DOMENICO ZEMA

MARCELLA GOMES DA COSTA PEREIRA

YURI GIL FERREIRA

APLICAÇÃO DE ARIMA EM PREVISÃO DE TRÁFEGO DE CURTO PRAZO

Trabalho de Formatura do Curso de Engenharia Civil apresentado à Escola Politécnica da Universidade de São Paulo

São Paulo

2020

APLICAÇÃO DE ARIMA EM PREVISÃO DE TRÁFEGO DE CURTO PRAZO

Trabalho de Formatura do Curso de Engenharia Civil apresentado à Escola Politécnica da Universidade de São Paulo

Orientador: Profº. Drº Claudio Luiz Marte

_____

São Paulo

2020

AGRADECIMENTOS

Ao Professor Dr. Cláudio Luiz Marte pela orientação. Por incentivar e apoiar o grupo ao longo de todo o estudo e em meio às dificuldades encontradas.

Ao Professor Olímpio Mendes de Barros pelo auxílio. Pela pesquisa que deu origem a este trabalho. Por estabelecer a ponte com a instituição fornecedora dos dados.

À STM pelo fornecimento das informações sem as quais este trabalho não seria possível.

Por fim, à Escola Politécnica da Universidade de São Paulo. À todos seus funcionários e docentes por proporcionar conhecimento ao grupo.

SUMÁRIO

RESUMO

A rápida evolução dos centros urbanos torna a previsão de tráfego de curto prazo uma importante ferramenta de gestão pública e de fornecimento de informação para sistemas ITS (Intelligent Transport Systems). Além do mais, é grande o volume de dados envolvidos neste processo, bem como é necessária uma forte capacidade de processamento computacional. Pensando nisso, esta pesquisa busca analisar os algoritmos para previsão de tráfego de curto prazo, sendo, desta forma, aplicadas ferramentas de *Data Science* (ARIMA) e métodos estatísticos. Os dados de radar do município de São Paulo foram acumulados por um período de três dias. Este trabalho de formatura tem como base o trabalho de qualificação de mestrado de Olímpio Mendes de Barros (2019): "Caracterização das condições de tráfego em tempo próximo ao real para o uso em sistemas de previsão de tráfego em cidades de grande porte". Foram aplicadas ferramentas de *Data Science* para análise de algoritmos de previsão de curto prazo, desde o pré-processamento de dados até as previsões de até 45 minutos, sem a utilização de simuladores de tráfego. A previsão demonstrou bons resultados nos pontos onde a quantidade de veículos é maior, enquanto cenários com variações abruptas de comportamento ou com baixíssimo tráfego tiveram desempenho inferior. No caso dos radares com maior volume de veículos, o erro percentual é da ordem de 24% para previsões de até 15 minutos (curto prazo) e aumenta progressivamente até atingir erros da ordem de 32% para as tentativas de prever a demanda em até 45 min. Conclui-se portanto que o ARIMA aparentemente possui potencial para ser utilizado na previsão de tráfego, afinal, obteve boas métricas nos pontos mais relevantes (por onde passam mais veículos) em cenários não atípicos.

Palavras-Chave: Engenharia de Transportes. Previsão de Tráfego de Curto Prazo. Linguagem Python. Séries temporais. Autoregressive Integrated Moving Average (ARIMA).

ABSTRACT

The rapid evolution of urban centers makes short-term traffic forecasting an important public management and information provision tool for ITS (Intelligent Transport Systems) systems. Furthermore, the volume of data involved in this process is high, as well as a strong computational processing capacity. With this in mind, this research seeks to analyze algorithms for short-term traffic forecasting, for which there is the application of data science tools (ARIMA) and statistical methods. Radar data from the Municipality of São Paulo was accumulated for a period of three days. This work is based on the master's qualification work of Olímpio Mendes de Barros (2019): "Characterization of traffic conditions in real time for use in traffic forecasting systems in large cities". *Data Science* tools were applied to analyze short-term forecasting algorithms, from data pre-processing to initial predictions of up to 45 minutes, without the use of traffic simulators. The forecast showed good results at the points where the number of vehicles is higher, while scenarios with abrupt variations in behavior or very low traffic had a lower performance. In the case of radars with greater flow, the percentage error is of the order of 24% for predictions of up to 15 minutes (short term) and increases progressively until reaching errors of the order of 32% for attempts to forecast up to 45 minutes. It is concluded, therefore, that ARIMA apparently has the potential to be used in traffic forecasting, after all, it obtained good metrics at the most relevant points (where more vehicles pass) in non-atypical scenarios.

Keywords: Transport Engineering. Short-Term Traffic Forecast. Python. Time series. Autoregressive Integrated Moving Average (ARIMA).

LISTA DE ABREVIATURAS

ACF	Auto-Correlation Function
ADF	Augmented Dickey-Fuller
AR	Autorregression
ARIMA	Autoregressive Integrated Moving Average
CET	Companhia de Engenharia de Tráfego
MA	Moving Average
MAPE	Mean Average Percentual Error
ML	Machine Learning
PACF	Partial Auto-Correlation Function
RMSE	Root Mean Squared Error
SARIMA	Seasonal ARIMA
SAS	Statistical Analysis System
SQL	Structured Query Language
STM	Secretaria de Transportes Metropolitanos

1 CONTEXTUALIZAÇÃO, JUSTIFICATIVA E OBJETIVOS

O incessante crescimento de centros urbanos torna a previsão de tráfego cada vez mais indispensável para a tomada de decisão em questões de gestão de cidades. O fluxo de veículos é afetado por diversos fatores que incluem desde as características das vias e demanda de mobilidade até adversidades como acidentes e intempéries. No entanto, como explicitado por Castro-Neto M. *et al*. (2009), a maioria da literatura a respeito de previsão de tráfego de curto prazo, foca em condições normais.

Dado o grande número de fatores envolvidos em previsões, existem diversos modelos de simulação de tráfego. Os modelos processam uma significativa gama de dados, provindos de diversas fontes em variados formatos. E esses podem ainda apresentar falhas. A fim de se obter uma calibração adequada dos modelos, os dados de entrada devem ser previamente tratados.

As informações históricas envolvidas em previsões precisam ser previamente analisadas e pré-processadas antes de serem submetidas a um modelo preditivo. Além do mais, é necessária uma alta capacidade computacional para processar a quantidade de dados envolvidos em previsão de tráfego. Esta é uma etapa complexa e indispensável para se obter uma boa qualidade final da previsão.

A motivação deste estudo provém do fato de que, na cidade de São Paulo, a Central de Operações da Companhia de Engenharia de Tráfego (CET/SP) dispõe de grande quantidade de informações úteis sobre o tráfego da região, mas ainda existe o desafio de gerir tais informações de maneira eficiente e eficaz. Buscando formas e ferramentas que otimizem os processos e parametrize decisões, Olímpio Mendes de Barros (2019) propõe uma metodologia que permite a organização, o tratamento e a análise de diversas fontes de dados, visando caracterizar as condições de tráfego em tempo próximo ao real. O presente trabalho de formatura visa auxiliar este trabalho de doutorado, discutindo a aplicabilidade de ARIMA na previsão de tráfego de curto prazo, sem a adoção de simuladores de tráfego.

Os algoritmos desenvolvidos podem ser acessados no GitHub por meio do link: <https://github.com/domenicozema/previsao_trafego_tcc>

1.1 Objetivos

O trabalho visa realizar previsões de curto prazo (até 45 minutos) para o volume de tráfego existente na cidade de São Paulo. A previsão tem como finalidade servir de ferramenta para tomada de decisão quanto à gestão do tráfego. Foram utilizados dados de apenas alguns radares e espera-se que o resultado seja generalizável para todos os radares da CET.

1.1.1 Objetivos Gerais

Analisar se o Autoregressive Integrated Moving Average (ARIMA) se mostra um bom modelo de previsão para a cidade de São Paulo, dadas algumas hipóteses para o comportamento do tráfego.
Ao se validar o ARIMA como um modelo razoável, pode-se utilizar o mesmo para projetar as demandas em uma curta janela temporal, auxiliando a CET a monitorar e controlar o tráfego na cidade.

1.1.2 Objetivos Específicos

- Prever o tráfego para janelas temporais específicas (3 a 45 minutos),
- Explicitar o erro esperado para as previsões,
- Procurar os melhores parâmetros para minimizar os erros,
- Garantir que o modelo treinado em poucos radares e em uma curta janela de tempo seja generalizável para outros radares e outras janelas temporais,
- Garantir estabilidade do modelo ao longo do tempo, ou que seja possível retreiná-lo sem grandes dificuldades.

2 REFERÊNCIAS BIBLIOGRÁFICAS

2.1 Python

Programação R, Structured Query Language (SQL) e Statistical Analysis System (SAS) são exemplos de ferramentas adotadas para analisar dados em previsões. Em comparação com as alternativas disponíveis, a linguagem Python foi escolhida pois trata-se da linguagem que apresenta maior familiaridade por parte dos autores. Além desse, há vários outros motivos, dentre eles:

- Usabilidade: por possuir sintaxe simples, é uma linguagem mais intuitiva e fácil de usar;
- Disponibilidade de bibliotecas: fornece grande número de bibliotecas de previsão, tais como ARIMA e Prophet;
- Artigos publicados: há grande quantidade de informações online disponível para consulta, tanto em pesquisas quanto em caso de dúvidas sobre o código em si;
- Escalabilidade: é escalável e de codificação rápida, pois há alta flexibilidade na resolução de problemas que seriam mais complexos em outras linguagens;
- Visualização dos dados: possui bibliotecas de fácil usabilidade, para visualizar dados graficamente ou em formato de tabelas;
- Código aberto: possui um modelo de desenvolvimento comunitário e aberto a qualquer um que deseje acessar o código.

2.2 Modelo de Greenshields

São três as abordagens básicas da análise de tráfego: macroscópica, microscópica e mesoscópica. Segundo Silva (1994), a abordagem macroscópica trabalha com as correntes de tráfego, a abordagem microscópica com a interação entre dois veículos consecutivos em uma corrente de tráfego e a abordagem mesoscópica com grupamentos de veículos que se formam nos sistemas viários. A abordagem macroscópica é satisfatória para validar os dados recebidos no presente estudo, uma vez que se aplica bem a situações de tráfego de alta densidade de veículos. (SILVA, 1994)

Além de permitir o entendimento da capacidade dos sistemas viários, a abordagem macroscópica também permite identificar o resultado de possíveis ocorrências que

provoquem pontos de estrangulamentos nas vias em análise, como afirma Silva (1994). São feitos paralelos com as Leis da Hidrodinâmica e, assim, trata-se de uma abordagem também reconhecida por Analogia Hidrodinâmica do Tráfego.

A fim de entender a abordagem macroscópica de tráfego é necessário primeiro definir três grandezas básicas: fluxo ou volume (F), concentração ou densidade (D) e velocidade (V).

Greenshields (1935) define o fluxo (F) como uma variável temporal que representa o número de veículos que passa numa determinada seção da via em um intervalo de tempo. Sendo T o intervalo de tempo definido, contam-se os n(x) veículos que passam por uma determinada seção ao longo de uma distância x da via em análise. O fluxo, formulado na Equação 1, é análogo à vazão de um fluido dentro de um duto na Hidrodinâmica.

Equação 1

$$F(x) = \frac{n(x)}{T}$$

A densidade (D) é uma variável espacial que representa a distribuição de veículos na via em um determinado comprimento X do trecho, durante um instante de tempo t. Ela é análoga à densidade de fluido na Hidrodinâmica e é expressa segundo a Equação 2.

Equação 2

$$D(t) = \frac{n(t)}{X}$$

Por fim, a velocidade é obtida pela relação do fluxo pela densidade, como indicada na Equação 3. A expressão resultante é conhecida com Equação Fundamental do Tráfego.

Equação 3

$$V = \frac{F(x)}{D(t)}$$

Modelos objetivam representar experimentalmente a realidade. Com base teórica nas leis da hidrodinâmica, um modelo pioneiro na teoria do fluxo de tráfego foi desenvolvido por Greenshields (1935). Para entendê-lo, definem-se:

Tabela 1: Variáveis macroscópicas do tráfego.

v_f	Velocidade de fluxo livre é aquela medida em uma seção como a máxima velocidade média praticada (v_f = Vmáx) pelos usuários quando restritos apenas pela via (suas características físicas e de controle de tráfego)
D_j	Densidade máxima, correspondente à situação de completo congestionamento
$F_{máx}$	Fluxo máximo que pode ser escoado a partir de uma fila sem interrupção
v_o	Velocidade ótima, correspondente ao ponto que se alcança o fluxo máximo
D_o	Densidade ótima, correspondente ao ponto em que se alcança o fluxo máximo

Fonte: adaptado de SILVA, P. C. M. (1994)

O modelo de Greenshields pressupõe linear a relação entre a velocidade e a densidade, assim como parabólicas as relações entre velocidade e fluxo e entre fluxo e densidade, como ilustrado a seguir:

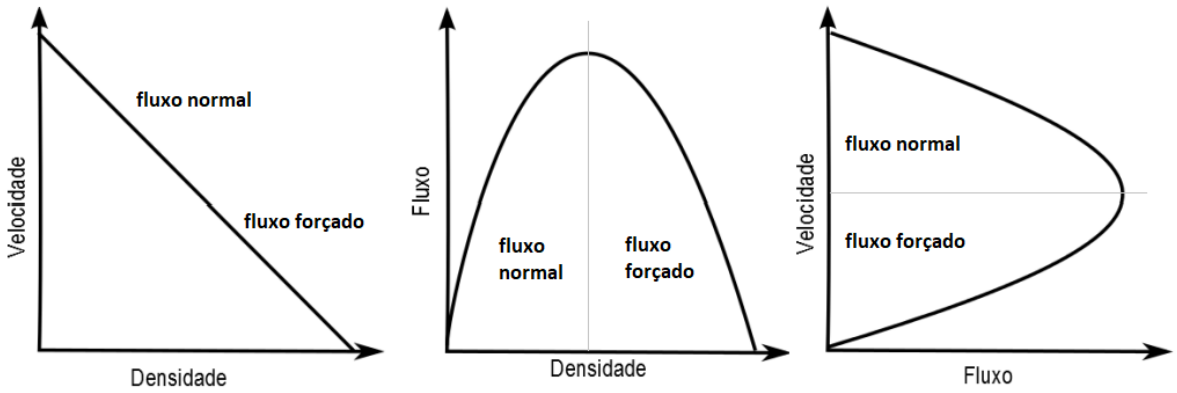

Figura 1: Representação do modelo de Greenshields (1935)

O modelo linear da relação entre velocidade e densidade tem representação gráfica mais detalhada como na Figura 2 e é formulado segundo a Equação 4.

