Mariana Santos Campos

Definição de tempos de percursos para linhas da Sociedade de Transportes Colectivos do Porto (S.T.C.P.)

Departamento de Matemática
Faculdade de Ciências da Universidade do Porto
2014

Mariana Santos Campos

Definição de tempos de percursos para linhas da Sociedade de Transportes Colectivos do Porto (S.T.C.P.)

Tese submetida à Faculdade de Ciências da Universidade do Porto para obtenção do grau de Mestre em Engenharia Matemática

Orientador: Prof.ª Doutora Ana Rita Gaio (FCUP)
Dr. Luís Giroto (STCP)
Coorientador: Prof.ª Doutora Teresa Mendonça (FCUP)

Departamento de Matemática
Faculdade de Ciências da Universidade do Porto
2014

Agradecimentos

Este projeto só foi possível graças à colaboração inestimável de algumas pessoas, a quem gostaria de agradecer de seguida.

À Dra. Luísa Campolargo por ter recebido com agrado a minha proposta de estágio, todas as palavras são poucas para agradecer esta oportunidade, muito obrigada por toda a confiança que sempre depositou em mim.

Ao Dr. Luís Giroto por ter aceite tão prontamente este projeto, pela sua grande disponibilidade, orientação e paciência que sempre demonstrou durante todo o meu estágio, foi uma experiência muito enriquecedora que jamais esquecerei por todas as vivências e aprendizagens diárias.

À Professora Rita Gaio pela liberdade científica que sempre me concedeu e por acreditar sempre em mim, sem esquecer a sua hábil orientação, simpatia e rigor.

À Professora Teresa Mendonça pela ajuda e orientação na procura de estágio mas também pela sua disponibilidade e simpatia.

A todos os colaboradores da STCP por me terem recebido e integrado tão bem na empresa. Em especial, às pessoas do Departamento de Operações, ao Sr. Ribeiro, ao Sr. Pereira, ao Sr. Serra, ao Sr. Amadeu, ao Sr. Nogueira e à D. Madalena por toda a simpatia e paciência na explicação de todas as etapas, e pelo sorriso diário com que me receberam (mesmo quando o Benfica ganhava).

À minha família por todo o apoio demonstrado durante esta longa viagem, em especial à Kiki, sem ela esta missão teria sido impossível .

Aos meus amigos e colegas, toda a amizade, paciência e ajuda.

Ao Tim, que na inocência dos seus três anos, tinha sempre o melhor abracinho do mundo pronto.

Ao João, por tudo, sempre.

Resumo

Atualmente as empresas de transportes necessitam de informar os passageiros, com o maior rigor possível, do horário a que os autocarros passam nas paragens. A Sociedade de Transportes Colectivos do Porto (S.T.C.P.), única empresa pública de transportes da área metropolitana do Porto, não é exceção. A S.T.C.P. tem uma vasta rede de linhas e paragens distribuídas por 6 concelhos. Para servir os clientes cada vez melhor, o erro do horário deve ser sempre baixo, não se permitindo que o autocarro tenha um atraso ou um adiantamento superior a 5 minutos.

Antes da resolução do problema da eventual definição de novos horários, foi necessário quantificar os atrasos e adiantamentos que estão a ser cometidos. Para isso foi criado um relatório automático de cumprimento de serviço. Esta ferramenta é bastante importante para a S.T.C.P por ser um indicador de fiabilidade e por detalhar numérica e graficamente todos os registos por nó e segmento em vários períodos horários e diferentes dias da semana.

A construção de novas propostas de horários baseou-se em cinco metodologias distintas: três de base empírica, outra baseada no Método dos Mínimos Quadrados Generalizados (MMQG), e outra que recorreu a Máquinas de Suporte Vectorial (MSV). No MMQG foram consideradas várias combinações entre diferentes funções de variância e diferentes estruturas de correlação, sendo que a melhor combinação foi escolhida tendo em conta essencialmente critérios de informação. Nas MSV foram considerados diferentes parâmetros para várias funções núcleo e através de validação cruzada foram selecionados os melhores parâmetros para cada núcleo; a melhor função núcleo foi escolhida de acordo com a apresentação do menor erro. Para cada um dos modelos considerados, foram analisados os erros máximos cometidos e a correlação das previsões com os tempos de viagem observados. Para a seleção da melhor metodologia foram tidas em consideração as seguintes medidas de desempenho: erro absoluto médio (EAM), o erro percentual absoluto médio (EPAM) e a raíz do erro quadrático médio (REQM).

