Análise das posições coletadas dos ônibus municipais de São Paulo pelo Olho Vivo

Danilo Lessa Bernardineli
Instituto de Física
Universidade de São Paulo
danilo.bernardineli@usp.br

13 de dezembro de 2015

Resumo

Desenvolveram-se scripts para adquirir e analisar dados de posição dos ônibus municipais de São Paulo, o que resultou em varreduras por toda cidade entre intervalos de 3 minutos, obtendo-se 80 milhões de dados em 5 semanas. Verificaram-se algumas características como picos de quantidade de ônibus ativos e velocidades menores nos horários de pico bem como diferenças na distribuição dos ônibus e a ausência do horário de pico matinal durante os finais de semana e ausência de pico na tarde aos sábados, bem como constância na quantidade de ônibus durante o dia nos finais de semana. Foi verificado também que é sempre mais rápido ir do que voltar ao centro e a zona leste de SP no que tange ônibus.

Sumário

1 Agradecimentos

Agradeço ao professor Zwinglio Guimarães, do Instituto de Física da Universidade de São Paulo, por seu enorme apoío na realização deste trabalho bem como por suas ideias e pela disposição da internet em sua sala no DAP, sem a qual dificilmente o volume de dados teria sido o atual.

2 Nota

As linhas de ônibus possuem normalmente uma rota da periferia ao centro, a qual será chamada de "ida"nesse trabalho, e outra do centro para a periferia, que será chamada de "volta".

3 Introdução

Este trabalho visa retirar conclusões qualitativas a respeito do transporte público paulistano, em específico no que tange os ônibus municipais, utilizando dados de posição dos ônibus obtidos através do sistema Olho Vivo da SPTrans.

Adicionalmente, foram desenvolvidos scripts para aquisição dos dados bem como processamento e análise destes, os quais podem ser obtidos e usados livremente em http://soc.if.usp.br/ danlessa/

3.1 Motivação

A motivação deste trabalho veio da curiosidade em obter conclusões sobre algumas perguntas-base:

- Qual é a faixa de horário que os ônibus estão mais rápidos?
- Se eu for pegar um ônibus, em que faixa de horário terei menor tempo de viagem?
- Como os ônibus se distribuem na cidade?
- A velocidade dos ônibus seguem alguma distribuição específica?
- Como os ônibus se comportam em horários de pico?

4 Método

4.1 Aquisição de dados

Foi criado um token de acesso na área de desenvolvedor da SPTrans para que se obtivesse acesso aos arquivos GTFS[1], os quais contém informações sobre as rotas e linhas dos ônibus, bem como acesso a API do Olho Vivo, a qual fornece informações sobre as posições dos ônibus.

Utilizando-se do acesso, foi então feito um script em Python para utilizar todas as funcionalidades da API do Olho Vivo bem como fornecer uma função de obter a posição de todos os ônibus da SPTrans em varreduras regulares (Foi definido um intervalo de 3 minutos entre cada varredura completa, sendo que normalmente esta era completada em menos de um minuto, ocorrendo assim um tempo ocioso). Os dados adquiridos pelo script são os códigos de linha, código de ônibus, latitude e longitude. O tempo utilizado era o timestamp do próprio computador e a velocidade instantânea era calculada utilizando a fórmula de haversine. Dados repetidos eram eliminados dinâmicamente[2].

A aquisição dos dados através inicialmente foi feita no servidor do projeto Sócrates do Instituto de Física da Universidade de São Paulo, porém não foi utilizados tais dados coletados devido ao servidor reiniciar diariamente e portanto gerar um volume de dados bastante descontínuo e em volume insuficiente.

Algum tempo depois de iniciado a aquisição de dados, foi cedido um lugar na sala do professor Zwinglio Guimarães para colocar um Raspberry Pi 2, de onde é oriundo todos os dados utilizados neste trabalho.

A necessidade de usar um computador dentro da Universidade de São Paulo se deu devido à sua conexão de Internet rápida e estável, pois apesar do script não realizar uma taxa de download alta, ele exige uma boa latência, algo inatingível com uma internet residencial ou servidor off-shore.

4.2 Processamento dos dados obtidos

Para o processamento de dados e geração de gráficos, foi feito outro script em Python3+NumPy, cujo código está dísponivel na seção de referências.

O processamento dos dados visava calcular o tempo de viagem com base na troca de linha de um mesmo ônibus (devido a estrutura da API, a ida e a volta são como se fossem linhas diferentes), obter quantidade de ônibus

[1]Sobre o GTFS: https://developers.google.com/transit/gtfs/

[2]O critério pode ser visualizado no código-fonte

ativos em intervalos do dia, e medianas do tempo de viagem e velocidade em intervalos específicos dos dias.

Utilizando os dados processados bem como os não-processados, foram feitos então vários[3] tipos de gráficos com diferentes variedades de filtros afim de validar diversas hípoteses acerca dos ônibus.

5 Resultados

Foram coletados uma quantidade de na ordem de 80 milhões de dados para este trabalho escrito. Normalmente, o script concluiu uma varredura completa nos ônibus de São Paulo em cerca de 100 segundos. Os dados são de 13/10/2015 (17:20) até 16/11/2015 (15:00)[4], o que resultou em aproximadamente cinco semanas de dados.

Através dos dados, foi constatado que há um total de 14492 códigos de ônibus distintos embora no site da SPTrans conste um total de 14812 veículos cadastrados. Foi verificado também a existência de 2689 linhas ativas, sendo que na doumentação consta 2710 linhas. A discrepãncia pode ser explicada devido a cada rota possuir duas linhas: uma de ida e outra de volta, porém algumas rotas são circulares e isso resulta de que na outra rota associada a linha se torna inativa.

Tabela 1: Números gerais dos dados

Quantidade de dados	aprox. 80 milhões (3.4GB)
Dados analisados	aprox. 70 milhões
Intervalo dos dados	13/10/2015 até 16/11/2015
Duração da varredura	aprox. 100s
Quantidade de ônibus	14715
Quantidade de ônibus (oficial)	14812
Linhas ativas	2689

Tabela 2: Indicadores estatísticos básicos

Tipo	Mediana	Média	Desvio Padrão	Skewness	Kurtosis
Velocidade (km/h)	13.03	14.29	9.09	1.28	3.15
Tempo de viagem (min)	45.02	52.53	35.55	2.44	10.18

[3]Ou em outras palavras, a seção de análise de dados inteira

[4]Em duas ocasiões, devido a quedas do servidor da SPTrans, o script de coleção de dados parou, o que ocasionou em 3 ou 4 dias sem dados nesse período

6 Análise de dados

6.1 Intervalo dos dados a serem considerados

Ao inspecionar os dados, algumas anomalias foram notadas: uma pequena porção deles apresentam velocidades absurdamente altas enquanto há um número substancial de velocidades nulas ou seguindo uma distribuição inesperada, conforme visto nos histogramas de velocidades[5]
Por inspeção, adotou-se então os dados no intervalo entre 1km/h e 70km/h para serem utilizados nas análises seguintes por apresentarem um comportamento mais suave[6]. Notou-se também pelo último histograma que a distribuição de velocidades apresentava um skewness positivo significante, e por isso o estimador utilizado nesse trabalho foi o da mediana.
Comentários adicionais sobre o comportamento das velocidades está na seção de discussões deste trabalho, na página 15.
Um comentário a respeito da porcentagem de exclusão: cerca de 8.92% dos dados estão com velocidade zero, 1.2% estão com velocidade superior a 70km/h e cerca de 11% estão com velocidade não-nula inferior a 1km/h.

6.2 Distribuição de velocidades

Para visualizar como as velocidades se distribuem, foram feitos gráficos de barras das medianas das velocidades em cada intervalo de hora do dia[7].
Notam-se intervalos de mínima velocidade na hora do rush matinal e vespertino bem como sutilmente um pico de almoço nos dias úteis, enquanto nos sábados e domingos não houveram picos matinais e houve um nítido pico de almoço (especialmente no domingo). Além disso, notou-se a ausência de um pico vespertino nos sábados, o que é surpreendente diante do esperado pelo senso comum.

6.3 Distribuição dos tempos de viagem

Foram feitos gráficos em barras das medianas dos tempos de viagem[8] por intervalo de hora do dia bem como um histograma.[9] O tempo foi calculado através do monitoramento de quando um ônibus trocava de linha, pois como

[5]Fig. 1, pág. 7

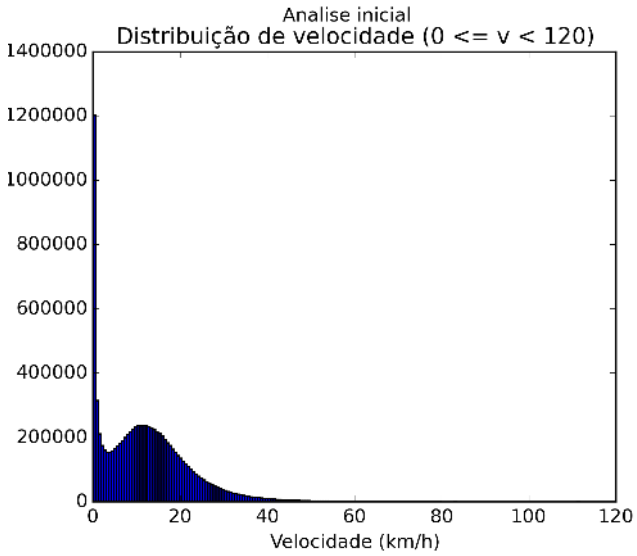
[6]Apesar de não ser o ideal, decidi por essa opção neste trabalho devido a dúvidas sobre a distribuição em velocidades baixas

