Universidade Federal de Campina Grande
Centro de Engenharia Elétrica e Informática
Coordenação de Pós-Graduação em Ciência da Computação

Previsão de horários dos ônibus do sistema de transporte público coletivo de Campina Grande

Matheus de Araújo Maciel

Dissertação submetida à Coordenação do Curso de Pós-Graduação em Ciência da Computação da Universidade Federal de Campina Grande - Campus I como parte dos requisitos necessários para obtenção do grau de Mestre em Ciência da Computação.

Área de Concentração: Ciência da Computação
Linha de Pesquisa: Sistemas de Computação

Nazareno Ferreira de Andrade
(Orientador)

Campina Grande, Paraíba, Brasil

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA CENTRAL DA UFCG

M152p Maciel, Matheus de Araújo.

Previsão de horários dos ônibus do sistema de transporte público coletivo de Campina Grande / Matheus de Araújo Maciel. – Campina Grande, 2016.

46 f. : il. color.

Dissertação (Mestrado em Ciência da Computação) – Universidade Federal de Campina Grande, Centro de Engenharia Elétrica e Informática, 2016.

"Orientação: Prof. Dr. Nazareno Ferreira de Andrade.

Referências.

1. Inteligência Artificial. 2. Aprendizado de Máquina. 3. Horários dos Ônibus – Previsão. 4. Sistemas de Transporte Inteligentes. I. Andrade, Nazareno Ferreira de. II. Titulo.

CDU 004.8(043)

REVISÃO DE HORÁRIOS DOS ÔNIBUS DO SISTEMA DE TRANSPORTE PÚBLICO COLETIVO DA CIDADE DE CAMPINA GRANDE"

MATHEUS DE ARAÚJO MACIEL

DISSERTAÇÃO APROVADA EM 09/09/2016

NAZARENO FERREIRA DE ANDRADE, D.Sc, UFCG
Orientador(a)

JOÃO ARTHUR BRUNET MONTEIRO, Dr., UFCG
Examinador(a)

JOSÉ ANTONIO FERNANDES DE MACÊDO, Dr., UFC
Examinador(a)

CAMPINA GRANDE - PB

Resumo

A previsibilidade dos serviços de transporte público é um aspecto central para a melhoria da experiência de seus usuários. Contudo, por funcionar dentro de um ambiente estocástico, essa previsibilidade é tipicamente prejudicada. Neste trabalho investigamos a possibilidade de tornar um sistema de transporte público mais previsível através do uso das informações históricas em um contexto onde não há disponível tecnologia de localização tempo real dos veículos ou informação atualizada sobre a operação do serviço. Embora GPS e outras tecnologias de *Automatic vehicle location* (AVL) em tempo real existam, muitos municípios brasileiros não as têm disponíveis. Considerando essa situação, utilizamos dados históricos de operação do sistema de ônibus da cidade de Campina Grande para avaliar o desempenho de quatro algoritmos de regressão na tarefa de prever no início do dia como os horários programados para os ônibus serão cumpridos. Os resultados apontam que embora a falta de informação em tempo real prejudique a capacidade preditiva dos algoritmos em determinadas situações, utilizá-los torna possível a previsão dos horários de saída reais dos ônibus com erro mediano de 28 segundos, e a previsão dos horários de fim de viagem com erro de mediano de -167 segundos.

Abstract

Predictability of public transport services is essential to improving its user experience. However, by working within a stochastic environment, predictability is typically impaired. In this work, we investigate the possibility of making a more predictable public transport system through the use of historical information, in a context where there is no available real-time vehicle location technology or updated information on the operation of the system. While *GPS* and other real-time *Automatic Vehicle Location* technologies (AVL) exists, many Brazilian cities do not have them available. Aware of this situation, we used data from the Campina Grande city bus system to evaluate the performance of four regression algorithms on the task of predicting, early in the day, how buses scheduled times will be fulfilled. Results show, although the lack of real time information may harm algorithms predictive ability in certain situations, using them makes it possible to forecast actual buses departure times with a median error of 28 seconds and buses arrival time with a median error of -167 seconds.

Agradecimentos

Primeiramente quero agradecer aos meus familiares pelo apoio incondicional. Aos meus amigos que palavras de motivação e encorajamento. Ao meu orientador Nazareno Andrade pelos conselhos e principalmente por não ter desistido de mim. Aos companheiros de laboratório pelas discussões bastante relevantes.

Também quero agradecer a Superintendência de Trânsito e Transportes Públicos representada na pessoa de Hélder Carlos por disponibilizar os dados utilizados nesse estudo e todo seu conhecimento da área.

Conteúdo

Lista de Símbolos

STTP - *Superintendência de Trânsito e Transportes Públicos*

ITS - *Intelligent Transportation System*

AVL - *Automatic Vehicle Location*

GPS - *Global Positioning System*

kNN - *k-Nearest Neighbors*

ANN - *Artificial Neural Networks*

SVM - *Support Vector Machine*

RF - *Random Forest*

RMSE - *Root Mean Squared Error*

ECDF - *Empirical Cumulative Distribution Function*

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

A pontualidade dos ônibus é um fator importante tanto para nível de satisfação do usuário quanto para a entidade gestora do serviço [8]. Ainda assim, pode ser complicado garantir um nível de pontualidade aceitável devido a aleatoriedade do sistema em que o serviço está inserido [22; 3]. Por esse motivo, o planejamento e a tomada de decisão tem um papel crucial tanto na gestão do sistema quanto na utilização dele. Nesse momento *Intelligent transportation systems* (ITS) usam TI para gerar informação útil para o cidadão e para o gestor. Alguns exemplos de ITS bastante utilizados são sistemas de navegação encontrados em carros, lombadas eletrônicas e localização automática de veículos usando sensores GPS.

No contexto do transporte público, especificamente no cenário dos ônibus, um dos problemas mais conhecidos é o do cumprimento dos horários estabelecidos. Por ser um problema que em muitos casos é causado por fatores que não podem ser controlados, nem sempre é possível evitar que ele aconteça. Vários estudos aplicaram aprendizado de máquina e estatística com o objetivo de prever a ocorrência desses eventos [2; 6; 18; 9]. Tal previsão permite que os interessados tenham essa informação com antecedência e possam decidir como contornar a situação.

1.1 Motivação

No caso de Campina Grande, os ônibus não possuem nenhum sistema de localização automática. A execução das viagens é coletada em pontos de fiscalização no início e no final da viagem. Além disso, toda a informação coletada durante o dia só é disponibilizada no

dia seguinte. Por isso, só é possível saber o horário em que o ônibus passou em um desses pontos até o dia anterior. Como todas as abordagens de previsão consolidadas utilizam basicamente dados de localização em tempo real, não existem evidências de que os métodos mais usados na previsão de horário dos ônibus conseguem resultados relevantes em cidades como Campina Grande.

Uma consideração central em nosso trabalho é a indisponibilidade de informação em tempo real ou de informação histórica detalhada sobre o posicionamento dos ônibus na cidade. Embora existam tecnologias para tanto – a exemplo do GPS – a administração de várias cidades brasileiras não possui acesso a uma infraestrutura implantada com GPS ou aos dados de controle da frota usada em suas cidades. Em maio de 2016, João Pessoa, Natal e Campina Grande ainda figuram como exemplos de municípios com essa limitação na infraestrutura de monitoramento. Além disso, na ocasião da disponibilidade dessa informação, os modelos baseados na informação histórica e menos detalhada que utilizamos neste estudo podem ser usados em complemento à informação de localização dos veículos em tempo real.

Neste contexto, a informação que consideramos disponível a partir do sistema de transporte público consiste apenas dos horários de saída e chegada executados pelos ônibus até a meia noite do dia anterior ao momento da previsão, e da informação de quantos passageiros foram transportados em cada viagem. Essas informações tipicamente estão disponíveis nas administrações públicas por necessitarem de um esforço de implantação menor, dependerem de tecnologias desenvolvidas relativamente simples, e por serem centrais à distribuição da renda do sistema entre as empresas concessionárias do serviço de transporte público. Embora em alguns contextos a informação do horário de início e fim das viagens de cada ônibus possa estar disponível em tempo real, consideramos aqui um cenário de pior caso e baseado na realidade de Campina Grande - PB no momento da condução deste trabalho.

1.2 Objetivos

O objetivo desse trabalho é analisar a qualidade das previsões do horário de início e fim de todas viagens de um determinado dia usando quatro métodos de predição comprovadamente eficientes em diversos domínios: *k-Nearest Neighbors* (k-NN), *Artificial Neural Networks* (ANN), *Support Vector Machines* (SVM) e *Random Forests* (RF). As previsões foram basea-

das unicamente em dados históricos. Os algoritmos foram avaliados com base nos seguintes fatores: quantidade de dias usados no treino, tamanho do histórico de uma viagem, dia da semana e mês do ano.

Especificamente, as tarefas de previsão que estudamos consistem dos dados: (i) o histórico de viagens executadas até a meia noite do dia anterior, (ii) informações sobre o dia atual tais como mês, dia da semana e clima, e (iii) uma viagem programada para o dia atual, com o objetivo de prever: (a) o horário de início da viagem tal qual ela será realizada e (b) o horário de seu fim. De posse dessas duas informações é possível interpolar previsões para em que momento o ônibus realizando a viagem estará em cada ponto da rota, contudo, essa etapa de interpolação não foi focada neste trabalho. Por fim, a medida de erro que consideramos que o modelo de previsão deve minimizar é a raiz quadrada do erro quadrático médio (RMSE).

1.3 Abordagem de pesquisa

No contexto de Campina Grande, procuramos avaliar o desempenho de aplicar algoritmos de regressão em dados históricos para a previsão do horário de início e fim das viagens de um dia, considerando que estes dados não possuem informação de localização em tempo real. O primeiro passo foi parear as viagens executadas com as viagens programadas de forma que pudéssemos utilizar o horário programado com vínculo entre viagens executadas em dias diferentes. Após o pareamento ocorreu uma etapa de pré-processamento para a criação dos atributos de cada viagem.

Para avaliar os algoritmos, selecionamos aleatoriamente 12 dias que seriam usados como conjunto de testes. Para cada dia selecionado foi criada uma combinação dos fatores listados na Tabela 1.1. Para cada uma dessas combinações foi treinado e testado um modelo de previsão. Cada modelo prevê o horário final e inicial de todas as viagens do dia, e a partir destas previsões são calculados os erros.

Esse formato permite que o desempenho dos algoritmos seja comparado em várias dimensões como a quantidade de dias usados no conjunto de treino, quantidade de dias usados no histórico recente, mês do ano, dia da semana e hora do dia. Dessa forma, podemos decidir qual modelo se encaixa melhor em determinadas situações e qual algoritmo pode ser usado de maneira geral.

Algoritmo	*Dias no Treino*	*Dias no histórico*
kNN	7	1
ANN	15	2
SVM	45	3
RF		

Tabela 1.1: Níveis de cada fator utilizado no experimento.

1.4 Resultados

Foi verificado que tanto para o início quanto para o fim da viagem a mediana dos erros foi de aproximadamente 28 e -167 segundos respectivamente. As análises indicam que esses resultados são afetados por fatores como quantidade de dias no conjunto de treino, quantidade de dias no histórico recente, mês do ano, dia da semana, hora do dia e linha.

Em geral, os modelos criados com *Support Vector Regression* obtiveram os melhores resultados com mediana de aproximadamente 10 segundos para o início da viagem e -13 segundos para o fim. Não foi possível definir evidentemente o melhor valor para a quantidade de dias usados no conjunto de treino, entretanto, usar 7 dias se mostrou uma opção promissora. O mesmo não pode ser afirmado para a quantidade de dias usados no histórico recente dado que nenhuma das opções se destacou.

Mesmo que a utilização de *Support Vector Regression* tenha se destacado, existem combinações de fatores em que os outros algoritmos conseguem obter resultados competitivos. Alguns exemplos são: *Random Forest* com 45 dias no treino e 2 dias no histórico recente, *kNN* como 7 dias no treino e 3 dias no histórico recente e *kNN* com 15 dias no treino e 3 dias no histórico recente.

As análises mostraram que o fator temporal também pode influenciar no desempenho dos modelos criados. Os meses de julho se destacou como melhor mês para a previsão usando *SVR* enquanto maio obteve os piores resultados, mostrando que pode ser necessário utilizar modelos diferentes em diferentes momentos do ano. A qualidade das previsões também muda no decorrer da semana onde os piores resultados foram obtidos na segunda-feira e o

melhores na quarta e quinta-feira. O mesmo comportamento dos meses do ano e dos dias da semana foi verificada nas horas do dia, onde os horários de início e fim do expediente obtiveram os resultados mais inconsistentes.

1.5 Estrutura do texto

O restante deste documento está organizado da seguinte forma. O Capítulo 2 descreve o funcionamento do sistema de transporte público coletivo de Campina Grande, utilizado como contexto da pesquisa. No Capítulo 3 comentamos estudos já realizados na previsão de horários de chegada dos ônibus como este estudo se relaciona com eles.

No Capítulo 4, descrevemos os dados e o processo de transformação dos dados brutos nos dados utilizados na execução do experimento. Detalhamos os resultados obtidos e suas implicações no Capítulo 5.

Finalmente, o Capítulo 6 apresenta as conclusões finais do trabalho e alguns direcionamentos para trabalhos futuros.

Capítulo 2

Contexto

Neste capítulo descrevemos como funciona o transporte público da cidade de Campina Grande. Além disso, comentamos sobre detalhes que influenciam no experimento executa neste trabalho.

2.1 Transporte público em Campina Grande

Na cidade de Campina Grande o transporte público é composto por três categorias: transporte coletivo, taxi e moto-taxi que é gerenciado pela Superintendência de Trânsito e Transportes Públicos – STTP [1]. O transporte coletivo da cidade é operado exclusivamente por ônibus e sua operação é delegada a empresas escolhidas por meio de licitação pública. Antes de descrever como o sistema funciona é importante alertar que na maioria das cidades do Brasil o que é chamado de rota em Campina Grande é conhecido com linha.

O serviço é realizado através de 12 linhas, com 49 rotas urbanas e 13 rotas distritais, responsáveis pelo deslocamento de aproximadamente 2.800.000 passageiros por mês e com uma frota cadastradas de 223 veículos.