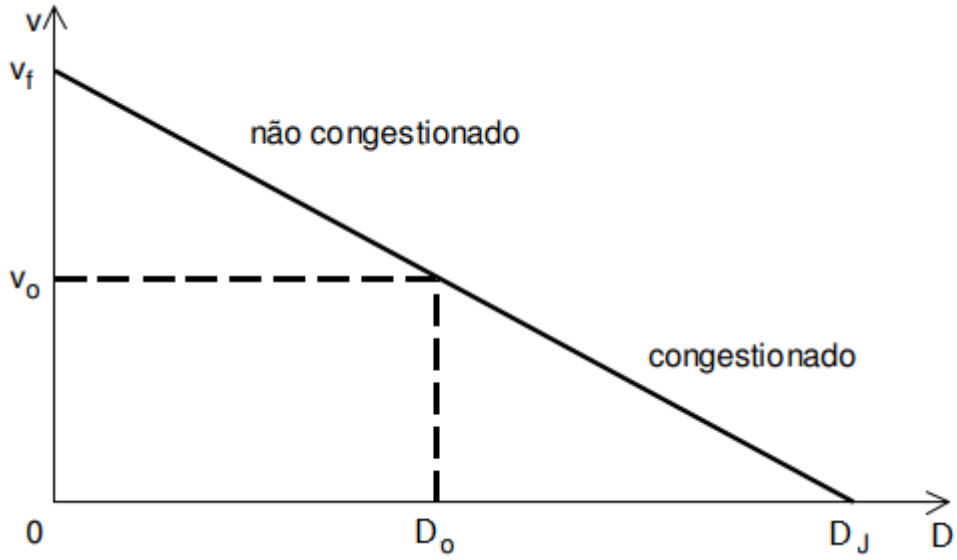

Figura 2: Relação linear entre velocidade e densidade. Greenshields (1935)

Equação 4

$$v = v_f\left(1 - \frac{D(t)}{D_j}\right)$$

É importante ressaltar que observações de campo indicam que essa relação linear é apenas válida para valores intermediários de v e D, como mostra a Figura 3.

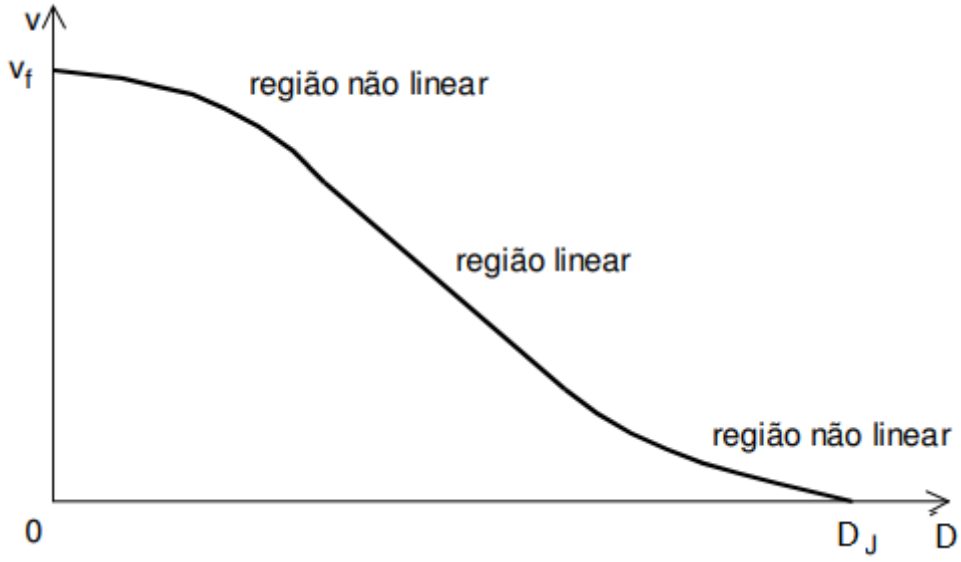

Figura 3: Observações de campo - relação velocidade e densidade. Greenshields (1935)

O modelo parabólico de Greenshields que relaciona fluxo e densidade está representado na Figura 4 e tem formulação segundo a Equação 5, a seguir representada. No ponto

correspondente a Fmáx, tem-se $D_o = D_j/2$ e, consequentemente, $v_o = v_f/2$ (Equação 4).

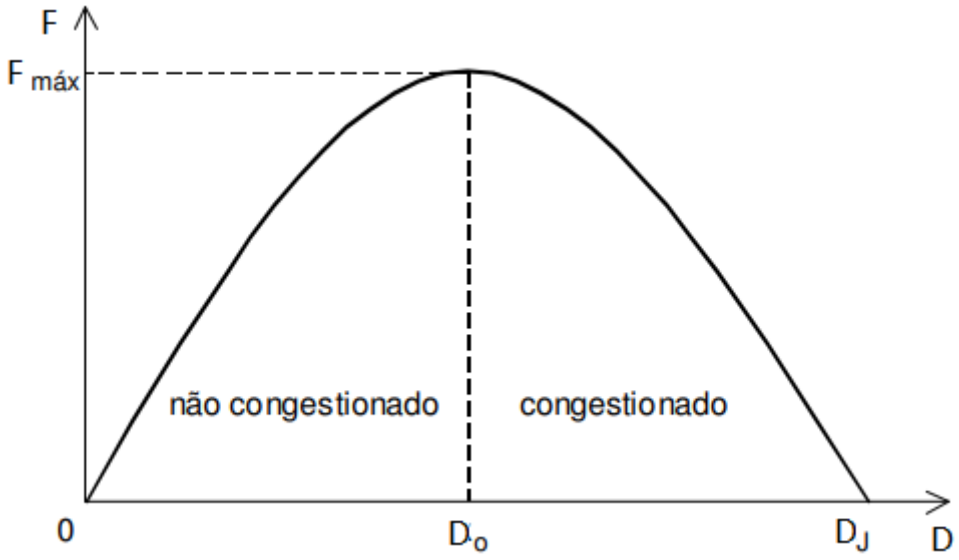

Figura 4: Curva parabólica da relação fluxo e densidade. Greenshields (1935)

Equação 5

$$F = v_f\left(D - \frac{D^2}{D_j}\right)$$

No entanto, observações de campo concluem que esta curva na verdade é assimétrica e está mais próxima daquela indicada na Figura 5.

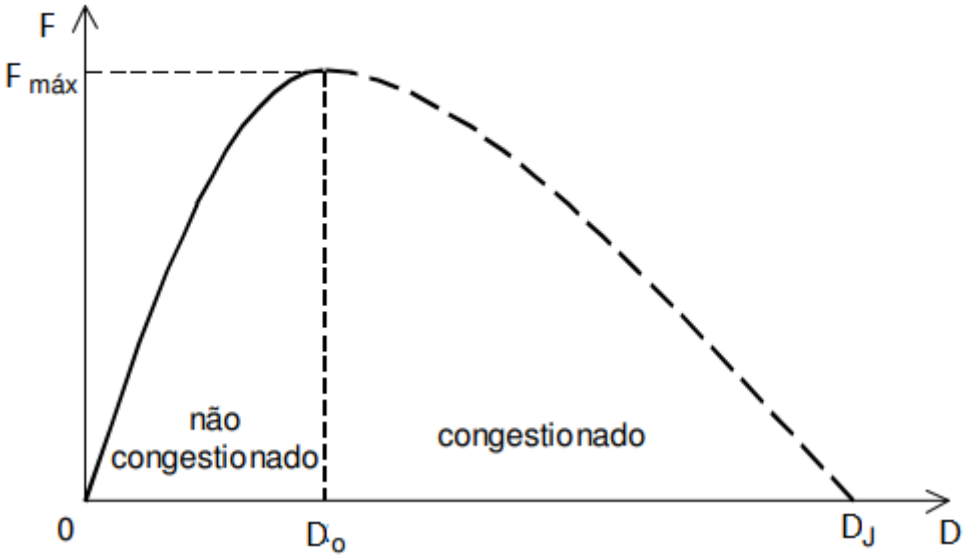

Figura 5: Observações de campo - fluxo e densidade. Greenshields (1935)

Greenshields relaciona a velocidade e o fluxo como representado na parábola da Figura 6. Essa curva tem sua formulação segundo a Equação 6, a seguir representada. No ponto correspondente a Fmáx, tem-se $D_o = D_j/2$ e consequentemente $v_o = v_f/2$ (Equação 4).

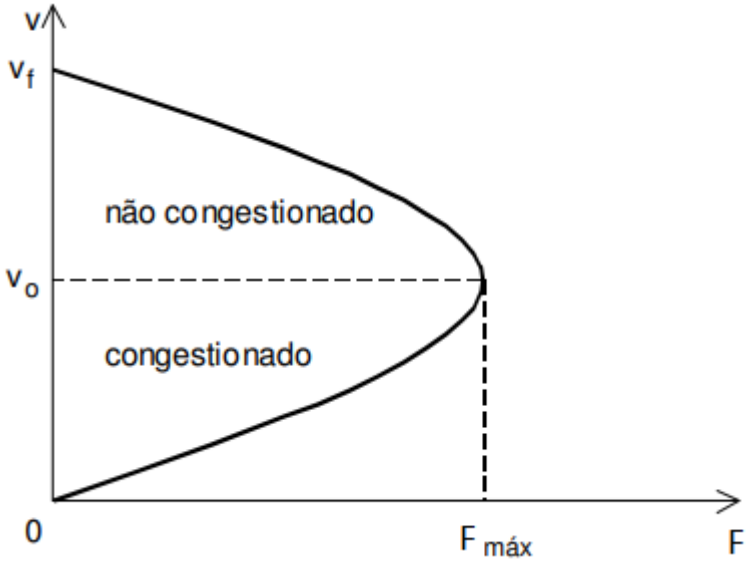

Figura 6: Relação parabólica entre velocidade e fluxo. Greenshields (1935)

Equação 6

$$F = D_j\left(v - \frac{v^2}{v_f}\right)$$

Esta não é uma hipótese com precisão aceitável em situações práticas, entretanto, permite uma análise qualitativa interessante para estudo da operação de tráfego. Portanto, será utilizada para este fim.

2.3 Séries temporais

Conforme Wooldridge (2000), uma série temporal é uma sequência de observações de uma variável ao longo do tempo. Ou seja, é um conjunto sucessivo de observações realizadas em intervalos uniformes durante um certo período.

Como afirma Moreira (2008) há muitos métodos de previsão, porém todos compartilham a característica de que o comportamento do passado observado interfere no comportamento do futuro estimado. Em outras palavras, previsões consideram somente comportamentos prévios e logo não são precisas. Assim, quanto maior o tempo de previsão, maior poderá ser o erro

envolvido. Além disso, os métodos de previsão estão sujeitos à disponibilidade de dados, tempo e recursos.

As séries temporais contam com três padrões básicos: tendência, sazonalidade e ciclo. A tendência informa como se comporta a série no longo prazo, ou seja, se cresce, decresce ou se mantém e como é a velocidade dessas alterações. Sazonalidade indica as oscilações recorrentes em determinado período. Ciclos são oscilações ao longo do tempo de forma suave e repetida ao longo da componente de tendência. A diferença fundamental entre ciclo e sazonalidade é que esta última é previsível e ocorre sempre no mesmo período, enquanto as oscilações de um ciclo podem ser irregulares (HYNDMAN, R. J.; ATHANASOPOLUS, G., 2013)

Ao estudar séries temporais os objetivos são principalmente dois. Um deles é analisar e modelar a série temporal, ou seja, descrever o fenômeno, checar características importantes e sua possível relação com outras séries. O segundo objetivo é realizar previsões, ou seja, estimar valores futuros baseados em valores históricos observados.

Existem diversos procedimentos de previsão que podem ser simples e intuitivos ou complexos e racionais. Dentre os modelos estatísticos de previsão, Gutierrez (2003) cita:

- “Modelos Univariados: Baseia-se em apenas uma série histórica. Alguns deles são:
    - Decomposição por componentes não observáveis, que foi o mais adotado até 1960;
    - Modelos automáticos que integram modelos de regressão, de médias móveis, ajustamento sazonal e alisamento exponencial;
    - Modelos univariados de Box & Jenkins (1970). Trata-se de uma classe geral de modelos lineares, mais conhecidos como ARIMA;
- Modelos de Função de Transferência: nos quais a série de interesse é explicada não só pelo seu passado histórico, como também por outras séries temporais não correlacionadas entre si.
- Modelos Multivariados: modelam simultaneamente duas ou mais séries temporais sem qualquer exigência em relação à direção da causalidade entre elas.”

Por fim, vale ressaltar que previsões não são um fim em si, mas precisam ser entendidas como parte integrante de um processo de tomada de decisão, visando a objetivos claros e bem definidos (MORETTIN; TOLOI, 2006). Neste caso, o fim é auxiliar entidades como a CET-SP na tomada de decisão quanto à gestão de tráfego.

2.4 ARIMA

O presente trabalho de formatura fez uso do método ARIMA para realizar predições de tráfego de curto prazo. Logo, será feita uma breve introdução teórica acerca do modelo a fim de esclarecer suas fundamentações teóricas, bem como vantagens e desvantagens. Para melhor entendimento, cada uma das técnicas será abordada isoladamente para então concluir como se comporta a combinação de todas elas. Vale ressaltar que é possível utilizar outras técnicas de previsão (como SARIMA ou modelos causais de regressão), no entanto, apenas o ARIMA será abordado.