[7]Fig. 5, pág. 9

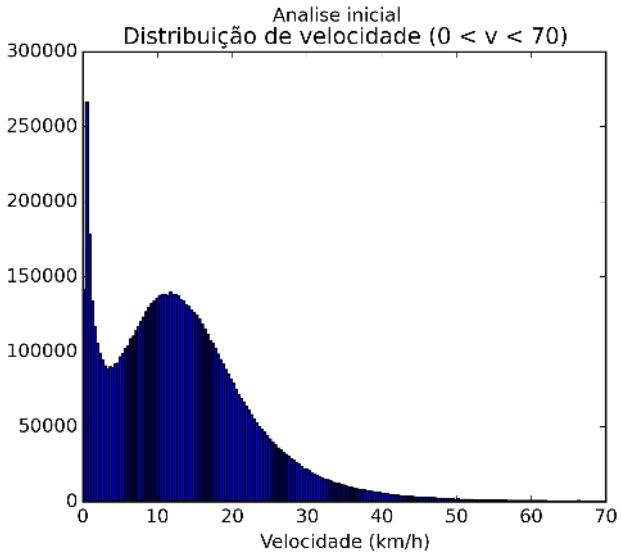
[8]Tempo de viagem caso você pegue um ônibus no horário em questão

[9]Fig. 7, pág. 11

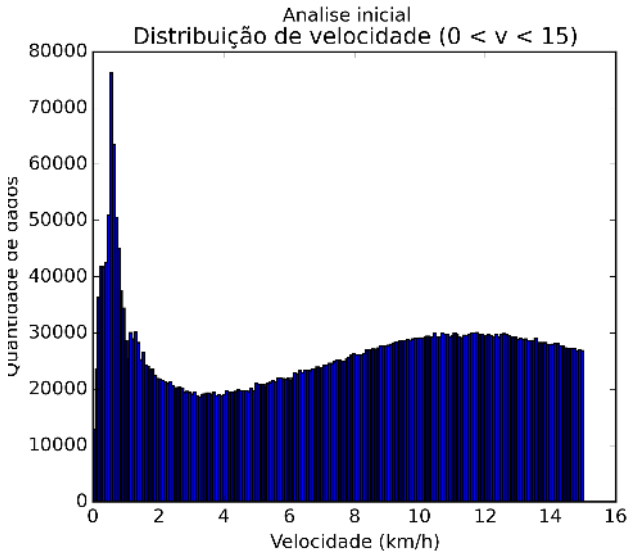
Figura 1: Histogramas de velocidades



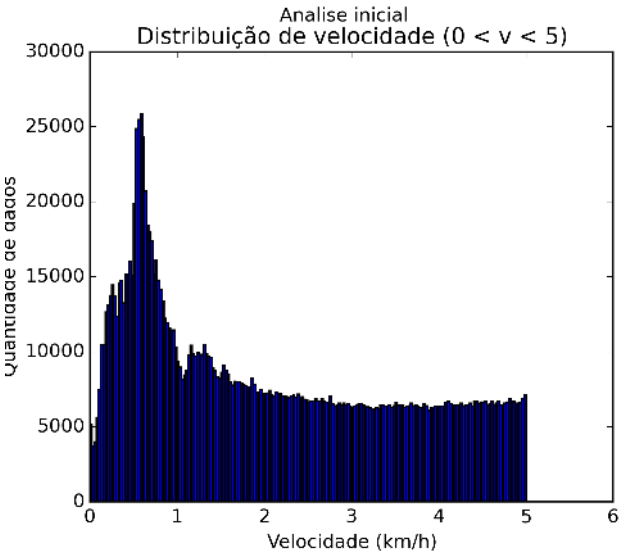
(a) Porcentagem de dados excluidos: 1.02%



(b) Porcentagem de dados excluidos: 10.37%

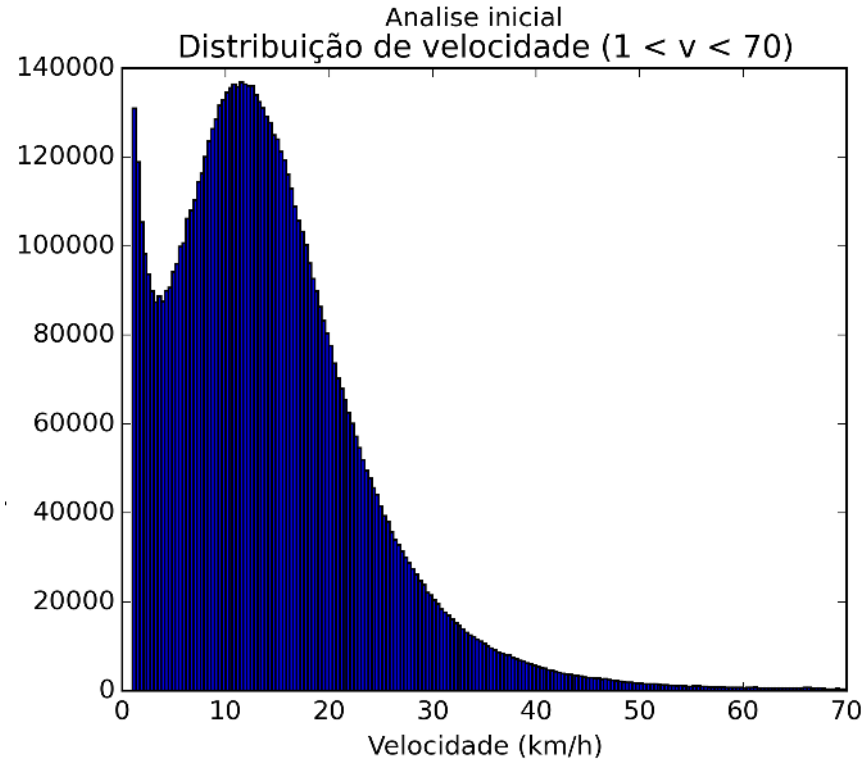


(c) Porcentagem de dados excluidos: 44.15%



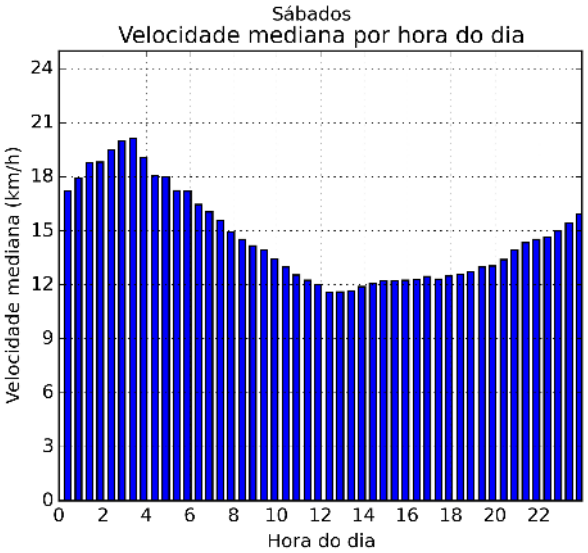
(d) Porcentagem de dados excluidos: 82.05%

Figura 3: Histograma das velocidades consideradas

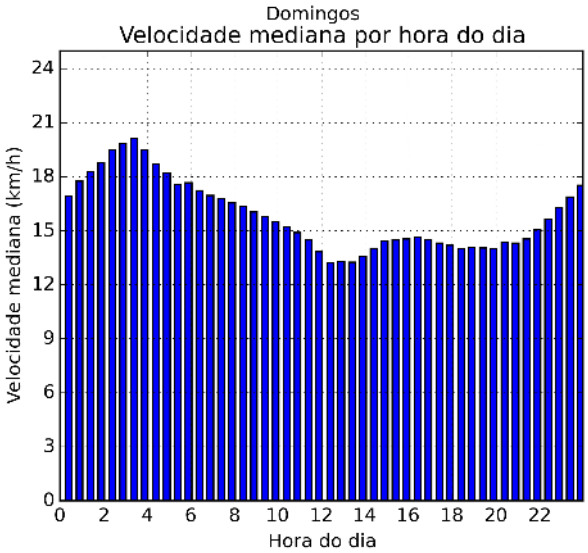


(a) Porcentagem de dados excluidos: 22.72%

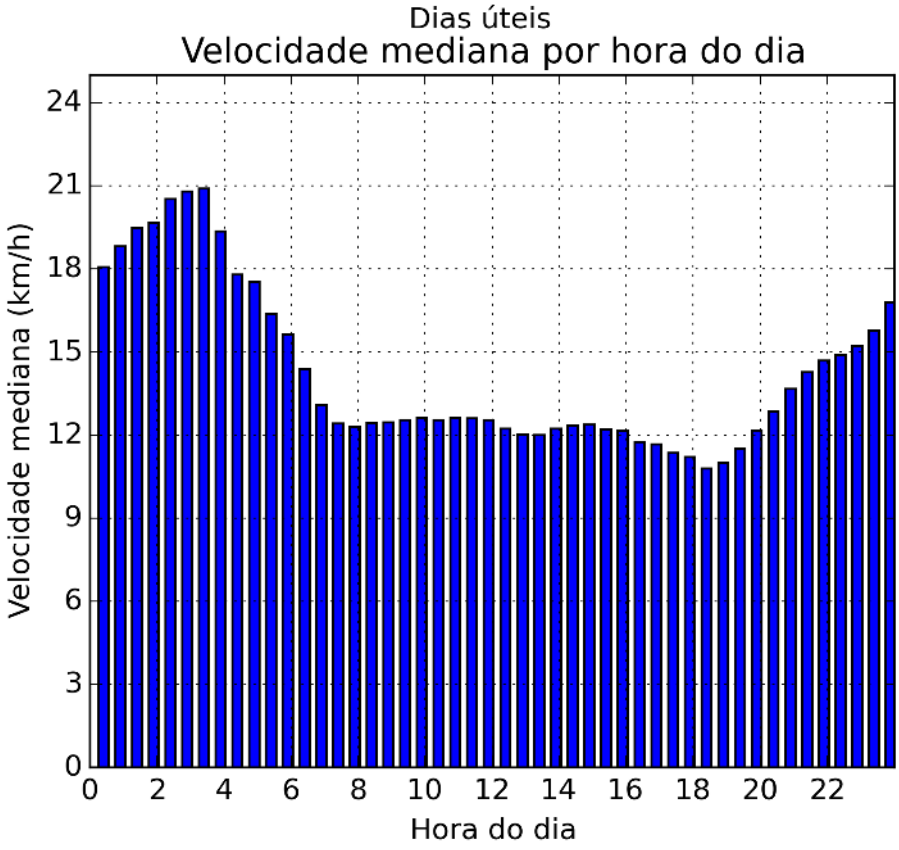
Figura 5: Velocidades medianas por intervalo do dia



