Fundação Getulio Vargas
Mestrado em Modelagem Matemática

Visualização da Informação

Explorando Visualizações de Redes com dados de Mobilidade Urbana no Rio de Janeiro

Aluna: Laura Sant'Anna

Professora: Asla Medeiros e Sá

Rio de Janeiro

Dezembro, 2016

1 Motivação

Andy Kirk introduz em seu livro *Data Visualisation: A Handbook for Data Driven Design* ideias de como formular um projeto de visualização de informações. Segundo o autor, precisamos escolher um tema que nos aguce a curiosidade por qualquer que seja o motivo. A partir daí definimos que tipo de informação queremos extrair a respeito deste tema, o que já temos e o que precisamos desenvolver para entendê-las.

Este projeto, em particular, se baseia no projeto de pesquisa e possível tema de dissertação que estou desenvolvendo desde o início do mestrado: mobilidade urbana. O principal objetivo é desenvolver visualizações que facilitem a compreensão da movimentação rotineira das pessoas no Rio de Janeiro e mostrem possíveis padrões regionais de escolha intermodal.

2 Dados

Neste projeto, escolhi explorar algumas bases de dados do Plano Diretor de Transporte Urbano da Região Metropolitana do Rio de Janeiro (PDTU-RMRJ). Este estudo foi realizado pela Secretaria Estadual de Transportes e a Companhia Estadual de Transportes e Logística em 2012 com o objetivo de auxiliar o Governo do Estado no desenvolvimento de políticas públicas de mobilidade urbana. O acesso aos dados foi disponibilzado através da FGV Projetos.

2.1 Descrição

Entre as muitas bases de dados geradas pelo PDTU, apenas as Pesquisas de Interceptação Individual serão abordadas neste trabalho.

Essas pesquisas foram realizadas nas ruas e estações de Metrô, Trem e Barcas da região metropolitana. Foram questionadas as seguintes informações sobre cada deslocamento interceptado: hora, motivo, sexo, faixa etária, duração da viagem, informações geográficas de origem e destino, método de pagamento, modal, modal utilizado antes e depois. No total foram entrevistadas 103.052 pessoas.

2.2 Tratamento

Os datasets de interceptação individual foram disponibilizados separadamente por modal, e por algum motivo, nem todas as informações aparecem em todos, então o trabalho teve que se restringir às informações em comum: origem, destino, motivo e hora para que os datasets pudessem ser agregados. O maior problema aqui foi a variável com o fator de expansão populacional: o dataset com as entrevistas realizadas em estações de Trem foi divulgado com duas colunas de fatores de expansão, enquanto os outros tinham apenas uma. Assim, a análise terá que se restringir à amostra.

Após juntar todos em um dataset único, foi necessário fazer a padronização das informações declaradas (os nomes de bairros e municípios foram escrito de inúmeras maneiras diferentes) e excluir linhas com informações não declaradas.

Ao final da limpeza o dataset ficou da seguinte maneira:

	Hora	Motivo	Modal	Ocupantes	Orig_Mun	Orig_Bairro	Orig_RA	Orig_AP	Dest_Mun	Dest_Bairro	Dest_RA	Dest_AP
0	9.0	Lazer	Auto	2.0	Rio de Janeiro	Madureira	XV RA - Madureira	AP III	Rio de Janeiro	Barra Da Tijuca	XXIV RA - Barra da Tijuca	AP IV
1	8.0	Lazer	Onibus	1.0	Rio de Janeiro	Madureira	XV RA - Madureira	AP III	Rio de Janeiro	Barra Da Tijuca	XXIV RA - Barra da Tijuca	AP IV
2	15.0	Lazer	Onibus	1.0	Rio de Janeiro	Madureira	XV RA - Madureira	AP III	Rio de Janeiro	Barra Da Tijuca	XXIV RA - Barra da Tijuca	AP IV
3	11.0	Lazer	Onibus	1.0	Rio de Janeiro	Madureira	XV RA - Madureira	AP III	Rio de Janeiro	Barra Da Tijuca	XXIV RA - Barra da Tijuca	AP IV
4	12.0	Lazer	Onibus	1.0	Rio de Janeiro	Madureira	XV RA - Madureira	AP III	Rio de Janeiro	Barra Da Tijuca	XXIV RA - Barra da Tijuca	AP IV

Figura 1: Variáveis mantidas no dataset limpo

Com a base de dados limpa, foram construídas matrizes origem-destino para cada modal que serviriam de input para as visualizações pretendidas.