O modelo ARIMA (*Autoregressive Integrated Moving Average*) é resultante da combinação de três técnicas de previsão: Auto-Regressão, Integração (Diferenciação) e Médias Móveis. Ele busca reconhecer padrões nas observações históricas para então realizar uma previsão do(s) valor(es) seguinte(s). Trata-se de uma generalização do já difundido ARMA, o qual combina apenas técnicas de auto-regressão e médias móveis.

2.4.1 Auto-regressão

A técnica da auto-regressão, em linhas gerais, busca encontrar os coeficientes de um polinômio que tem como variáveis as medidas já observadas da série temporal. Como afirma Adhikari (2009), os modelos *AR(p)* partem da hipótese de que o valor a ser previsto depende linearmente das *p* observações precedentes e de um termo estocástico. Quanto maior for a ordem (*p*), mais dados históricos serão utilizados na previsão.

Desse modo, um modelo *AR(p)* pode ser formulado da seguinte maneira (Equação 7):

Equação 7

$$y_t = c + \sum_{i=1}^{p} \varphi_i y_{t-i} + \varepsilon_t$$
$$= c + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \cdots + \varphi_p y_{t-p} + \varepsilon_t$$

Onde y são as observações da série histórica, φ_k são os parâmetros do modelo e ε é o erro de previsão. Para realizar uma previsão, basta encontrar a combinação de parâmetros que minimizam o erro. A forma mais simples deste modelo é o modelo de ordem 1, AR(1), o qual utiliza apenas a observação mais recente (Equação 8).

Equação 8 $$AR(1): y_t = c + \varphi_1 y_{t-1} + \varepsilon_t$$

Os modelos *AR(p)* de ordens baixas são equações relativamente simples de serem solucionadas. No entanto, à medida que a ordem aumenta, o que normalmente resulta em resultados mais precisos, a complexidade da equação também é elevada. É muito comum o auxílio de linguagens de programação (R ou Python) para solucioná-las.

Modelos de ordem mais alta também se mostram muito mais precisos nos dados de treino (calibração), que são os dados que o modelo já viu. Enquanto na população de teste (ou nos dados do futuro) apresentam erros maiores. Isso se dá ao fenômeno conhecido como *overfitting.* O ideal é a população de treino (calibração) e teste terem erros semelhantes, o que garante que o modelo está generalizando para populações as quais ele não viu e que estão em uma janela temporal diferente da de treino (calibração).

2.4.2 Média Móvel

Como afirma Hyndman (2013), diferentemente da auto-regressão, a média móvel é uma técnica que envolve o processo de amortecimento. Isto é, a regressão é feita não somente baseada nos valores anteriores, mas sim no erro em relação a elas. Utilizar a diferença entre os diferentes pontos da série histórica é uma técnica para conseguir identificar a tendência nela existente. Devido a isso, tal técnica é comumente utilizada para prever – com boa precisão – séries nas quais há uma tendência não muito acentuada e com pouca variação abrupta.

Conforme afirma Adhikari (2009), a formulação do modelo de médias móveis é muito semelhante ao auto regressivo, como pode ser observado na Equação 9.

Equação 9

$$y_t = \mu + \sum_{j=1}^{q} \theta_i \varepsilon_{t-j} + \varepsilon_t$$
$$= \mu + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_p \varepsilon_{t-q} + \varepsilon_t$$

Onde θ são os parâmetros do modelo, μ é a média da série e ε os erros de previsão passados (diferença entre o valor do instante *t-j* e a média). Assume-se a hipótese de que esses erros são variáveis aleatórias independentes e identicamente distribuídas (i.i.d.) com média nula e variância σ^2. Logo, a média móvel trata-se – em essência – de uma regressão linear da observação presente em relação às diferenciações dos termos precedentes.

Para efeitos ilustrativos, o equacionamento de um modelo de média móvel de primeira ordem *MA(1)* é (Equação 10):

Equação 10

$$MA(1): y_t = \mu + \theta_1 \epsilon_{t-1} + \varepsilon_t$$

Para médias móveis de baixa ordem (1 ou 2), é possível provar – segundo Hyndman (2013) que os parâmetros do modelo devem obedecer a certas restrições:

- Para o modelo MA(1): $-1 < \theta_1 < 1$
- Para o modelo MA(2): $-1 < \theta_2 < 1$, $\theta_2 + \theta_1 > -1$, $\theta_1 - \theta_2 < 1$

Ao aumentar a ordem do modelo, relações mais complexas devem ser obedecidas. Nos pacotes de previsão do Python, os modelos pré-definidos garantem que tais restrições serão respeitadas, não sendo necessário entrar em detalhes para restrições existentes nos modelos de ordem 3 ou superior.

2.4.3 Modelo ARMA

Combinando os métodos de auto regressão e média móvel, é possível constituir o chamado modelo ARMA(p, q). Enquanto a média móvel captura a tendência da série em estudo, a auto regressão captura a dispersão da série, tornando a junção de ambos capaz de prever com maior acurácia ambos os fenômenos. Nota-se que o modelo possui tanto a ordem da regressão

(p) quanto da média móvel (q), sendo apenas a soma dos dois modelos (KRUK, 2002) – Equação 11.

Equação 11

$$y_t = c + \sum_{i=1}^{p} \varphi_i y_{t-i} + \sum_{j=1}^{q} \theta_i \varepsilon_{t-j} + \varepsilon_t$$

De acordo com Valipour *et al.* (2013), mesmo já sendo difundido, principalmente nas áreas econômica e hidrológica, o ARMA é ainda um método relativamente simples por contar única e exclusivamente com a série histórica, sem a possibilidade de diferenciá-la para garantir a estacionariedade. Caso o termo ε seja variável ao longo do tempo, o modelo acaba por perder consistência ao prever um intervalo mais longo à frente (JAIN, 2017). A variação, na maioria dos casos, ocorre no longo prazo. Caso a mudança num curto espaço de tempo seja pequena – cenário mais comum - o método continua sendo efetivo no curto prazo (cinco minutos a uma hora, segundo Kumar, 2015).

2.4.4 Diferenciação

A diferenciação é a técnica que separa os modelos ARMA e ARIMA. Como dito anteriormente, o modelo ARMA é sensível a alterações do termo estocástico. Quando o termo não varia, a série é dita estacionária. Para obter uma maior acurácia em séries não estacionárias – também chamadas de séries integrais (SILVA, 2017) – utiliza-se a técnica da diferenciação – Equações 12 e 13.

Equação 12

$$y'_t = y_t - y_{t-1}$$

Equação 13

$$y''_t = (y_t - y_{t-1}) - (y_{t-1} - y_{t-2})$$

As equações 12 e 13 mostram, respectivamente, as diferenciações de primeiro e segundo grau. Diferenciar uma séria é uma técnica comumente utilizada para obter uma série estacionária (MILLS, 1991). Caso a série possua um grau de integração I, serão necessárias I diferenciações para torná-la estacionária.

2.4.5 O Modelo ARIMA

Combinando as técnicas de auto regressão (AR), média móvel (MA) e diferenciação, obtemos o modelo ARIMA(p, d, q). Como discutido anteriormente, tal modelo foi desenvolvido para eliminar a limitação do modelo ARMA de ser efetivo apenas em séries estacionárias. A única diferença entre tais modelos é, de fato, a diferenciação. Pode-se ver, pelas equações 11 e 14 que a formulação de ambos os métodos é muito similar.

Equação 14

$$y'_t = c + \sum_{i=1}^{p} \varphi_i y'_{t-i} + \sum_{j=1}^{q} \theta_i \varepsilon_{t-j} + \varepsilon_t$$

Neste caso, entretanto, a equação é construída com termos diferenciados (y'_t), podendo se tratar de qualquer nível de diferenciação. Por ser o modelo mais geral dentre os discutidos anteriormente, é possível obtê-los simplificando os parâmetros do ARIMA. Por exemplo:

- ARIMA(p, 0, 0) = AR(p)
- ARIMA(0, 0, q) = MA(q)

De acordo com Pavlyuk (2016), este é um dos métodos mais comumente utilizados na previsão de tráfego. Além disso, trata-se de um método já implementado em Python (existente em bibliotecas estatísticas, como a *statsmodels*). Por tais razões, optou-se por usá-lo a fim de ter bases mais sólidas para avaliação e comparação dos resultados. Serão estudados os melhores parâmetros (p, d, q) de modo a minimizar o erro de previsão.

2.4.6 Definição dos Parâmetros

Para que o modelo apresente resultados aceitáveis, é necessário que sejam previamente definidos os parâmetros de cada um dos componentes do modelo ARIMA. Eles dependem do comportamento da série temporal a qual está sendo estudada e, através dela, podem ser

determinados com o auxílio do teste de Dickey-Fuller aumentado (ADF) e das funções de autocorrelação (ACF) e autocorrelação parcial (PACF) (BOX *et al*, 1976).

2.4.6.1 Teste de Dickey-Fuller Aumentado (ADF)

Para que o ARIMA possa ser aplicado à determinada série temporal, é necessário que seja estacionária. Caso não haja taxa de crescimento significativa, a série pode ser considerada estacionária (e caso haja, deve-se diferenciá-la a fim de atingir um estado de estacionaridade). Apenas observar a série para inferir suas propriedades não é uma metodologia correta ou aceitável.

Uma maneira mais precisa de verificar tal propriedade, numa sequência de observações, é utilizar o teste Augmented Dickey-Fuller (ADF). Ele verifica se cada termo y_t está diretamente relacionado com seu anterior y_{t-1} utilizando a técnica de regressão linear (FULLER, 1976) – Equação 15

Equação 15
$$\Delta y = \alpha + \beta t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \delta_2 \Delta y_{t-2} + \cdots$$

O teste verifica, então, se o coeficiente do termo y_{t-1} (γ) é ou não zero. Em caso positivo, não podemos rejeitar a hipótese nula e a série pode ser considerada estacionária. Caso contrário, a hipótese nula de estacionariedade é rejeitada e a série deverá ser diferenciada antes da aplicação do modelo ARIMA. Caso seja necessária a diferenciação, deve-se ainda definir o grau de diferenciação necessário (parâmetro d), o que é feito com o auxílio da ACF, conforme descrita a seguir.

2.4.6.2 Função de Autocorrelação (ACF)

Assim como a correlação mede a intensidade da relação linear entre duas variáveis, a autocorrelação mede a relação linear entre os valores defasados de uma série temporal (HYNDMAN, 2013). Plotando a *Autocorrelation Function (*ACF), cada valor da série corresponde ao quão relacionados estão os valores da série em relação aos seus antecedentes. Por exemplo, o termo r_1 mede a relação entre os termos y_t e y_{t-1}, enquanto o termo r_2 mede a relação entre y_t e y_{t-2}. Os valores da ACF são calculados pela Equação 16.

Equação 16

$$r_k = \frac{\sum_{t=k+1}^{T} (y_t - E(y))(y_{t-k} - E(y))}{\sum_{t=1}^{T} (y_t - E(y))^2}$$

Assim sendo, o primeiro termo da curva – ACF(0) – é sempre 1, tendo em vista que os valores da série temporal são sempre perfeitamente correlacionados consigo mesmos. Usualmente, a função é inicialmente decrescente até que os valores ao longo do eixo x se estabilizem variando em torno de 0, afinal, valores muito distantes no passado normalmente possuem pouca ou nenhuma correlação com os valores atuais (salvo em caso de sazonalidade).

2.4.6.3 Função de Autocorrelação Parcial (PACF)

A ACF carrega a autocorrelação intermediária entre dois termos da série temporal, o que pode resultar em interpretações errôneas. Em alguns casos, um valor da série está fortemente correlacionado com seu antecedente, e essa forte relação entre os valores causará a impressão de que a observação atual também está relacionada a todas as anteriores. Para que tal efeito seja contornado, é também utilizada a Partial Auto-Correlation Function (PACF).

A autocorrelação parcial é calculada levando em conta a autocorrelação e variâncias condicionais da série temporal, de modo a eliminar o efeito de termos intermediários entre dois valores da curva (ROMER, 2019). O primeiro termo da PACF é igual ao da ACF, tendo em vista que não há termos intermediários entre eles. Do segundo termo em diante, possuímos a seguinte métrica de cálculo – Equações 17, 18 e 19:

Equação 17

$$PAC(1) = AC(1)$$

Equação 18

$$PAC(2) = \frac{Cov(y_t, y_{t-2}|y_{t-1})}{\sqrt{Var(y_{t-1})Var(y_{t-2}|y_{t-1})}}$$

Equação 19

$$PAC(3) = \frac{Cov(x_t, x_{t-3}|x_{t-1}, x_{t-2})}{\sqrt{Var(x_{t-1}, x_{t-2})Var(x_{t-3}|x_{t-1}, x_{t-2})}}$$

Para exemplificar a diferença entre ambas as funções, foram construídas as curvas de ACF e PACF para dois dia (06/09/2018, quinta-feira e 21/09/2018, sexta-feira) de volume de tráfego do radar 4314, agregados em intervalos de três minutos e localizado no cruzamento da Av. Salim Farah Maluf (Bairro/Centro) com a Av. Vila Ema. Pela imagem, pode-se ver que a autocorrelação parcial decai rapidamente, enquanto a correlação pode ser observada entre 100 lags (número de observações passadas), puramente por efeito dos primeiros valores.

