Processando 6 Bilhões de Posições GPS de Ônibus da Cidade do Rio de Janeiro

Leonardo Murta
leomurta@ic.uff.br

Baseado nos slides de Marcos Lage e Bruno Schettino

Instituto de Computação
uff

– Contexto –

Grandes cidades

Vivemos em cidades

Em torno de

50%

da população mundial
vive em regiões urbanas

United Nations 2012, “World Urbanization Prospects: The 2011 Revision”

As cidades estão crescendo

Até 2050,

70%

da população mundial viverá em regiões urbanas

United Nations 2012, "World Urbanization Prospects: The 2011 Revision"

Rio é um dos maiores centros urbanos do Brasil

No Brasil,
a região mais urbanizada é
a Região Sudeste.

Rio é um dos maiores centros urbanos do Brasil

Rio de Janeiro

6,5M

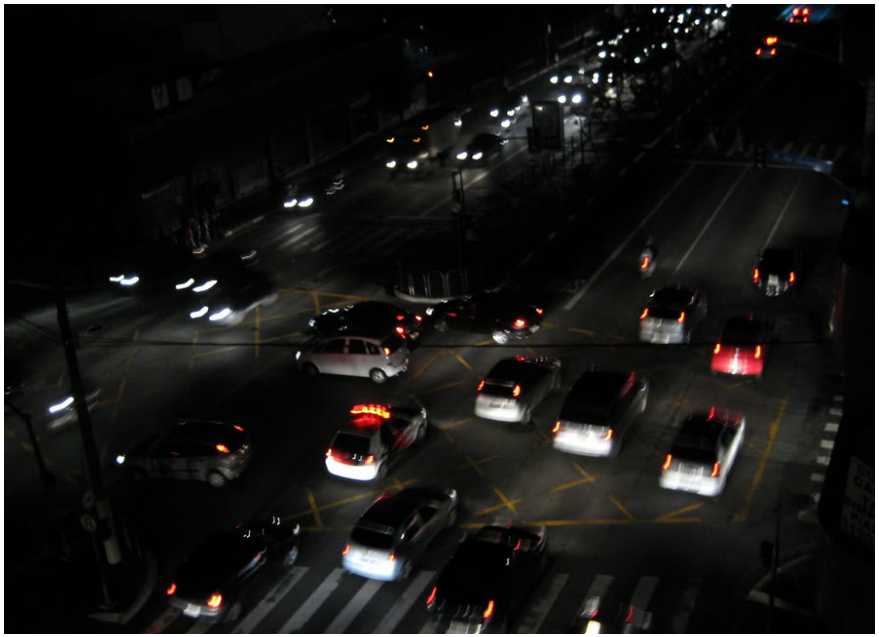
de habitantes!

http://www.cidades.ibge.gov.br

Expectativas

Eficiência, resiliência e sustentabilidade

Transporte



Eletricidade

Expectativas

Eficiência, resiliência e sustentabilidade

Deslizamentos

Expectativas

Eficiência, resiliência e sustentabilidade

Reciclagem

– Oportunidade 1 –

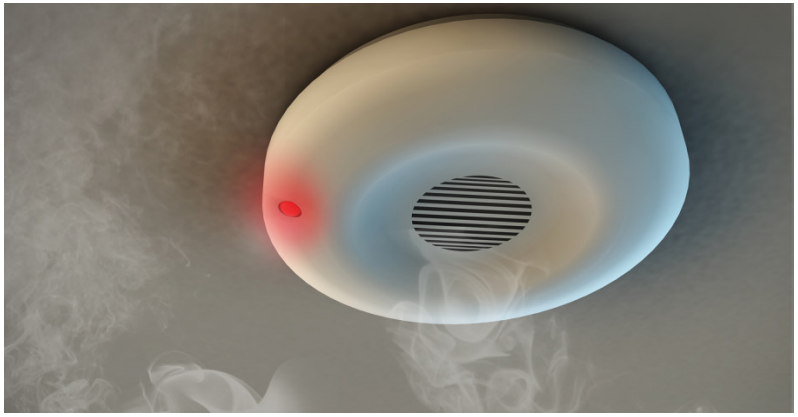
Grande disponibilidade de dados

Câmeras

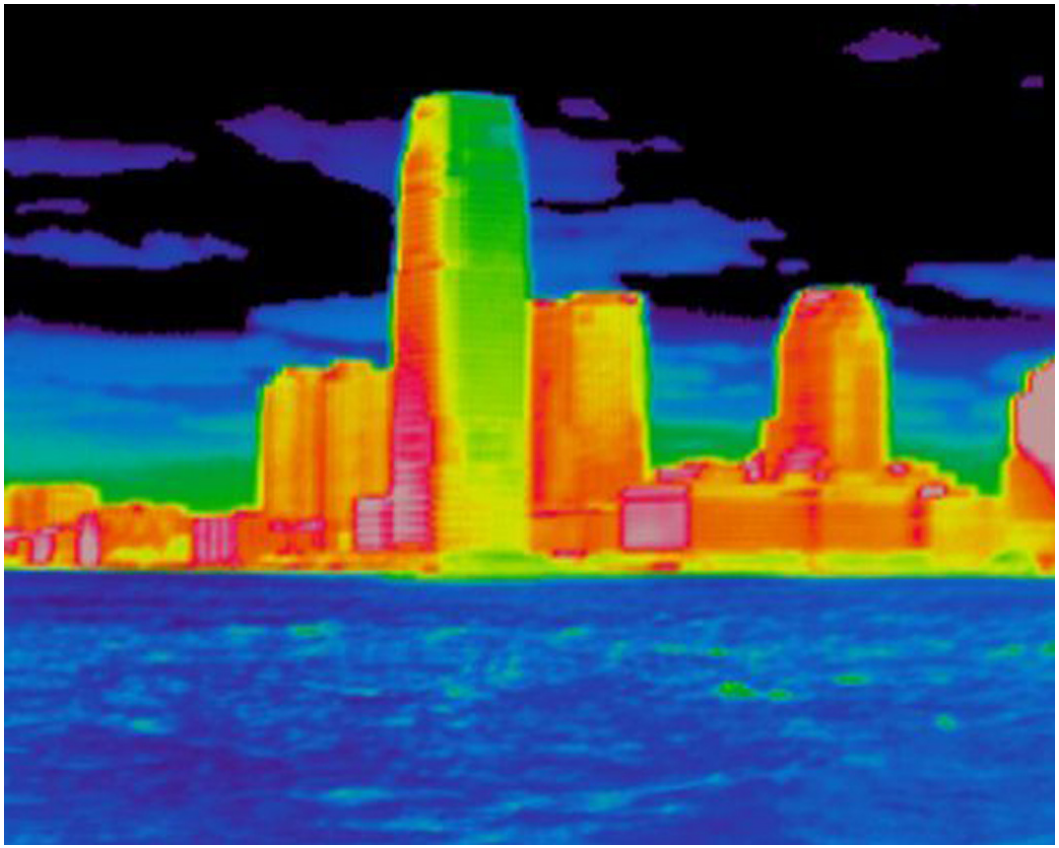
560

Câmeras de segurança
na cidade do Rio de Janeiro

Presença e fumaça



Temperatura e ruído

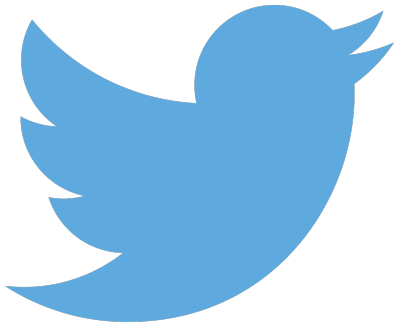


Celulares

271M

Celulares habilitados no Brasil em 2013 (ANATEL)

Redes sociais



76M

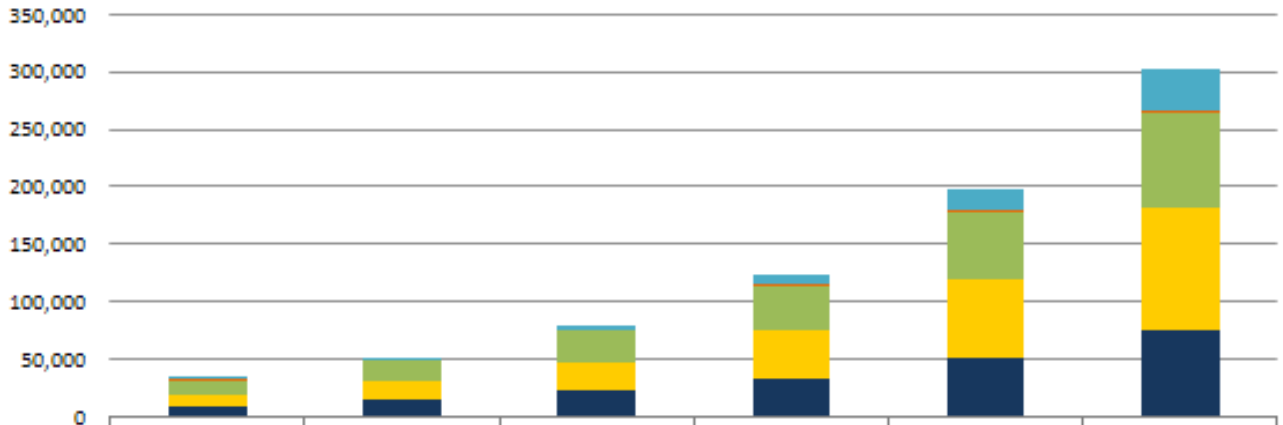
de brasileiros ativos no Facebook

– Oportunidade 2 –

Grande disponibilidade computacional

Capacidade crescente de armazenamento

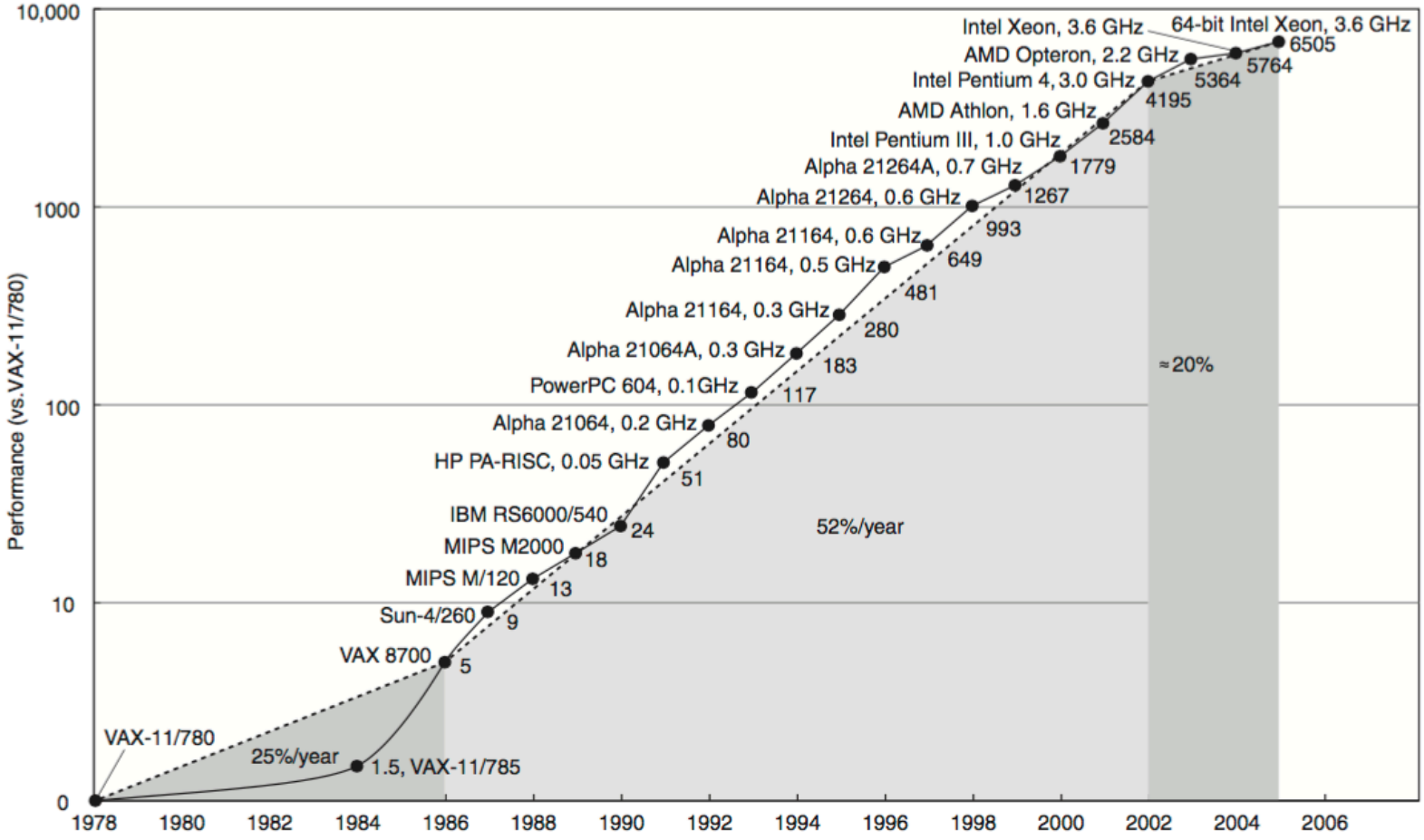
Total Worldwide Digital Archive Capacity, by Media Type, 2010-2015 (Petabytes)



	2010	2011	2012	2013	2014	2015
Cloud	768	1,708	3,676	8,033	17,908	37,846
Optical	303	410	537	714	968	1,247
Tape	12,784	18,939	27,116	39,362	58,220	81,562
External disk	9,712	16,053	25,637	41,501	68,430	106,830
Internal disk	9,650	14,881	22,185	33,547	51,707	75,510