Toda a etapa de tratamento dos dados foi feita utilizando Excel e Python (numpy, pandas, matplotlib).

2.3 Análise Exploratória

Ao fim da etapa de limpeza, comecei a explorar visualmente os dados no Tableau e no Python para entender a amostra e procurar inspiração para o projeto de visualização final.

Para começar, olhei para o tamanho das regiões como origem e destino:

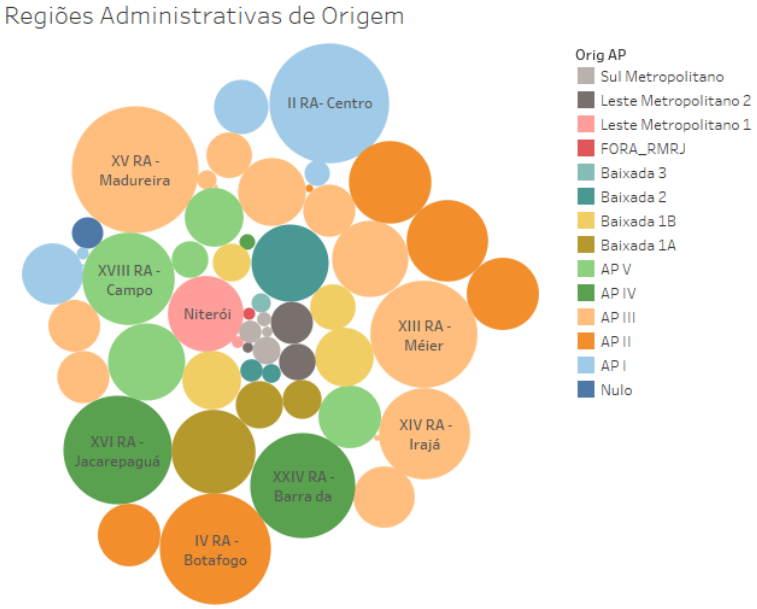

Figura 2: Frequência de cada região como origem

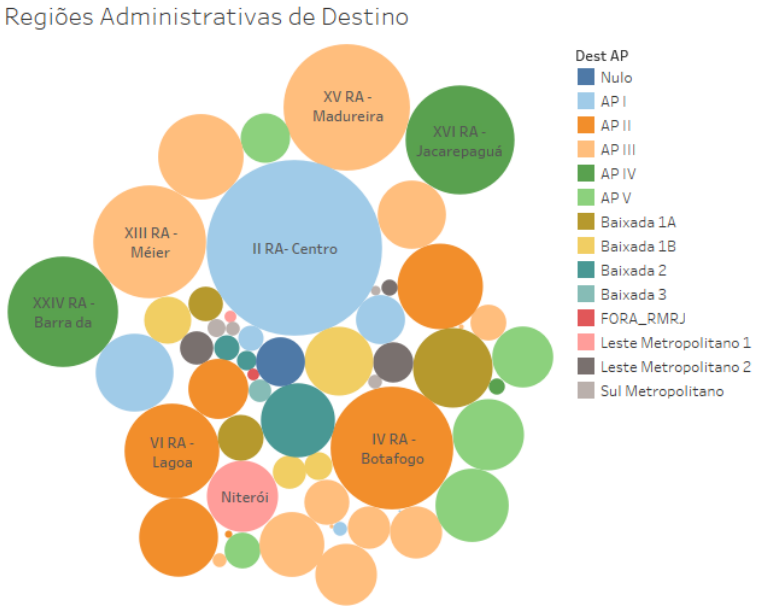

Figura 3: Frequência de cada região como destino

Podemos ver que a região do Centro aparece com grande frequência tanto como origem quanto destino, mas tem maior representatividade como destino. As regiões de Madureira, Botafogo, Méier, Barra da Tijuca, Jacarepaguá e Niterói também aparecem com destaque

como origem e destino. Campo Grande e Irajá parecem ter grande frequência como origem e nem tanta como destino.