Cada rota é representada por um código alfanumérico de 4 dígitos, por exemplo "0404". Todas as rotas que operam nas mesmas zonas são agrupadas em linhas que são identificadas por um nome e uma cor. Na maioria dos casos o nome da linha é referente a sua cor como pode ser visto na Figura 2.1 1. Em Campina Grande, todas as rotas são circulares então o comportamento usual das rotas é que o ponto inicial e final de uma viagem sejam o mesmo

[1] http://sttpcg.com.br/

ponto. Existem casos em que isso não acontece, porém, esses casos só ocorrem sob demanda da STTP.

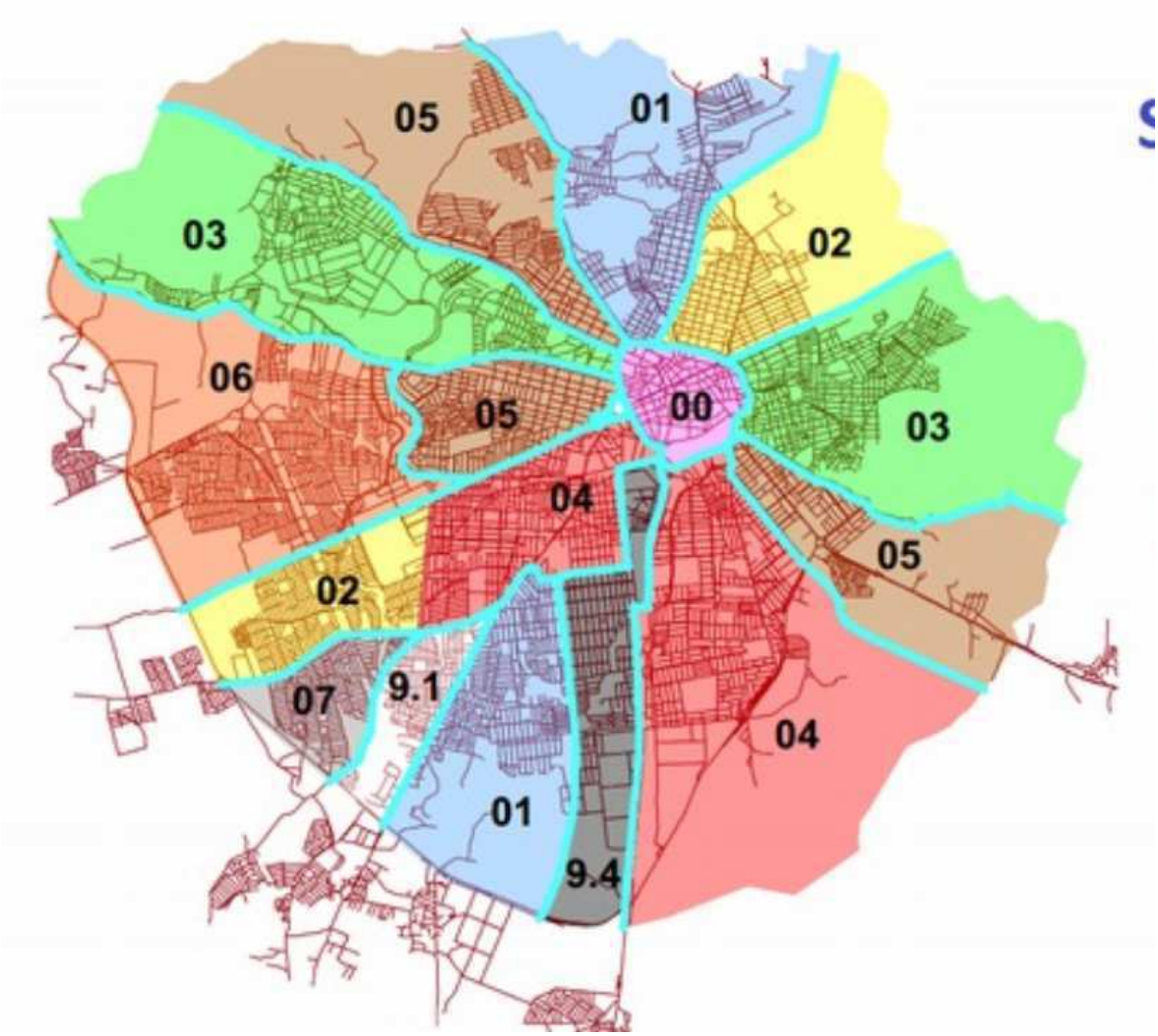


Figura 2.1: Mapa de linhas do transporte coletivo de Campina Grande.

2.1.1 Tabela de horários

Além de definir o percurso que o ônibus deve cumprir a rota também define o itinerário que ele deve cumprir. A demanda dos ônibus de uma rota é especificada pela STTP com a quantidade de ônibus operando e o tempo entre viagens, por exemplo, a rota *x* precisa de 5 veículos operando no período de 7h30 às 10h com um intervalo de 15 minutos entre as viagens. Para atender a demanda estipulada a empresa responsável pela rota define o itinerário dos ônibus dessa rota. O resultado final desse processo é uma tabela que define

além dos horários programados, o tipo de dia, quantidade de viagens previstas para o dia, duração média de uma viagem, quantidade mínima de carros necessários e quilometragem de uma viagem completa como poder ser visto na Figura 2.2. A partir dessa tabela também é possível saber a quantidade de carros que estão operando em um determinado momento do dia usando a coluna *Carro*.

Pela tabela podemos notar que em um dia considerado *dia útil* a primeira viagem da rota *263B* possui 5 horários a serem cumpridos, começando na parada *Chico Mendes* às 05h40 e terminando no mesmo ponto às 06h55.

Linha: 263B - Tipo de Quadro de Horário: Dias Úteis

Número de viagens previstas: 59
Tempo médio previsto das viagens:74'
Frota prevista: 5 carros
Quilometragem da viagem completa: 25,07 km

#	Carro	CHICO MENDES	INTEGRACAO	MARINGA	INTEGRACAO	CHICO MENDES	Tempo Viagem	Quilometragem
1	1	05:40	06:08	06:10	06:23	06:55	75'	20,10 km
2	2	05:55	06:23	06:25	06:38	07:10	75'	20,10 km
3	3	06:10	06:38	06:40	06:53	07:25	75'	20,10 km
4	4	06:25	06:53	06:55	07:08	07:40	75'	20,10 km
5	5	06:40	07:10	07:12	07:25	07:55	75'	20,10 km
6	1	06:55	07:28	07:30	07:43	08:10	75'	20,10 km
7	2	07:10	07:43	07:45	07:58	08:25	75'	20,10 km
8	3	07:25	07:58	08:00	08:13	08:40	75'	20,10 km
9	4	07:40	08:13	08:15	08:28	08:55	75'	20,10 km
10	5	07:55	08:28	08:30	08:43	09:10	75'	20,10 km
11	1	08:10	08:43	08:45	08:58	09:25	75'	20,10 km
12	2	08:25	08:58	09:00	09:13	09:40	75'	20,10 km
13	3	08:45	09:15	09:17	09:30	09:55	70'	21,40 km
14	4	09:10	09:40	09:42	09:52	10:25	75'	20,10 km
15	1	09:25	09:58	10:00	10:13	10:40	75'	20,10 km
16	2	09:45	10:18	10:20	10:30	10:55	70'	20,10 km
17	3	10:05	10:35	10:37	10:50	11:25	80'	20,10 km
18	4	10:25	10:58	11:00	11:13	11:40	75'	20,10 km
19	1	10:40	11:13	11:15	11:28	11:55	75'	20,10 km
20	2	10:55	11:28	11:30	11:43	12:10	75'	20,10 km
21	5	11:10	11:43	11:45	11:58	12:25	75'	20,10 km
22	3	11:25	11:58	12:00	12:13	12:40	75'	20,10 km
23	4	11:40	12:13	12:15	12:28	12:55	75'	20,10 km
24	1	11:55	12:28	12:30	12:43	13:10	75'	20,10 km
25	2	12:10	12:43	12:45	12:58	13:25	75'	20,10 km
26	5	12:25	12:58	13:00	13:13	13:40	75'	20,10 km
27	3	12:40	13:13	13:15	13:28	13:55	75'	20,10 km
28	4	12:55	13:28	13:30	13:43	14:10	75'	20,10 km
29	1	13:10	13:43	13:45	13:58	14:25	75'	20,10 km
30	2	13:25	13:58	14:00	14:13	14:40	75'	20,10 km
31	5	13:40	14:13	14:15	14:28	14:55	75'	20,10 km
32	3	13:55	14:28	14:30	14:43	15:10	75'	20,10 km
33	4	14:10	14:43	14:45	14:58	15:25	75'	20,10 km
34	1	14:25	14:58	15:00	15:13	15:40	75'	20,10 km
35	2	14:40	15:13	15:15	15:28	15:52	72'	20,10 km
36	5	14:55	15:28	15:30	15:43	16:10	75'	20,10 km
37	3	15:10	15:43	15:45	15:58	16:25	75'	20,10 km
38	4	15:30	16:03	16:05	16:17	16:40	70'	20,10 km
39	1	15:52	16:26	16:28	16:40	17:10	78'	20,10 km
40	2	16:10	16:43	16:45	16:58	17:25	75'	20,10 km
41	3	16:25	16:58	17:00	17:13	17:40	75'	20,10 km
42	4	16:40	17:13	17:15	17:28	17:55	75'	20,10 km
43	5	16:55	17:28	17:30	17:43	18:10	75'	20,10 km
44	1	17:10	17:43	17:45	17:58	18:25	75'	20,10 km
45	2	17:25	17:58	18:00	18:13	18:40	75'	20,10 km

Figura 2.2: Tabela de horários programados da rota *263B* em um dia útil.

2.1.2 Operação do serviço

A maior parte da operação do serviço é registrada através de um validador existente em cada ônibus. O validador é o equipamento utilizado para o pagamento da tarifa além de controlar a catraca do ônibus. Dessa forma, mesmo que o usuário pague a tarifa em dinheiro, é necessário que o validador seja acionado para que a catraca libere a entrada do passageiro.

Além de registrar os passageiros que entram no ônibus, o validador é utilizado para registrar o início e fim da viagem. Para registrar um desses eventos é necessário que o motorista, cobrador ou fiscal acionem o validador registrando o horário de tal evento. No final do expediente, o ônibus volta para a garagem onde tudo que foi registrado nos validadores é coletado e enviados para a fiscalização da STTP.

Um aspecto negativo dessa estratégia é que não é possível identificar quando o passageiro sai do ônibus, inviabilizando o rastreio dos usuários dentro do sistema. Esse problema é agravado no terminal de integração onde o usuário paga apenas para entrar no terminal, logo, não se sabe qual ônibus o usuário entrou. Isso diminui a qualidade da informação de lotação existente.

2.1.3 Fiscalização

Após receber os dados de operação do dia a as viagens executadas são pareadas com as viagens programadas para que suas durações e horários sejam comparados. A STTP determina um limiar de 15 minutos de variação para que a viagem seja classificada como atrasada ou adiantada. Só após esse processo a STTP toma as medidas necessárias. Isso cria um atraso de no mínimo dois dias na solução dos problemas.

Como foi mostrado na Figura 2.2, a rota em questão possui 5 horários programados por viagem enquanto o sistema só é capaz de registrar os horários de início e fim. A fiscalização oficial da STTP é feita apenas considerando os horários de início e fim da viagem, entretanto, a STTP possui um grupo de fiscais que durante o dia se deslocam entre paradas internas da viagem e que registram o horário de passagem dos ônibus. Dessa forma, utilizando o exemplo da rota *263B* podemos dizer que durante o expediente dos ônibus existe a possibilidade de que um fiscal esteja na parada *Integração* ou na parada *Maringá* anotando os horários que os ônibus desta rota estão executando. Tal prática é mais frequente em rotas com muitos

atrasos em dias anteriores ou quando a STTP recebe alguma reclamação dos usuários.

Capítulo 3

Literatura

Previsão é um tema bastante estudado no contexto dos sistemas de trânsito e transporte. Isso inclui previsão de fluxo de tráfego, duração de viagens (ir de A para B), atrasos e horários. A partir da revisão desses trabalhos identificamos duas práticas dominantes: utilização de informação de localização em tempo real obtidas usando *GPS* e aplicação de *Aprendizado de Máquina* na previsão de tais variáveis.

Não conseguimos encontrar estudos com a mesma abordagem que utilizamos nesta pesquisa. Por isso, procuramos por práticas consolidadas na previsão de horários em sistemas de transporte com o objetivo de avaliar a sua utilização no contexto da cidade de Campina Grande.

3.1 Objetivo da previsão

Duas abordagens se destacam quanto ao objetivo da previsão. A primeira delas é a previsão de variáveis relacionadas ao tráfego que podem de alguma forma serem utilizadas na operação do sistema. A previsão da velocidade do tráfego é uma das formas de indiretamente calcular os horários dos ônibus e utilizar técnicas de *Aprendizado de Máquina* pode ser uma opção viável para tal tarefa [21].

De maneira semelhante, técnicas de estatística podem ser aplicadas com alguns aprimoramentos na previsão do fluxo de tráfego em curto prazo [20]. A mesma variável também foi prevista com a alcunha de volume de tráfego e com a utilização de *Redes Neurais* como técnica de previsão [24].

A segunda abordagem é a previsão do horário de passagem do ônibus ou como acontece em certos casos a previsão do tempo de viagem. Nesse contexto predomina a utilização de informação de localização em tempo real. Uma abordagem comum é a utilização desses dados em conjunto com técnicas de estatística como média móvel [9], *método de Monte Carlo* [11], *redes bayesianas* [12] e conjuntos difusos [14].

Além disso, muitos estudos tentaram aplicar *Aprendizado de Máquina* e *Redes Neurais* como o objetivo de melhorar os resultados alcançados até o momento. Foi possível identificar 4 algoritmos bastante presentes na literatura, sendo eles: *k-Nearest Neighbors*, *Artificial Neural Networks*, *Support Vector Machines* e *Kalman filters*. Neste estudo decidimos abordar apenas os três primeiros além de utilizar *Random Forests* como uma nova solução.

3.1.1 *k-Nearest Neighbors*

k-Nearest Neighbors é um dos algoritmos mais simples de *Aprendizado de Máquina*, contudo, sua utilização obteve resultados promissores. Uma característica importante desse algoritmo é a desnecessidade de analisar profundamente os dados. Em contrapartida, dados com muito ruído podem afetar o desempenho do modelo. Sabendo disso, remover este ruído usando *Singular Spectrum Analysis* (SSA) conseguiu melhorar o desempenho das previsões [10]. Para solucionar o problema de dados faltantes, interpolação dos horários se mostrou praticável [2; 16].