(a) Sábados



(b) Domingos



(c) Dias úteis

a rota de ida e volta são representadas como linhas diferentes, um ônibus nunca pode continuar na mesma linha após terminar o trajeto.
A primeira viagem detectada de cada ônibus monitorado foi eliminada para evitar pegar um tempo "pela metade". Foi aplicado também uma exclusão de tempos de viagem menores que 10min, pois além de ser intuitivamente muito improvável, há também a consideração dos "teletransportes"[10], onde um ônibus pode subitamente mudar de linha e causar tempos de viagens abrutamente baixos.
Ao visualizar o histograma, nota-se uma distribuição com skewness significantemente positiva, o que novamente justifica a adoção da mediana como estimador de tendência central.
Os tempos de viagem estão dentro do esperados, sendo possível imediatamente associa-los inversamente com as barras de velocidade por intervalos em horas do dia. Repare que os tempos de viagem começam a aumentar um pouco antes dos horários de picos observados pelos gráficos de velocidade.
Note que as barras aparentam estar com alturas somente em múltiplos de 3. Isso se deve ao script de aquisição esperar 3 minutos após o ínicio de uma iteração para recomeçar novamente. Como a varredura normalmente é feita em cerca de 1min, a consequência é que os tempos de viagem se concentram perto dos múltiplos de 3.

6.4 Distribuição de ônibus ativos

Para cada intervalo de tempo em cada dia com dados, foi obtida a quantidade de ônibus distintos operando e foi então calculada a mediana para o intervalo. O resultado é um gráfico em barras da quantidade de ônibus ativos por intervalos de hora do dia (fig.9, pág. 12.
Nos dias úteis, a maior quantidade de ônibus operando ocorre nos horários de picos matinais e vespertinos, o que é algo esperado porém não menos curioso, enquanto que que os dias de final de semana não apresentam a inflexão no meio do dia igual no dia de semana, sinal de que a frequência de ônibus é aproximadamente constante durante todo o dia.

6.5 Ônibus de ida x ônibus de volta

As linhas municipais apresentam sempre, até onde foi verificado manualmente, um destino "central"e um destino "periférico"que são facilmente obtidos através da numeração da linha (os de ida possuem um código entre 0 e 10000, e os de volta possuem código entre 30000 e 40000).

[10]Ver seção de discussões, pág. 15

Figura 7: Histograma e barras de tempos de viagem por intervalo de hora

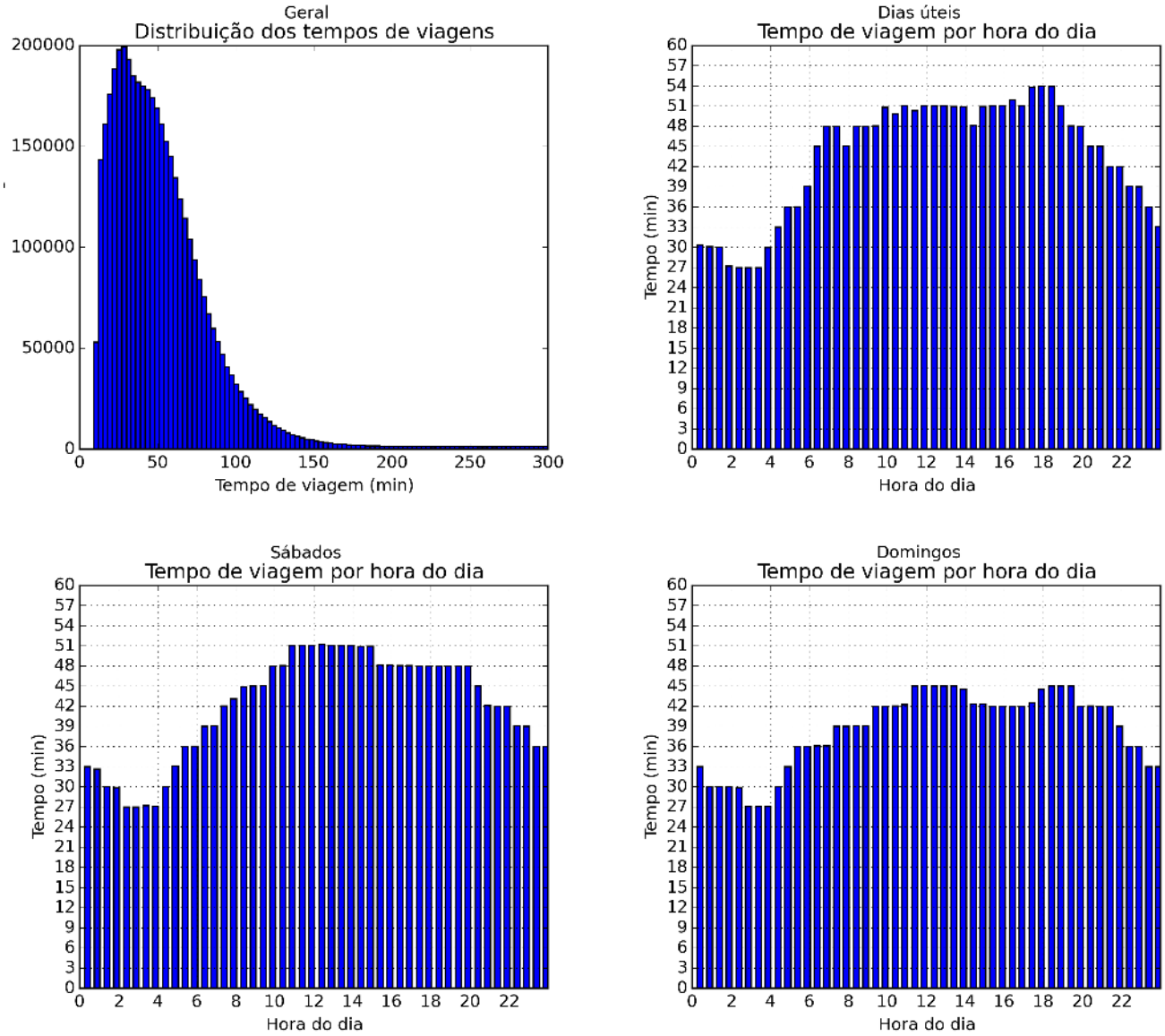


Figura 9: Barras da mediana dos ônibus ativos por intervalos de hora

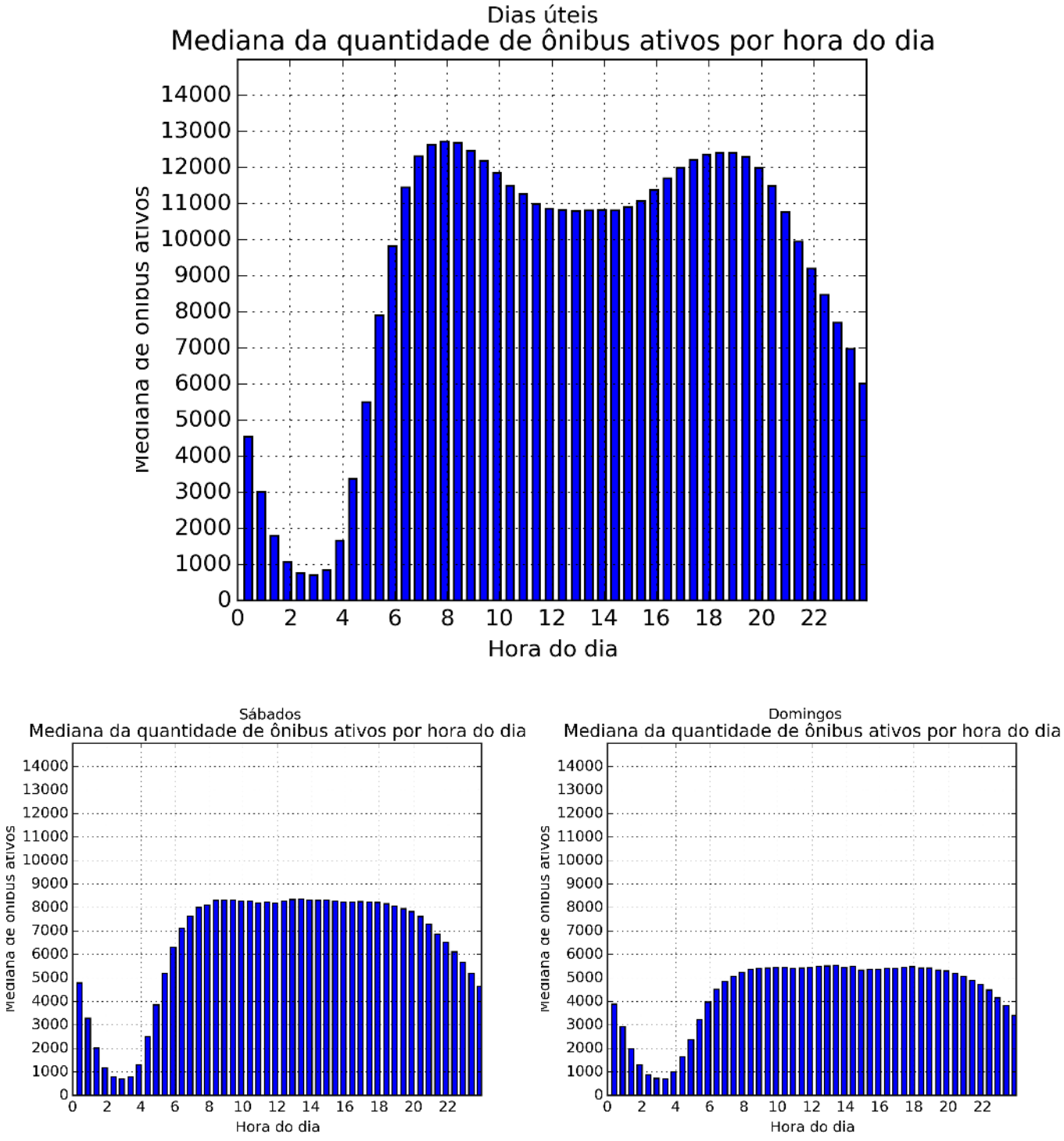
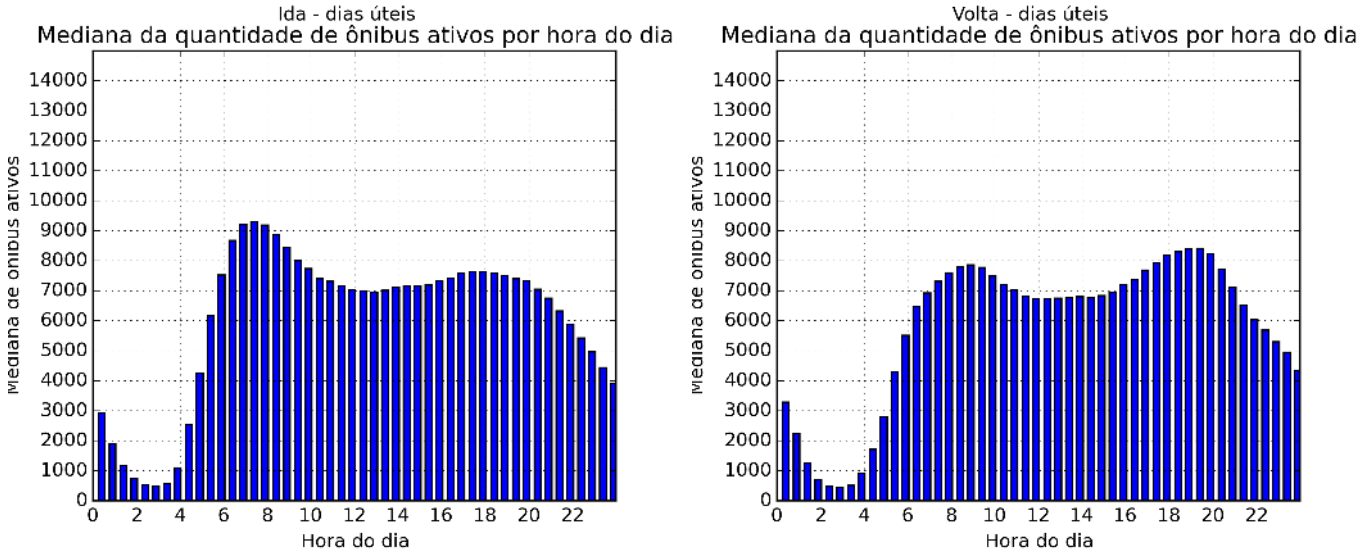