Espera-se que o modelo final seja relevante para a empresa, conseguindo aumentar significativamente o nível de cumprimento de serviço.

Palavras-chave: PREVISÃO; TEMPO DE VIAGEM; CUMPRIMENTO DE SERVIÇO; MÉTODO DOS MÍNIMOS QUADRADOS; MÁQUINAS DE SUPORTE VECTORIAL; FUNÇÕES NÚCLEO; VALIDAÇÃO CRUZADA.

Abstract

Currently bus operating companies need to inform passengers, with the highest possible accuracy, about the time that buses run at stops. Sociedade de Transportes Colectivos do Porto (S.T.C.P.), the only public transport bus company in the metropolitan area of Porto, is not an exception. S.T.C.P. has an extensive network of lines and stops spread across 6 counties. To better serve their customers, failures in the fullfillment of the time schedules should be in very small numbers; in particular, time delays or advances have to be shorter than 5 minutes.

In order to evaluate the need for a definition of new time schedules, current bus delays and advances had to be readily quantified. For this an automatic service fullfillment report was created. This new reliability tool is thought to be quite important to STCP as it friendly details, numerically and graphically, current trajectory times by node, segment, period of the day, and period of the week.

The definition of new bus time schedules was obtained by five different methodologies: three based on empirical grounds, the other using the Generalized Least Squares (GLS) method, and the third applied Support Vector Machines (SVM). In the GLS method, different combinations of variance functions and correlation structures were considered, and the choice of the best model was based on information criteria. The application of SVM considered different core parameters and kernel functions, and used cross-validation and evaluation of the prediction errors for selection of the best model. The evaluation of the different time schedule proporsals was based on the following performance measures: mean absolute error (MAE), mean absolute percentage error (MAPE) and the root mean square error (RMSE).

We expect the work that has been develloped in this thesis to be of great relevance to STCP, and that it may raise the company levels of service commitment.

Keywords: PREDICTION; TRAVEL TIME; FULFILLMENT OF SERVICE; GENERALIZED LEAST SQUARES; SUPPORT VECTOR MACHINES; KERNEL FUNCTIONS; CROSS-VALIDATION.

Conteúdo

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

1.1 Apresentação da Empresa

A designação do Serviço de Transportes Colectivos do Porto (S.T.C.P.) surge em 1946, com o Resgate da Concessão feita pela Câmara Municipal do Porto a um grupo de empresários, para o transporte de pessoas. Esta concessão durou 40 anos. Os transportes coletivos nesta cidade remontam, no entanto, a 1872, ano em que a ”Companhia Carril Americano do Porto” foi iniciadora em Portugal da sua exploração.

Um ano depois, forma-se a ”Companhia Carris de Ferro do Porto”. Estas duas empresas concorrentes fundem-se, em 1893, numa só e decide-se manter o nome desta última - C.C.F.P.

Passam-se 13 anos, e é então que a concessão atrás referida é outorgada, surgindo, como consequência, e um ano mais tarde, (1907), a ”Companhia de Viação Eléctrica do Porto” que, tendo durado apenas 1 ano, não chega a dar início a qualquer atividade.

Em 1908 é absorvida pela ”Companhia Carris de Ferro do Porto”, que vem a beneficiar da concessão.
A ”Companhia Carris de Ferro do Porto” durou, com esta designação, 73 anos. Ainda hoje, quase 40 anos depois de ter desaparecido, há muita gente que, seguindo o velho hábito, ao referir-se à S.T.C.P. continua a chamar-lhe ”Carris”.

Em 1994 dá-se a passagem a sociedade anónima (de capitais exclusivamente públicos) passando a designar-se Sociedade de Transporte Colectivos do Porto, S.A.

Figura 1.1: Logotipo da STCP

Em 2012 completou 140 anos de história. É a maior empresa de transportes públicos coletivos de passageiros da Área Metropolitana do Porto, com 94 milhões de passageiros transportados por ano.

