PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO DE JANEIRO
DEPARTAMENTO DE ECONOMIA

MONOGRAFIA DE FINAL DE CURSO

DETERMINANTES ESPACIAIS DA CRIMINALIDADE NA CIDADE DO RIO DE JANEIRO

Igor da Silva Carvalho
N° de matrícula: 1211897

Orientador: Pedro Carvalho Loureiro de Souza

Junho 2017

PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO DE JANEIRO
DEPARTAMENTO DE ECONOMIA

MONOGRAFIA DE FINAL DE CURSO

DETERMINANTES ESPACIAIS DA CRIMINALIDADE NA CIDADE DO RIO DE JANEIRO

Igor da Silva Carvalho
N° de matrícula: 1211897

Orientador: Pedro Carvalho Loureiro de Souza

Junho 2017

Declaro que o presente trabalho é de minha autoria e que não recorri, para realizá-lo, a nenhuma forma de ajuda externa, exceto quando autorizado pelo professor tutor.

As opiniões expressas neste trabalho são de responsabilidade única e exclusiva do autor.

Agradecimentos

Aos meus pais, Marlene Maria da Silva Carvalho e Jocemar da Silva Carvalho, pelo carinho, pela atenção e pelo incentivo ao logo da minha vida acadêmica.

A minha irmã e amiga Ana Teresa da Silva Carvalho, advogada brilhante que sempre me aconselhou, me apoiou e me motivou a vida toda.

A minha namorada, Bruna Scoralick Sousa Lisbôa, futura graduada em Letras, que em tantas oportunidades me ajudou a manter o rigor formal da língua portuguesa ao longo deste trabalho; a ela, agradeço de modo especial pela companhia e pela paciência ao longo da graduação.

Ao meu orientador, Pedro Carvalho Loureiro de Souza, pela disponibilidade, e pela oportunidade que me proporcionou ao participar deste projeto.

Sumário

Lista de figuras

Lista de tabelas

Introdução

No Brasil, em especial no Rio de Janeiro, a criminalidade figura entre as grandes mazelas sociais. O elevado índice de crimes do município carioca interfere diretamente no direito de ir e vir dos indivíduos, diminui a qualidade de vidas dos mesmos e impacta negativamente áreas como educação e comércio. Em linhas gerais, a criminalidade é um fator que retarda o desenvolvimento social na cidade do Rio de Janeiro.

Embora a instituição responsável por prover segurança pública tenha restrições no que diz respeito a recursos humanos, o enfrentamento e a redução do índice de criminalidade não se limita em ampliar o número de policiais militares nas ruas. Tal desafio requer, primeiramente, ferramentas que auxiliem a definição de ações estratégicas eficientes. Sob esta perspectiva, apresenta-se como questão chave o modo de alocação dos agentes de segurança nos diversos pontos da cidade.

Neste contexto, o exercício de analisar o histórico de crimes do município do Rio de Janeiro e entender, ainda que de maneira superficial, o processo gerador destes dados, aparenta ser uma tarefa valorosa que pode proporcionar uma intuição do que potencialmente acontecerá no futuro, o que ao fim ao cabo poderá servir como sugestão da maneira como agentes de segurança deverão ser alocados na cidade.

O objetivo desta monografia é desenvolver um modelo econométrico capaz de gerar previsões um passo à frente de eventos criminais em unidades de área de 250m x 250m para a cidade do Rio de Janeiro. De modo particular, perseguirei a hipótese de que defasagens espaciais são relevantes para fins de explicar a criminalidade em um determinado ponto do espaço; sendo assim, a expectativa é de que, dentre os modelos descritos ao longo deste trabalho, aqueles que incluem defasagens espaciais, ou seja, a incidência de crimes em pontos do espaço adjacentes àquele em análise, como variável explicativa, sejam os melhores para fins de prever eventos criminais um passo à frente.

Duas instituições engajadas na temática de segurança pública participaram do desenvolvimento deste trabalho, o Instituto de Segurança Pública do Estado do Rio de Janeiro (ISP) e o Instituto Igarapé.

O primeiro, trata-se de uma autarquia vinculada à Secretaria de Estado de Segurança do Rio de Janeiro cuja missão é produzir informações, pesquisas e análises tendo em vista a implementação de políticas públicas de segurança. O segundo, o

Instituto Igarapé, trata-se de uma entidade dedicada a produzir e a difundir conhecimentos e estratégias acerca da temática da segurança pública, considerada por esta entidade como fundamental para fins de desenvolvimento social.

A participação destas duas instituições no desenvolvimento deste trabalho se deu em etapas diferentes. A primeira etapa, que diz respeito ao tratamento preliminar dos dados, tal como será descrito do item 1.2, foi desenvolvida junto ao ISP. Já a segunda etapa, que diz respeito ao georreferenciamento dos dados dos incidentes criminais, bem como o desenvolvimento do modelo preditivo, tal como serão descritos, respectivamente, nos itens 1.3 e 2, foi desenvolvida junto ao instituto Igarapé.

Outras instituições já se propuseram a desenvolver uma ferramenta de previsão de eventos criminais em unidades refinadas de espaço para a cidade do Rio de Janeiro, no entanto, dado o caráter privado destas instituições, a metodologia por trás das ferramentas por estas desenvolvidas não está publicamente disponível para acesso.

A obra a seguir será apenas a etapa inicial de um projeto, e sobre este ponto é importante ressaltar que os modelos aqui descritos foram desenvolvidos tendo em vista os dados disponíveis; nas próximas etapas deste projeto, tais modelos serão aperfeiçoados e análises alternativas e adjacentes serão exploradas. Para além do desenvolvimento de um modelo preditivo, este trabalho propõe a documentação cuidadosa das técnicas econométricas empregadas, afim de que esteja publicamente à disposição da sociedade.

1. Análise preliminar e tratamento de dados

1.1 Dados utilizados

Por se tratar de previsão de incidentes criminais, os dados a serem utilizados na construção do modelo preditivo dizem respeito aos endereços onde tais incidentes ocorreram. Dado o caráter condicional dos modelos a serem desenvolvidos, a análise será feita com base nos endereços de incidentes criminais dos últimos mil dias de dados disponíveis, o que corresponde aproximadamente a 3 anos de dados, tendo início em 11/07/2013 e fim em 05/04/2016; a opção por este recorte temporal se deve ao fato de que dados defasados para além de 3 anos não são relevantes para fins de modelar a criminalidade no município do Rio de Janeiro.

A utilização destes endereços em um modelo econométrico preditivo requer, por sua vez, o georreferenciamento dos mesmos, que consiste no processo de atribuir coordenadas geográficas a cada um destes endereços.

Nosso primeiro contato com os dados de incidentes criminais se deu através do ISP, que nos disponibilizou uma base amostral de endereços de incidentes criminais. Entre as variáveis contidas na base, segue a listagem daquelas que são relevantes para fins de georreferenciamento:

- Tipo de logradouro, que indica se um dado endereço diz respeito a uma avenida, rua, estrada, praça, dentre outros;
- Logradouro, que indica o nome de uma dada avenida, rua, praça, dentre outros;
- Número, que indica o número do endereço;
- Bairro, que indica o bairro do endereço descrito;
- Município, que indica o município do incidente criminal;
- Complemento, que indica o complemento do endereço quando necessário;
- Referência, que indica um ponto de referência próximo ao endereço descrito.

1.2 Tratamento dos dados

1.2.1 Relevância e justificativa do tratamento

As informações contidas na base de dados disponibilizada pelo ISP são oriundas dos registros de ocorrência dos incidentes criminais. A qualidade com a qual a informação (endereço de ocorrência do crime) é prestada e registrada interfere diretamente no quão bem esta informação será georreferenciada. Por isso, antes de partir efetivamente para a estimação de um modelo preditivo, faz-se necessário um esforço inicial para entender o quão bem estes endereços estão especificados para fins de georreferenciamento.

Em reunião com a equipe do ISP, em 04 de outubro de 2016, tomamos conhecimento do método e das malhas cartográficas que os mesmos utilizam na geocodificação dos endereços de incidentes criminais. Nesta ocasião, também tomamos conhecimento das características dos endereços que são adequadamente geocodificados e daqueles que, em contra partida, não o são, e que, por isso, são considerados residuais. Em linhas gerais, os endereços geocodificados de maneira adequada possuem tipo de logradouro, nome do logradouro e número do logradouro bem especificados, tal como os exemplos da Tabela 1.

Tabela 1: Exemplos de endereços bem especificados.

Tipo de Logradouro	Nome do Logradouro	Número do Logradouro	Bairro	Munícipio	Referencia
Avenida	Presidente Vargas	1733	Centro	Rio de Janeiro	Central do Brasil
Rua	Bebeto	48	Cosmorama	Mesquita	-
Rua	Macaé	13	Promorá	Angra dos Reis	-
Rua	São João	372	Centro	Niterói	-
Rua	Geranios	23	Pantanal	Parati	-

Já endereços não geocodificados possuem má especificação em alguma das informações como tipo, nome e número do logradouro. Na Tabela 2, podemos observar exemplos de endereços com má especificação para número do logradouro e que, por isso, tornam-se residuais no processo de geocodificação.

Tabela 2: Exemplos de endereços mal especificados.

Tipo de Logradouro	Nome do Logradouro	Número do Logradouro	Bairro	Município	Referência
Ladeira	São Felipe	0	Morro da Glória 2	Angra dos Reis	Próximo ao número 560
Sem tipo	Lincon Guimaraes	0	Palmeiras	Cabo Frio	Próximo ao posto de saúde
Rua	Rodolfo Bruno	0	Nogueira	Petrópolis	Em frente ao nº 1015
Rua	José Bento	00	Centro	Armação dos Búzios	Praça Santo Dumont
Rodovia	RJ-124	0	Rio do Limão	Araruama	Próximo ao CBMERJ

Em um modelo preditivo, a quantidade de observações disponíveis da variável que será predita é extremamente relevante para fins de determinação do modelo a ser utilizado, o que justifica um esforço no sentido de recuperar estes endereços considerados residuais.

Uma estratégia para recuperar esses endereços seria, tal como sugerido pelo ISP, utilizar a coluna, da base de dados por eles disponibilizada, que diz respeito ao ponto de referência destes endereços. Uma releitura da Tabela 2 nos permite observar que, embora não tenhamos o número preciso do local onde estes incidentes criminais tenham ocorrido, temos uma indicação aproximada destes locais através da coluna "Referência"; recorrendo às informações deste campo, é possível transformar uma localização vaga, como, por exemplo, "Rodovia RJ-124", tal como na última linha da Tabela 2, em algo mais preciso, uma vez que sabemos que o incidente se deu "Próximo ao CBMERJ".

Assim, ao final da reunião do dia 04 de Outubro, o ISP se comprometeu em nos enviar uma base amostral de endereços que, por má especificação de alguma informação, não puderam ser geocodificados e, em contra partida, nos comprometemos em desenvolver uma ferramenta capaz de extrair do campo "Referência", desta base de dados, informações que se refiram a localidades na expectativa de que, ao utilizá-las no processo de geocodificação, possamos recuperar endereços até então considerados residuais.

1.2.2 Algoritmo utilizado no tratamento dos dados

O passo seguinte, após a reunião com a equipe do ISP, foi desenvolver um algoritmo capaz de separar, para cada uma das entradas do campo "Referência", as informações que tratam de uma localidade específica e eliminar palavras que indicam posicionamento relativo, como "Próximo a(o)" e "Em frente a(o)". Procedemos desta forma pois acreditávamos que palavras as quais designavam posicionamento não eram relevantes no processo de georreferenciamento.

Para desenvolver este algoritmo, contamos com a ajuda de Felipe Luiz Rocha, monografando do Professor Hélio Lopes, ambos do departamento de Informática da PUC-Rio. O *script* construído por Felipe Luiz foi desenvolvido em *Python* (uma linguagem de programação de alto nível) através de uma biblioteca chamada NLTK (*Natural Language Toolkit)*, própria para realizar classificação textual. A funcionalidade desta biblioteca está em sua capacidade de classificar gramaticalmente as palavras da língua portuguesa presentes em um determinado *string* (uma cadeia de caracteres), sendo possível essa classificação através de um treinamento prévio da biblioteca com base em um conjunto de artigos, denominado *mac_morpho*, publicados no jornal Folha de São Paulo, em 1994, que contém mais de um milhão de palavras anotadas com etiquetas classificadoras.

Para fins de descrição, podemos separar o algoritmo de tratamento em três etapas. A primeira etapa, que se refere a um pré-tratamento, torna todos os caracteres dos *strings* minúsculos em maiúsculos e remove todos os acentos gráficos, tanto no campo "Referência" quanto no conjunto de artigos, através do qual é realizado o treinamento da biblioteca NLTK, conforme descrito. É importante lembrar que a base de dados que contém o campo "Referência" é oriunda dos registros de ocorrência de incidentes criminais; sendo assim, é convenientes que acentos sejam eliminados, uma vez que não sabemos se todos os responsáveis por fazer os registros usam adequadamente acentuação gráfica. Além disso, nesta etapa, são feitas expansões de abreviações, nas quais *strings* como "AV" ou "AV." são substituídos por "AVENIDA", por exemplo. Nesta etapa, também são feitas decomposições de conjunções, que se referem à substituição de "DO" por "DE O", por exemplo; e, enfim, também é realizado o treinamento da biblioteca NLTK com base no conjunto de artigos *mac_morpho*.

Na segunda etapa do algoritmo de tratamento, dado o treinamento prévio, a biblioteca NLTK classifica gramaticalmente os *strings* da coluna "Referência". A partir

desta classificação, os dados do campo "Referência" serão analisados sob a forma de etiquetas, as quais, cada uma destas, possui uma palavra que compõe uma dada referência e sua classificação gramatical. Tomando, como exemplo, o endereço da segunda linha da Tabela 2, o campo "Referência", após realizada a classificação gramatical, será representado como no exemplo abaixo:

Exemplo 1:

[('PROXIMO', 'ADJ'), ('AO', 'PREP'), ('NUMERO', 'N'), ('560', 'NUM')]

Os quais, "ADJ", significa um adjetivo, "PREP", indica uma preposição, "N", um substantivo, e "NUM", um numeral.

