizada em estatística e que também possui farta biblioteca de pacotes para tratamento de dados disponíveis, porém Python pode ser considerada uma ferramenta mais completa para propósitos generalistas (BOWLES, 2015).

III.1.3 Google Earth

Google Earth é uma plataforma desenvolvida e mantida pela empresa *Google* que permite a visualização e interação com mapas 2D e 3D de maneira simples e intuitiva. A ferramenta não possui todos os recursos de uma ferramenta profissional de GIS, como *ArcGIS* ou *QGIS*, mas atende as necessidades da pesquisa em questão com bastante eficiência. O *Google Earth* utiliza informação de GNSS integrando imagens coletadas pelos satélites e informação georreferenciada, adotando sistema geodésico WGS84 e formato de arquivo KML *(Keyhole Markup Language)* (LOPES et al., 2015). A ferramenta é disponibilizada em duas versões, a *Free* e a *Professional (utilizada no trabalho)* incluindo funcionalidades apresentadas

[15] https://docs.python.org/2/library/json.html
[16] https://docs.python.org/2/library/sqlite3.html
[17] http://initd.org/psycopg/
[18] https://docs.python.org/2/library/datetime.html

na Tabela III.1 e que desde 2015 também pode ser utilizada gratuitamente tanto para uso pessoal quanto empresarial[19].

Tabela III.1 Comparação entre *Google Earth Pro* e *Google Earth*
Fonte: https://support.google.com/earth/answer/189188?hl=pt-BR&ref_topic=2376762

Recursos	**Google Earth**	**Google Earth Pro**
Resolução de impressão	1.000 pixels	4.800 pixels
Importação de dados SIG		ESRI .shp, MapInfo .tab
Importação de endereços em massa	Geolocalizar manualmente cada endereço	Geolocalizar automaticamente até 2500 por vez
Importação de arquivos grandes de imagens	Limitado ao tamanho da textura	Superposições de imagens
Camadas suplementares		Informações demográficas, lotes, contagem de volume de tráfego
Criação de filmes de qualidade para exportação		HD 1920x1080
Ferramentas de medição	Linha, caminho	Linha, caminho, polígono, círculo, caminho 3D, polígono 3D

Seguem algumas funcionalidades disponíveis na versão *Pro*.

- **Medidas avançadas:** medição de área poligonal ou determinação do raio afetado com medida de circunferência;
- **Impressão de alta resolução:** capacidade de impressão de fotos com resolução de 4.800 x 3.200 pixels;
- **Camadas de dados exclusivas:** informações demográficas, lotes e contagem de volume de tráfego;
- **Importação de planilhas:** capacidade de inserção de até 2.500 endereços por vez, atribuindo marcas de local e modelos de estilo em massa;
- **Importação de SIG:** visualização de arquivos em formato ESRI (.shp) e MapInfo (.tab);
- **Movie-Maker:** exportação de filmes em alta definição, em formato Windows Media e Quicktime, com até 1.920 x 1.080 pixels de resolução.

III.1.4 Google Fusion Tables

Além do *Google Earth*, outra ferramenta que nos suportará na visualização dos dados é *Google Fusion Tables*. *Fusion Tables* é uma plataforma *web* também desenvolvida e mantida pelo *Google* e que faz parte do *Google Drive*. As grandes facilidades proporcionadas pela

[19] https://support.google.com/earth/answer/189188?hl=pt-BR&ref_topic=2376762

ferramenta são a possibilidade de visualização de dados geográficos tabelados de maneira muito simples e sua fácil colaboração com a equipe do projeto, permitindo uma grande facilidade de compartilhamento de atividades e elevado nível de sinergia por parte do grupo de pesquisa (ZULAR et al., 2011). A possibilidade de editar o nível de privacidade dos arquivos também é um ponto forte, simplificando a publicação dos resultados obtidos nas análises para o público geral. Assim como *Google Earth,* o *Fusion Tables* também utiliza KML e permite além das funcionalidades geoespaciais, a elaboração de diferentes tipos de gráficos e ferramenta para visualização de grafos. A ferramenta possui limitação quanto ao *upload* de arquivos tanto no tamanho (100MB) quanto no formato de arquivos (CSV, KML, ou formatos de planilha mais utilizados).

III.2 Conjunto de Dados

O *dataset* escolhido para o trabalho é o principal insumo de todo processo, pois a partir desta grande quantidade de dados, derivarão análises complexas que demandarão grande esforço de processamento. Neste experimento são utilizados dados disponibilizados pela Prefeitura Municipal do Rio de Janeiro (PMRJ) em seu portal de dados abertos[20]. Esses dados são gerados a partir de dispositivos GPS equipados em ônibus que atendem a RMRJ e que informam a cada minuto sua localização em termos de coordenadas geográficas, latitude e longitude, velocidade instantânea e hora da coleta segundo GPS. Conforme visto no Capítulo II, esses ônibus formam uma Rede Veicular *Ad Hoc*, pois enquadram-se exatamente na definição proposta por FONSECA e VAZÃO, 2013, que utilizam uma Rede de Sensores Sem Fio (GPS) para transmitir os dados necessários pelo sistema.

A PMRJ disponibiliza os dados em formato JSON e CSV, sendo possível a utilização de ambos os formatos conforme verificaremos mais adiante. Ela também disponibiliza outros dados que podem vir a ser interessantes em trabalhos futuros, tais como, informações sobre estações de ônibus e BRT, pontos do trajeto das diversas linhas em operação, GTFS dos ônibus (*General Transit Feed Specification* ou Especificação Geral de *Feeds* de Transporte) e dados de outros modais como barcas, metro e trem. A seguir analisaremos os atributos contidos no *dataset*.

III.2.1 Estrutura dos Dados

Este conjunto de dados possibilita acesso às informações sobre a posição e velocidade dos ônibus da cidade do Rio de Janeiro, por linha de ônibus e pelo número identificador do carro, através da internet, num dado momento[21]. Esses dados são atualizados a cada minuto o

[20] http://data.rio/group/transporte-e-mobilidade
[21] http://dadosabertos.rio.rj.gov.br/apitransporte/apresentacao/pdf/documentacao_gps.pdf

que nos permite analisar a evolução da movimentação dos veículos ao longo do tempo. Se contabilizarmos a quantidade de registros gerados a cada ano, chegaremos a um valor próximo a 4,2 x 10^9, o que faz com que haja uma preocupação de se buscar formas otimizadas para armazenamento e tratamento destes dados.

Apesar dos dados estarem disponíveis a cada minuto, a coleta para o experimento será feita respeitando intervalo de dez minutos, as justificativas para tal procedimento serão esclarecidas na seção que abordará o tratamento dos dados. A estrutura do *dataset* é apresentada na Tabela III.2.

Tabela III.2 Descrição do *Dataset* – Conjunto GPS ônibus
Fonte: http://dadosabertos.rio.rj.gov.br/apitransporte/apresentacao/pdf/documentacao_gps.pdf

Metadados associados ao dataset – GPS ônibus			
Campo	**Descrição**	**Tipo**	**Tamanho**
DataHora	Data e hora da coleta do dado	DATETIME	23
Ordem	Identificação alfanumérica encontrada na lateral dos ônibus	VARCHAR	7
Linha	Linha do ônibus	VARCHAR	7
Latitude	Latitude do ônibus na coleta (GPS, WGS84)	NUMERIC	11
Longitude	Longitude do ônibus na coleta (GPS, WGS84)	NUMERIC	11
Velocidade	Velocidade do ônibus na hora do coleta do dado	NUMERIC	6

- **DataHora**: Informação do dia e horário em que a coleta dos dados foi feita utilizando como parâmetro os dispositivos GPS de cada ônibus, sendo esta informação formatada no padrão *Datetime* (DD-MM-AAAA HH:MM:SS);
- **Ordem:** Código que identifica cada um dos ônibus de onde as coletas são realizadas, geralmente para RMRJ este código é alfanumérico e composto de uma letra e cinco algarismos;
- **Linha:** Representa a linha (trajeto) que cada um dos ônibus está associado em um determinado momento. Importante ressaltar que diferente da Ordem, esta informação pode variar ao longo do tempo, pois um ônibus pode ser designado para rodar em linhas diferentes em períodos diferentes;
- **Latitude e Longitude:** Descreve as coordenadas geográficas no momento da coleta em formato geodésico de graus decimais;
- **Velocidade**: Descreve a velocidade instantânea no momento da coleta dos dados em km/h.

A seguir na Tabela III.3 são apresentados alguns exemplos de registros do *dataset* GPS ônibus.

Tabela III.3 Exemplo de registros – Conjunto GPS ônibus
Fonte: http://data.rio/dataset/gps-de-onibus

dataHora	ordem	*linha*	latitude	longitude	velocidad…
04-08-2016 14:25:14	B19565	SVA665	-22.814819	-43.341141	0
04-08-2016 10:44:17	B73133	SVA665	-22.81616	-43.34177	0
04-08-2016 14:10:58	C27053	SV940	-22.81723	-43.30117	0
04-08-2016 14:19:48	B10138	SV901	-22.788816	-43.162197	0
04-08-2016 14:02:05	D58645	SV777	-22.890657	-43.450222	0.56
04-08-2016 13:59:54	C27198	SV774	-22.81485	-43.3032	31
04-08-2016 12:51:22	C27195	SV774	-22.827129	-43.319641	0
04-08-2016 11:40:38	C27116	SV774	-22.80876	-43.328331	37
04-08-2016 11:03:50	C27114	SV774	-22.807249	-43.3274	0
04-08-2016 14:12:38	C20228	SV390	-22.943722	-43.387341	0
04-08-2016 14:07:52	C20134	SV390	-22.944387	-43.387989	0
04-08-2016 14:25:22	B19602	SV376	-22.902121	-43.195221	0
04-08-2016 14:04:49	B19612	SV376	-22.901011	-43.19466	0
04-08-2016 14:25:31	B25577	SV249	-22.89786	-43.312069	0
04-08-2016 13:48:13	D53682	SR397	-22.901699	-43.554272	0

Os pontos referentes à coleta de uma semana da linha 455 são representados graficamente na Figura III.1. A plotagem dos dados foi feita utilizando a ferramenta *Google Fusion Tables*.

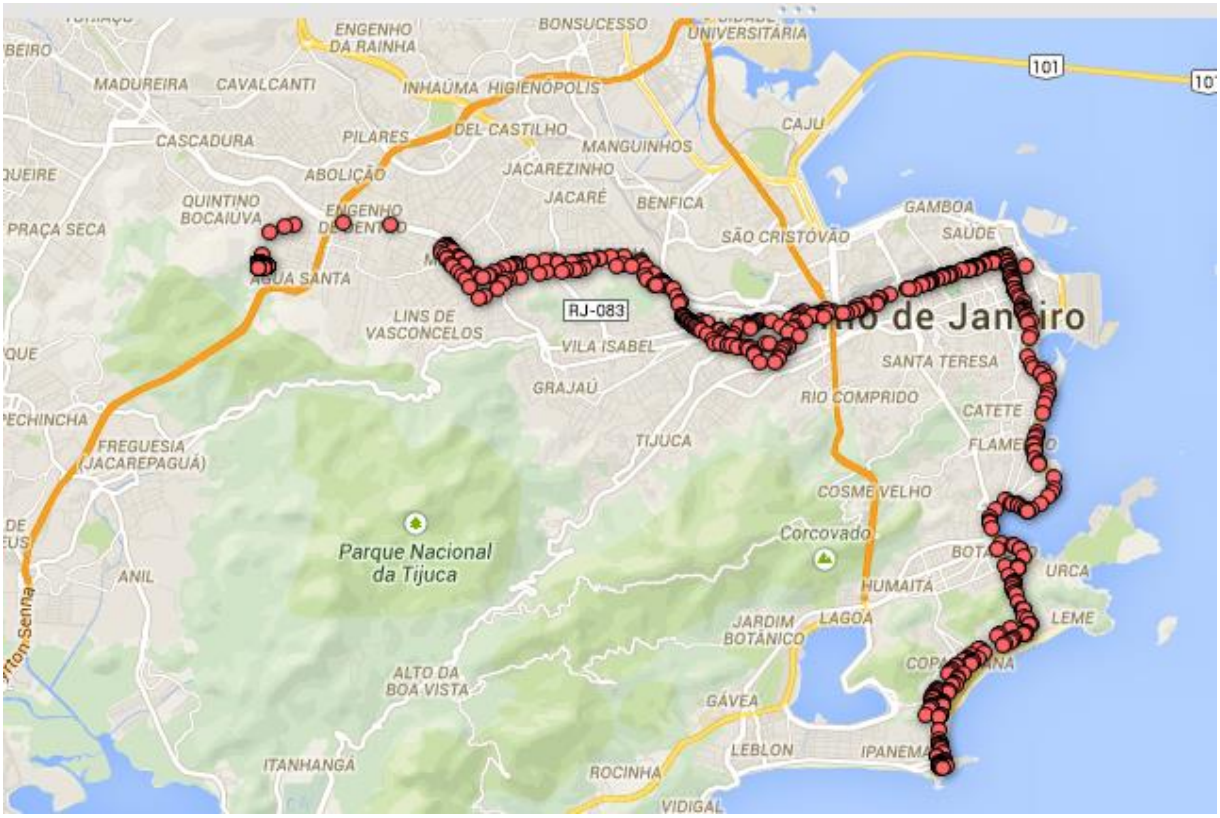


Figura III.1 Representação Espacial do *Dataset* GPS ônibus – Linha 455
Fonte: https://goo.gl/v5d2BU *(Google Fusion Tables)*

III.3 Processo

Nesta seção abordaremos o processo estabelecido pelo método proposto neste trabalho para identificação de estrangulamentos no trânsito com base em dados georreferenciados coletados a partir de ônibus da RMRJ equipados com dispositivos GPS. A Figura III.2 apresenta o processo utilizando um fluxograma.

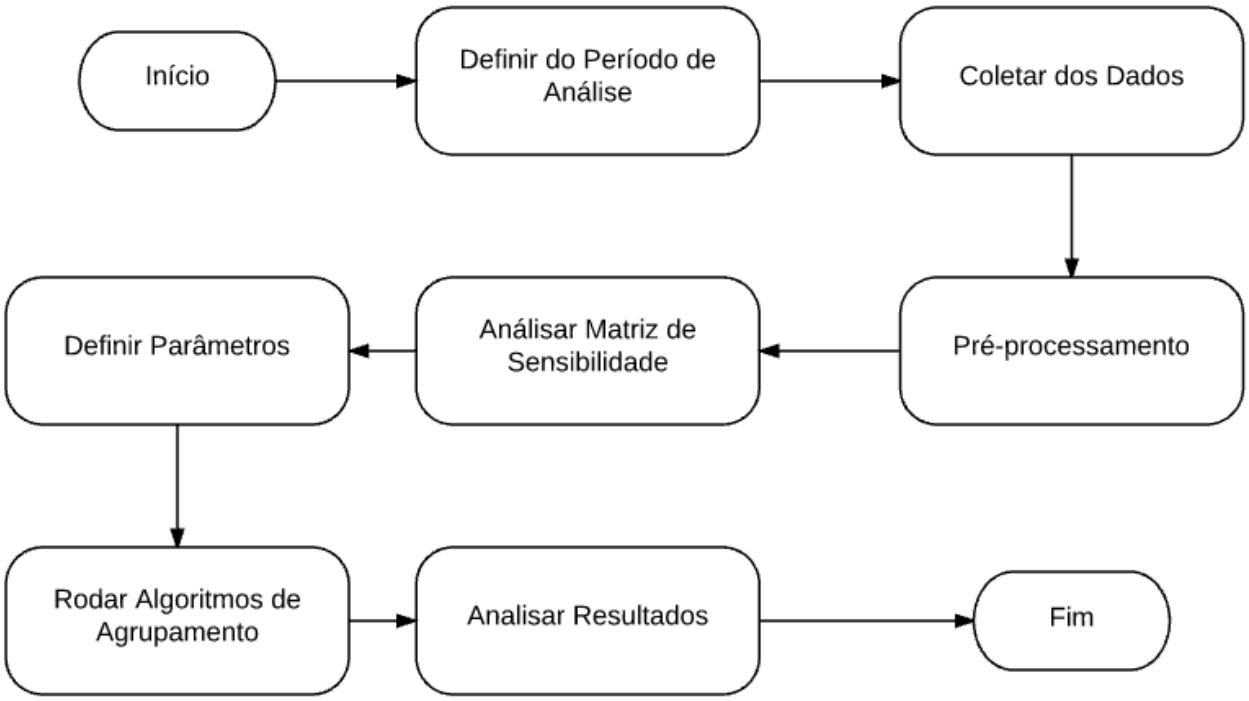


Figura III.2 Processo de identificação de estrangulamentos

1. Definição do período de análise: O primeiro passo é estabelecer o período de análise dos dados e o intervalo entre cada coleta. A definição dessas premissas é fundamental para a estimativa dos parâmetros, pois a quantidade total de observações vai orientar a formação dos *clusters*. No experimento serão utilizadas seis coletas realizadas a cada dez minutos, com cada coleta contendo aproximadamente oito mil observações. Maiores detalhes na Seção III.4;
2. Coleta dos dados: O passo seguinte é a coleta dos dados referentes ao período definido disponibilizados no servidor do LAB-MOB. Esse servidor recebe e armazena continuamente os arquivos do servidor da PMRJ, registrando o histórico de movimentações dos ônibus desde abril/2014;
3. Pré-processamento: Após definir o *dataset* preliminar é necessário que sejam aplicados procedimentos que garantam a integridade dos dados e a confiabilidade dos resultados. Este passo será explorado com mais detalhes na Seção III.5;
4. Análise de sensibilidade: Esta etapa serve como preparação para a etapa de agrupamento, analisando a evolução da quantidade de *clusters* gerados conforme variação dos parâmetros de entrada e será detalhada na Seção III.7;
5. Definição dos parâmetros: Uma vez que as características do experimento foram definidas e a análise de sensibilidade realizada, prosseguimos com a escolha dos parâmetros de entrada para algoritmos de agrupamento (Seção III.7);

6. Algoritmo de agrupamento: As técnicas de agrupamento selecionados para o trabalho são detalhadas nas Seções II.8 e II.9;
7. Analisar resultados: Os resultados obtidos são analisados utilizando como parâmetro situações de referência em que o sistema apresente baixa saturação. Em caso de identificação de gargalos fora dos padrões verificados comumente, devemos analisar as prováveis origens de tais estrangulamentos no trânsito e seus impactos diretos nas circunvizinhanças.

III.4 Definição do período de análise

A definição do período de análise dos dados levou em consideração dois aspectos fundamentais. Primeiro a estabilidade na rota dos ônibus que atendem a RMRJ, de modo a reduzir variações no comportamento dos agrupamentos por conta exclusiva de mudanças na roteirização do trânsito. A partir de 2015 diversas linhas de ônibus foram criadas, extintas ou tiveram seus trajetos alterados com objetivo de racionalizar o uso do transporte rodoviário pelos cidadãos, implantando corredores BRT pela cidade e de novas linhas alimentadoras às principais estações[22]. Como estas mudanças ocorreram de forma lenta ao longo do ano, a utilização de dados de 2015 poderia comprometer análises de séries históricas e nos levar a falsas conclusões. Portanto decidimos pela utilização de dados referentes ao ano de 2014.

Outro aspecto considerado foi a escolha de dias em que fossem reportados eventos extraordinários e que tivessem relevante impacto no trânsito da cidade. Foram selecionadas datas de jogos de Copa do Mundo, dias de greve dos rodoviários e de acidentes que comprometessem importantes vias de acesso da cidade. Deste modo, buscamos avaliar a eficácia do método na identificação de possíveis estrangulamentos na mobilidade urbana.

III.5 Etapa de pré-processamento

A etapa de pré-processamento tem como objetivo adequar os dados coletados a partir dos servidores da PMRJ a utilização nos *scripts* desenvolvidos para este trabalho. Por adequação entenda-se a exclusão de registros que não representem a realidade, principalmente por problemas apresentados durante a transmissão dos dados entre dispositivo GPS e a PMRJ. Alguns filtros devem ser aplicados ao *dataset* original para que sejam identificados os registros que apresentem “ruídos” e devidamente expurgados da base de dados final que será submetida às etapas de agrupamento. Seguem as justificativas para os filtros propostos:

[22] Site do Portal G1 de 02/10/2015 as 05:00hs

- **Posicionamento**: alguns registros podem apresentar localizações geográficas não coerentes com a realidade, como por exemplo, observações fora da RMRJ. Para reduzir o impacto nestes casos utilizamos uma área limite para análise dos dados, e qualquer observação fora deste limite é descartada. Esta área está compreendida entre os pontos (longitude, latitude) -43.8, -23.1 e -43.0, -22.6. É possível também que algumas observações coincidam com áreas de oceano ou lagoas, nestes casos poderíamos utilizar a função *is.land* do pacote *Basemap (Python),* porém isso acarretaria na eliminação de alguns registros próximos a região litorânea que por consequência do erro esperado do dispositivo GPS pudessem ter sido reconhecidos como submersos. Portanto, nestes casos, aceitaremos essas observações como válidas, mas sempre avaliando os possíveis impactos sobre os resultados;
- **Velocidade:** existem registros que apresentam valor de velocidade instantânea pouco provável de se observar na prática, o que incita dúvida quanto ao correto funcionamento do dispositivo GPS. O valor de corte para este parâmetro é de 150km/h, e sob tais condições a redução no tamanho da amostra é de aproximadamente 0,1%;
- **Horário:** durante as coletas foram identificados registros que não apresentavam o atributo DataHora coerente com o horário em que tais coletas foram feitas. Esse fato também levanta a possibilidade de mau funcionamento dos GPS instalados nos ônibus. Portanto, optou-se por excluir tais registros de modo a não aumentar indevidamente a quantidade de observações do *dataset*;
- **Linha:** dependendo da finalidade do experimento, o atributo Linha pode ser incluído na base de filtros com objetivo de identificar os veículos que não apresentam nenhuma linha de ônibus regular no momento da coleta. Isso pode significar, por exemplo, que o veículo se encontra na garagem ou a caminho de e não disponível para utilização.

III.6 Garagens dos ônibus

Durante a etapa de avaliação experimental, os resultados dos agrupamentos dos eventos alvo serão confrontados com os dos eventos referência, isto para que sejam identificados os casos que se apresentam como “pontos fora da curva”. Para facilitar a interpretação dos dados plotados, são acrescentadas as localizações geográficas de quarenta garagens de empresas de ônibus do Rio de Janeiro. Deste modo esperamos que falsos positivos não sejam apontados durante o experimento como, por exemplo, identificar um estrangulamento na madrugada que na verdade nada mais seria uma garagem repleta de veículos. As localizações das garagens também devem servir como uma espécie de

treinamento para o método, podendo ser rodado durante a madrugada com intuito de verificar se no posicionamento das garagens são formados *clusters* dos veículos estacionados. Após este teste foram identificadas algumas garagens que não apresentaram agrupamentos, e posteriormente confirmada a informação que estas garagens realmente estavam desativadas. Também foram identificados locais que oficialmente não são garagens de empresas de ônibus, mas que servem para pernoite de veículos. Nos gráficos do experimento as garagens são identificadas pelos círculos cinza espalhados pela cidade do Rio de Janeiro.

III.7 Análise de Sensibilidade

Um ponto fundamental para uma boa aplicação das técnicas de agrupamento propostas é a correta seleção dos parâmetros de que servirão de *input* nos algoritmos. No caso do DBSCAN, *MinPts* e ε, e para *Grid Growing, MinPts e* $I(n,n)$. Para ambos os algoritmos propomos uma análise de sensibilidade inovadora que nos permitirá entender a formação de *clusters* conforme variação dos parâmetros de entrada.

A análise de sensibilidade é um processo pelo qual obteremos os parâmetros para os algoritmos que tenham melhores condições de identificar possíveis estrangulamentos. O processo inicia com a elaboração da matriz sensibilidade que é um gráfico do tipo mapa de calor o qual seu gradiente de cores varia conforme a quantidade de *clusters* formados e composto por uma matriz bidimensional com $I(n,n)$ ou ε no eixo y e *MinPts* no eixo x. Deste modo, podemos comparar as análises de diferentes dias e escolher os parâmetros que melhor identifiquem os fenômenos observados. A Figura III.3 apresenta um exemplo de uma análise de sensibilidade em mapa de calor.