As cores nesses dois gráficos foram atribuídas de acordo com a Área de Planejamento identificada pelo Governo Estadual. A princípio, acreditava que essa era uma boa escolha, mas depois vi que resultava em um número muito grande de grupos que dificultavam mais do que ajudavam a compreensão da visualização. No trabalho final, optei por agrupar as regiões administrativas de outra maneira, que será explicada na respectiva sessão.

Outros aspectos que mereciam atenção eram os motivos da viagem declarados pelos entrevistados, a hora e qual modal utilizavam quando foram interceptados. Na amostra, esses aspectos apareceram da seguinte maneira:

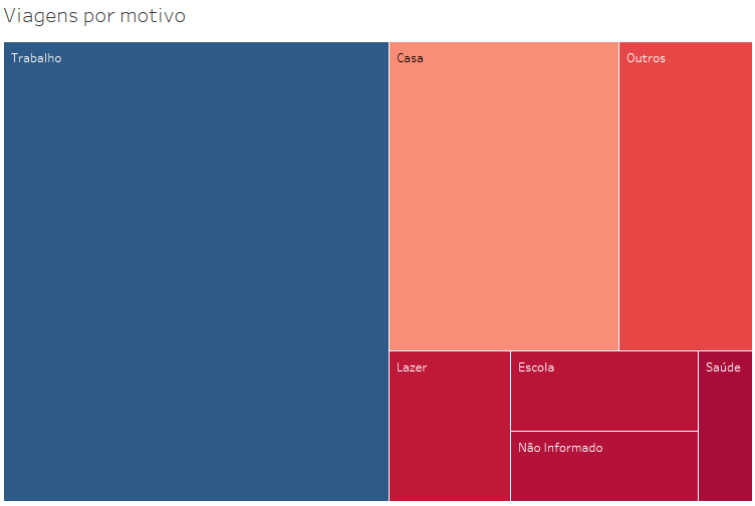

Figura 4: Frequência relativa de entrevista por motivo declarado

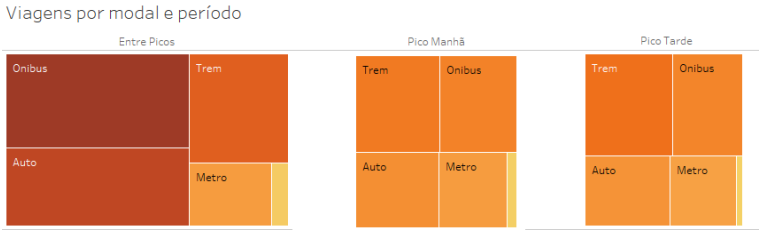

Figura 5: Frequência relativa de entrevistas por faixa de horário de modal

Como já era esperado, o motivo declarado com maior frequência foi Trabalho seguido de

Casa. Essa informação faz sentido junto à observação do gráfico anterior em que a região do Centro aparece mais como destino do que como origem. Provavelmente esses fatores aparecem devido à amostragem. Na prática, a maior parte das pessoas faz o trajeto de ida e volta e isso poderia ser ajustado com o fator de expansão populacional.
A segunda visualização nos dá uma informação interessante: nos horários de pico, os modais Trem, Ônibus, Automóveis e Metro são utilizados em uma proporção bem parecida, mas nos horários entre picos a utilização relativa de ônibus e automóveis é substancialmente maior. As barcas tem frequência relativa menor em qualquer horário, já que sua utilização é de fato limitada a poucas regiões.