Revisitando nosso objetivo, o algoritmo aqui descrito pretende extrair do campo "Referência" informações que sejam relevantes para georreferenciar endereços mal especificados. Tendo em vista este objetivo, a partir da classificação gramatical das palavras que compõem os *strings* do campo "Referência", foi possível criar dois grupos de regras para capturar estas informações de interesse.

O primeiro conjunto de regras tem como objetivo separar os termos que compõem o campo "Referência" e, por isso, são denominadas regras de quebra. Uma quebra no *string* "Referência" deve ocorrer caso:

1) Uma preposição não seja precedida por uma palavra pertencente à lista chamada *nonBreakingWords* (cuja descrição, com detalhes, encontra-se na Tabela 3, ao final desta seção), ou, não seja precedida por um substantivo próprio, ou, não seja precedida por outra preposição;

2) Um artigo não seja parte de uma conjunção;

3) Um determinado objeto do *string* faça parte da lista *separators* (cuja descrição, com detalhes, encontra-se na Tabela 3, ao final desta seção).

Tomando como referência o Exemplo 1, esperamos que, após a aplicação das regras de quebra, obtenhamos o resultado abaixo.

Exemplo 2:

[('PROXIMO', 'ADJ')]

[('NUMERO', 'N'), ('560', 'NUM')]

Neste exemplo, a quebra se dá pelo fato da preposição “AO” não ser precedida por uma palavra que pertence à lista *nonBreakingWords*, ou também, pelo fato de não ser precedida por um substantivo próprio.

Já o segundo conjunto de regras tem como objetivo filtrar os resultados obtidos após a aplicação das regras de quebra. Esta filtragem é feita com base em duas listas de palavras; a primeira é denominada *forbiddenWords*, e a segunda *forbiddenSingletonWords* (ambas descritas com detalhes na Tabela 3 ao final desta seção). Para uma lista de termos já quebrados, tal como no Exemplo 2, a filtragem elimina um dado termo caso:

1) Este seja formado por uma única etiqueta que contenha uma palavra pertencente à lista *forbiddenSingletonWords*;

2) Este seja formado por uma ou mais etiquetas que contenha(m) palavra(s) pertencente(s) à lista *forbiddenWords*;

3) Este seja formado por etiquetas que não contenham substantivos, tanto próprios quanto comuns.

Tomando como referência o Exemplo 2, esperamos que, após a aplicação das regras de filtragem, obtenhamos o resultado abaixo.

Exemplo 3:

[(‘NUMERO’, ‘N’), (‘560’, ‘NUM’)]

Neste caso, a filtragem eliminou a primeira linha da lista de termos já quebrados, uma vez que era formada por uma única etiqueta, e a palavra nela contida pertencia à lista *forbiddenSingletonWords*. Também é valido observar que não houve eliminação da segunda linha da lista de termos já quebrados, uma vez que pelo menos uma das palavras que o compõem trata-se de um substantivo.

Por último, na terceira etapa do algoritmo de tratamento é realizada a entrega daquilo que sobrou após a aplicação das regras de quebra e de filtragem, que corresponde ao que acreditamos ser a informação relevante para nos auxiliar na geocodificação de endereços mal especificados. Nesta etapa, a(s) etiqueta(s) sobrevivente(s) da etapa anterior são desfeitas, as conjunções, antes decompostas, são reconstruídas, e, enfim, estas informações são transformadas em um *string*. Dessa forma, ao final do algoritmo, é gerado um *output*, tal como o exemplo abaixo.

Exemplo 4:

PROXIMO AO NUMERO 560

[(‘PROXIMO’, ‘ADJ’), (‘AO’, ‘PREP’), (‘NUMERO’, ‘N’), (‘560’, ‘NUM’)]

> [(‘PROXIMO’, ‘ADJ’)]
> [(‘NUMERO’, ‘N’), (‘560’, ‘NUM’)

>>> “NUMERO 560”

Podemos observar que este resultado corresponde a um resumo dos resultados das etapas descritas anteriormente. Na seção 1.2 do apêndice, foram listados outros exemplos de *outputs* gerados pelo algoritmo de tratamento para que o leitor tenha melhor percepção dos resultados.

Tabela 3: Lista de palavras utilizadas no algoritmo de tratamento.

Nome da lista	Objetos da lista	Descrição da lista
placeTypes	SEM TIPO, RUA, FAVELA, SEM, RODOVIA, AVENIDA, OUTROS, LADEIRA, ESTRADA, BECO, PRAIA, MORRO, PRACA, VIA, RAMAL, LARGO, VIELA, CAMPO, TRAVESSA, LOTEAMENTO, RETORNO, SERVIDAO, ACESSO, ALAMEDA, PARQUE, ARCO, ENTRADA, VILA, FAZENDA, ESCADARIA, CONDOMINIO, TERMINAL, TUNEL, CAMINHO, TREVO, PATIO, CONJUNTO, PONTE, ESCADA, TERRENO BALDIO, SITIO, SUBIDA, VALE, REPRESA, VIADUTO, PASSARELA, COMUNIDADE, COMPLEXO, CIDADE, ZONA, ESQUINA, LOGRADOURO, TRAVESSIA, MATA, ESTACAO, ESCOLA, CASA, POSTO, DISTRITO, BARRACAO.	Lista de palavras possuem significado de lugar
nonBreakingWords	RUA, FAVELA, RODOVIA, AVENIDA, ESTRADA, PRAIA, MORRO, PRACA, VIA, LARGO, CAMPO, TRAVESSA, ALAMEDA, PARQUE, ILHA, VILA, FAZENDA, CONDOMINIO, TERMINAL, TUNEL, CAMINHO, TREVO, PATIO, CONJUNTO, PONTE, SITIO, SUBIDA, VALE, VIADUTO, PASSARELA, COMUNIDADE, COMPLEXO, CIDADE, ZONA, PONTO, BECO, LOJA, FABRICA, BAR, LADEIRA, PEDRA, BAIRRO, ESTACAO, CHACARA.	Lista de palavras que não geram quebra nos *strings* do campo "Referencia", ou seja, que são preservadas por serem consideradas relevantes.
forbiddenWords	CONHECIDA, CONHECIDO, DENOMINADA, DENOMINADO, FATO, LOCALIZADA, LOCALIZADO, INFORMADA, INFORMADO, INFORMAR, INFORMOU, CITADO.	Lista de palavras que, quando acompanhadas ou solitárias após a aplicação das regras de quebra, devem ser eliminadas no processo de filtragem.

forbiddenSingletonWords	Lista ***placeTypes*** somada aos seguintes objetos: FRENTE, AREIA, INTERIOR, NOME, PROXIMO, SN, S/N, LOCALIDADE, MURO, AC, ALTURA, LOCAL, FATO, PROX, BAIRRO, NOME, DIRECAO, PROXIMIDADE, PROXIMIDADES, ARREDORES, INICIO, FINAL, PONTO, OBS.	Lista de palavras que, quando solitárias após a aplicação das regras de quebra, devem ser eliminadas no processo de filtragem.
separators	‘(’, ‘)’, ‘/’, ‘:’, ‘\’, ‘\\’’, ‘"’, ‘""’, ‘`’, ‘```’, ‘.’, ‘,’, ‘-’	Lista de sinais que, quando presentes em um *string*, geram quebra.

1.2.3 Evolução dos resultados gerados pelo algoritmo

Da primeira versão do algoritmo até a versão parcial descrita na subseção anterior, alguns ajustes foram realizados no algoritmo a fim de refinar o resultado gerado.

Dos *outputs* das versões anteriores do algoritmo descrito na seção 1.2.2, vislumbramos oportunidades de melhora no código, como por exemplo:

- Inclusão de novas palavras na lista *forbiddenWords*, como ‘FATO’, cujas melhoras nos *outputs* podem ser observadas no Exemplo 1.b e no Exemplo 3.b da seção 1.2 do Apêndice.
- Inclusão de novas palavras na lista *forbiddenSingletonWords*, como ‘FINAL’, cujas melhoras nos *outputs* podem ser observadas no Exemplo 2.b da seção 1.2 do Apêndice.

Analisando cuidadosamente os *outputs* exemplificados na seção 1.2 do Apêndice podemos observar que, embora o algoritmo seja eficiente ao que se propõe fazer, o mesmo nem sempre é capaz de recuperar todas as informações que, de fato, interessam.

No Exemplo 3.b da seção 1.2 do Apêndice - “LOCAL DE O FATO; RUA DE A LIBERDADE - PROX. 27 - VISTA ALEGRE” -, podemos observar que “RUA DA

LIBERDADE" não foi devidamente extraída do campo "Referência", quando o deveria. Este tipo de imprecisão é consequência do fato de que os *strings* que compõem o campo "Referência" são expressões quase regulares, o que exige constante observação dos resultados gerados no sentido de evitar eventuais perdas de informação. No caso citado anteriormente, a imprecisão do algoritmo se dá uma vez que o sinal ";" ainda não pertence à lista *separators*, pois, caso o fosse, uma quebra seria feita neste sinal e a informação "RUA DA LIBERDADE" seria devidamente extraída.

Vale registrar que, para todas as versões do algoritmo até aquela descrita na subseção 1.2.2, partimos da premissa de que *outputs* com palavras como "BECO", "VALAO" ou "BAR" não seriam relevantes para fins de georreferenciamento, e que poderiam ser descartadas.

1.2.4 Retorno ao ISP e próximas etapas

De posse de uma versão parcial do algoritmo, que julgamos estar gerando resultados satisfatórios, realizamos uma nova reunião com a equipe do Instituto de Segurança Pública no dia 16 de novembro de 2016. Nesta reunião apresentamos os critérios, bem como os resultados parciais, obtidos com a aplicação do algoritmo de tratamento, os mesmos foram criticados pela equipe do ISP que fez sugestões sobre o que falta incorporar ao *script* apresentado, e também definimos os próximos passos neste processo de tratamento de dados.

Dentre as críticas colocadas pela equipe do ISP, vale destacar aquela que diz respeito à eliminação de termos como "PROXIMO" e "ESQUINA", presentes na lista *forbiddenSingletonWords*, bem como localizações que pretendíamos eliminar como "BECO" e "BAR". Ao contrário do que acreditávamos, segundo a equipe do ISP, estas informações são relevantes para fins de geocodificação, o que abre margem para melhorias do *script* de tratamento. À parte das críticas feitas, a equipe do ISP avaliou, de forma geral, os resultados como satisfatórios.

Da reunião do dia 16 de novembro, ficou sob responsabilidade da equipe da PUC:

1) Flexibilizar alguns critérios de filtragem a fim de preservar palavras que indicam posicionamento relativo como "PROXIMO" ou "ESQUINA";

2) Apresentar o *output* final do algoritmo no formato de um conjunto com três de informações, no qual, a primeira entrada do conjunto indicará um posicionamento

relativo, a segunda entrada será um bloco de texto com a informação de interesse, e a terceira será uma categorização deste bloco de texto, segundo os critérios: Logradouro, Bairro, Cidade, Número, Estabelecimento ou Outros. Tomando como referência o Exemplo 4, a expectativa é que o *output* final tenha a estrutura do exemplo a seguir.

Exemplo 5:

>>> [“PROXIMO”, “560”, “NUMERO”]

Também ficou sob responsabilidade da equipe da PUC:

3) Aplicar o algoritmo de tratamento ao campo “Complemento” da base de dados a fim de extrair deste informações relevantes quanto ao número de logradouro.

Já a equipe do ISP, ficou responsável por:

1) incorporar o algoritmo de tratamento nos procedimento usuais do ISP, no sentido de aplicá-lo à coluna “Referência” de uma base maior de endereços residuais na expectativa de que, ao georreferenciá-los novamente, seja possível recuperar endereços, até então, considerados residuais. O feedback desta aplicação é importante, uma vez que este algoritmo foi desenvolvido com base em uma amostra de endereços, assim, interessa saber como o algoritmo responde a um conjunto maior de dados.

2) Enviar, à equipe da PUC, uma lista de bairros por delegacia policial e outras referências úteis para que seja feita a categorização dos *outputs*;

3) Enviar, à equipe da PUC, uma base de dados, com campo “Referência”, aleatorizada; a base que utilizamos até o momento parece ser oriunda de um grupo particular de indivíduos responsáveis por registrar os incidentes criminais, o que, potencialmente, pode fazer com que o algoritmo, quando aplicado, não responda bem a um conjunto maior de endereços.

A versão do algoritmo de tratamento, realizada as pendencias descritas anteriormente, e entregues ao ISP no dia 09 de dezembro de 2016, está descrito na seção 1.1 do Apêndice.

1.3 Georreferenciamento dos dados

O processo de georreferenciamento dos endereços dos incidentes criminais será feito com base na API(*Application Programming Interface*) do Google Maps. A API trata-se de uma interface que permite acessar funções pré-programadas definidas por um determinado fornecedor, que neste caso é o Google.

Neste processo utilizaremos a API do Google Maps para atribuir à cada um dos endereços de incidentes criminais uma coordenada geográfica; esta coordenada será composta por uma longitude, a posição de um ponto do espaço em relação ao meridiano de Greenwich, e uma latitude, a posição de um ponto do espaço em relação a linha do equador. Assim, estaremos transformando informações que até então encontravam-se sob a forma de *strings* em pares de números decimais, o que viabilizará a utilização destes dados em um modelo matemático.