Além dos dados de entrada, a definição dos parâmetros de aprendizado também pode influenciar o desempenho das previsões. Para o k verificamos que em geral o valor escolhido é de aproximadamente 10 mas que isso pode variar com a mudança de contexto [2; 16]. Com relação à função de distância a opção mais comum é usar distância Euclidiana [2; 10], contudo, existe outras abordagens como usar *clustering* [16].

Por último, é necessário escolher a função de previsão com base nos vizinhos escolhidos. A opção mais usada é média ponderada [2; 9; 16].

3.1.2 *Artificial Neural Networks*

Redes Neurais são bastante conhecidas pela habilidade de encontrar padrões em problemas complexos usando apenas os dados. Essa capacidade vem ao custo de alto tempo de treina-

mento e dificuldade de entender como o modelo funciona.

A estrutura da rede neural utilizada varia com o contexto dos horários que serão previstos, entretanto, redes *feedforward* multi-camadas são comumente utilizadas [13; 15]. A utilização do histórico da variável resposta com entrada da rede neural se mostrou eficiente [13; 21], ainda assim, usar outras variáveis relacionadas ao meio também pode ser um caminho viável [15].

3.1.3 *Support Vector Regression*

A habilidade de generalização das *Support Vector Machines* e sua resistência a *overfitting* faz com que elas sejam uma opção interessante na previsão dos horários dos ônibus.

A escolha da *kernel function* é um passo essencial para a utilização de *SVMs*. Geralmente *Linear* e *Radial Basis* são as opções de obtiveram melhores resultados [17; 23], destacando que a primeira perde desempenho quando o tamanho da amostra diminui. Após a escolha da *kernel function* é necessário definir seus parâmetros de aprendizado. Para isso, a estratégia mais comum é faze-lo empiricamente [17].

Tal solução já foi comparada com previsões utilizando estatística e redes neurais. No primeiro caso, *SVM* conseguiu obter melhores resultados que o concorrente [23]. Quando comparado com a aplicação de redes neurais, *SVM* se destacou quando o conjunto de treino era de apenas alguns dias, enquanto seu desempenho foi superado quando o tamanho do conjunto de treino aumentou [21].

3.2 Nossa contribuição

A partir do que foi apresentado neste capítulo, podemos identificar que até o momento em que este estudo foi realizado a previsão de horários de ônibus tem como principal fundamento a utilização de algoritmos de *Aprendizado de Máquina* aplicados em dados de localização em tempo real. Não foi possível identificar a melhor solução indicando que o contexto em que tais algoritmos são utilizados tem grande influência nos resultados. Notamos também que os bons resultados obtidos usando dados de localização em tempo real fez com que poucas abordagens utilizem outras fontes de informação relacionadas ao sistema. Por último, as soluções mais estudadas consideram apenas a previsão de horários em futuro próximo

considerando apenas uma porção do sistema.

Por isso, este estudo tem como objetivo atacar alguns dos aspectos que ainda não foram estudados. O primeiro deles é a falta de informação em tempo real e atualizada. O segundo é comparação sistemática de várias soluções incluindo uma que não conseguimos encontrar aplicação da mesma no contexto dos horários de ônibus. O terceiro aspecto é a avaliação de previsões de médio prazo, nesse caso um dia inteiro de viagens. Por último, este estudo considerou todo o sistema de transporte coletivo da cidade de Campina Grande.

Embora o foco deste estudo seja a indisponibilidade de informação de localização em tempo real, a aplicação dos resultados encontrados não está restrita a esses casos. Os modelos analisados neste estudo podem ser usados em conjunto com modelos que utilizam dados de localização em tempo real.

Capítulo 4

Metodologia

Este capítulo detalha os dados e métodos utilizados para criar os conjuntos de dados utilizados no treino e teste dos modelos de previsão criados. Também é explicado como as viagens executas são pareadas com as viagens programadas além de como essas viagens foram filtradas.

4.1 Dados utilizados

Os dados utilizados no estudo foram obtidos em uma parceria com a Superintendência de Trânsito e Transportes Públicos de Campina Grande (STTP-CG). Esses dados consistem de dois tipos de informação sobre o serviço de ônibus da cidade: o quadro de horários determinados pela STTP-CG e os horários executados pelo ônibus em cada viagem realizada. O primeiro especifica os horários de início e fim que uma determinada viagem deve cumprir. Esses valores são determinados com base na demanda dos locais que a viagem atende como também no conhecimento prévio dos responsáveis pela gerência e planejamento do serviço.

Em Campina Grande, os horários programados são divididos em três tipos de dia: dias úteis, sábado e domingo. Essa divisão estabelece que existem três itinerários diferentes dependendo do dia em que o serviço está sendo executado. O primeiro tipo define como o sistema deve operar de segunda a sexta-feira enquanto os outros dois especificam como o sistema deve ser executado no sábado e domingo.

Como Campina Grande não possui informação sobre a localização dos ônibus em tempo real, a única maneira de obter informação sobre as viagens é a partir dos registros de início

e fim de viagem realizados pelos motoristas, cobradores ou fiscais. Essas informações são coletadas através do validador usado para o pagamento da tarifa. Sempre que uma viagem é iniciada ou finalizada o motorista, o cobrador ou o fiscal tem a obrigação de registrar o fato no validador. Com essa prática é possível obter horário de inicio e fim de cada viagem, além de informações como motorista, quantidade de passageiros, quantidade de passageiros que pagaram meia passagem e quantidade de passageiros gratuitos.

4.2 Pareamento de viagens

Nos dados de horários de início e fim das viagens realizadas não há uma marcação que especifique qual viagem do quadro de horários corresponde a uma viagem feita por um ônibus. Por isso, um primeiro passo em nossa análise é parear as viagens executadas aos horários programados. Foi utilizado para isso o mesmo algoritmo de pareamento usado pela STTP-CG. O algoritmo define que uma viagem programada deve ser pareada à viagem realizada com menor diferença de tempo entre seus horários de início, atendendo a condição de que tal diferença não pode exceder o valor do *headway*. O *headway* de uma viagem **x** é a diferença do horário de início de **x** e de sua viagem seguinte. Também é necessário salientar que uma viagem programada só pode ser pareada com uma realizada e vice-versa.

O pareamento é feito de forma ordenada começando na primeira viagem programada do dia. Sempre que um horário programado é pareado com uma viagem executada, as duas são removidas de seus respectivos conjuntos de dados. O resultado do pareamento é um conjunto com três tipos de viagens: pareada, não executada e extra.

As viagens não executadas são viagens programadas que não foram pareadas com nenhuma viagem executada. É importante deixar claro que isso não quer dizer que não existiram viagens realizadas com o objetivo de cumpri-las. O não pareamento da viagem programada mostra que nenhuma viagem realizada atendeu aos parâmetros de tempo especificados pela STTP-CG. O inverso de uma viagem não executada é a viagem extra. Todas as viagens que não se adequaram às exigências da STTP-CG são consideradas viagens extras.

O foco desse estudo são as viagens pareadas pois apenas com elas é possível obter o histórico de operação de uma determinada viagem.

4.3 Viagens fechadas incorretamente

Como já foi explicado, a coleta dos horários de início e fim dependem do registro por parte do motorista, do cobrador ou do fiscal no validador existente no veículo. Contudo, existem situações em que as viagens não são finalizadas devidamente. O caso mais comum é que o encarregado só registra o início da primeira viagem e o fim da última viagem de seu expediente. Isso reflete na existência de viagens com duração fora do normal que consequentemente gera grande diferença de tempo entre o horário final programado e executado. Para remover tais registros, todas as viagens executadas com duração 2,5 vezes maior que a duração programada foram removidas dos conjuntos de dados.

4.4 Criação de atributos

Parte dos atributos usados no treinamento dos modelos já existem nativamente nos dados enquanto outros precisam ser gerados a partir dos atributos existentes ou coletados de fontes externas.

4.4.1 Categóricos

Estão nos dados viagens de 46 rotas que podem ser agrupadas em 17 linhas. Seis empresas foram responsáveis pela execução dessas rotas utilizando um total de 469 veículos. Cada viagem também é classificada pelo dia da semana, dia normal, férias e se ela foi realizada a noite. Para a STTP-CG são considerados dias normais terça, quarta e quinta-feira dado que é esperado comportamento diferente do sistema na segunda e sexta-feira devido especificidades da cidade. Definimos também que todas as viagens que ocorreram em janeiro, junho, julho e dezembro pertencem à categoria de realizadas no período de férias. Por último, se a viagem teve início após às 17:59 ela é categorizada como tendo acontecido à noite.

A Figura 4.1 mostra a quantidade de viagens existentes em cada uma das categorias. Como esperado, existe um desbalanceamento dentro das categorias com destaque para a linha e o período do dia. No caso das linhas, Figura 4.1a, a diferença entre a quantidade de viagens pode ser explicada por dois fatores: demanda de passageiros e tamanho da viagem. A linhas com menos viagens normalmente atendem a distritos ou bairros mais distantes do

centro da cidade. Já as linhas com mais viagens geralmente operam em bairros centrais da cidade com maior demanda de locomoção. A linha *Branca* é uma exceção, contudo, ela é a principal opção dos estudantes da Universidade Federal de Campina Grande e por isso possui grande quantidade viagens.

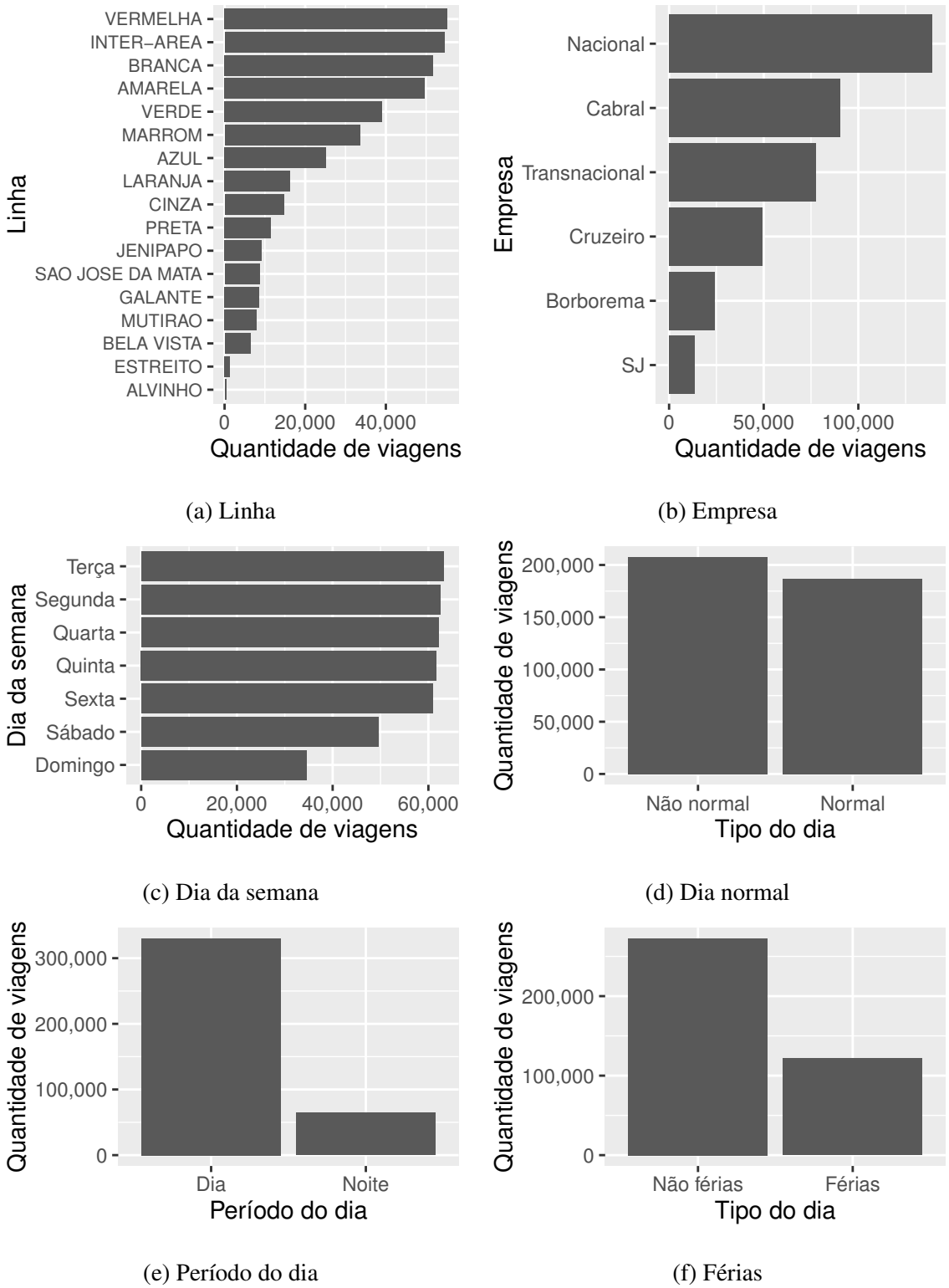
No caso do período do dia, a baixa quantidade de viagens a noite pode ser explicado pela queda na demanda após o horário de pico das *17h* às *19h*. Outra fator que deve ser considerado é que boa parte das viagens que rodam à noite podem ter iniciado antes das *17h* e como a rotulagem usa o horário de início da viagem, elas acabam sendo classificadas como viagens diurnas, como pode ser visto na Figura 4.1e.

Quando analisamos os dias da semana, o comportamento encontrado foi o aguardado com uma demanda similar entre os dias da semana e com uma queda no final de semana, Figura 4.1c. Por Campina Grande ser uma cidade universitária e boa parte desses estudantes não serem naturais da cidade é esperado que na segunda e sexta-feira a demanda seja inconstante. Por esse motivo, a STTP-CG utiliza a classificação de dia normal explicada acima.

4.4.2 Lotação

A quantidade de pessoas dentro do ônibus durante a viagem pode influenciar seu cronograma tanto diretamente quanto indiretamente. O ônibus tende a ter mais tempo parado se necessitar embarcar e desembarcar um número elevado de passageiros. Além disso, a lotação pode ser um indicador da demanda de locomoção da cidade no momento da viagem, por isso, se a demanda é alta muitos veículos devem estar circulando o que pode gerar demoras no trânsito.