Figura 11: Barras de ônibus ativos por intervalos de hora (ida ou volta)

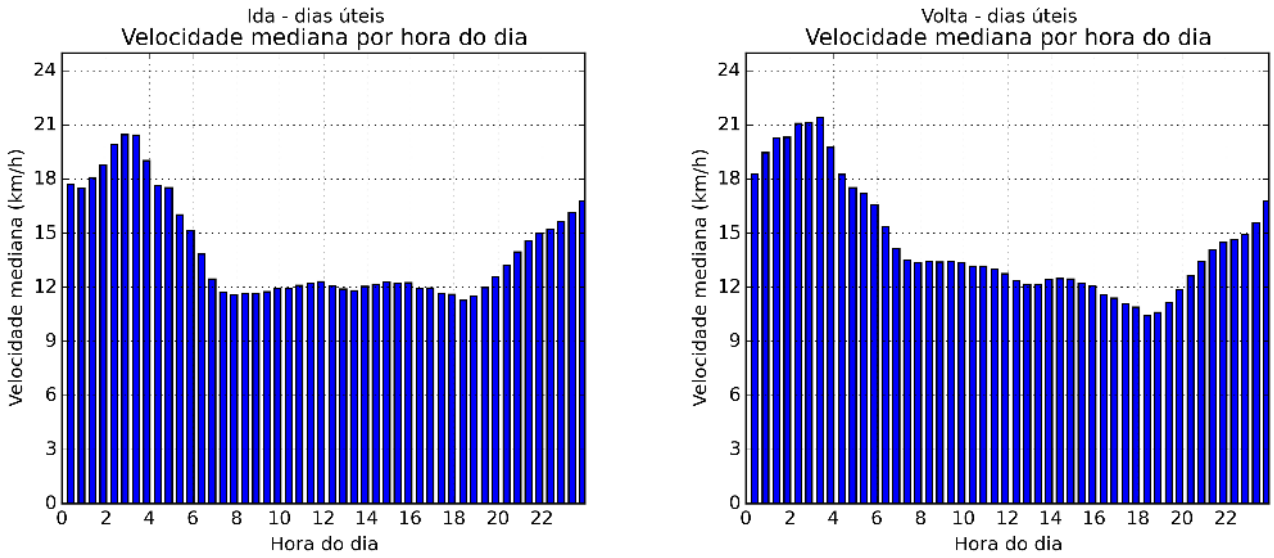


Para verificar a tal a afirmação, basta a visualização dos gráficos de ônibus ativos separados pelo código. Nota-se que as linhas de "ida"possuem um máximo no começo da manhã, que é quando, pelo senso comum, os ônibus de ida estão mais lotados. O análogo acontece com as linhas de "volta", porém no final da tarde.
É feito também barras de velocidade mediana por intervalo de hora para as linhas de "ida"e de "volta", e nota-se que as mesmas apresentam velocidades menores respectivamente na hora do rush da manhã e da tarde em relação uma a outra. O que pode ser índício de menor velocidade do trânsito no sentido e que é válido conforme o esperado.
Foi criado também um gráfico de barras da diferença da velocidade dos ônibus de ida e os de volta por intervalo de hora.

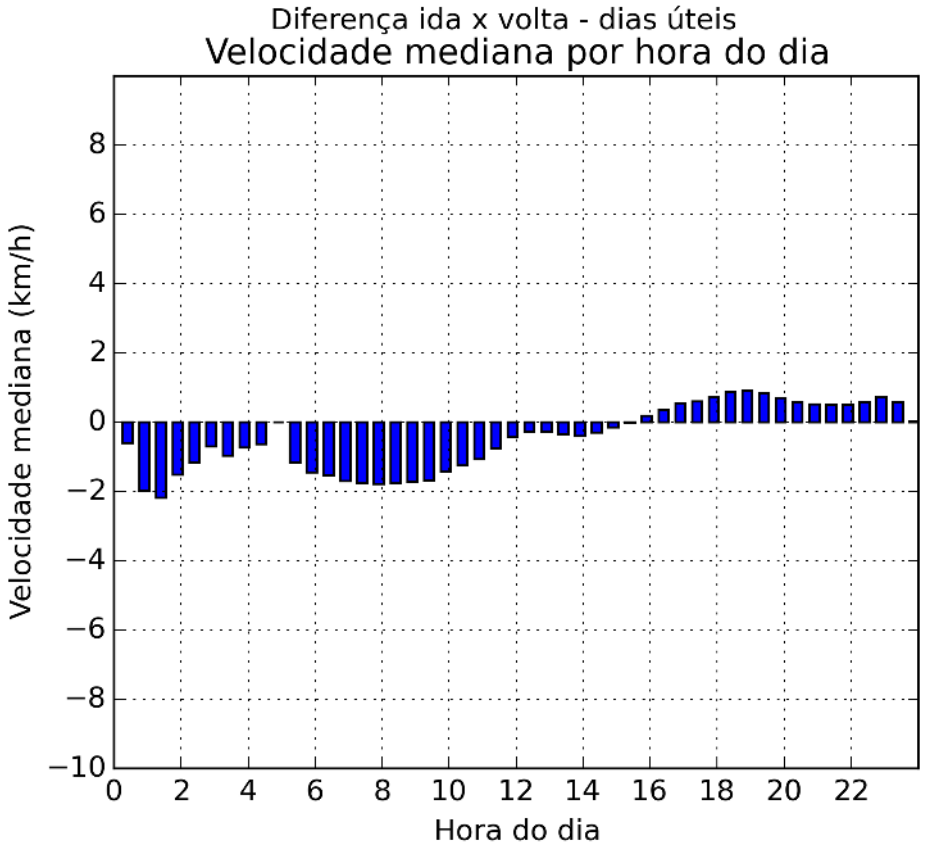
6.6 Análise de regiões específicas

Foram feitos para algumas regiões gráficos de barras de velocidade por intervalo de hora, barras de quantidade de ônibus ativos por intervalo de hora, barras de diferença de velocidade entre linhas de ida e volta por intervalo de hora, mapa matricial de distribuição normalizada de ônibus e os mapas matriciais de diferença entre linhas de ida e volta nos horários de pico como na seção anterior.
As regiões escolhidas foram:

Figura 13: Barras de velocidades medianas por intervalos de hora (ida ou volta)



(a) Diferença entre a vel. mediana da ida e a da volta



- Centro, por ser uma região muito movimentada
- Zona leste, por ser populososo e periférico
- Zona oeste, por ser perto da USP e melhor conhecido pelo autor
- Extremo sul, por ser um caso mais extremo

As regiões são delimitadas pelo quadrado delimitado pela permutação entre os pontos [-23.50, -23.63] x [-46.37, -46.61] para a zona leste, [-23.50, -23.60] x [-46.60, -46.70] para o centro, [-23.52, -23.62] x [-46.67, -46.77] para a zona oeste e [-23.65, -23.85] x [-46.65, -46.85] para o extremo sul.
Os gráficos estão no apêndice por fins de formatação, porém o resultado para a Zona oeste, leste e sul é análogo com o das seções anteriores, com algumas observações ao observar as barras de velocidade.