Embora já satisfeita com essas informações extraídas do dataset, eu ainda queria encontrar uma maneira de visualizar os dados de acordo com a sua natureza: um grafo direcionado. Até então, só conseguia pensar na visualização padrão de um grafo que aprendemos na aula de Modelagem Mineração de Dados utilizando o Gephi. Mas a rede formada por esses dados é muito densa e a visualização gerada pelo Gephi não diz absolutamente nada:

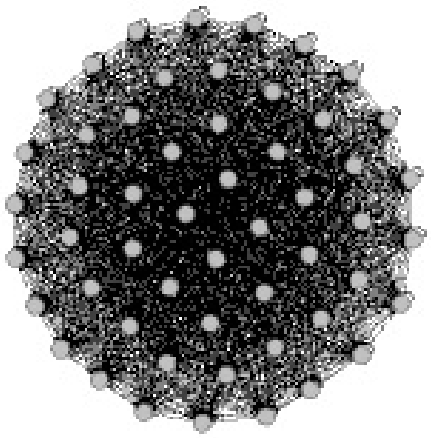

Figura 6: Grafo gerado pela matriz origem-destino dos dados usando Gephi(ForceAtlas2)

Decepcionada com o grafo, comecei a buscar por alternativas de visualizações que representassem de maneira eficiente a matriz origem destino, isto é, que conseguisse mostrar com clareza a intensidade das viagens. Então finalmente conheci os *heatmaps* e os *chord diagrams* que pareciam perfeitos para o que eu queria mostrar. Usando apenas Python ou Tableau,

não consegui chegar a uma primeira versão de *chord diagram*, mas consegui elaborar um primeiro *heatmap* usando o pacote Seaborn:

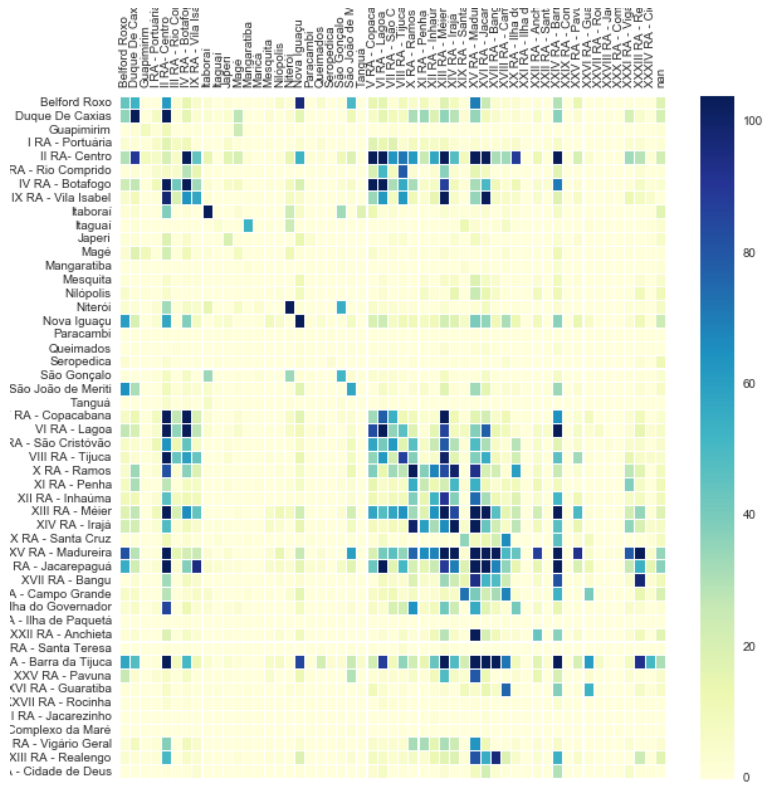

Figura 7: Matriz origem-destino das viagens de ônibus na Região Metropolitana do Rio de Janeiro

Para uma versão inicial, gostei do resultado, mas acho que a informação passada pelo gráfico ainda não ficou muito clara. Alguns aspectos que precisam ser melhorados são:

- A organização da matriz por ordem alfabética não revela padrões característicos dos agrupamentos de regiões - por exemplo, seria interessante que os municípios da região metropolitana formassem um grupo e as zonas da cidade do Rio de Janeiro outros.
- Os nomes das regiões estão muito poluídos e não acrescentam informação (X RA não diz nada, basta aparecer Santa Cruz, por exemplo).
- Está difícil relacionar a célula colorida às respectivas origem e destino. Seria legal que o número de viagens correspondente à célula aparecesse quando o mouse passa por cima.