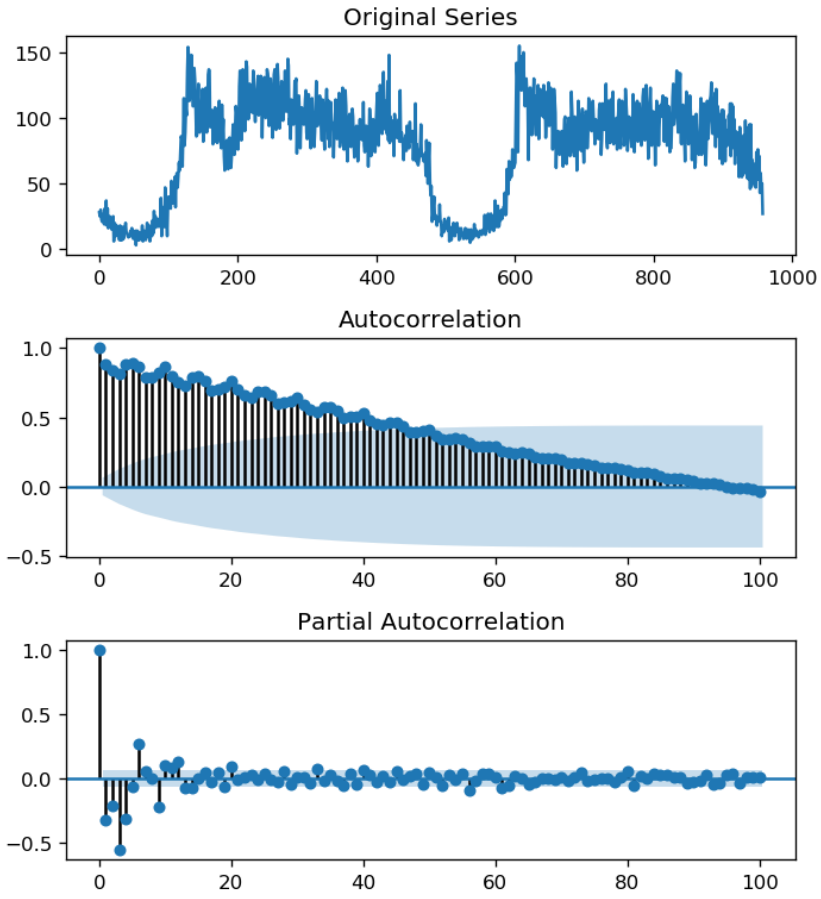

Figura 7: Curvas de ACF e PACF

A questão ainda não respondida é como tais curvas e testes podem auxiliar na determinação de quais são os parâmetros adequados de um modelo ARIMA. Seguiremos a sequência de passos proposta por Brownlee (2017), composta por quatro etapas:

1. **Verificar Estacionariedade:** primeiramente é realizado um teste ADF para verificar se a série é estacionária ou não. Caso seja necessário diferenciá-la, a etapa 2 é realizada para determinar quantas vezes é necessário fazê-lo.
2. **Grau de Diferenciação:** caso o valor "p", obtido no teste, seja inferior ao nível de significância desejado (usualmente 0.05), será necessário diferenciar a série. Para

determinar o grau de diferenciação, deve-se plotar a ACF para a série com diversas ordens de diferenciação e selecionar aquela que antecede uma ACF com queda abrupta para valores negativos, indicando que a série está super diferenciada.

3. **Ordem da Auto-regressão:** por se tratar de uma regressão feita sobre os valores anteriores da série, deve-se ver até que ponto há correlação entre os *lags*, que é definido através da PACF. Pelo plot podemos observar até que ponto a correlação está acima do nível de confiança especificado. Qualquer ordem do modelo AR até esse ponto pode agregar informação, fornecendo um *threshold* máximo, ou seja, um limite máximo aceitável (área em azul mais claro na Figura 7) para a ordem do modelo.
4. **Ordem da Média Móvel:** analogamente à etapa três, o plot da ACF fornece um *threshold* superior para o ordem do modelo MA. Vale ressaltar que esse valor pode ser muito elevado, então deve-se sempre levar em conta tanto o fator de performance do modelo quanto custo computacional, evitando o uso de parâmetros muito elevados em quaisquer modelos.

2.4.7 Aplicação

O método ARIMA pode ser aplicado na previsão de características do tráfego – como demanda, velocidade, etc – desde que elas obedeçam a algumas hipóteses comumente adotadas por pesquisadores da área, como Pavlyuk (2016).

1. Relação linear entre os valores predecessores e atuais: assumimos que há linearidade entre os diferentes pontos do tempo. Normalmente, é uma hipótese não absurda para predições no curto prazo (o qual é o tema do presente trabalho de formatura).
2. Independência das características do tráfego no tempo: possibilidade de utilizar um modelo distinto para cada uma das características do tráfego. Sem essa hipótese, seria necessário utilizar uma regressão múltipla em várias séries para estimar o valor de uma única característica.
3. Fluxo estacionário: podem ocorrer fenômenos – como grandes eventos, acidentes, intempéries – capazes de gerar grande impacto no comportamento do tráfego, mas sem relações de causalidade com o mesmo. Assumir um fluxo estacionário significa analisar o cenário “padrão”, onde não há a ocorrência de tais anomalias.

4. Independência espacial entre as observações: suposição de que o ocorrido em determinado ponto não afeta os demais. Isso evita, mais uma vez, que seja necessária a regressão de múltiplas séries em apenas uma. Sabe-se que tal hipótese é muitas vezes falha e, por isso, pode-se optar pela utilização de novas técnicas, como matrizes de rede – visando estimar a relação entre dois pontos da rede de tráfego. Mas, tais técnicas não fazem parte do escopo desse trabalho de formatura.

O modelo ARIMA, logo, pode ser utilizado, mas mantendo-se sempre atento às hipóteses levantadas – ver Figura 8. Caso sejam inválidas (por exemplo, prever dois pontos muito próximos como se fossem independentes), é necessária uma reavaliação do modelo utilizado.

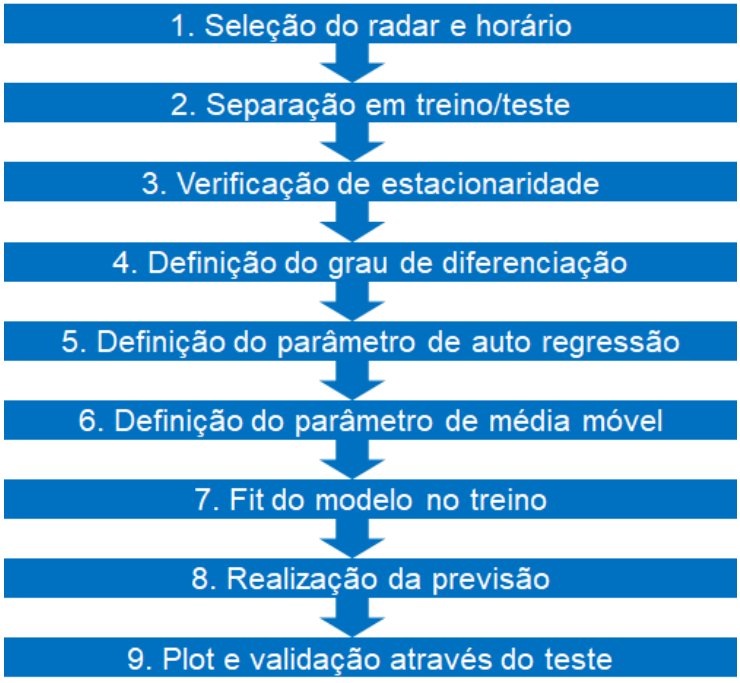

Figura 8: Fluxograma para modelagem em ARIMA

3 DADOS UTILIZADOS E METODOLOGIA

3.1 Coleta dos dados

Para o presente trabalho de formatura foram disponibilizados três dias de medições dos radares da cidade São Paulo. Os dias foram: (09/05/2018, quarta-feira, 06/09/2018, quinta-feira e 21/09/2018, sexta-feira).

3.2 Preparação dos dados

Recebemos as bases de dados dos radares de São Paulo, cordialmente cedidos pela STM. Os dados fornecidos eram inicialmente vários arquivos ".txt", emitidos individualmente por cada um dos 1280 radares. Os dados foram recebidos, pré-processados e em seguida foi feita a compilação e anonimização todos esses arquivos em formato ".csv", separados por dia e cada um deles contendo cerca de 6.6 milhões de linhas. Ver Figura 9.

Ao tentar processar tais arquivos, inicialmente houve problemas de encoding (codificação), causados por caracteres especiais contidos no endereço de algumas localidades. Foi realizado, então, um tratamento de tais linhas, no qual foram removidos tais caracteres para que o arquivo pudesse ser processado sem maiores problemas. Em seguida, foi realizado um tratamento de valores nulos, nos quais pontos com medidas nulas de velocidade ou de localização foram descartados. O número de linhas removidas nessa etapa foi extremamente pequeno, não ultrapassando 0,5% da base, um indício de que os dados possuem um bom nível de qualidade.

Em seguida, foi realizado o agrupamento dos dados em períodos de três minutos, computando a quantidade de veículos que passaram nesse intervalo de tempo e também a velocidade média em que estavam – ver Figura 10. Assim, foram constituídas as séries temporais para os três dias de dados fornecidos.

	Lote	Data	Hora	Local	Faixa	EntreFaixa	Registro	TipoRegistro	Placa	EspecieVeiculo	ClasseVeiculo	Comprimento	VelocidadePontual
0	L2	20180508	235603	4326	5	0	596620	0.0	NaN	1.0	0.0	32.0	142.0
1	L2	20180508	235607	4326	4	0	596621	0.0	NaN	0.0	0.0	0.0	0.0
2	L2	20180508	235607	4326	4	0	596622	0.0	NaN	0.0	0.0	0.0	0.0
3	L2	20180508	235608	4326	5	0	596623	0.0	NaN	0.0	1.0	191.0	119.0
4	L2	20180508	235643	4326	2	0	596624	0.0	NaN	1.0	0.0	32.0	42.0

TempoOcupacao	VelocidadeMedia	Empresa	Local2	Descricao	Latitude	Longitude	LongitudeLatitude	ValorAgrupado
536.0	0.0	L2	4326	Av. Alcântara Machado (Centro/Bairro) a mais 1...	-23.549602	-46.605001	-46.60500096 -23.54960196	10301.0
787.0	0.0	L2	4326	Av. Alcântara Machado (Centro/Bairro) a mais 1...	-23.549602	-46.605001	-46.60500096 -23.54960196	10301.0
926.0	0.0	L2	4326	Av. Alcântara Machado (Centro/Bairro) a mais 1...	-23.549602	-46.605001	-46.60500096 -23.54960196	10301.0
2229.0	0.0	L2	4326	Av. Alcântara Machado (Centro/Bairro) a mais 1...	-23.549602	-46.605001	-46.60500096 -23.54960196	10301.0
1832.0	0.0	L2	4326	Av. Alcântara Machado (Centro/Bairro) a mais 1...	-23.549602	-46.605001	-46.60500096 -23.54960196	10301.0

Figura 9: Dados originados pelos radares

	Data	Local	VelMed	Qtde
0	2018-05-09 00:00:00	4326	90.000000	2
1	2018-05-09 00:03:00	4326	117.611111	18
2	2018-05-09 00:06:00	4326	94.000000	35
3	2018-05-09 00:09:00	4326	113.052632	19
4	2018-05-09 00:12:00	4326	104.000000	17

Figura 10: Dados agrupados por Local

3.3 Validação dos dados

Com os dados em mãos, foi necessário validar a qualidade deles Para tal, foram selecionados 55 radares localizados na região de Pinheiros. A partir das coordenadas geográficas, os radares foram locados em um mapa, utilizando o programa QGIS. A validação espacial mostrou-se satisfatória, como ilustra a Figura 11.

Figura 11: Radares analisados para validação espacial

Para validação qualitativa, segundo as teorias de fluxo de tráfego, os dados foram analisados segundo o modelo de Greenshields. Embora o modelo seja aplicável apenas para vias expressas e rodovias, onde o fluxo é contínuo, a sua aplicação neste caso é suficientemente aceitável para validação dos dados. Observa-se que os pontos obtidos seguem o modelo de Greenshields, pois a relação entre velocidade e fluxo é parabólica como demonstrado na Figura 12. Conclui-se que os dados possuem qualidade para serem aplicados ao modelo desenvolvido.

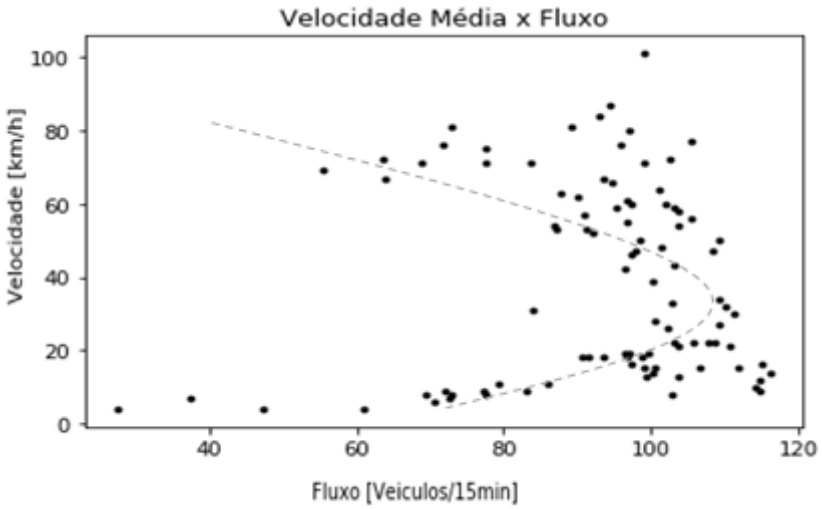

Figura 12: Validação dos dados segundo o modelo de Greenshields

Ao analisar a série histórica de alguns pontos, percebeu-se que o primeiro dia de dados fornecidos possuía comportamento anômalo, com volumetria muito inferior e picos em horários diferentes. Analisando mais a miúde, nota-se que se trata de uma data na qual os

caminhoneiros estavam em greve, o que comprometeu a representatividade de tais dados. Assim sendo, esse dia foi descartado e seguiram-se as análises com os dois últimos dias apenas.

3.4 Seleção do Modelo

Como o objetivo do presente trabalho é realizar predições de curto prazo sobre volumes de tráfego, é preciso definir qual a técnica de previsão que será utilizada, tendo em vista que há de necessidade uma preparação dos dados para que eles estejam no formato adequado para a aplicação do modelo. Dessa forma, pondera-se sobre quais dados estão disponíveis, qual o custo computacional necessário para cada técnica e qual é a técnica recomendada para os dados em questão.