- Os horários de picos matinal e vespertino da zona oeste são mais acentuados que o de outras regiões ou da cidade como todo
- O horário de pico de almoço do extremo sul é quase equivalente a um horário de pico tradicional
- É mais rápido ir do que voltar para o centro e a zona leste

Os gráficos de diferença de velocidade demonstram o último item, enquanto a análise dos gráficos de velocidade mediana demonstram o primeiro e segundo.

7 Discussões

7.1 Velocidades extremas (muito baixas e muito altas)

A investigação que foi feita sobre as velocidades altas, indicam que as vezes os ônibus se "teletransportam"para algum lugar (normalmente, mas nem sempre, para o destino ou origem da linha) e retornam imediatamente para a posição original. Há também casos de ônibus que ficam "permanentemente"em algum ponto (como por exemplo os circulares 2 que atendem a USP) que podem contribuir para a abundância de dados com velocidade nula e/ou baixas velocidades.
Adicionalmente, não é bem caracterizado o erro na velocidade dos ônibus, que podem vir e serem combinados de diversas origens, como por exemplo:

- Erro associado ao aparelho de GPS, cuja precisão pode variar significativamente para cada modelo[1]

- Atrasos durante a transmissão de posição do ônibus
- Atraso do servidor em transmitir a última posição
- Erro devido ao intervalo de atualização de posição dos ônibus
- Dados anômalos devido a ônibus parado ligando ou desligando
- Aparelho de GPS advinhando a posição com base no último dado ou pela torre de celular

No momento atual, não há conhecimento sobre como cada um desses itens afeta os dados e portanto uma futura extensão do trabalho pode se dar no âmbito de caracterizar estes e analisar a influência dos mesmos na análise feita.

8 Conclusões

A respeito das distribuições, concluiu-se que as velocidades dos ônibus seguem uma distribuição assimétrica com skewness positivo, o que é intuitivo e esperado. Porém nada se pode afirmar a respeito da distribuição em velocidades abaixo de 1km/h sem estudos adicionais. A distribuição dos tempos de viagem parece ser similiar ao das velocidades dos ônibus, porém a falta de resolução adicional devido ao intervalo da aquisição de dados impede mais investigação.
No que tange a análise temporal, foi verificada a variação de velocidade conforme esperado pelo senso comum (velocidades noturnas maiores que diurnas), foi verificado a existência de intervalos com velocidades menores que a vizinhança: começo da manhã (nos dias úteis), almoço e final da tarde, notou-se que os ônibus ativos possuem maior quantidade durante os picos matinais e vespertinos durante dias úteis (que é intuitivo, porém não trivial) ao mesmo tempo que permanecem relativamente constantes durante os finais de semana. Os tempos de viagem parecem estarem associados inversamente com a velocidade mediana, porém novamente a falta de resolução evita mais conclusões[11]. Adicionalmente, foi verificada a queda de velocidade e maior número de ônibus no sentido centro ou periféria dependendo do sentido da linha.
Porém durante os sábados, surgiu o resultado inesperado de haver somente um horário de pico: durante o almoço.

[11]Há a conclusão de que todo paulistano sabe: não pegue ônibus no horário de pico da tarde, pode demorar 10min a mais que o normal no dia

Ao avaliar regiões específicas, foram obtidos os resultados interessantes de que em mediana, é sempre mais rápido ir do que sair da região central e da zona leste de São Paulo, e também que nos horários de pico da zona oeste os ônibus são mais lentos que o de outras regiões.

9 Futuro do trabalho e sugestões

9.1 Aplicações

O código escrito para a realização desse trabalho é facilmente adaptável para uma gama significante de usos, desde como sendo um wrapper para as funcionalidades do Olho Vivo para até como ser uma biblioteca para análise dinâmica dos dados de ônibus. Por consequência, há algumas possibilidades de aplicação.
A mais interessante em minha opinião, e que também é um possível projeto de férias, é a criação de um site público informativo para fornecer estatísticas dinâmicas para quem quiser acessar. È possível fornecer todos os gráficos e estatísticas feitas neste trabalho de forma dinâmica e personalizada (como por exemplo, fazer somente referente a uma linha, ou somente a uma região específica).
Há também a possibilidade do uso das análises ou dos scripts como forma de auxiliar em análise geossocial. Uma ideia nesse âmbito por exemplo seria a de mesclar o mapa de densidade de ônibus com um mapa de densidade demográfica. Ou então associar densidade de ônibus x densidade de moradores possuindo carros ou determinadas condições sócio-econômicas. Desse modo seria possível obter e relacionar os ônibus com determinados indicadores.
O código desse trabalho é público e livre para usar, dessa forma há também a pretensão de auxiliar o trabalho de futuras pesquisas que envolvam análise posicional dos ônibus de São Paulo

9.2 Potenciais desenvolvimentos

Assim como a SPTrans, a EMTU também possui um serviço de rastreamento de ônibus, porém não há nenhuma área de desenvolvedor. Ou seja: não há API nem documentação. Porém é possível fazer engenharia reversa no site de rastreamento para fazer um wrapper e coletor de dados igual o feito com o do Olho Vivo.
Originalmente, uma das ideías do projeto era de calcular os horários de saída do ônibus através dos dados coletados, analisar sua dispersão e comparar com os horários oficiais para verificar a pontualidade. Outra ideía era de

calcular o tempo de viagem médio para cada horário de saída. Isso não foi feito para este trabalho devido a dificuldade maior que o esperado para essa ideía, porém é uma possibilidade de análise para uma segunda etapa.

9.3 Sugestões e críticas

Há análises potencialmente interessantes que podem a vir gerar mais conclusões, como por exemplo extender a análise dos horários de picos para o horário do almoço bem como os dias de sábado e domingo. Foi desenvolvido uma gama de filtros e representações que ainda podem ser utiizadas de maneiras diferentes.
A distribuição dos ônibus em velocidades baixas apresenta potencial de estudo, pois suas fontes de erros não estão bem caracterizadas.
A análise espacial desse trabalho teve resultados aquém das expectativas. Apesar de possuir amplas aplicações, os mapas matriciais não foram eficazes em demonstrar o fluxo centro-periferia nos horários de picos. Talvez haja alguma forma melhor de demonstrar esse fluxo, e essa é uma sugestão de melhora.

10 Referências

Referências

[1] Wing, Michael G.1; Eklund, Aaron2; Kellogg, Loren D., Consumer-Grade Global Positioning System (GPS) Accuracy and Reliability, Journal of Forestry, Volume 103, Number 4, June 2005, pp. 169-173(5),

[2] SPTrans, Documentação da API do Olho Vivo, `http://www.sptrans.com.br/desenvolvedores/APIOlhoVivo/Documentacao.aspx?1`

[3] William J. Hughes Technical Center, Global Positioning System (GPS) Standard Positioning Service (SPS) Performance Analysis Report, `http://www.nstb.tc.faa.gov/reports/PAN86_0714.pdf#page=22`