No conjunto de dados usados no experimento existem três indicadores de lotação dos ônibus. O primeiro é a quantidade total de passageiros que entraram no ônibus durante a viagem, o segundo é quanto desse total pagou meia entrada e o terceiro é quanto do total entrou gratuitamente. É importante deixar claro que o nenhum dos três valores pode ser usado para representar a quantidade de passageiros que estavam, simultaneamente, dentro do ônibus durante a viagem. O experimento realizado utilizou apenas a quantidade total de passageiros como indicador de lotação.



(a) Linha

(b) Empresa

(c) Dia da semana

(d) Dia normal

(e) Período do dia

(f) Férias

Figura 4.1: Viagens agrupadas pelas variáveis categóricas.

4.4.3 Meteorológicos

As condições meteorológicas no momento da viagem podem ter relação com as condições das vias, tempo de embarque e desembarque de passageiros e quantidade de passageiros. Todos esses fatores influenciam a execução do serviço por parte dos ônibus. Por esses motivos, três dos atributos de uma viagem referem as condições climáticas do dia em que a viagem ocorreu, são eles: precipitação total, temperatura média e umidade média. Esses dados estão disponíveis no banco de dados observacionais do Centro de Previsão de Tempo e Estudos Climáticos [7].

A Figura 4.2, mostra o comportamento das três variáveis citadas durantes os dias existentes nos dados. É importante citar que não existe informação no período de 14/03/2015 à 11/04/2015, fazendo com que os valores das 3 variáveis sejam iguais a 0 no período em questão.

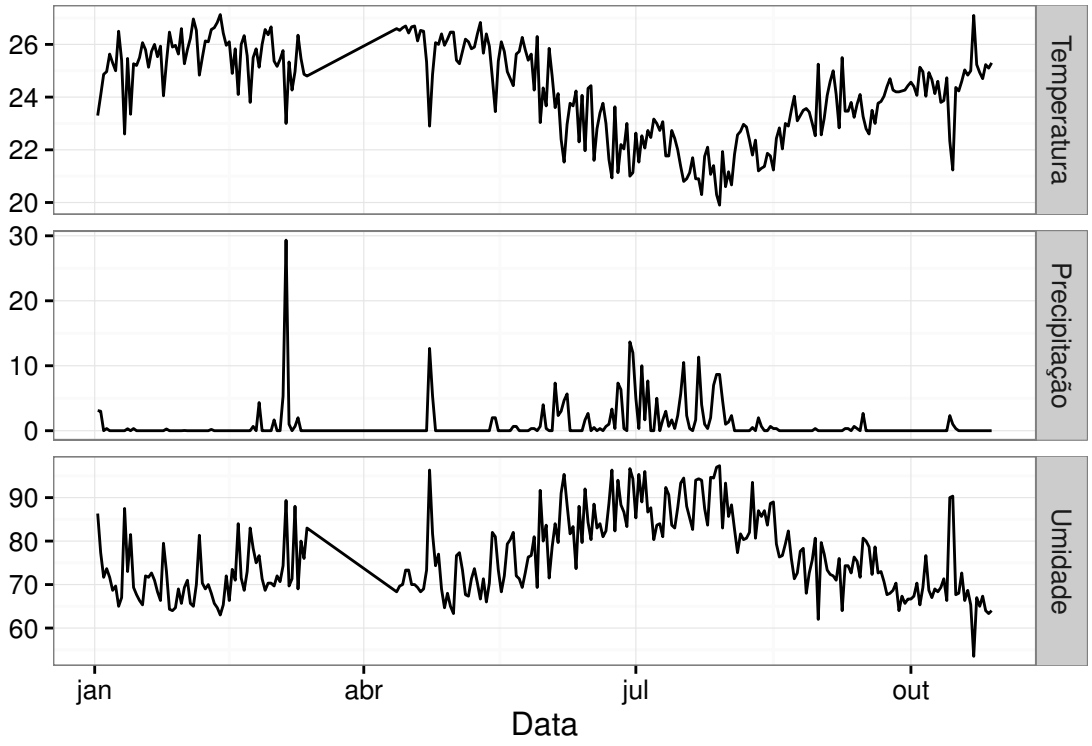


Figura 4.2: Comportamento das variáveis meteorológicas durante os dias usados no experimento.

4.4.4 Temporais

Parte dos atributos de uma viagem são cronológicos: dia do mês, mês do ano e o horário de início programado. Cada viagem também possui dados referentes as viagens respectivas em dias anteriores. Viagens respectivas são aquelas que foram pareadas com o mesmo horário programado. Sendo $v_{r,i,d}$ a viagem do dia d da rota r que foi pareada com a viagem programada i, $v_{r,i,d}$ possui horário de início/fim observado, quantidade de passageiros, atraso inicial e final, precipitação, temperatura e umidade das viagens $v_{r,i,(d-k)}$ onde $k \in 1, 2, 3$.

A utilização desse histórico de viagens permite representar o desempenho do sistema no passado recente de maneira que eventos recentes que possam influenciar a operação do sistema sejam considerados no treinamento dos modelos.

4.5 Variável resposta

Os modelos treinados usando os atributos citados acima tem como objetivo prever dois valores: a pontualidade referente ao início da viagem e a pontualidade do final da viagem. Ao observar a variação desses valores nos dados é possível notar que em mediana os ônibus começam as viagens com 1 minuto de atraso e terminam aproximadamente 3 minutos atrasados. Nas duas situações a execução dos ônibus está dentro do previsto na perspectiva da STTP, que define uma margem de 15 minutos para que uma viagem seja classificada com atrasada.

Mesmo que em mediana o serviço esteja funcionando corretamente existem casos que os horários executados pelos ônibus são distantes do programado como pode ser visto na Figura 4.3. Por exemplo, a viagem com a maior variação do horário inicial começou com um pouco mais de 5 horas de atraso, e a viagem com maior variação final terminou aproximadamente 6 horas após o horário previsto. Ainda existem casos em que as viagens começam ou terminam antes do horário previsto, chamadas de adiantadas. As viagens adiantadas são tão problemáticas quanto as atrasadas e também possuem uma margem de 15 minutos para serem punidas. A viagem mais adiantada teve início aproximadamente 5 horas antes do programado e a viagem que terminou com quase 5 horas a menos que o programado foi a viagem com maior adiantamento no horário final.

Uma viagem que iniciou com 5 horas de atraso leva a crer que pode existir algum erro

nos dados ou no pareamento. Entretanto, essas situações não são estranhas no contexto de Campina Grande. Por exemplo, se uma rota possui dois horários de início programados: *9h* e *15h*, e por algum motivo a primeira viagem não foi executada e a segunda viagem teve início às *14h*, o pareamento final será entre a viagem programada para as *9h* e a viagem executadas às *14h* gerando um atraso de 5 horas para o início da viagem. Essas situações são predominantes nas linhas que operam nos distritos de Campina Grande como a linha *Alvinho*.

Os atrasos e adiantamentos no início totalizam 28.852 viagens representando pouco mais de 7% das viagens pareadas. O cumprimento do horário final das viagens é mais inconsistente sendo 125.585 consideradas atrasadas ou adiantadas, quase 32% das viagens.

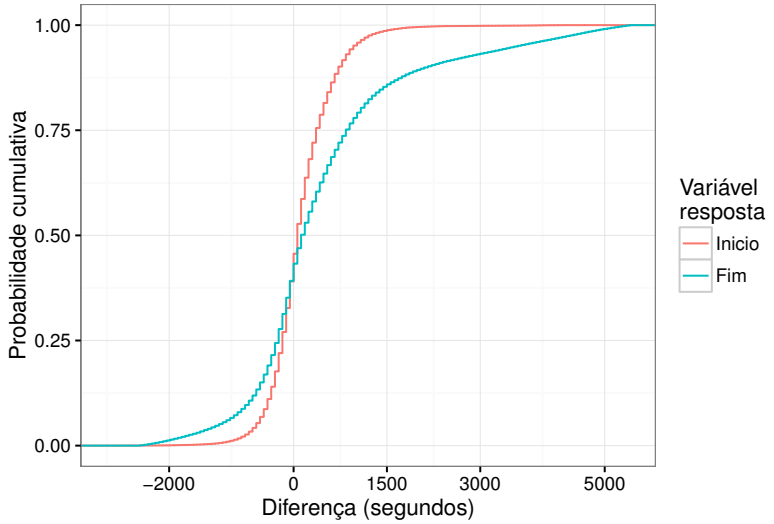
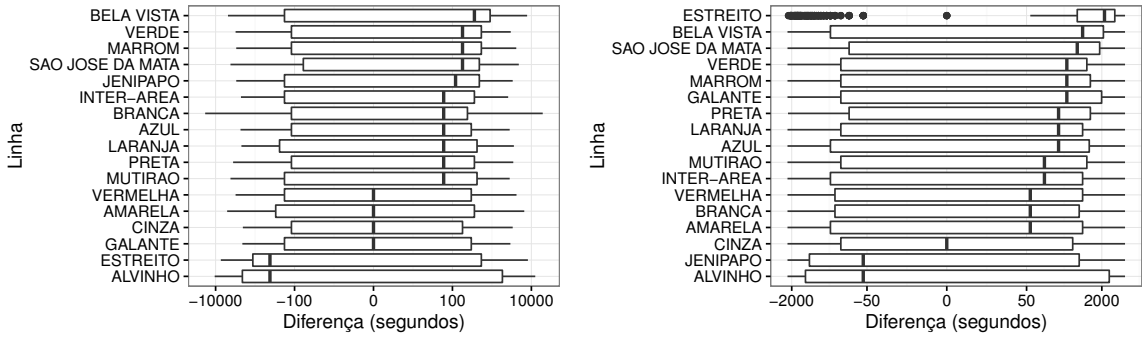


Figura 4.3: *ECDF* da diferença entre horários programados e executados pelos ônibus.

Avaliando a pontualidade das linhas é possível notar na Figura 4.4a que as linhas *Vermelha, Amarela, Cinza e Galante* são bastante pontuais com relação ao início da viagem, com medianas próximas a 0. Apenas duas linhas possuem mediana menor que 0, entretanto, todas as linhas possuem ocorrências com início adiantado.

A Figura 4.4b mostra que das 4 linhas com melhor desempenho no início da viagem apenas a linha *Cinza* consegue manter o mesmo nível de pontualidade no final. Ainda é possível verificar que a linha *Estreito* é a mais impontual com 75% de seus registros entre 420 e 3900 segundos.



(a) Inicio da viagem

(b) Fim da viagem

Figura 4.4: *Boxplot* das diferenças entre horário programado e executado, agrupados por linha.

Capítulo 5

Resultados

Neste capítulo apresentamos o desempenho de quatro algoritmos de regressão: kNN, ANN, SVM e Random Forest, sendo os três primeiros bastante utilizados na literatura para a previsão de horário de ônibus [13; 16; 17; 21]. Na comparação também foi considerado como *baseline* um algoritmo que chamamos de *schedule* que prevê a diferença entre o horário programado e o observado sempre como 0. Este algoritmo equivale à ausência do uso de qualquer informação além da programação criada pela STTP.

O desempenho dos modelos criados foi comparado através do intervalo de confiança da mediana do erro. Os intervalos de confiança foram calculados usando *bootstraping*. Para visualizar os resultados usando a escala *log* foi acrescentado 1 segundo a todas as previsões com erro igual a zero. Além disso, a transformação para *log* em valores negativos foi aplicada ao valor absoluto do erro em seguida o sinal original foi aplicado ao valor.

5.1 Experimento

Cada ensaio do experimento divide os dados em treino e teste. O conjunto de teste é o equivalente a um dia completo de viagens pareadas com os horários estabelecidos. Para o treino foram utilizados os 7, 15 ou 45 dias imediatamente predecessores do dia usado no teste. Além disso, dentre os atributos de cada viagem existe o histórico recente das viagens respectivas em 1, 2 ou 3 dias anteriores. Usando este formato, foram escolhidos 12 dias aleatórios entre 01/01/2015 até 28/10/2015 e para cada um deles o experimento foi realizado. Em cada um destes dias todas as combinações dos demais fatores são exercitadas como parte

do experimento.

Não existiu nenhuma restrição na definição dos dias selecionados como pode ser notado na Figura 5.1. O conjunto de dias escolhidos totaliza 15.269 viagens de 46 rotas agrupadas em 17 linhas executadas por 6 empresas do total de 394.295 viagens do conjunto de dados usado no experimento.

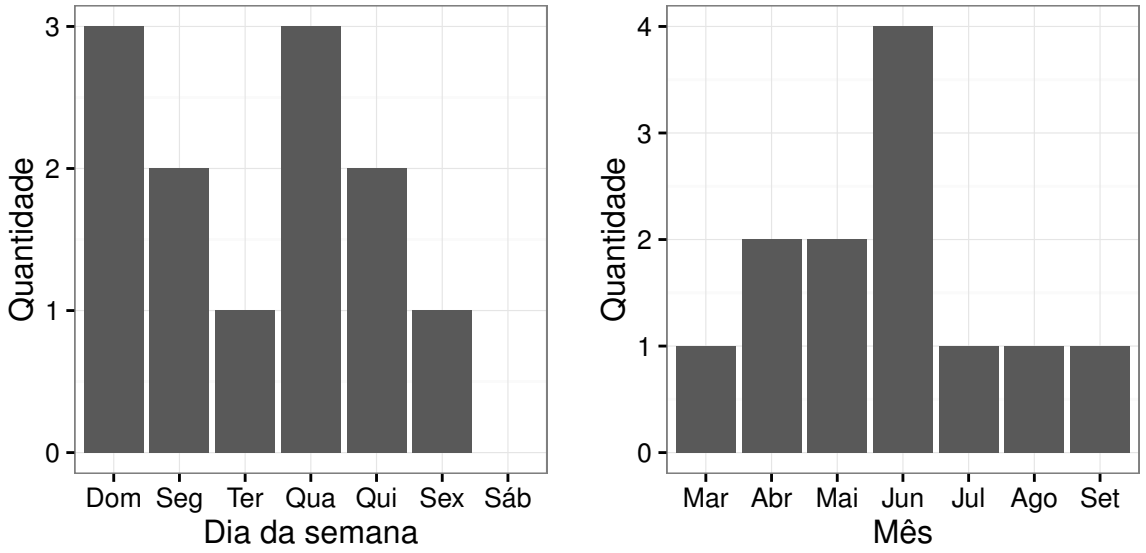


Figura 5.1: Datas escolhidas para o conjunto de teste agrupadas por dia da semana e mês do ano.