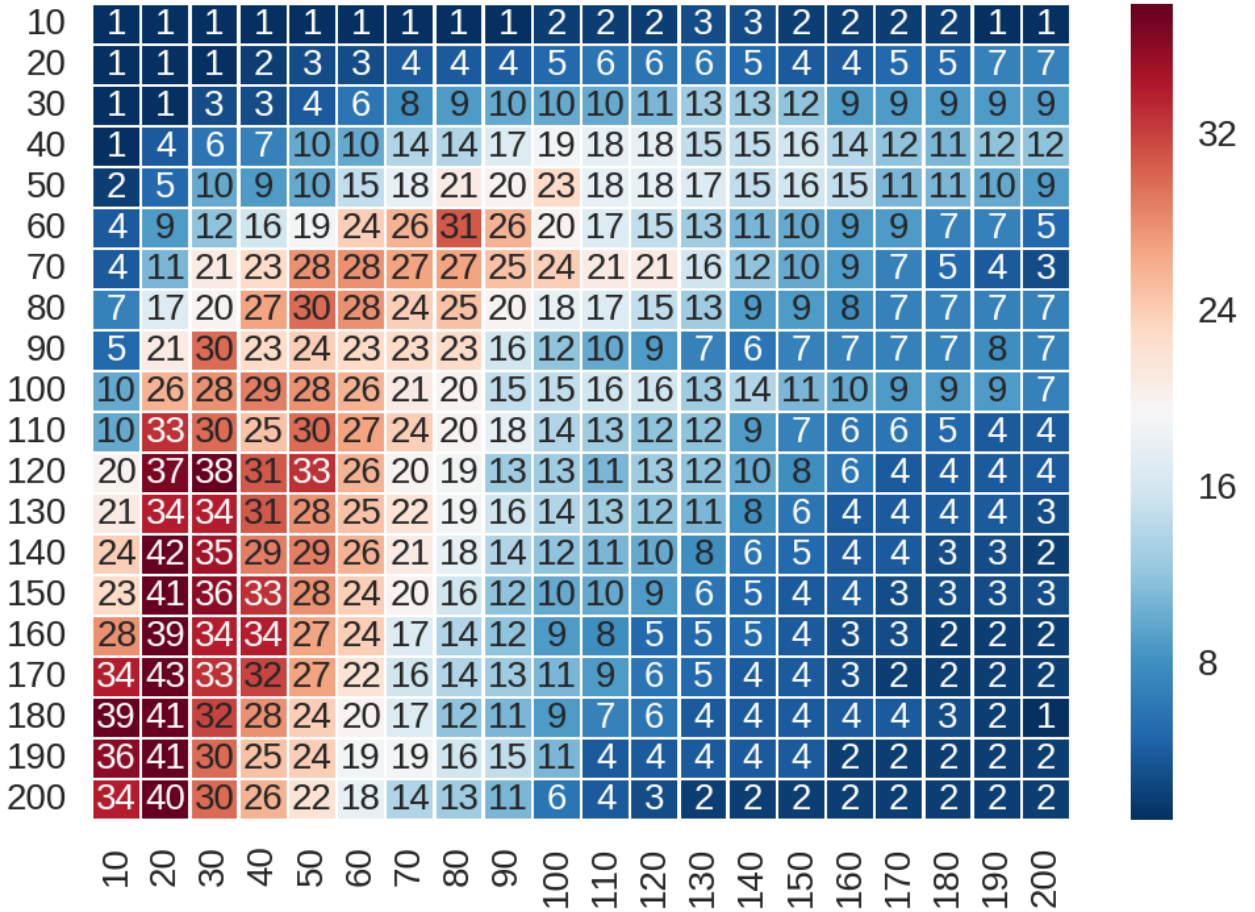


Figura III.3 Matriz para Análise de Sensibilidade

Uma boa avaliação do gráfico é fundamental para a correta interpretação dos resultados obtidos. É necessário que se conheça previamente os padrões de cada dia da semana e de cada período do dia, possibilitando a identificação de variações no comportamento padrão que apontem na direção de eventos excepcionais. Mas não só a quantidade de *clusters* formados é relevante em uma avaliação prévia, mas também as características destes clusters. Por exemplo, existe alguma semelhança entre os resultados obtidos a partir da técnica de *Grid Growing* com parâmetros 50x20 (Figura III.4) e 160x140 (Figura III.5)? Em ambos os casos o resultado de agrupamentos é o mesmo, ou seja, cinco. Mas as características de cada processo de agrupamento são bem distintas.

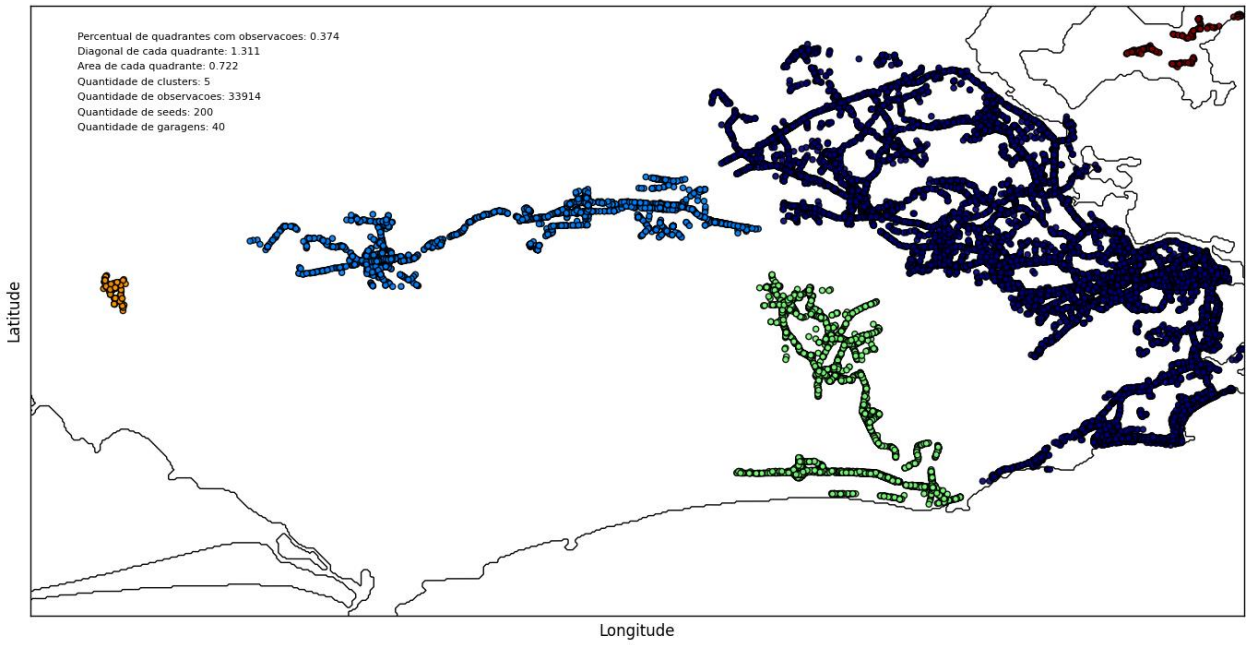


Figura III.4 Agrupamento *Grid Growing* com Parâmetros 50x20

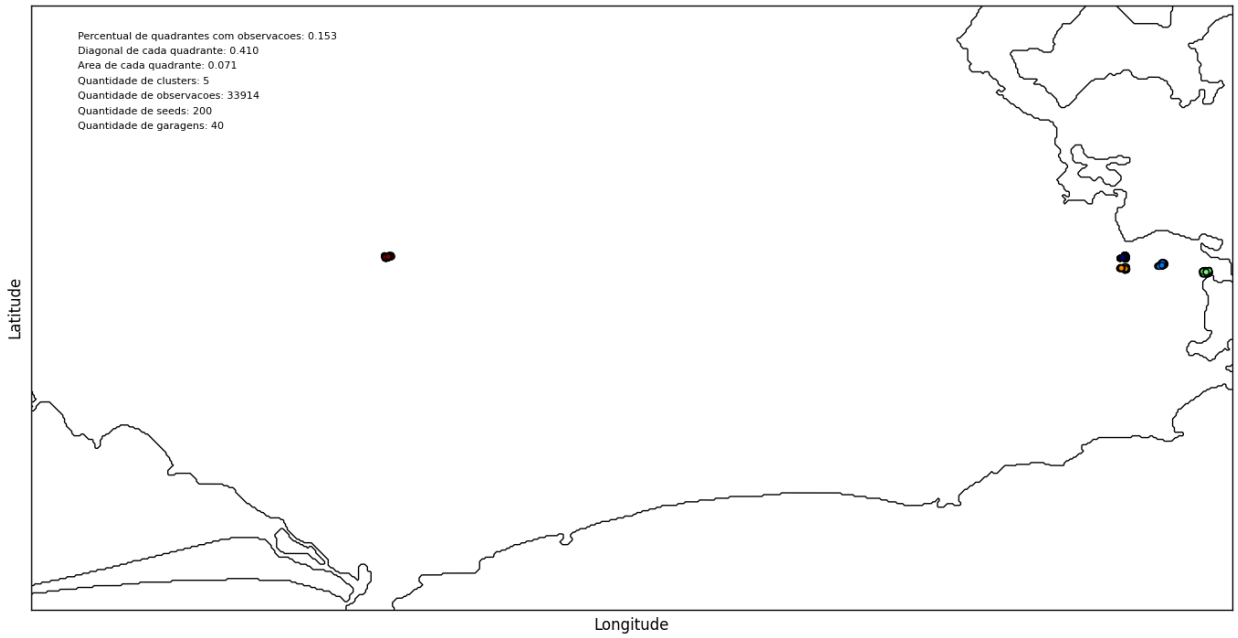


Figura III.5 Agrupamento *Grid Growing* com Parâmetros 160x140

Na Figura III.4 observamos cinco grupos bem extensos em área de abrangência e numerosos em quantidade de observações, completamente diferente do apresentado na Figura III.5. Essa diferença se dá pela menor restrição na seleção dos quadrantes adjacentes elegíveis a formação de *clusters* no primeiro caso (maior área por malha e menor quantidade mínima de pontos). Já no segundo caso, os critérios bem mais restritivos impossibilitam a formação de *clusters* “gigantes”. Os dois casos podem ser úteis, desde que saibamos o que estamos procurando e apliquemos a configuração adequada para esta finalidade.

O método de avaliação das condições de trânsito apresentado neste trabalho, propõe uma forma inovadora de determinação dos parâmetros de entrada para os algoritmos de

agrupamento utilizados. Esta utiliza uma combinação de análises qualitativas e quantitativas com intuito de maximizar os resultados esperados. Durante a análise qualitativa dos parâmetros, devemos dividir o gráfico de calor em quatro quadrantes (Figura III.6) e avaliar as principais características de cada um deles. Importante ressaltar que a avaliação de cada quadrante deve ser feita com base na quantidade aproximada de pontos e especificidades da geografia analisada, devendo sempre ser revista quando tais parâmetros forem alterados. As análises a seguir foram feitas com base características identificadas no experimento do Capítulo IV -

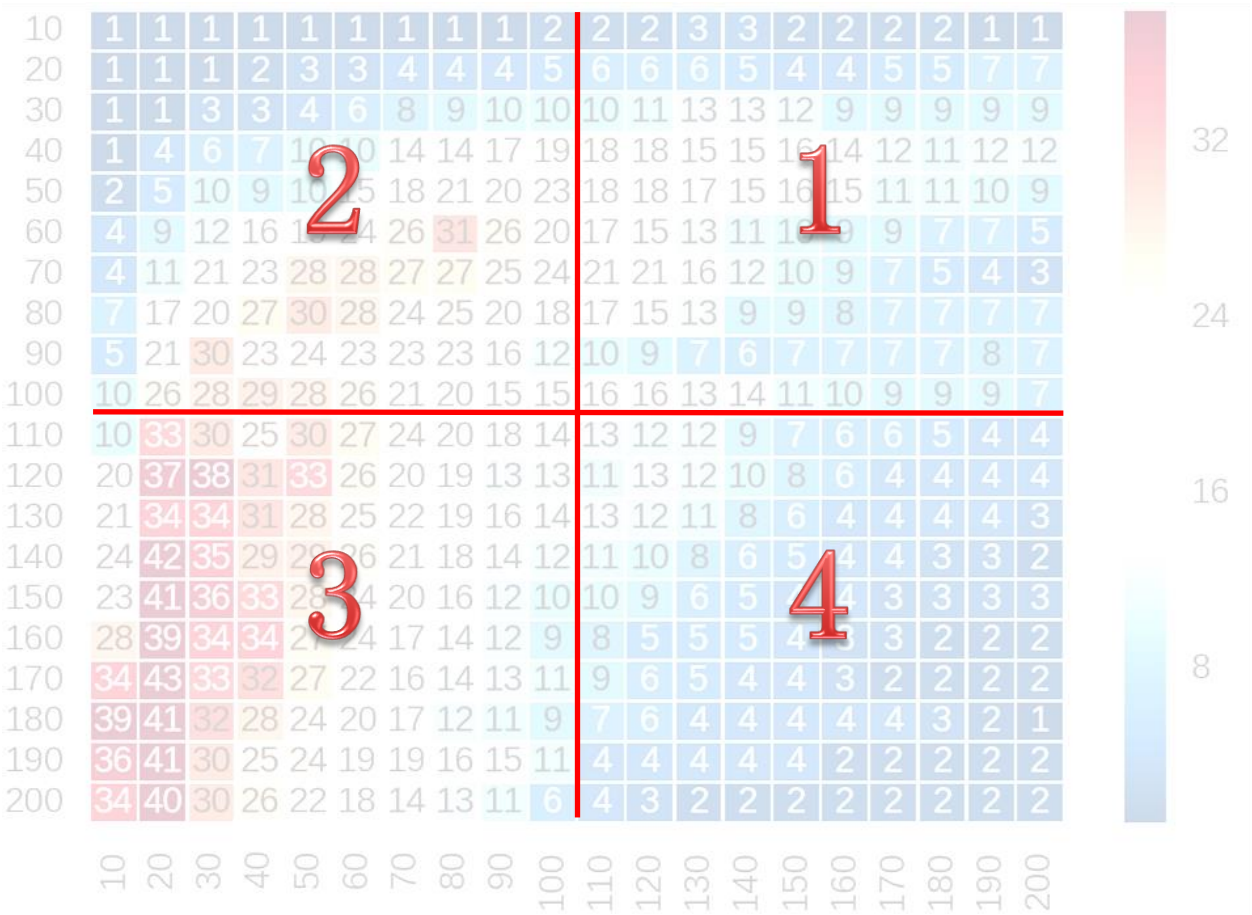


Figura III.6 Análise de Sensibilidade por Quadrante

- Primeiro quadrante: pouco restritivo com relação à área, porém mais seletivo quanto a quantidade mínima de observações para formação de *clusters*, pode ser uma alternativa para identificação de gargalos em áreas de menor densidade populacional (Zona Oeste) ou fora do horário de pico (6hs as 9hs e 17hs as 20hs);
- Segundo quadrante: pouco restritivo tanto na área quanto na quantidade mínima de observações para formação de *clusters*, por isso facilita a formação de agrupamentos muito extensos a qualquer horário do dia, não sendo ideal para identificação de congestionamentos localizados;

- Terceiro quadrante: restritivo quanto a área analisada e menos rígido quando ao limite mínimo de observações. Em horários de pico é o quadrante com maior quantidade de agrupamentos formados, sendo útil na identificação de estrangulamentos na maior parte da cidade, incluindo áreas de grande densidade demográfica, como Centro e Zona Sul;
- Quarto quadrante: maior restrição quanto à área e quantidade mínima de observações para formação de *clusters*. Tende a formar poucos e pequenos agrupamentos sendo útil na identificação de pontos com altíssima concentração de veículos, tais como, garagens, terminais rodoviários ou grandes congestionamentos.

Após análise qualitativa para escolha de qual quadrante contém os parâmetros mais adequados para o experimento, prosseguimos com uma análise quantitativa. Durante este passo, devemos selecionar eventos que possam ser comparáveis (dia da semana e horário) e analisar as variações apresentadas em seus resultados (evento alvo e evento referência). Dois questionamentos surgem neste momento: quais são os parâmetros que me oferecem a maior variação na quantidade de clusters? E quais parâmetros me oferecem a maior variação de observações agrupadas? Respondendo a estas duas perguntas teremos condições de calibrar os algoritmos de agrupamento para apresentar as variações mais contundentes entre os eventos alvo e referência.

No que se refere ao DBSCAN, os próprios desenvolvedores do algoritmo (ESTER et al., 1996) propõem um método heurístico de definição de ε e *MinPts,* outros algoritmos derivados também endereçam a questão da definição das variáveis de entrada, GDBSCAN (SANDER et al., 1998), DBCLASD (XU et al., 1998), OPTICS (ANKERST et al., 1999), VDBSCAN (LIU et al., 2007) e ISDBSCAN (CASSISI et al., 2012). Com relação aos algoritmos baseados em *grid,* os desenvolvedores do STING (WANG et al., 1997) também abordaram o problema, ZHAO et al. (2015) autores do GG fizeram uma análise da quantidade de *clusters* conforme variação no tamanho das malhas e na quantidade do *seed*, mas não trataram da questão da quantidade mínima de pontos. Porém, apesar da vasta quantidade de publicações que tratam do tema, optamos pelo desenvolvimento de uma forma particular para de determinação dos parâmetros.

A resposta a primeira pergunta proposta pelo método sugere a criação de uma nova matriz de sensibilidade com a diferença entre os valores de formação de *clusters* dos eventos escolhidos. Esta matriz também será apresentada em forma de mapa de calor e nos indicará quais parâmetros apresentam as maiores variações na formação de agrupamentos, e como critério de escolha selecionaremos os valores acima do 80-percentil. A Figura III.7 apresenta os valores selecionados para o 2º quadrante da matriz utilizada nos exemplos anteriores (destacados em amarelo).

C_A = *Quantidade de clusters formados no evento alvo*

C_R = *Quantidade de clusters formados no evento referência*

Δ_C = *Variação na quantidade de clusters formados entre alvo e referência*

$$\Delta_C = C_A - C_R$$

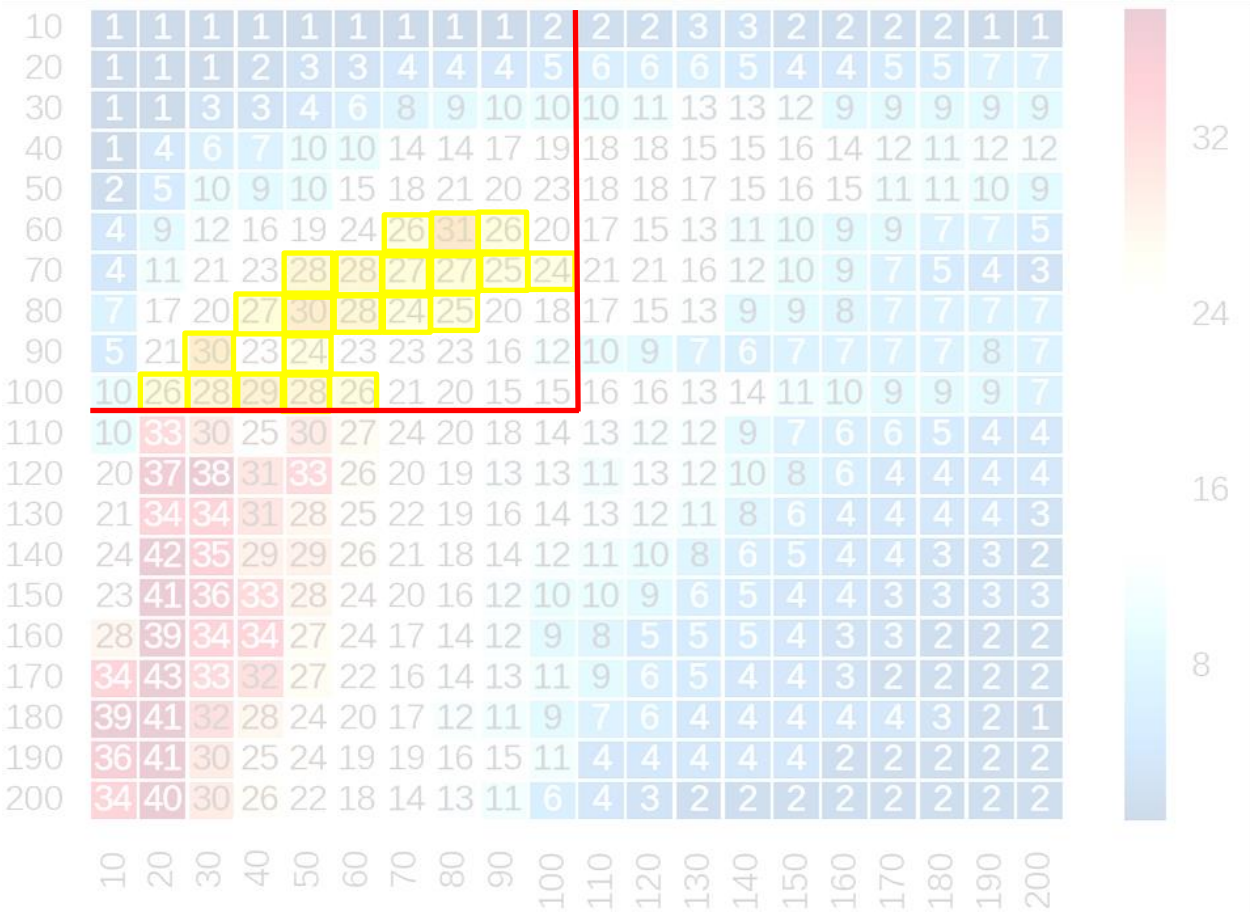


Figura III.7 Análise de Sensibilidade – Cálculo de Δ_C

A resposta a segunda pergunta é a criação de uma outra matriz de sensibilidade, desta vez considerando não mais a variação na quantidade de *clusters* formados, mas a variação na relação entre observações que foram agrupadas e quantidade total de observações. A Figura III.8 apresenta resultados obtidos no exemplo proposto (destacados em verde).

O_A = *Quantidade total de observações do evento alvo*

O_R = *Quantidade total de observações do evento referência*

O_{CA} = *Quantidade de observações agrupadas do evento alvo*

O_{CA} = *Quantidade de observações agrupadas do evento referência*

R_A = *Percentual de observações agrupadas no evento alvo*

R_R = *Percentual de observações agrupadas no evento referência*

Δ_R = *Variação de percentual de observações agrupados entre alvo e referência*

$$R_A = O_{CA} \,/\, O_A$$

$$R_R = O_{CR} \,/\, O_R$$

$$\Delta_R = R_A - R_R$$

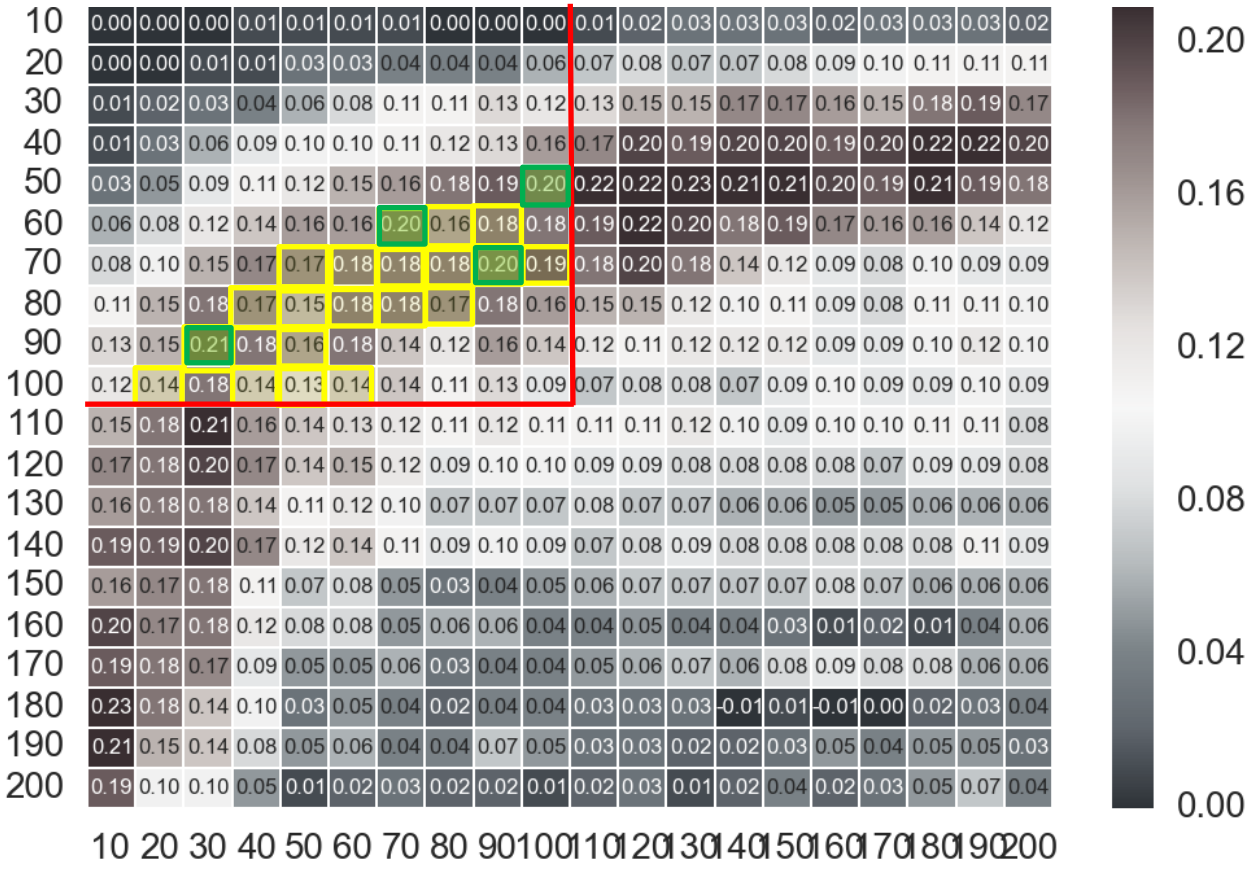


Figura III.8 Análise de Sensibilidade – Cálculo de Δ_R

A escolha pelo 80-percentil foi feita com base na quantidade da amostra disponível (cem pares de parâmetros por quadrante) e na quantidade de vezes que entendemos como razoável rodar o algoritmo de agrupamento (até quatro vezes). Com isto, economizamos processamento e tempo de análise, com intuito de tornarmos o método eficiente capaz de responder com rapidez as demandas que lhe são impostas. Ao final desta etapa temos condições de indicar parâmetros com razoáveis de condições de alimentar nossos algoritmos com objetivo de gerar agrupamentos que possam servir como base para uma análise crítica sobre a condição da mobilidade urbana da cidade. No exemplo apresentado, foram selecionados os parâmetros 30x90, 70x60, 90x70 e 100x50.