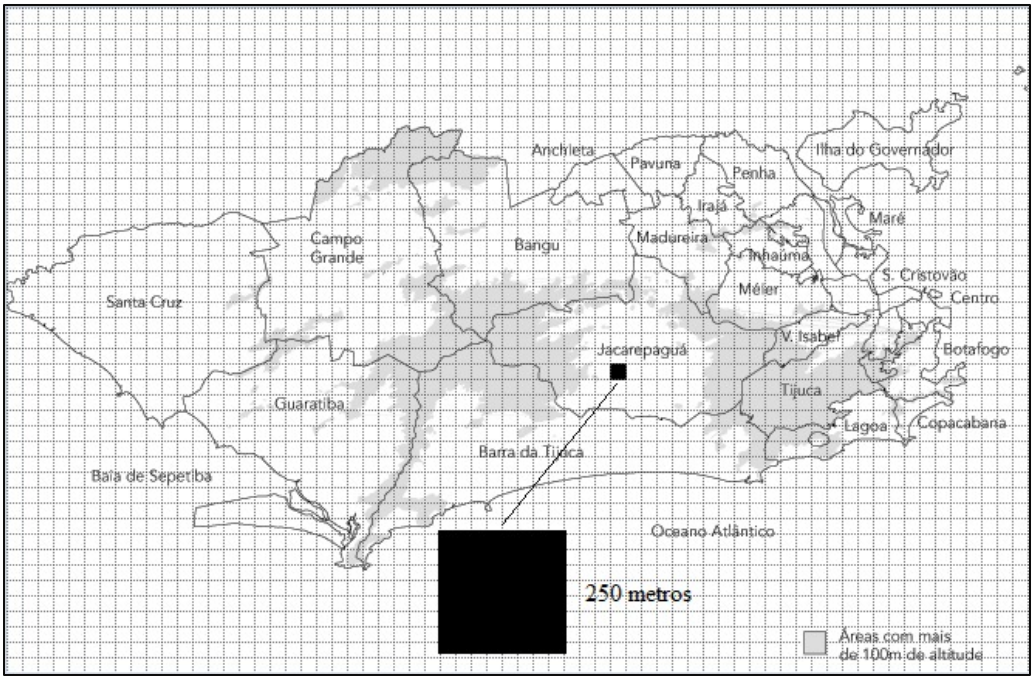
Esta etapa do trabalho foi desenvolvida em parceria com o Instituto Igarapé que possui uma licença, junto ao Google Maps, possibilitando georreferenciar uma grande quantidade de endereços de uma só vez.

2. Desenvolvimento dos modelos preditivos

Os modelos preditivos que serão descritos a seguir têm como objetivo prever o número de incidentes criminais em unidades refinadas de espaço para uma fração de horas do dia, ou período do dia, à frente.

Por se tratar de previsões em unidades refinadas de espaço, o território da cidade do Rio de Janeiro será dividido em quadrados cujo lado mede 250 metros; assim, as previsões de incidentes criminais para um período à frente serão tantas quantos forem o número de quadrados que compõem o território da cidade. A figura a seguir ilustra as unidades de área(ou setores) para as quais as previsões serão geradas.

Figura 1: Divisão da cidade do Rio de Janeiro.



Dada a divisão, o Rio de Janeiro será analisado como uma matriz com 200 linhas e 279 colunas.

A unidade de tempo para o qual as previsões serão geradas diz respeito a uma fração de horas, período, que compõem um dia, ou seja, madrugada(de 00:00h às 06:00h), manhã(de 06:00h às 12:00h), tarde(de 12:00h às 18:00h), ou noite(de 18:00h às 00:00h). Assim, dado que utilizaremos um recorte temporal de mil dias, que vão de 11/07/2013 à 05//04/2016, conforme mencionado na seção 1.1, estaremos trabalhando, ao todo, com quatro mil observações para cada unidade de área da cidade.

2.1 Descrição dos modelos

Em uma análise preliminar, tomando como referência unidade de área hachurada na Figura 1, é plausível supor que a incidência crimes em um determinado período do dia está condicionada aos incidentes criminais ocorridos no período do dia imediatamente anterior; nesta análise, defasagens de tempo são consideradas relevantes para fins de explicar o que acontece no presente e o que potencialmente acontecerá no futuro em um determinado ponto do espaço.

Uma análise mais cuidadosa, por outro lado, nos permite supor que a incidência de crimes na unidade de área em questão está condicionada, também, à incidência de crimes nas unidades de área vizinhas, representadas na Figura 1 pelos quadrados em torno daquele hachurado. Dada esta suposição adicional, estamos considerando defasagens de espaço como relevantes para fins de explicar o que acontece e potencialmente acontecerá em um determinado ponto do espaço.

Em linhas gerais, os modelos descritos a seguir partem da suposição básica de que a incidência de crimes em um determinado ponto do espaço está condicionada à defasagens de tempo, incidência de crimes no período do dia anterior, e à defasagens de espaço, incidência de crimes em pontos do espaço vizinhos.

Modelo 1: Autoregressivo com uma defasagem de tempo(AR1). Este modelo básico será tratado como um *benchmark* para fins de comparação com os demais modelos.

$$y_{i,j,t} = \varphi + \rho_1 y_{i,j,t-1} + \varepsilon_{i,j,t} \qquad \textbf{(1)}$$

Modelo 2: Autoregressivo com quatro defasagens de tempo(AR4).

$$y_{i,j,t} = \varphi + \rho_1 y_{i,j,t-1} + \rho_2 y_{i,j,t-2} + \rho_3 y_{i,j,t-3} + \rho_4 y_{i,j,t-4} + \varepsilon_{i,j,t} \qquad \textbf{(2)}$$

Modelo 3: Autoregressivo com uma defasagem de tempo, e média dos últimos quatro períodos do dia como regressores(AR1Block).

$$y_{i,j,t} = \varphi + \rho_1 y_{i,j,t-1} + \alpha \frac{1}{4} \sum_{s=1}^{4} y_{i,j,t-s} + \varepsilon_{i,j,t} \qquad \textbf{(3)}$$

Modelo 4: Autoregressivo com três defasagens de tempo, e média dos últimos quatro períodos do dia como regressores(AR3Block).

$$y_{i,j,t} = \varphi + \rho_1 y_{i,j,t-1} + \rho_2 y_{i,j,t-2} + \rho_3 y_{i,j,t-3} + \alpha \frac{1}{4} \sum_{s=1}^{4} y_{i,j,t-s} + \varepsilon_{i,j,t} \quad \textbf{(4)}$$

Modelo 5: Autoregressivo com uma defasagem de tempo, variáveis *dummy* de dia de semana, e variáveis *dummy* de período de dia como regressorres(AR1D).

$$y_{i,j,t} = \rho_1 y_{i,j,t-1} + D_t \gamma + H_t \tau + \varepsilon_{i,j,t} \quad \textbf{(5)}$$

Modelo 6: Autoregressivo com quatro defasagens de tempo, variáveis *dummy* de dia de semana, e variáveis *dummy* de período de dia como regressores(AR4D).

$$y_{i,j,t} = \rho_1 y_{i,j,t-1} + \rho_2 y_{i,j,t-2} + \rho_3 y_{i,j,t-3} + \rho_4 y_{i,j,t-4} + D_t \gamma + H_t \tau + \varepsilon_{i,j,t} \quad \textbf{(6)}$$

Modelo 7: Autoregressivo com uma defasagem de tempo, média dos últimos quatro períodos do dia, variáveis *dummy* de dia de semana, e variáveis *dummy* de período do dia como regressores(AR1BlockD).

$$y_{i,j,t} = \rho_1 y_{i,j,t-1} + \alpha \frac{1}{4} \sum_{s=1}^{4} y_{i,j,t-s} + D_t \gamma + H_t \tau + \varepsilon_{i,j,t} \quad \textbf{(7)}$$

Modelo 8: Autoregressivo com três defasagens de tempo, média dos últimos quatro períodos do dia, variáveis *dummy* de dia de semana, e variáveis *dummy* de período do dia como regressores.(AR3BlockD).

$$y_{i,j,t} = \rho_1 y_{i,j,t-1} + \rho_2 y_{i,j,t-2} + \rho_3 y_{i,j,t-3} + \alpha \frac{1}{4} \sum_{s=1}^{4} y_{i,j,t-s} + D_t \gamma + H_t \tau + \varepsilon_{i,j,t} \quad \textbf{(8)}$$

Modelo 9: Autoregressivo com uma defasagem de tempo, média dos últimos quatro períodos do dia, variáveis *dummy* de dia de semana, e variáveis *dummy* de período do dia, e uma defasagem de espaço como regressores(AR1BlockDNeigh).

$$y_{i,j,t} = \rho_1 y_{i,j,t-1} + \alpha \frac{1}{4} \sum_{s=1}^{4} y_{i,j,t-s} + D_t \gamma + H_t \tau + Y_{1,t-1} \beta + \varepsilon_{i,j,t} \quad \textbf{(9)}$$

Modelo 10: Autoregressivo com três defasagens de tempo, média dos últimos quatro períodos do dia, variáveis *dummy* de dia de semana, variáveis *dummy* de período do dia, e uma defasagem de espaço como regressores(AR3BlockDNeigh).

$$y_{i,j,t} = \rho_1 y_{i,j,t-1} + \rho_2 y_{i,j,t-2} + \rho_3 y_{i,j,t-3} + \alpha \frac{1}{4} \sum_{s=1}^{4} y_{i,j,t-s} + D_t \gamma + H_t \tau + Y_{1,t-1} \beta + \varepsilon_{i,j,t} \quad \textbf{(10)}$$

Modelo 11: Modelo que assume o dado observado em t como previsão para t+1(asLag1).

$$y_{i,\,j,\,t} = y_{i,\,j,\,t-1} \qquad \textbf{(11)}$$

Modelo 12: Modelo que assume o dado observado em t-3 como previsão para t+1(asLag4).

$$y_{i,\,j,\,t} = y_{i,\,j,\,t-4} \qquad \textbf{(12)}$$

Nos modelos acima, $Y_{1,\,t-1}$, é uma matriz Nx8 cujo as colunas dizem respeito às variáveis explicativas $y_{i-1,\,j-1,\,t-1}$, $y_{i,\,j-1,\,t-1}$, $y_{i+1,\,j-1,\,t-1}$, $y_{i-1,\,j,\,t-1}$, $y_{i+1,\,j,\,t-1}$, $y_{i-1,\,j+1,\,t-1}$, $y_{i,\,j+1,\,t-1}$, $y_{i+1,\,j+1,\,t-1}$; β é uma matriz 8x1 que contém os parâmetros β_1, β_2, ..., β_8; D_t é uma matriz Nx6 cujo as colunas dizem respeito às variáveis *dummy* D_t^{ter}, D_t^{qua}, D_t^{qui}, D_t^{sex}, D_t^{sab}, D_t^{dom}; γ trata-se de uma matriz 6x1 que contém os parâmetros γ_1, γ_2, ..., γ_6; H_t é uma matriz Nx4 cujo as colunas dizem respeito às variáveis *dummy* H_t^{madr}, H_t^{manh}, H_t^{tard}, H_t^{noit}; por fim, τ trata-se de uma matriz 4x1 que contém os parâmetros τ_1, τ_2, τ_3, τ_4.

Uma vez que estamos analisando a cidade do Rio de Janeiro como uma matriz os subscritos i e j das variáveis dos modelos indicam respectivamente a linha e a coluna da unidade de área em análise. O subscrito t, por sua vez, indica o instante do tempo para o qual os dadas estão sendo analisados. Variações em i e j, como por exemplo i-1 ou j-1, indicam unidades de área vizinhas àquela em análise; já as variações em t, como t-1, indicam defasagens de um período do dia do dado em questão.

A variável dependente $y_{i,\,j,\,t}$, comum a todos os modelos**,** trata-se de um vetor coluna que contém informações a respeito da unidade de área localizada na linha i e coluna j, e cada linha deste vetor contém o número de crimes ocorridos nesta unidade de área para cada instante de tempo observado.

Dado o caráter autoregressivo dos modelos especificados acima, as variáveis explicativas utilizadas nos mesmos correspondem a variável dependente defasada no tempo e/ou no espaço. Assim, a variável explicativa $y_{i,\,j,\,t-1}$ corresponde à variável dependente com dados defasados em um período do dia, já a variável explicativa $y_{i-1,j,t-1}$ também corresponde a variável dependente, no entanto, com dados defasados em um

período do dia e em uma unidade de área para a esquerda. A maior defasagem observada entre os subscritos i e j define a ordem de proximidade da unidade de área analisada na variável explicativa em ralação àquela analisada na variável dependente, sendo assim, a variável explicativa $y_{i-1,\, j,\, t-1}$ contém informações de incidentes criminais defasados em um período do dia para um vizinho de primeira ordem, ou seja, um vizinho que faz fronteira diretamente com a unidade de área em análise. Para evitar equações extensas, as informações referentes aos vizinhos de primeira ordem estão organizadas na matriz $Y_{1,\, t-1}$.

A variável explicativa $\frac{1}{4}\sum_{s=1}^{4} y_{i,\, j,\, t\, -\, s}$, por sua vez, trata-se de um vetor coluna que contém a média da incidência de crimes dos últimos quatro períodos do dia, o que corresponde ao número médio de crimes ocorridos no último dia móvel; a inclusão desta variável explicativa se dá pela crença de que a incidência de crimes em um período do dia está potencialmente condicionada ao número médio de incidentes ocorridos no último dia.

Uma outra variável explicativa utilizada nos modelos, a *dummy* D_t^{ter}, diz respeito a um vetor coluna que assume valor 1 quando um determinado instante do tempo observado é uma terça-feira. O mesmo vale para as variáveis D_t^{qua}, D_t^{qui}, D_t^{sex}, D_t^{sab}, D_t^{dom}, que indicam se um dado instante do tempo trata-se, respectivamente, de uma quarta, quinta, sexta, sábado ou domingo. A inclusão destas variáveis no modelo preditivo se dá pela possível existência de sazonalidade na incidência de crimes entre os dias da semana. Observe que estas variáveis estão organizadas na matriz D_t.

A *dummy* H_t^{madr}, por sua vez, trata-se de um vetor coluna que assume valor 1 quando um determinado instante do tempo observado diz respeito a uma madrugada. De modo semelhante as variáveis H_t^{manh}, H_t^{tard}, H_t^{noit}, indicam se um dado instante do tempo trata-se, respectivamente, de uma manhã, tarde ou noite. A inclusão destas variáveis no modelo se dá pela possível existência de sazonalidade na incidência de crimes entre os períodos do dia. Observe que as *dummies* de período do dia estão organizadas na matriz H_t.

Os termos que multiplicam as variáveis explicativas e que não possuem os subscritos i, j, e t são os parâmetros dos modelos a serem estimados. Estes parâmetros medem o efeito de uma dada variável explicativa sobre a variável dependente.