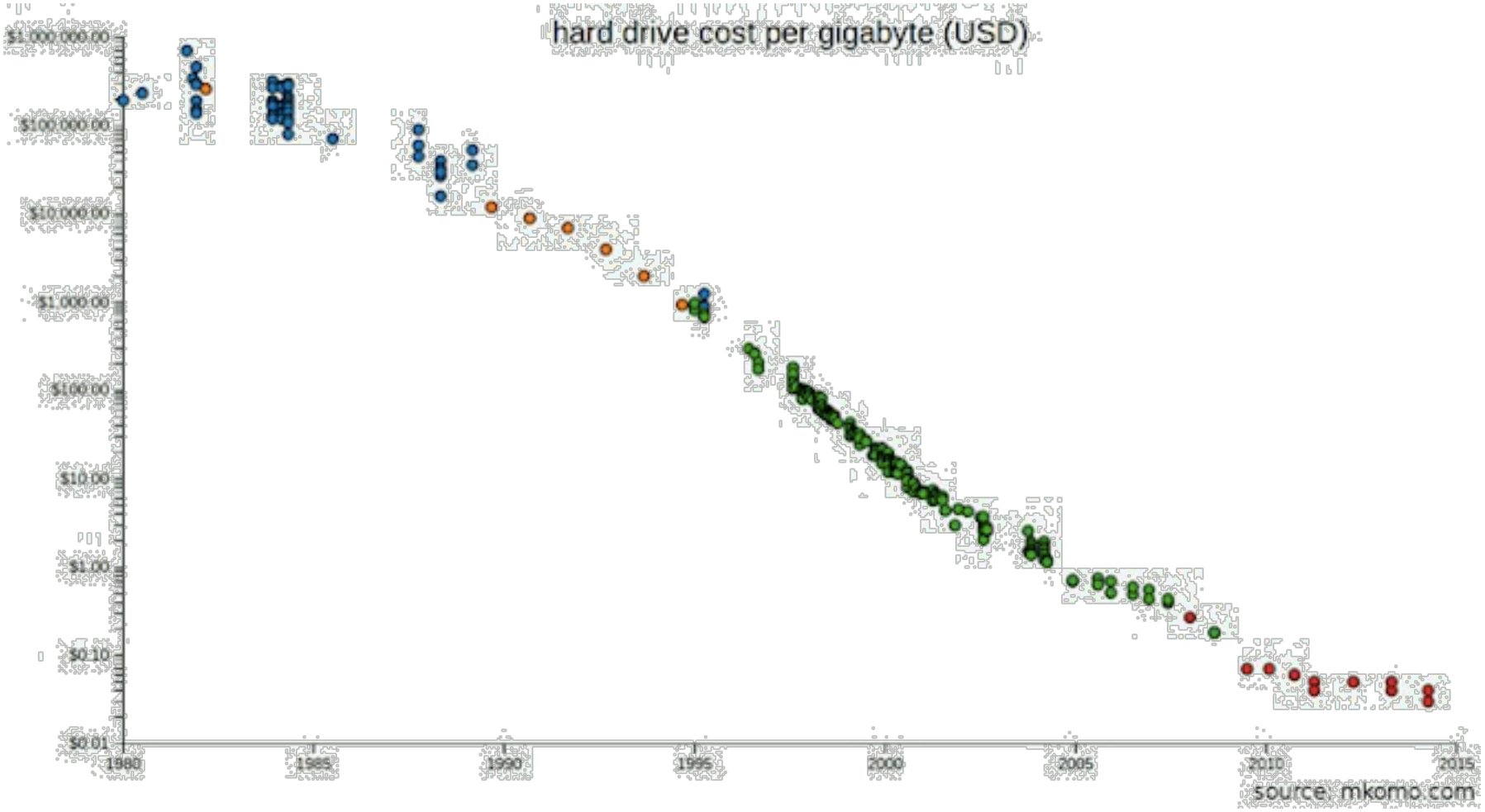
Source: Enterprise Strategy Group

Capacidade crescente de processamento



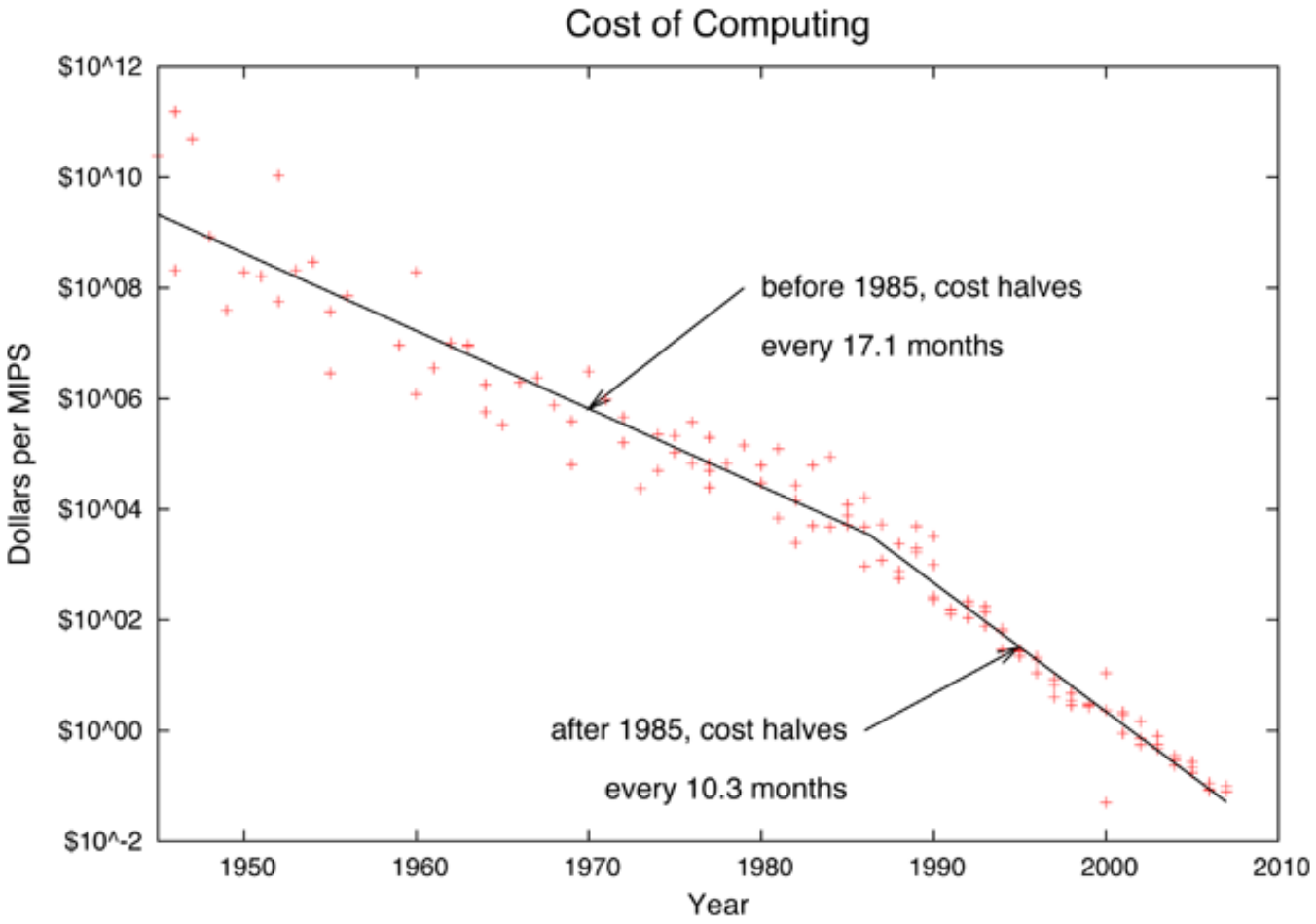
http://www.cs.columbia.edu/~sedwards/classes/2012/3827-spring/advanced-arch-2011.pdf

Preço decrescente de armazenamento



http://www.mkomo.com/cost-per-gigabyte-update

Preço decrescente de processamento



http://www.frc.ri.cmu.edu/users/hpm/book97/ch3/processor.list.txt

Supercomputador #1 em 11/1996 *versus* PC doméstico #1 em 10/2015

CP-PACS, 2.048 núcleos @ 150 MHz
128 GB memória, 529 GB disco

368,2 Gflops
$ 22 milhões

Core i7 5960X, 8 núcleos @ 3.0 GHz
32 GB memória, 2 TB disco

354 Gflops
$ 3 mil

Supercomputador #1 em 10/2015

100 mil vezes mais rápido que o Core i7 5960X

Tianhe-2, 3.120.000 núcleos @ 2,4 GHz
1,4 PB memória, 12,4 PB disco

Obs.: 1 Peta = 1.000 Tera = 1.000.000 Giga

– Possibilidades –

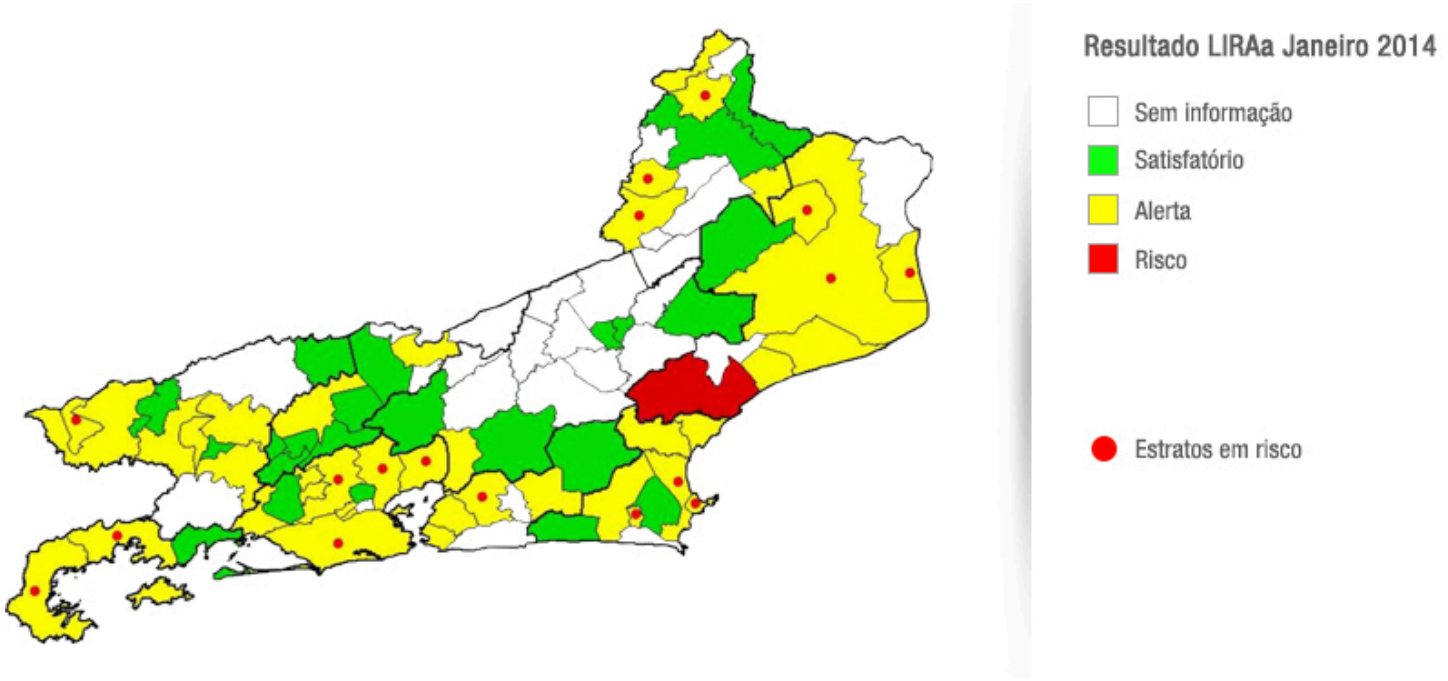
Aplicações para cidades inteligentes

Monitoramento e controle

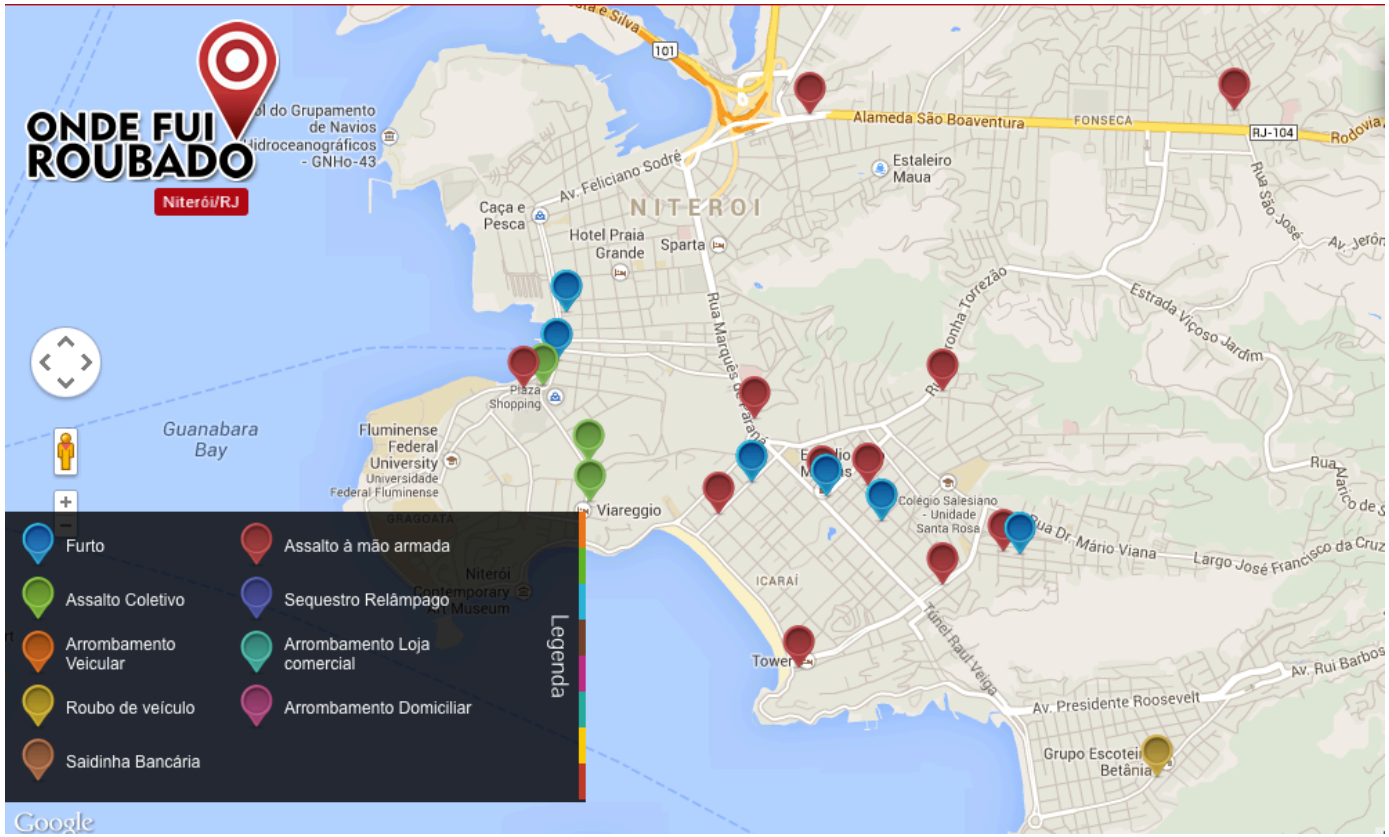
Monitoramento de estruturas

Planejamento de novas construções

Monitoramento de epidemias



Mapa de criminalidade



Análise de rotas de taxis

– Estudo de caso –

Análise de rotas de ônibus na cidade do Rio de Janeiro

Por que analisar o trânsito do Rio?

A cidade do Rio de Janeiro tem o 30º pior trânsito do mundo...

... cariocas perdem em torno de 100 horas no trânsito por ano em viagens diárias de 30 minutos (que demandam 56 minutos).

http://www.tomtom.com/en_gb/trafficindex/#/city/RIO

Por que analisar ônibus?

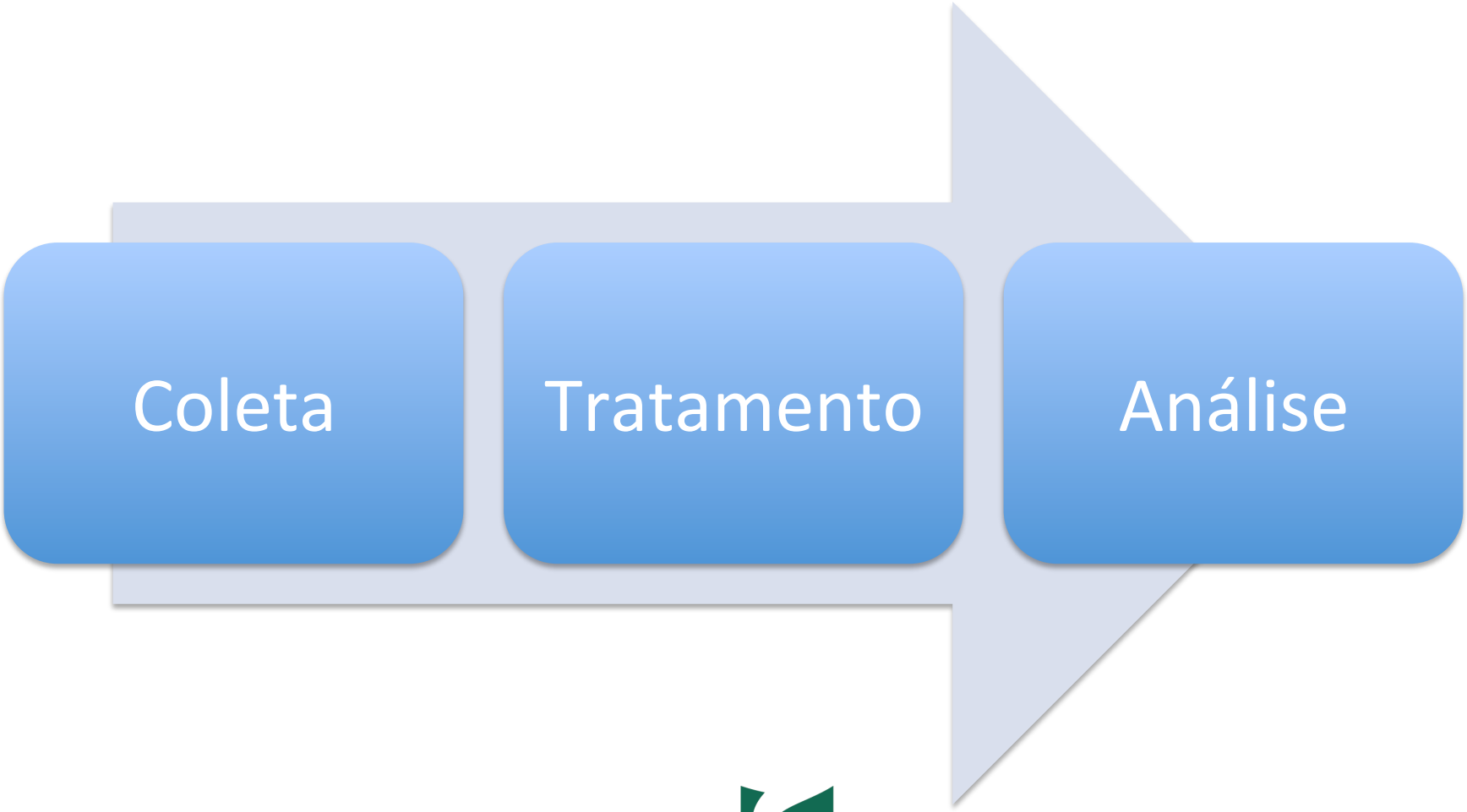
60 carros = 1 ônibus

Pôster feito pelo departamento de planejamento da cidade de Münster na Alemanha

Por que analisar ônibus?