Capítulo IV - Avaliação Experimental

Esse capítulo tem como objetivo apresentar os resultados obtidos durante os experimentos realizados utilizando método proposto no capítulo anterior. Serão utilizados estudos de casos reais para verificar a eficácia do método e se as respostas ao problema alvo foram devidamente endereçadas. Conforme já antecipado, com intuito de buscarmos uma estabilidade da malha rodoviária e das linhas regulares que atendem a RMRJ, de modo, que as interferências às análises e impactos aos resultados sejam mitigados, foram escolhidas datas anteriores às mudanças nas linhas de ônibus propostas pela PMRJ a partir do segundo semestre de 2015. Foram pesquisadas datas referentes a diferentes naturezas eventos que afetaram o trânsito usual da cidade (greve de ônibus, jogo de copa do mundo e acidentes graves de trânsito).

IV.1 Caso 1: Estreia do Brasil na Copa do Mundo

O primeiro estudo de caso apresentado é a estreia da seleção brasileira de futebol na Copa do Mundo de 2014. Apesar do jogo não ter ocorrido na cidade do Rio de Janeiro, os impactos de um jogo da seleção do Brasil em Copas do Mundo no cotidiano das pessoas são sentidos em praticamente todo país. O jogo ocorreu em 12/06/2014 (quinta-feira) as 17:00 e o *dataset* coletado entre as 18:00 e 18:50, durante a realização da partida e este será considerado nosso evento referência. Para o evento alvo utilizaremos os dados obtidos em 05/06/2014 (quinta-feira anterior ao jogo) como parâmetro de um dia de trânsito sem casos extraordinários que impactem na mobilidade urbana do Rio de Janeiro.

O resultado esperado para este experimento, é que devido ao interesse quase que absoluto da população por assistir ao jogo, os congestionamentos usualmente observados na cidade durante horário de *rush* da tarde não ocorram, e que os ônibus que transitam a cidade estejam em grande parte parados em garagens, pontos finais e terminais rodoviários, sendo um excelente referencial para medirmos os estrangulamentos que surgirão na análise do evento alvo.

IV.1.1 Caso 1: DBSCAN

Uma vez que o período de análise foi definido, os dados coletados e pré-processados, devemos prosseguir com a etapa de análise de sensibilidade dos parâmetros de cada um dos algoritmos. As Figura IV.1 e Figura IV.2 apresentam as análises qualitativas do algoritmo DBSCAN para os eventos alvo e referência (conforme explicado na Seção III.7).

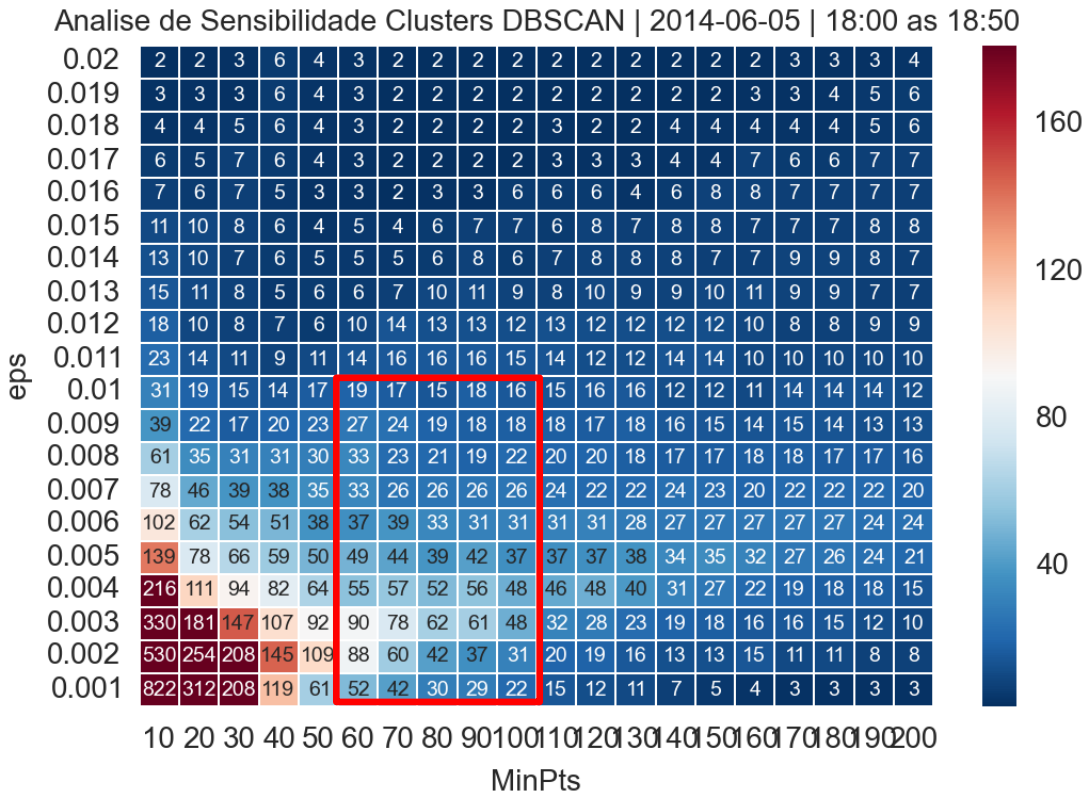


Figura IV.1 Análise de sensibilidade DBSCAN – 05/06/2014 (quinta-feira)

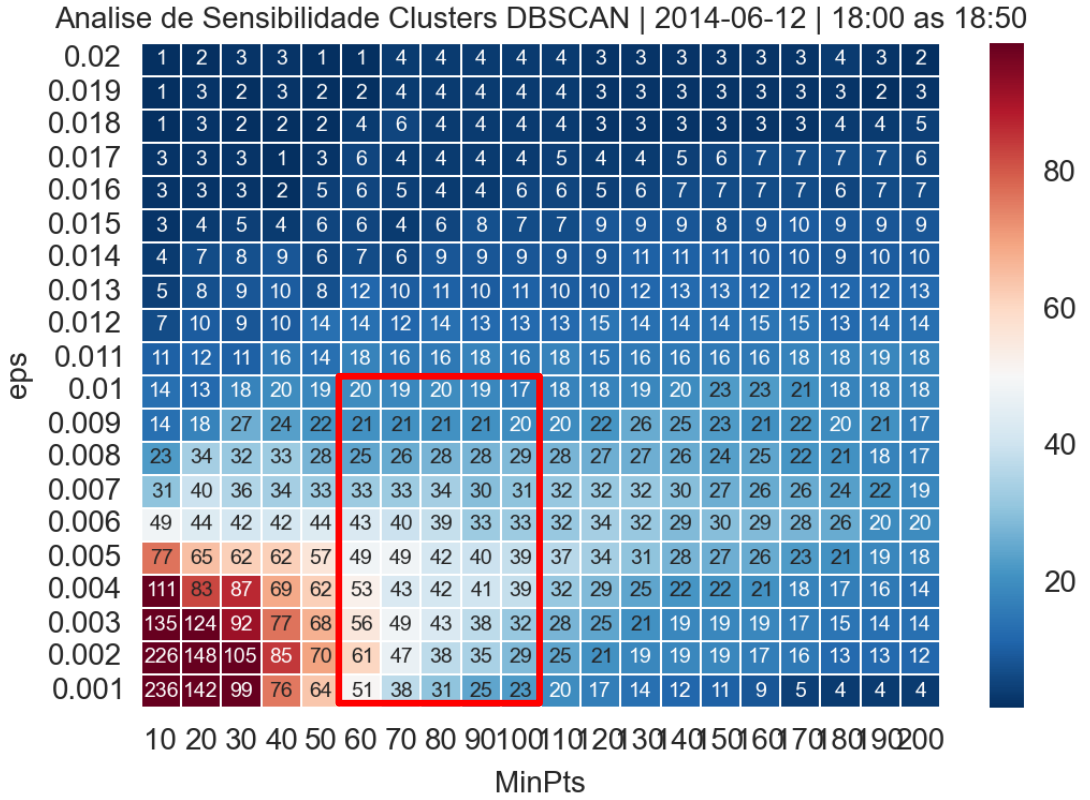


Figura IV.2 Análise de sensibilidade DBSCAN – 12/06/2014 (quinta-feira)

Conforme processo estabelecido no Capítulo III, consideramos o terceiro quadrante o mais adequado para realização da análise quantitativa, por estar em linha com as características do problema apresentado (identificação de estrangulamentos por toda cidade em horário de pico). Devido a baixa capacidade restritiva das colunas com quantidade mínima de pontos inferior a cinquenta, optamos por utilizar as colunas com *MinPts* entre 60 e 100, reduzindo a quantidade de *clusters*, e buscando identificar locais onde realmente houvessem uma concentração acentuada de observações. As Figura IV.3 e Figura IV.4 apresentam a análise quantitativa do estudo de caso para o algoritmo DBSCAN.

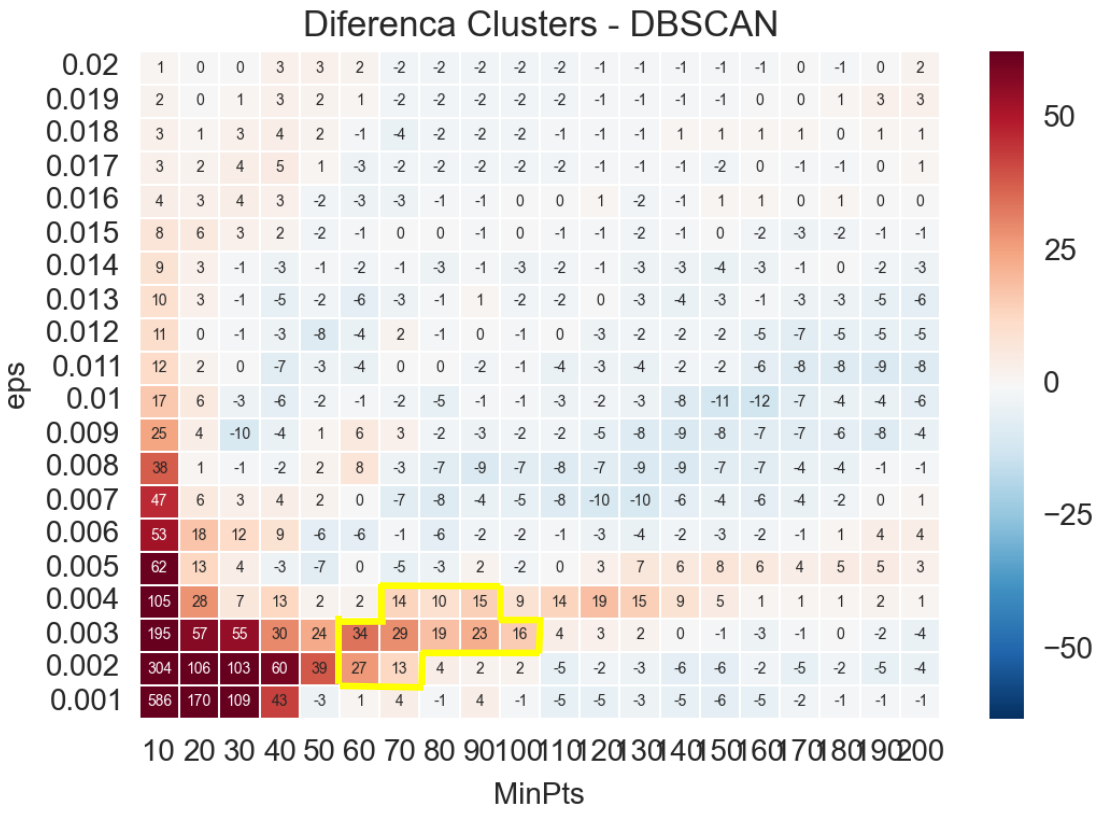


Figura IV.3 Cálculo de Δ_C para DBSCAN (05/06/14 e 12/06/14)

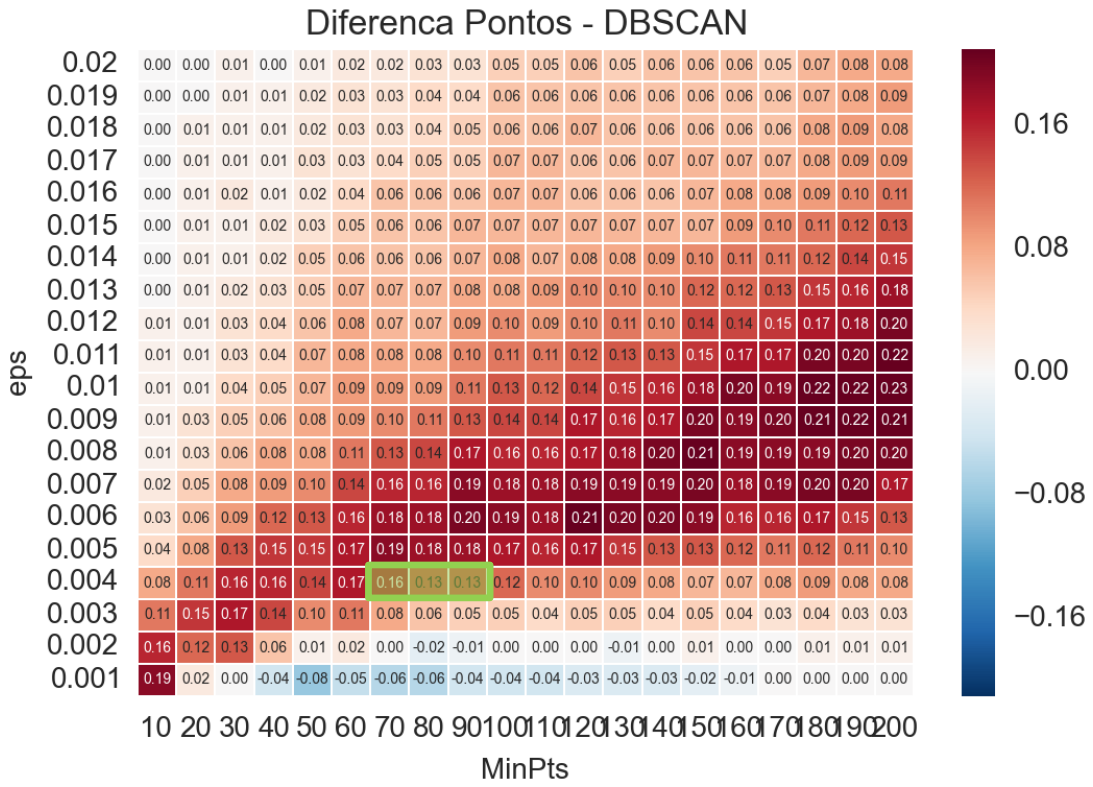


Figura IV.4 Cálculo de Δ_R para DBSCAN (05/06/14 e 12/06/14)

Aplicando-se as regras para análise quantitativa estabelecidas no capítulo anterior, obtemos uma primeira restrição na matriz Δ_C destacada em amarelo na Figura IV.3, e o resultado final apresentado na matriz Δ_R em verde conforme Figura IV.4. Deste modo, os parâmetros sugeridos para a fase de agrupamento no DBSCAN são 70x0.004, 80x0.004 e 90x0.004. Por conta dos resultados obtidos serem adjacentes e não haver diferença considerável na formação de clusters entre eles, apresentaremos apenas o agrupamento utilizando o parâmetro mais restritivo, 90x0.004 (Figura IV.6 e Figura IV.5).

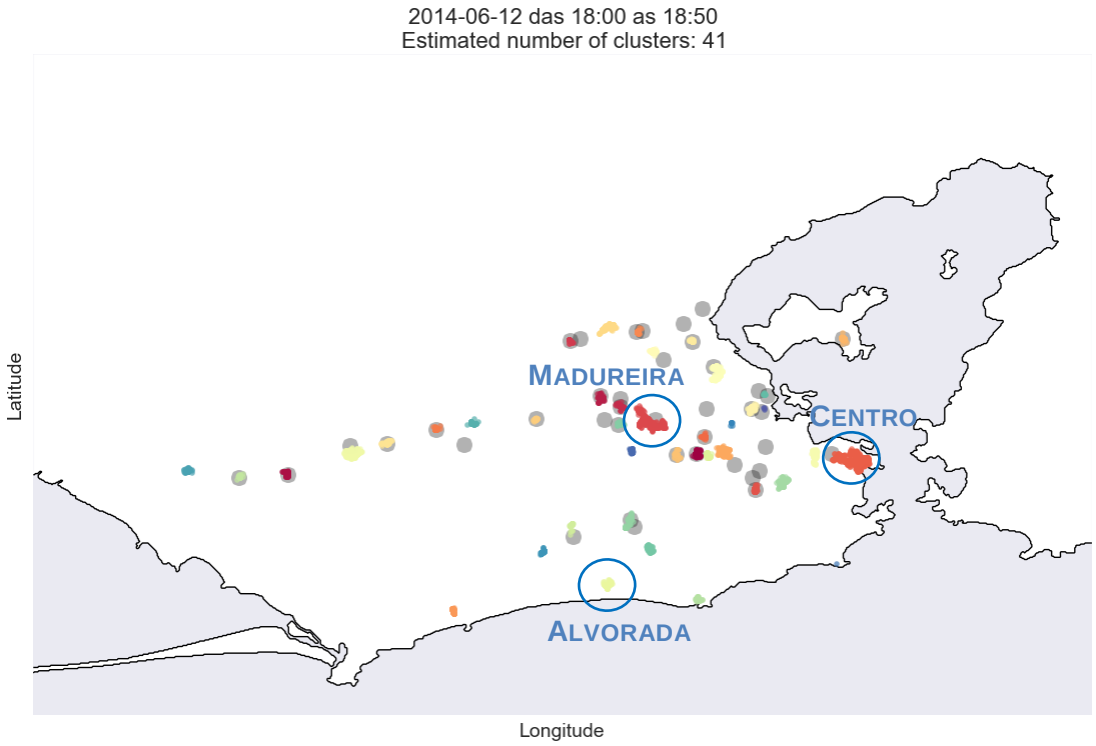


Figura IV.5 DBSCAN *MinPts* = 90 e ε = 0.004 – 12/06/14

Podemos observar que os *clusters* no dia da estreia do Brasil são formados majoritariamente por garagens de empresas de ônibus (zonas cinza), terminais rodoviários e locais utilizados como estacionamento provisório para os ônibus (destacados pelos círculos azuis). Isso demonstra uma baixíssima utilização do sistema de ônibus durante a realização da partida, o que já seria esperado. Portanto, este evento é apropriado para servir de referência na medição de estrangulamentos quando comparados a eventos que queremos analisar.

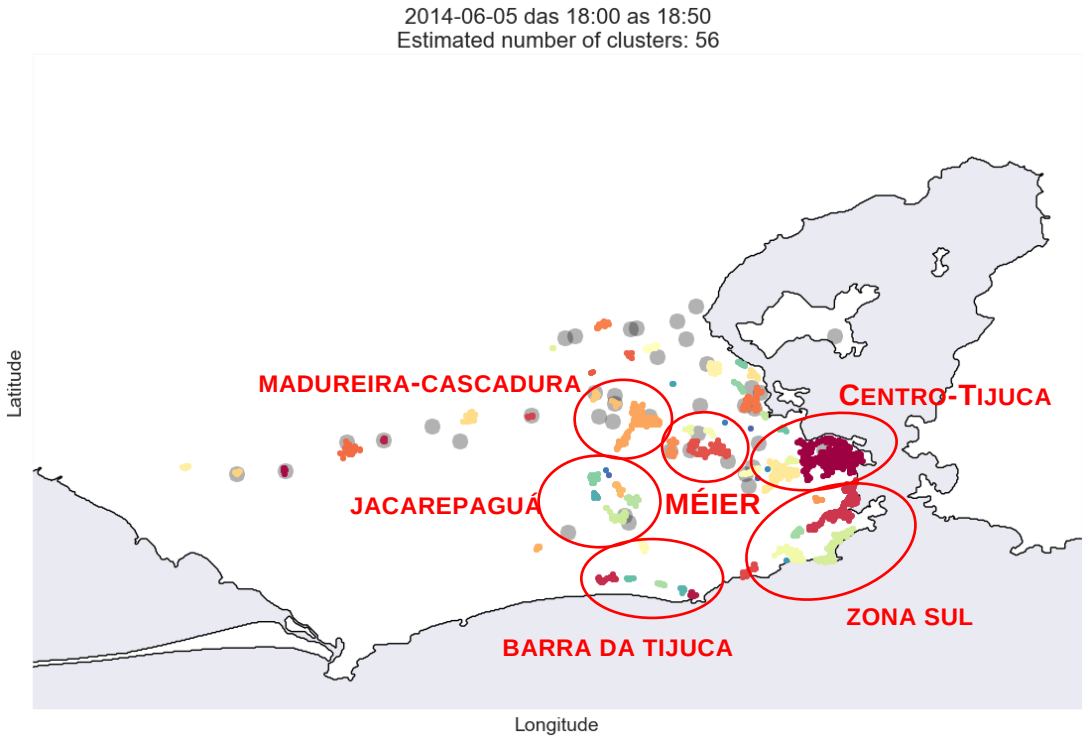


Figura IV.6 DBSCAN *MinPts* = 90 e ε = 0.004 – 05/06/14

Analisando os resultados obtidos após rodar DBSCAN para o evento alvo, identificamos uma quantidade bem maior de agrupamentos (56 contra 41), bem mais densos também, o que denota uma maior quantidade de ônibus operando em linhas. Classificamos os principais *clusters* por área (Figura IV.6) e vamos comparar com os resultados do GG e selecionaremos uma das áreas para uma análise mais detalhada.

IV.1.2 Caso 1: *Grid Growing*

Nesta seção prosseguiremos com raciocínio semelhante ao aplicado para DBSCAN, porém utilizando *Grid Growing*. As Figura IV.7 e Figura IV.8 apresentam a área da matriz de sensibilidade definida como ponto de partida para análise.