3 Ideias e Referências

Assim que acabei a versão do *heatmap* apresentada na sessão anterior, me lembrei da visualização que vimos em sala desenvolvida por Mike Bostock para a coocorrência de personagens na obra *Os Miseráveis* de Victor Hugo:

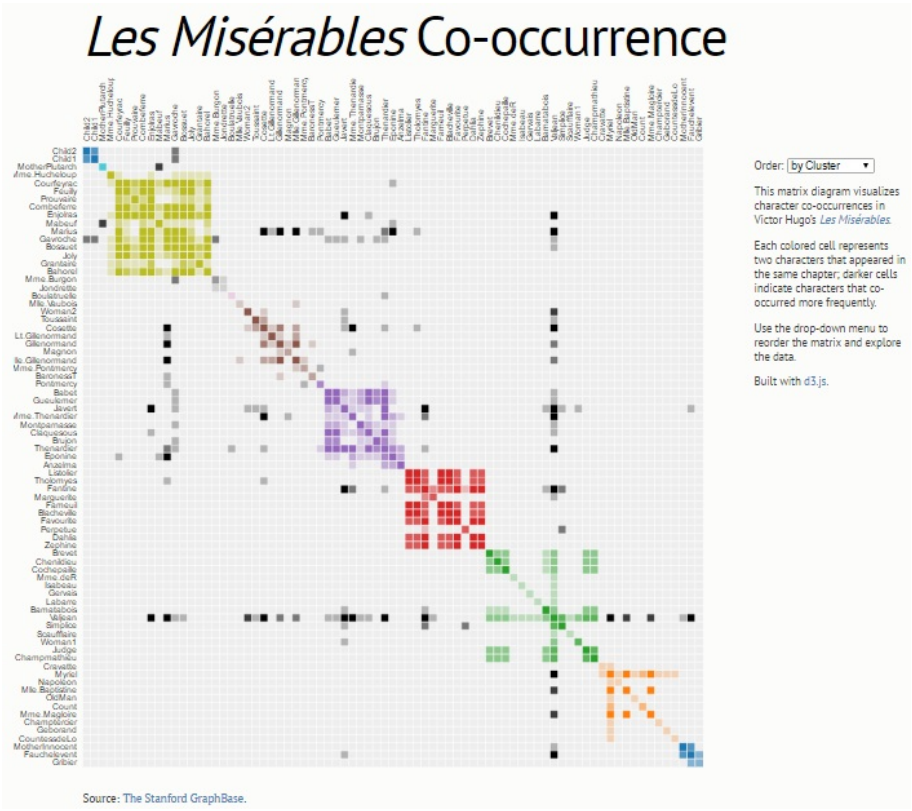

Figura 8: Matriz de coocorrência de personagens em *Os Miseráveis* desevolvida por Mike Bostock

Este exemplo é interessante porque atribui cor aos grupos dentro do enredo e permite interatividade na ordenação: o usuário pode escolher se quer ordenar a matriz por nome, frequência ou por grupos.

Mesmo não conseguindo fazer uma versão inicial de *chord diagram* no Python e no Tableau, procurei exemplos desenvolvidos no Processing ou D3. Então, encontrei o tutorial *Chord Diagrams in D3* escrito por Steven Hall, e gostei principalmente deste exemplo:

Figura 9: Viagens de Uber entre bairros de São Francisco desenvolvida por Steven Hall

Esta alternativa me chamou a atenção porque atribui cores de acordo com a região de localização do bairro e também é interativa: ao posicionar o mouse em um bairro, aparecem apenas as viagens com origem e destino naquele bairro e o número correspondente em uma caixa de texto, assim o usuário pode obter detalhes de viagens que sejam de seu interesse.