O parâmetro ρ_1, por exemplo, trata-se do efeito dos crimes ocorridos na unidade de área situada na linha i e coluna j em t-1, sobre os crimes ocorridos na mesma unidade de área em t. E de modo semelhante, os parâmetros ρ_2, ρ_3, ρ_4, recuperam o efeito dos crimes ocorridos, respectivamente, em t-2, t-3, e t-4, sobre os crimes ocorridos em t, para uma unidade de área situada na linha i e coluna j. O parâmetro α, por sua vez, diz respeito ao efeito do número médio de crimes ocorridos no último dia móvel sobre os crimes ocorridos na unidade de área em análise no período t.

Já os parâmetros γ_1, γ_2, ..., γ_6, contidos na matriz γ, capturam o efeito dos dias da semana sobre a incidência de crimes na unidade de área em análise. Estes parâmetros nos possibilitam entender a sazonalidade dos incidentes criminais, ou seja, como o número de crimes varia ao longo de uma semana para uma determinada unidade de área.

De modo similar, os parâmetros τ_1, τ_2, τ_3, τ_4, contidos na matriz τ, recuperam o efeito dos períodos do dia sobre a incidência de crimes na unidade de área em análise, o que irá possibilitar entender a sazonalidade dos incidentes criminais ao longo de um dia.

Os parâmetros β_1, β_2, ..., β_8, contidos na matriz β, tratam-se do efeito dos crimes ocorridos nos vizinhos de primeira ordem(unidades de área fronteiriças) em t-1, sobre os crimes ocorridos na unidade de área em análise em t.

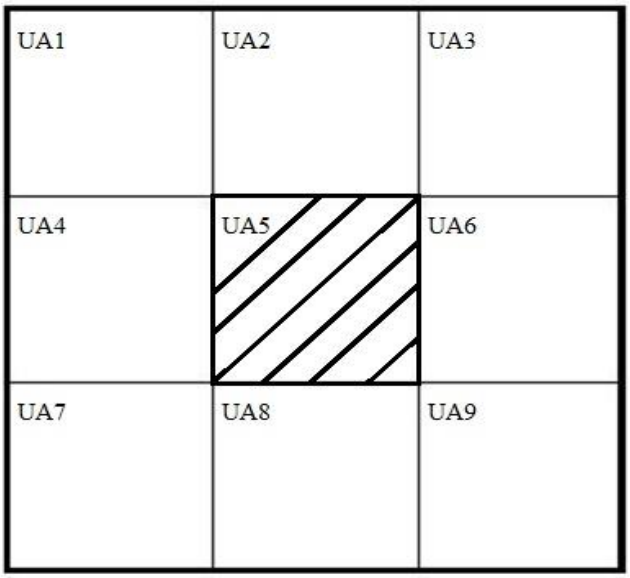
Por fim os parâmetros φ e ε são, respectivamente, o intercepto e o termo de erro dos modelos preditivos; este último deve ser interpretado como o efeito que provém de variáveis explicativas que, uma vez não observáveis, não foram inclusas nos modelos.

2.2 Os modelos adotados frente a modelos alternativos

Nos modelos descritos até o momento as defasagens espaciais são tratadas como variáveis explicativas relevantes, mas até então omitidas, para fins de explicar incidentes criminais em um determinado ponto do espaço. Uma metodologia alternativa para modelar criminalidade seria utilizar os modelos clássicos de econometria espacial que, embora também considerem defasagens espaciais como variáveis explicativas importantes, apresentam duas desvantagens em relação a modelagem descrita na seção 2.1. Para justificar tais desvantagens, simplificaremos o modelo 9 e o compararemos com um modelo econométrico espacial equivalente; para facilitar a análise, vamos supor que ambos os modelos são adequados para descrever a criminalidade na cidade

hipotética Z, cujo território é composto por nove unidades de área, tal como representado na Figura 2.

Figura 2: Divisão da cidade Z



Modelo A:

$$y_{i,\,j,\,t} = \varphi + \rho_1 y_{i,\,j,\,t-1} + Y_{1,\,t-1}\beta + \varepsilon_{i,\,j,t} \qquad \textbf{(13)}$$

Modelo B(modelo econométrico espacial clássico):

$$y_{i,\,t} = \beta_0 + \beta_1 y_{i,t-1} + \beta_2 W_{i,t-1} y_{i,t-1} + \varepsilon_{i,\,t} \qquad \textbf{(14)}$$

O modelo A corresponde a uma simplificação do modelo 9 descrito na seção 2.1. No modelo B, $y_{i,t}$ trata-se de um vetor coluna no qual cada linha contém o número de incidentes criminais ocorridos em uma determinada unidade de área no período t, e $y_{i,t-1}$ trata-se do mesmo vetor coluna defasado em um dia. Já $W_{i,\,t-1}$, trata-se de uma matriz quadrada de pesos na qual o número linhas e colunas é dado pelo número unidades de área que serão analisadas; dada a disposição espacialmente das nove unidades de área que compõem o território da cidade Z, a matriz de pesos, $W_{i,\,t-1}$, deverá ser representada tal como:

Figura 3: Matriz de peso espacial

	UA1	UA2	UA3	UA4	UA5	UA6	UA7	UA8	UA9
UA1	0	1	0	1	1	0	0	0	0
UA2	1	0	1	1	1	1	0	0	0
UA3	0	1	0	0	1	1	0	0	0
UA4	1	1	0	0	1	0	1	1	0
UA5	1	1	1	1	0	1	1	1	1
UA6	0	1	1	0	1	0	0	1	1
UA7	0	0	0	1	1	0	0	1	0
UA8	0	0	0	1	1	1	1	0	1
UA9	0	0	0	0	1	1	0	1	0

Na matriz de pesos acima, a linha 1 recebe valor 1 para a coluna 2 o que reflete o fato de UA2(unidade de área 2) ser um vizinha de UA1(unidade de área 1); de modo similar, todas as unidades de área observadas nas colunas que forem vizinhas daquelas observadas nas linhas da matriz de pesos recebem valor 1, e caso contrário recebem 0.

A primeira desvantagem do modelo B em relação modelo A, diz respeito ao processo de estimação dos parâmetros. No modelo B, a estimação do efeito dos vizinhos de primeira ordem sobre a unidade de área 5, hachurada na Figura 2, requer a resolução de um problema de otimização matemática que envolve matrizes, e que por isso exigirá maior poder computacional e tempo de execução quando comparado com a estimação deste mesmo efeito no modelo A.

A outra desvantagem da metodologia exposta pelo modelo B diz respeito ao significado do parâmetro estimado. Dado a forma matricial do modelo B, o parâmetro β_2 estimado para este modelo trata-se de um efeito médio de vizinhos de primeira ordem sobre uma dada unidade de área, em outras palavras, este efeito nos diz que independente da unidade de área que estejamos tratando o efeito proveniente dos vizinhos fronteiriços é o mesmo. No entanto, ao comparar duas unidades de área disjuntas, supondo a primeira cercada por montanhas e a segunda por regiões urbanizadas, é plausível admitir que o efeito proveniente dos vizinhos de primeira ordem é diferente para cada uma destas unidades de área; dessa forma, o modelo A parece mais adequado para recuperar o efeito heterogêneo proveniente dos vizinhos de primeira ordem, visto que tal modelo deverá ser estimado para cada uma das unidades de área em questão; como consequência, estamos assumindo também que cada unidade de área terá o seu melhor modelo para fins de previsão.

Um aspecto importante dos modelos descritos até o momento é que as previsões de incidentes criminais para um período do dia à frente são condicionadas aos crimes ocorridos na data corrente, o que requer a atualização da base de incidentes criminais em tempo real. Entretanto, uma vez que tais incidentes são registrados e consolidados á nível de batalhão para posteriormente serem submetidos ao ISP, os dados mais recentes que teremos a disposição são defasados em alguns meses; este fato justificaria abandonar os modelos preditivos condicionais até então analisados e dedicar atenção à modelos incondicionais, nos quais a previsão para um período a frente é feita com base na média do número de incidentes criminais ocorridos no mesmo período para o qual se pretende prever de anos anteriores. Por outro lado, ao utilizar um modelo condicional para prever muitos passos à frente, obtém-se previsões que tendem para aquelas que seriam obtidas por um modelo incondicional; dessa forma, ao optar pelo modelagem descrita na seção 2.1, estaremos adotando um modelo geral o suficiente que se adequará bem tanto ao cenário considerado ideal, em que os dados que pretendemos prever são atualizados com alta frequência, quanto ao cenário em que estes mesmos dados são atualizados com baixa frequência.

2.3 A estimação dos parâmetros dos modelos

Os parâmetros φ, ρ_1, ..., ρ_4, α, β, γ, τ, e ε dos modelos desenvolvidos serão estimados através do método de mínimos quadrados ordinários. O método de MQO(mínimos quadrados ordinários) ou OLS(*Ordinary Least Squares*) é uma técnica de otimização matemática que procura encontrar o melhor parâmetro que minimiza a soma dos quadrados dos resíduos de um dado modelo de regressão, sendo estes resíduos a diferenças entre o valores observados e os valores estimados da variável dependente.

Tomando como referência o modelo 1, a estimação do parâmetro ρ através do método de mínimos quadrados ordinário pode ser expresso como:

$$\min_{\rho_1} \sum_{i=1}^{n} (y_{i,j,t} - \varphi - \rho_1 y_{i,j,t-1})^2 \qquad \textbf{(15)}$$

De um modo mais simplificado temos,

$$\min_{\rho_1} \sum_{i=1}^{n} (\varepsilon_{i,\,j,\,t})^2 \qquad \textbf{(16)}$$

Os parâmetros do modelo já estimados, ou seja, conhecidos, são representados com acento circunflexo.

Considerando que os modelos descritos na seção 2.1 tenham sido adequadamente especificados, espera-se que as seguintes hipóteses sejam atendidas para que a estimação dos parâmetros via MQO seja válida: 1) dada a estimação dos parâmetro, o resíduo, $\hat{\varepsilon}_{i,\,j,\,t}$, deve ser normalmente distribuído, ou seja, deve ter média zero e variância constante; esta hipótese também implica que os valores do resíduo sejam independentes e identicamente distribuídos. 2) O resíduo não deve ser correlacionado com as variáveis explicativas do modelo. 3) O modelo deve ser linear nos parâmetros, ou seja, deve existir uma relação linear entre a variável dependente e as variáveis explicativas do modelo. 4) Nenhuma das variáveis explicativas podem ser uma combinação linear das demais, ou seja, não pode existir colinearidade entre as mesmas.

2.4 Avaliação dos modelos preditivos

Para analisar quão boas são as previsões realizadas pelos modelos desenvolvidos faremos uma análise *out of the sample*, que consiste em duas etapas. Na primeira, para cada unidade de área analisada, os parâmetros dos modelos serão estimados via mínimos quadrados ordinários utilizando os primeiros 3/4 de dados disponíveis, que corresponde as três mil primeiras observações de cada setor.

Realizada a primeira etapa da análise *out of the sample*, as previsões de incidentes criminais para um dia a frente serão dadas por:

Previsão dada estimação do modelo 1

$$\hat{y}_{i,\,j,\,t+1} = \hat{\varphi} + \hat{\rho}_1 y_{i,\,j,\,t} \qquad \textbf{(17)}$$

Previsão dada estimação do modelo 2

$$\hat{y}_{i,\,j,\,t+1} = \hat{\varphi} + \hat{\rho}_1 y_{i,\,j,\,t} + \hat{\rho}_2 y_{i,\,j,\,t-1} + \hat{\rho}_3 y_{i,\,j,\,t-2} + \hat{\rho}_4 y_{i,\,j,\,t-3} \qquad \textbf{(18)}$$

Previsão dada estimação do modelo 3

$$\hat{y}_{i,\,j,\,t+1} = \hat{\varphi} + \hat{\rho}_1 y_{i,\,j,\,t} + \hat{\alpha}\frac{1}{4}\sum_{s=0}^{3} y_{i,\,j,\,t-s} \qquad \textbf{(19)}$$

Previsão dada estimação do modelo 4

$$\hat{y}_{i,\,j,\,t+1} = \hat{\varphi} + \hat{\rho}_1 y_{i,\,j,\,t} + \hat{\rho}_2 y_{i,\,j,\,t-1} + \hat{\rho}_3 y_{i,\,j,\,t-2} + \hat{\alpha}\frac{1}{4}\sum_{s=0}^{3} y_{i,\,j,\,t-s} \qquad \textbf{(20)}$$

Previsão dada estimação do modelo 5

$$\hat{y}_{i,\,j,\,t+1} = \hat{\rho}_1 y_{i,\,j,\,t} + D_{t+1}\hat{\gamma} + H_{t+1}\hat{\tau} \qquad \textbf{(21)}$$

Previsão dada estimação do modelo 6

$$\hat{y}_{i,\,j,\,t+1} = \hat{\rho}_1 y_{i,\,j,\,t} + \hat{\rho}_2 y_{i,\,j,\,t-1} + \hat{\rho}_3 y_{i,\,j,\,t-2} + \hat{\rho}_4 y_{i,\,j,\,t-3} + D_{t+1}\hat{\gamma} + H_{t+1}\hat{\tau} \qquad \textbf{(22)}$$

Previsão dada estimação do modelo 7

$$\hat{y}_{i,\,j,t+1} = \hat{\rho}_1 y_{i,\,j,\,t} + \hat{\alpha}\frac{1}{4}\sum_{s=0}^{3} y_{i,\,j,\,t-s} + D_{t+1}\hat{\gamma} + H_{t+1}\hat{\tau} \qquad \textbf{(23)}$$

Previsão dada estimação do modelo 8

$$\hat{y}_{i,\,j,t+1} = \hat{\rho}_1 y_{i,\,j,\,t} + \hat{\rho}_2 y_{i,\,j,\,t-1} + \hat{\rho}_3 y_{i,\,j,\,t-2} + \hat{\alpha}\frac{1}{4}\sum_{s=0}^{3} y_{i,\,j,\,t-s} + D_{t+1}\hat{\gamma} + H_{t+1}\hat{\tau} \qquad \textbf{(24)}$$

Previsão dada estimação do modelo 9

$$\hat{y}_{i,\,j,t+1} = \hat{\rho}_1 y_{i,\,j,\,t} + \hat{\alpha}\frac{1}{4}\sum_{s=0}^{3} y_{i,\,j,\,t-s} + D_{t+1}\hat{\gamma} + H_{t+1}\hat{\tau} + Y_{1,t}\hat{\beta} \qquad \textbf{(25)}$$

Previsão dada estimação do modelo 10

$$\hat{y}_{i,\,j,\,t+1} = \hat{\rho}_1 y_{i,\,j,\,t} + \hat{\rho}_2 y_{i,\,j,\,t-1} + \hat{\rho}_3 y_{i,\,j,\,t-2} + \hat{\alpha}\frac{1}{4}\sum_{s=0}^{3} y_{i,\,j,\,t-s} + D_{t+1}\hat{\gamma} + H_{t+1}\hat{\tau} + Y_{1,t}\hat{\beta} \qquad \textbf{(26)}$$

Previsão dada estimação do modelo 11

$$\hat{y}_{i,\,j,\,t+1} = y_{i,\,j,\,t} \qquad \textbf{(27)}$$

Previsão dada estimação do modelo 12

$$\hat{y}_{i,\,j,\,t+1} = y_{i,\,j,\,t-3} \qquad \textbf{(28)}$$

Na segundo etapa, de posse dos modelos já estimados, serão realizadas previsões um passo à frente para o mesmo horizonte de tempo dos demais 1/4 dos dados disponíveis que restaram, respeitando o seguinte procedimento: faz-se a primeira previsão um passo à frente, esta previsão será inclusa na base de dados, a partir desta nova base o modelo será reestimado e uma nova previsão um passo à frente será gerada; este processo será realizado mil vezes, o que corresponde a fazer mil previsões um passo à frente.