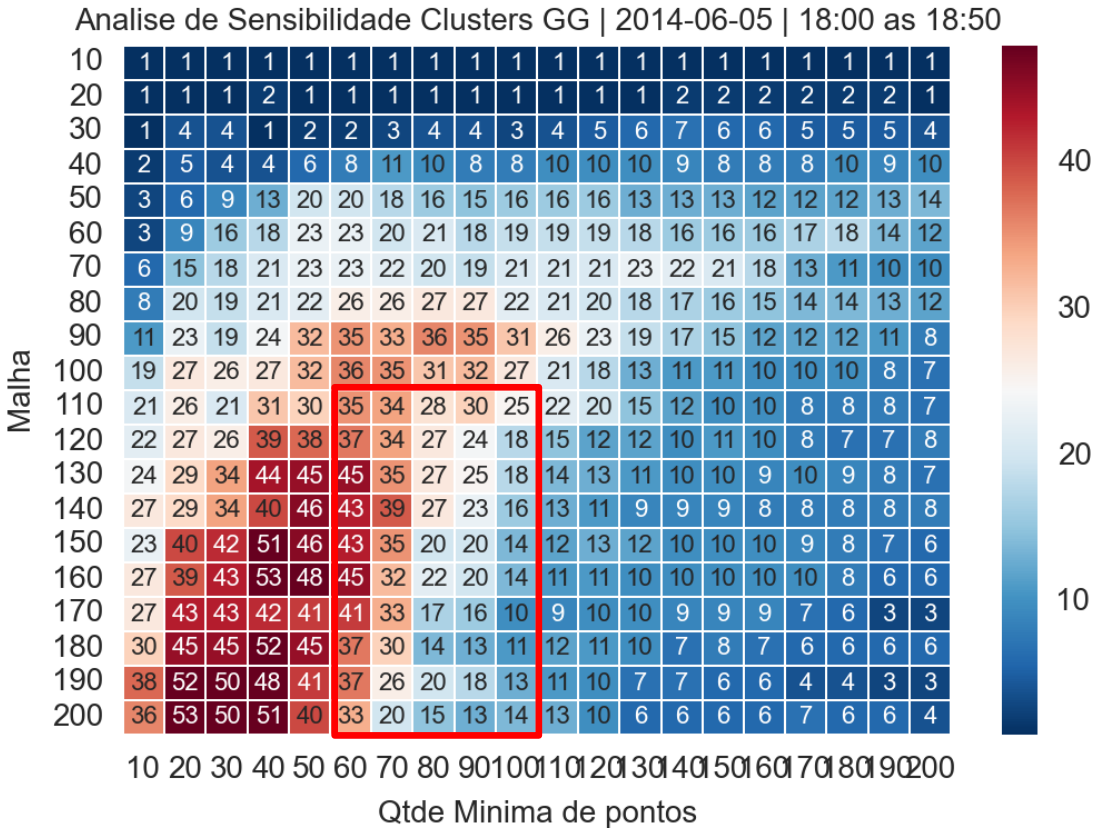


Figura IV.7 Análise de sensibilidade GG – 05/06/2014 (quinta-feira)

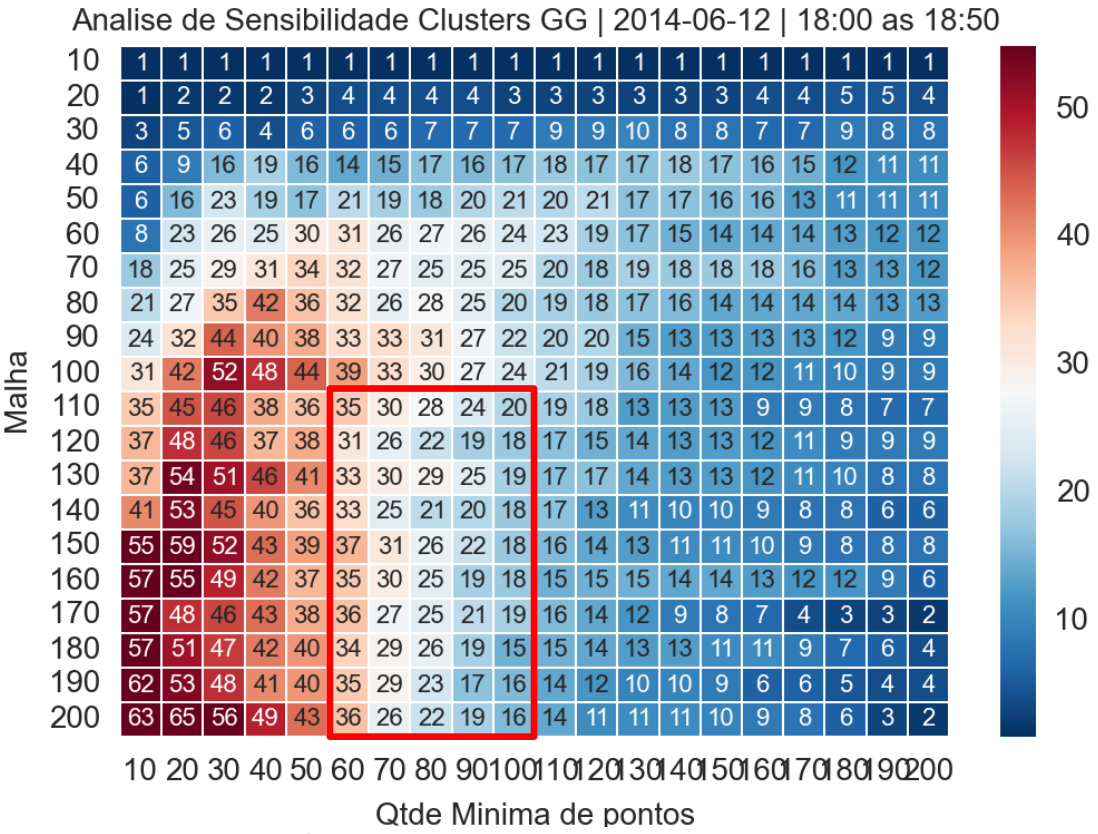


Figura IV.8 Análise de sensibilidade GG – 12/06/2014 (quinta-feira)

As premissas assumidas anteriormente se mantêm, pois as considerações feitas durante a análise de sensibilidade do DBSCAN são igualmente válidas para GG. O terceiro quadrante é o mais adequado para pesquisarmos parâmetros para o agrupamento, e a utilização apenas de metade do quadrante tem como objetivo sermos mais criteriosos na formação de *clusters*. As Figura IV.9 e Figura IV.10 demonstram a análise de sensibilidade do GG com a parte quantitativa.

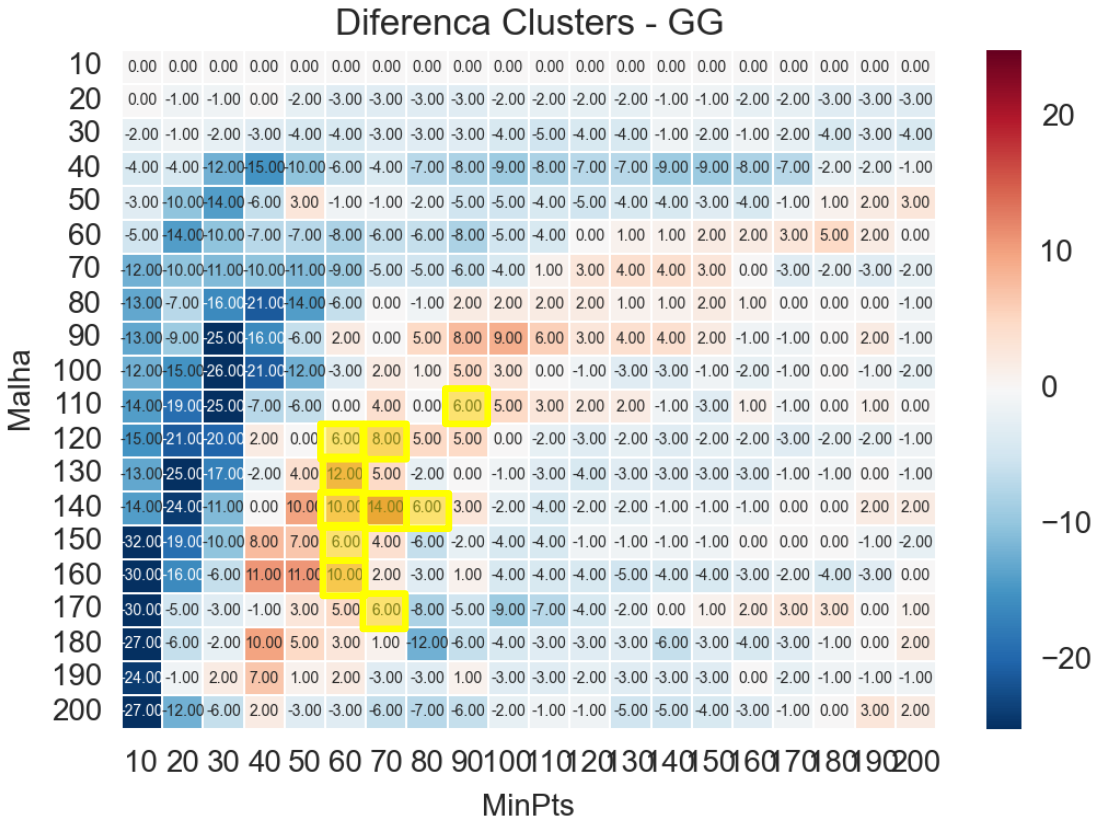


Figura IV.9 Cálculo de Δ_C para GG (05/06/14 e 12/06/14)

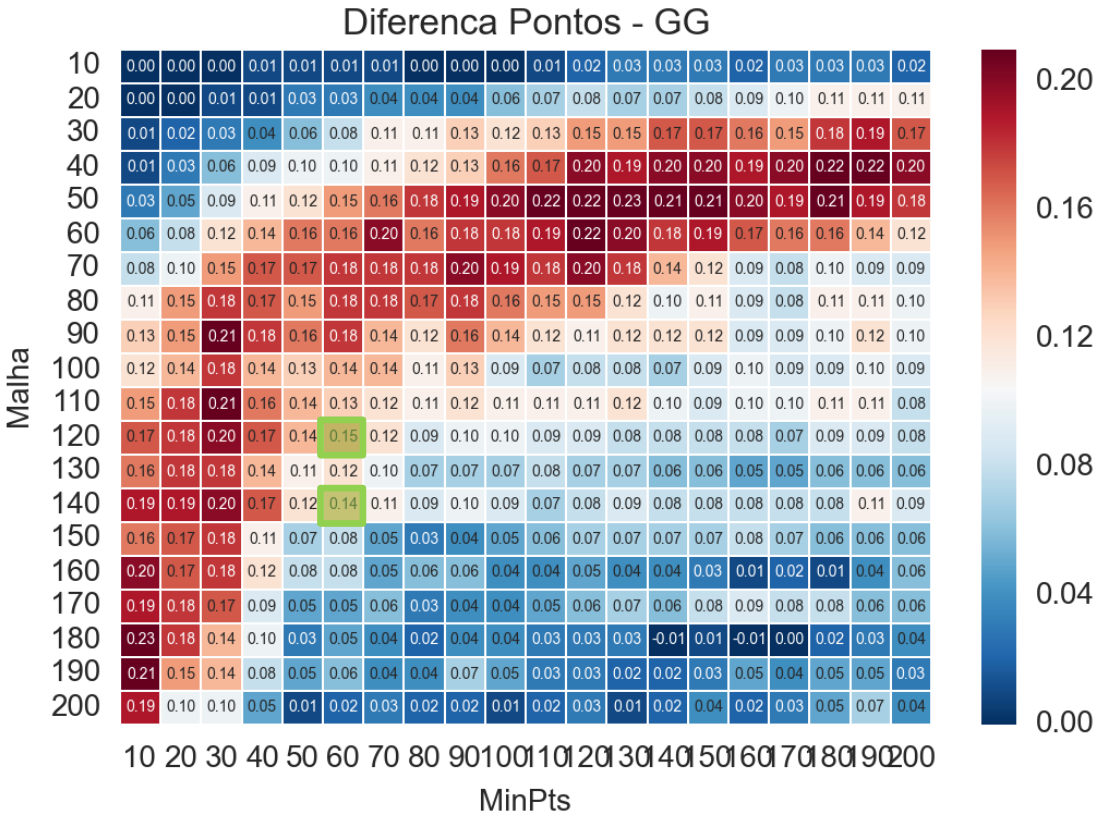


Figura IV.10 Cálculo de Δ_R para GG (05/06/14 e 12/06/14)

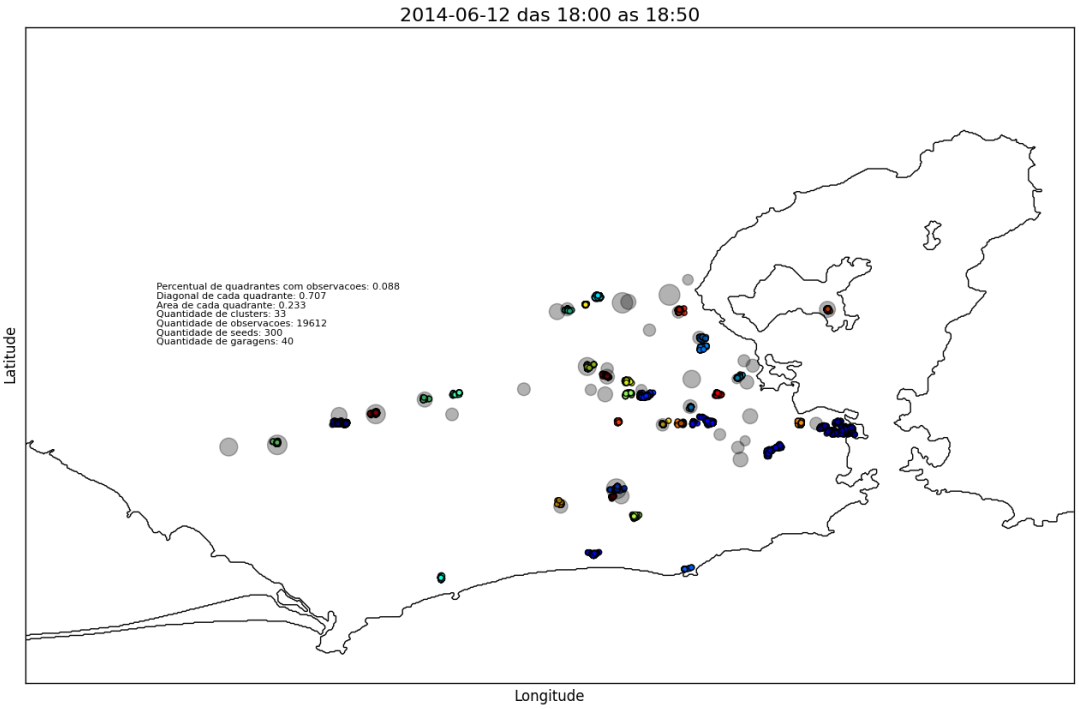


Figura IV.11 GG *MinPts* = 60 e Malha = 140 – 12/06/14

Os resultados do agrupamento utilizando GG *dataset* de 12/06/2014 foram muito próximos aos obtidos via DBSCAN, com pequena variação na quantidade de clusters formados, ou seja, também podemos utilizar este evento como referência em comparação ao evento alvo escolhido (05/06/2014).

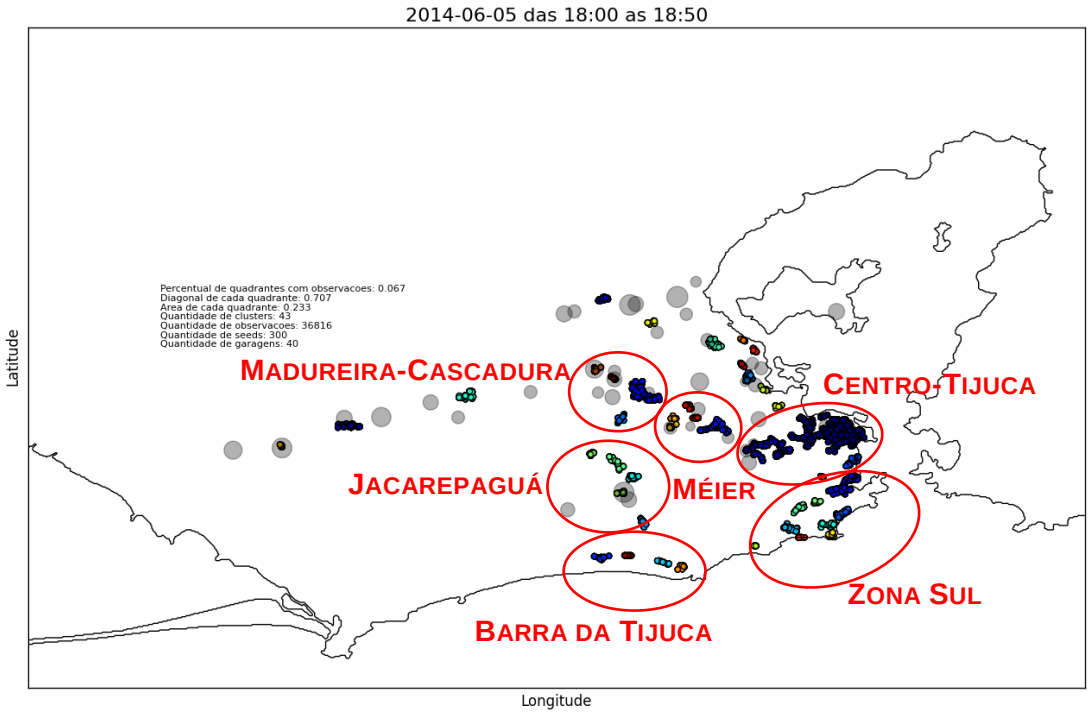


Figura IV.12 GG *MinPts* = 60 e *Malha* = 140 – 05/06/14

Comparando os resultados obtidos entre os dois algoritmos, novamente observamos bastante similaridade, porém com uma menor quantidade de clusters obtidos pelo GG, muito por conta da fusão de clusters menores identificados no DBSCAN como, por exemplo, na área do Centro e Tijuca. Conseguimos identificar na Figura IV.12 as mesmas áreas de concentração observadas na Figura IV.6 (Centro-Tijuca, Zona Sul, Barra da Tijuca, Madureira-Cascadura, Méier e Jacarepaguá). A Figura IV.13 apresenta em uma escala maior os *clusters* formados na Zona Sul

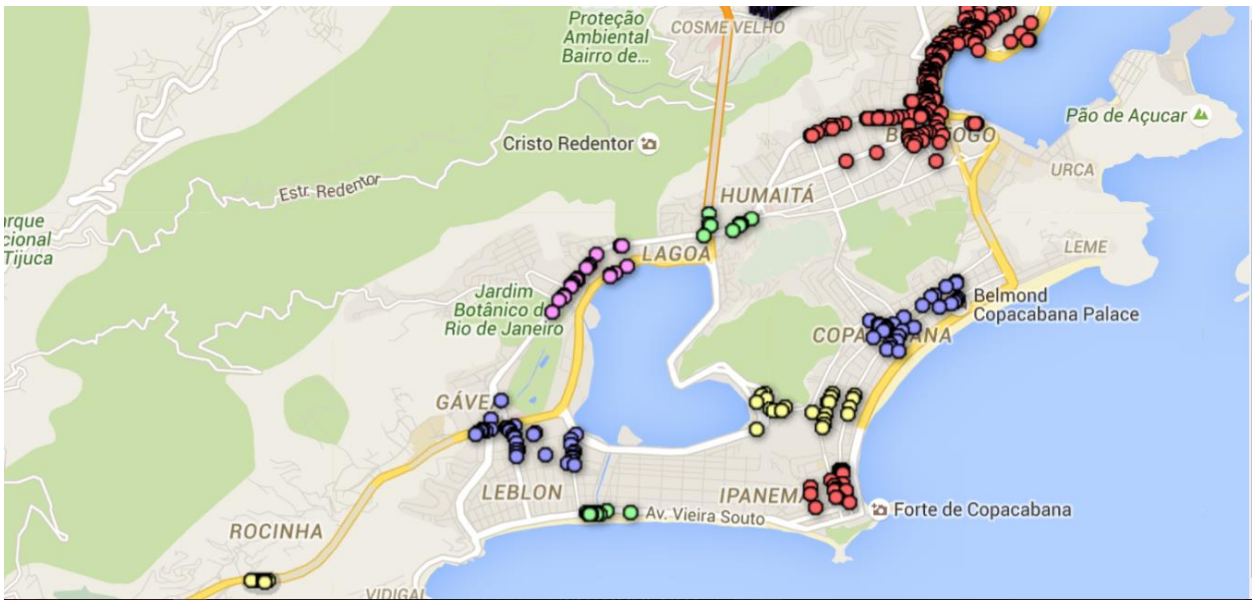


Figura IV.13 GG *MinPts* = 60 e *Malha* = 140 – 05/06/14 (Zona Sul)

Tradicionalmente a Zona Sul do Rio de Janeiro é uma área bastante populosa e com intenso fluxo de veículos ao longo de todo dia, muito devido a sua natureza turística e sua proximidade com Centro. Alguns dos agrupamentos formados referem-se a congestionamentos bem conhecidos pela população carioca.

- Rua Jardim Botânico: Das 17hs as 21hs uma das faixas da pista sentido Humaitá se torna reversível[23], causando longos engarrafamentos;
- Humaitá (acesso ao Túnel Rebouças): Túnel Rebouças é uma das principais vias de acesso entre a Zona Sul e Zona Norte, sendo bastante concorrido durante os horários de pico;
- Praia de Botafogo/Rua São Clemente: Devido a grande quantidade de colégios principalmente situados nas proximidades das ruas São Clemente e Voluntários da Pátria são formados grandes congestionamentos com os reflexos até a Praia de Botafogo;
- Gávea (acesso ao Túnel Acústico): Este *cluster* é formado pelo acumulo de veículos em três diferentes pontos, Rua Mário Ribeiro (próximo a PUC), Rua Bartolomeu Mitre (próximo ao Hospital Miguel Couto) e Afrânio de Melo Franco (próximo ao Clube de Regatas do Flamengo);
- Jardim de Alah: A orla do Leblon é muito utilizada principalmente por aqueles que querem subir a Avenida Niemayer sentido São Conrado;
- Copacabana: Também muito utilizada por quem deseja acessar Leblon, Ipanema, São Conrado ou bairros da Zona Oeste. Durante horário de pico apresentou três clusters (Forte de Copacabana, Corte do Cantagalo e Copacabana Palace), dependendo do valor de ε (DBSCAN) ou do tamanho da malha (GG) poderiam ser agrupados como um único *cluster*.
- São Conrado (acesso ao Túnel Zuzu Angel): Pode ser considerado uma continuação do *cluster* da Gávea, porém devido a indisponibilidade do serviço GPS dentro dos túneis, eles são agrupados separadamente pelos algoritmos.

IV.2 Caso 2: Greve do Rodoviários do Rio de Janeiro

Outro estudo de caso proposto, refere-se a análise do trânsito de um dia de greve dos rodoviários da cidade do Rio de Janeiro ocorrida em 13/05/14 (terça-feira). Essa paralisação teve adesão de boa parte dos profissionais, e reduziu drasticamente a quantidade de ônibus em circulação[24]. Compararemos os *datasets* referentes aos dias da greve com os do dia 20/05/14, um dia supostamente sem excepcionalidades no trânsito, em que a frota de veículos

[23] http://www.rio.rj.gov.br/web/smtr/exibeconteudo?id=2801886
[24] Portal EBC – Agência Brasil publicado em 13/05/2014 as 08:00

estava plenamente disponível a população e assim como no dia da paralisação é uma terça-feira.

Uma avaliação preliminar ao processamento dos dados nos aponta no sentido de que provavelmente devido a menor quantidade de ônibus nas ruas, menos agrupamentos serão formados, passando uma impressão de um trânsito menos congestionado, porém esta suposição não leva em consideração o aumento da quantidade de veículos particulares com objetivo de suprir esta deficiência temporária do transporte público. Também seria razoável esperarmos por uma maior concentração de veículos nas garagens durante horários de *rush*, caracterizando a adesão dos rodoviários ao movimento grevista. A seguir faremos as análises para os dias acima citados entre os horários das 8:00 e 8:50, início do horário de pico na parte da manhã (LOPES et al. 2015). A partir deste estudo de caso, buscando maior objetividade, apresentaremos de forma contínua as figuras com as análises de sensibilidade e mapas com as divisões dos *clusters* sem comentários entre eles e faremos as considerações pertinentes a cada agrupamento ao fim de cada seção.