4 Resultado

Com a inspiração das visualizações citadas acima, o trabalho restante envolvia apenas pensar em adaptações para os dados de mobilidade urbana e mudar a estrutura dos dados. Para as duas visualizações foi necessário adaptar a forma e converter os dados de csv para json - no primeiro caso para um dicionário e no seguo para um array. Ambas as visualizações foram desenvolvidas usando D3.js.

Na visualização de *heatmap*, além da opção de ordenação, decidi incluir a opção de escolha do modal. Isso porque os modais disponíveis em cada região do Rio de Janeiro variam bastante, e pode ser do interesse do usuário observar cada matriz separadamente.

Além, disso, como a matriz ficou muito grande (53x53), fica difícil olhar a origem e destino

referentes a uma célula específica, então decidi colorir os respectivos nomes de vermelho e uma caixa de texto com as informações de origem, destino e valor no *mouseover* para referenciar e detalhar.

As cores foram atribuídas de acordo com a classificação da região de destino: zonas georeferenciadas do Rio de Janeiro (Central, Norte, Sul ou Oeste), municípios da região metropolitana ou municípios fora da região metropolitana. A legenda para as cores aparece no mapa do lado direito (construído no QGis). A luminância representa a intensidade de viagens entre a origem e o destino. As cores foram escolhidas usando a plataforma Adobe Color CC.

A visualização está disponível em `http://lauragualda.github.io/trabviz/heatmap.html`