Realizada a segunda etapa da análise *out of the sample*, torna-se possível avaliar qual modelo gera as previsões um passo à frente que melhor se aproximam do número

observado de incidentes criminais, ou seja, que gera o menor erro de previsão. Esta avaliação será feita com base nas funções objetivo MAE(*Mean Absolute error*) e MSE(*Mean Squared Error*) que pretendemos minimizar, expressas por:

$$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n} |\hat{y}_{i,j,t} - y_{i,j,t}| \qquad \textbf{(29)}$$

$$\text{MSE} = \frac{1}{n}\sum_{t=1}^{n} (\hat{y}_{i,j,t} - y_{i,j,t})^2 \qquad \textbf{(30)}$$

Onde $\hat{y}_{i,j,t}$ trata-se de um vetor coluna que contém as previsões de incidentes criminais para uma determinada unidade de área, e $y_{i,j,t}$ é o vetor coluna que contém os valores de fato observados para esta variável; é importante observar que ambos vetores contém informações para o mesmo horizonte de tempo.

Assim, para cada unidade de área, o modelo que apresentar o menor MAE ou MSE será considerado o melhor modelo para fins de previsão de eventos criminais um passo à frente. A respeito deste ponto deve-se fazer a ressalva de que ao tomar como referência as funções objetivo descritas acima, o melhor modelo para fins de previsão não necessariamente será o melhor modelo para fins de representar o processo gerador de dados de incidentes criminais.

Os processos de estimação e avaliação dos modelos preditivos, tal como descrito nas seções 2.3 e 2.4, exigiram grande capacidade computacional, por isso contamos com a participação de Felipe Luiz, cujo *know-how* em computação nos propiciou as condições necessárias para que tivéssemos eficiência na execução de tais processos.

2.5 Implementação e resultados dos modelos

Tal como o tratamento dos dados descrito no primeiro capítulo, os modelos descritos na seção 2.1 foram implementados em *Python*, através de um plataforma de desenvolvimento chamada *Zeppelin* que congrega banco de dados com linguagens de programação. Esta plataforma foi estruturada por Felipe Luiz.

Para facilitar a análise, os modelos foram implementados para um conjunto reduzido de setores(unidades de área). Dado a grande incidência de crimes, decidimos centralizar a análise na Central do Brasil, que está localizado na latitude = -22,903751 e longitude = -43,191250, e que na divisão da cidade em quadrados de 250m de lado corresponde ao setor i = 38 e j =243. A partir deste setor foram contados 25 setores para a direita e para esquerda, e 25 setores para cima e para baixo; dessa forma, os modelos foram implementados para um conjunto de 2601 setores.

É importante observar que os modelos nos quais variáveis *dummy* foram inclusas, o termo de intercepto, representado pelo parâmetro φ, foi suprimido para evitar colinearidade entre as variáveis explicativas; pelo mesmo motivo, a variável *dummy* que indica uma segunda-feira também foi suprimida do conjunto de *dummies* de dia da semana. O problema de colinearidade surge quando alguma variável explicativa pode ser escrita como combinação de linear de outras variáveis explicativas; no caso dos modelos descritos na seção 2.1, poderíamos escrever o intercepto como a soma das variáveis *dummy* de período de dia; de modo similar, poderíamos escrever a *dummy* que indica uma segunda-feira como a soma das variáveis *dummy* de período de dia, menos, a soma das demais variáveis *dummy* de dia da semana.

Embora estejamos analisando o entorno da Central do Brasil, no qual há grande incidência de crimes, alguns dos setores analisados possuem todos os dados iguais a zero. Os modelos 9 e 10, em contra partida, incluem os vizinhos de primeira ordem como regressores, dessa forma, para evitar que setores com todos os dados iguais a zero sejam utilizados como regressores(variáveis explicativas), apenas os setores com algum dado diferente de zero poderão ser considerados vizinhos de primeira ordem; caso contrário, ao utilizar um vetor coluna, apenas com zeros, como variável explicativa incorreremos em colinearidade visto que este vetor poderá ser escrito como combinação linear de quaisquer outras variáveis explicativas.

Para cada um dos modelos implementados foram geradas duas planilhas para fins de avaliar performance preditiva, uma contendo o MAE e a outra contendo o MSE de

cada um dos setores em análise. Os mapas de calor a seguir resumem a performance de cada um dos modelo descritos na seção 2.1.

Ao analisar a desempenho dos doze modelos, temos que, de acordo com as Figuras 4 e 5, as melhores previsões segundo a função objetivo MAE são geradas pelo modelo asLag1, cujo as previsões um passo à frente tratam-se do número de incidentes criminais ocorridos um períodos de dia atrás.

De acordo com as Figuras 6 e 7, é possível observar que, segundo a função objetivo MSE, a performance dos modelos autoregressivos melhora, no entanto, isto já era de se esperar dado que o método de estimação dos parâmetros dos modelos minimiza erros quadráticos.

Figura 4: Performance dos modelos via MAE

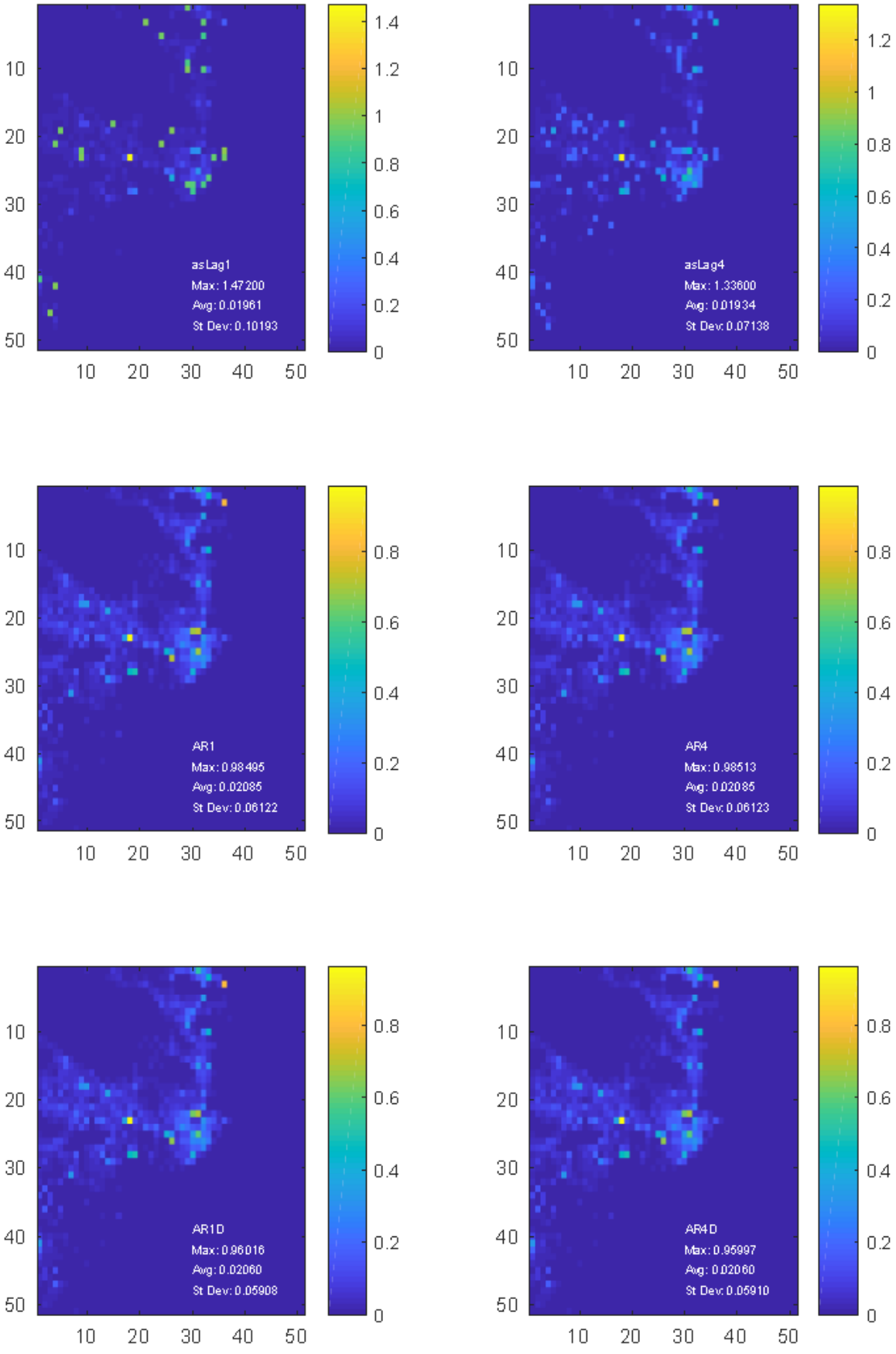


Figura 5: Performance dos modelos via MAE(Continuação da Figura 4)