IV.2.1 Caso 2: DBSCAN

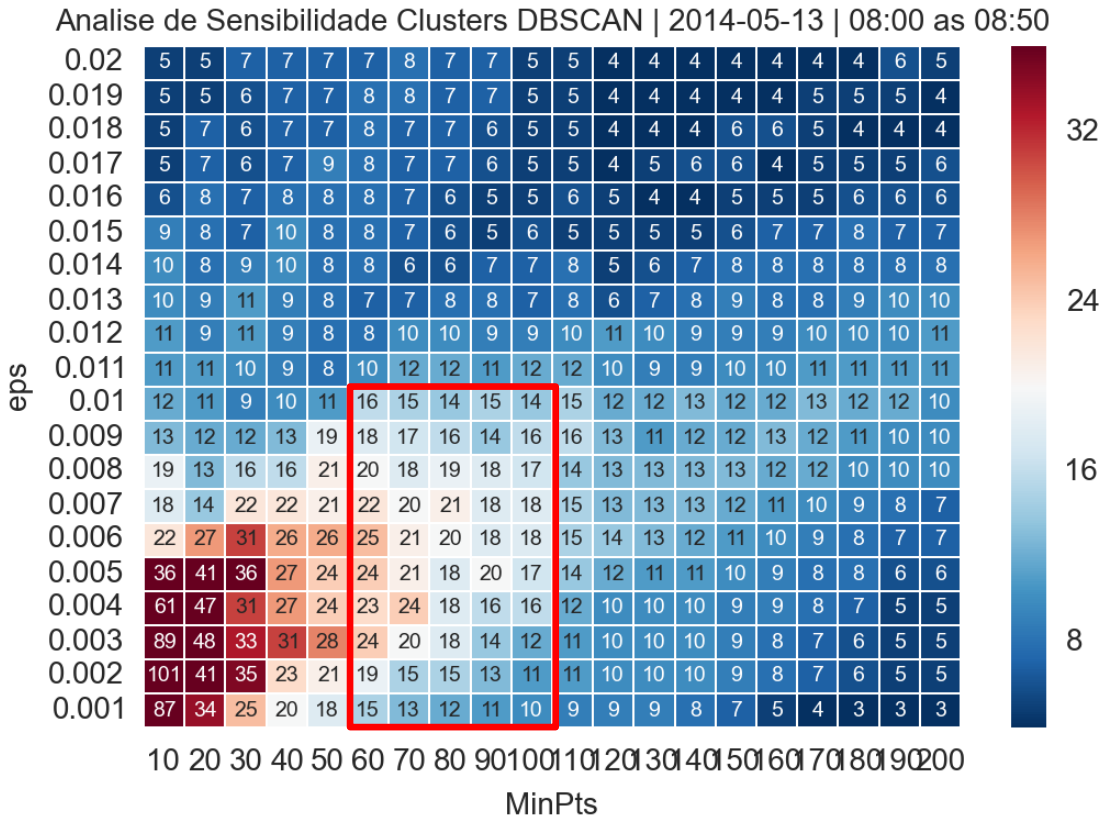


Figura IV.14 Análise de sensibilidade DBSCAN – 13/05/2014 (terça-feira)

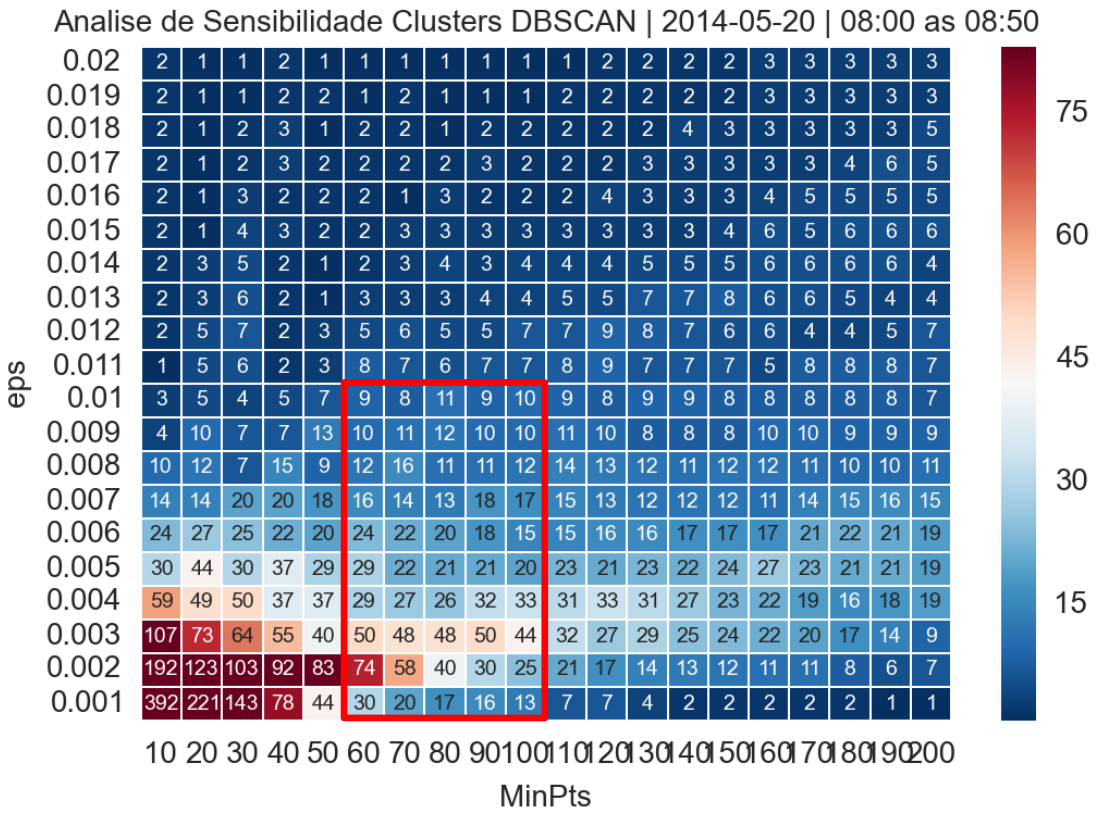


Figura IV.15 Análise de sensibilidade DBSCAN – 20/05/2014 (terça-feira)

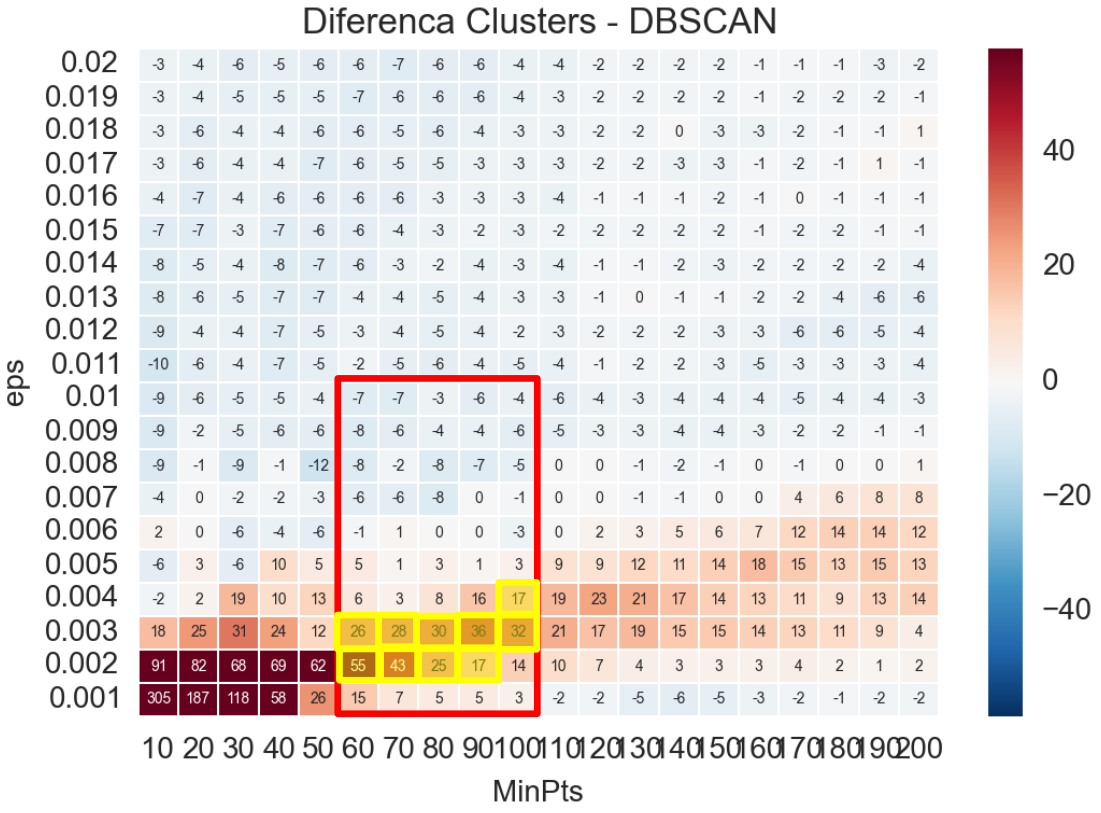


Figura IV.16 Cálculo de Δ_C para DBSCAN (13/05/14 e 20/05/14)

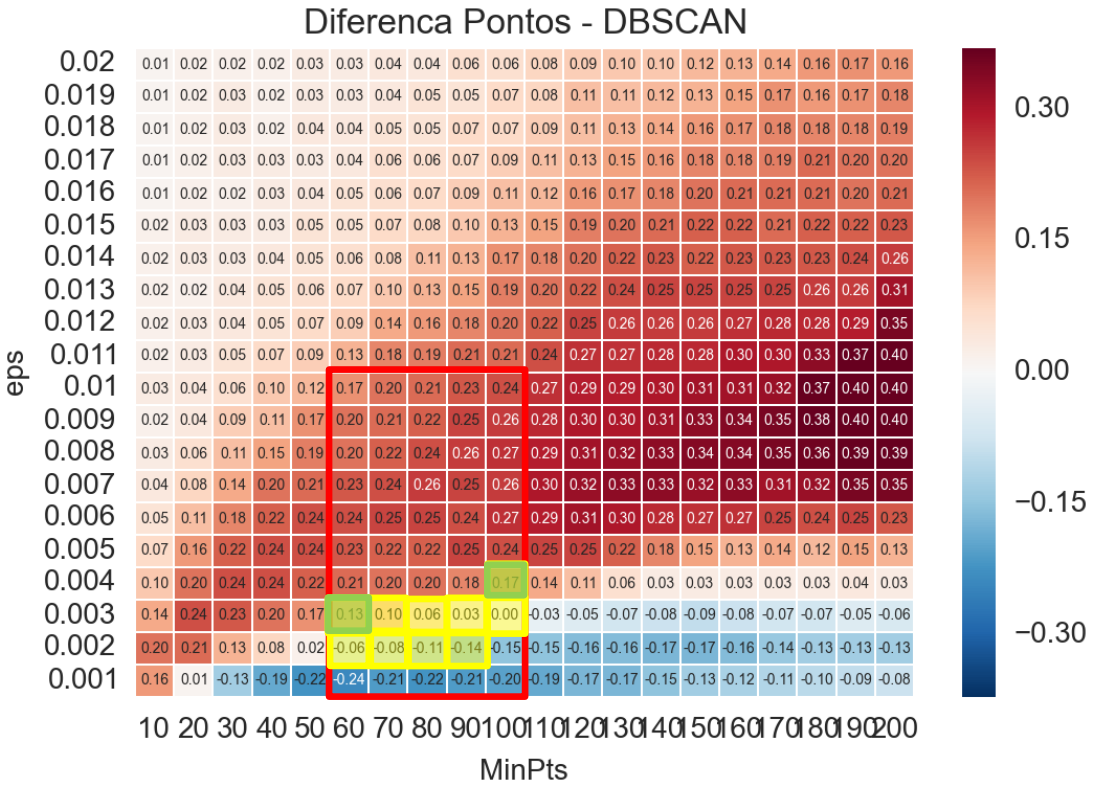


Figura IV.17 Cálculo de Δ_R para DBSCAN (13/05/14 e 20/05/14)

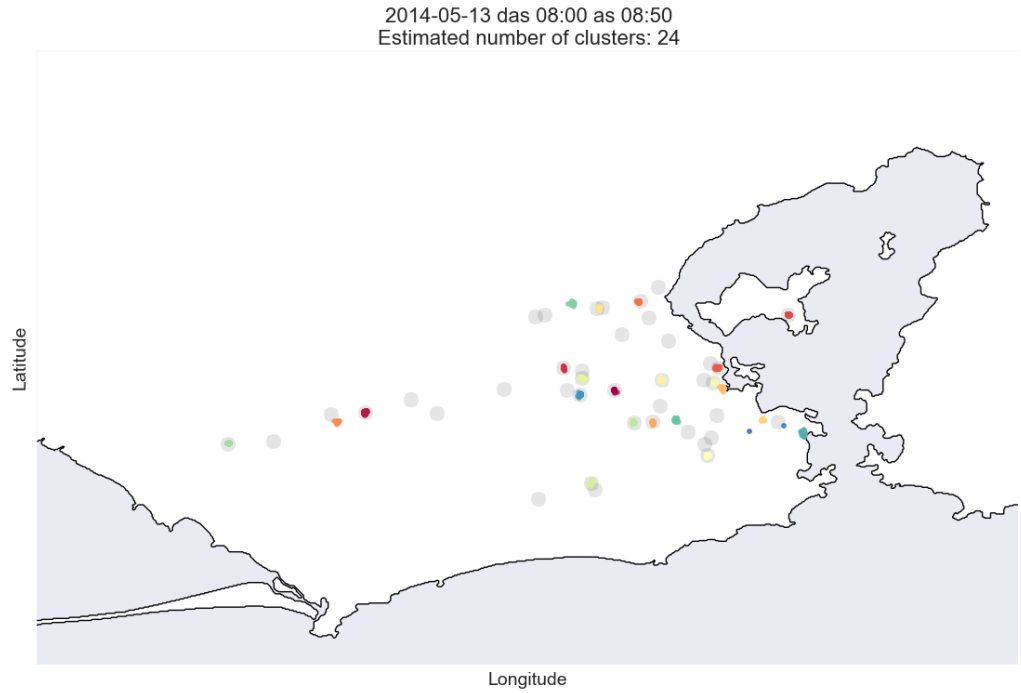


Figura IV.18 DBSCAN *MinPts* = 60 e ε = 0.003 – 13/05/14

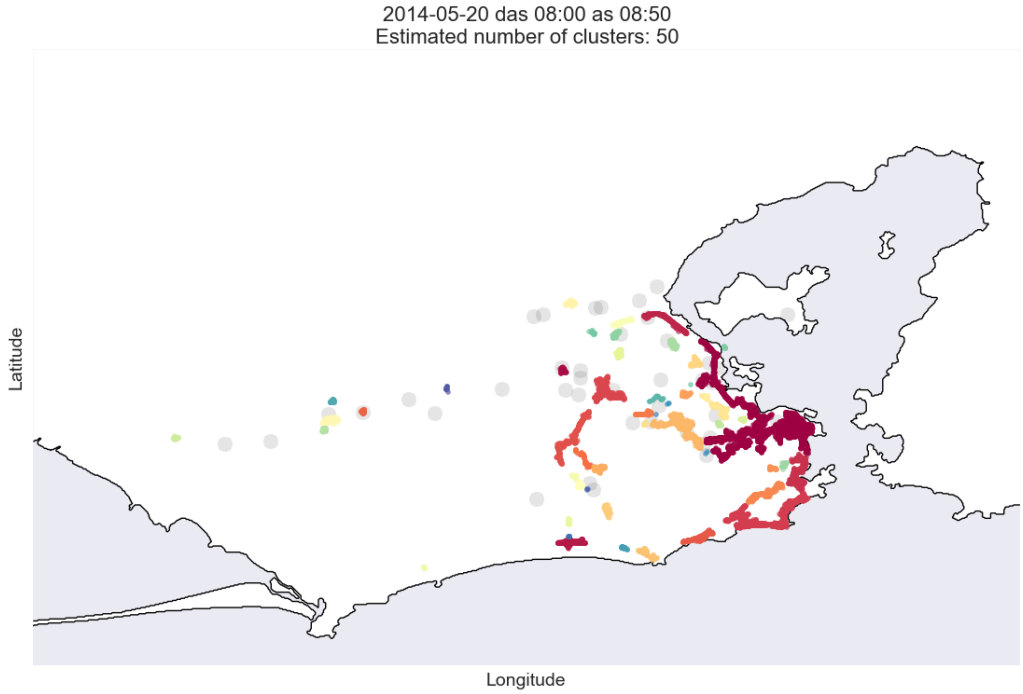


Figura IV.19 DBSCAN *MinPts* = 60 e ε = 0.003 – 20/05/14

A análise de sensibilidade apontou dois pares de parâmetros de entrada para DBSCAN 0,004x100 e 0,003x60. Resolvemos utilizar o par 0,003x60 neste estudo de caso com objetivo de não viciar a escolha dos parâmetros (no estudo de caso anterior utilizamos 0,004x90) e proporcionar um rodízio que nos possibilite avaliar a ferramenta sob diferentes circunstâncias.

O resultado do agrupamento do dia 13/05/2014 confirma as expectativas de um cenário em que o sistema de transporte rodoviário encontra-se demasiadamente reduzido, pois são raros os *clusters* fora de garagens, sendo adequado para utilizado como evento referência. Já os dados do dia 20/05/2014, demonstram um cenário oposto, com muitos *clusters*, alguns bem extensos (consequência de um *MinPts* mais baixo). Prosseguiremos com o mesmo processo utilizando GG e análise dos estrangulamentos de uma determinada área.

IV.2.2 Caso 2: *Grid Growing*

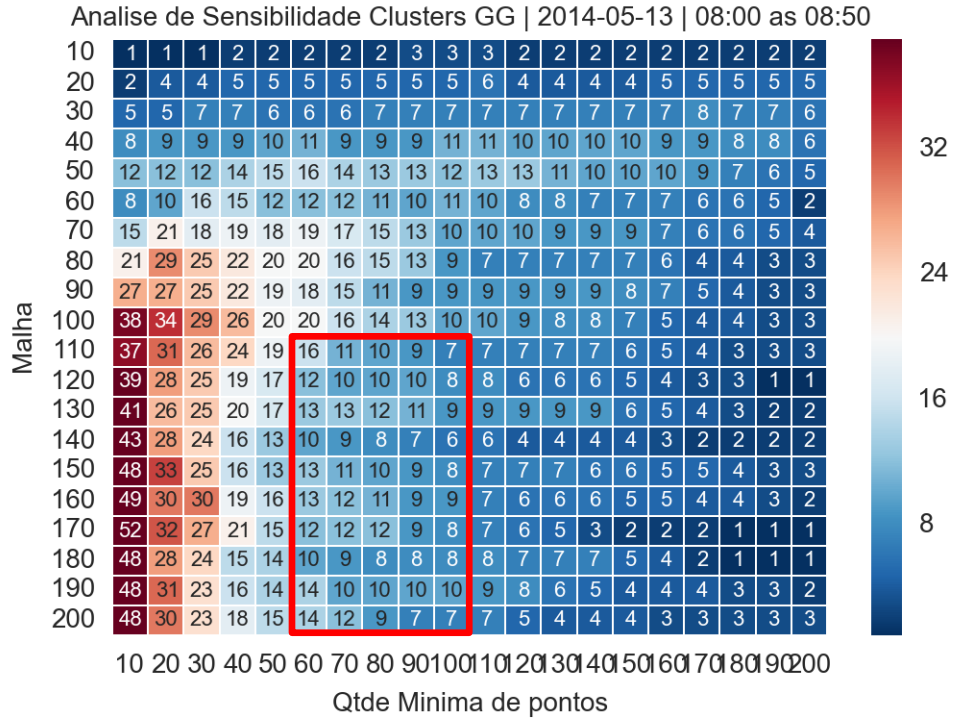


Figura IV.20 Análise de sensibilidade GG – 13/05/2014 (terça-feira)

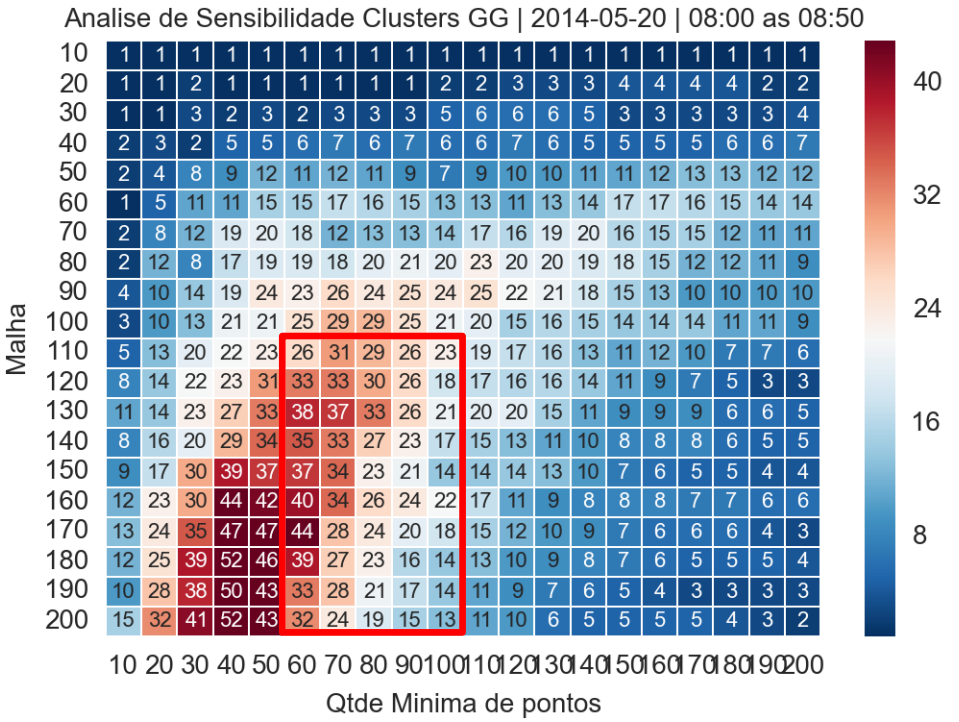


Figura IV.21 Análise de sensibilidade GG – 20/05/2014 (terça-feira)

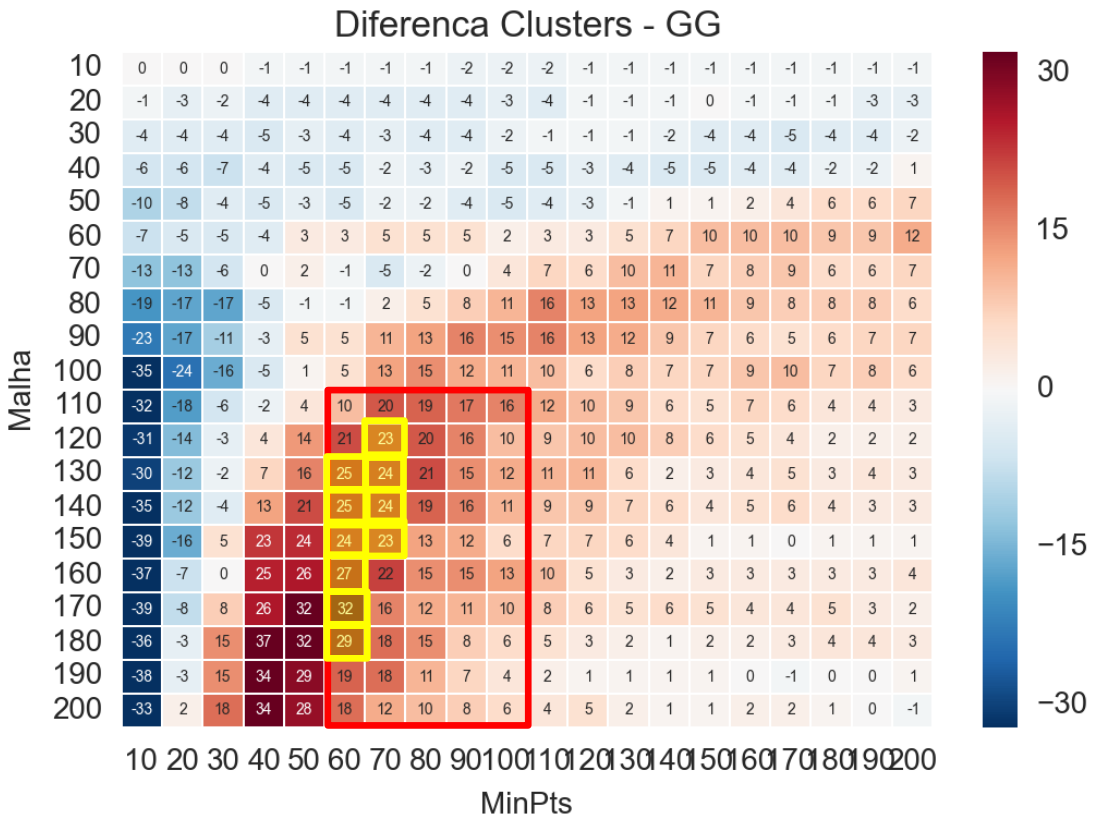


Figura IV.22 Cálculo de Δ_C para GG (13/05/14 e 20/05/14)

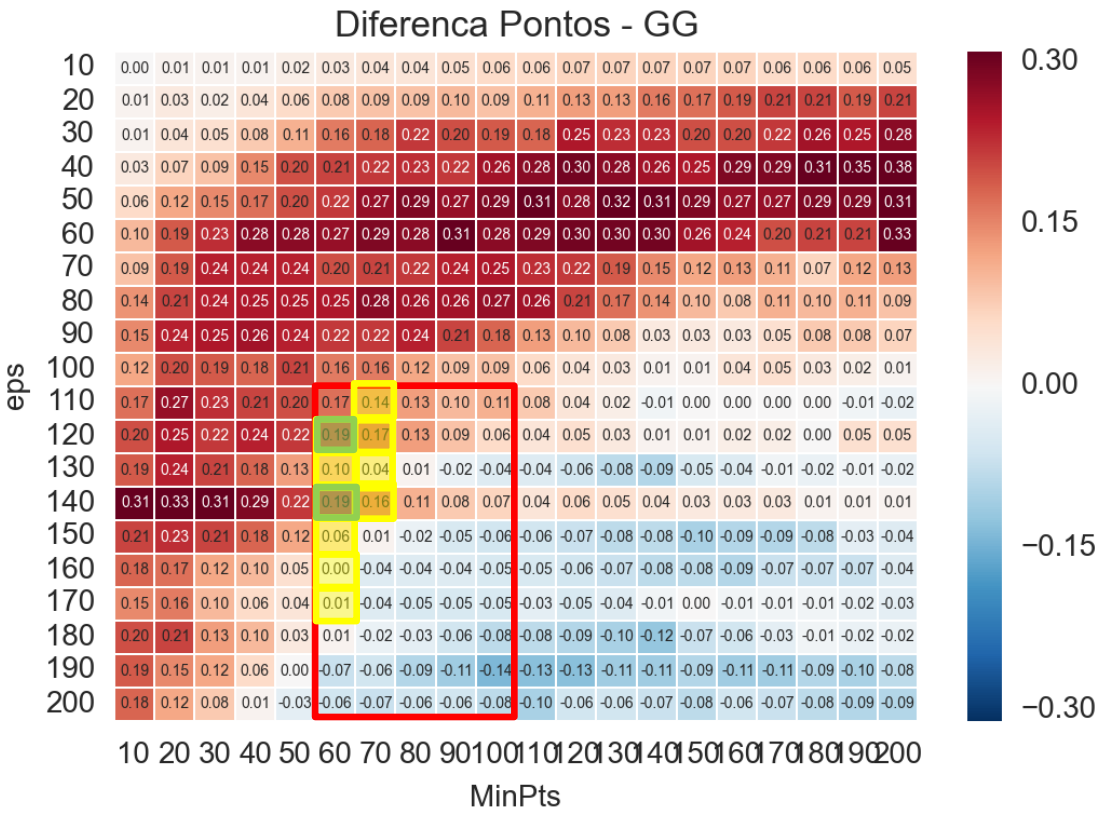


Figura IV.23 Cálculo de Δ_R para GG (13/05/14 e 20/05/14)