Atualmente, há uma larga gama de modelos computacionais utilizados tanto em ambientes profissionais quanto competições do *Kaggle* (plataforma online na qual há competições de Machine learning pela criação do melhor modelo dado um problema – www.kaggle.com) ou trabalhos acadêmicos. A grande maioria desses modelos são algoritmos de *machine learning* (ML), uma área do conhecimento que está em ascensão nos últimos anos. Além de tais modelos, há também técnicas mais tradicionais com foco na análise de séries temporais, como médias móveis, amortecimento exponencial, etc.

A grande vantagem da utilização de algoritmos de ML (como *Random Forest, XGBoost*, *LightGBM*, redes neurais, etc) é a possibilidade de se inserir variáveis exógenas na previsão (MORANTZ, 2004). Por exemplo, no caso de tráfego, sabe-se que intempéries são fenômenos extremamente relevantes no tempo de viagem em grandes cidades. Utilizar a pluviometria poderia, então, gerar resultados mais precisos pois traz informações não contidas na série temporal por si só.

No cenário desse trabalho de formatura, entretanto, há a disponibilidade de apenas dois dias de dados do banco de radares fornecidos pela CET-SP. Com tais dados, é possível a criação de uma série temporal para cada um dos radares. Como o grupo não dispõe de nenhum tipo de banco de dados externo confiável (e pensando numa solução produtiva, seria também muito complexo atualizar tais informações em tempo próximo ao real), optamos por utilizar uma técnica exclusivamente de *time series*.

Dentre as técnicas cotadas, estão: médias móveis (MA), auto-regressão (AR), amortecimento exponencial, ARIMA e SARIMA (Seasonal-ARIMA). Como apenas dois dias de dados úteis foram disponibilizados, não seria de grande valia utilizar o modelo com sazonalidade, pois apesar de ser o mais complexo entre os cinco, a quantidade de dados é insuficiente para o modelo captar intempéries. Assim, optamos pela utilização do clássico ARIMA: ele possui as vantagens da regressão e média móveis, pois é uma combinação de ambos os modelos e utiliza exatamente o tipo de dados que possuímos, além de não ser oneroso do ponto de vista computacional.

3.5 Calibração do Modelo

Ao criar o modelo, é necessário segregar o conjunto de dados (*dataset*) em treino (calibração) e teste, para que seja possível avaliar a previsão adequadamente. Uma divisão não muito criteriosa pode inserir informações as quais deveriam ser exclusivas do teste na base de treino (calibração) - fenômeno conhecido como *data leakage* (BROWNLEE, 2016). Isso implica que utilizaremos informações futuras para prever a si mesmas, o que é ilógico e também impossível no cenário real, no qual nunca saberemos qual será a demanda real no futuro. A segregação adequada entre treino (calibração) e teste garante que parte dos dados disponíveis sejam invisíveis para o modelo criado. Logo, caso o modelo consiga prever adequadamente o conjunto de teste (ou validação), ele passa a ser um modelo confiável.

Dessa forma, o *dataset* de treino (calibração) do modelo foi definido com base no horário. Por exemplo, caso o objetivo seja realizar uma previsão a partir das 18h00, uma certa quantidade N - de amostras anteriores a esse horário - serão incluídas no treino (calibração). As demais podem ser usadas como teste, a depender da quantidade de intervalos a serem previstos. Torna-se necessário, então, definir o valor de N – ou seja, o tamanho do treino (calibração). Ver Figura 13.

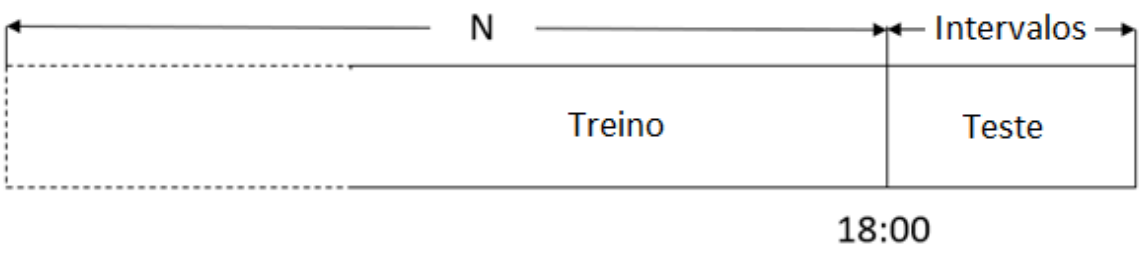

Figura 13: Parâmetros para o treinamento do modelo de previsão

A quantidade de observações utilizadas para realizar o fit do modelo (definição dos coeficientes internos do modelo baseado na base de treino calibração) foi estudada de maneira iterativa. Isto é, inicia-se a previsão com um treino (calibração) reduzido (50 observações) e aumenta-se progressivamente seu tamanho até que todas as amostras anteriores ao horário previsto estivessem nele contidas. Para cada iteração, foi computado o *Root Mean Squared Error (*RMSE) com o intuito de encontrar uma faixa de valores que minimize o erro.

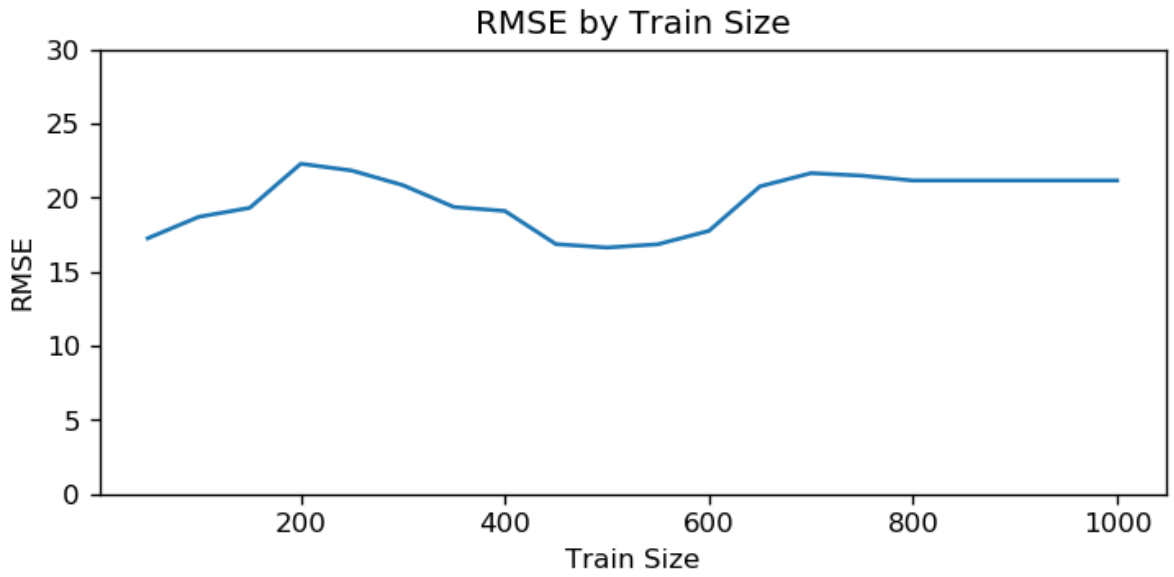

Figura 14: Radar na Av. Alcântara Machado - Erro computado na definição dos parâmetros de previsão (curva de aprendizado)

Como pode-se observar na Figura 14, um treino (calibração) com 50 observações possui um RMSE similar ao uso de toda a série temporal (estabilização da curva próximo de 800 observações). Esse procedimento foi repetido para vários radares em diversos horários e viu-se que em nenhum dos casos houve mudança significativa na métrica de erro. No gráfico da Figura 14 foi utilizado como exemplo o radar localizado na Av. Alcântara Machado (Centro/Bairro), com volume médio de 125 veículos a cada três minutos. Dessa forma, um RMSE de aproximadamente 20 significa um erro percentual abaixo de 20%. Para reduzir custos computacionais e, ao mesmo tempo, para disponibilizar uma quantidade satisfatória ao modelo ARIMA, optou-se por manter um tamanho fixo de 100 amostras para o treino (calibração).

3.6 Seleção de Parâmetros

Para selecionar os parâmetros do modelo ARIMA, foi utilizada a técnica proposta por Brownlee (2017), a qual está detalhada na Seção 2.4.6. Aqui, será apenas apresentado um exemplo de como foram selecionados os parâmetros para o mesmo radar citado anteriormente (localizado na Av. Alcântara Machado). Essa etapa antecede a previsão em si e é de suma importância para assegurar a qualidade do modelo.

É necessário que o modelo possua parâmetros adequados para a série temporal em questão. Uma seleção adequada garante resultados mais precisos, porém deve-se sempre considerar que modelos com maiores parâmetros de auto-regressão (p) ou médias móveis (q) são mais onerosos computacionalmente. Por isso, é sempre interessante tentar mantê-lo o mais simples possível, e estar atento onde um aumento na complexidade do modelo cause pequenos ganhos de precisão versus grande perda de performance computacional.

Ainda utilizando como exemplo os dados do mesmo radar da seção anterior, aqui será descrita, etapa por etapa, como é feita a seleção de parâmetros do modelo ARIMA. A primeira etapa é verificar se a série é estacionária ou não. Isso pode ser feito com o teste de Dickey-Fuller aumentado (ADF). Em Python, utilizamos a função “adfuller” contida na biblioteca “statsmodels”. Tal função recebe a série temporal e retorna o “valor p” a ela associado, que no exemplo em questão foi de 0.018676 (inferior ao nível de confiança de 0.05, significando que a série pode ser considerada estacionária).

Para validar o resultado do teste ADF, foi também feita uma análise visual das funções de autocorrelação para a série original e também com diferenciação de primeira e segunda ordem. Espera-se que a série original tenha uma curva ACF decrescente enquanto as outras possuam uma queda abrupta pois, segundo o teste ADF, elas já serão diferenciadas.

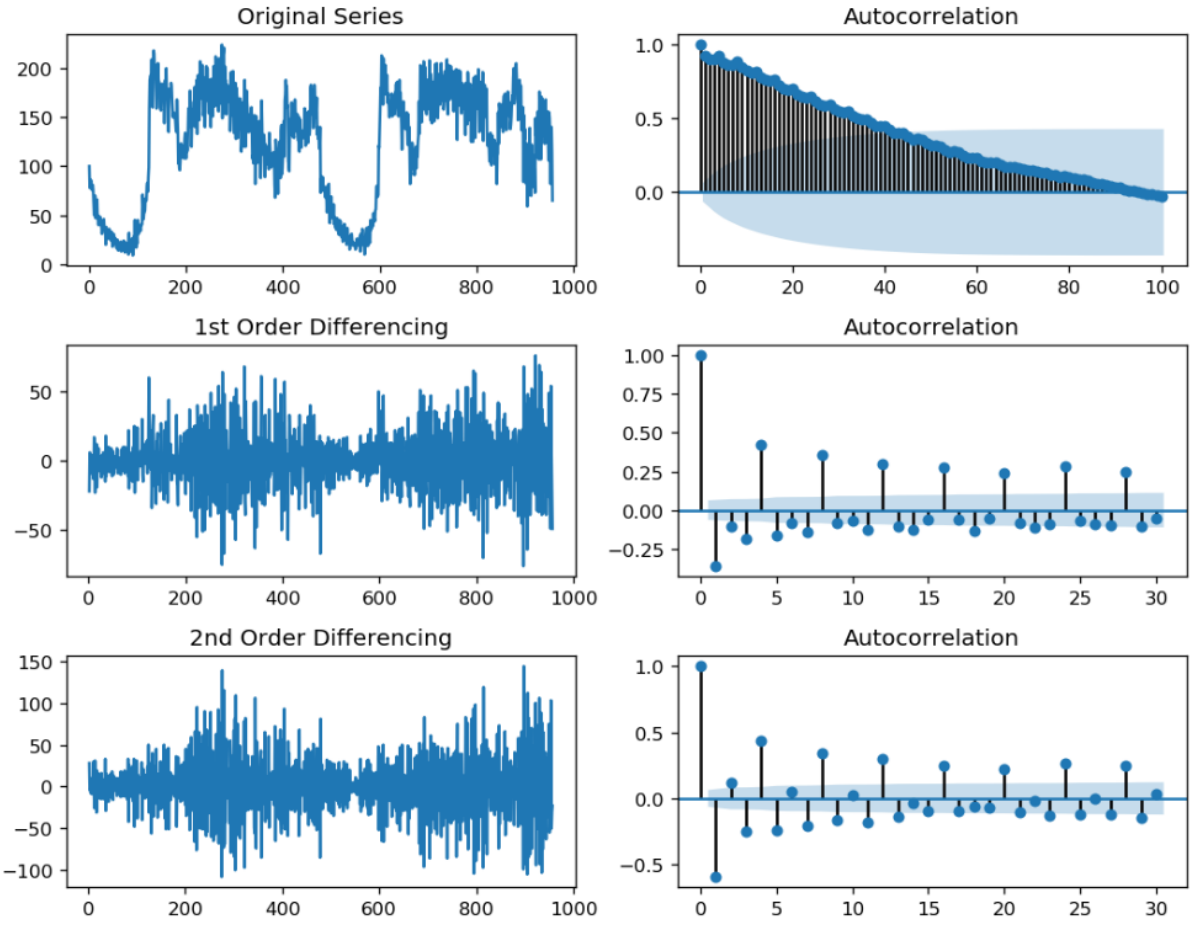

Figura 15: Análise visual das funções de autocorrelação

Como pode ser observado na Figura 15, o comportamento é realmente o esperado. A queda abrupta nas curvas de autocorrelação das séries diferenciadas indicam que não há relação alguma entre o valor presente e os passados, indicando que não há necessidade de diferenciar a série para torná-la estacionária, como havia sido indicado pelo teste de hipótese. Isso significa que o termo de diferenciação (d) será 0.

A etapa seguinte consiste em encontrar a ordem do modelo de auto-regressão. Para isso, será utilizada a função de autocorrelação parcial. Espera-se que ela evidencie quantos *lags* estão relacionados ao valor presente, logo, definindo até que ordem a auto-regressão realmente agregará informações ao modelo.