O valor previsto para cada viagem é a diferença de tempo, em segundos entre o horário programado e o horário que irá acontecer. Por exemplo, para uma viagem programada para às 17h, um modelo pode ter como resposta -1.080s (= −18min), implicando em um horário previsto para essa viagem de $17h - 18min = 16h42$, 18 minutos antes do programado.

5.1.1 Infraestrutura

Dado a experiência prévia, todo o experimento, desde o pré-processamento até a análise dos resultados, foi executado usando o ambiente de computação estatística *R*[1]. Para o pré-processamento foram usados os pacotes *dplyr*[2], *tidyr*[3], *reshape2*[4] e *lubridate*[5]. Para a criação

[1]https://www.r-project.org/

[2]https://cran.r-project.org/web/packages/dplyr/index.html

[3]https://cran.r-project.org/web/packages/tidyr/index.html

[4]https://cran.r-project.org/web/packages/reshape2/index.html

[5]https://cran.r-project.org/web/packages/lubridate/index.html

e validação dos modelos foram usados os pacotes *caret*[6] e *doMC*[7]. Por último, na análise dos resultados e criação das visualizações foram usados os mesmos pacotes usados no pré-processamento mais o pacote *ggpplo2*[8].

5.1.2 Parâmetros de aprendizado

A definição dos parâmetros de aprendizado de cada um dos algoritmos testados foi feita empiricamente através a função *train* do pacote *caret*. Essa função se encarrega de treinar vários modelos usando o algoritmo passado como parâmetro da função e usar validação cruzada repetida para encontrar quais os valores dos parâmetros de aprendizado que minimizam a métrica de avaliação escolhida, no caso desse experimento a raiz quadrada do erro quadrático médio (RMSE).

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(c_i - \bar{c}_i)^2}$$

Para a criação dos modelos usando *k-Nearest Neighbors*, o método passa para a função *train* foi *knn*. Esse método define que o parâmetro de aprendizado que será ajustado é o *k* que define a quantidade de vizinhos que serão utilizados para a previsão [1]. O k escolhido foi 9 para os modelos de previsão do início e fim das viagens.

O parâmetro *ann* define que os modelos treinados devem usar uma *Rede Neural* com uma camada escondida como algoritmo de previsão dos horários. Tal método define os valores dos parâmetros de aprendizado *size* e *decay*. O primeiro representa a quantidade de neurônios na camada escondida e o segundo é o fator de decadência dos pesos de cada neurônio [19]. Para o início da viagem, a quantidade de neurônios na camada escondida escolhida foi 1 enquanto para o final da viagem o valor escolhido foi 3. O fator de decadência escolhido para o início da viagem foi 0 e para o final foi 0,1.

Os modelos de *Support Vector Machines* possuem um parâmetro de aprendizado que não é avaliado pela função *train* do *caret*, a *kernel function*. Com base na literatura, função de base radial tende a obter os melhores resultados na previsão de horários [17; 23], logo ela foi selecionada como *kernel function*. Além do parâmetro C *SVMs* que utilizam função de

[6] http://topepo.github.io/caret/index.html

[7] https://cran.r-project.org/web/packages/doMC/index.html

[8] http://ggplot2.org/

base radial necessitam do parâmetro γ. Tanto para o início quanto para o final da viagem os mesmos valores para C e γ foram selecionados, 1 e 0,0067 respectivamente.

Para os modelos criados usando Árvores aleatórias dois atributos foram considerados. O primeiro é a quantidade de árvores criadas, parâmetro esse que tem influência direta no custo de treinamento dos modelos. A quantidade de árvores escolhida foi 10 para a previsão do início e fim das viagens. O segundo parâmetro é o m_{try} que define a quantidade de atributos que serão utilizadas na criação de cada árvore de decisão [5]. O m_{try} selecionado para a previsão do início da viagem foi 2 enquanto para o final da viagem o valor selecionado foi 11.

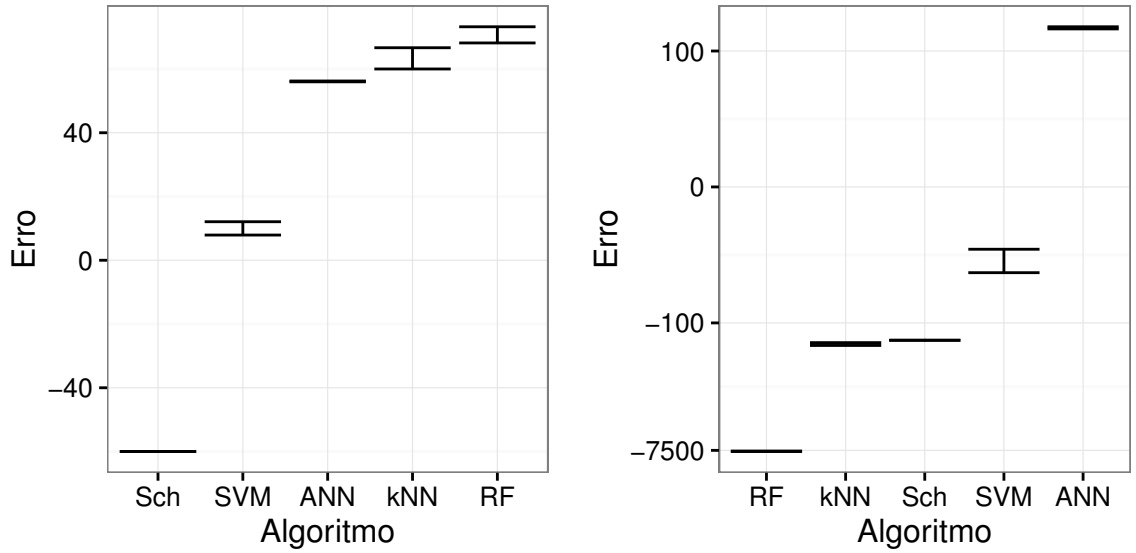
5.2 Comparação entre algoritmos

Os resultados mostram que os erros na previsão do horário de início da viagem foram menores do que na previsão do horário final. Esse comportamento era esperado dado que as viagens são pareadas a partir do horário de início. Os erros para o início da viagem ficaram entre -900 e 780 segundos com mediana de aproximadamente 28 segundos. A variação do erro para o final da viagem foi maior estando entre -38870 e 1649 segundos com mediana de quase -167 segundos.

A diferença entre os erros agrupados pelos algoritmos pode ser vista na Figura 5.2. É visível que os modelos treinados usando *SVM* obtiveram os melhores resultados dado a proximidade do 0. Ainda é possível apontar que para o início da viagem apenas o algoritmo *schedule* tende a errar para menos. Já para o final da viagem o comportamento o inverso tendo apenas o *ANN* errando para mais. A tabela 5.1 apresenta o ranking dos algoritmos considerando o valor absoluto da mediana do erro.

5.3 Tamanho do conjunto de treino

Um dos fatores avaliados foi o tamanho do conjunto de dados utilizado no treinamento dos modelos. O tamanho desse conjunto é definido pela quantidade de dias, imediatamente anteriores ao dia previsto, que serão usadas no treinamento. Esse valor varia entre 7, 15 e 45 dias. A Figura 5.3 apresenta a quantidade de viagens usadas em cada dia previsto.



(a) Inicio da viagem

(b) Fim da viagem

Figura 5.2: Intervalos de confiança da mediana do erro agrupado por algoritmos.

Algoritmo	*Momento*	*Mediana do erro (s)*
SVM	Inicio	10,09
SVM	Fim	-13,60
ANN	Inicio	56,04
kNN	Inicio	60,00
Schedule	Inicio	-60,00
RF	Inicio	70,00
Schedule	Fim	-180,00
kNN	Fim	-206,66
ANN	Fim	224,66
RF	Fim	-7735

Tabela 5.1: Ranking da mediana do erro agrupado por algoritmo.

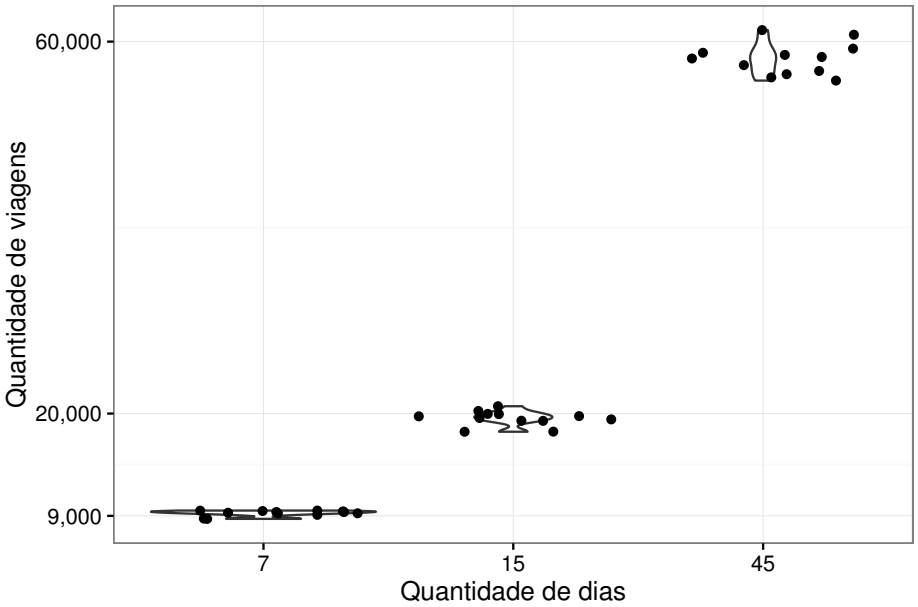


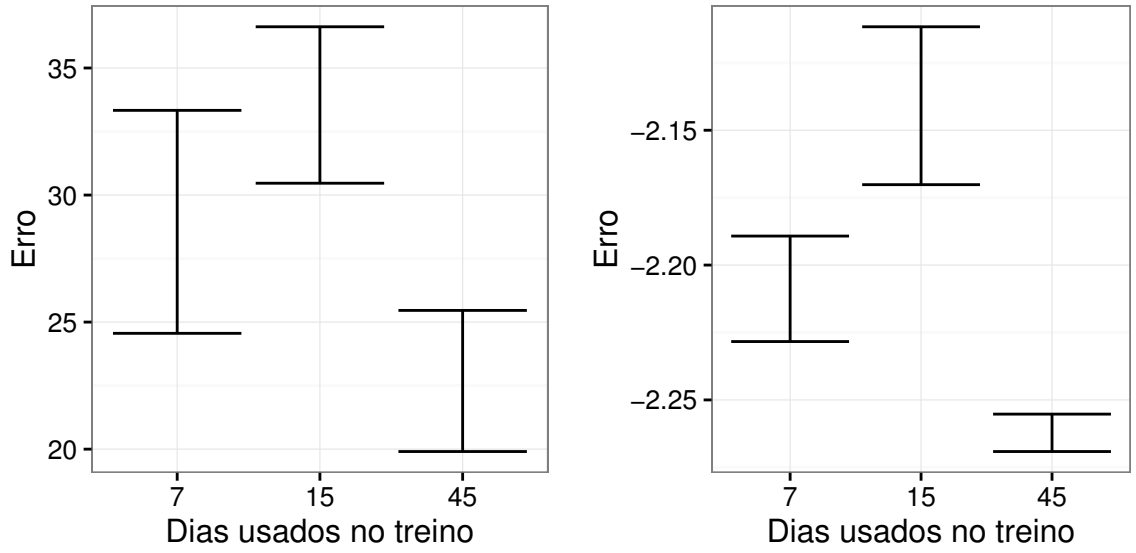
Figura 5.3: Quantidade de viagens por conjunto de treino.

Os resultados mostram que a variação do tamanho do conjunto de treino pode influenciar no desempenho dos modelos criados tanto no início da viagem quanto no final, como pode ser visto na Figura 5.4. Para o início da viagem não é possível afirmar qual o melhor desempenho entre a utilização de 7 e 15 dias, entretanto, treinar os modelos com 45 dias tende a obter os melhores resultados. No final da viagem a diferença entre os níveis é mais clara sendo 15 dias a melhor opção seguido de 7 e 45 dias, nessa ordem.

Avaliando tal fator considerando cada um dos algoritmos, é possível notar que a variação identificada globalmente é originada principalmente dos modelos criados usando florestas aleatórias. Nos demais algoritmos a variação da quantidade de dias usados no treino não gera variação tão forte nos resultados, como pode ser notado na Figura 5.5. Está claro que utilizar 45 dias no treinamento influencia consideravelmente o desempenho da RF. Essa influência é positiva no início da viagem e negativa no final.