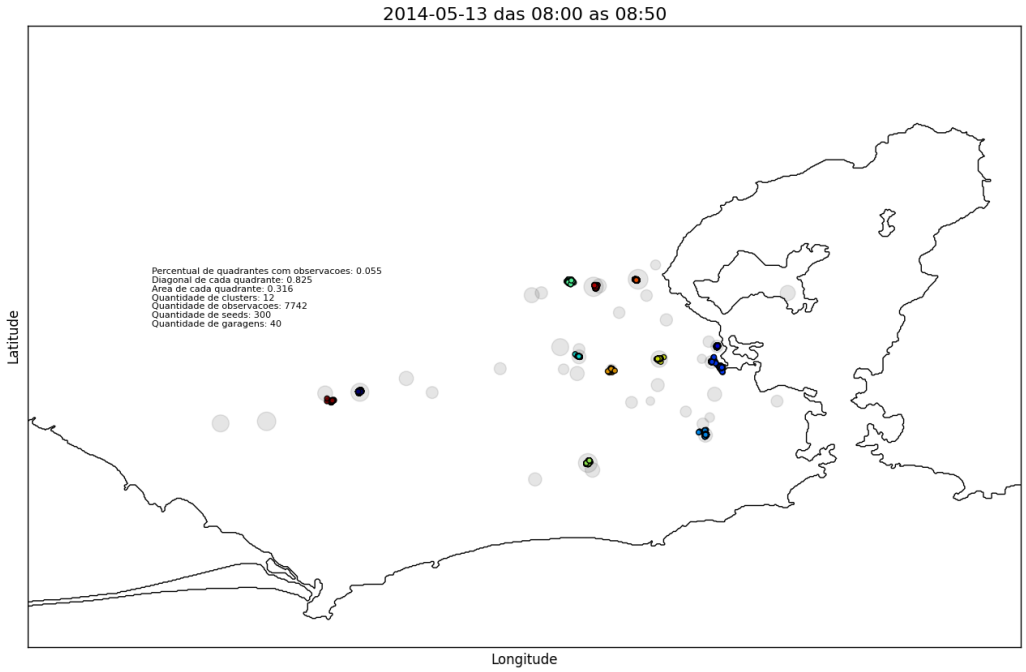


Figura IV.24 GG *MinPts* = 60 e *Malha* = 120 – 13/05/14

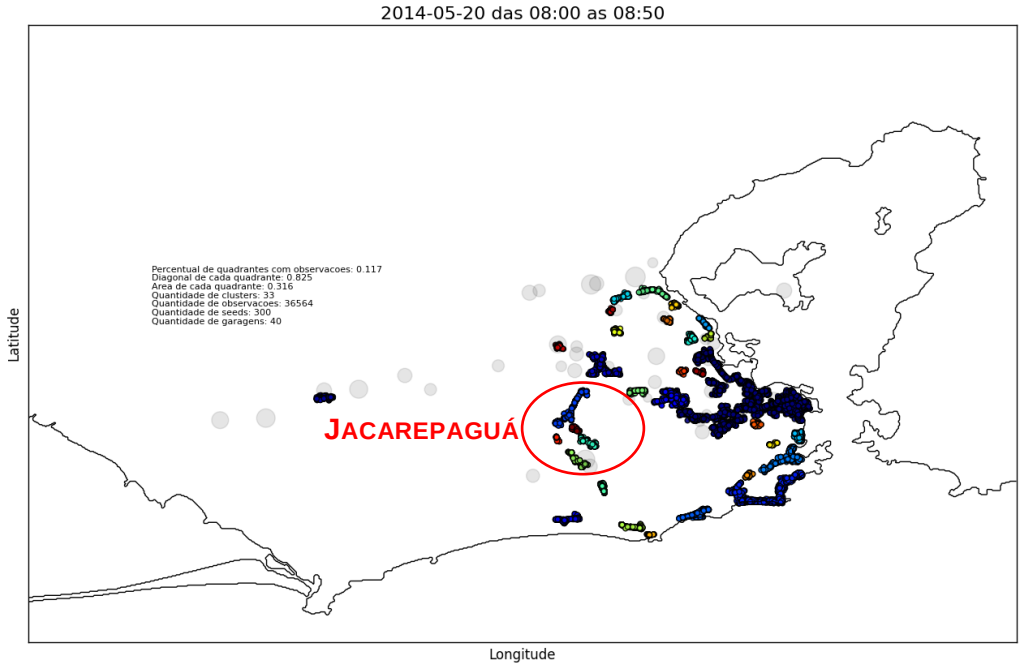


Figura IV.25 GG *MinPts* = 60 e *Malha* = 120 – 20/05/14

Cruzando as saídas de ambos algoritmos para o dia 20/05/2014, assim como no exemplo anterior, identificamos resultados bastante equivalentes. Nos debruçaremos sobre a análise da Região de Jacarepaguá conforme apresentado na Figura IV.26.

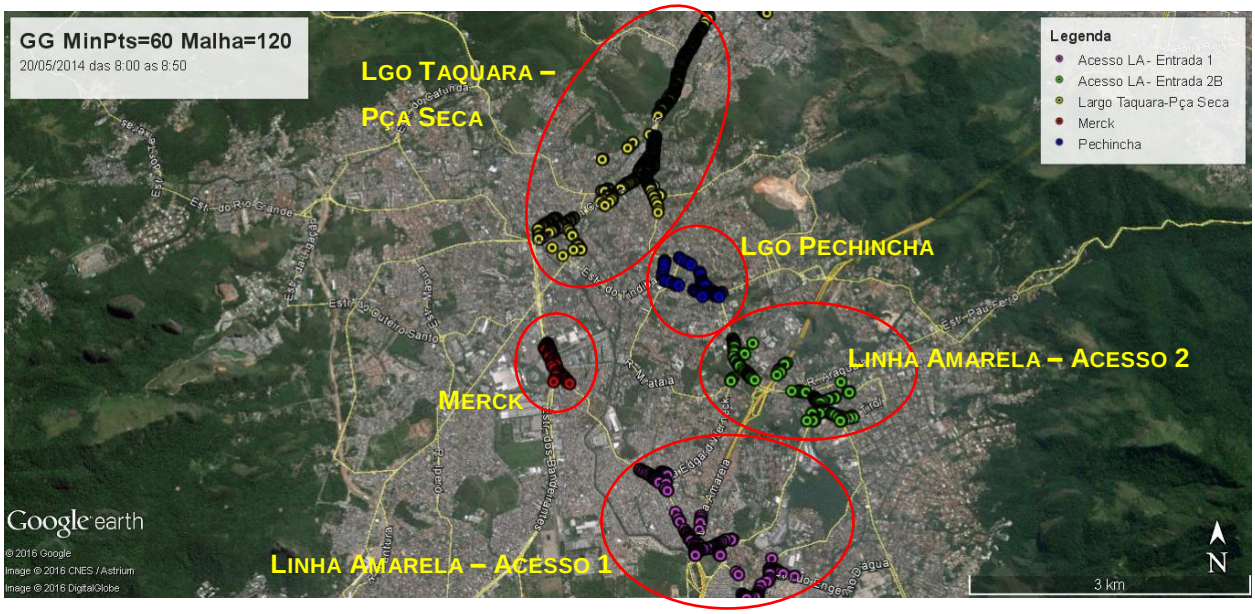


Figura IV.26 GG *MinPts* = 60 e *Malha* = 120 – 20/05/14 (Região Jacarepaguá)

- Cidade de Deus (acesso 1 Linha Amarela): A linha amarela é uma das principais vias que cruzam o bairro servindo de acesso para o Centro e para Barra da Tijuca, o que garante grande fluxo de veículos em ambos sentidos a qualquer hora do dia, principalmente no horário de *rush*;
- Freguesia (acesso 2 Linha Amarela): Outro acesso à Linha Amarela a partir de uma via bastante utilizada no Bairro, a Avenida Geremário Dantas;
- Largo do Pechincha: Junção de duas vias de acesso com fluxo intenso de veículos, Estrada do Pau-ferro e Geremário Dantas, ambas fundamentais nas rotas para Centro. De certo modo pode ser considerado uma continuação do *cluster* da Freguesia;
- Largo da Taquara/Pça Seca: Maior *cluster* formado na região, há tempos se apresenta como grande gargalo da mobilidade urbana do bairro, pois concentra todo tráfego de veículos destinado a Zona Norte;
- Estrada dos Bandeirantes (Merck): Confluência de duas vias de grande tráfego, Estrada dos Bandeirantes que traz veículos da Curicica, Vargem Grande e Recreio dos Bandeirantes, e da Estrada Miguel Salazar Mendes de Moraes com veículos oriundos majoritariamente da Cidade de Deus e Barra da Tijuca.

IV.3 Caso 3: Acidente no Viaduto dos Marinheiros

O último estudo de caso proposto neste trabalho, retrata um acidente de trânsito em uma importante via de saída do Centro do Rio de Janeiro. O acidente ocorreu em 23/09/2014, terça-feira, uma colisão entre um ônibus e um reboque no Viaduto dos Marinheiros que liga a movimentada Avenida Presidente Vargas à Avenida Radial Oeste, um dos acessos mais utilizados por quem vai em direção à Zona Norte[25]. O trânsito foi fechado parcialmente por aproximadamente duas horas, tendo sido liberado pela Companhia de Tráfego as 20h42m.

Espera-se que com esse fechamento parcial seja observado um congestionamento fora do comum nas proximidades do acidente, e em vias alternativas que liguem o Centro à Zona Norte. Para este exemplo definimos o dia 23/09/2014 das 19:00 as 19:50 como evento alvo, e o dia 16/09/2014 das 19:00 as 19:50 como evento referência. Diferente dos outros casos estudados até então, é provável que o evento alvo apresente discrepâncias locais (do acidente) em relação ao evento referência, não devendo haver grandes variações quanto a formação dos *clusters* em outras regiões ou quantidade total de *clusters* gerados. Para tentarmos identificar tais estrangulamentos nas proximidades do acidente, utilizaremos o quarto quadrante da nossa matriz de sensibilidade, ao invés, do terceiro quadrante que utilizamos nos exemplos anteriores. A seguir daremos prosseguimento a mesma dinâmica de análise de sensibilidade e agrupamento aplicada nas seções anteriores.

[25] Portal UOL dia 23/09/2014

IV.3.1 Caso 3: DBSCAN

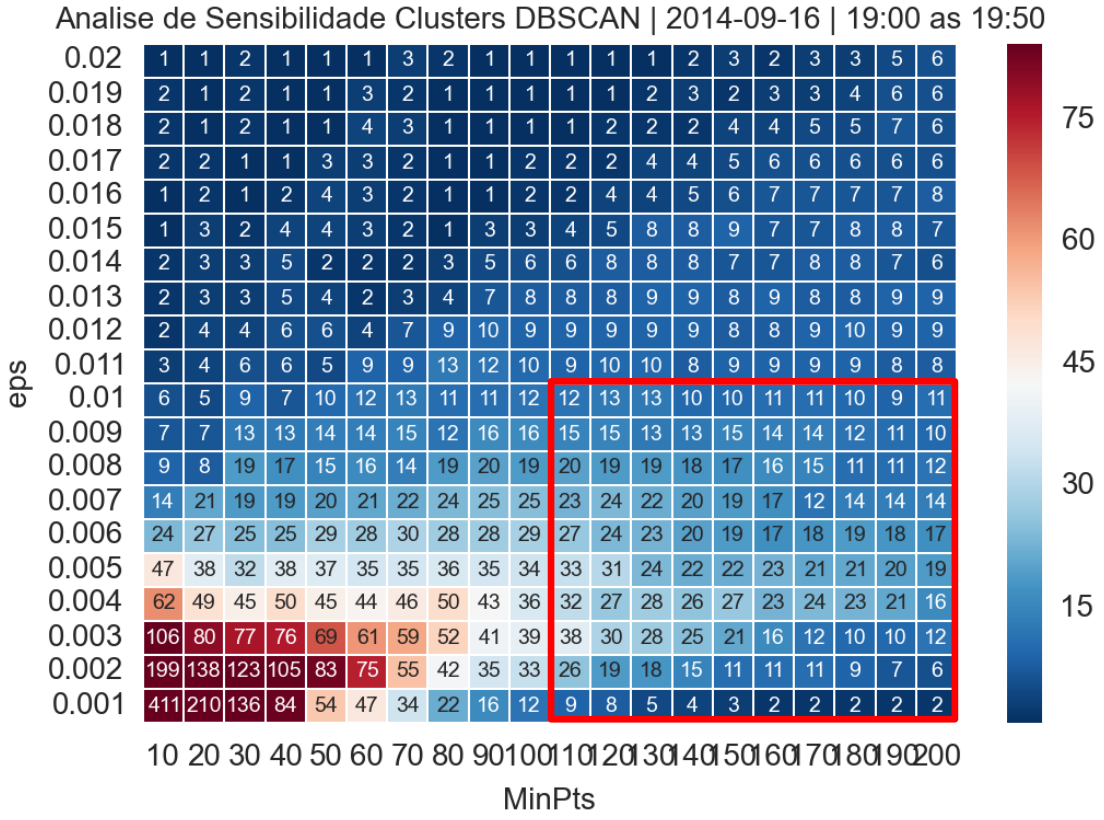


Figura IV.27 Análise de sensibilidade DBSCAN – 16/09/2014 (terça-feira)

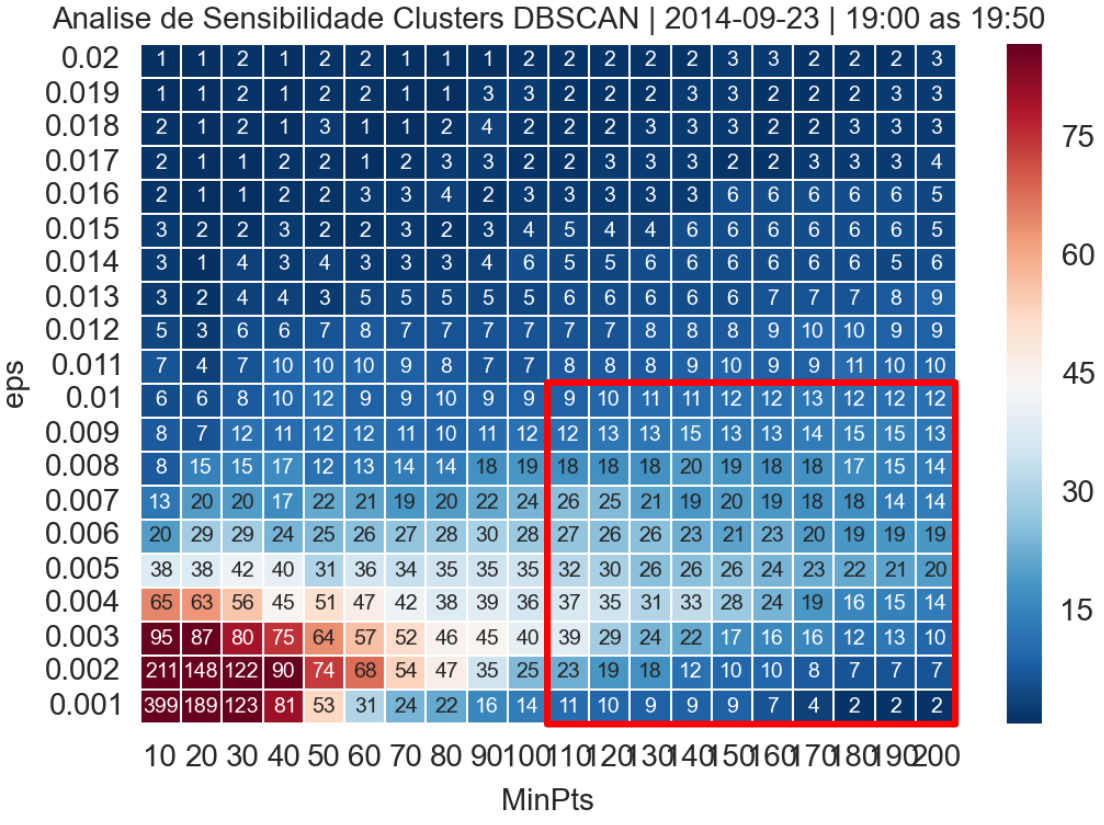


Figura IV.28 Análise de sensibilidade DBSCAN – 23/09/2014 (terça-feira)

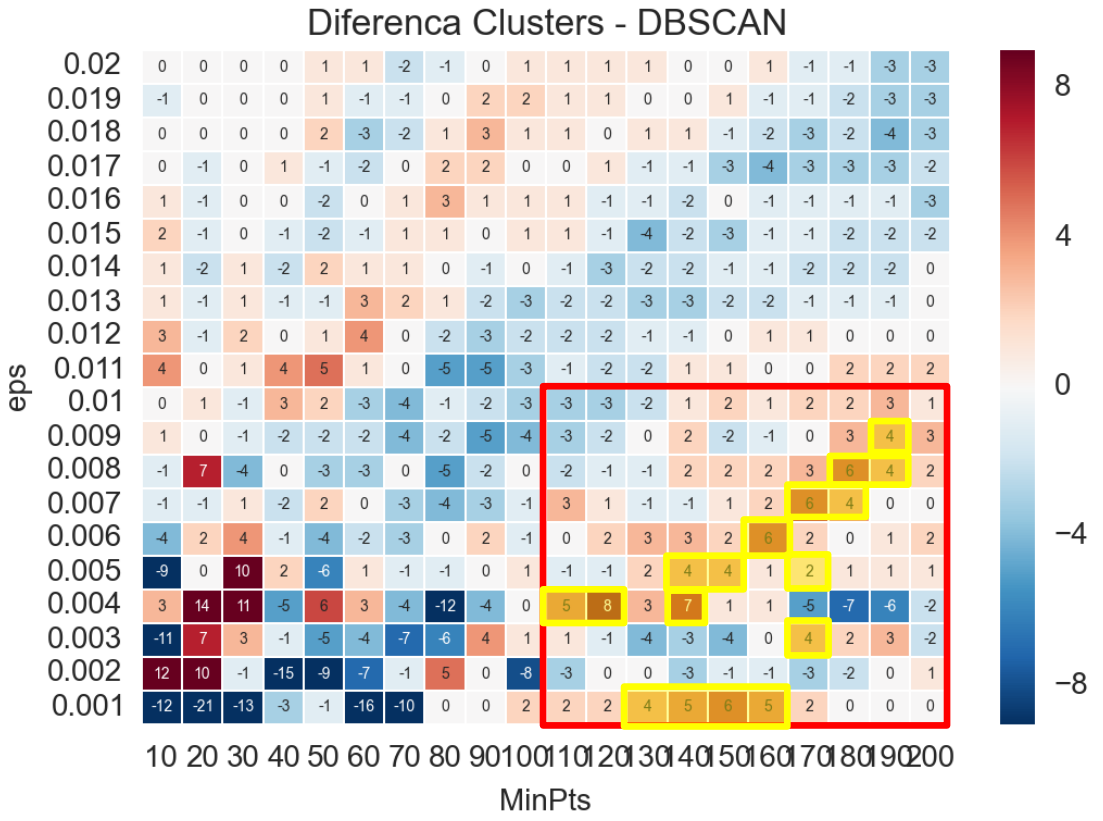


Figura IV.29 Cálculo de Δ_C para DBSCAN (16/09/14 e 23/09/14)

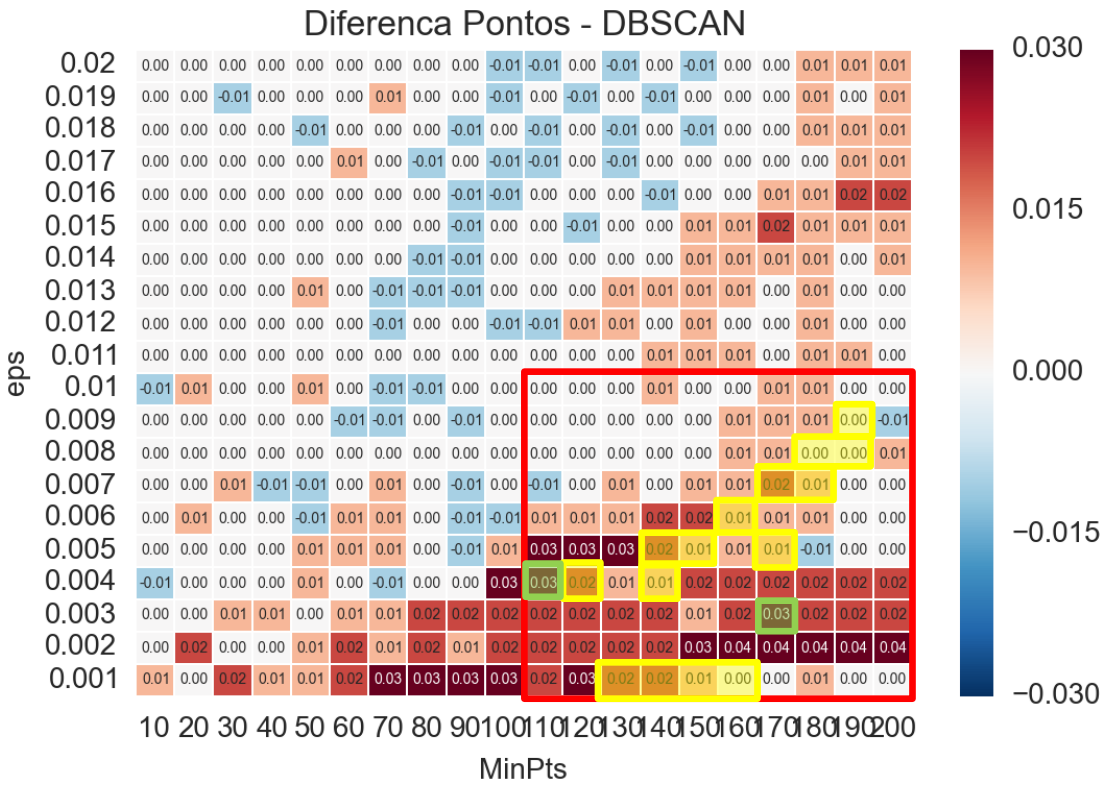


Figura IV.30 Cálculo de Δ_R para DBSCAN (16/09/14 e 23/09/14)

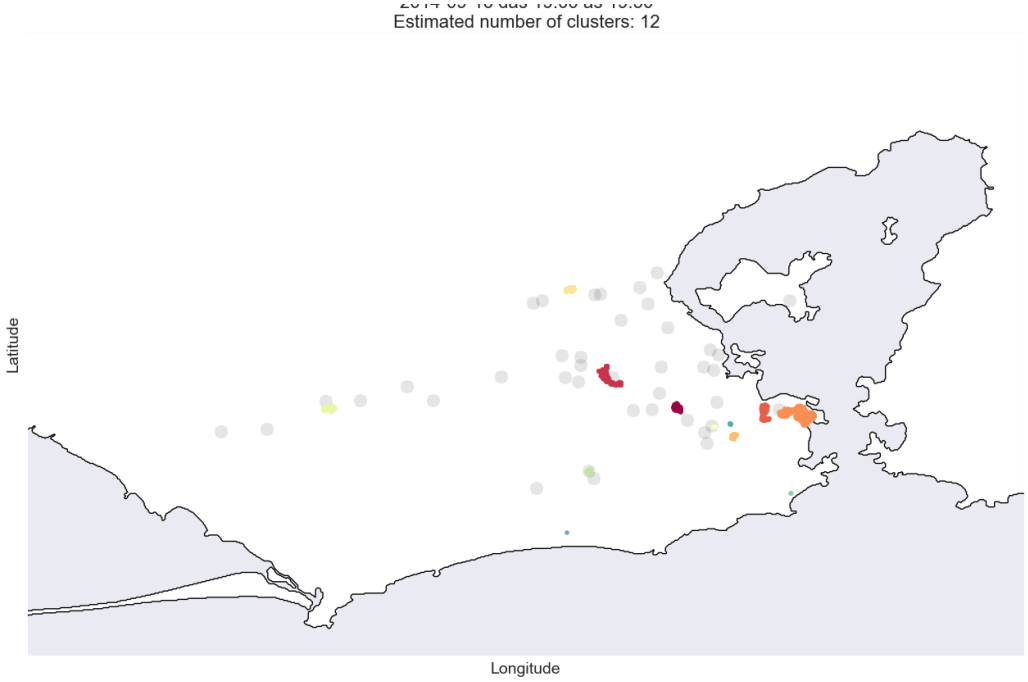


Figura IV.31 DBSCAN *MinPts* = 170 e ε = 0.003 – 16/09/14

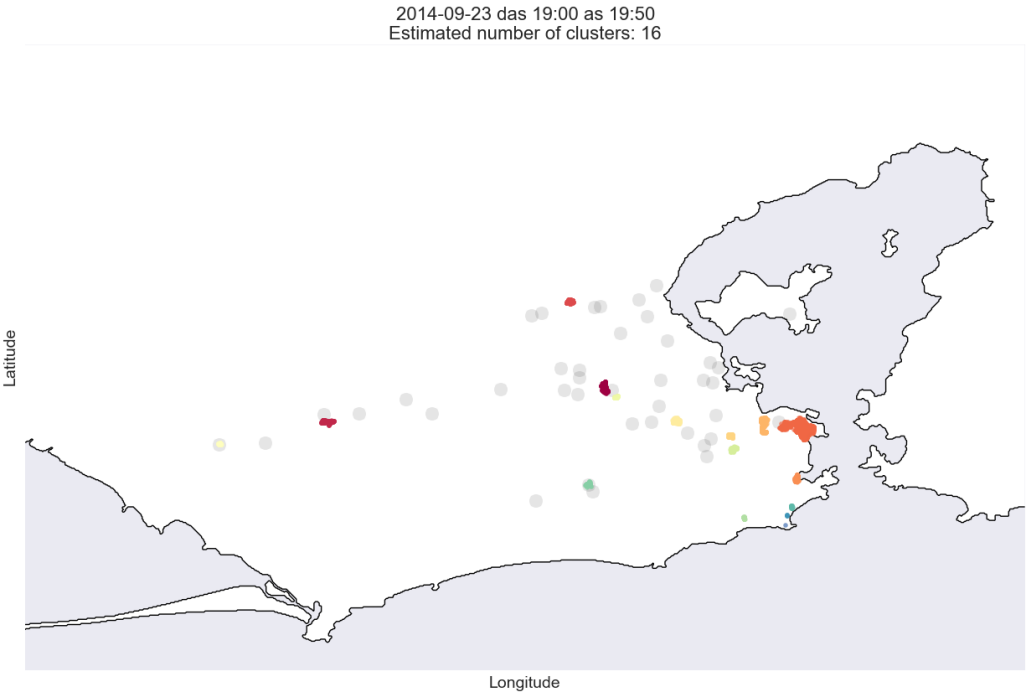


Figura IV.32 DBSCAN *MinPts* = 170 e ε = 0.003 – 23/09/14

A análise de sensibilidade para DBSCAN nos apontou dois resultados possíveis, sendo que optamos pelo par de parâmetros mais restritivos, pois conforme justificado no início da seção, nos interessa nesta rodada que sejam formados *clusters* com alta concentração de veículos, ocasionados pelo acidente do evento alvo. Comparando os resultados dos eventos alvo e referência, identificamos praticamente os mesmos *clusters* na área do Centro, local do acidente, porém diversos pequenos agrupamentos foram identificados no dia 23/09/2014 na

Zona Sul (Botafogo, Copacabana, Ipanema e Gávea). É possível que tais acúmulos de veículos tenham relação com o acidente, uma vez que moradores da Barra da Tijuca e Zona Oeste podem ter alterado seu trajeto usual pela Avenida Presidente Vargas, para um trajeto via orla da Zona Sul. Porém, é difícil prosseguirmos com tal afirmação sem mais informações, portanto, vamos recorrer ao GG para tentar solucionar este problema.