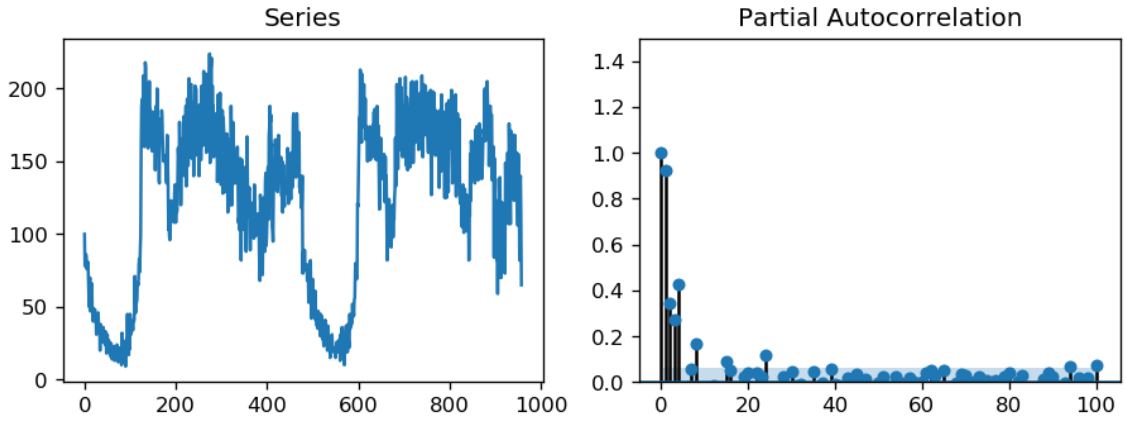

Figura 16: Autocorrelação parcial

O gráfico da Figura 16 mostra que os quatro primeiros valores possuem correlação acima do *threshold* mínimo (definido pela área em azul). O primeiro ponto – de valor 1 – indica que o valor presente é perfeitamente correlacionado consigo mesmo, enquanto os próximos quatro pontos indicam uma relação real com a série histórica. Para maximizar o ganho de informação, será utilizada uma ordem de regressão p = 4.

Por fim, define-se a ordem das médias móveis utilizando, novamente, a curva de autocorrelação. Como pode ser observado na Figura 15, a ACF decai lentamente, se mantendo acima do intervalo de confiança até aproximadamente 40 termos. Infelizmente, tal valor não é de grande auxílio pois nos dá um *threshold* muito alto, contendo inclusive valores extremamente incomuns (normalmente, médias móveis de até terceira ordem são utilizadas, não sendo comum o uso de ordens superiores no meio profissional). Por isso, optou-se por um modelo mais simples para então analisar qual o desempenho para um modelo ARIMA(4, 0, 1).

3.7 Previsão

Tendo a ordem do modelo definida, basta construir o modelo utilizando os dados de treino (calibração) selecionados e os parâmetros definidos. Para efeitos de comparação, foi também utilizado o modelo “Prophet”, desenvolvido pela equipe do Facebook. Tal modelo foi concebido com o intuito de realizar previsões de maneira rápida e simples, podendo ser utilizado mesmo por pessoas com pouca ou nenhuma experiência de previsão. Por se tratar de uma biblioteca nova, não há muita literatura acerca de seus resultados, porém ele será utilizado apenas como uma referência.

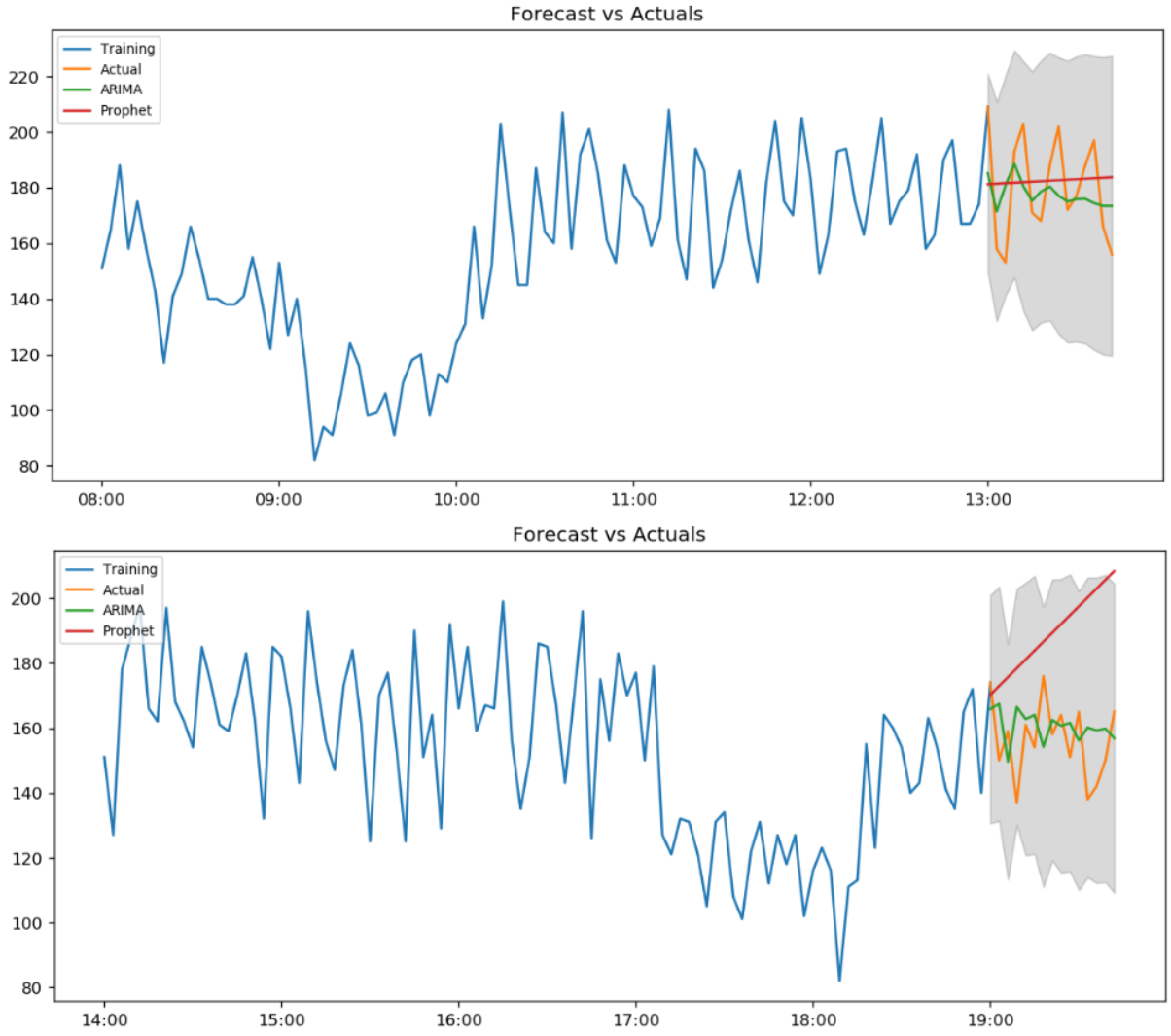

Figura 17: Previsão de 45 minutos com ARIMA e Prophet

A Figura 17 mostra os resultados da previsão realizada em dois horários distintos: às 13:00 e às 19:00. No primeiro horário, ambas as previsões do ARIMA (4, 0, 1) e Prophet ficaram muito semelhantes, e ambas relativamente próximas da série real (caso desconsideremos a flutuação existente no volume de tráfego). Já no segundo horário, o modelo ARIMA manteve-se próximo da série real enquanto o Prophet previu um padrão de crescimento na quantidade de veículos que não ocorreu.

4 RESULTADOS E CONCLUSÕES

4.1 Análise qualitativa das predições

Durante a construção dos modelos, foram realizadas plotagens para a verificação visual dos resultados. Tais plotagens nos deram uma ideia inicial da qualidade da previsão (apesar de não serem suficientes para assegurar sua qualidade).

Após gerar gráficos de diferentes radares em diferentes horários, a primeira impressão dos resultados é de que é possível gerar previsões com bons resultados. A qualidade da previsão é comprometida em momentos onde há variação muito abrupta do volume de tráfego. Nesses casos, o treino (calibração) reduzido (100 *lags*) torna a previsão impossível. Isso ocorre pois no período utilizado para construir o modelo (treino ou calibração) tal variação não ocorria.

Outro fato observado é que as previsões para radares com menor volume de tráfego são mais imprecisas, tendo em vista que a variação relativa nesses locais é mais crítica que a variação em pontos mais movimentados. Isso ocorre pois, por exemplo, num cenário onde o volume médio é de 20 veículos a cada 3 minutos, a variação de 2 veículos representa uma alteração de 10%. Caso o volume médio fosse de 2 veículos, a alteração seria de 100%. Isso torna a série histórica mais volátil e, portanto, mais difícil de se prever.

Para todos os casos foram geradas previsões num horizonte de 45 minutos. Isto é, foram feitas previsões até 15 intervalos no futuro, sendo cada uma delas relativa ao volume de tráfego previsto para aquele ponto num intervalo de três minutos. Abaixo, seguem algumas imagens com o intuito de exemplificar os *insights* iniciais que o grupo teve ao iniciar as predições.

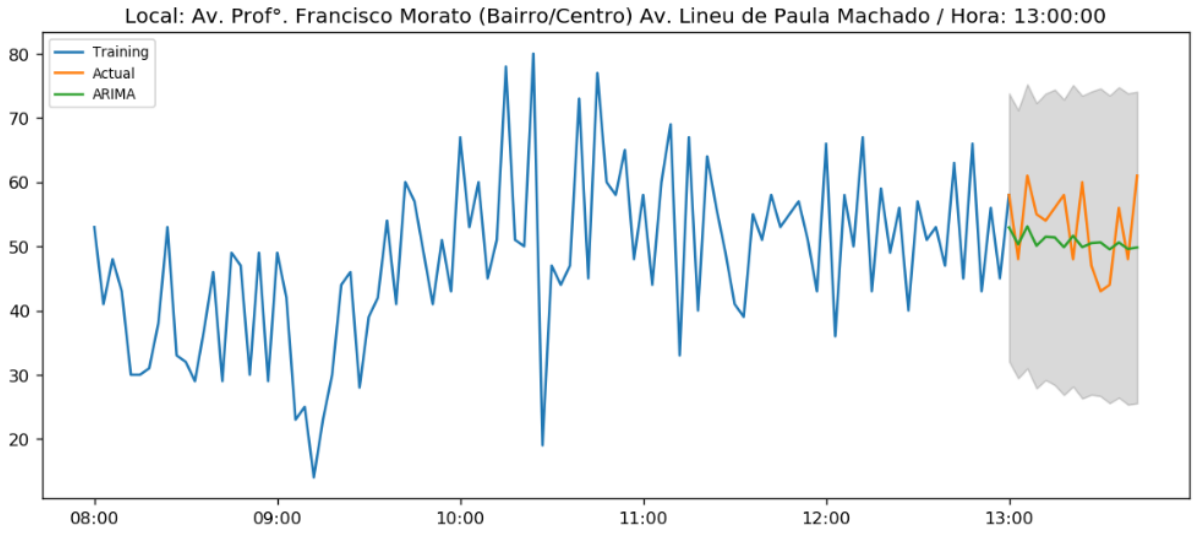

Figura 18: 1ª Previsão de 45 minutos

A Figura 18 indica um caso no qual o modelo apresenta um desempenho relativamente bom. A A área em cinza representa o intervalo de confiança do ARIMA. Apesar de o modelo não ser capaz de captar exatamente a flutuação real, ele apresentou valores próximos dos observados e também identificou uma leve tendência de queda, a qual realmente ocorre.

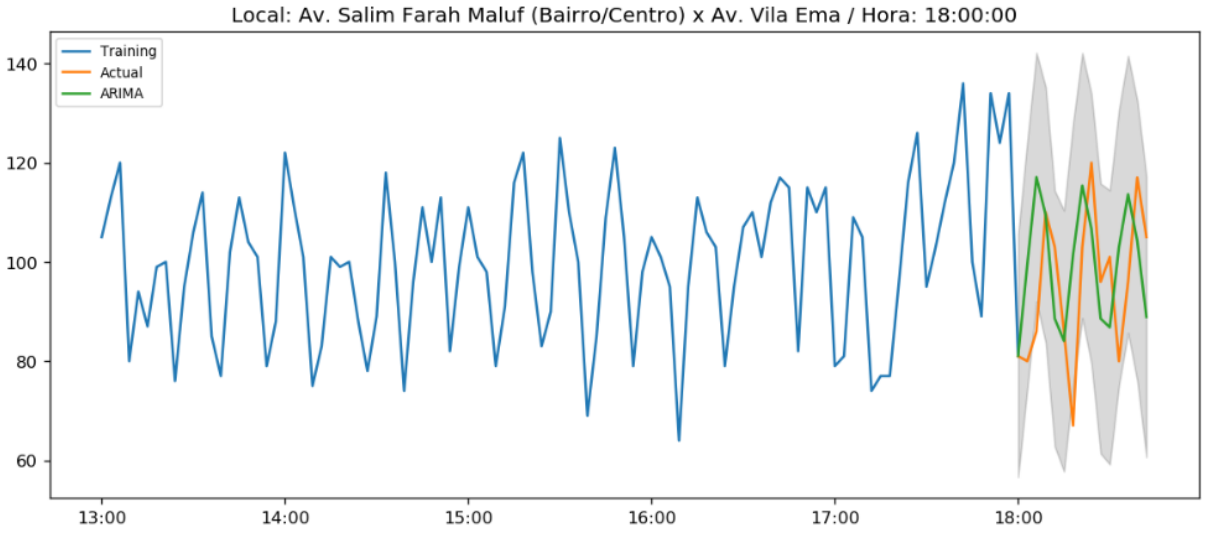

Figura 19: 2ª Previsão de 45 minutos

Em alguns casos, o modelo consegue inferir até mesmo a flutuação existente no tráfego com precisão aceitável. A Figura 19 mostra um cenário onde a previsão obteve um RMSE de 14.57, erro de aproximadamente 13% tendo em vista que o volume médio da via a cada três minutos é de cerca de 110 veículos. Até mesmo a previsão no horizonte mais longínquo (45 minutos adiante) obteve uma precisão razoável.