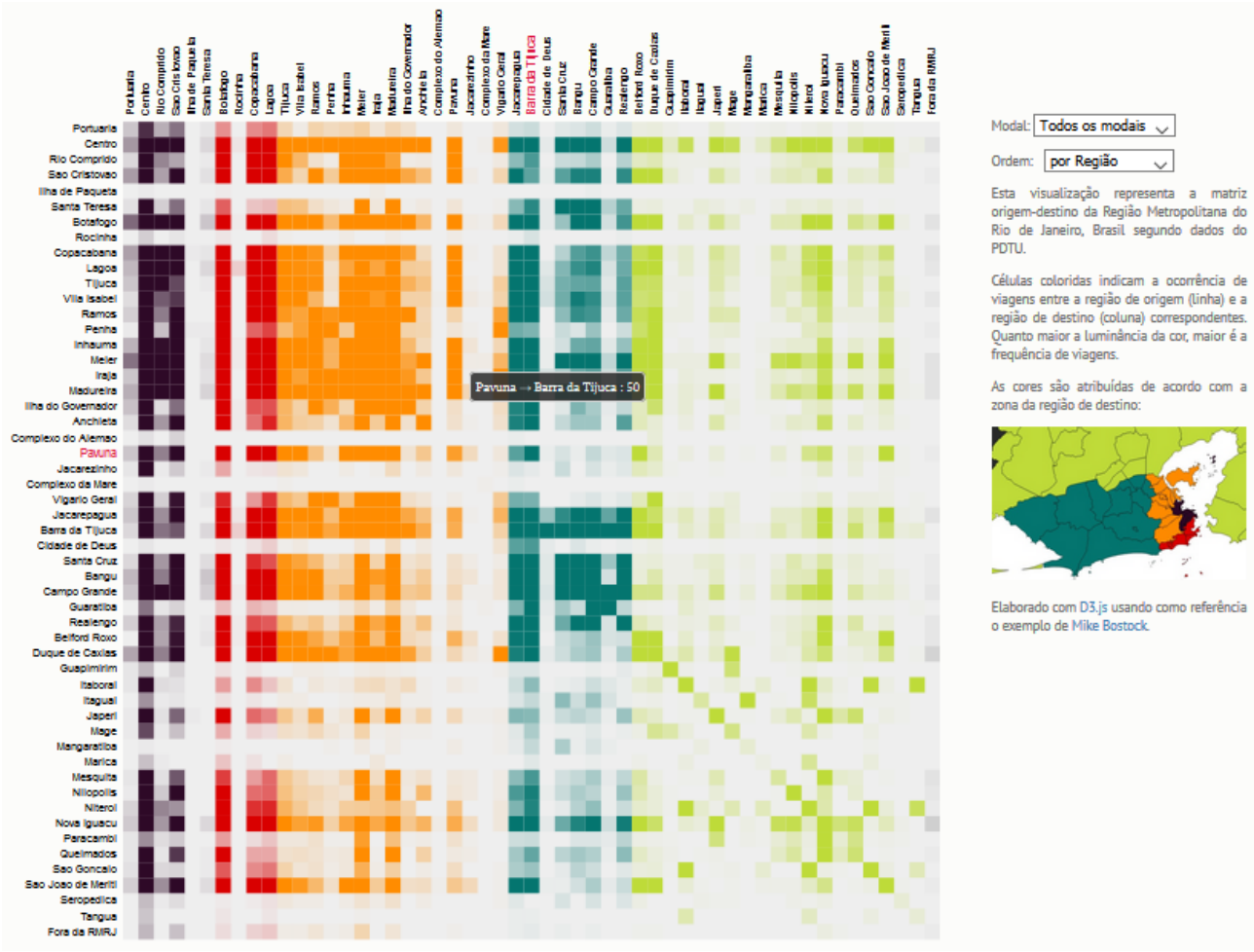

Figura 10: Matriz origem-destino de todos os modais na Região Metropolitana do Rio de Janeiro ordenada por grupos

Embora tenha gostado bastante do resultado final do *heatmap*, a visualização de *chord* parece

ter sido mais fácil de entender - sobretudo para não matemáticos. Em particular, o fato de a visualização inteira caber na tela já ajuda bastante.

Neste gráfico, quanto mais espesso o link entre duas regiões, maior é a frequência de viagens entre elas. A cor é atribuída de acordo com a zona da região de onde saem mais viagens. Por exemplo, o Centro tem muito mais viagens como destino do que como origem, por isso quando todos os links relacionados à região são de outras cores.

Ao passar o mouse sobre alguma região o gráfico foca apenas nas viagens que tiveram aquela região como origem ou destino e ao passar o mouse sobre algum link, aparece uma caixa de texto com a frequência de viagens relativa àquele link.

A visualização está disponível em `http://lauragualda.github.io/trabviz/chord.html`