IV.3.2 Caso 3: *Grid Growing*

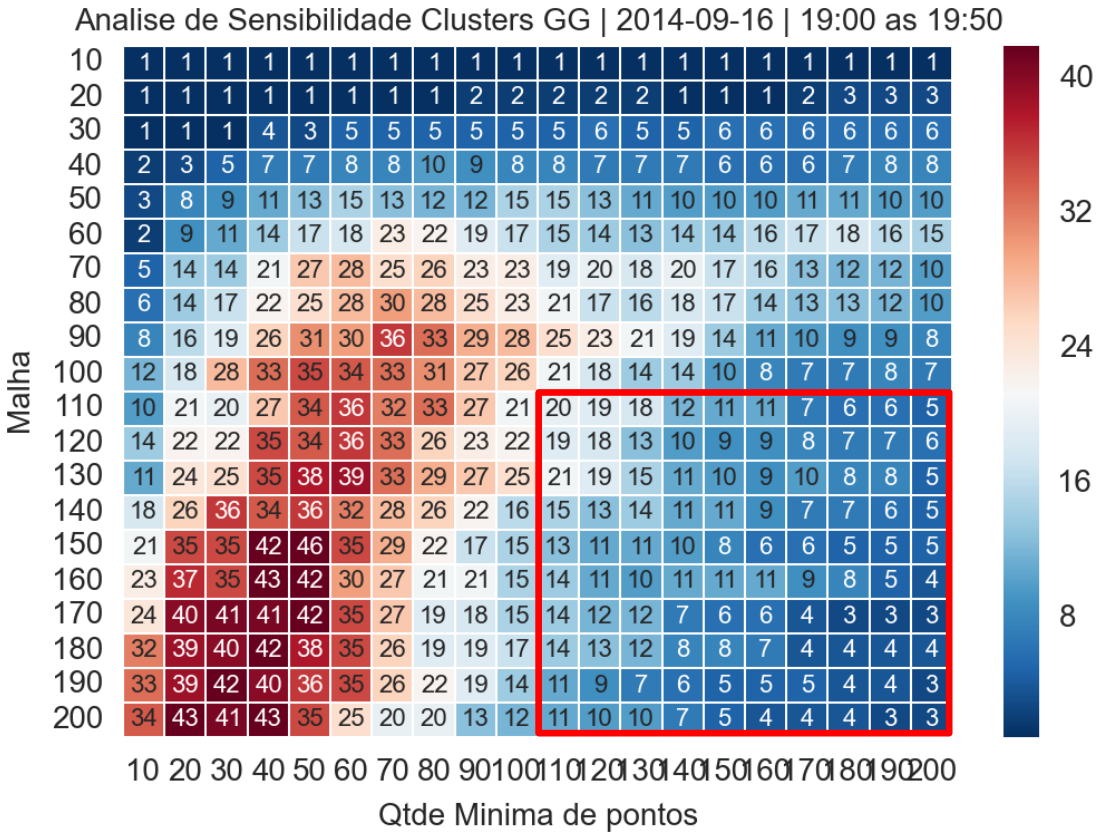


Figura IV.33 Análise de sensibilidade GG – 16/09/2014 (terça-feira)

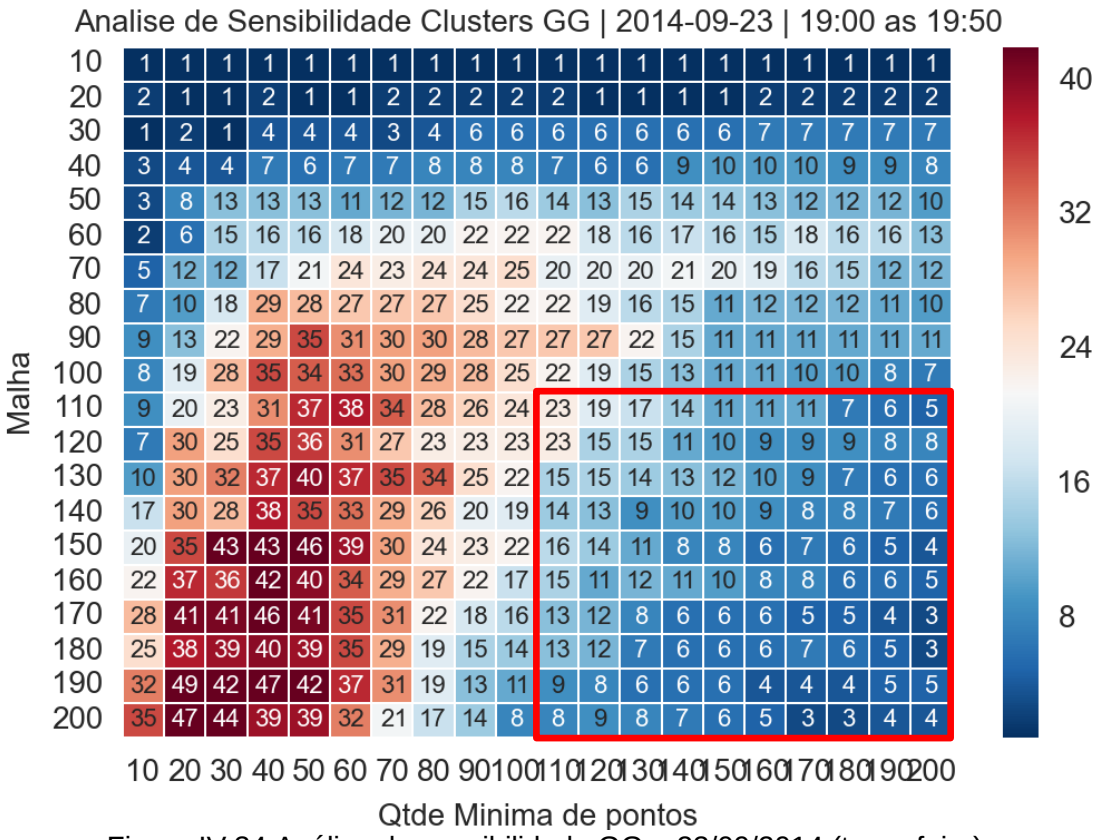


Figura IV.34 Análise de sensibilidade GG – 23/09/2014 (terça-feira)

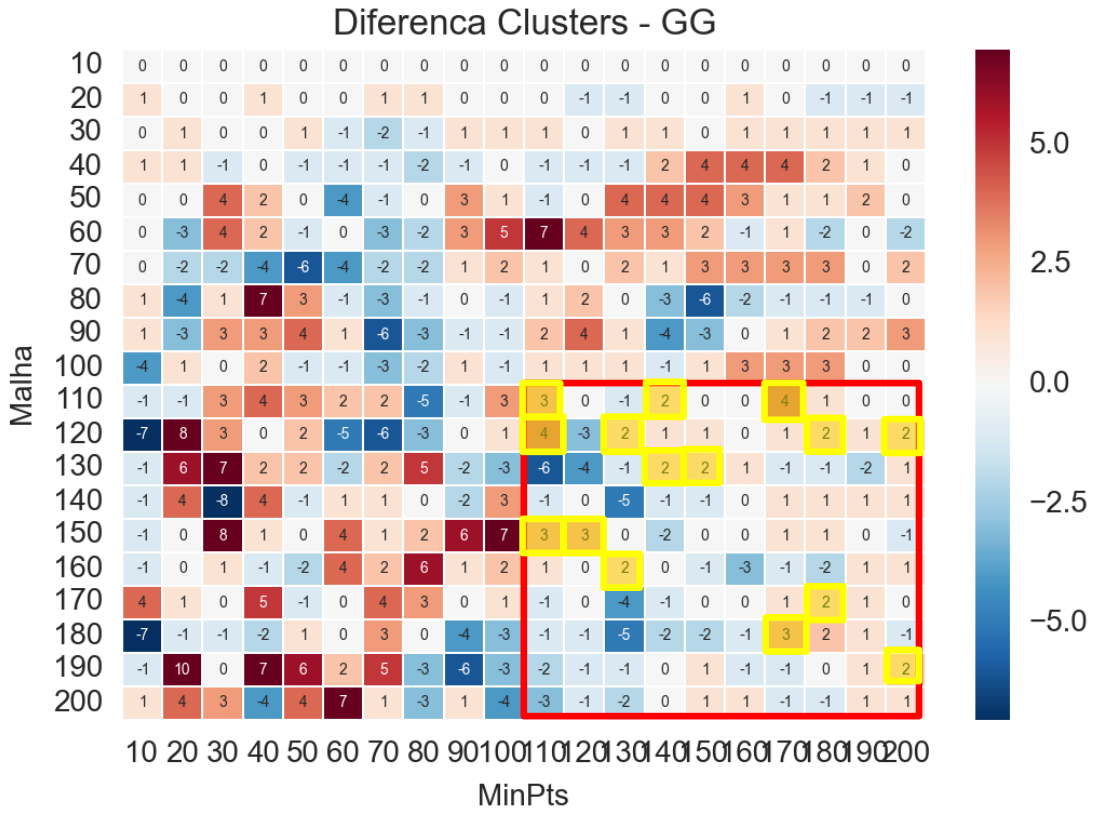


Figura IV.35 Cálculo de Δ_C para GG (16/09/14 e 23/09/14)

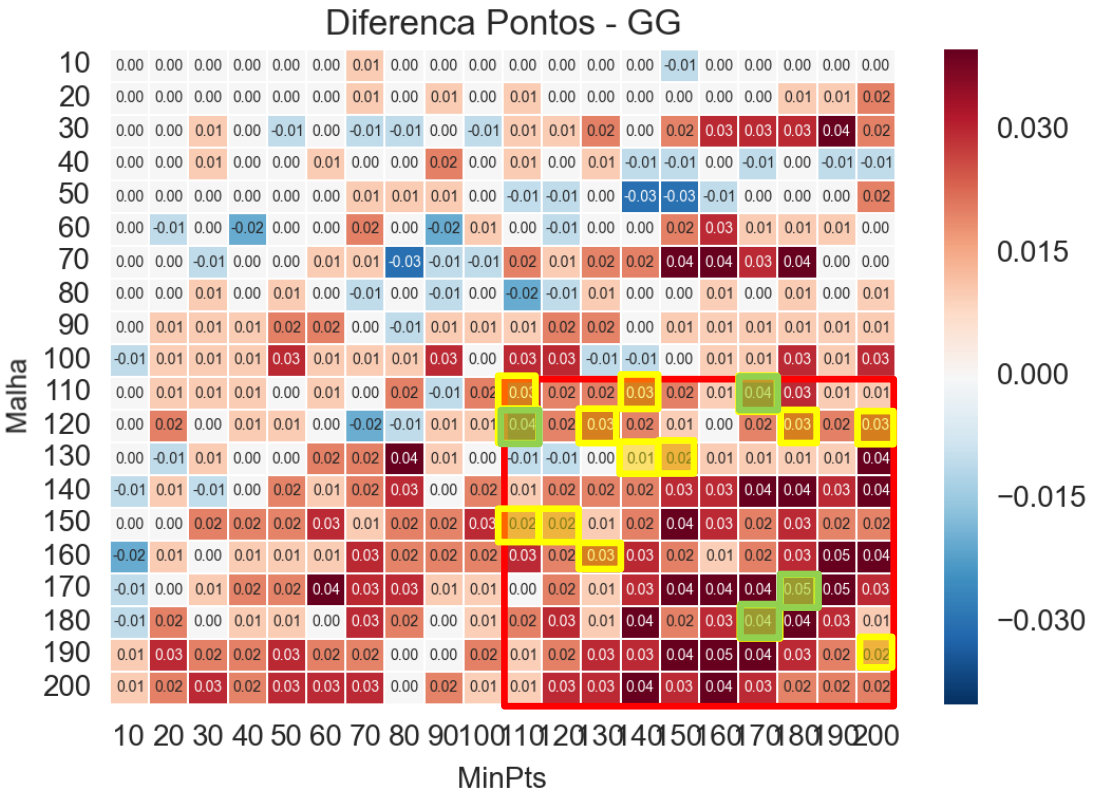


Figura IV.36 Cálculo de Δ_R para GG (16/09/14 e 23/09/14)

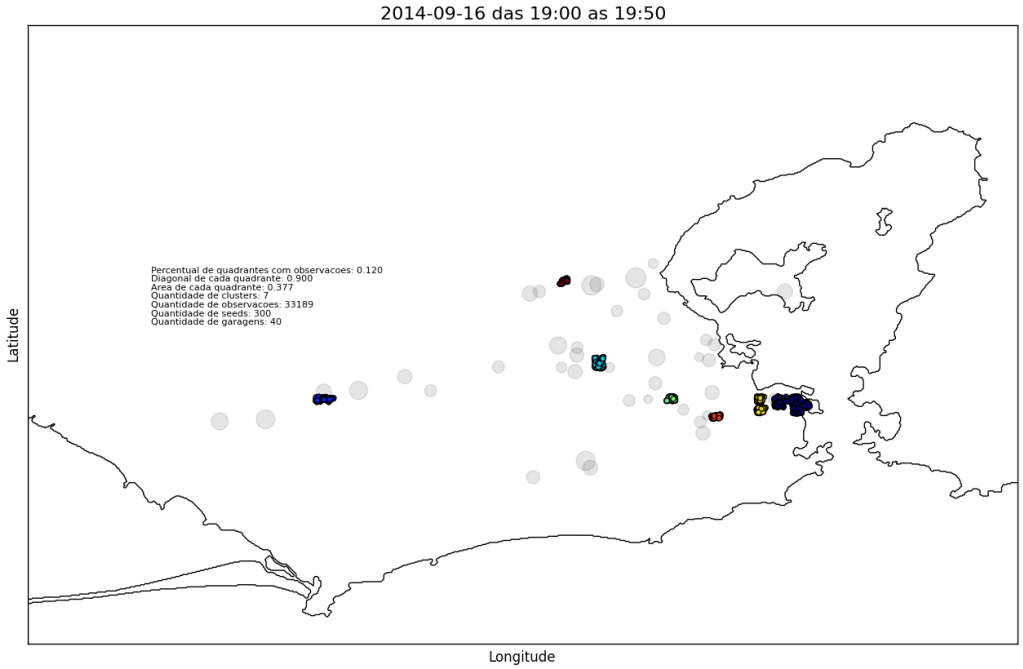


Figura IV.37 GG *MinPts* = 170 e *Malha* = 110 – 16/09/14

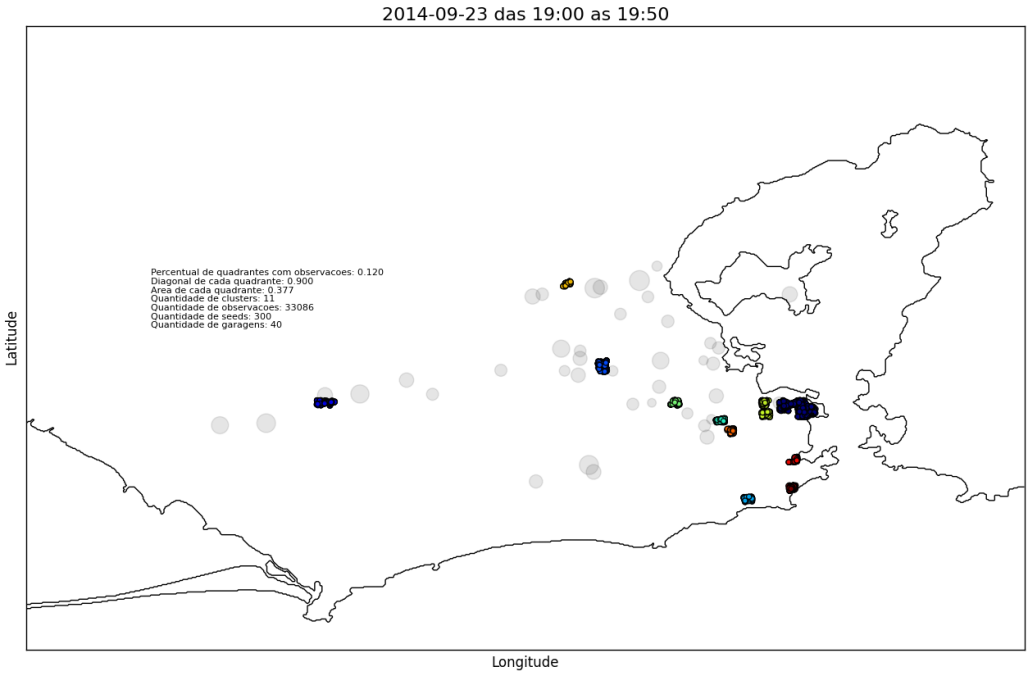


Figura IV.38 GG *MinPts* = 170 e *Malha* = 110 – 23/09/14

A análise de sensibilidade do GG referendou quatro pares de parâmetros, dos quais decidimos arbitrariamente utilizar o parâmetro 170x110. Assim como identificado pós DBSCAN, o resultado do GG também apontou *clusters* extraordinários pela Zona Sul no dia 23/09/2014. Com relação a região do Centro, vamos plotar as observações em uma escala maior para tentar identificar eventuais anomalias causadas pelo acidente (Figura IV.39 e Figura IV.40).



Figura IV.39 GG *MinPts* = 170 e *Malha* = 110 – 16/09/14 (Região Centro-Tijuca)

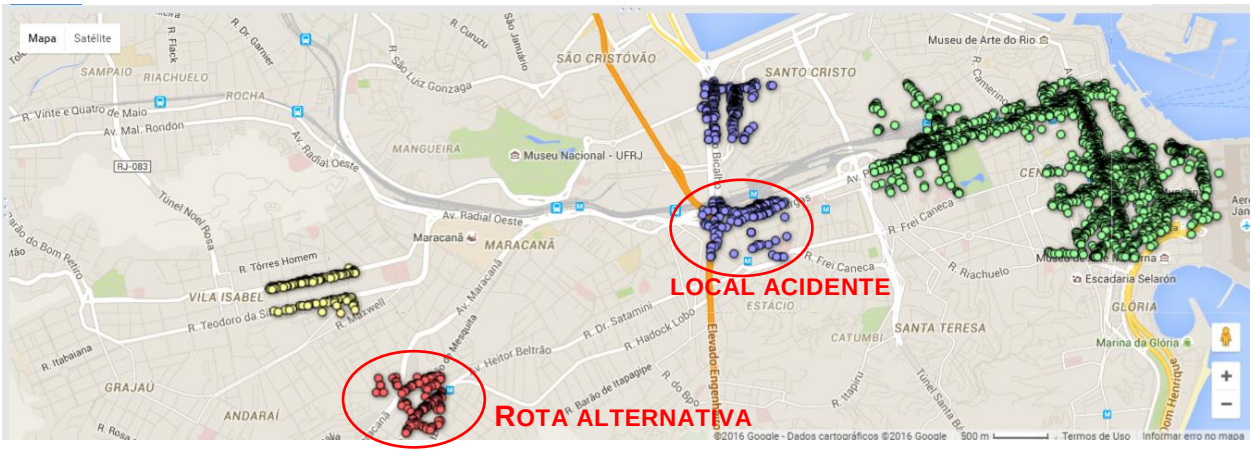


Figura IV.40 GG *MinPts* = 170 e *Malha* = 110 – 23/09/14 (Região Centro-Tijuca)

Comparando os resultados dos eventos alvo e referência na área do acidente, identificamos o surgimento de um *cluster* nas proximidades da Rua Conde de Bonfim altura da Praça Saens Pena, o que pode caracterizar uma rota alternativa adotada por motoristas que pretendiam escapar do trânsito na Avenida Presidente utilizando as Ruas Frei Caneca, Doutor Satamini e Avenida Heitor Beltrão, culminando em um aumento do trânsito na região da Tijuca (Figura IV.40).

Utilizamos os demais pares de parâmetros resultado da análise de sensibilidade, porém nenhum deles contribuiu com informações diferentes da que já havíamos levantado. Então decidimos estender as análises para alguns parâmetros que haviam sido descartados na última etapa da análise quantitativa (Cálculo de Δ_R). Ao rodar GG com parâmetros 110x150 nos deparamos com o seguinte cenário (Figura IV.41 e Figura IV.42).

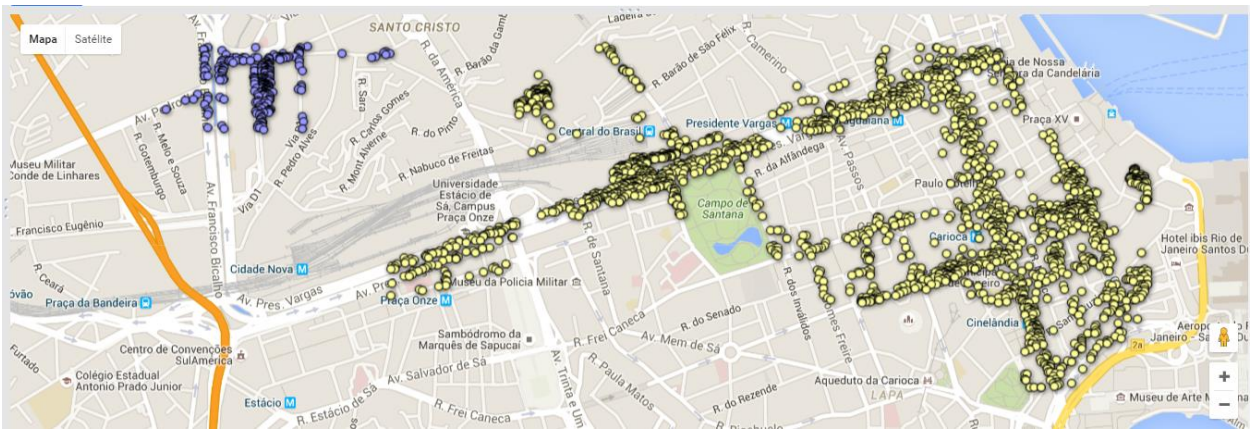


Figura IV.41 GG *MinPts* = 110 e *Malha* = 150 – 16/09/14 (Região Centro)



Figura IV.42 GG *MinPts* = 110 e *Malha* = 150 – 23/09/14 (Região Centro)

A Figura IV.42 mostra a formação de um *cluster* na Avenida Francisco Bicalho, bem próximo ao local do acidente que não foi identificado no evento alvo (Figura IV.41). Esse fato pode sugerir uma opção de desvio feita pelos motoristas que ao invés de acessar a Radial Oeste diretamente pelo Viaduto dos Marinheiros, decidiram realizar este acesso via retornado na Francisco Bicalho. Outro argumento que corrobora com esta hipótese é o fato deste novo *cluster* não se estender até o *cluster* da Rodoviária (destacada em azul), o que poderia caracterizar, por exemplo, um engarrafamento mais pesado nesta região que teria se prolongado até a Avenida Presidente Vargas. No próximo capítulo apresentaremos as considerações finais e potenciais oportunidades futuras de prolongamento desta pesquisa

Capítulo V - Conclusões

O principal objetivo desta dissertação é desenvolver um método para análise de dados geoespaciais com a finalidade de identificar gargalos no trânsito de cidades. Conforme observado no Capítulo I, o maior motivador para este estudo é o provável agravamento da situação da mobilidade urbana em grandes metrópoles devido ao crescimento da população urbana frente a população rural.