Entender o **trafego de ônibus** permite entender indiretamente o **trânsito da cidade**

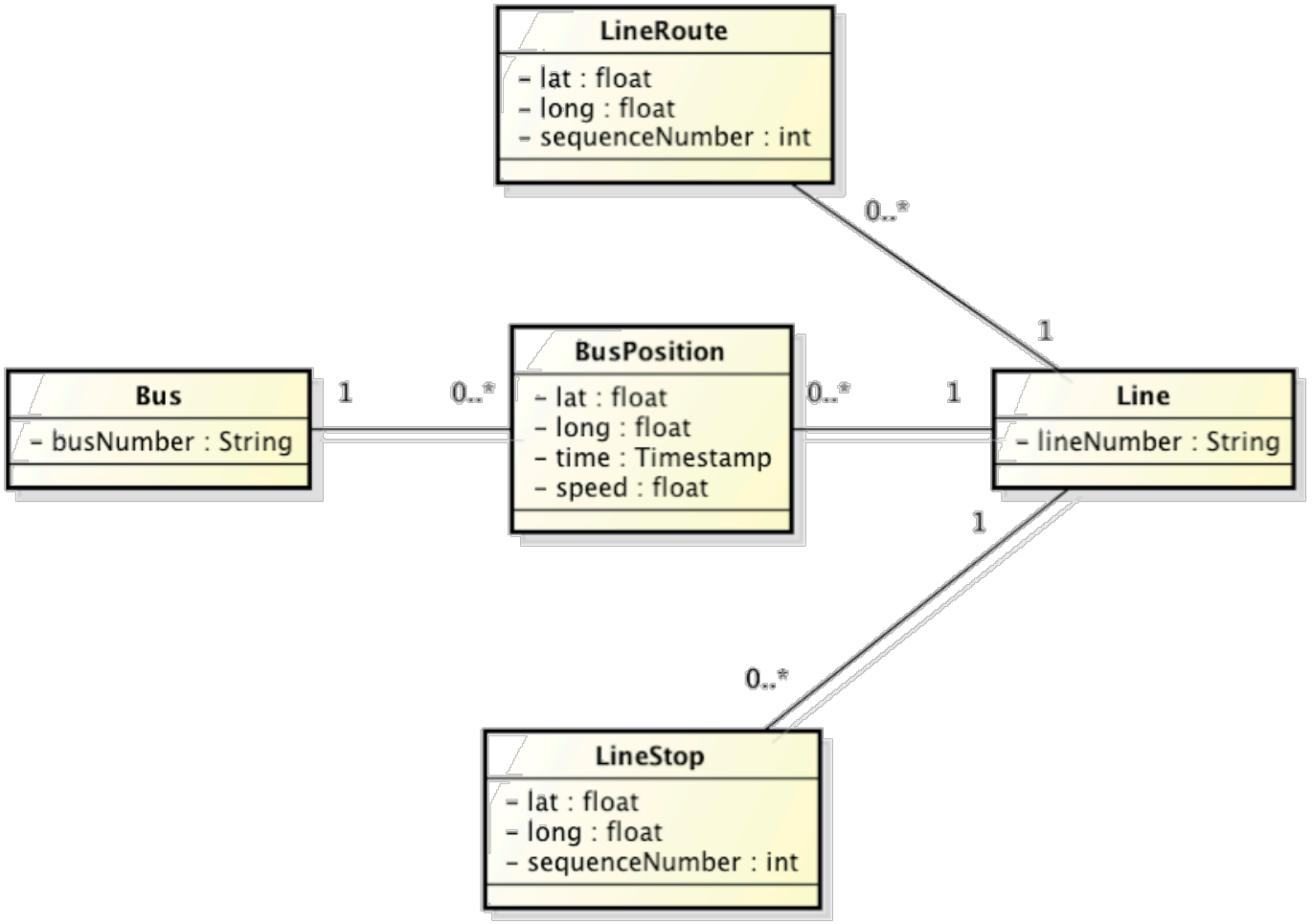
Processo



Coleta

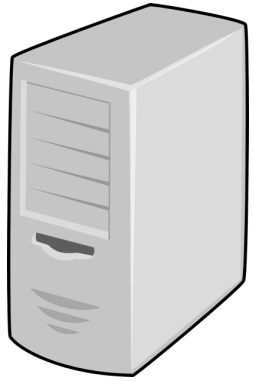
- Base de dados aberta com
  - Trajeto das linhas
  - Pontos de parada das linhas
  - GPS dos ônibus

Coleta

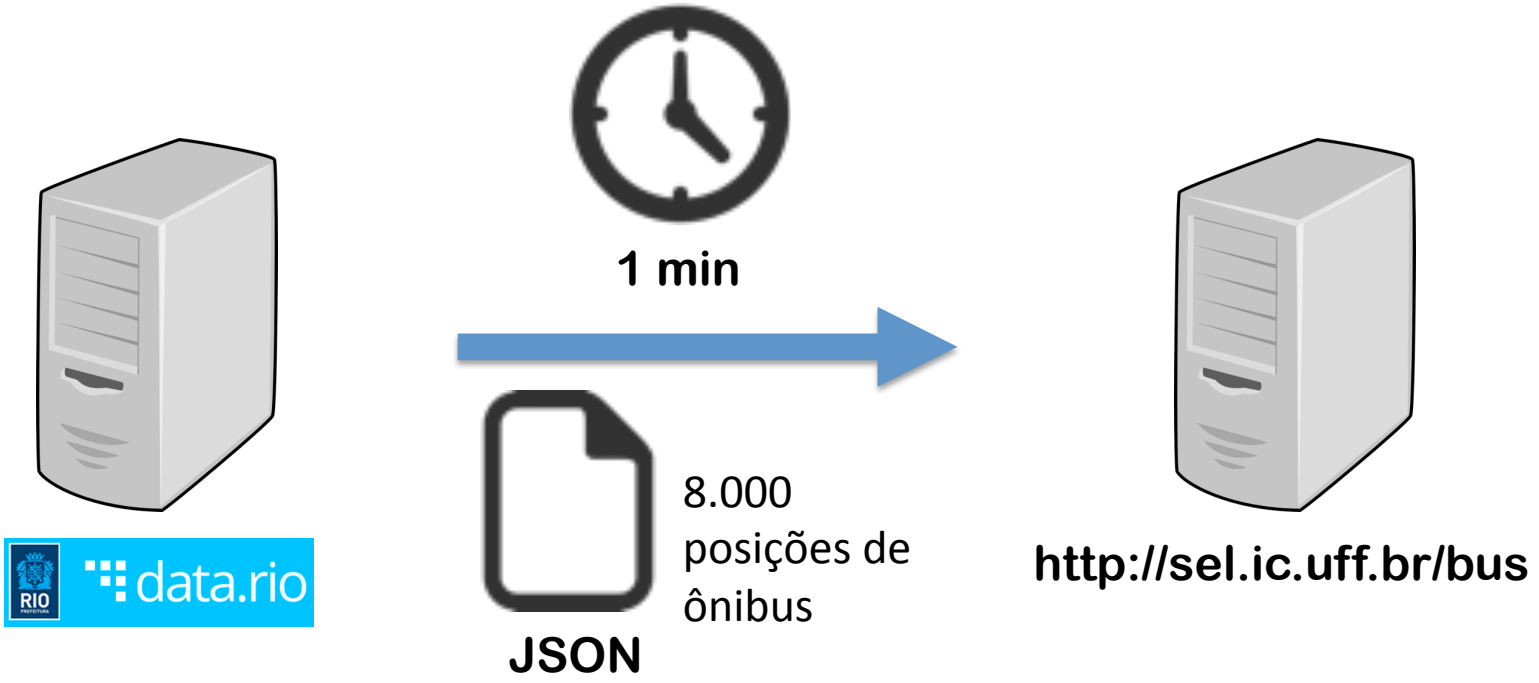

LineRoute
- lat : float
- long : float
- sequenceNumber : int
0..*
1
Bus
- busNumber : String
1
0..*
BusPosition
- lat : float
- long : float
- time : Timestamp
- speed : float
0..*
1
Line
- lineNumber : String
1
0..*
LineStop
- lat : float
- long : float
- sequenceNumber : int

Coleta

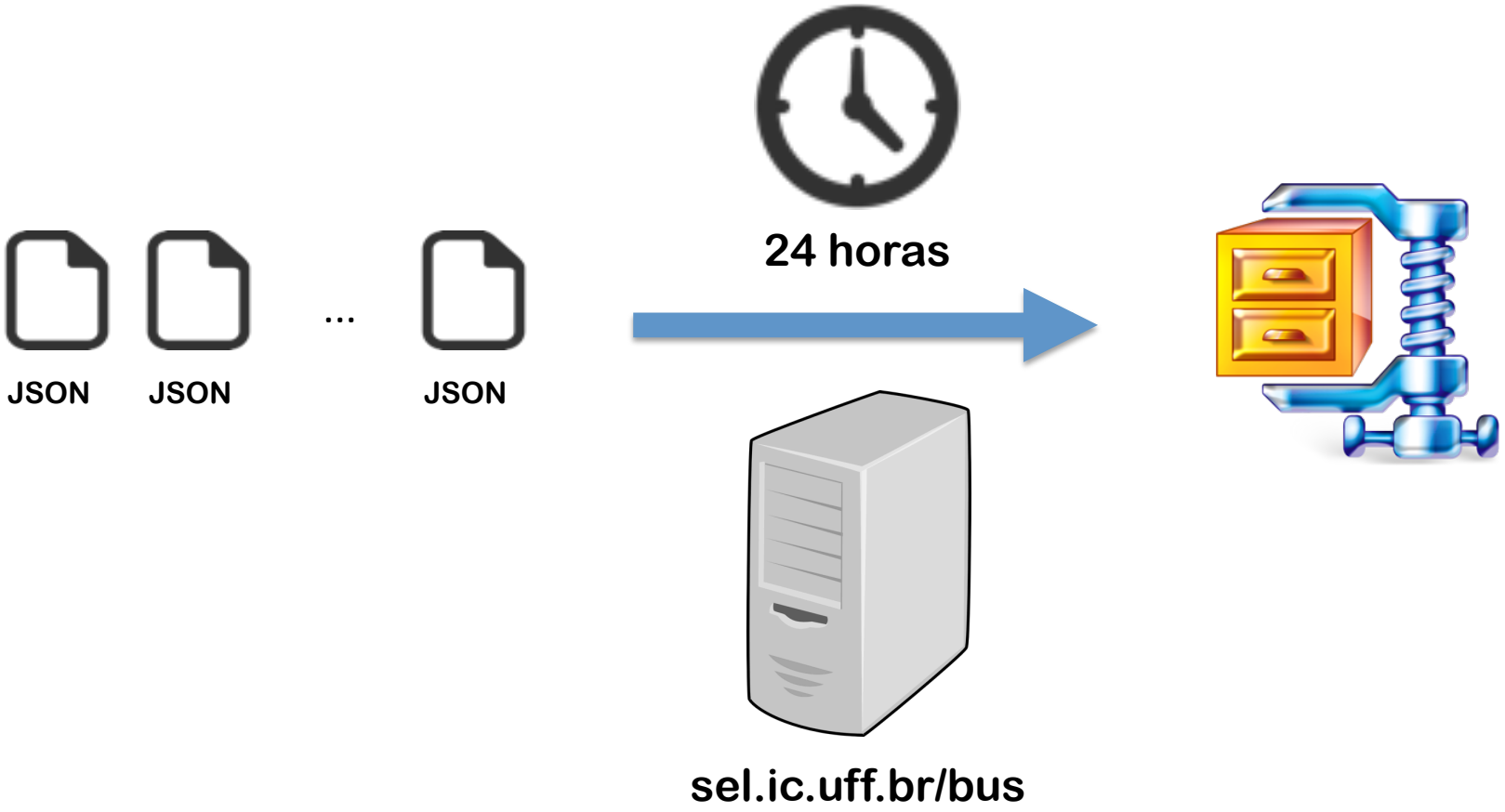
8.000 ônibus



Coleta


1 min
data.rio
8.000
posições de
ônibus
JSON
http://sel.ic.uff.br/bus

Coleta



Números da Coleta

- Período de coleta (1,5 anos)
  - 16/4/2014 a 8/10/2015
- 1 minuto
  - 8 mil posições GPS
  - 450 KB
- 1 dia
  - 11,5 milhões de posições GPS
  - 650 MB
- 1,5 anos
  - 6,2 bilhões de posições GPS
  - 350 GB

Tratamento

Importação dos zips para BD
e filtragem de dados inválidos

BD
(PostGIS)

Tratamento
1ª tentativa

- Ruby on Rails
- 1 JSON por minuto
  - Índices atualizados a cada *commit*
  - Construção do JSON em memória
- Levaria 1,5 anos para processar tudo

Tratamento
implementação atual

- Java
- 300 JSON por minuto
  - Índices desligados e reconstruídos ao final
  - Processamento do JSON por *stream*
  - Otimização das estruturas de dados
  - Reuso de objetos em memória
- Leva 2 dias para processar tudo

Tratamento filtragens

- Registro repetido
- Velocidade incompatível
- Sem linha
- Parado na garagem
- Fora da rota

Tratamento
registro repetido

- Últimos 10 registros de cada ônibus armazenado em memória
- Comparação rápida entre um novo registro e o últimos registros do ônibus (*timestamp*)

Tratamento velocidade incompatível

- Detecção de outliers
  - 3º quartil (valor aos 75% da amostra ordenada) somado a 1,5 da distância interquartil (distância entre o valor aos 25% e aos 75% da amostra ordenada): Q3 + 1.5 (Q3 - Q1)
  - Velocidade de corte: 85,57 km/h
- Critérios de descarte
  - Velocidade instantânea
  - Velocidade média considerando a distância euclidiana percorrida entre duas posições GPS e o tempo gasto
  - [estamos implementando o calculo da distância projetada na rota]