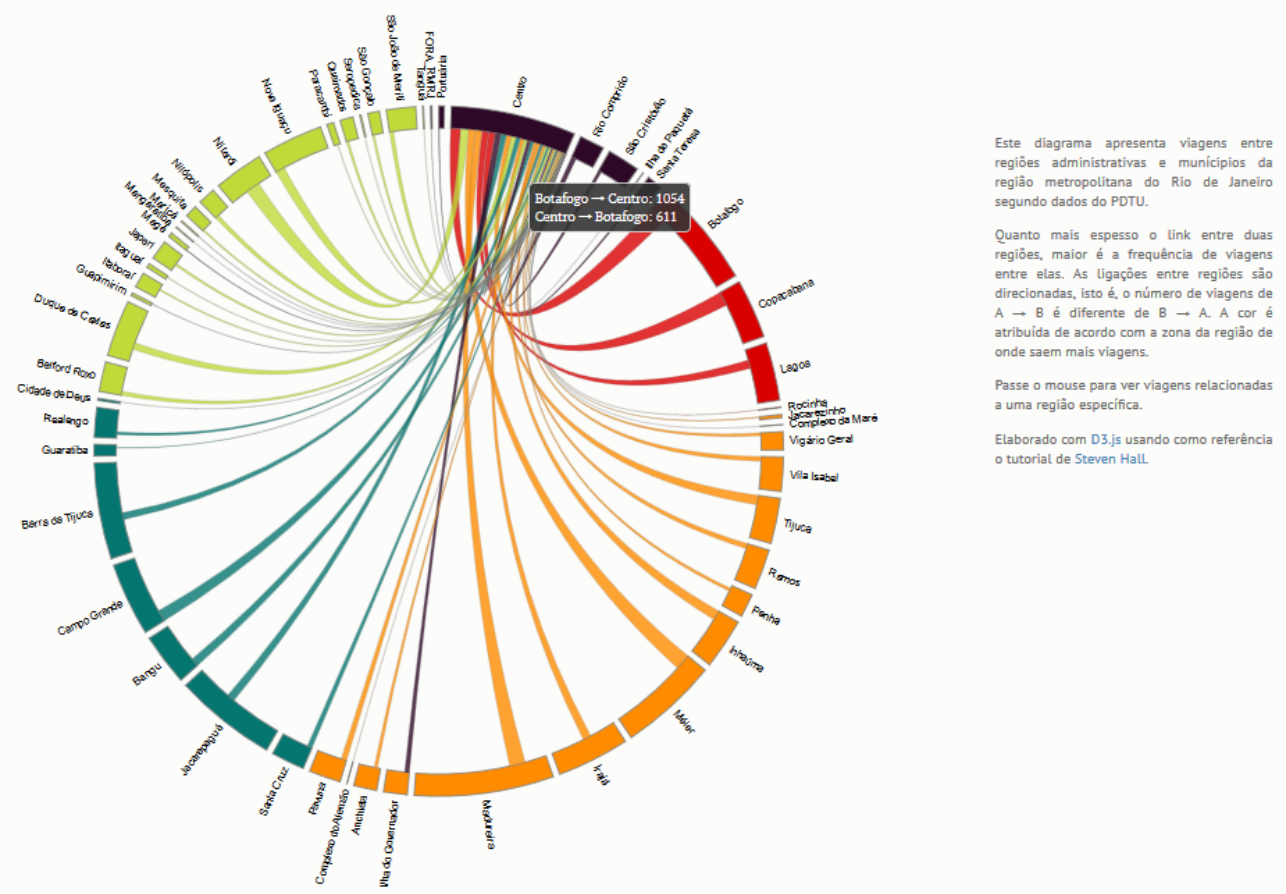

Figura 11: *Chord diagram* com as viagens de todos os modais na Região Metropolitana do Rio de Janeiro

5 Conclusão

Este foi o trabalho mais divertido que já fiz desde os tempos de escola. Sempre gostei muito de extrair informações a partir de todos os tipos de dados, mas nunca tinha empenhado tempo em pensar como apresentá-los para que qualquer pessoa pudesse compreendê-los.
Essa oportunidade foi muito proveitosa tanto para começar a entender o que devo levar em conta sobre a percepção das pessoas, mas também para aprender a construir visualizações realmente bonitas com um novo ferralmental - até então nunca tinha usado o Tableau, o QGis e muito menos programado em JavasSript. Essa última parte, apesar de dolorosa, também valeu totalmente o esforço.
Com certeza essas são só as primeiras visualizações desenvolvidas com o aprendizado adquirido durante o curso.

6 Referências

1. KIRK, A. *Visualising Data: A Handbook for Data Driven Design* - Página oficial do livro disponível em `http://book.visualisingdata.com/`.

2. Hall, S. *Chord Diagrams in D3.* Disponível em `http://www.delimited.io/blog/2013/12/8/chord-diagrams-in-d3`.

3. Bostock, M. *Les Miserables Co-occurrence.* Disponível em `https://bost.ocks.org/mike/miserables/`.

4. D3.js - Página oficial. Disponível em: `http://d3js.org/`.

5. Tableau - Disponível em: `http://www.tableau.com/pt-br`.

6. Gephi - Disponível em: `https://gephi.org/`.

7. QGis - Disponível em: `http://www.qgis.org/en/site/`.