11 Apêndices

11.1 Gráficos de algumas regiões da cidade

Figura 15: Centro

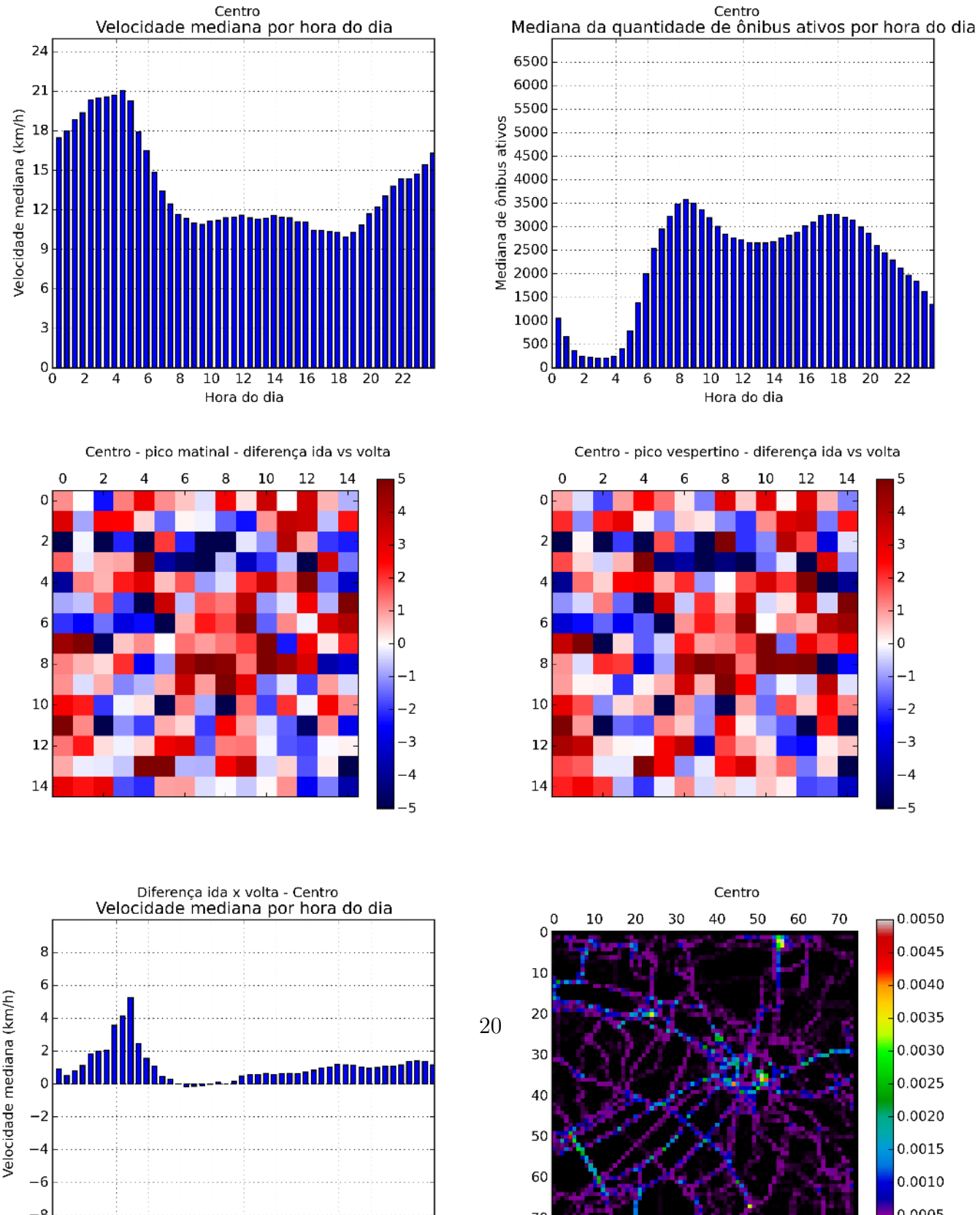
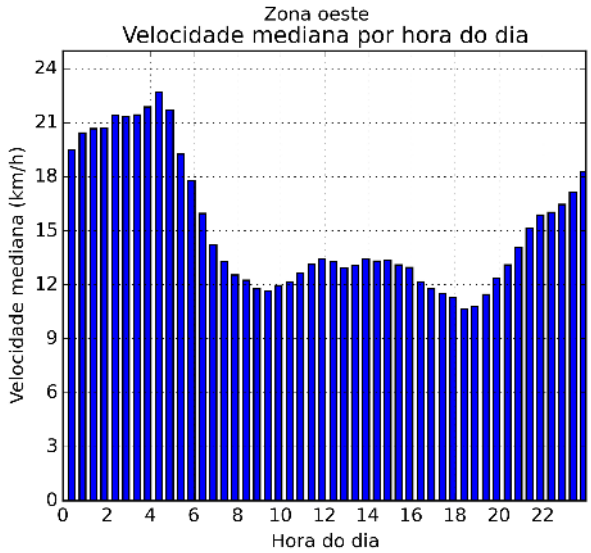
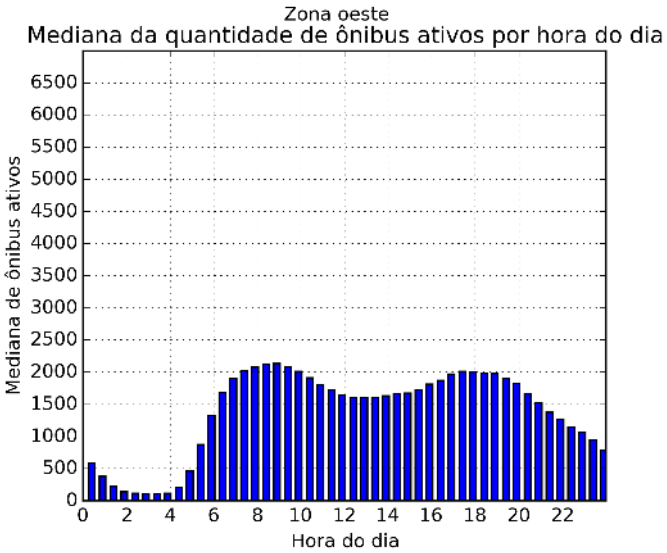
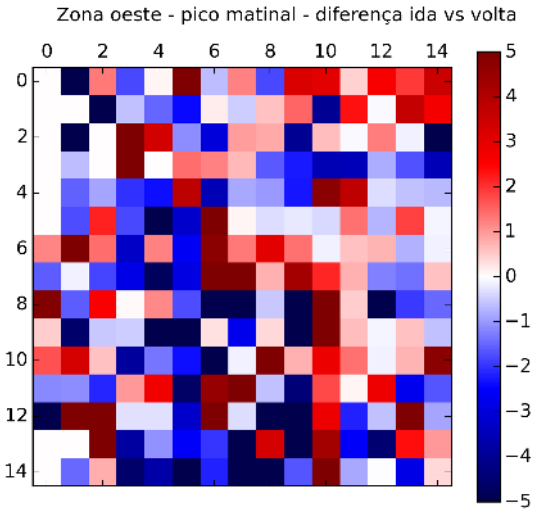
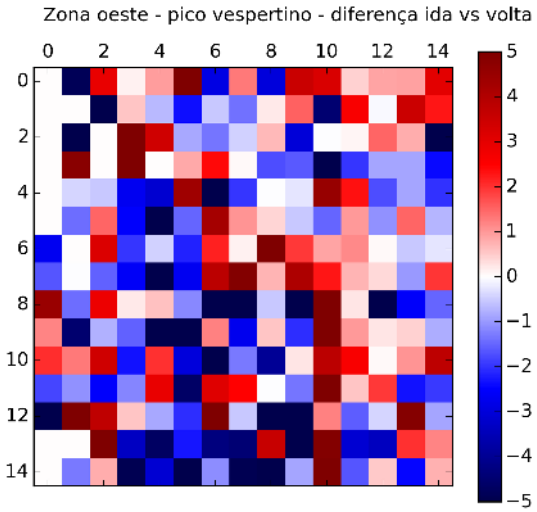


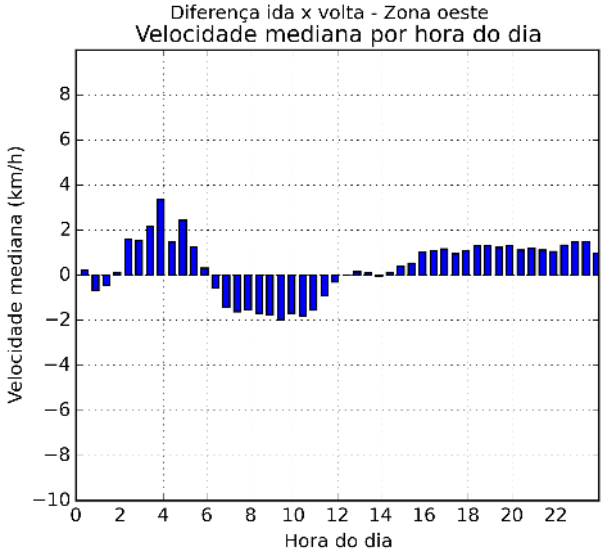
Figura 17: Zona oeste











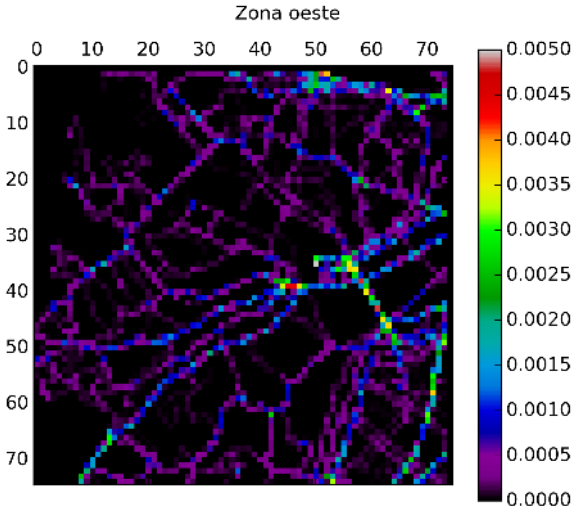
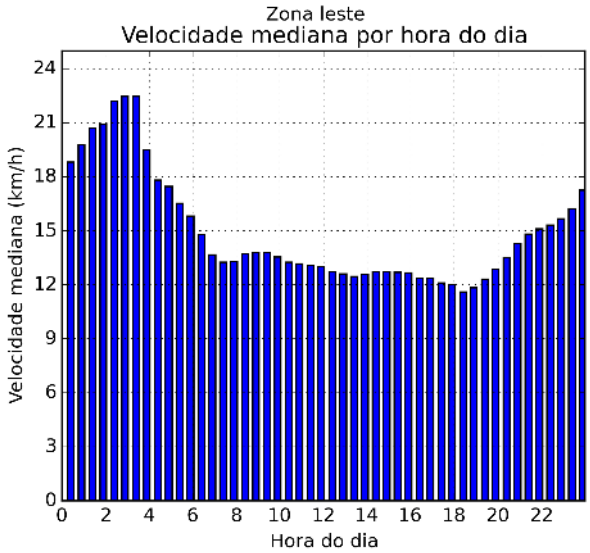
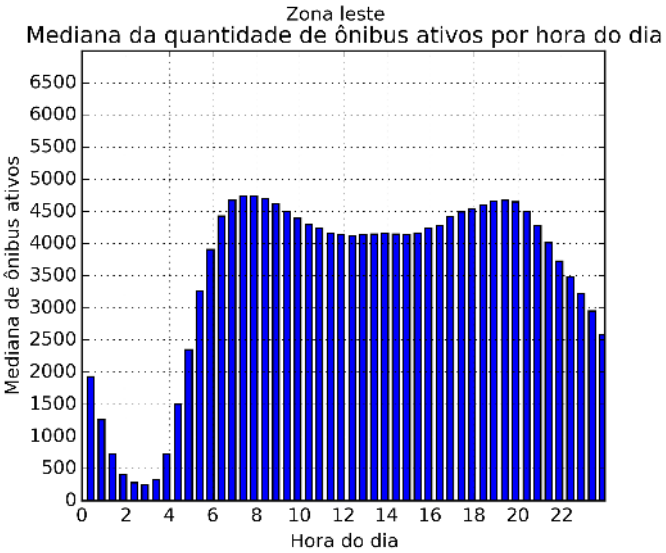
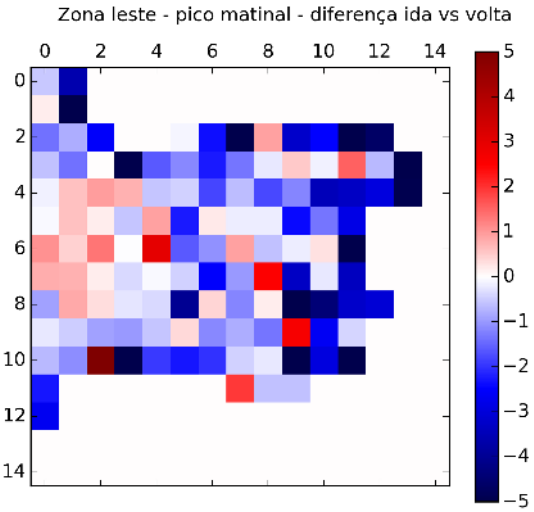
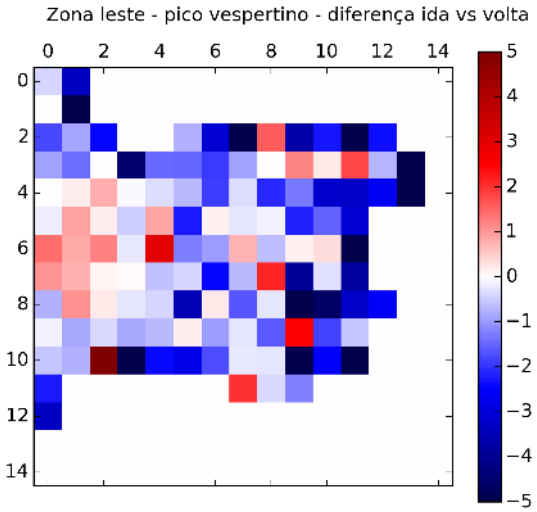