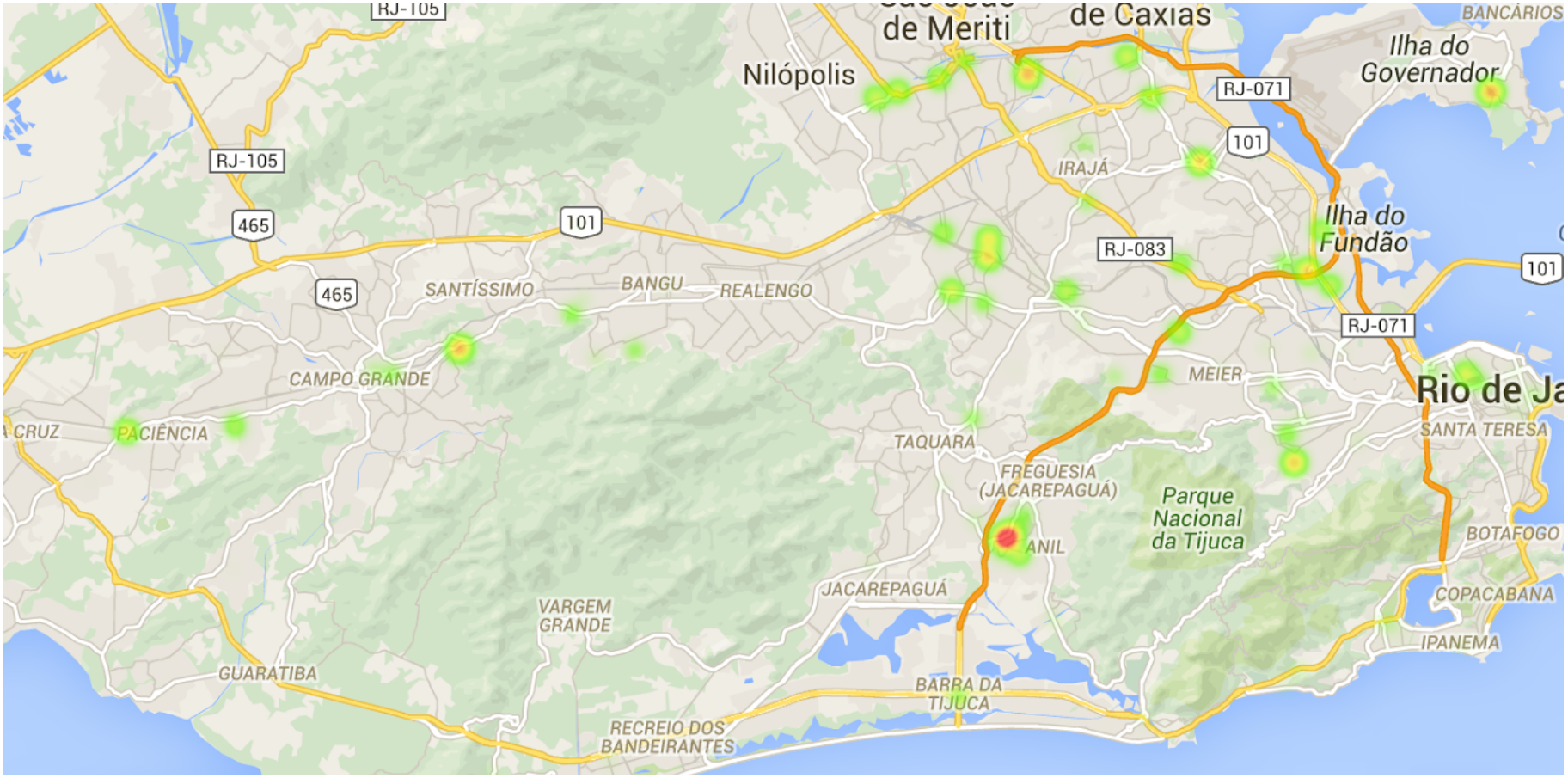
Tratamento sem linha

- Registro não contém informação da linha na qual está circulando
- [estamos implementando a detecção automática da linha do ônibus]

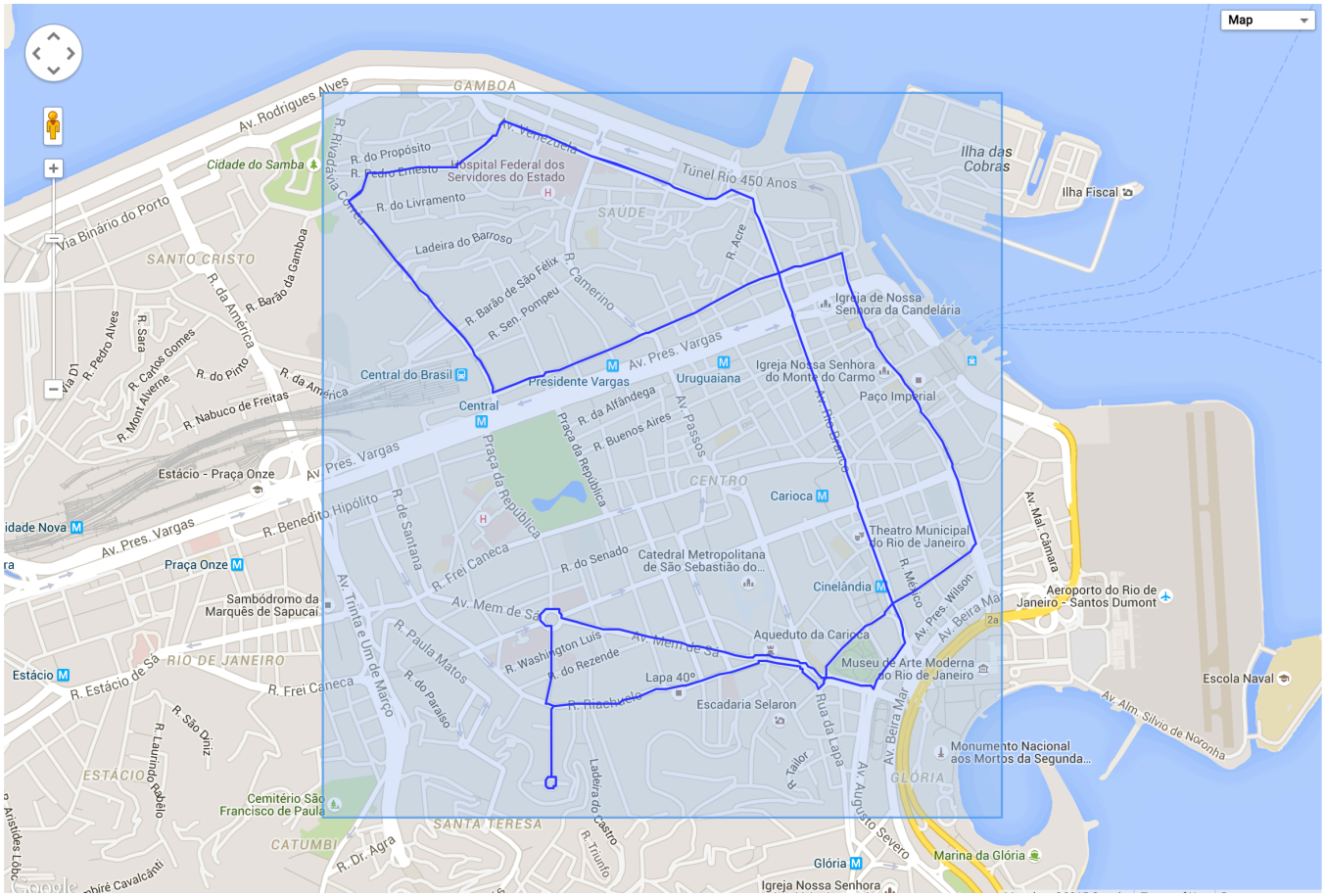
Tratamento garagem/fora da rota

- Detecção manual das garagens
  - Análise de clusters de ônibus com 0 Km/h de madrugada
  - [estamos implementando detecção automática de garagens]
- Detecção automática de deslocamento entre as garagens e a linha
  - Bounding box ao redor da linha e descarte de posições fora da bounding box

Tratamento garagem



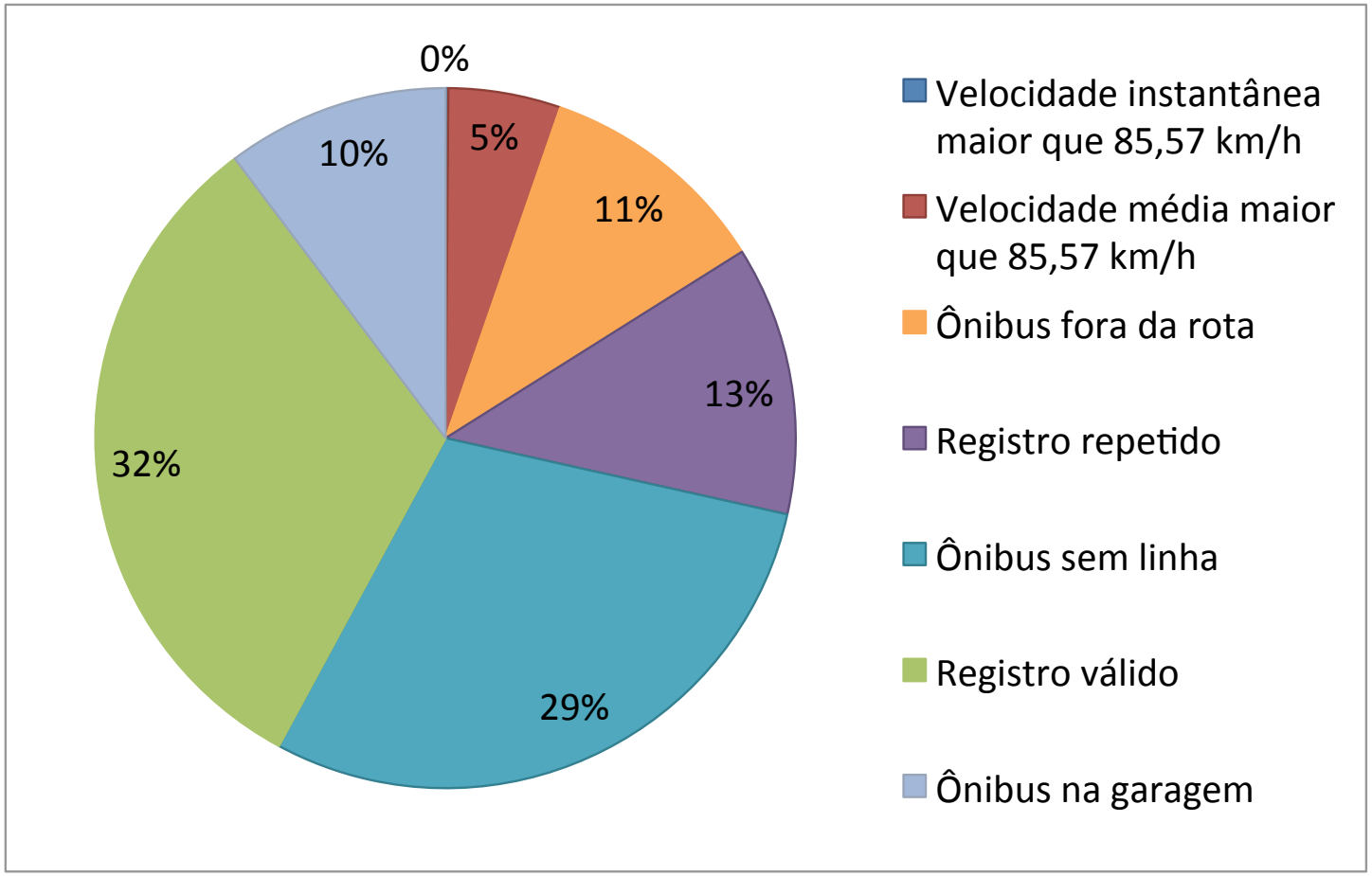
Tratamento fora da rota



Análise
questões de pesquisa

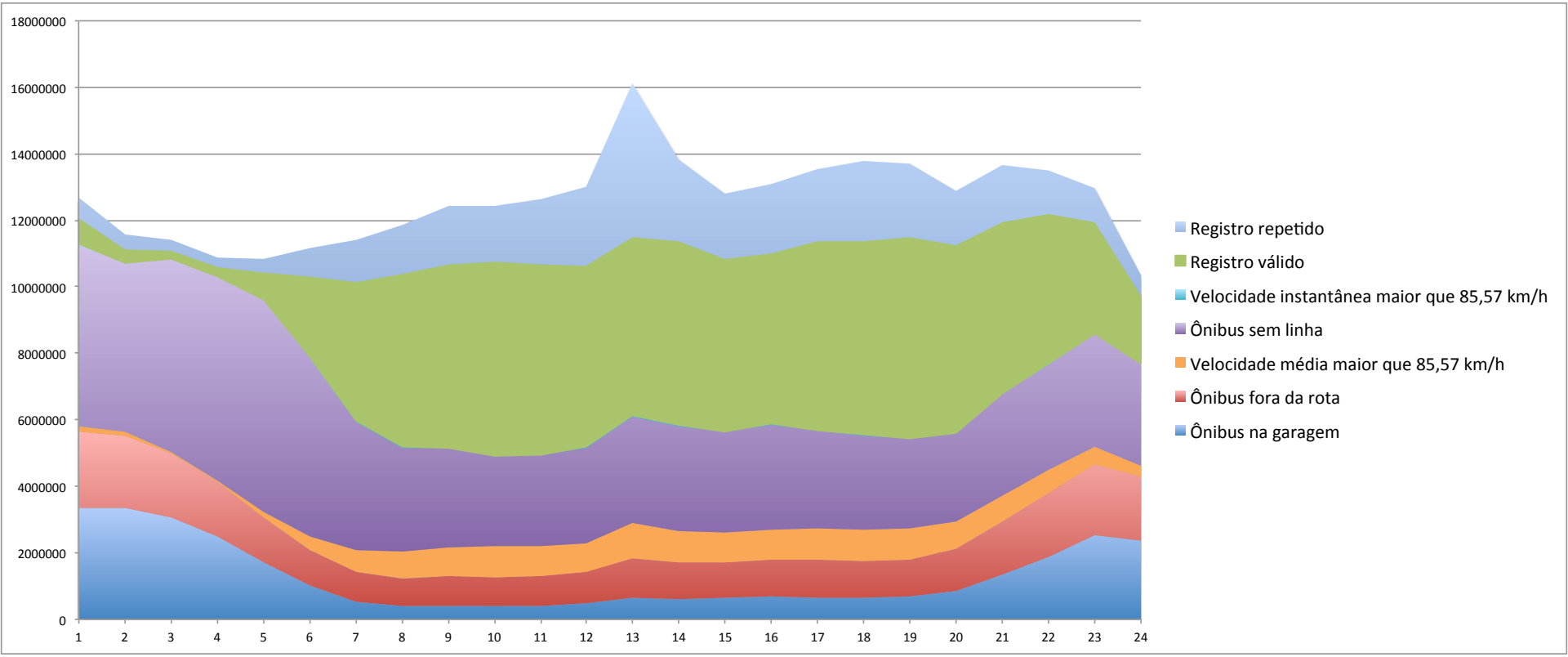
- RQ1: Há dados inválidos em que proporção?
- RQ2: Como os dados inválidos variam durante o dia?
- RQ3: Como as condições do trânsito variam nos diferentes dias da semana?
- RQ4: Como as condições do trânsito variam nas diferente horas do dia?
- RQ5: Como as condições do trânsito se comparam em diferentes dias da semana?
- RQ6: Como as condições do trânsito variam em diferentes meses?
- RQ7: Como são as condições do trânsito em dias de feriados?
- RQ8: Como são as condições do trânsito em dias de grandes eventos?

Análise – RQ1: Há dados inválidos em que proporção?



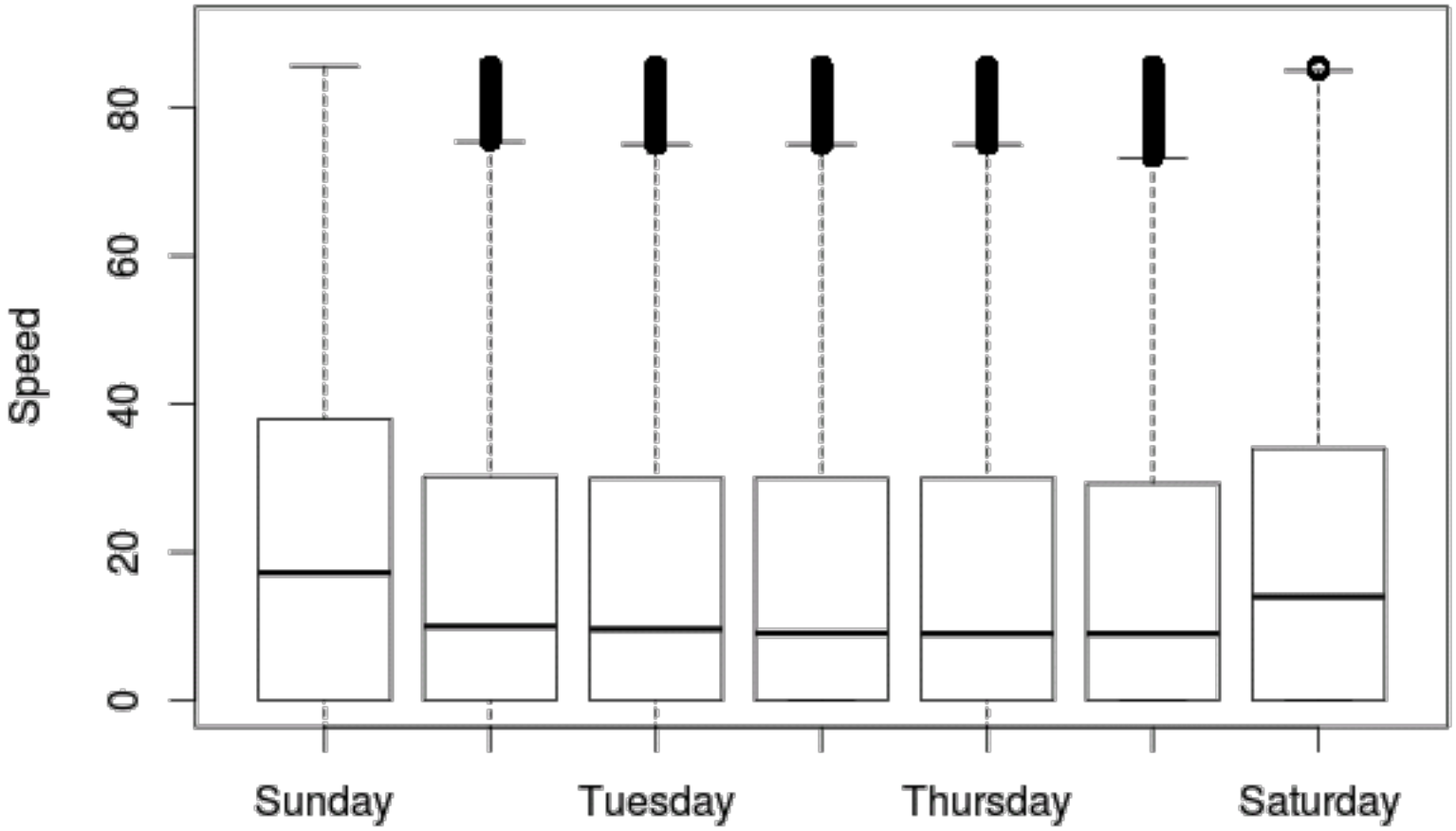
Amostra de 01/05/2015 a 04/06/2015

Análise – RQ2: Como os dados inválidos variam durante o dia?