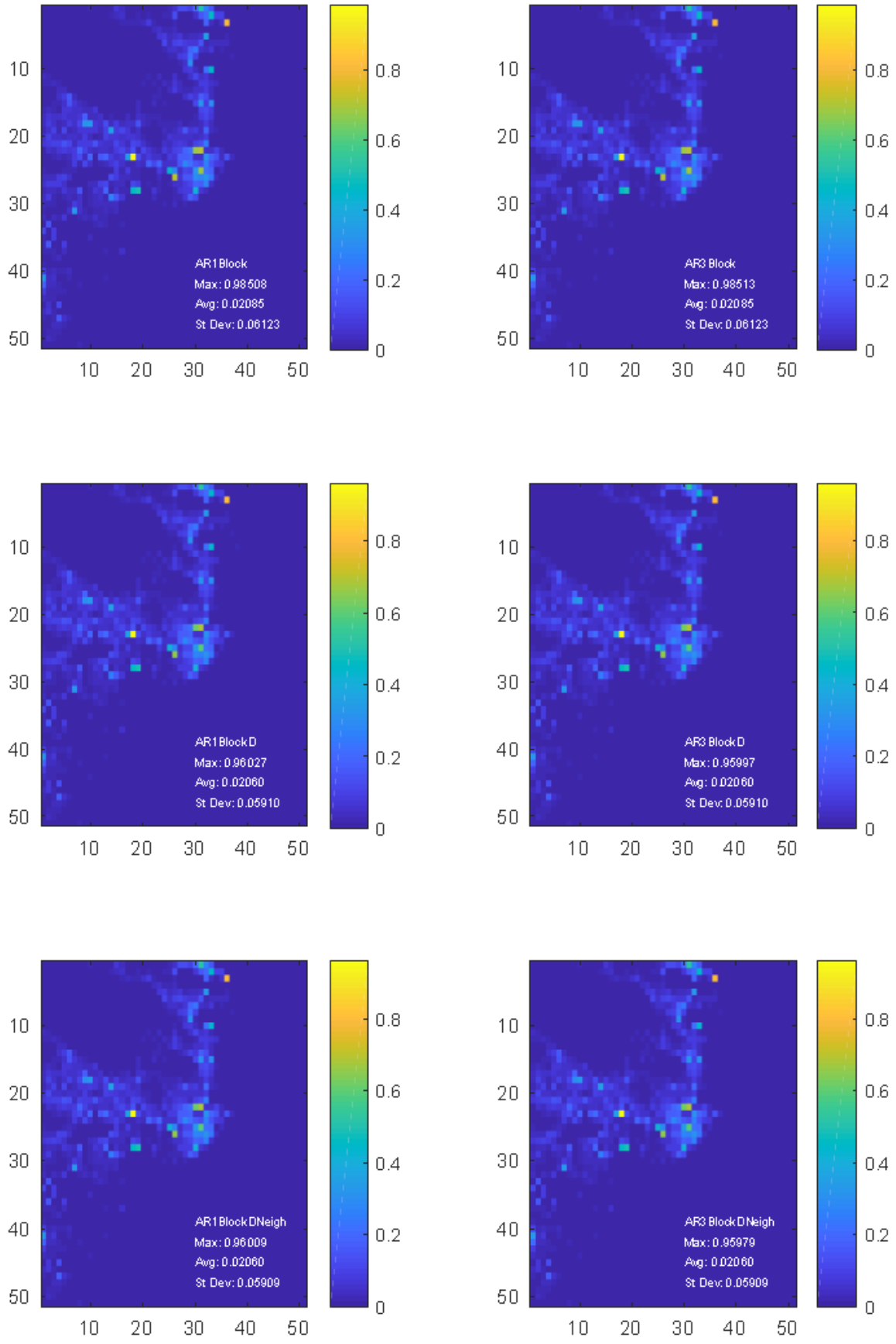


Figura 6: Performance dos modelos via MSE

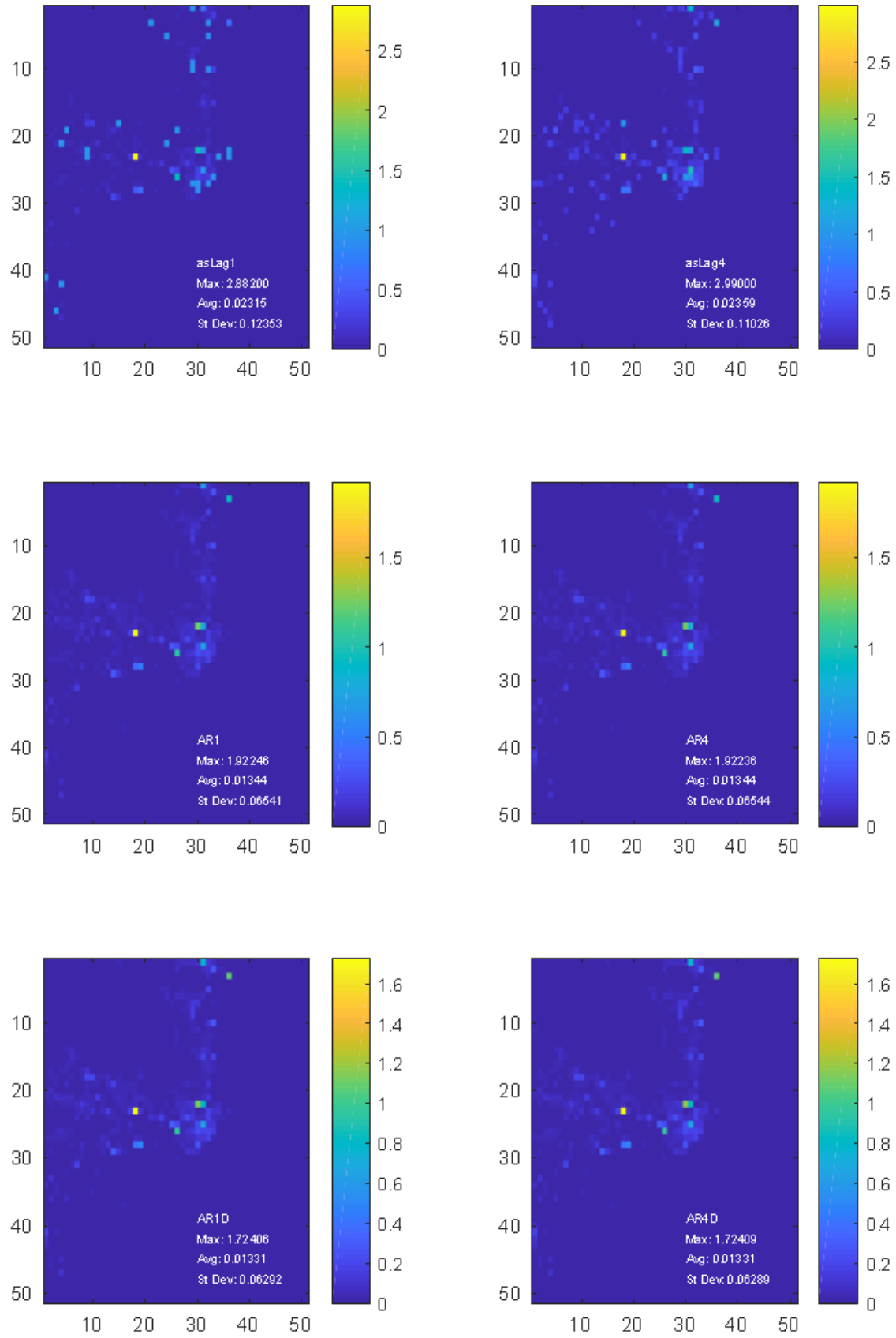
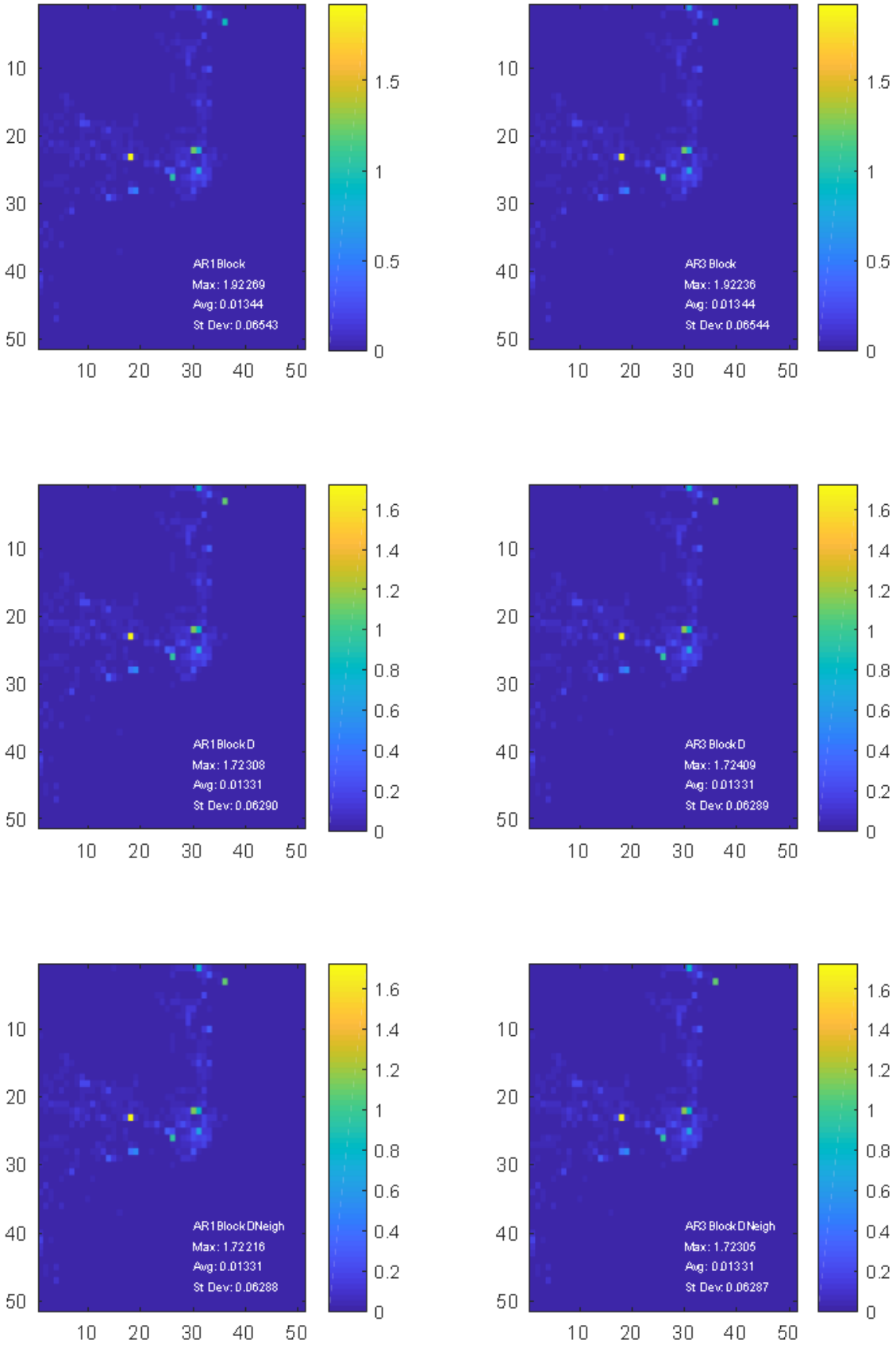


Figura 7: Performance dos modelos via MSE(Continuação da Figura 6)



Complementar aos mapas de calor apresentados anteriormente a Tabela 4, apresenta, tanto para a função objetivo MAE quanto MSE, a frequência com que cada um dos modelos são selecionados como o melhor para fins de previsão.

Tabela 4: Frequência de seleção dos modelos

Modelo	MAE	MSE
AR1	17	244
AR4	1	169
AR1Block	0	65
AR3Block	1	4
AR1D	16	112
AR4D	0	49
AR1BlockD	2	48
AR3BlockD	0	9
AR1BlockDNeigh	2	70
AR3BlockDNeigh	2	86
asLag1	2559	1745
asLag4	1	0

Tal como ilustrado nas Figuras 4 e 5, ao utilizar a função objetivo MAE para selecionar o melhor modelo preditivo de cada setor em análise, perceber-se-á o predomínio do modelo asLag1 como o melhor para fins de previsão um passo a frente.

Outra informação interessante a ser extraída desta tabela é que, mesmo optando por selecionar o melhor modelo para cada setor via MSE, observar-se-á o predomínio do modelo asLag1.

É importante observar que para os setores que possuem todos as quatro mil observações iguais zero, o modelo asLag1 será considerado como melhor modelo para fins de previsão, e consequentemente o MAE e o MSE para este modelo serão iguais a zero.

Os mapas de calor a seguir agregam, para cada setor, os diferentes modelos implementados por MAE mínimo e por MSE mínimo.

Figura 8: Menor MAE para cada setor

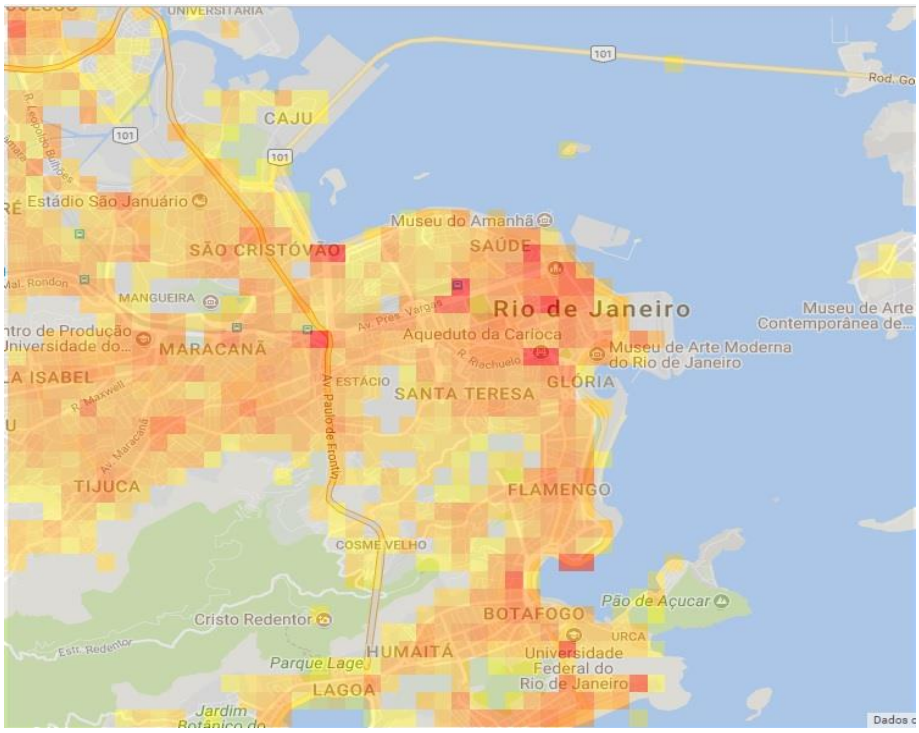
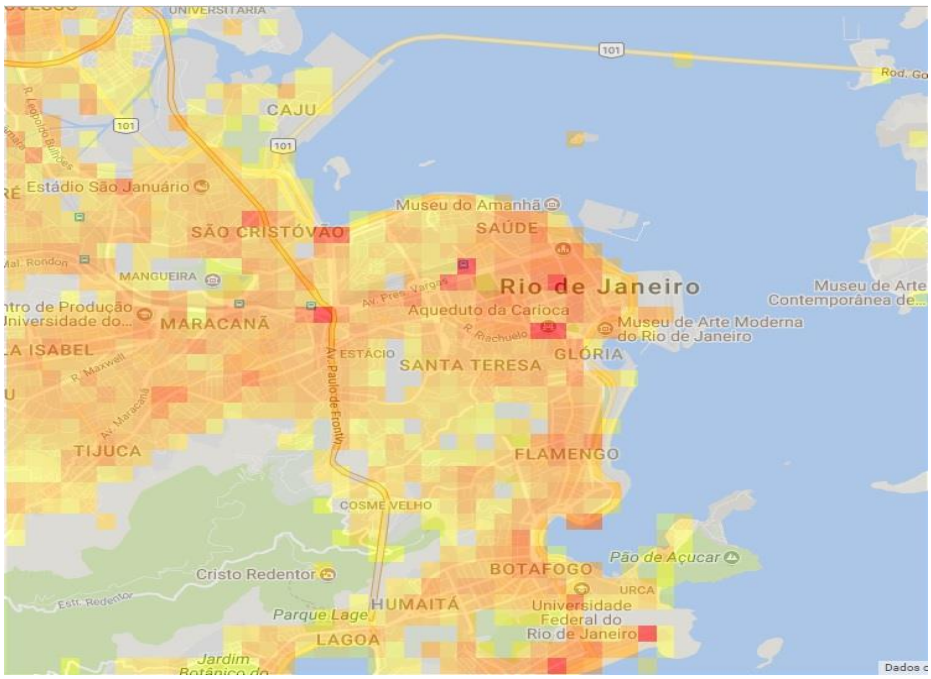
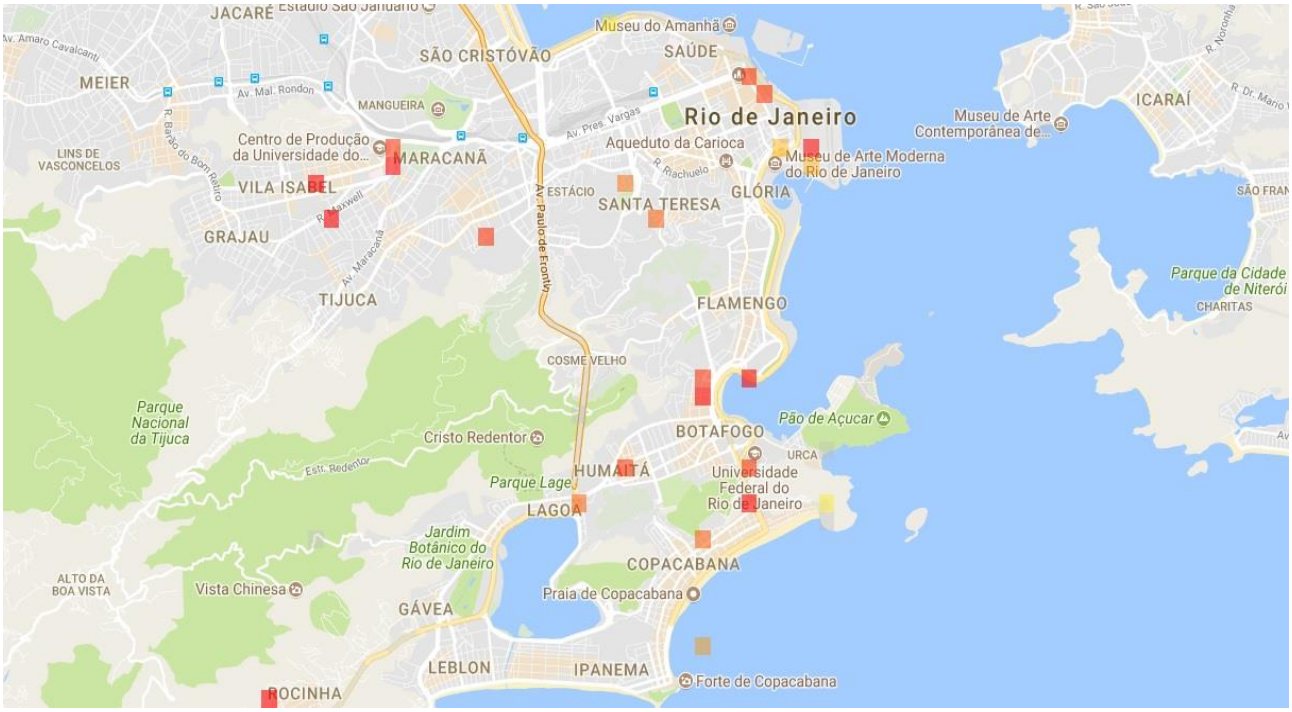