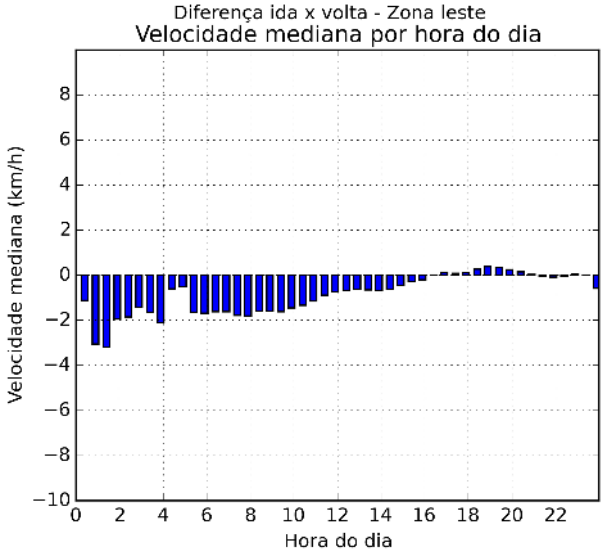
Figura 19: Zona leste











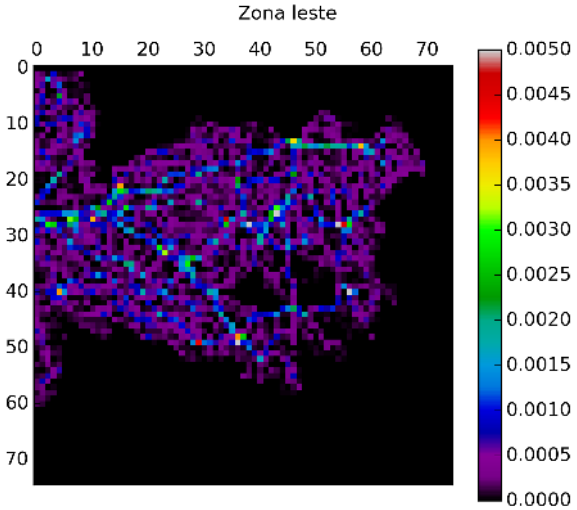
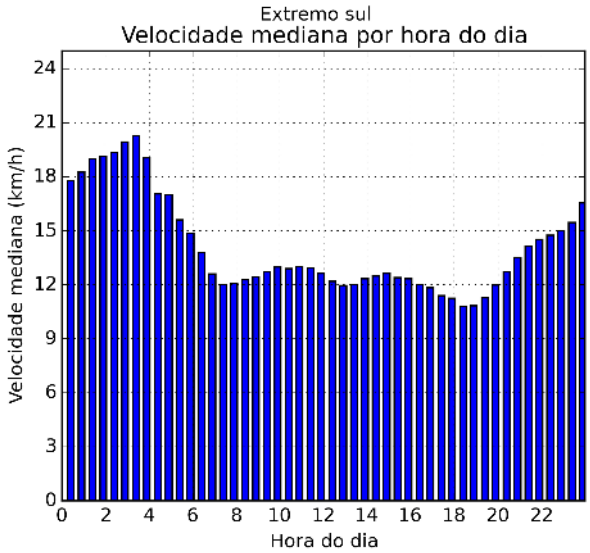
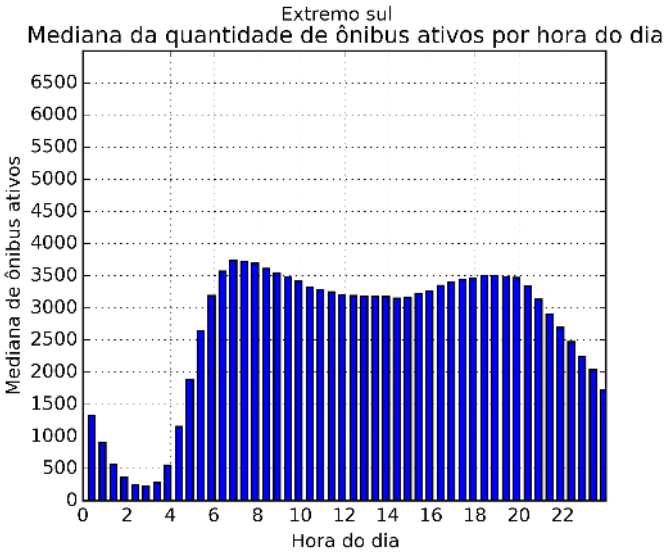
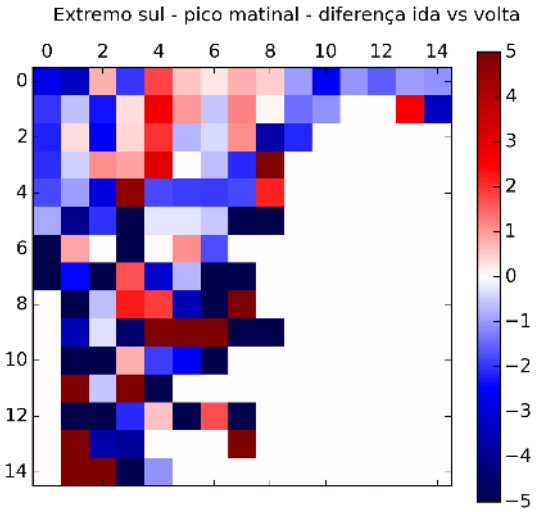
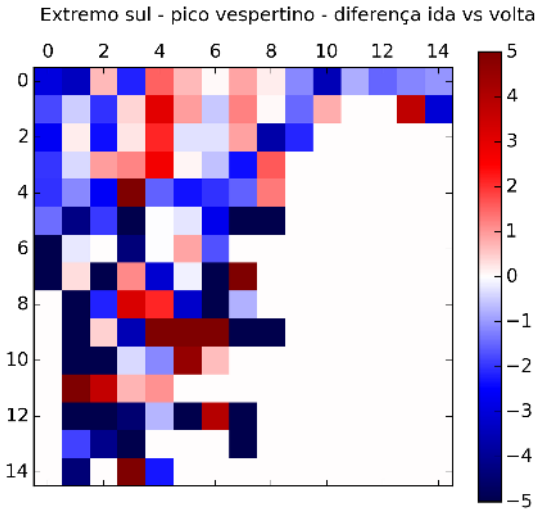