5.4 Tamanho do histórico

O terceiro fator considerado no experimento foi o tamanho do histórico recente da viagem. Como explicado na Seção 4.4.4, o histórico varia de 1 a 3 dias. Na Figura 5.6 é possível notar que a influência da variação do histórico usado é ainda mais aparente do que no caso



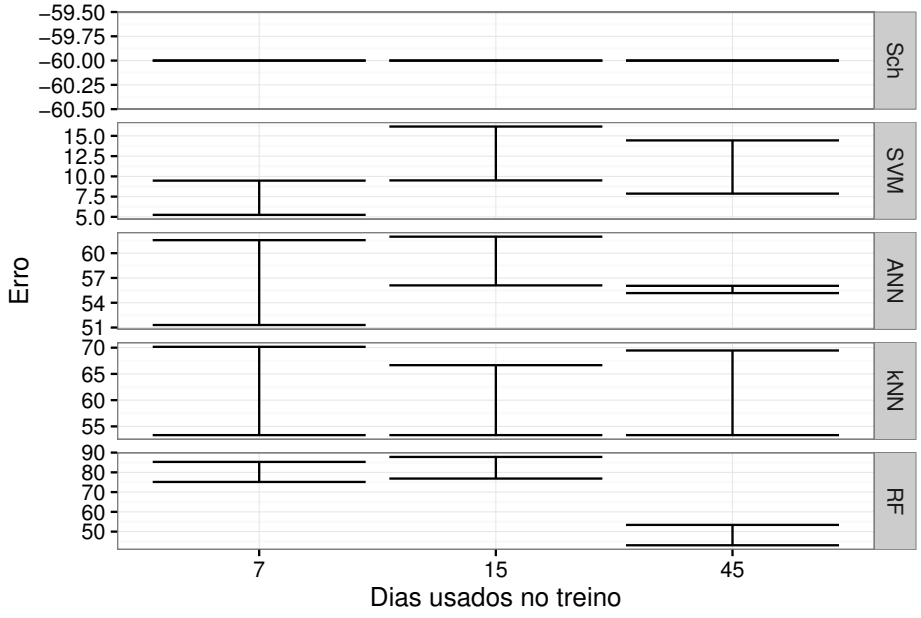
(a) Inicio da viagem

(b) Fim da viagem

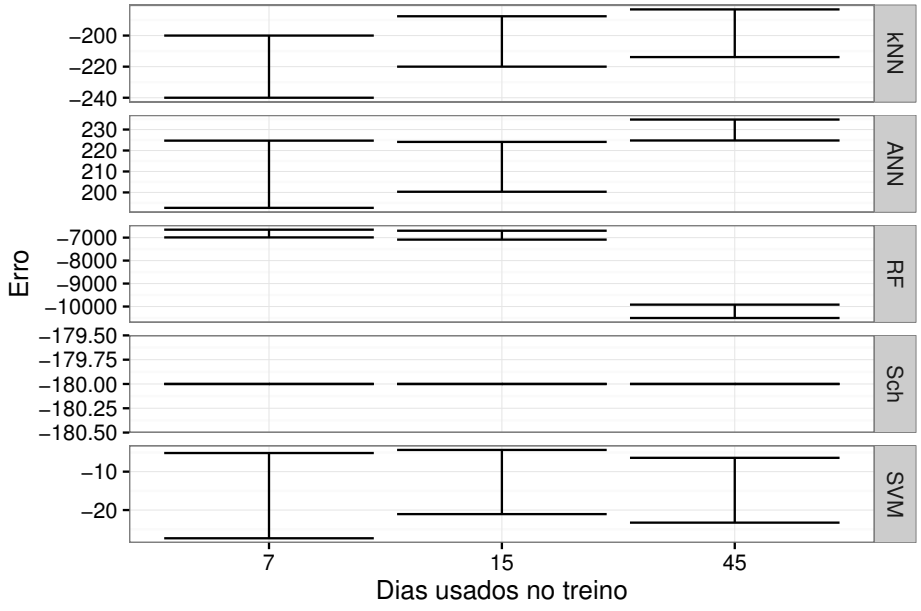
Figura 5.4: Intervalos de confiança da mediana do erro agrupado por quantidade de dias usados no conjunto de treino.

do tamanho do conjunto de treino. Para o início da viagem a melhor opção parece ser a utilização de 2 dias, porém, ainda existe uma pequena sobreposição entre os intervalos de confiança. No caso do final da viagem a diferença é ainda mais perceptível tendo a utilização de 3 dias como melhor opção.

Seguindo o mesmo princípio, foi analisado como esse fator se comporta quando separado por algoritmo. Considerando o início da viagem, apenas o *kNN* e o *RF* mostram indícios de que o tamanho do histórico pode afetar os resultados obtidos. Nos dois casos utilizar 2 dias se mostrou melhor que usar apenas 1 dia, contudo, adicionar o terceiro dia só gerou melhoras para o *kNN*. Algo semelhante pode ser afirmado para o final da viagem onde apenas *kNN* e o *RF* mostram ser influenciados. O *kNN* tem o mesmo comportamento tanto para o início quanto para o final da viagem, já o *RF* obteve melhores resultados usando apenas 1 dia de histórico. Lembrando que não foi possível identificar diferença concreta com a variação do tamanho do histórico para os demais algoritmos.

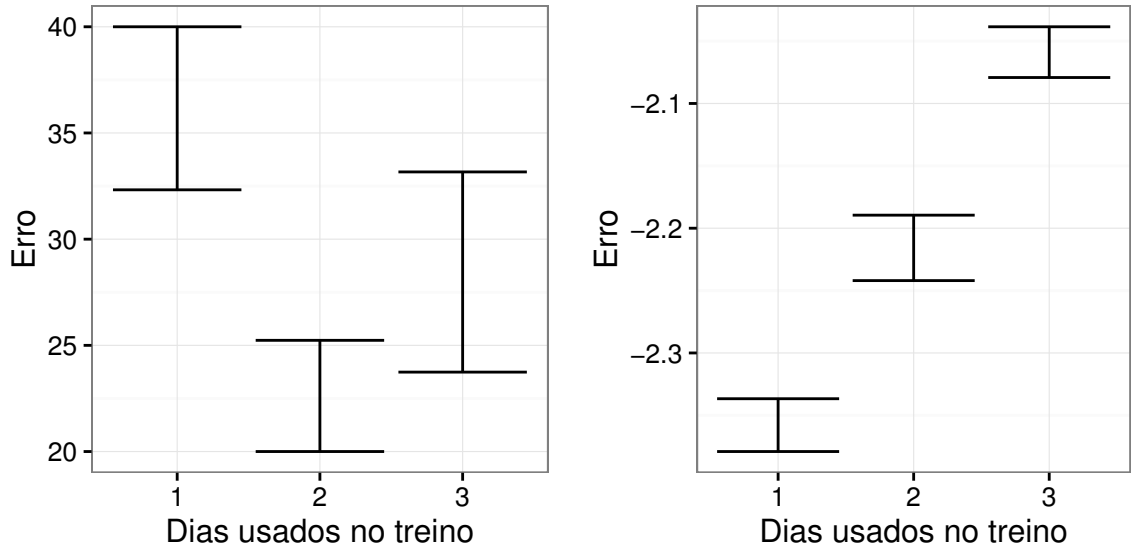


(a) Inicio da viagem



(b) Fim da viagem

Figura 5.5: Intervalos de confiança da mediana do erro agrupado por quantidade de dias usados no conjunto de treino para cada algoritmo.



(a) Inicio da viagem

(b) Fim da viagem

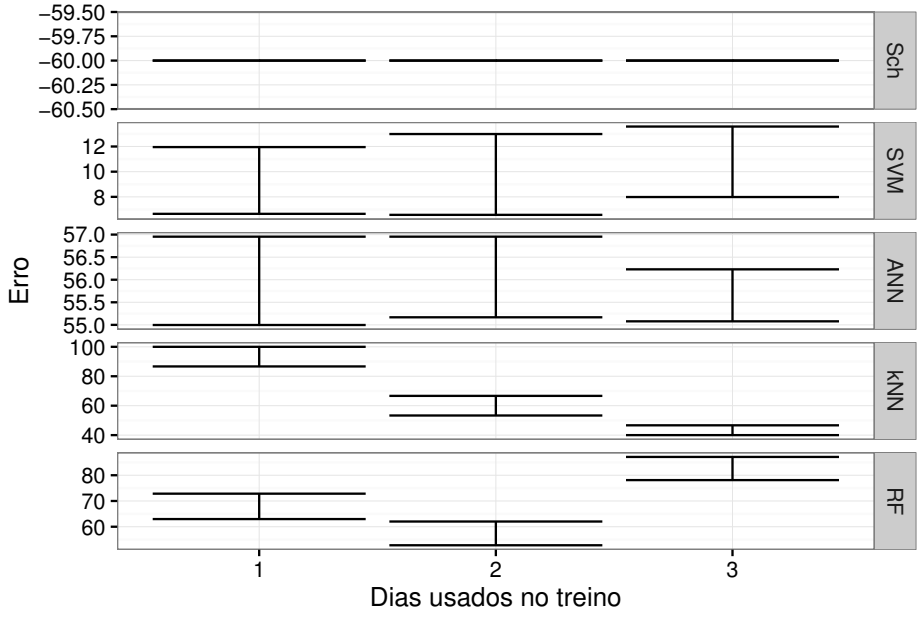
Figura 5.6: Intervalos de confiança da mediana do erro agrupado por quantidade de dias usados no histórico recente da viagem.

5.5 Resultado geral

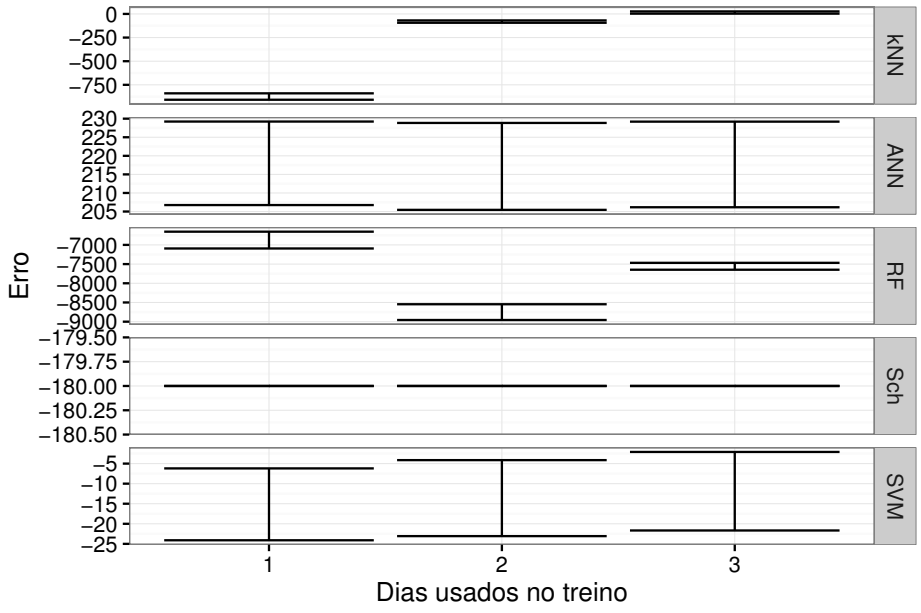
Fechando a primeira parte dos resultados, foi avaliada a interação entre os 3 fatores considerados no experimento: algoritmo, tamanho do conjunto de treino e tamanho do histórico recente. A Figura 5.8 apresenta o ranking com as combinações dos fatores ordenado a partir do valor absoluto do erro.

Como apresentado anteriormente *SVM* obteve os melhores resultados tanto para o início quanto para o final da viagem. Mesmo assim, algumas combinações usando os outros algoritmos conseguiram se aproximar dos resultados alcançados usando *SVM* como *RF-45-2* para o início da viagem e *kNN-7-3*, *kNN-15-3* e *kNN-45-3* para o final da viagem. Também é importante citar que mesmo não sendo o algoritmo mais acurado, *ANN* se mostrou a opção bastante precisa na previsão das duas variáveis resposta.

Além dos fatores usados no treinamento dos modelos é possível comparar o desempenho dos algoritmos utilizando caraterísticas dos conjuntos de teste como mês, dia da semana, linha, entre outros. Para facilitar a visualização dos próximos resultados, a partir deste ponto serão considerados apenas os modelos criados usando *SVM*.

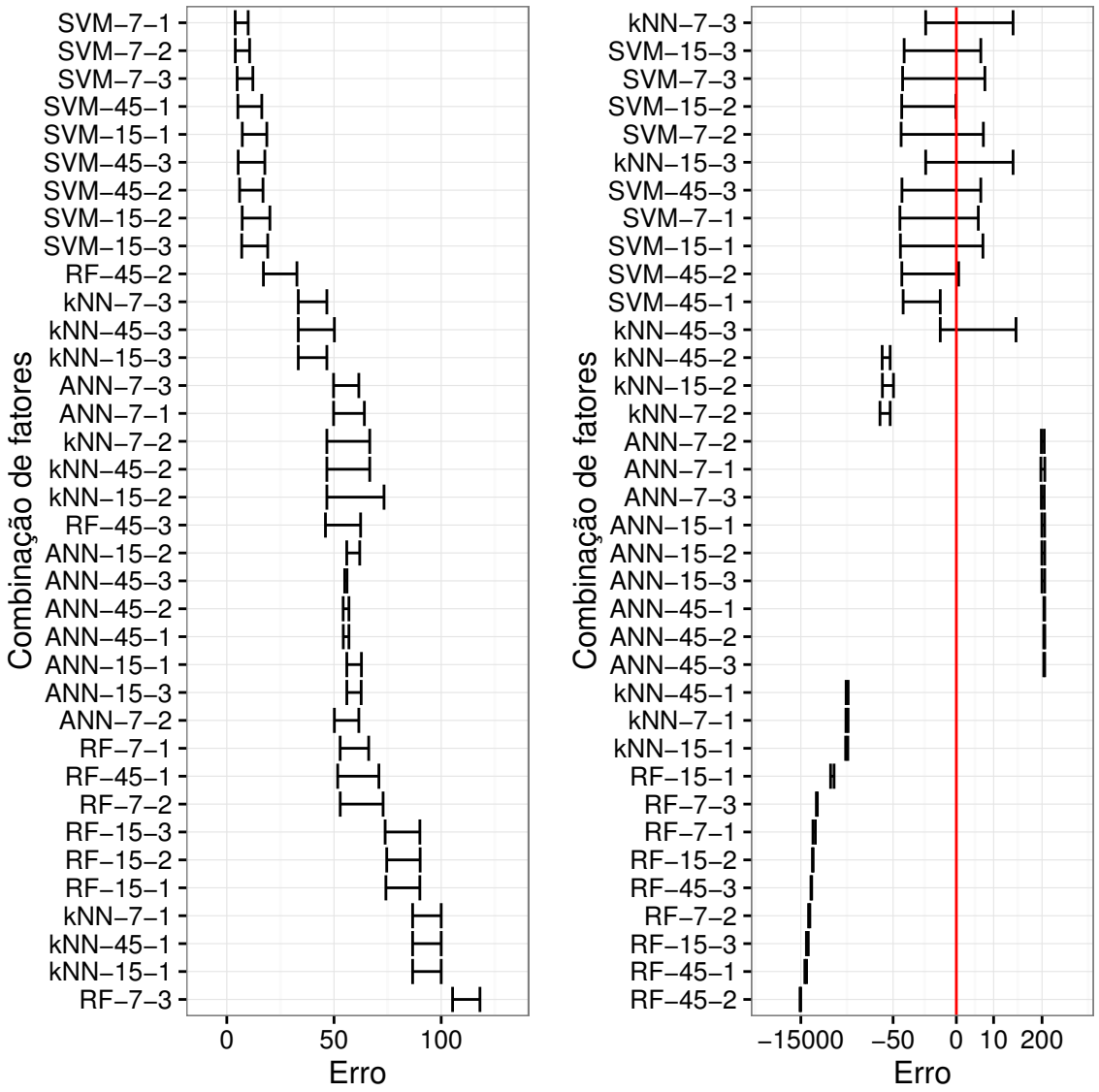


(a) Inicio da viagem



(b) Fim da viagem

Figura 5.7: Intervalos de confiança da mediana do erro agrupado por quantidade de dias usados no histórico recente da viagem para cada algoritmo.