Figura 9: Menor MSE para cada setor



Considerando o compromisso inicial de realizar previsões tendo em vista os dados disponíveis, e utilizando a função objetivo MAE para selecionar o melhor modelo preditivo para cada setor, as previsões um passo para criminalidade no Rio de Janeiro estão representadas no mapa a seguir.

Figura 10: Previsão para eventos criminais um passo à frente



O dado mais recente observado para os setores em análise diz respeito a noite do dia 05/04/2016, sendo assim, as previsões representadas no gráfico acima referem-se a madrugada do dia 06/04/2016.

3. Conclusão

Embora valoroso e enriquecedor, o exercício de gerar previsões um passo à frente apresenta-se como tarefa árdua e repleta de percalços.

Neste trabalho encontra-se registrado a tentativa de gerar previsões de criminalidade para um período de dia à frente em unidades refinadas de espaço. Após definir a metodologia a ser utilizada, o processo de estimação e avaliação dos modelos preditivos se encerraram em um desafio computacional tendo em vista o volume de dados e as unidade de área para as quais estes são observados.

Em linhas gerais, com base na análise comparativa dos modelos descritos ao longo deste trabalho, é possível observar que, uma vez utilizadas como variáveis explicativas, defasagens espaciais não geram melhoras significativas no desempenho preditivo de um modelo.

Esta conclusão, no entanto, não derruba hipótese de que defasagens espaciais sejam importantes para modelar criminalidade; é importante salientar que os modelos aqui descritos foram implementados tendo em visto os dados disponíveis, sendo assim, analisar e caracterizar previamente os dados pode servir como uma guia para adoção de uma metodologia mais adequada.

No todo, dado que esta monografia diz respeito a etapa inicial de um projeto, a modelagem até o momento desenvolvida será atualizada e aperfeiçoada na expectativa de que alcancemos modelos mais acurados para fins de previsão.

Um potencial plano de ação para as próximas etapas deste projeto consistirá em:

1) Validar os dados, que corresponde a checar a qualidade e a veracidade dos dados disponíveis. Este processo consistirá em cruzar a base de dados oriunda do ISP com outras bases de dados, como por exemplo, a de letalidade do SUS.

2) Desenvolver métodos para detecção de *outliers*, observação aberrantes. Assim, ao invés de analisarmos modelos para criminalidade média, tal como aqueles descritos na seção 2.1, poderemos, de nodo alternativo, analisar modelos que capturam apenas os picos de incidentes criminais.

3) Explorar modelagem alternativa àquela utilizada até o momento. Visto que o número de incidentes por período do dia para cada setor são baixos, os dados que temos disposição são típicos de contagem, e por isso, a modelagem adequada não é necessariamente linear; modelos de contagem, por sua vez, não possuem solução

fechada e irão requerer a maximização numérica de uma função(por período *out of the sample*, e por setor), o que aumentará consideravelmente o custo computacional.

4) Estimar os parâmetros dos modelos via LASSO (*Least Absolute Shrinkage and Selection Operator*). Ao longo deste trabalho, os modelos foram estimados via OLS, o que implica que uma vez definida as variáveis explicativas estas serão inclusas no modelo não importando o quão relevantes são para modelar a variável dependente; a estimação via LASSO, por outro lado, seleciona, dentre as variáveis explicativas, aquelas que de fato são relevantes para modelar a variável dependente. Este método alternativo de estimação possibilitará entender quais são as variáveis relevantes, dentre aquelas utilizadas até o momento, para fins de modelagem, e também possibilitará incluir vizinhos de ordem maior, bem como outras variáveis externas nos modelos.

Bibliografia

- DIEBOLD, Francis. **Econometrics**. University of Pennsylvania, 2016. Disponível em: http://www.ssc.upenn.edu/~fdiebold/Teaching104/Econometrics.pdf. Acessado em 25/04/2017.
- LESAGE, James, and Robert Kelley PACE. **Introduction to Spatial Econometrics**. CRC Press, 2009.

Apêndice

1. Apêndice referente ao capítulo ‘1. Analise preliminar e tratamento de dados’

1.1 Versão final do algoritmo entregue ao ISP no dia 09/12/2016

Tal como na seção 1.3.2, para fins de descrição, iremos separar o algoritmo de tratamento em quatro etapas. A descrição detalhada, bem como a relação dos objetos, das listas citadas a seguir encontram-se na Tabela 4 ao final desta seção.

A primeira etapa do algoritmo, que se refere a um pré-tratamento, torna todos os caracteres dos *strings* minúsculos em maiúsculos e remove todos os acentos gráficos das palavras do campo “Referência”; são feitas expansões de abreviações, nas quais *strings* como “AV” ou “AV.” são substituídos por “AVENIDA”, por exemplo; são realizadas decomposições de contrações, que se referem à substituição de “DO” por “DE O”, por exemplo; e, enfim, também é realizado o treinamento da biblioteca NLTK com base no conjunto de artigos *mac_morpho*.

Tomando, como exemplo, o endereço da segunda linha da Tabela 2, o campo “Referência”, após realizada a primeira etapa do algoritmo, será representado como no exemplo abaixo:

Exemplo 1:

PROXIMO AO NUMERO 560

Na segunda etapa do algoritmo de tratamento, dado o treinamento prévio, a biblioteca NLTK classifica gramaticalmente os *strings* da coluna “Referência”. A partir desta classificação, os dados do campo “Referência” serão analisados sob a forma de etiquetas, as quais, cada uma destas, possui uma palavra que compõe uma dada referência e sua classificação gramatical. Tomando como referência o Exemplo 1, esperamos que, após a segunda etapa do algoritmo, obtenhamos o resultado abaixo.

Exemplo 2:

[(‘PROXIMO’, ‘ADJ’), (‘AO’, ‘PREP’), (‘NUMERO’, ‘N’), (‘560’, ‘N|AP’)]

Os quais, “ADJ”, significa um adjetivo, “PREP”, indica uma preposição, “N”, um substantivo, e “N|AP”, um numeral.

Revisitando nosso objetivo, o algoritmo aqui descrito pretende extrair do campo “Referência” informações que sejam relevantes para georreferenciar endereços mal especificados. Tendo em vista este objetivo, e dada a classificação gramatical das palavras que compõem os *strings* do campo “Referência”, a terceira etapa do algoritmo de tratamento contempla cinco regras, ou conjunto de regras, para capturar estas informações de interesse.

O primeiro conjunto de regras tem como objetivo reclassificar as palavras que indicam posicionamento relativo. Neste processo, as palavras de um dado *string* que pertencem a lista *qualifierList* (cuja descrição, com detalhes, encontra-se na Tabela 4, ao final desta seção) terão sua classificação gramatical substituídas pela letra ‘Q’, que indica que tais palavras são qualificadores. Tomando como referência o Exemplo 2, após a aplicação desta regra, obteremos o resultado abaixo.

Exemplo 3:

[(‘PROXIMO’, ‘Q’), (‘AO’, ‘PREP’), (‘NUMERO’, ‘N’), (‘560’, ‘N|AP’)]

O segundo conjunto de regras tem como objetivo separar os termos que compõem o campo “Referência” e, por isso, são denominadas regras de quebra. Uma quebra no *string* “Referência” deve ocorrer caso:

1) Uma preposição não seja precedida por uma palavra pertencente à lista *nonBreakingWords*, ou, não seja precedida por um substantivo próprio.

2) Um artigo não seja parte de uma contração;

3) Um determinado objeto do *string* faça parte da lista *separators*.

4) Uma das palavras que compõem o *string* seja uma conjunção.

Tomando como referência o Exemplo 3, esperamos que, após a aplicação das regras de quebra, obtenhamos o resultado abaixo.

Exemplo 4:

[(‘PROXIMO’, ‘Q’)]

[(‘AO’, ‘PREP’), (‘NUMERO’, ‘N’), (‘560’, ‘N|AP’)]

Neste exemplo, a quebra se dá pelo fato da preposição “AO” não ser precedida por uma palavra que pertence à lista *nonBreakingWords*, ou também, pelo fato de não ser precedida por um substantivo próprio.

Já o terceiro conjunto de regras tem como objetivo filtrar os resultados obtidos após a aplicação das regras de quebra. No processo de filtragem, se uma dada parte quebrada:

1) For composta por um única palavra e a mesma é classificada como um qualificador, então esta será preservada;
2) Possui pelo menos uma das palavras classificada como numeral, então esta será preservada;
3) For composta por um única palavra e a mesma pertence a lista *forbiddenSingletonWords*, então esta será descartada;
4) For composta por uma única palavra, não sendo esta classificada como nome próprio, mas pertencente a lista *bigList*, então esta será descartada.
5) Possui pelo menos uma das palavras que pertencente a lista *forbiddenWords*, então esta será descartada;
6) Não possui nenhum substantivo, comum ou próprio, então esta será descarta;
7) Possui um único substantivo comum e o mesmo pertence a lista *placeTypeSet*, então esta será descarta.

Tomando como referência o Exemplo 4, esperamos que, após a aplicação das regras de filtragem, obtenhamos o resultado abaixo.

Exemplo 5:

[(‘PROXIMO’, ‘Q’)]

[(‘AO’, ‘PREP’), (‘NUMERO’, ‘N’), (‘560’, ‘N|AP’)]

No exemplo acima, dado a execução das regras de filtragem, temos que a primeira parte quebrada foi preservada uma vez que esta é composta por uma única palavra e a mesma é classificada como qualificador; já a segunda parte quebrada foi preservada uma vez que pelo menos um das palavras que a compõem é classificada como numeral.

No quarto conjunto de regras é realizada a limpeza de cada uma das partes quebradas que sobreviveram à execução das regras de filtragem. Assim as regras de limpeza são responsáveis por:

1) Eliminar preposições soltas no início de uma dada parte quebrada;

2) Eliminar artigos soltos no início de uma dada parte quebrada.

Tomando como referência o Exemplo 5, esperamos que, após a aplicação das regras de limpeza, obtenhamos o resultado abaixo.

Exemplo 6:

[('PROXIMO', 'Q')]
[('NUMERO', 'N'), ('560', 'N|AP')]

No exemplo acima, dado a execução das regras de limpeza, temos que, para a primeira parte quebrada nenhum termo foi eliminado visto que o mesmo não começava com artigo ou preposição; já na segunda parte quebrada, temos que a preposição 'AO' foi eliminada.

Enfim, o quinto conjunto de regras tem como finalidade pós-processar as informações que sobreviveram à aplicação das regras descritas anteriormente. Neste processo, para cada parte quebrada, as regras de pós-processamento serão responsáveis por:

1) Remover palavras referentes a lugar consideradas não importantes, relacionadas na lista *wordsToTrimList*, no início e no fim de cada uma das partes quebradas.
2) Juntar as contrações que foram decompostas na primeira etapa do algoritmo.

Tomando como referência o Exemplo 6, esperamos que, após a aplicação das regras de pós-processamento, obtenhamos o resultado abaixo.

Exemplo 7:

[('PROXIMO', 'Q')]
[('560', 'N|AP')]

No exemplo acima, dada a aplicação do quinto conjunto de regras, é possível observar que, para a segunda parte quebrada, a palavra "NUMERO" foi eliminada visto que a mesma pertence a lista *wordsToTrimList*.

Por fim, na quarta e última etapa do algoritmo, para as informações que sobreviveram a terceira etapa, tal como no Exemplo 7, as etiquetas classificadoras são desfeitas e é realizada a entrega destas que são as informações consideradas relevantes para fins de georreferenciamento.

Tal como solicitado pelo próprio ISP, o *output* final do algoritmo é entregue no formato de um conjunto com três informações. A primeira entrada do conjunto indicará um posicionamento relativo, que diz respeito as palavras classificadas como qualificador, caso estas existam; a segunda entrada será um bloco de texto com a informação de interesse, aquela que sobrou após a execução do algoritmo de tratamento; já a terceira entrada do conjunto tratar-se-á da categorização da informação que se encontra na segunda entrada do conjunto, com base nos critérios: Comunidade, Rodovia, Logradouro, Número, Bairro, Munícipio, ou UF. Tomando como referência o Exemplo 7 e dada a execução das etapas anteriores do algoritmo, o *output* final será entregue tal como a última linha exemplo abaixo.