Apesar de muitos casos apresentarem bons resultados, há ainda cenários mais complexos nos quais o modelo é incapaz de prever corretamente a tendência da série. Por exemplo em vias nas quais há variação abrupta de tráfego no horário de pico, como representado na Figura 20. O que comprova que o método não é apropriado para esse tipo de variações. Muitos autores indicam que a precisão da previsão é para os casos típicos e não para os atípicos, quando ocorrem essas variações. Mesmo que fossem utilizados mais lags, possivelmente esse tipo de variação continuaria gerando grande imprecisão na previsão. E, nestes casos, vê-se também que as hipóteses necessárias para a aplicação do ARIMA não são satisfeitas.

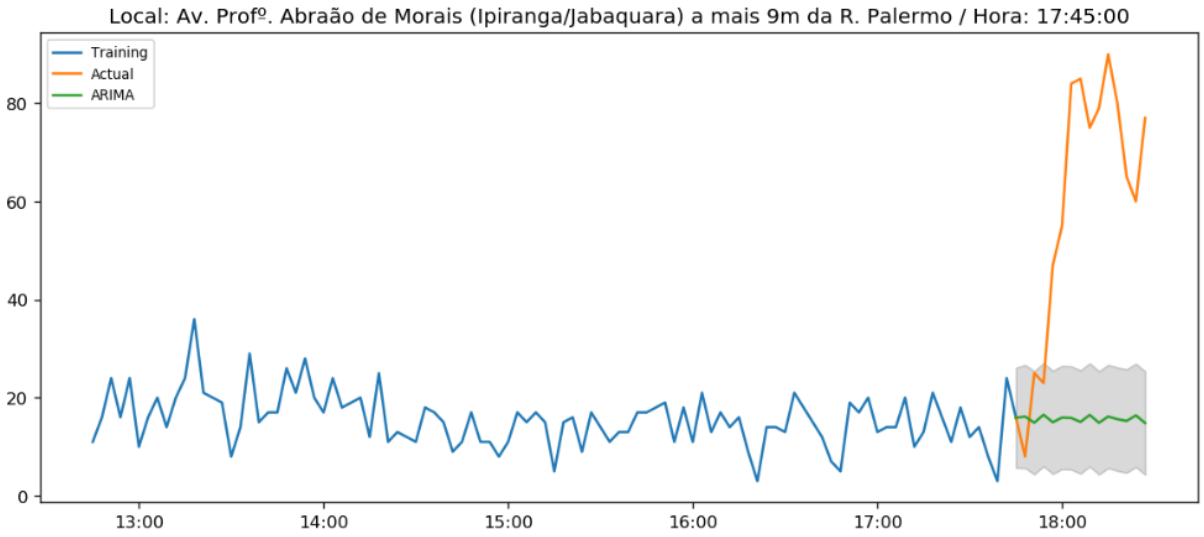

Figura 20: 3ª Previsão de 45 minutos

Apesar de nesses casos o resultado ser extremamente insatisfatório, uma análise mais cuidadosa mostra que o modelo passa a aderir melhor à tendência de crescimento à medida que o treino (calibração) passa a incorporar mais informações. Por exemplo, mantendo o mesmo radar da Av. Prof. Abraão Morais e realizando a previsão para as 18:00, quando parte do acréscimo de volume já passa a ser compreendido no treino (calibração), o resultado se torna aparentemente melhor – Figura 21. Mesmo com um desempenho ainda não muito bom, vemos que o ARIMA é capaz de se adequar rapidamente à medida que o treino (calibração) engloba o novo comportamento da série. Portanto, o método utilizado embora não "preveja" variações abruptas, consegue se auto-ajustar após alguns passos. Isto tem relação com os parâmetros "pdq" do ARIMA, com o método empregado e com o fato de não ocorrer

linearidade dos dados (premissa 1) no momento do início do pico. Portanto, seriam necessários aqui mais testes, por exemplo variando os parâmetros “pdq” do ARIMA.

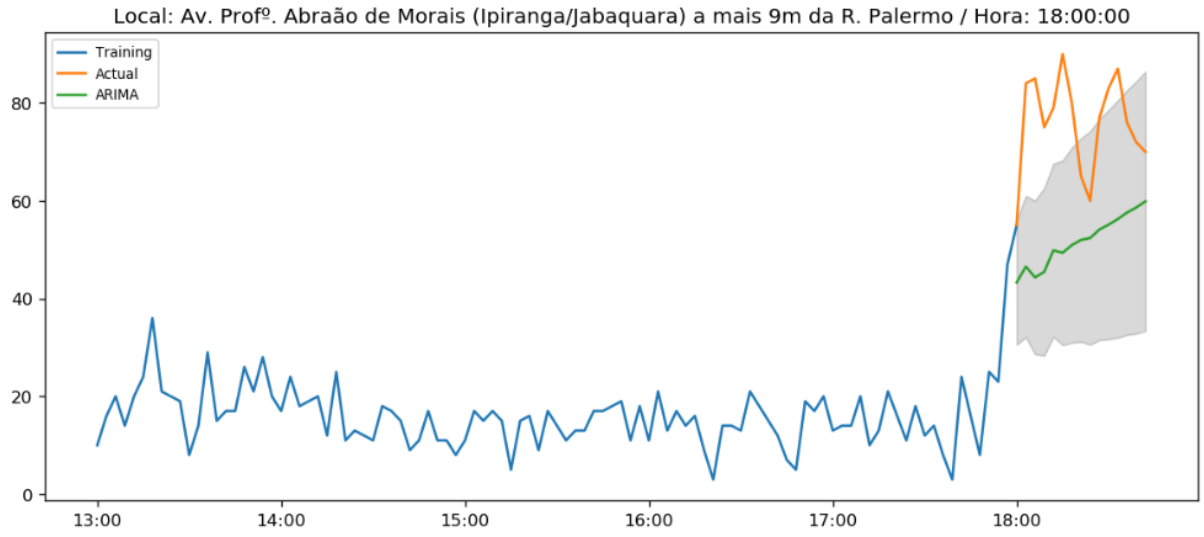

Figura 21: 4ª Previsão de 45 minutos

Os casos com pior desempenho, por fim, são os casos nos quais o volume de tráfego é extremamente baixo. Por exemplo, há radares nos quais há um fluxo de pouco mais de um veículo a cada três minutos. Isso significa que uma variação de três carros pode significar um acréscimo de 300% no volume, e tal volatilidade acaba por comprometer o desempenho do modelo. A Figura 22 mostra um desses casos, nos quais o erro fica próximo de 50%.

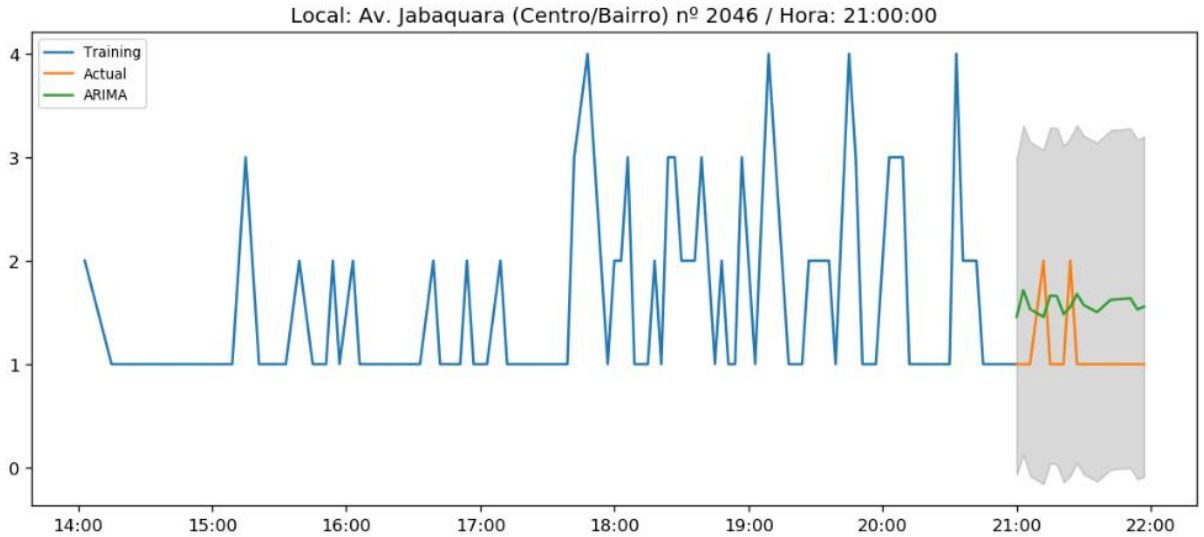

Figura 22: 5ª Previsão

4.2 Análise quantitativa das predições

Como a base de radares possui observações de 1280 pontos espalhados ao longo de toda a cidade de São Paulo, torna-se inviável a análise visual de todos eles. Além do mais, garantir que os resultados são aceitáveis em todos os radares não é suficiente: deve-se também verificar os resultados ao longo de diversas horas do dia. Para tanto, utilizou-se a seguinte técnica para computar o erro de previsão:

1. **Seleção de radares**: como visto anteriormente, a previsão é insatisfatória para pontos com volume de tráfego muito baixo. Além disso, é impossível o cálculo do erro percentual para casos onde há observações com valor nulo. Por isso, foram selecionados 171 radares os quais possuem volume médio acima de 20 veículos a cada três minutos. Anteriormente foram utilizados 55 radares para a validação espacial dos dados, aqui os 171 radares foram utilizados para determinar o erro de previsão, pois estes são os radares com maior fluxo de tráfego a cada 3 min.
2. **Diversos horários**: para verificar se a hora de previsão realmente afeta o seu desempenho, foram realizadas previsões em todas as horas fechadas, desde as 10:00 até as 22:00. Isso significa que foram realizadas 13 predições para cada um dos 171 radares selecionados.
3. **Modelo simplificado**: devido à grande quantidade de modelos a serem construídos e predições a serem realizadas, optou-se por simplificar o ARIMA utilizado nessa etapa. Durante as análises manuais de parâmetros, observou-se que grande parte das séries temporais resultavam em modelos com parâmetros (2, 0, 1). Por isso, optou-se por aplicá-lo em todos os pontos com o intuito de reduzir o custo computacional.
4. **Análise dos intervalos**: em cada caso, será previsto o volume do minuto +3 até o minuto +45. Espera-se que o erro aumente com o horizonte de previsão. Por isso, os resultados foram também segregados para cada um dos intervalos, podendo avaliar a evolução do erro.

Seguindo essas quatro etapas, foram executadas um total de 2.223 previsões e seus erros foram computados por hora, para cada um dos diferentes intervalos. O resultado obtido está representado na Tabela 2.

Tabela 2: Previsões e respectivos erros computados

	Horizonte de previsão														
Hora	3min	6min	9min	12min	15min	18min	21min	24min	27min	30min	33min	36min	39min	42min	45min
10:00	18%	18%	27%	19%	20%	21%	22%	22%	25%	29%	29%	32%	29%	30%	30%
11:00	19%	16%	18%	18%	19%	20%	20%	21%	23%	22%	24%	23%	24%	25%	25%
12:00	56%	55%	42%	52%	55%	71%	59%	68%	57%	55%	61%	61%	60%	56%	56%
13:00	20%	15%	20%	19%	22%	20%	20%	20%	23%	22%	24%	25%	22%	23%	23%
14:00	24%	21%	30%	21%	24%	20%	21%	26%	22%	22%	22%	21%	23%	23%	24%
15:00	25%	17%	22%	19%	20%	21%	21%	22%	23%	24%	23%	24%	24%	26%	26%
16:00	19%	20%	23%	21%	24%	22%	24%	23%	24%	24%	27%	27%	28%	28%	38%
17:00	42%	24%	26%	26%	35%	27%	26%	27%	27%	27%	29%	28%	27%	29%	27%
18:00	37%	17%	23%	21%	23%	26%	31%	30%	29%	31%	33%	35%	36%	35%	34%
19:00	19%	16%	20%	18%	20%	20%	22%	21%	24%	24%	25%	23%	24%	24%	30%
20:00	21%	17%	22%	23%	23%	22%	25%	28%	34%	28%	26%	30%	27%	29%	36%
21:00	25%	25%	23%	23%	25%	24%	26%	28%	31%	29%	31%	31%	34%	31%	35%
22:00	19%	15%	23%	16%	19%	18%	21%	19%	23%	21%	24%	27%	31%	25%	27%

Pela Tabela 2 vê-se que os erros para os horizontes mais próximos são, no geral, menores. O que de fato é esperado. No entanto, há valores que fogem do padrão observado como por exemplo os erros obtidos ao meio dia, que ao contrário dos demais horários, basicamente todos os valores estão acima de 50%. Inicialmente, pensou-se que as séries apresentariam comportamento atípico nesse horário (por exemplo, pico na quantidade de veículos) e por isso o erro é tão elevado. No entanto, outros horários também considerados “de pico” possuem métricas próximas dos demais, invalidando a hipótese inicial. A segunda hipótese encontrada é de que algumas das séries estudadas podem possuir volume muito baixo próximo às 12h00, e como o erro calculado é percentual, um divisor baixo gera erros extremamente elevados. Aqui haveria a necessidade de uma verificação mais detalhada.

Para tentar validar alguma das hipóteses levantadas, foi construído um histograma do volume de tráfego para os 171 radares estudados. O esperado é que a distribuição seja normal, com uma grande quantidade de observações próximas da média (aprox 20). No entanto, ao gerar o histograma, notou-se que há uma grande concentração de valores próximos de 0 (vide figura 23). Isso dá suporte a segunda hipótese de que tais pontos com baixíssimo volume estão aumentando a métrica de erro percentual pois a divisão por um número próximo de zero gera números muito elevados.