No Capítulo II identificamos um interesse latente da comunidade científica sobre o problema da mobilidade urbana em grandes centros e analisamos o crescimento na quantidade de publicações na área durante a última década. Verificamos que esse interesse sobre o tema não deve ser encarado como um fenômeno efêmero, mas sim como uma tendência duradoura suportada principalmente pelas novas tecnologias que permitem nos orientar quanto as principais linhas de pesquisa e ferramentas adotadas pelos especialistas na proposição de soluções para as questões sobre sistemas inteligentes de transporte.

No capítulo seguinte detalhamos o método proposto por este trabalho, apresentamos alguns conceitos necessários para seu pleno entendimento, as ferramentas computacionais utilizadas nos experimentos, o conjunto de dados disponível, as técnicas de mineração de dados selecionadas e modo como foram aplicadas ao problema escolhido e a elaboração de uma proposta inédita para determinação dos parâmetros de entrada dos algoritmos de agrupamento.

No Capítulo IV foram apresentados os resultados experimentais de três estudos de casos que utilizaram os processos estabelecidos anteriormente e os dados disponibilizados pela Prefeitura Municipal do Rio de Janeiro e que foram gerados a partir de dispositivos GPS instalados na frota de ônibus que atendem a população da cidade. Cada um dos casos analisados foi minuciosamente dissecado, tendo sido apresentadas todas as etapas previstas pelo método de identificação dos gargalos de trânsito, o qual teve sua eficácia avaliada por tais situações.

No primeiro caso foram selecionados um dia atípico na rotina da cidade, estreia do Brasil na Copa do Mundo, e um dia útil comum. O objetivo era verificar a partir da comparação entre os dois dias selecionados, a formação de *clusters* pela cidade e validar se os resultados da fase de mineração de dados refletiam a realidade cotidiana vivida pela população. Os dois algoritmos apresentaram resultados bastante semelhantes, sendo possível observar em ambos os casos as principais regiões da cidade com seus agrupamentos principais bem definidos. Selecionamos umas das regiões, a Zona Sul, para uma análise mais detalhada e verificamos

um a um os *clusters* identificados e a consistência em relação aos congestionamentos usualmente conhecidos para o horário da análise. Os resultados se apresentaram bastante coerentes, sinalizando positivamente quanto a eficácia do método.

Assim como o caso anterior, o segundo exemplo tem como objetivo avaliar a identificação genérica de agrupamentos pela cidade. Para tal, foram escolhidos mais uma vez dois eventos, um dia de greve dos rodoviários (evento referência) e um dia útil sem excepcionalidades conhecidas que impactassem fortemente o trânsito da cidade (evento alvo). Novamente os resultados obtidos pelos dois algoritmos de agrupamento foram bem semelhantes, apesar da diferença entre a quantidade total de clusters formados entre DBSCAN e GG, esse fato deve-se basicamente a aglutinação de alguns *clusters* menores em outros maiores. Já com relação a variação da quantidade de *clusters* entre os dias analisados, pudemos perceber uma concentração de observações no evento alvo em relação ao evento referência, caracterizando os congestionamentos típicos de um dia de semana. Selecionamos outra parte da cidade para avaliar cada um dos agrupamentos propostos pelas técnicas de mineração de dados, a região de Jacarepaguá e todas as áreas identificadas como estrangulamentos no experimento coincidem com áreas que apresentam tais problemas na prática.

E por fim, apresentamos um estudo de caso referente a um acidente ocorrido no horário de *rush* de fim de tarde em uma das vias de acesso mais movimentadas da cidade, que liga o Centro à Zona Norte. Diferente dos exemplos anteriores, o objetivo desta análise era detectar possíveis variações no trânsito das imediações do local acidente. Analisando os resultados do DBSCAN pudemos observar uma manutenção da situação no trânsito do Centro entre evento alvo (dia do acidente) e evento referência (dia sem anormalidades). Porém, foram detectados novos *clusters* na região da Zona Sul, o que pode ser reflexo do acidente no Centro, uma vez que a mudança no trajeto pode ter sido uma solução escolhida pelos motoristas. Já em relação aos resultados do GG pudemos observar além dos *clusters* da Zona Sul, alguns agrupamentos bem próximos ao local do acidente, que caracterizariam pequenos desvios feitos pelos motoristas durante a volta para casa.

Consideramos os resultados obtidos em todos os casos analisados bastante satisfatórios, uma vez que houve coerência entre os resultados dos agrupamentos formados e a dinâmica do trânsito na cidade do Rio de Janeiro. Os algoritmos apresentaram cenários bem semelhantes, com uma sútil prevalência do GG que além de identificar bem as variações de trânsito por toda cidade (estudos de caso 1 e 2), conseguiu propor um cenário mais completo para análise das variações pontuais (estudo de caso 3). Certamente o método proposto nesta dissertação se apresenta como promissor, sendo importante submetê-lo a novas avaliações que contemplem cenários não explorados nesta pesquisa.

Além do objetivo principal definido no início deste trabalho, outras contribuições puderam ser observadas ao longo de seu desenvolvimento. Os dois algoritmos de agrupamento tiveram suas aplicações adaptada às necessidades do problema proposto. No caso do GG, a proposta inicial feita por ZHAO et al. (2015) contemplava identificar pontos de interesse utilizando dados de embarque e desembarque de passageiros em taxis. Estendemos esta aplicabilidade, atribuindo uma nova possibilidade de uso para algoritmos baseados em *grid*. Foi desenvolvido uma proposta para determinação dos parâmetros de entrada para DBSCAN e GG não identificado na literatura até então, que congrega uma etapa preliminar para análise qualitativa com base na natureza do problema e uma etapa de análise quantitativa que leva em consideração a variação na formação dos *clusters* e nas observações entre evento alvo e referência. Outra contribuição relevante foi a revisão literária sobre o tema proposto, utilizamos uma grande quantidade de publicações que retratam as tendências de aplicação do uso da tecnologia com objetivo de responder a problemas de mobilidade urbana e de sistemas inteligentes de transporte ao redor do mundo.

Concluindo, deixamos como sugestão a indicação de potenciais iniciativas que deem prosseguimento ao trabalho apresentado. Conforme já comentado anteriormente, seria importante a submissão do método de identificação de estrangulamentos a novos casos que apresentem características distintas às dos experimentos realizados, inclusive se possível utilizando outra cidade para fins de comparação. Com relação a condução do método, alguns pontos podem ser melhor explorados, como por exemplo, uma análise de sensibilidade que contemple também variações na quantidade do *seed*, de modo a garantir plenamente que todos os *clusters* elegíveis foram identificados (método *Grid Growing*). Outro ponto seria o aumento na quantidade de algoritmos avaliados, incluindo nas comparações métodos de agrupamento por densidade e baseados em *grid*, tais como, ISDBSCAN, VDBSCAN, GDBSCAN, OPTICS, DBCLASD e STING com objetivo de buscar as alternativas que melhor respondam ao problema levantado. Também poderíamos incluir nas métricas de avaliação critérios relacionados a performance do processamento e custo computacional da solução. Por fim, a fase de pré-processamento e limpeza poderia incluir novos parâmetros para seleção do conjunto de dados, possibilitando a exclusão de registros que eventualmente não estejam sendo atualizados corretamente pelo GPS do veículo observado.

Referências Bibliográficas

ALVES, J. M. E. Modelaçao do campo da gravidade utilizando métodos de reduçao do terreno e anomalias de densidade. **Boletim da Sociedade de Geografia de Lisboa**, v. 104, n. 7, p. 5–112, 1986.

ANAGNOSTOPOULOS, C.; HADJIEFTHYMIADES, S.; KOLOMVATSOS, K. Time-optimized user grouping in Location Based Services. **Computer Networks**, v. 81, p. 220–244, abr. 2015.

ANKERST, M. et al. **OPTICS: Ordering Points To Identify the Clustering Structure**, 1999.

ASTARITA, V.; GUIDO, G.; GIOFRÈ, V. P. Co-operative ITS: Smartphone based Measurement Systems for Road Safety Assessment. **Procedia Computer Science**, v. 37, p. 404–409, 2014.

BACHMANN, C. et al. A comparative assessment of multi-sensor data fusion techniques for freeway traffic speed estimation using microsimulation modeling. **Transportation Research Part C: Emerging Technologies**, v. 26, p. 33–48, jan. 2013.

BALI, R. S.; KUMAR, N.; RODRIGUES, J. J. P. C. Clustering in vehicular ad hoc networks: Taxonomy, challenges and solutions. **Vehicular Communications**, v. 1, n. 3, p. 134–152, jul. 2014.

BEAZLEY, D.; JONES, B. K. **Python cookbook: [recipes for mastering Python 3]**. 3. ed ed. Bejing: O'Reilly, 2013.

BILAL, S. M.; BERNARDOS, C. J.; GUERRERO, C. Position-based routing in vehicular networks: A survey. **Journal of Network and Computer Applications**, v. 36, n. 2, p. 685–697, mar. 2013.

BORGIA, E. The Internet of Things vision: Key features, applications and open issues. **Computer Communications**, v. 54, p. 1–31, dez. 2014.

BOUKERCHE, A. et al. Vehicular Ad Hoc Networks: A New Challenge for Localization-Based Systems. **Computer Communications**, v. 31, n. 12, p. 2838–2849, jul. 2008.

BOWLES, M. **Machine learning in Python: essential techniques for predictive analysis**. Indianapolis, Ind: Wiley, 2015.

CALABRESE, F. et al. Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. **Transportation Research Part C: Emerging Technologies**, v. 26, p. 301–313, jan. 2013.

CAO, Q. et al. **A grid-based clustering method for mining frequent trips from large-scale, event-based telematics datasets**. Systems, Man and Cybernetics, 2009. SMC 2009. IEEE International Conference on. **Anais**...IEEE, 2009Disponível em: <http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5345924>. Acesso em: 21 abr. 2016

CAO, X.; CONG, G.; JENSEN, C. S. Mining significant semantic locations from GPS data. **Proceedings of the VLDB Endowment**, v. 3, n. 1-2, p. 1009–1020, 2010.

CARREL, A. et al. Quantifying transit travel experiences from the users' perspective with high-resolution smartphone and vehicle location data: Methodologies, validation, and example analyses. **Transportation Research Part C: Emerging Technologies**, v. 58, p. 224–239, set. 2015.

CASSISI, C. et al. **Enhancing density-based clustering: Parameter reduction and outlier detection**, 2012.

CHEN, L. et al. A personal route prediction system based on trajectory data mining. **Information Sciences**, v. 181, n. 7, p. 1264–1284, 1 abr. 2011.

CHEN, X.; GONG, H.; WANG, J. BRT Vehicle Travel Time Prediction Based on SVM and Kalman Filter. **Journal of Transportation Systems Engineering and Information Technology**, v. 12, n. 4, p. 29–34, ago. 2012.

CUI, J. et al. Identifying mismatch between urban travel demand and transport network services using GPS data: A case study in the fast growing Chinese city of Harbin. **Neurocomputing**, v. 181, p. 4–18, mar. 2016.

DANTAS, A.; TACO, P.; YAMASHITA, Y. Sistemas de Informação Geográfica em Transporte: O Estudo do Estado da Arte. **Proceedings of the X Congresso da Associação Nacional de Pesquisa e Ensino em Transportes (ANPET)**, p. 211–222, 1996.

DUNCAN, M. J.; BADLAND, H. M.; MUMMERY, W. K. Applying GPS to enhance understanding of transport-related physical activity. **Journal of Science and Medicine in Sport**, v. 12, n. 5, p. 549–556, set. 2009.

ESTER, M. et al. **A density-based algorithm for discovering clusters in large spatial databases with noise.** Kdd. **Anais**...1996Disponível em: <http://www.aaai.org/Papers/KDD/1996/KDD96-037>. Acesso em: 19 abr. 2016

FENG, Y.; HOURDOS, J.; DAVIS, G. A. Probe vehicle based real-time traffic monitoring on urban roadways. **Transportation Research Part C: Emerging Technologies**, v. 40, p. 160–178, mar. 2014.

FLORIN, R.; OLARIU, S. A survey of vehicular communications for traffic signal optimization. **Vehicular Communications**, v. 2, n. 2, p. 70–79, abr. 2015.

FONSECA, A.; VAZÃO, T. Applicability of position-based routing for VANET in highways and urban environment. **Journal of Network and Computer Applications**, v. 36, n. 3, p. 961–973, maio 2013.

GOODRICH, M. T.; TAMASSIA, R.; GOLDWASSER, M. H. **Data Structures and Algorithms in Python**, mar. 2013.

GUBBI, J. et al. Internet of Things (IoT): A vision, architectural elements, and future directions. **Future Generation Computer Systems**, v. 29, n. 7, p. 1645–1660, set. 2013.

HAGE, R.-M. et al. Unscented Kalman filter for urban network travel time estimation. **Procedia - Social and Behavioral Sciences**, v. 54, p. 1047–1057, out. 2012.

HAN, J.; KAMBER, M.; PEI, J. **Data Mining Concepts and Techniques**. third edition ed. [s.l.] Morgan Kaufmann, 2011.

HAO, J.; ZHU, J.; ZHONG, R. The rise of big data on urban studies and planning practices in China: Review and open research issues. **Journal of Urban Management**, v. 4, n. 2, p. 92–124, dez. 2015.

HARA, Y.; KUWAHARA, M. Traffic Monitoring immediately after a major natural disaster as revealed by probe data – A case in Ishinomaki after the Great East Japan Earthquake. **Transportation Research Part A: Policy and Practice**, v. 75, p. 1–15, maio 2015.

HOPCROFT, J.; KANNAN, R. **Foundations of Data Science**. [s.l: s.n.].

JABBARPOUR, M. R.; NOOR, R. M.; KHOKHAR, R. H. Green vehicle traffic routing system using ant-based algorithm. **Journal of Network and Computer Applications**, v. 58, p. 294–308, dez. 2015.

JIMÉNEZ-MEZA, A.; ARÁMBURO-LIZÁRRAGA, J.; DE LA FUENTE, E. Framework for Estimating Travel Time, Distance, Speed, and Street Segment Level of Service (LOS), based on GPS Data. **Procedia Technology**, v. 7, p. 61–70, 2013.

JOSSE, G. et al. **A framework for computation of popular paths from crowdsourced data**. Data Engineering (ICDE), 2015 IEEE 31st International Conference on. **Anais**...IEEE, 2015Disponível em: <http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=7113393>. Acesso em: 5 mar. 2016

KAFI, M. A. et al. A study of Wireless Sensor Network Architectures and Projects for Traffic Light Monitoring. **Procedia Computer Science**, v. 10, p. 543–552, 2012.

KERAMAT JAHROMI, K. et al. Simulating human mobility patterns in urban areas. **Simulation Modelling Practice and Theory**, v. 62, p. 137–156, mar. 2016.

KHALEGHI, B. et al. Multisensor data fusion: A review of the state-of-the-art. **Information Fusion**, v. 14, n. 1, p. 28–44, jan. 2013.

KONG, X. et al. Urban traffic congestion estimation and prediction based on floating car trajectory data. **Future Generation Computer Systems**, dez. 2015.

KURZHANSKIY, A. A.; VARAIYA, P. Traffic management: An outlook. **Economics of Transportation**, v. 4, n. 3, p. 135–146, set. 2015.

LIU, P.; ZHOU, D.; WU, N. **VDBSCAN: Varied Density Based Spatial Clustering of Applications with Noise**, 2007.

LIU, X. et al. Revealing travel patterns and city structure with taxi trip data. **Journal of Transport Geography**, v. 43, p. 78–90, fev. 2015.

LIU, X.; KARIMI, H. A. Location awareness through trajectory prediction. **Computers, Environment and Urban Systems**, v. 30, n. 6, p. 741–756, nov. 2006.

LOPES, A. F.; CORDEIRO, M. C.; PINTO, N. L. **Desenvolvimento de uma metodologia para cálculo do índice de circulação de frota dos ônibus no município do Rio de Janeiro**, nov. 2015.

LUTZ, M. **Learning Python: [powerful object-oriented programming ; covers Python 2.6 and 3.x]**. 4. ed., [Nachdr.] ed. Beijing: O'Reilly, 2009.

MANNINI, L. et al. On the Short-term Prediction of Traffic State: An Application on Urban Freeways in ROME. **Transportation Research Procedia**, v. 10, p. 176–185, 2015.

MA, X. et al. Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. **Transportation Research Part C: Emerging Technologies**, v. 54, p. 187–197, maio 2015.

MAZLOUMI, E.; CURRIE, G.; ROSE, G. Using GPS data to gain insight into public transport travel time variability. **Journal of Transportation Engineering**, v. 136, n. 7, p. 623–631, 2009.

MEI, Y.; TANG, K.; LI, K. Real-time identification of probe vehicle trajectories in the mixed traffic corridor. **Transportation Research Part C: Emerging Technologies**, v. 57, p. 55–67, ago. 2015.

MITCHELL, A. **The ESRI Guide to GIS Analysis - Volume 1 - Geographic Patterns and Relationships**. [s.l.] ESRI Press, 1999. v. 1

MONTOLIU, R.; GATICA-PEREZ, D. **Discovering human places of interest from multimodal mobile phone data**. Proceedings of the 9th International Conference on Mobile and Ubiquitous Multimedia. **Anais**...ACM, 2010Disponível em: <http://dl.acm.org/citation.cfm?id=1899487>. Acesso em: 21 abr. 2016

MOREIRA-MATIAS, L. et al. Time-evolving O-D matrix estimation using high-speed GPS data streams. **Expert Systems with Applications**, v. 44, p. 275–288, fev. 2016.

NEIROTTI, P. et al. Current trends in Smart City initiatives: Some stylised facts. **Cities**, v. 38, p. 25–36, jun. 2014.

NICULESCU, D.; NATH, B. DV based positioning in ad hoc networks. **Telecommunication Systems**, v. 22, n. 1-4, p. 267–280, 2003.

PAN, G. et al. Land-Use Classification Using Taxi GPS Traces. **IEEE Transactions on Intelligent Transportation Systems**, v. 14, n. 1, p. 113–123, mar. 2013.

PAN, H.-H.; WANG, S.-C.; YAN, K.-Q. An integrated data exchange platform for Intelligent Transportation Systems. **Computer Standards & Interfaces**, v. 36, n. 3, p. 657–671, mar. 2014.

PARK, H. J. et al. Data Mining Strategies for Real-time Control in New York City. **Procedia Computer Science**, v. 32, p. 109–116, 2014.

POUKE, M. et al. Practical simulation of virtual crowds using points of interest. **Computers, Environment and Urban Systems**, v. 57, p. 118–129, maio 2016.

QURESHI, K. N.; ABDULLAH, A. H. A survey on intelligent transportation systems. **Middle-East Journal of Scientific Research**, v. 15, n. 5, p. 629–642, 2013.

RAO, K.; GOVARDHAN, A.; RAO, K. V. C. Spatiotemporal Data Mining: Issues, Tasks And Applications. **International Journal of Computer Science & Engineering Survey**, v. 3, n. 1, p. 39–52, 29 fev. 2012.

RIGAUX, P.; SCHOLL, M.; VOISARD, A. **Spatial Databases with Application to GIS**. [s.l: s.n.].

RIVEST, S. et al. SOLAP technology: Merging business intelligence with geospatial technology for interactive spatio-temporal exploration and analysis of data. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 60, n. 1, p. 17–33, dez. 2005.

RORIZ JUNIOR, M.; ENDLER, M.; SILVA, F. J. DA S. E. An on-line algorithm for cluster detection of mobile nodes through complex event processing. **Information Systems**, jan. 2016.

ROUHIEH, B.; ALECSANDRU, C. Adaptive route choice model for public transit systems: An application of Markov decision processes. **Canadian Journal of Civil Engineering**, v. 39, n. 8, p. 915–924, ago 2012.

SALIM, F.; HAQUE, U. Urban computing in the wild: A survey on large scale participation and citizen engagement with ubiquitous computing, cyber physical systems, and Internet of Things. **International Journal of Human-Computer Studies**, v. 81, p. 31–48, set. 2015.

SANDER, J. et al. **Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications**, 1998.

SANTOS, A. R.; LOUZADA, F. L. R. O.; EUGENIO, F. C. **ArcGIS 9.3 Total - Aplicações para Dados Espaciais**. 2ª. ed. [s.l.] Mundo da Geomática, 2010.

ŠEDŠNKA, J.; GASTI, P. **Privacy-preserving distance computation and proximity testing on earth, done right**. ACM Press, 2014Disponível em: <http://dl.acm.org/citation.cfm?doid=2590296.2590307>. Acesso em: 7 abr. 2016

SHAREF, B. T.; ALSAQOUR, R. A.; ISMAIL, M. Vehicular communication ad hoc routing protocols: A survey. **Journal of Network and Computer Applications**, v. 40, p. 363–396, abr. 2014.

SINGHAL, M.; SHUKLA, A. Implementation of Location based Services in Android using GPS and Web Services. **International Journal of Computer Science Issues**, v. 9, n. 1, p. 237–242, jan. 2012.

STEENBRUGGEN, J.; TRANOS, E.; NIJKAMP, P. Data from mobile phone operators: A tool for smarter cities? **Telecommunications Policy**, v. 39, n. 3-4, p. 335–346, maio 2015.

ŞTEFĂNESCU, P. et al. Trip Planners Used in Public Transportation. Case Study on the City of Timişoara. **Procedia - Social and Behavioral Sciences**, v. 124, p. 142–148, mar. 2014.

SUN, J. (DANIEL); LIU, Q.; PENG, Z. Research and Analysis on Causality and Spatial-Temporal Evolution of Urban Traffic Congestions—A Case Study on Shenzhen of China. **Journal of Transportation Systems Engineering and Information Technology**, v. 11, n. 5, p. 86–93, out. 2011.

TANG, J. et al. Uncovering urban human mobility from large scale taxi GPS data. **Physica A: Statistical Mechanics and its Applications**, v. 438, p. 140–153, nov. 2015.

TANG, L.; THAKURIAH, P. (VONU). Ridership effects of real-time bus information system: A case study in the City of Chicago. **Transportation Research Part C: Emerging Technologies**, v. 22, p. 146–161, jun. 2012.

TOOLE, J. L. et al. The path most traveled: Travel demand estimation using big data resources. **Transportation Research Part C: Emerging Technologies**, v. 58, p. 162–177, set. 2015.

UNITED NATIONS; DEPARTMENT OF ECONOMIC AND SOCIAL AFFAIRS; POPULATION DIVISION. **World urbanization prospects: the 2014 revision : highlights**. [s.l: s.n.].

VLAHOGIANNI, E. I.; KARLAFTIS, M. G.; GOLIAS, J. C. Short-term traffic forecasting: Where we are and where we're going. **Transportation Research Part C: Emerging Technologies**, Special Issue on Short-term Traffic Flow Forecasting. v. 43, Part 1, p. 3–19, jun. 2014.

VON FERBER, C. et al. Public transport networks: empirical analysis and modeling. **The European Physical Journal B**, v. 68, n. 2, p. 261–275, mar. 2009.

WANG, W.; YANG, J.; MUNTZ, R. **STING : A Statistical Information Grid Approach to Spatial Data Mining**, 1997.

WHAIDUZZAMAN, M. et al. A survey on vehicular cloud computing. **Journal of Network and Computer Applications**, v. 40, p. 325–344, abr. 2014.

XIA, D. et al. A distributed spatial–temporal weighted model on MapReduce for short-term traffic flow forecasting. **Neurocomputing**, v. 179, p. 246–263, fev. 2016.

XINGHAO, S. et al. Predicting Bus Real-time Travel Time Basing on both GPS and RFID Data. **Procedia - Social and Behavioral Sciences**, v. 96, p. 2287–2299, nov. 2013.

XIN, J.; CHEN, S. Bus Dwell Time Prediction Based on KNN. **Procedia Engineering**, v. 137, p. 283–288, 2016.

XU, F. et al. Assessing the Impact of Rainfall on Traffic Operation of Urban Road Network. **Procedia - Social and Behavioral Sciences**, v. 96, p. 82–89, nov. 2013.

XU, X. et al. **A Distribution-Based Clustering Algorithm for Mining in Large Spatial Databases**, 1998.

YONG-CHUAN, Z. et al. Traffic Congestion Detection Based On GPS Floating-Car Data. **Procedia Engineering**, v. 15, p. 5541–5546, 2011.

ZHAO, Q. et al. A grid-growing clustering algorithm for geo-spatial data. **Pattern Recognition Letters**, v. 53, p. 77–84, fev. 2015.

ZHONG, C. et al. Inferring building functions from a probabilistic model using public transportation data. **Computers, Environment and Urban Systems**, v. 48, p. 124–137, nov. 2014.

ZHOU, H.; HIRASAWA, K. Traffic Density Prediction with Time-Related Data Mining using Genetic Network Programming. **The Computer Journal**, v. 57, n. 9, p. 1395–1414, 1 set. 2014.

ZULAR, A. et al. A UTILIZAÇÃO DO GOOGLE FUSION TABLES PARA COMPARTILHAR E INTEGRAR DADOS DE SISTEMAS DEPOSICIONAIS QUATERNÁRIOS COSTEIROS. 2011.