Exemplo 8:

PROXIMO AO NUMERO 560

[(‘PROXIMO’, ‘ADJ’), (‘AO’, ‘PREP’), (‘NUMERO’, ‘N’), (‘560’, ‘N|AP’)]

> [(‘PROXIMO’, ‘Q’)]
> [(‘NUMERO’, ‘N’), (‘560’, ‘N|AP’)]

>>> (‘PROXIMO’, ‘560’, ‘NUMERO’)

Na última linha do Exemplo 8, é possível observar que a palavra “PROXIMO” indica um posicionamento relativo e por isso encontra-se na primeira entrada do conjunto; já o numeral “560” diz respeito a informação relevante para fins de georreferenciamento; e “NUMERO” trata-se da categoria a qual esta informação se enquadra. Também é possível notar que o resultado exposto no Exemplo 8 corresponde a um resumo dos resultados das etapas descritas anteriormente.

Tabela 5: Versão final da lista de palavras utilizadas no algoritmo de tratamento.

Nome da lista	Objetos da lista	Descrição da lista
placeTypes	SEM TIPO, RUA, FAVELA, SEM, RODOVIA, AVENIDA, OUTROS, LADEIRA, ESTRADA, BECO, PRAIA, MORRO, PRACA, VIA, RAMAL, LARGO, VIELA, CAMPO, TRAVESSA, LOTEAMENTO, RETORNO, SERVIDAO, ACESSO, ALAMEDA, PARQUE, ARCO, ENTRADA, VILA, FAZENDA, ESCADARIA, CONDOMINIO, TERMINAL, TUNEL, CAMINHO, TREVO, PATIO, CONJUNTO, PONTE, ESCADA, TERRENO BALDIO, SITIO, SUBIDA, VALE, REPRESA, VIADUTO, PASSARELA, COMUNIDADE, COMPLEXO, CIDADE, ZONA, ESQUINA, LOGRADOURO, TRAVESSIA, MATA, ESTACAO, ESCOLA, CASA, POSTO, DISTRITO, BARRACAO, NUMERO, NUMEROS, BAIRRO.	Lista de palavras possuem significado de lugar
nonBreakingWords	RUA, FAVELA, RODOVIA, AVENIDA, ESTRADA, PRAIA, MORRO, PRACA, VIA, LARGO, CAMPO, TRAVESSA, ALAMEDA, PARQUE, ILHA, VILA, FAZENDA, CONDOMINIO, TERMINAL, TUNEL, CAMINHO, TREVO, PATIO, CONJUNTO, PONTE, SITIO, SUBIDA, VALE, VIADUTO, PASSARELA, COMUNIDADE, COMPLEXO, CIDADE, ZONA, PONTO, BECO, LOJA, FABRICA, BAR, LADEIRA, PEDRA, ESTACAO, CHACARA, PADARIA, CASA, POSTO.	Lista de palavras que não geram quebra nos *strings* do campo "Referencia", ou seja, que são preservadas por serem consideradas relevantes.
separators	'.', ',', '-', '(', ')', '/', ':', ';', '\'', '\'\'', '"', '""', '`', '``'	Lista de sinais que, quando presentes em um *string*, geram quebra.

forbiddenWords	CONHECIDA, CONHECIDO, DENOMINADA, DENOMINADO, FATO, LOCALIZADA, LOCALIZADO, INFORMADA, INFORMADO, INFORMAR, INFORMOU, CITADO.	Lista de palavras que, quando acompanhadas ou solitárias após a aplicação das regras de quebra, geram a eliminação de uma dada parte quebrada.
forbiddenSingletonWords	Lista ***placeTypes*** somada aos seguintes objetos: FRENTE, AREIA, INTERIOR, NOME, PROXIMO, SN, S/N, LOCALIDADE, MURO, AC, ALTURA, LOCAL, FATO, PROX, BAIRRO, NOME, DIRECAO, PROXIMIDADE, PROXIMIDADES, ARREDORES, INICIO, FINAL, PONTO, OBS.	Lista de palavras que, quando solitárias após a aplicação das regras de quebra, geram a eliminação de uma dada parte quebrada.
bigList	-	Lista que contém 29858 palavras da língua portuguesa.
qualifierList	ESQUINA, PROXIMO, PERTO, FRENTE, ATRAS, FINAL, INTERIOR, LATERAL, ULTIMO, ULTIMA, PRIMEIRO, PRIMEIRA.	Lista de palavras que devem ser classificadas como qualificador.
wordsToTrimList	NUMERO, NUMEROS, BAIRRO.	Lista de palavras consideradas não importantes, e que por isso são eliminadas no pós-processamento.

1.2 Exemplos de *outputs* gerados pelo algoritmo de tratamento.

Exemplo 1

a. *Output* de uma versão do algoritmo anterior àquele descrito na seção 1.3.2:

O FATO OCORREU EM O INTERIOR DE A LINHA FERREA (ESTACAO DE OLINDA/NILOPOLIS)

[(‘O’, ‘ART’), (‘FATO’, ‘N’), (‘OCORREU’, ‘V’), (‘EM’, ‘PREP|+’), (‘O’, ‘ART’), (‘INTERIOR’, ‘N’), (‘DE’, ‘PREP|+’), (‘A’, ‘ART’), (‘LINHA’, ‘N’), (‘FERREA’, ‘ADJ’), (‘(‘, ‘(‘), (‘ESTACAO’, ‘N’), (‘DE’, ‘NPROP’), (‘OLINDA/NILOPOLIS’, ‘NPROP’), (‘)’, ‘)’)]

> [(‘FATO’, ‘N’), (‘OCORREU’, ‘V’)]
> [(‘INTERIOR’, ‘N’)]
> [(‘LINHA’, ‘N’), (‘FERREA’, ‘ADJ’)]
> [(‘ESTACAO’, ‘N’), (‘DE’, ‘NPROP’), (‘OLINDA/NILOPOLIS’, ‘NPROP’)]

>>> “FATO OCORREU”
>>> “LINHA FERREA”
>>> “ESTACAO DE OLINDA/NILOPOLIS”

b. *Output* da versão do algoritmo descrito na seção 1.3.2:

O FATO OCORREU EM O INTERIOR DE A LINHA FERREA (ESTACAO DE OLINDA/NILOPOLIS)

[(‘O’, ‘ART’), (‘FATO’, ‘N’), (‘OCORREU’, ‘V’), (‘EM’, ‘PREP|+’), (‘O’, ‘ART’), (‘INTERIOR’, ‘N’), (‘DE’, ‘PREP|+’), (‘A’, ‘ART’), (‘LINHA’, ‘N’), (‘FERREA’, ‘ADJ’), (‘(‘, ‘(‘), (‘ESTACAO’, ‘N’), (‘DE’, ‘NPROP’), (‘OLINDA/NILOPOLIS’, ‘NPROP’), (‘)’, ‘)’)]

> [(‘FATO’, ‘N’), (‘OCORREU’, ‘V’)]
> [(‘INTERIOR’, ‘N’)]
> [(‘LINHA’, ‘N’), (‘FERREA’, ‘ADJ’)]
> [(‘ESTACAO’, ‘N’), (‘DE’, ‘NPROP’), (‘OLINDA/NILOPOLIS’, ‘NPROP’)]

>>> “LINHA FERREA”
>>> “ESTACAO DE OLINDA/NILOPOLIS”

c. *Output* da versão do algoritmo descrito na seção 4.1.1:

O FATO OCORREU EM O INTERIOR DE A LINHA FERREA (ESTACAO DE OLINDA / NILOPOLIS)

[(‘O’, ‘ART’), (‘FATO’, ‘N’), (‘OCORREU’, ‘V’), (‘EM’, ‘PREP|+’), (‘O’, ‘ART’), (‘INTERIOR’, ‘N’), (‘DE’, ‘PREP|+’), (‘A’, ‘ART’), (‘LINHA’, ‘N’), (‘FERREA’, ‘ADJ’), (‘(‘, ‘(‘), (‘ESTACAO’, ‘N’), (‘DE’, ‘NPROP’), (‘OLINDA’, ‘NPROP’), (‘/’, ‘/’), (‘NILOPOLIS’, ‘NPROP’), (‘)’, ‘)’)]

> [(‘FATO’, ‘N’), (‘OCORREU’, ‘V’)]
> [(‘INTERIOR’, ‘Q’)]
> [(‘LINHA’, ‘N’), (‘FERREA’, ‘ADJ’)]
> [(‘ESTACAO’, ‘N’), (‘DE’, ‘NPROP’), (‘OLINDA’, ‘NPROP’)]
> [(‘NILOPOLIS’, ‘NPROP’)]

>>> (‘INTERIOR’, ‘LINHA FERREA’, None)
>>> (None, ‘ESTACAO DE OLINDA’, None)
>>> (None, ‘NILOPOLIS’, ‘MUNICIPIO’)

Exemplo 2

a. *Output* de uma versão do algoritmo anterior àquele descrito na seção 1.3.2:

FINAL DE A RUA B

[(‘FINAL’, ‘N’), (‘DE’, ‘PREP|+’), (‘A’, ‘ART’), (‘RUA’, ‘N’), (‘B’, ‘NPROP’)]

> [(‘FINAL’, ‘N’)]
> [(‘RUA’, ‘N’), (‘B’, ‘NPROP’)]

>>> “FINAL”
>>> “RUA B”

b. *Output* da versão do algoritmo descrito na seção 1.3.2:

FINAL DE A RUA B

[(‘FINAL’, ‘N’), (‘DE’, ‘PREP|+’), (‘A’, ‘ART’), (‘RUA’, ‘N’), (‘B’, ‘NPROP’)]

> [(‘FINAL’, ‘N’)]
> [(‘RUA’, ‘N’), (‘B’, ‘NPROP’)]

>>> “RUA B”

c. *Output* da versão do algoritmo descrito na seção 4.1.1:

FINAL DE A RUA B

[(‘FINAL’, ‘N’), (‘DE’, ‘PREP|+’), (‘A’, ‘ART’), (‘RUA’, ‘N’), (‘B’, ‘NPROP’)]

> [(‘FINAL’, ‘Q’)]
> [(‘RUA’, ‘N’), (‘B’, ‘NPROP’)]

>>> (‘FINAL’, ‘RUA B’, ‘LOGRADOURO’)

Exemplo 3

a. *Output* de uma versão do algoritmo anterior àquele descrito na seção 1.3.2:

LOCAL DE O FATO; RUA DE A LIBERDADE - PROX. 27 - VISTA ALEGRE

[(‘LOCAL’, ‘N’), (‘DE’, ‘PREP|+’), (‘O’, ‘ART’), (‘FATO’, ‘N’), (‘;’, ‘;’), (‘RUA’, ‘N’), (‘DE’, ‘PREP|+’), (‘A’, ‘ART’), (‘LIBERDADE’, ‘N’), (‘-’, ‘-’), (‘PROX’, ‘NPROP’), (‘.’, ‘NPROP’), (‘27’, ‘NPROP’), (‘-’, ‘-’), (‘VISTA’, ‘NPROP’), (‘ALEGRE’, ‘ADJ’)]

> [(‘LOCAL’, ‘N’)]
> [(‘FATO’, ‘N’), (‘;’, ‘;’), (‘RUA’, ‘N’), (‘DE’, ‘PREP|+’), (‘A’, ‘ART’), (‘LIBERDADE’, ‘N’)]
> [(‘PROX’, ‘NPROP’)]
> [(‘27’, ‘NPROP’)]
> [(‘VISTA’, ‘NPROP’), (‘ALEGRE’, ‘ADJ’)]

>>> “FATO ; RUA DA LIBERDADE”
>>> “27”
>>> “VISTA ALEGRE”

b. *Output* da versão do algoritmo descrito na seção 1.3.2:

LOCAL DE O FATO; RUA DE A LIBERDADE - PROX. 27 - VISTA ALEGRE

[(‘LOCAL’, ‘N’), (‘DE’, ‘PREP|+’), (‘O’, ‘ART’), (‘FATO’, ‘N’), (‘;’, ‘;’), (‘RUA’, ‘N’), (‘DE’, ‘PREP|+’), (‘A’, ‘ART’), (‘LIBERDADE’, ‘N’), (‘-’, ‘-’), (‘PROX’, ‘NPROP’), (‘.’, ‘NPROP’), (‘27’, ‘NPROP’), (‘-’, ‘-’), (‘VISTA’, ‘NPROP’), (‘ALEGRE’, ‘ADJ’)]

> [(‘LOCAL’, ‘N’)]
> [(‘FATO’, ‘N’), (‘;’, ‘;’), (‘RUA’, ‘N’), (‘DE’, ‘PREP|+’), (‘A’, ‘ART’), (‘LIBERDADE’, ‘N’)]
> [(‘PROX’, ‘NPROP’)]
> [(‘27’, ‘NPROP’)]
> [(‘VISTA’, ‘NPROP’), (‘ALEGRE’, ‘ADJ’)]

>>> “27”
>>> “VISTA ALEGRE”

c. *Output* da versão do algoritmo descrito na seção 4.1.1:

LOCAL DE O FATO; RUA DE A LIBERDADE - PROX. 27 - VISTA ALEGRE

[(‘LOCAL’, ‘N’), (‘DE’, ‘PREP|+’), (‘O’, ‘ART’), (‘FATO’, ‘N’), (‘;’, ‘;’), (‘RUA’, ‘N’), (‘DE’, ‘PREP|+’), (‘A’, ‘ART’), (‘LIBERDADE’, ‘N’), (‘-’, ‘-’), (‘PROX’, ‘NPROP’), (‘.’, ‘NPROP’), (‘27’, ‘NPROP’), (‘-’, ‘-’), (‘VISTA’, ‘NPROP’), (‘ALEGRE’, ‘ADJ’)]

> [(‘LOCAL’, ‘N’)]
> [(‘FATO’, ‘N’)]
> [(‘RUA’, ‘N’), (‘DE’, ‘PREP|+’), (‘A’, ‘ART’), (‘LIBERDADE’, ‘N’)]
> [(‘PROX’, ‘NPROP’)]
> [(‘27’, ‘NPROP’)]
> [(‘VISTA’, ‘NPROP’), (‘ALEGRE’, ‘ADJ’)]

>>> (None, ‘RUA DA LIBERDADE’, ‘LOGRADOURO’)
>>> (None, ‘27’, None)
>>> (None, ‘VISTA ALEGRE’, ‘BAIRRO’)