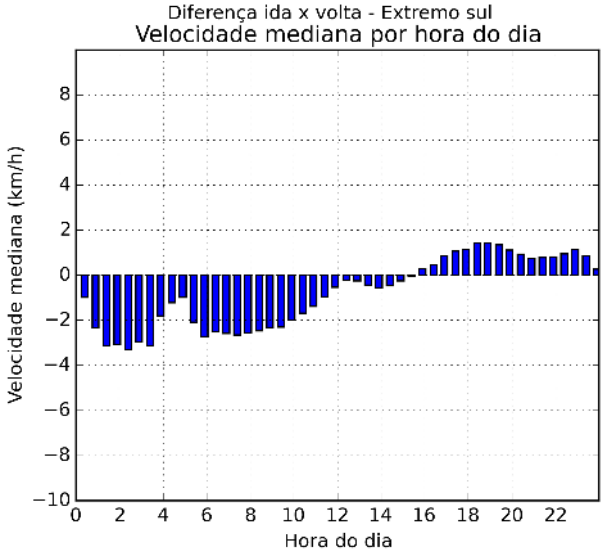
Figura 21: Extremo sul

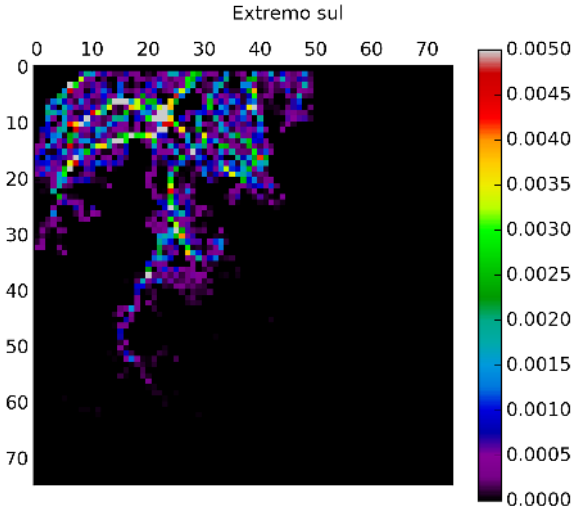












11.2 Horários de pico - mapas matriciais

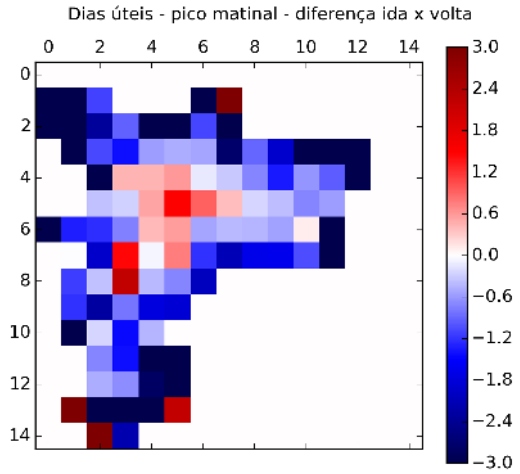
Nessa seção, busca-se visualizar e verificar espacialmente os congestionamentos ou atrasos decorrentes dos horários de picos matinais.
O horário de pico matinal será considerado como sendo das 7:30h até 8:30h, enquanto o horário de pico vespertino será das 17:30h até 18:30h, sendo que ambos os horários podem ser visualizados como sendo os de menor velocidade mediana na fig. **??**, pág 12.
Fez-se então mapas matriciais, tanto para o horário de pico da manhã e quanto para o pico da tarde, de diferença entre da velocidade mediana a "ida"e a velocidade mediana da "volta"para cada quadrado de coordenadas em intervalos tal que todos os dados estejam contidos dentro de uma matriz 15x15.[12]
Os dois mapas são relativamente parecidos e homogêneos em sua distribuição. As diferenças parecem estarem acentuadas no horário de pico da manhã, porém seguem essencialmente a mesma distribuição durante a tarde.

11.2.1 Discussão

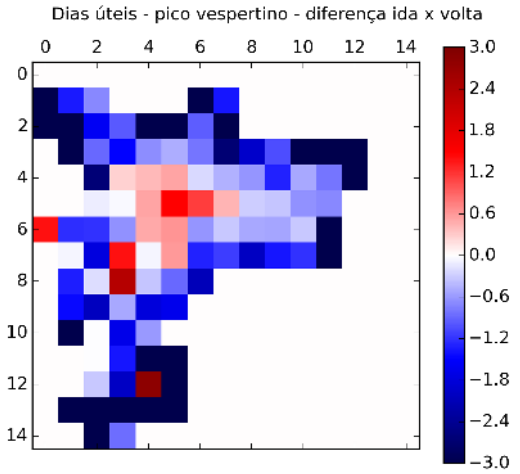
As barras de diferença de velocidade mediana entre os ônibus de ida e volta demonstram haver uma diferença na ordem de 2km/h para o horário de pico da manhã e -1km/h no horário de pico da tarde. Ao fazer os mapas de diferença, imagina-se que haja uma inversão em relação do mapa da manhã ao da tarde, pois na manhã deve haver um movimento predominante rumo ao centro, enquanto na tarde deve ter rumo a periferia. Porém isso não acontece.
Espacialmente, a distribuição de diferença na maior parte da cidade é idêntica tanto de manhã quanto de tarde, o que contradiz o esperado de haver inversão na distribuição.
O que causa isso? Talvez o fluxo de ônibus não é tão unidirecional quanto o esperado. Sendo assim, espacialmente, na maior parte da cidade, a lentidão deve ocorrer de forma homôgenea tanto no sentido de ida quanto o de volta, o que torna díficil uma análise por meio de mapas matriciais.
Porém o fluxo centro-periferia é vísivel a partir de análises de barras de velocidades em regiões específicas[13]

[12] Foi feito também dentro de uma matriz 125x125, ver apêndice
[13] ver apêndice de regiões específicas, pág 21

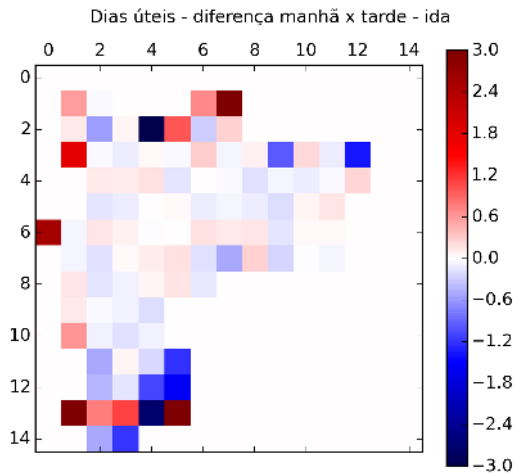
Figura 23: Mapas de diferenças de velocidade



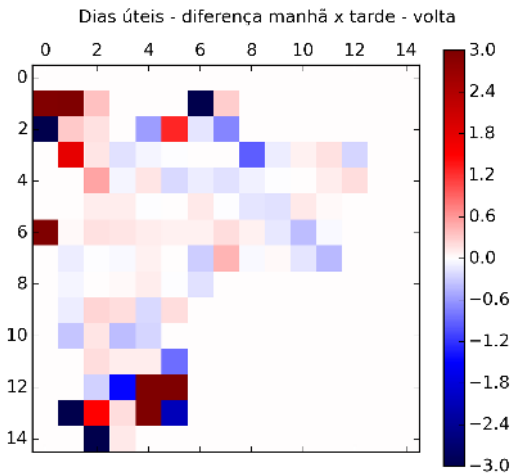
(a) Diferença no horário de pico matinal



(b) Diferença no horário de pico vespertino

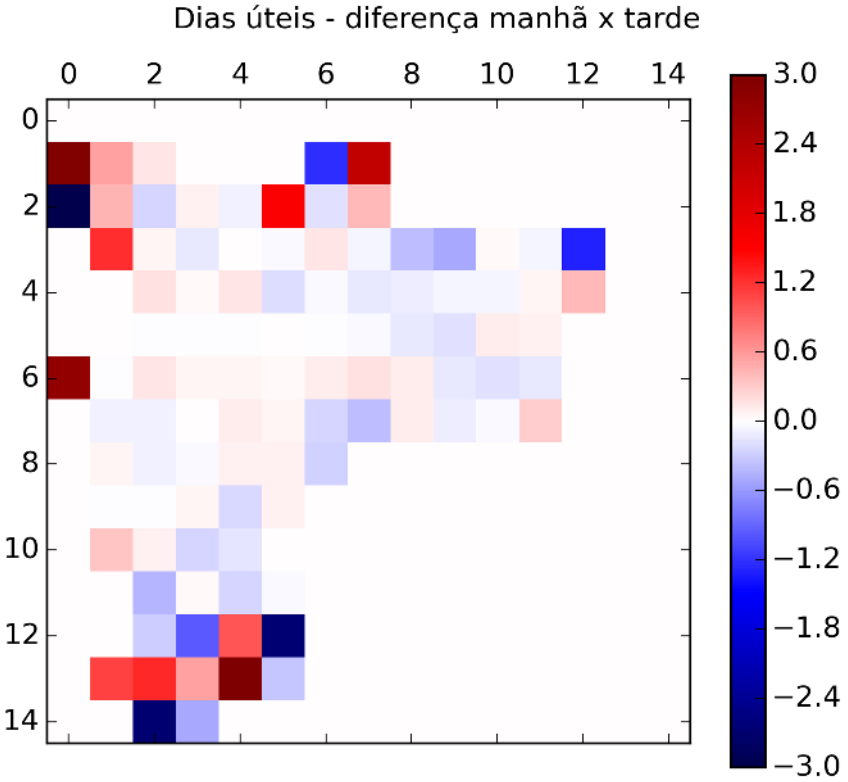


(c) Diferença nas linhas de ida

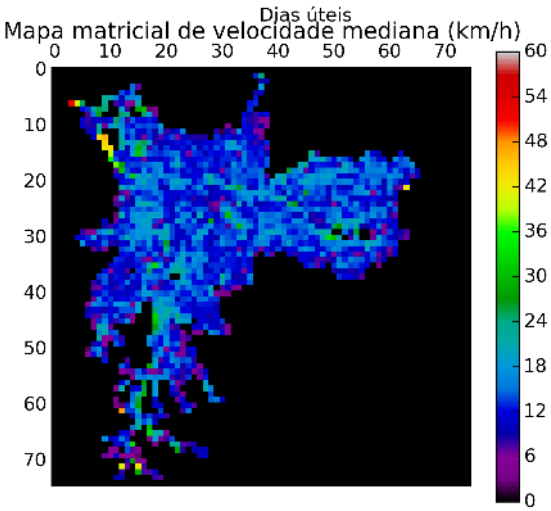


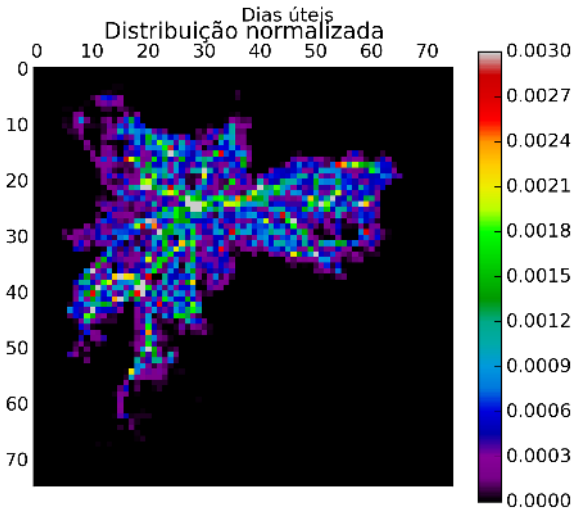
(d) Diferença nas linhas de volta

(a) Diferença entre a velocidade mediana de manhã e a da tarde



11.3 Alguns mapas gerados


Dias úteis
Mapa matricial de velocidade mediana (km/h)
0 10 20 30 40 50 60 70
0 10 20 30 40 50 60 70
60 54 48 42 36 30 24 18 12 6 0


Dias úteis
Distribuição normalizada
0 10 20 30 40 50 60 70
0 10 20 30 40 50 60 70
0.0030 0.0027 0.0024 0.0021 0.0018 0.0015 0.0012 0.0009 0.0006 0.0003 0.0000