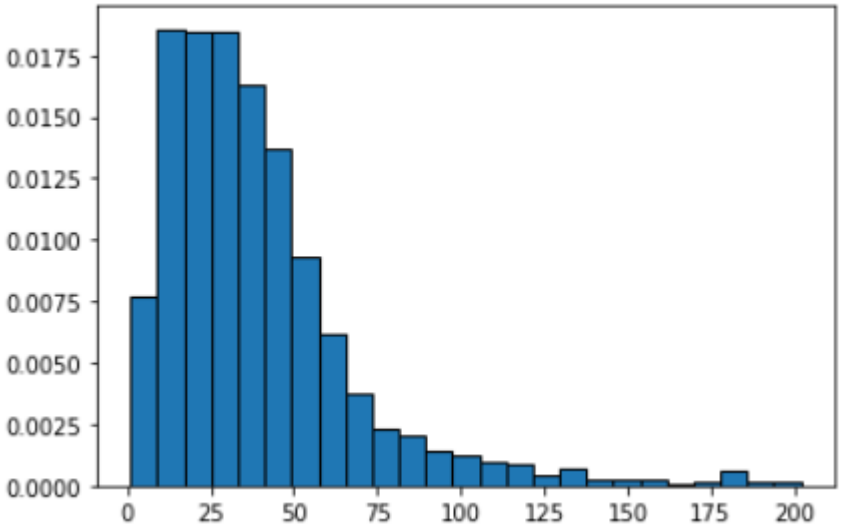

Figura 23: Histograma de Volume de Tráfego dos 171 radares analisados entre 12:00 e 12:45

Para verificar se o erro realmente aumenta com o horizonte de previsão, como havia sido inferido, foi calculado o erro médio de todos os horários do dia para cada um deles. A Figura 23 contém os valores calculados, e a partir dela pode-se ver que o erro realmente aumenta. Entretanto, ele varia cerca de dez pontos percentuais entre o horizonte de três e quarenta e cinco minutos. Tal variação é pequena, tendo em vista que estamos prevendo quinze etapas adiante do presente. Em Kumar (2015) são alcançadas predições para 5min até 1h com boa precisão.

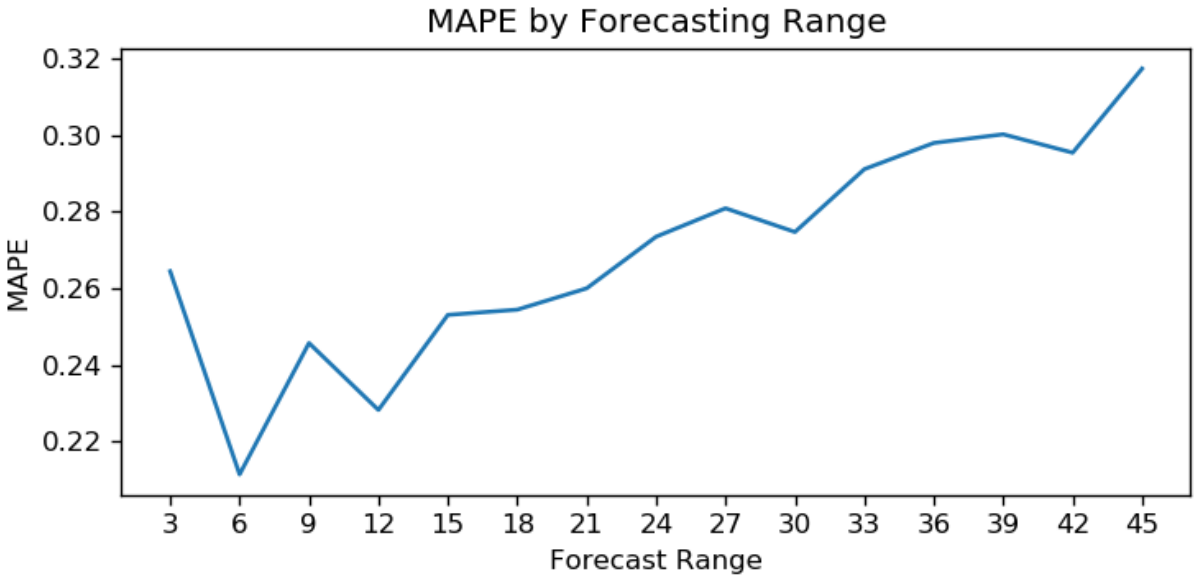

Figura 24: Mean Average Percentual Error - Erro médio da previsão em todos os horários

Por fim, foi verificado se o horário previsto realmente é um fator determinante para o desempenho. De todos os horários analisados, o único com comportamento longe do padrão (erro percentual entre 20% e 30%) é o horário das 12:00. Como discutido anteriormente, acredita-se que uma pequena parcela dos dados possua comportamento anômalo nessa parte do dia e tenham sido responsáveis por tornar a métrica de erro maior do que ela realmente seja. Dessa maneira, vê-se que o horário do dia pode afetar o desempenho da previsão, tendo em vista que os valores são mais altos às 18:00 e 21:00, horários onde há a movimentação de trabalhadores retornando para suas residências e estudantes de faculdades noturnas se locomovendo, respectivamente. Ver Figura 25.

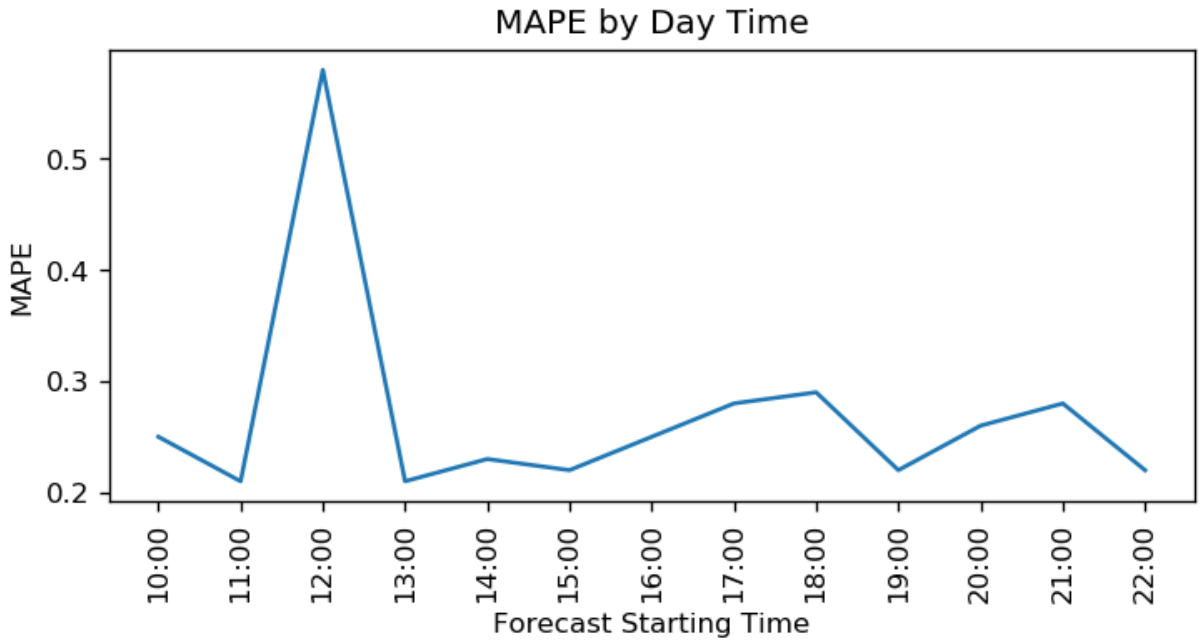

Figura 25: Erro médio da previsão entre 10h e 22h

4.3 Considerações Finais e Recomendações

Tendo em vista todos os resultados apresentados e discutidos, as seguintes conclusões foram inferidas:

- O modelo ARIMA tem potencial para ser utilizado para a previsão de tráfego no curto prazo (3 a 5 min) e talvez até mesmo no médio prazo (45 min até 1h). Erros da ordem de 20% são aceitáveis e a previsão pode contribuir para o planejamento imediato do tráfego.

- Infelizmente o modelo possui performance inaceitável em casos com volume de tráfego extremamente baixo. No entanto, tais pontos são menos críticos, necessitando de pouco ou nenhum planejamento. Isso significa que a ausência de previsão para tais casos é pouco crítica.
- Tanto o horário da previsão quanto o seu horizonte são fatores que afetam o desempenho do modelo, porém menos que o esperado.
- É difícil a implementação dessa solução pois ela necessita de alimentação de dados em tempo próximo ao real. Tendo em vista que ainda não há uma base com dados de radares consolidadas e atualizada em tempo próximo ao real (como um banco de dados no SQL, por exemplo), pode ainda não ser algo viável.
- O tamanho do treino (calibração) não se mostrou um fator determinante para o desempenho do modelo. Acredita-se que isso ocorreu pois temos apenas dois dias de dados fornecidos, sendo impossível incorporar a sazonalidade no modelo com apenas dois ciclos.

Além do mais, sabe-se que o modelo ARIMA é, atualmente, considerado relativamente simples e muitas melhorias podem ser implementadas. Para dar continuidade ao presente trabalho de formatura recomenda-se:

- Buscar um histórico maior de dados e tentar utilizar técnicas com sazonalidade, como por exemplo o SARIMA.
- Buscar informações externas, que possam ser inseridas num modelo causal que combine dados históricos e informações auxiliares (como pluviometria, quantidade de acidentes, feriados, etc).
- Estudar a correlação entre os fluxos nos diversos radares e verificar a possibilidade de fazer regressões múltiplas com o intuito de obter resultados mais precisos e compreender melhor o funcionamento da malha viária.

REFERÊNCIAS BIBLIOGRÁFICAS

ADHIKARI, R., AGRAWAL, R. K. (2009) - **An Introductory Study on Time Series Modeling and Forecasting.** LAP Lambert Academic Publishing, Germany

BARROS, O. M. **Caracterização das condições de tráfego em tempo próximo ao real para o uso em sistemas de previsão de tráfego em cidades de grande porte**. 2019. 116 f. Texto para Qualificação de Mestrado – Departamento de Transportes, Escola Politécnica, Universidade de São Paulo, São Paulo, 2019.

BOX, G. E. P., JENKINS, G., **Time Series Analysis: Forecasting and Control.** 1976. Holden-Day.

BROWNLEE, J. **Data Leakage in Machine Learning**. 2016. Disponível em: < https://machinelearningmastery.com/data-leakage-machine-learning>. Acesso em: 15 de Jan de 2020.

BROWNLEE, J. **Introduction to Time Series Forecasting with Python: How to Prepare Data and Develop Models to Predict the Future**. 2017. Machine Learning Mastery.

CASTRO-NETO M.; JEONG Y.S.; KEEJEONG M.; LEE D.H. (2009). **Online-SVR for short-term traffic flow prediction under typical and atypical traffic conditions.**

FULLER, W. A. **Introduction to Statistical Time Series**. 1976. New York: John Wiley and Sons.

GUTIÉRREZ, J. L. **Monitoramento da Instrumentação da Barragem de Corumbá-I por Redes Neurais e Modelos de Box & Jenkins**. 2003

GREENSHIELDS, B. D. **A Study of Traffic Capacity**. Highway Research Board, 14, 448-447. 1935.

HYNDMAN, R. J.; ATHANASOPOLUS, G. **Forecasting: principles and practice**. 2013. Disponível em: <https://www.otexts.org/fpp/>. Acesso em: 21 de Ago de 2019.

JAIN, S. (2017)**. Time Series Analysis for Financial Data IV – ARMA Models**. Disponível em<https://medium.com/auquan/time-series-analysis-for-finance-arma-models-21695e14c999> acesso no dia 21/07/2019.

KUMAR, S.V., VANAJAKSHI, L. **Short-term traffic flow prediction using seasonal ARIMA model with limited input data**. 2015. *Eur. Transp. Res. Rev.* **7,** 21 doi:10.1007/s12544-015-0170-8

KRUK, R. (2002) **A general spatial ARMA model: Theory and application.** ERSA (European Regional Science Association) Conference.

MILLS, T. C. (1991). **Time Series Techniques for Economists**. Cambridge University Press.

MORANTZ, B. **Time Series Forecasting Real World Data: Auto Regression Using a Neural Network Forecaster with Weighted Windows**. 2004. IEEE Computational Intelligence Society.

MOREIRA, D.A., **Administração da Produção e Operações**. 2.ed.São Paulo: Cengage, Learning, 2008

MORETTIN, P. A.; TOLOI, C. M. C. **Análise de séries temporais**. 2 ed. São Paulo: Edgard Blücher, 2006.

NAU, R. **Statistical Forecasting: notes on regression and time series analysis**. 2019. Fuqua School of Business, Duke University.

PAVLYUK, D. (2016). **Short-Term Traffic Forecasting Using Multivariate Autoregressive Models**. 16thConference on Reliability and Statistics in Transportation and Communication.

ROMER, R., HECKARD, R., FRICKS, J. (2019). **Stat 510**. The Pennsylvania State University, Eberly College of Science.

SHIKIDA, Pery Francisco Assis e MARGARIDO, Mario Antonio. **Uma análise econométrica de sazonalidade dos preços da cana-de-açúcar, estado do Paraná, 2011-2007**. In: Informações Econômicas, SP, v.39, n.2, fev. 2009. Disponível em: <ftp://ftp.sp.gov.br/ftpiea/publicacoes/IE/2009/tec7-0209.pdf>. Acesso em: 25 de agosto de 2012.

SILVA, L. (2017). **Análise da Aplicação do Modelo ARIMA: Estudo em uma Instituição de Ensino Superior**. Santa Maria – RS.

SILVA, P. C. M. (1994). **Teoria do fluxo de tráfego.** Universidade de Brasília.

SLACK, N.; CHAMBERS, S.;JOHNSTON, R. **Administração da Produção**.3 ed. São Paulo: Atlas, 2009.

VALIPOUR, M. BANIHABIB, M. E., BEHBAHNI, S. M. R., (2013). **Comparison of the ARMA, ARIMA, and the autoregressive artificial neural network models in forecasting the monthly inflow of Dez dam reservoir**. Elsevier, Journal of Hidrology.

WOOLDRIDGE, Jeffrey M. **Introductory Econometrics: a Modern Approach**. 2000, South-Western College Publishing, a division of Thomson Learning. ISBN 0-538-85013-2.