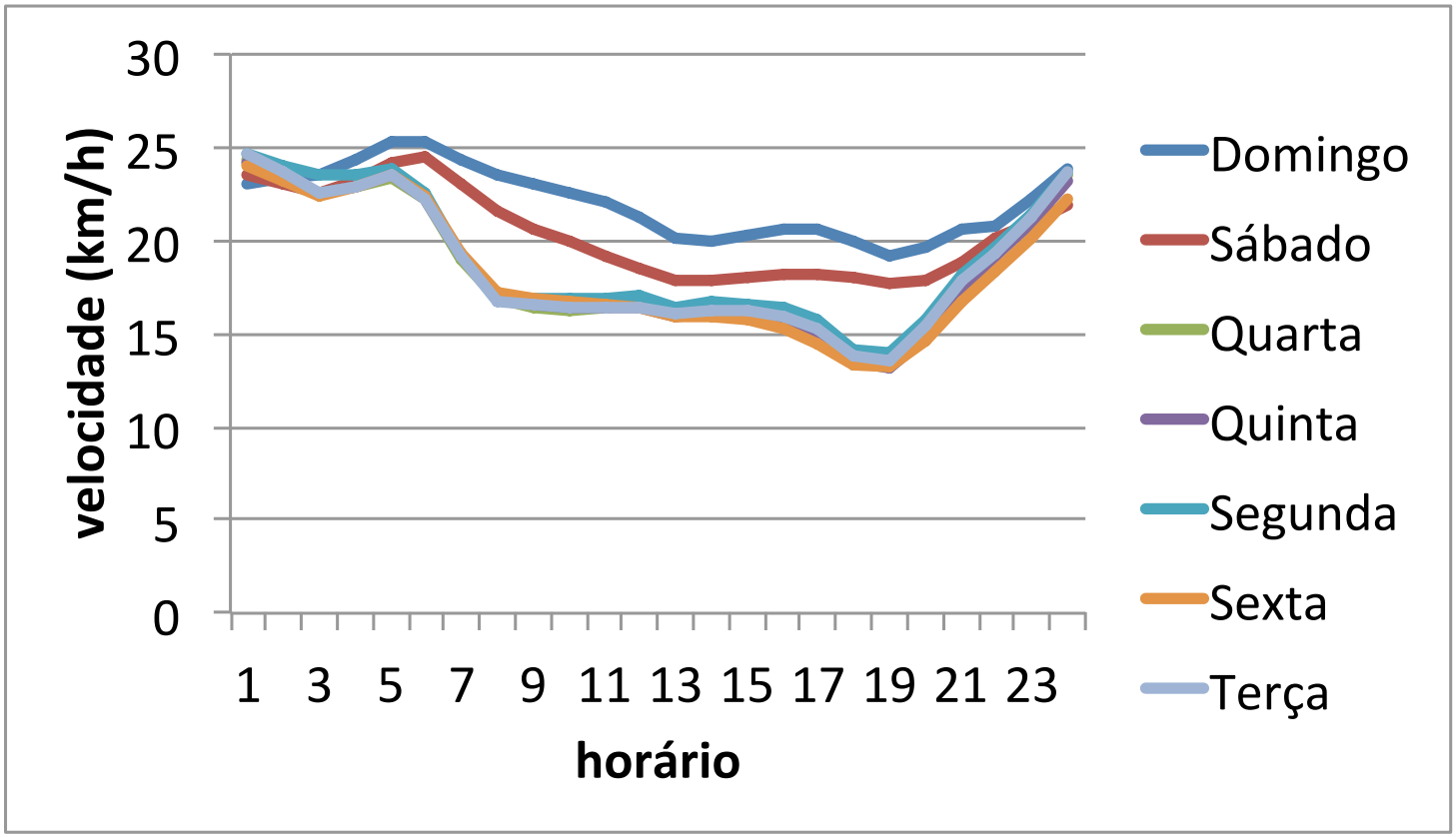


Amostra de 01/05/2015 a 04/06/2015

Análise - RQ3: Como as condições do trânsito variam nos diferentes dias da semana?



Análise – RQ4: Como as condições do trânsito variam nas diferente horas do dia?

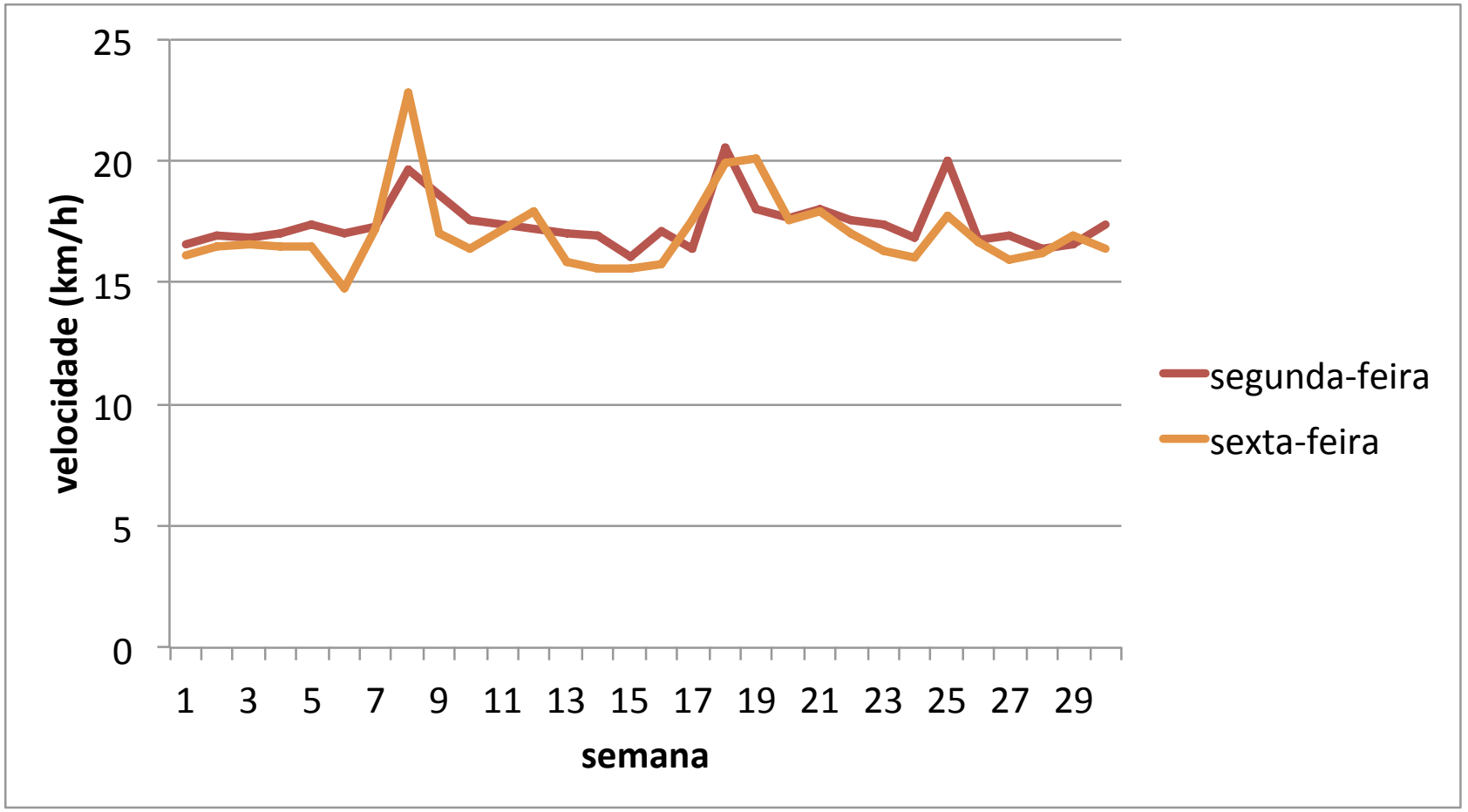


Análise – RQ5: Como as condições do trânsito se comparam em diferentes dias da semana?

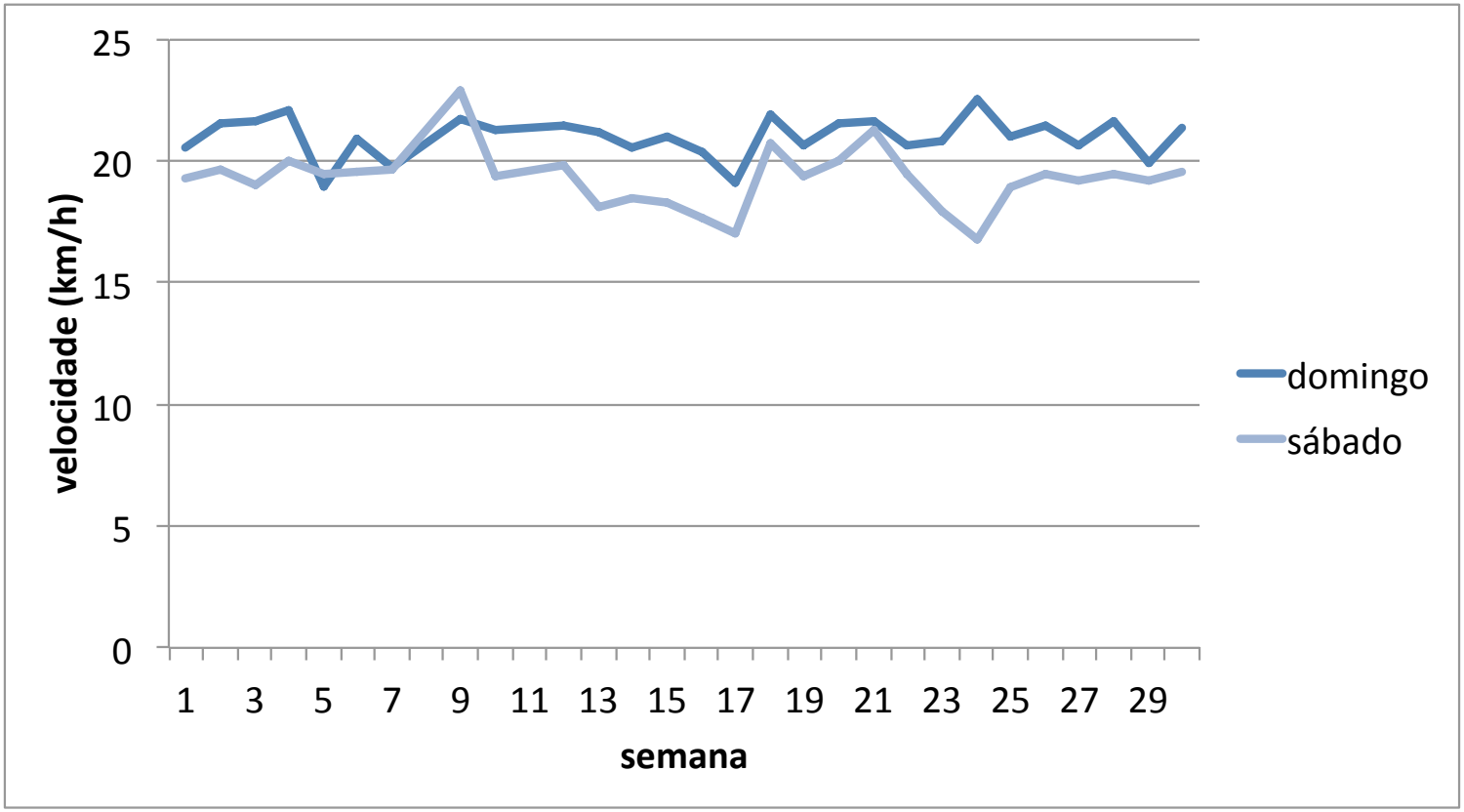
Correlação

	Domingo	Segunda	Terça	Quarta	Quinta	Sexta
Segunda	0,19					
Terça	0,20	0,45				
Quarta	-0,12	0,62	0,50			
Quinta	0,02	0,47	0,14	0,26		
Sexta	0,05	0,73	0,13	0,27	0,42	
Sábado	0,28	0,42	0,42	0,23	0,07	0,35

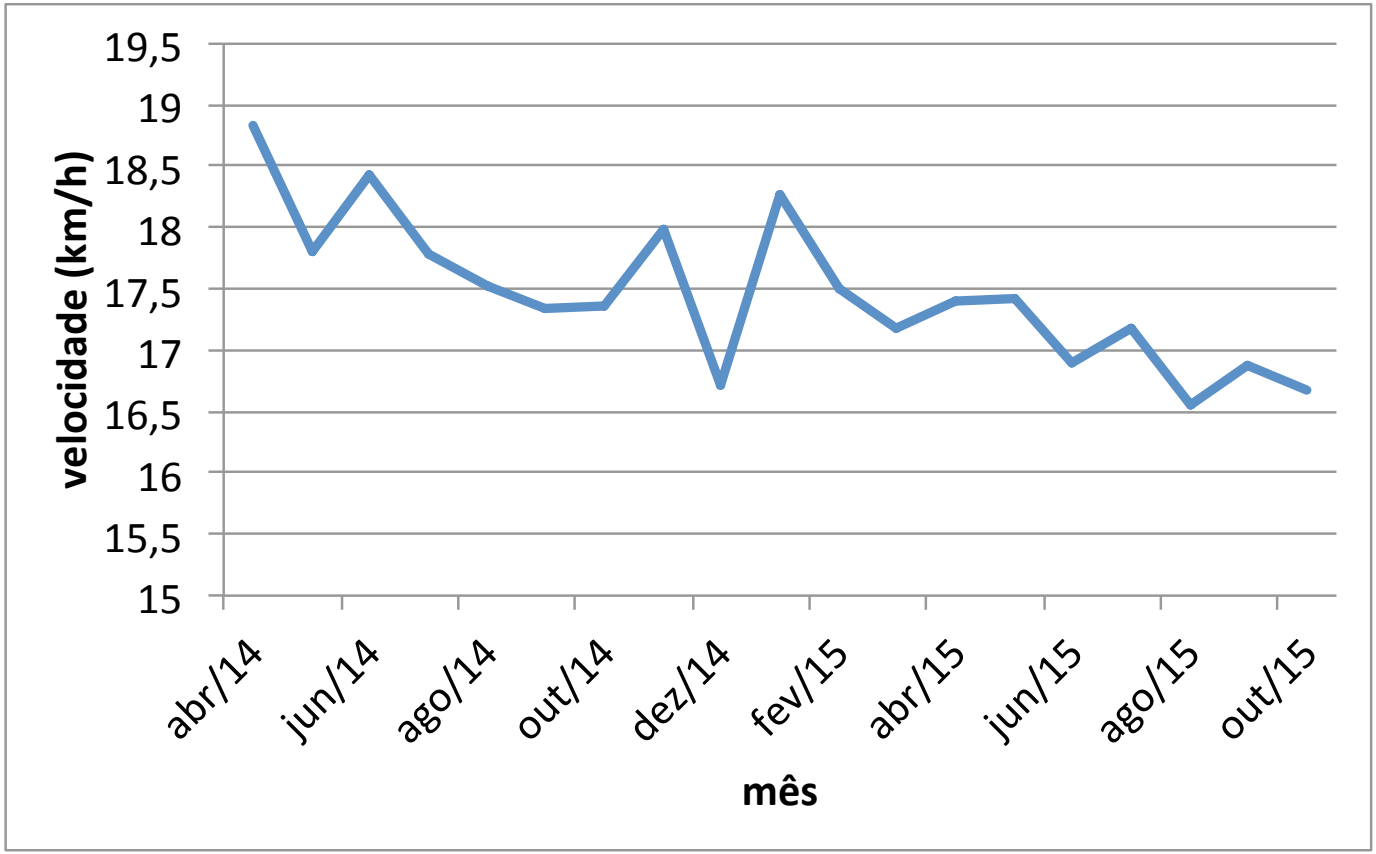
Análise – RQ5: Como as condições do trânsito se comparam em diferentes dias da semana?