(a) Inicio da viagem

(b) Fim da viagem

Figura 5.8: Intervalos de confiança da mediana do erro para cada combinação dos fatores: algoritmo, quantidade de dias usados no conjunto de treino e quantidade de dias usados no histórico recente.

5.6 Comparação por períodos do ano

A época do ano em que a viagem ocorreu pode ser um aspecto importante na previsão dos horários. O mês do ano pode ser usado para identificar eventos que podem influenciar a operação do serviço como período de férias escolares [4] ou ocorrência de eventos, como o Maior São João do mundo no caso de Campina Grande. Essas situações podem influenciar tanto a demanda por locomoção quanto as condições de operação dos ônibus e consequentemente a previsibilidade do serviço.

A Figura 5.9 mostra que o comportamento esperado não pode ser totalmente confirmado. Os valores mais incomuns foram obtidos no mês de maio com o aumento do erro mediano. Já em junho que é o começo do período de férias e dos São João, ocorre uma melhora no desempenho das previsões. Esse comportamento pode ser explicado pela importância do São João fazendo com que o planejamento e fiscalização do serviço seja mais rigorosa tornando o serviço mais confiável e previsível. O que ocorre no mês de maio pode ser um efeito colateral da implementação do planejamento realizado para o mês de junho.

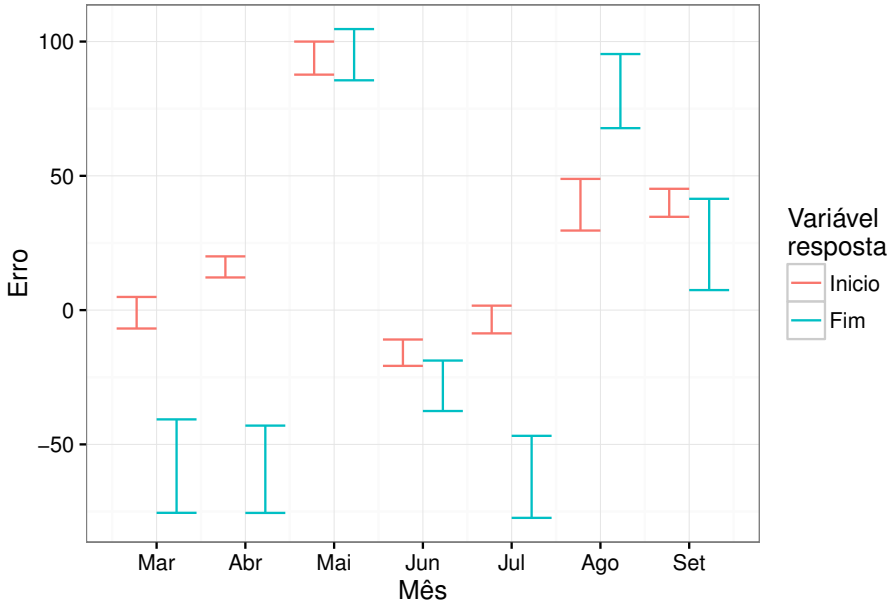


Figura 5.9: Intervalo de confiança da mediana do erro do algoritmo *SVM* agrupado por mês do ano.

5.7 Comparação por dia da semana

A operação dos ônibus durante a semana é um lado importante do serviço. Existe uma variação esperada de demanda durante o final de semana que pode afetar o desempenho dos modelos. Essa variação pode ser explicada principalmente pelo período letivo das escolas e da universidade existente na cidade. É esperado que na sexta-feira a demanda comece a diminuir por causa dos alunos que voltam para suas cidades e na segunda ocorra o aumento da demanda por causa da volta desses alunos. Essa variação no comportamento pode comprometer as previsões.

Como esperado a qualidade dos resultados cai no Domingo e tem seus melhores resultados no meio da semana (quarta e quinta-feira). Os melhores resultados para o início da viagem foram verificados na segunda e sexta enquanto a quarta e quinta são os dias com melhores previsões para o final da viagem. Em linhas gerais, a previsão do horário de início da viagem foi mais acurada com exceção da quarta e quinta onde o desempenho das previsões do horário final conseguiram se aproximar.

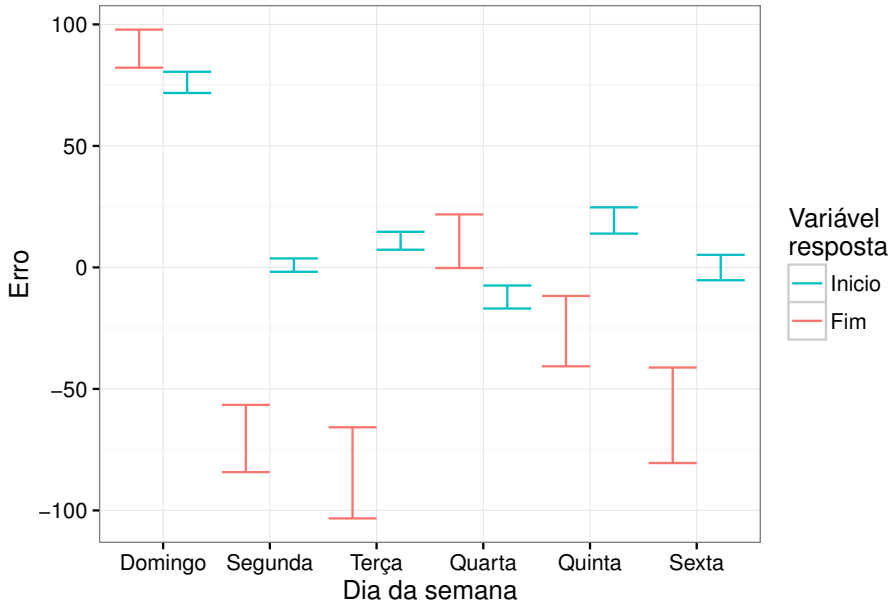


Figura 5.10: Intervalo de confiança da mediana do erro do algoritmo *SVM* agrupado por dia da semana.

5.8 Comparação por hora do dia

O desempenho dos modelos também foi comparado na perspectiva das horas do dia. Na Figura 5.11 é possível visualizar o comportamento dos erros durante as horas do dia. Fica claro que durante os horários de início e final de expediente os erros são mais inconsistentes. A linha vermelha define o limiar de classificação de uma viagem como viagem realizada à noite ou durante o dia. Esse mesmo horário é considerado pela STTP-CG como horário de pico. Nota-se que na vizinhança da linha vermelha existe uma variação maior nos intervalos de confiança.

É importante citar que os modelos obtiveram bom desempenho no horário de pico de *11h* às *13h*. Esse período do dia possui grande demanda do serviço e por isso é relevante para o usuário que as previsões nesse momento do dia sejam eficientes.

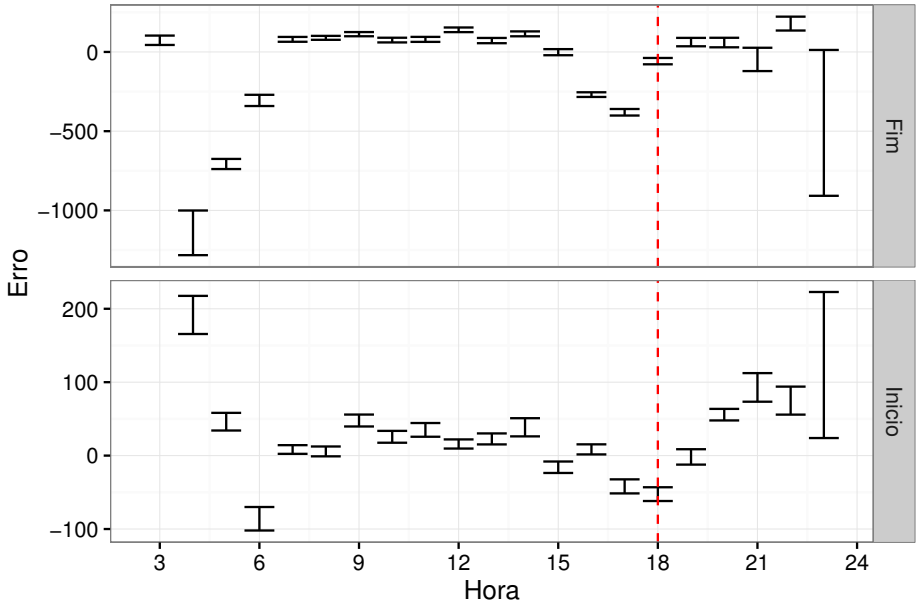


Figura 5.11: Intervalo de confiança da mediana do erro do algoritmo *SVM* agrupado por hora do dia.

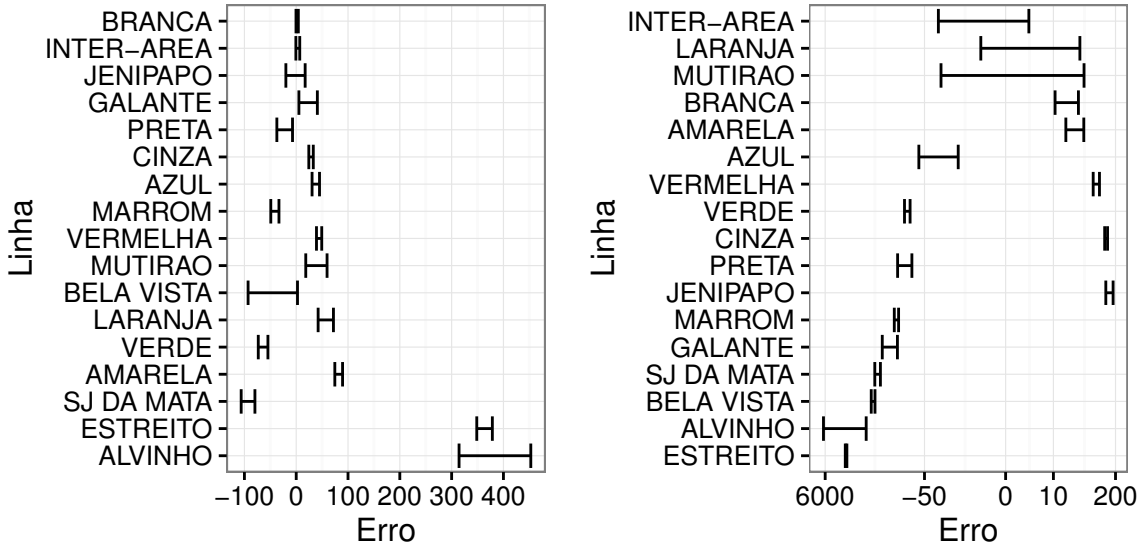
5.9 Comparação por linha

A partir dos dados utilizados a única maneira de representa o percurso da viagem é através da rota ou da linha do ônibus. Essa informação pode ser usada para avaliar se existem

locais da cidade onde a previsão dos horários é melhor ou pior. Por questões de visualização utilizamos as linhas como variável de comparação.

Os rankings das linhas podem ser vistos na Figura 5.12. A partir dos rankings é possível identificar linhas como *Galante* que obteve bons resultados para o início da viagem e resultados não tão bons para o final da viagem. O inverso do comportamento encontrado para *Galante* pode ser visto nas linhas *Amarela* e *Laranja*.

Duas linhas se destacaram por obter os piores resultados tanto para o início quanto para o final da viagem, foram as linhas *Estreito* e *Alvinho*. Essas linhas são exemplos de linhas que operam nos distritos da cidade. O destaque positivo é a linha *Inter-área* que aparece no *top 2* dos dois rankings.



(a) Inicio da viagem

(b) Fim da viagem

Figura 5.12: Intervalo de confiança da mediana do erro do algoritmo *SVM* agrupado por linha.

5.10 Importância dos atributos

Por último, foram selecionados os melhores modelos de cada um dos algoritmos com o intuito de verificar a importância dos atributos no resultado da previsão. Essa tarefa foi realizando usando a função *varImp* do *caret* que avalia a importância das variáveis independentes

Algoritmo	*Inicio*	*Fim*
kNN	7-3	7-3
ANN	7-3	7-2
SVM	7-1	15-3
RF	45-2	15-1

Tabela 5.2: Melhores combinações de quantidade de dias usado no conjunto do treino e quantidade de dias no histórico recente para cada algoritmo e variável resposta.

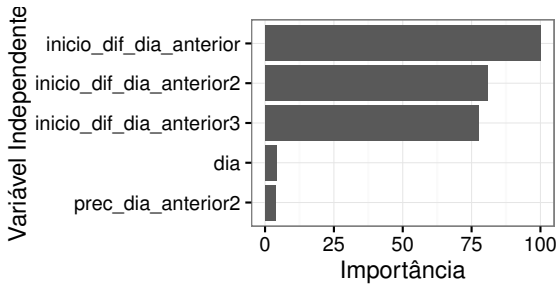
usando métodos específicos para cada algoritmo [9]. É possível ver na Tabela 5.2 a combinação de tamanho do conjunto de teste e tamanho do histórico recente escolhida para cada algoritmo em cada variável resposta.

É possível notar na Figura 5.13 que a variação no horário em dias anteriores é importante para quase todos os modelos avaliados. No caso de *kNN* e *SVM* é possível afirmar que a importância das demais variáveis chega a ser insignificante. O dia do mês, a rota, a linha e o fato da viagem ser executada a noite também aparecem em vários dos modelos avaliados.

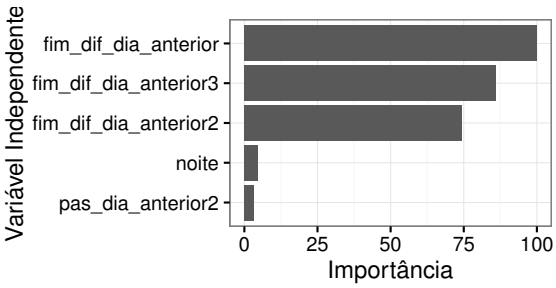
Um aspecto interessante é que os modelos criados com *ANN* e *RF* conseguiram se aproveitar mais de outras variáveis independentes além da diferença de horários em dias anteriores, enquanto os resultados de *kNN* e *SVM* são basicamente provenientes da diferença de horários em dias anteriores.