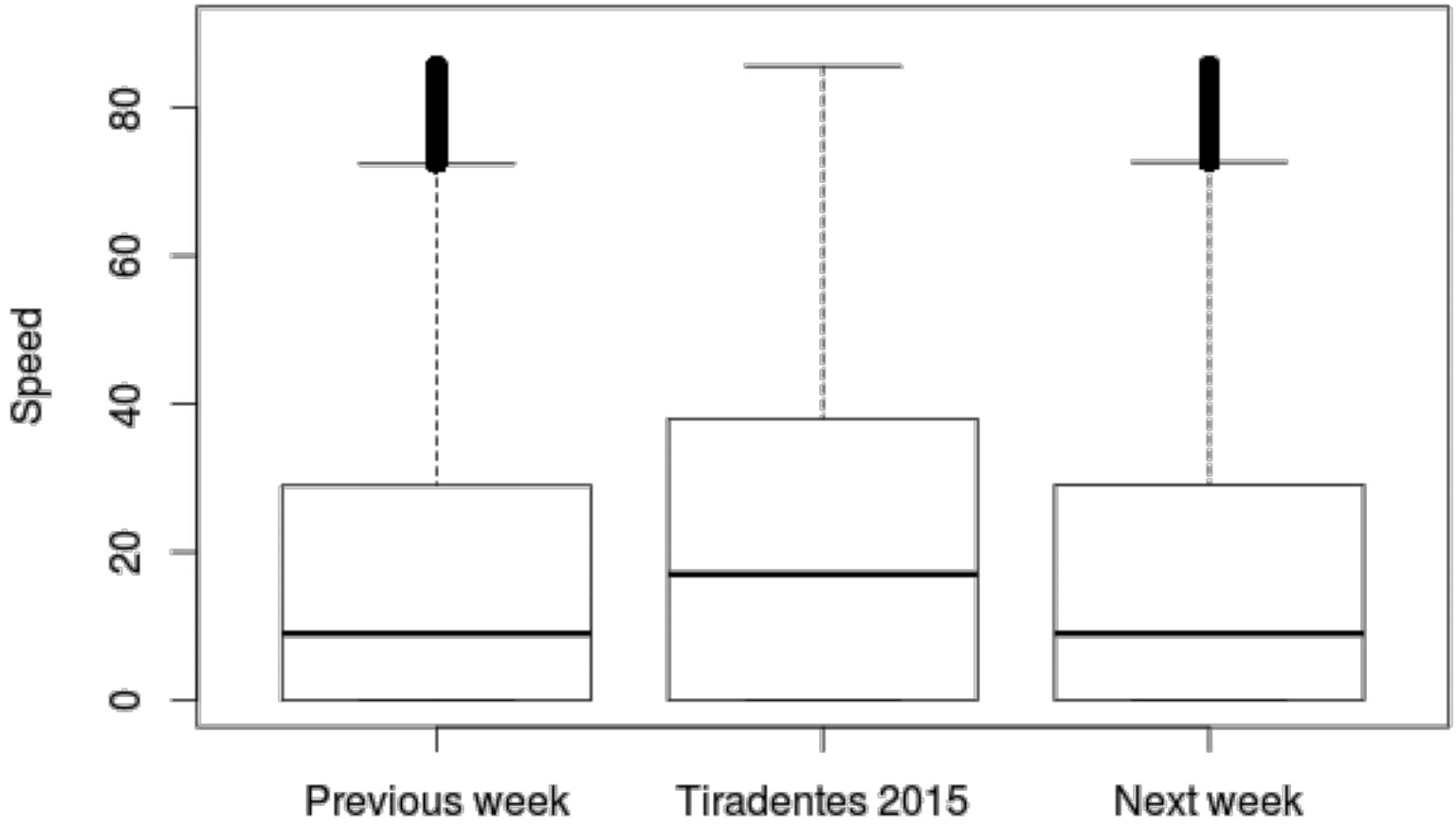
Análise – RQ5: Como as condições do trânsito se comparam em diferentes dias da semana?



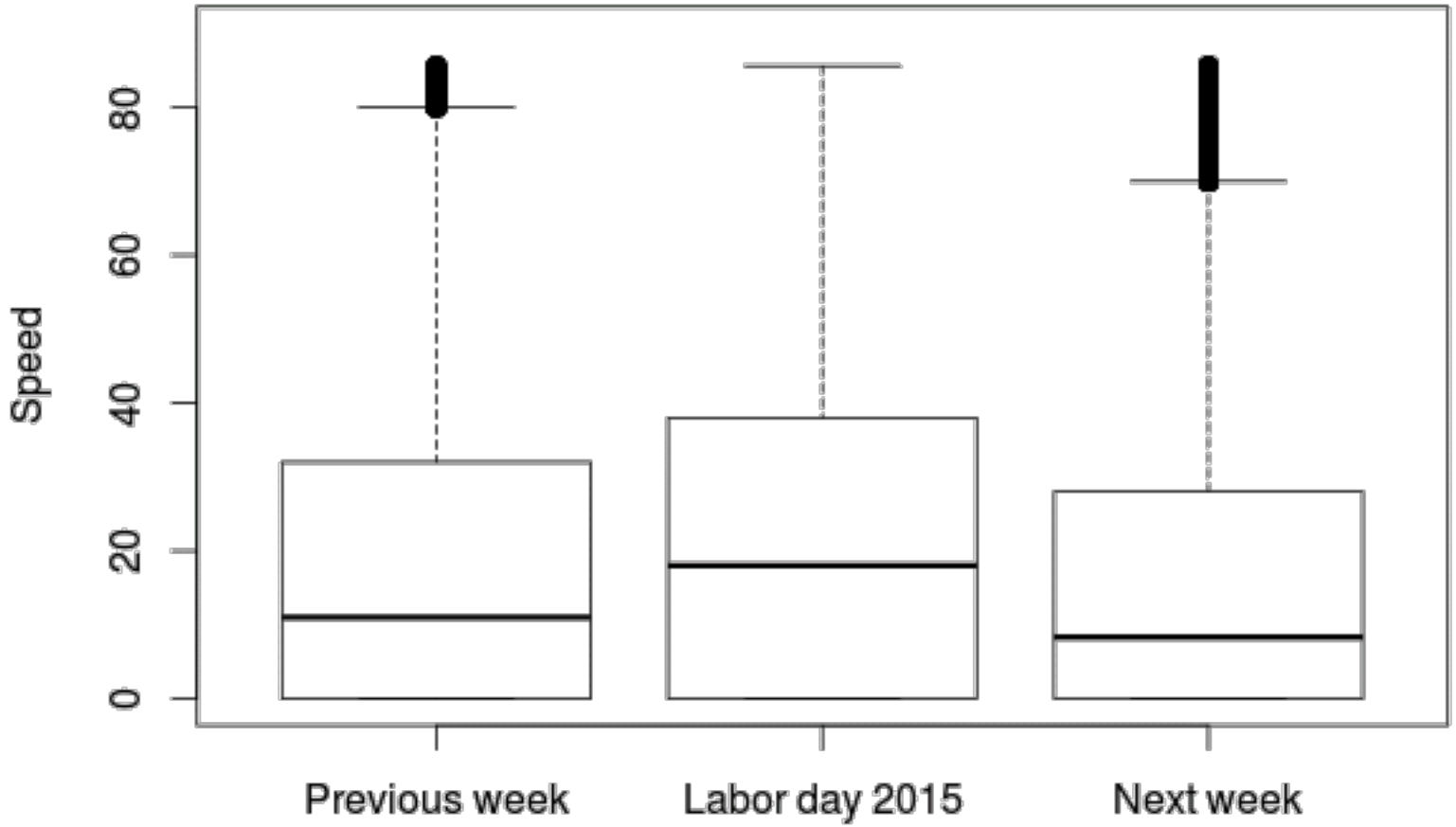
Análise – RQ6: Como as condições do trânsito variam em diferentes meses?



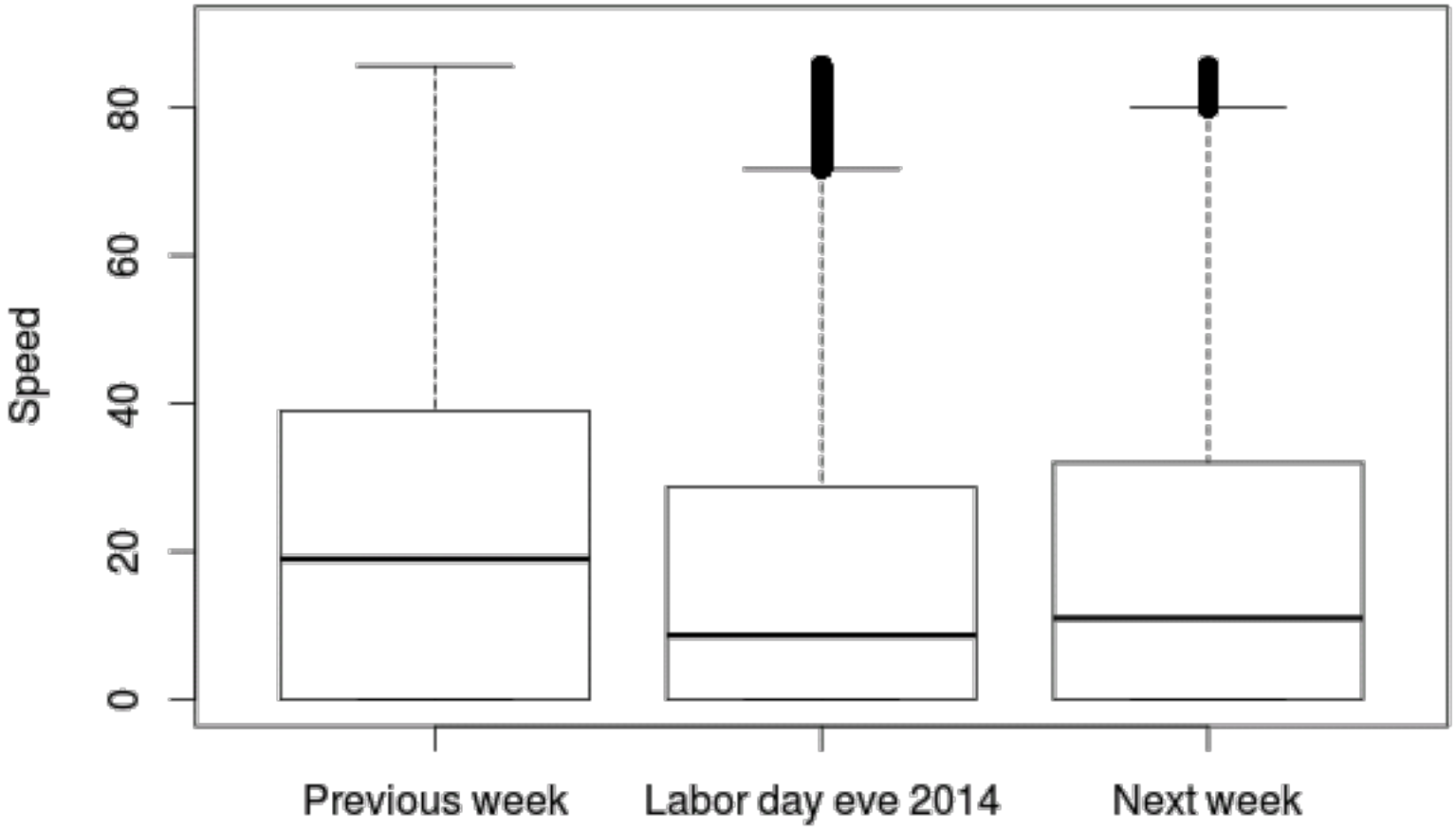
Análise – RQ7: Como são as condições do trânsito em dias de feriados?



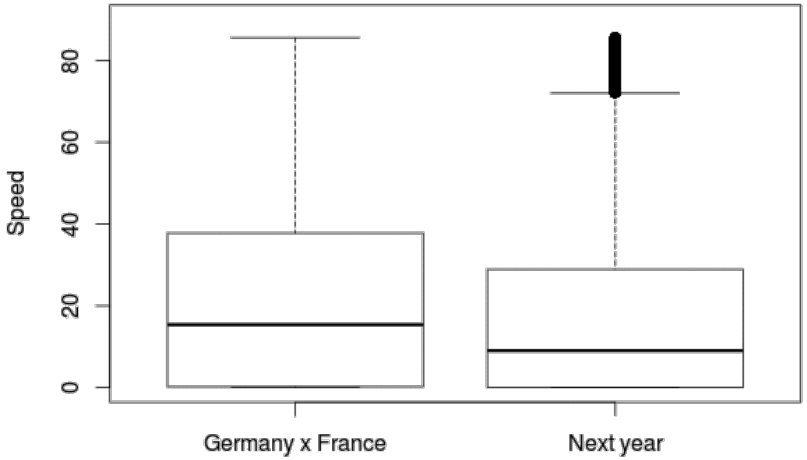
Análise – RQ7: Como são as condições do trânsito em dias de feriados?

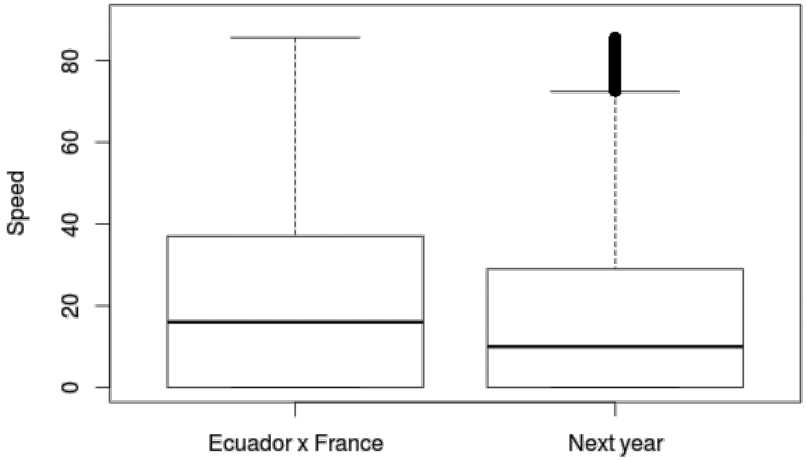


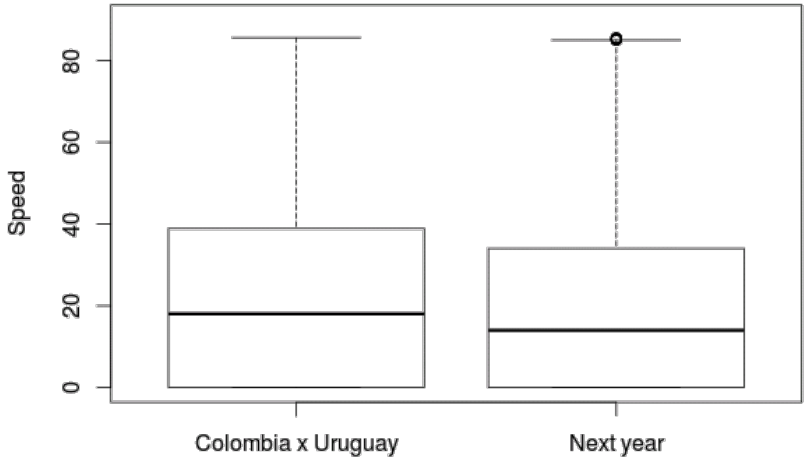
Análise – RQ7: Como são as condições do trânsito em dias de feriados?



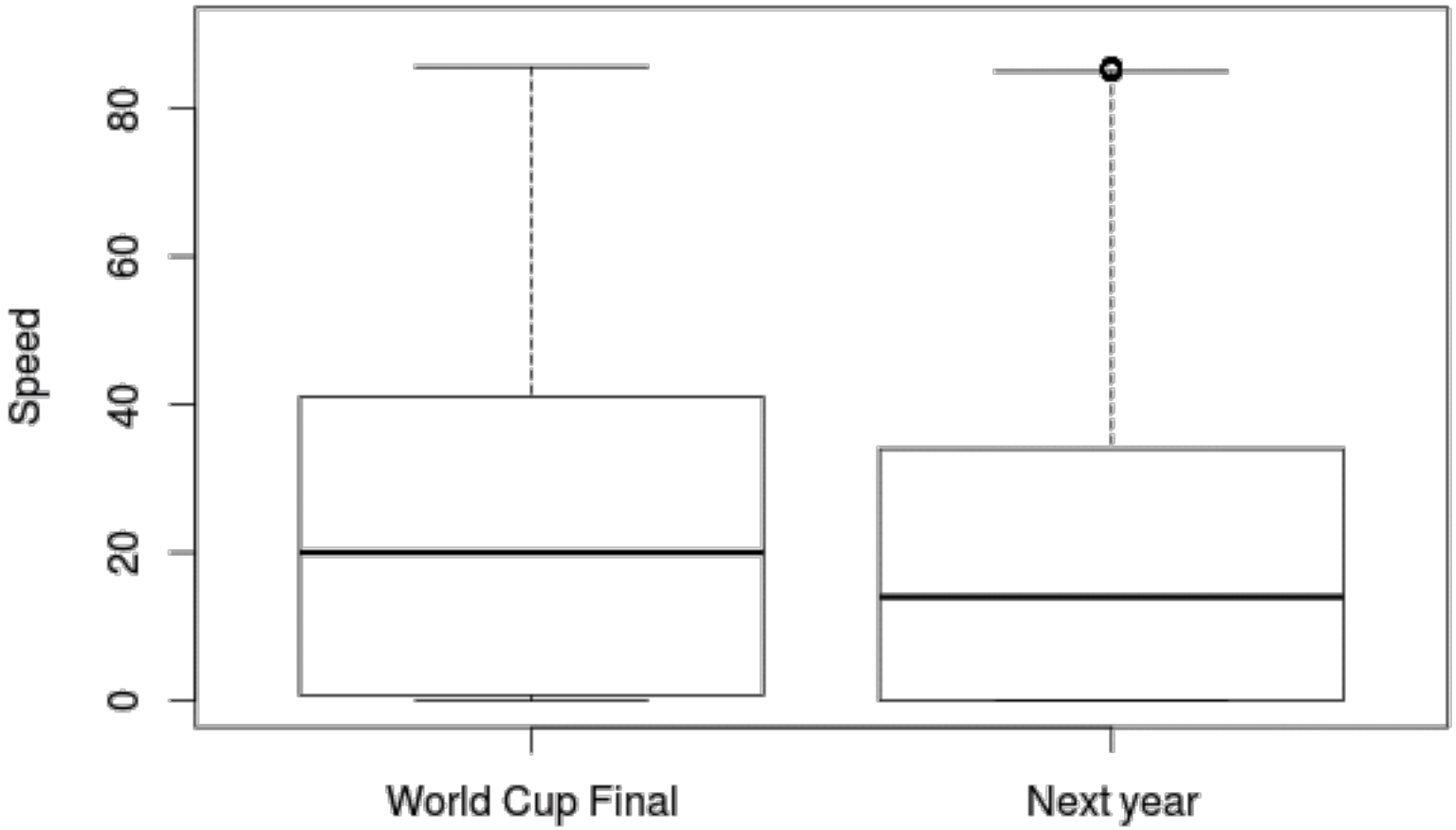
Análise – RQ8: Como são as condições do trânsito em dias de grandes eventos?







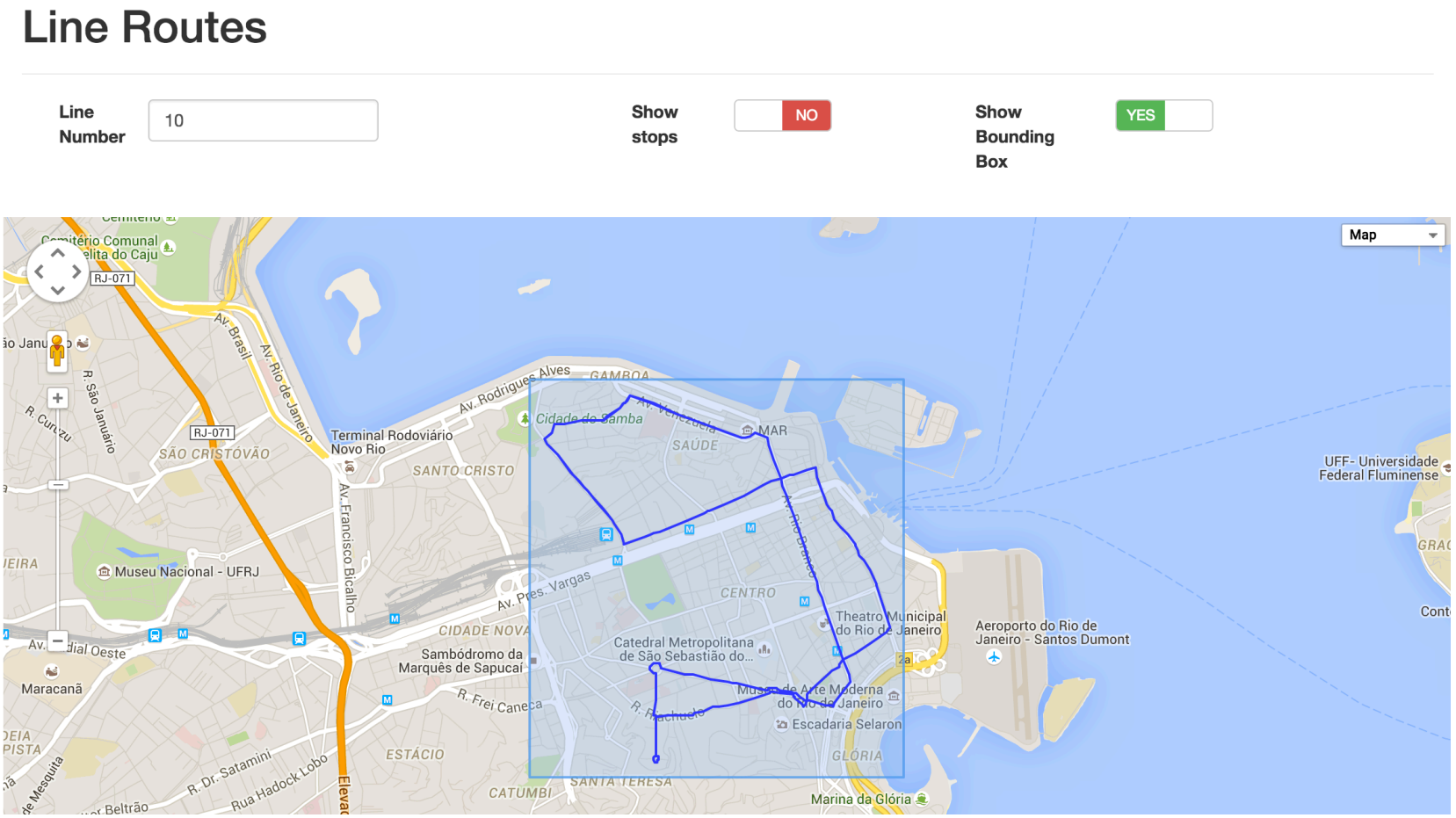
Análise – RQ8: Como são as condições do trânsito em dias de grandes eventos?



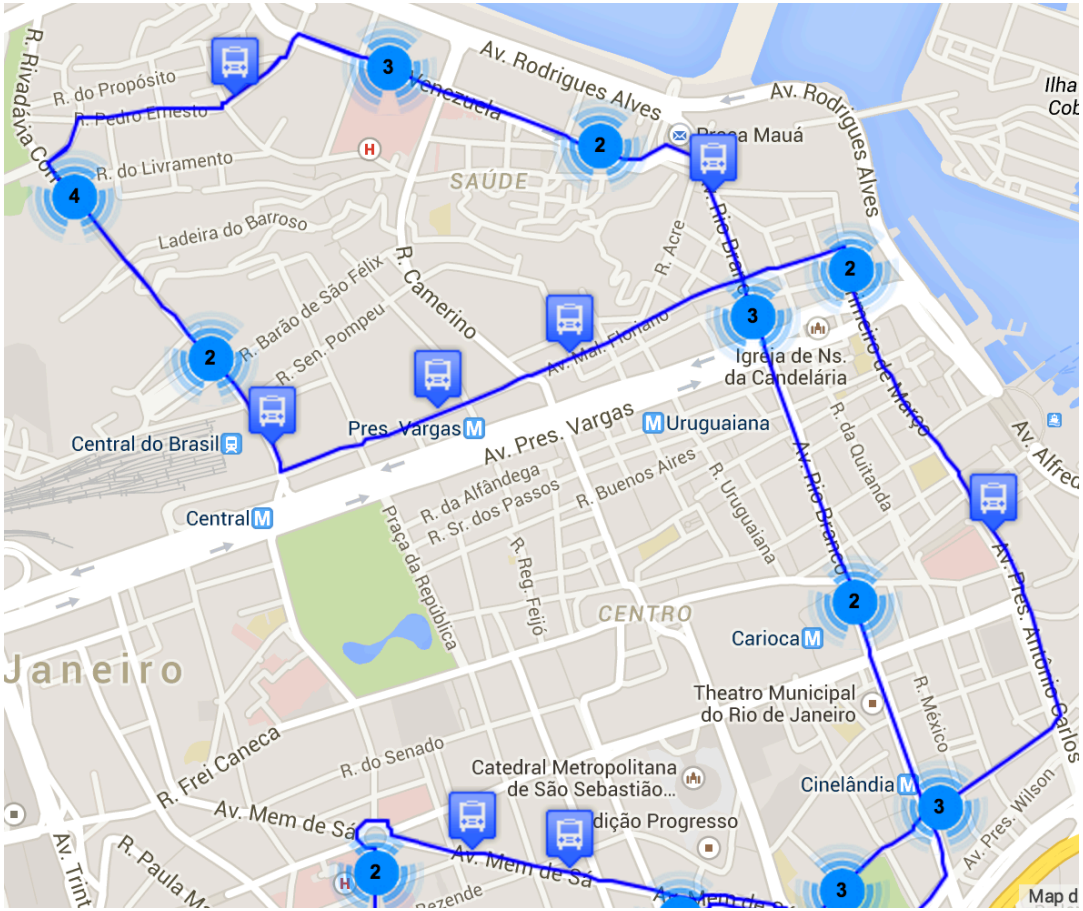
Análise
visualização dos dados

- Rotas
- Pontos de ônibus
- Congestionamentos
- Comparações entre dias e horários
- Densidades