Como identificado, *SVM* obteve os melhores resultados usando basicamente uma variável independente para o início da viagem e três para o final. Isso mostra que em casos onde a informação disponível é limitada talvez o *SVM* seja a melhor opção. Enquanto isso quando a disponibilidade de informação histórica é diversificada modelos criados usando *ANN* ou *RF* podem ser uma opção extra.

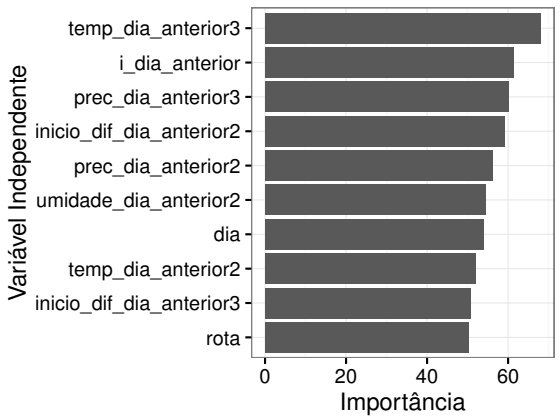
[9] http://topepo.github.io/caret/varimp.html



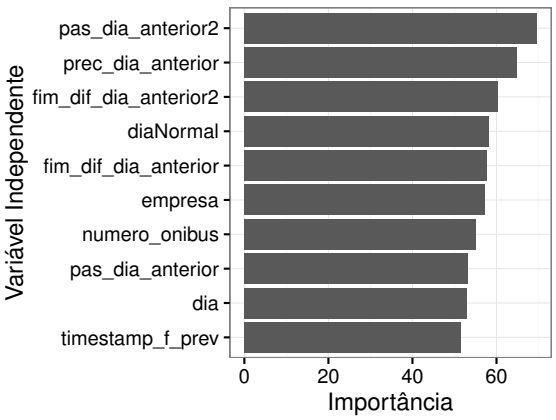
(a) kNN-7-3 Inicio



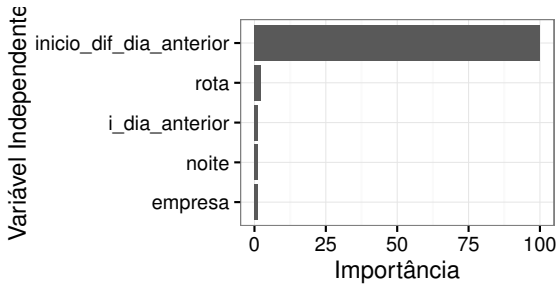
(b) kNN-7-3 Fim



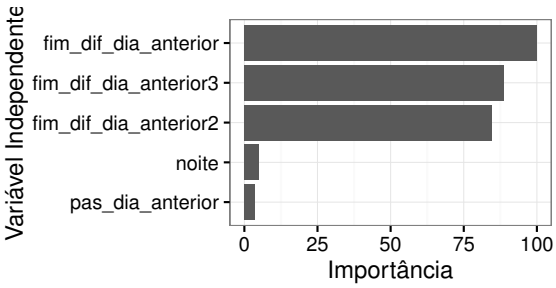
(c) ANN-7-3 Inicio



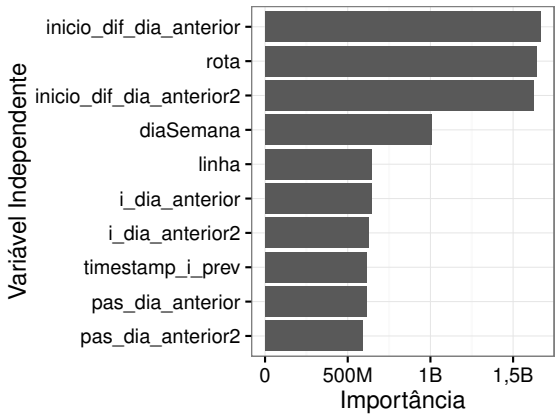
(d) ANN-7-2 Fim



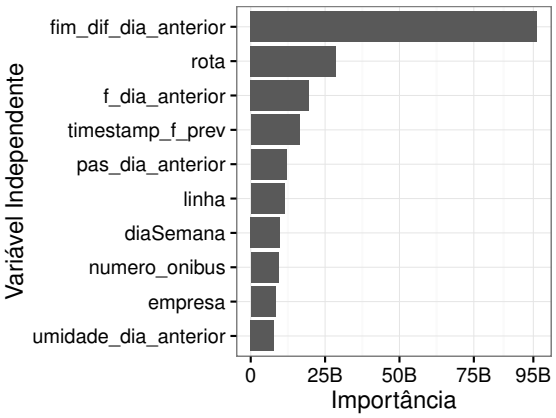
(e) SVM-7-1 Inicio



(f) SVM-15-3 Fim



(g) RF-45-2 Inicio



(h) RF-15-1 Fim

Figura 5.13: Ranking de importância das variáveis independentes para cada uma das combinações listadas na Tabela 5.1. Dado a baixa influencia da maioria das variáveis apenas o *top 5* foi mostrado para *kNN* e *SVM*. Para as demais foi mostrado o *top 10*.

Capítulo 6

Conclusões

Neste trabalho foi analisado o desempenho de 4 algoritmos, sendo três deles considerados estado da arte na previsão de horários de ônibus usando informação de localização em tempo real, quando aplicados no contexto de Campina Grande que não possui tal informação. Para suprir a falta de dados mais precisos e atualizados foram utilizados dados históricos de 3 tipos: categóricos, lotação e meteorológicos. Até onde sabemos esse é o primeiro estudo que trabalhou com dados que não possuem informação em tempo real para prever os horários dos ônibus.

O desempenho dos algoritmos foi avaliado considerando aspectos como quantidade de dias usados no treinamento, quantidade de dias usados no histórico recente de uma viagem, linha, mês do ano, dia da semana e hora do dia com o objetivo de prever a diferença dos horários de início e fim realizados e programados de um dia viagens. Além disso esses algoritmos foram comparados a um valor *baseline* sempre previa essa diferença igual a 0 – horário programado executado perfeitamente. Este experimento no revelou que algumas das combinações de fatores obtiveram melhores resultados que a utilização do valor *baseline*, mostrando que é possível tornar o serviço de ônibus mais previsível.

De maneira geral, notamos que a variação desses fatores pode afetar o desempenho dos modelos criados e que certos níveis de fatores interagem melhor. Ainda podemos identificar que os erros foram maiores para a previsão do horário final do que para o horário inicial da viagem, com medianas de -167 e 28 segundos respectivamente. Esses valores mostram que é possível obter resultados semelhantes aos encontrados na literatura que ficam em torno de 100 segundos. A melhor solução como um todo foi a utilização de *Support Vector Machine*,

entretanto, existem casos em que os outros algoritmos obtiveram resultados semelhantes ou melhores.

Ainda foi possível notar que o fator temporal é bastante importante na previsão dos horários, dado que foi verificada variação significante nos resultados com mudança do mês, dia da semana e hora do dia em que a viagem aconteceu.

Por último avaliamos da importância das variáveis preditivas para os melhores modelos de cada algoritmo, e verificamos que a variável que mais se destacou foi a varição do horário executado em dias anteriores, principalmente para os modelos criados usando *SVM* e *kNN*. Também foi visto que outras variáveis como rota, dia da semana, noite e quantidade de passageiros em dias anteriores podem ser consideradas importantes na previsão dos horários de início e fim das viagens.

A partir dessas conclusões podemos afirmar que é possível tornar o sistema de transporte público coletivo de Campina Grande mais previsível, e que isso pode ser alcançado utilizando técnicas de *Aprendizado de Máquina* em conjunto com os dados disponibilizados pelo sistema. De maneira mais direta, prever os horários de início e fim das viagens de um dia usando *Support Vector Regression* com 7 dias no conjunto de treino e 3 dias de histórico recente poderá melhorar a qualidade de informação disponível tanto para o usuário quanto para o gestor.

6.1 Limitações

Todos os resultados apresentados são apenas um ponto de partida para a solução da previsão de horários de ônibus sem informação de localização em tempo real e atualizada. Este trabalho atacou uma de duas partes que envolve a solução desse problema. Enquanto os resultados da previsão de horários de inicio e fim das viagens se mostraram promissores, o cálculo dos horários durante a viagem ainda precisa ser melhorados.

Quanto ao desempenho dos algoritmos, mostramos que não foi encontrada uma solução geral. Logo calibrar cada algoritmo para cada particularidade do sistema pode ser um caminho favorável. Esse caminho ainda pode ser continuado com a utilização de simultânea de vários modelos.

O pareamento de viagens e a identificação de irregularidades nas viagens é outro aspecto

deste estudo que pode ser melhorado através de técnicas de *Aprendizado de Máquina*. Tais procedimentos foram executados com base na prática do órgão gestor do serviço e em certos casos não são as mais indicadas.

Por último, o contexto deste estudo foi totalmente quantitativo. Acreditamos que o fator qualitativo pode ajudar na aplicação dos modelos. Particularmente na questão do cálculo do erro a opinião do usuário do serviço pode ser um fator importante na definição de métricas de qualidade dos modelos de previsão utilizados.

Bibliografia

[1] David W. Aha, Dennis Kibler, and Marc K. Albert. Instance-based learning algorithms. *Machine Learning*, 6(1):37–66, 1991.

[2] A. T. Baptista, E. P. Bouillet, and P. Pompey. Towards an uncertainty aware short-term travel time prediction using gps bus data: Case study in dublin. In *2012 15th International IEEE Conference on Intelligent Transportation Systems*, pages 1620–1625, Sept 2012.

[3] Ana L.C. Bazzan and Franziska Klügl. Introduction to intelligent systems in traffic and transportation. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 7(3):1–137, 2013.

[4] A. I. Bejan, R. J. Gibbens, D. Evans, A. R. Beresford, J. Bacon, and A. Friday. Statistical modelling and analysis of sparse bus probe data in urban areas. In *Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on*, pages 1256–1263, Sept 2010.

[5] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[6] P. Chen, X. p. Yan, and X. h. Li. Bus travel time prediction based on relevance vector machine. In *2009 International Conference on Information Engineering and Computer Science*, pages 1–4, Dec 2009.

[7] CPTEC/INPE. Portal de Tecnologia da Informação para Meteorologia. http://bancodedados.cptec.inpe.br/, 2016.

[8] Laura Eboli and Gabriella Mazzulla. A methodology for evaluating transit service

quality based on subjective and objective measures from the passenger's point of view. *Transport Policy*, 18(1):172 – 181, 2011.

[9] J. Gong, M. Liu, and S. Zhang. Hybrid dynamic prediction model of bus arrival time based on weighted of historical and real-time gps data. In *2013 25th Chinese Control and Decision Conference (CCDC)*, pages 972–976, May 2013.

[10] F. Guo, R. Krishnan, and J. W. Polak. Short-term traffic prediction under normal and incident conditions using singular spectrum analysis and the k-nearest neighbour method. In *Road Transport Information and Control (RTIC 2012), IET and ITS Conference on*, pages 1–6, Sept 2012.

[11] A. Hadachi, S. Mousset, and A. Bensrhair. Practical testing application of travel time estimation using applied monte carlo method and adaptive estimation from probes. In *Intelligent Vehicles Symposium (IV), 2012 IEEE*, pages 1078–1083, June 2012.

[12] A. Hofleitner, R. Herring, P. Abbeel, and A. Bayen. Learning the dynamics of arterial traffic from probe data using a dynamic bayesian network. *IEEE Transactions on Intelligent Transportation Systems*, 13(4):1679–1693, Dec 2012.

[13] R. Jeong and R. Rilett. Bus arrival time prediction using artificial neural network model. In *Intelligent Transportation Systems, 2004. Proceedings. The 7th International IEEE Conference on*, pages 988–993, Oct 2004.

[14] Yanying Li and M. McDonald. Link travel time estimation using single gps equipped probe vehicle. In *Intelligent Transportation Systems, 2002. Proceedings. The IEEE 5th International Conference on*, pages 932–937, 2002.

[15] Hao Liu, Ke Zhang, Ruihua He, and Jing Li. A neural network model for travel time prediction. In *Intelligent Computing and Intelligent Systems, 2009. ICIS 2009. IEEE International Conference on*, volume 1, pages 752–756, Nov 2009.

[16] T. Liu, J. Ma, W. Guan, Y. Song, and H. Niu. Bus arrival time prediction based on the k-nearest neighbor method. In *Computational Sciences and Optimization (CSO), 2012 Fifth International Joint Conference on*, pages 480–483, June 2012.

[17] J. n. Wang, X. m. Chen, and S. x. Guo. Bus travel time prediction model with v - support vector regression. In *2009 12th International IEEE Conference on Intelligent Transportation Systems*, pages 1–6, Oct 2009.

[18] R. P. S. Padmanaban, K. Divakar, L. Vanajakshi, and S. C. Subramanian. Development of a real-time bus arrival prediction system for indian traffic conditions. *IET Intelligent Transport Systems*, 4(3):189–200, September 2010.

[19] Kevin L Priddy and Paul E Keller. *Artificial neural networks: an introduction*, volume 68. SPIE Press, 2005.

[20] Shiliang Sun, Guoqiang Yu, and Changshui Zhang. Short-term traffic flow forecasting using sampling markov chain method with incomplete data. In *Intelligent Vehicles Symposium, 2004 IEEE*, pages 437–441, June 2004.

[21] L. Vanajakshi and L. R. Rilett. A comparison of the performance of artificial neural networks and support vector machines for the prediction of traffic speed. In *Intelligent Vehicles Symposium, 2004 IEEE*, pages 194–199, June 2004.

[22] Simon P Washington, Matthew G Karlaftis, and Fred Mannering. *Statistical and econometric methods for transportation data analysis*. CRC press, 2010.

[23] Chun-Hsin Wu, Jan-Ming Ho, and D. T. Lee. Travel-time prediction with support vector regression. *IEEE Transactions on Intelligent Transportation Systems*, 5(4):276–281, Dec 2004.

[24] Y. Zhengxiang, X. Guimin, and W. Jinwen. Transport volume forecast based on grnn network. In *Future Computer and Communication (ICFCC), 2010 2nd International Conference on*, volume 3, pages V3–629–V3–632, May 2010.