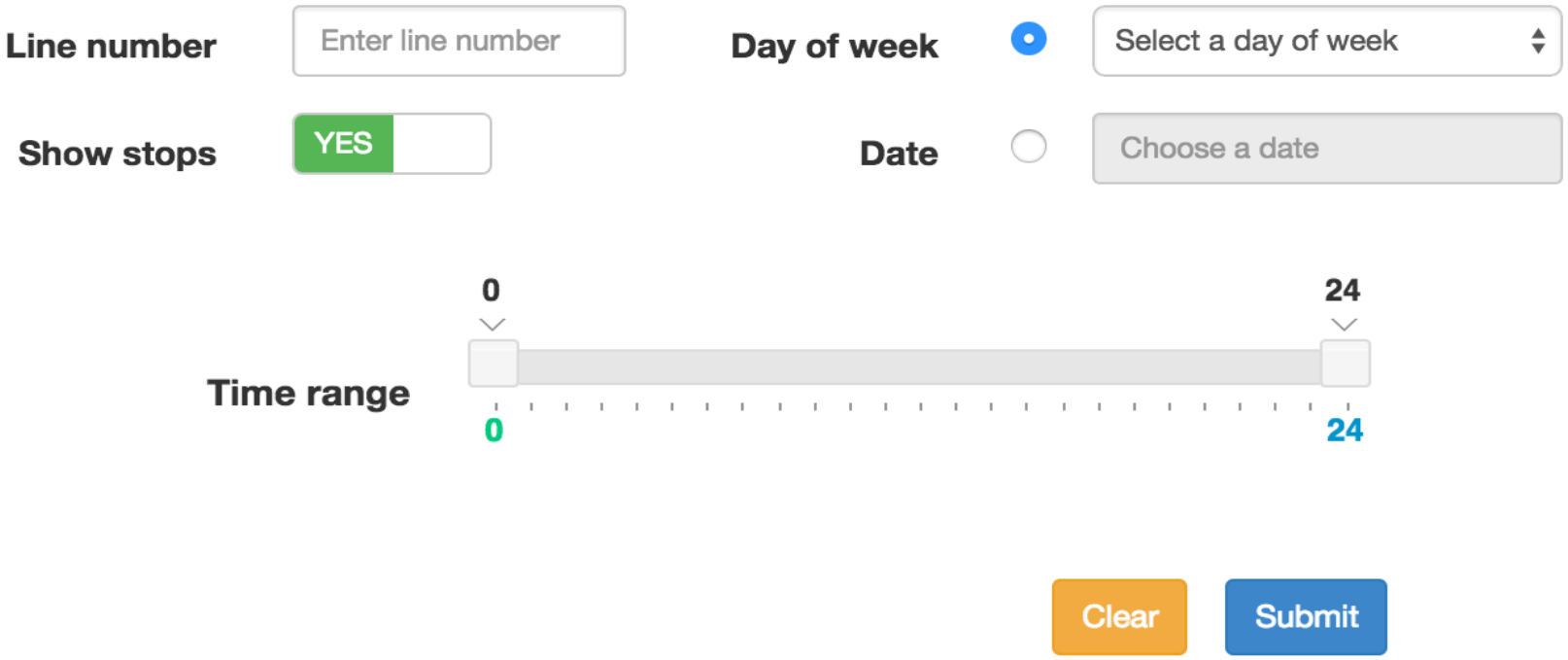
Análise rotas



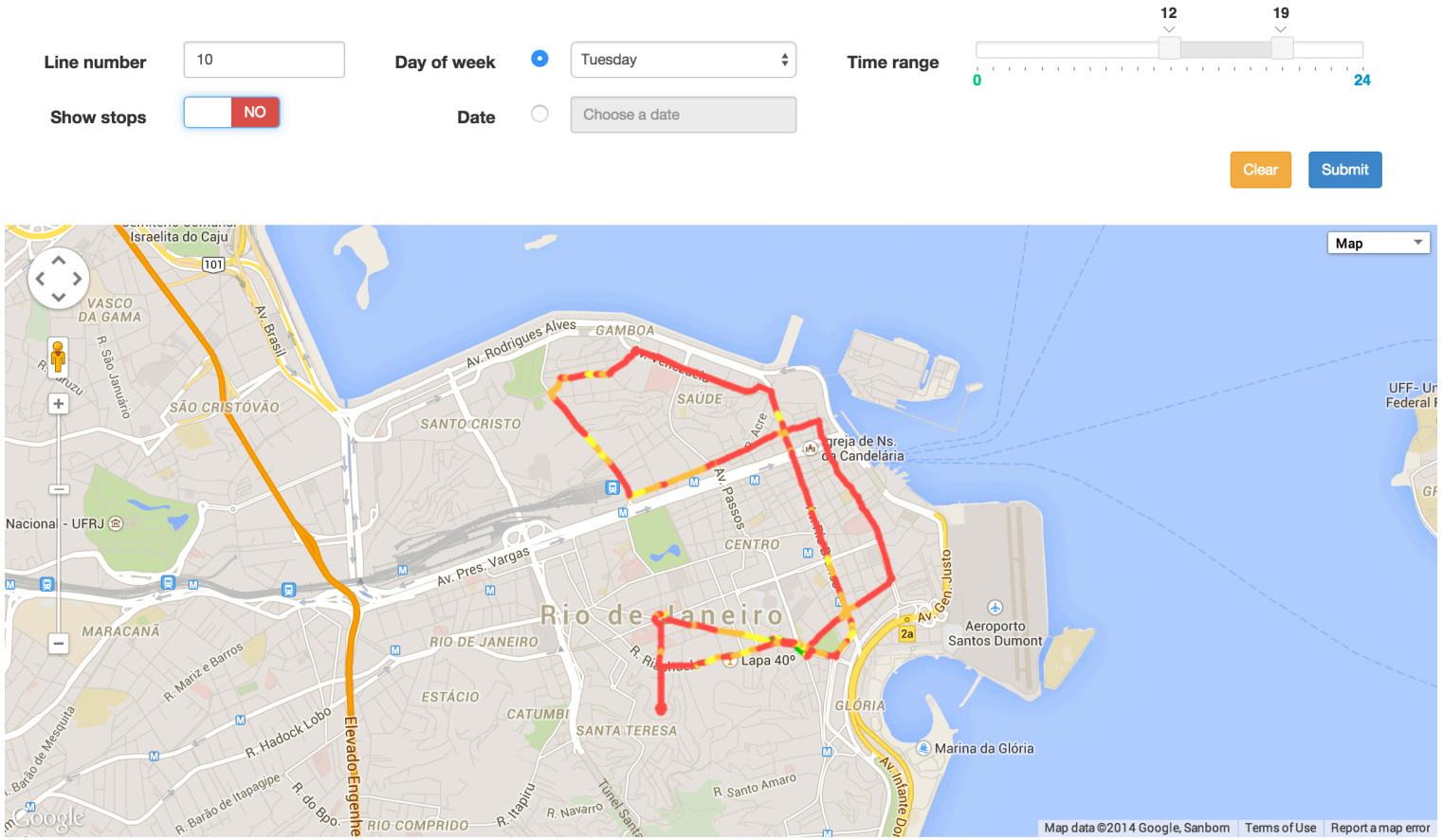
Análise pontos



Análise congestionamento



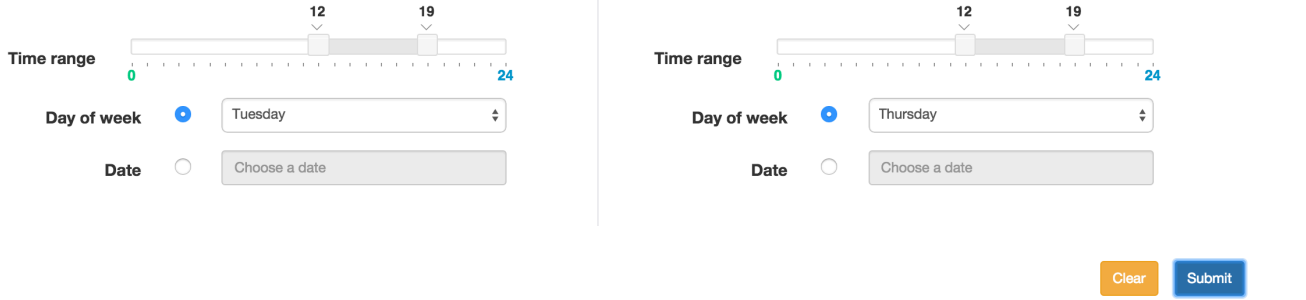
Análise congestionamento

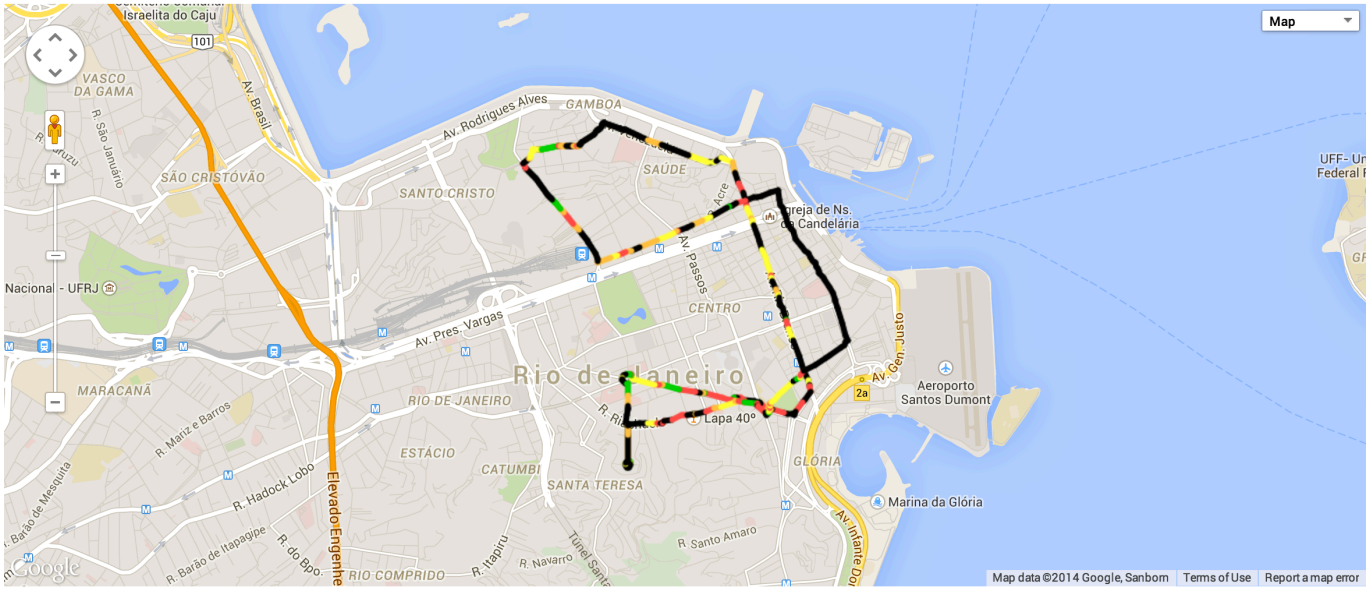


Análise congestionamento

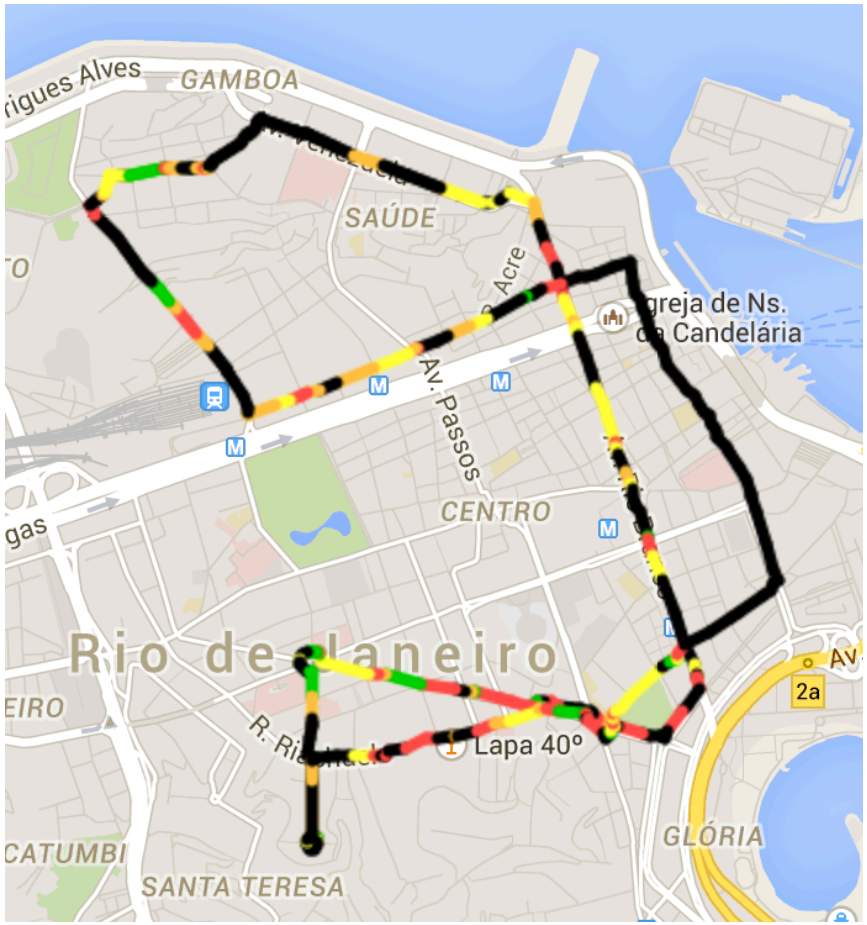
- .Menor que 15 km/h
- .Entre 15 e 30 km/h
- .Entre 30 e 60 km/h
- .Maior que 60 km/h

Análise comparações



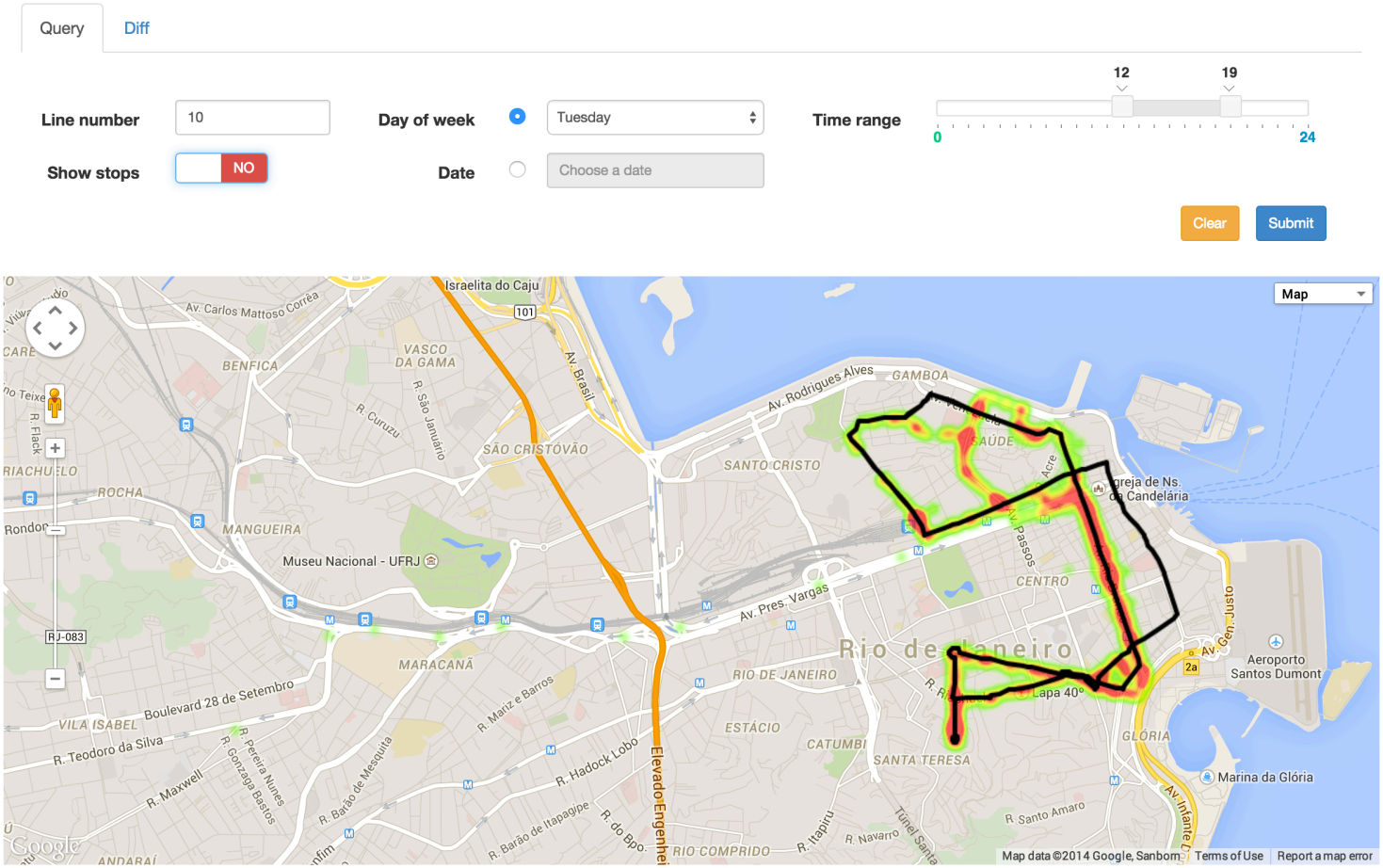


Análise comparações

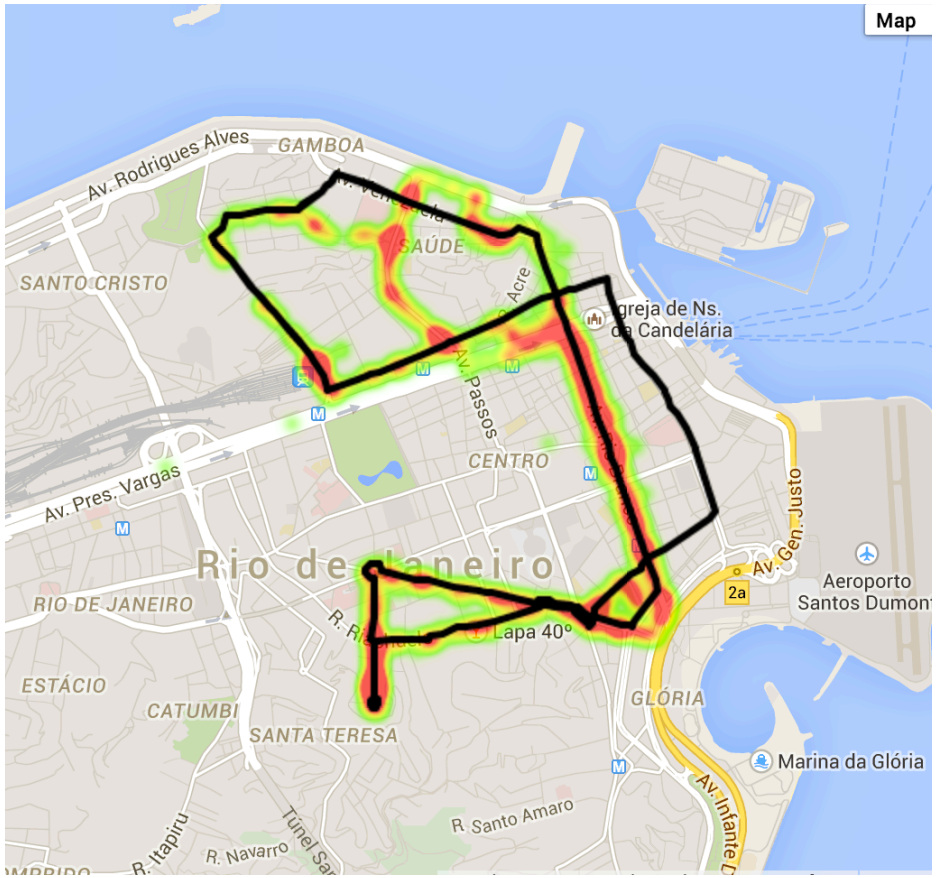


- .Menor que -20 km/h
- .Entre-20 e -5 km/h
- .Entre -5 e 5 km/h
- .Entre 5 e 20 km/h
- .Maior que 20 km/h

Análise densidade



Análise densidade



Trabalhos em andamento

- Melhoria nas análises
  - Detecção automática de garagens
  - Detecção automática de linhas
  - Cálculo de distância projetada na rota
- Cruzamento com dados de outras fontes
- Sugestão de horário de saída para usuário final
- Analise de outros tipos de dados urbanos

– Equipe –

Professores, alunos e parceiros

Professores

Esteban Clua

Computação Visual

Leonardo Murta

Engenharia de Software

Marcos Lage

Computação Visual

Vanessa Braganholo

Banco de Dados

Alunos

Alessandro Campello

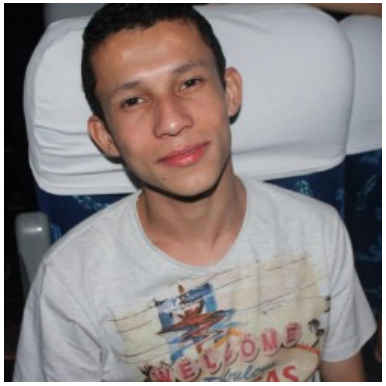
Bruno Schettino

Clarissa Tuxen

Daniel Campagna

Matheus Rodrigues

Pedro Fadel



Thiago de Paula

Parceiros

NYU/CUSP

Intercâmbio acadêmico:
Visitas docentes, pós-doc, mini-cursos, pesquisa ...

Processando 6 Bilhões de Posições GPS de Ônibus da Cidade do Rio de Janeiro

Leonardo Murta
leomurta@ic.uff.br

Baseado nos slides de Marcos Lage e Bruno Schettino

Instituto de Computação
uff