A S.T.C.P. tem como missão prestar um serviço de transporte público urbano de passageiros na Área Metropolitana do Porto (AMP), em articulação concertada com os demais operadores rodoviários, ferroviário e de metro ligeiro. Desta forma, contribui para a efetiva mobilidade das pessoas, disponibilizando uma alternativa competitiva ao transporte individual privado e gerando, pela sua atividade, benefícios sociais e ambientais num quadro de racionalidade económica e na busca da melhoria contínua do seu desempenho.

A S.T.C.P. presta serviços comerciais em seis concelhos: Porto, Gaia, Gondomar, Maia, Matosinhos, Valongo, e mais precisamente em 51 freguesias com cerca de 900 mil habitantes. Com um total de 70 linhas de autocarros, 59 na rede diurna e 11 na rede madrugada, e 3 linhas de elétrico. Na figura 1.2 pode ser visto o Mapa da Rede Geral da S.T.C.P.

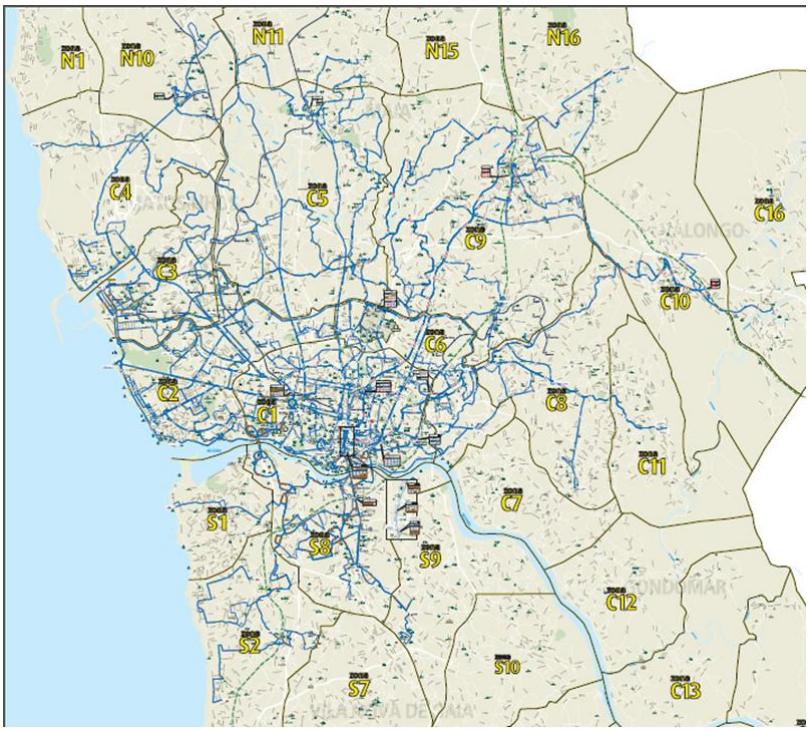


Figura 1.2: Mapa da Rede

A extensão da rede é de 485 quilómetros com 2458 paragens, 482 quilómetros de rede de autocarro e 9 quilómetros de rede de carro elétrico. Para a prestação destes serviços a S.T.C.P. possui 6 eléctricos e 475 autocarros, dos quais 254 são viaturas a gás natural e as restantes a gasóleo. Da frota dos autocarros há 4 tipos: standard, articulados, minis e 2 pisos. Os minis e 2 pisos apenas existem com o combustível gasóleo. Na figura 1.3 está representada a frota descrita anteriormente.

Figura 1.3: Frota

No ano de 2012 a empresa transportou 93 milhões de pessoas e conseguiu uma taxa média de ocupação por autocarro de 15,1%, transportando em média, por dia útil, 316 mil passageiros e percorreu 25,7 milhões de quilómetros. O serviço prestado pela S.T.C.P. no ano de 2012 correspondeu a uma poupança de cerca de 20.000 toneladas de CO_2 devido à tipologia da frota e às viaturas ligeiras que retira da cidade (STCP, 2012).

1.1.1 Processamento de Horários

O processamento dos horários é realizado no *software* GIST (Gestão Integrada de Sistema de Transportes). Este *software* é comercializado por uma outra empresa, a OPT, Optimização e Planeamento de Transportes, S.A., que tem como área nuclear de atividade a gestão operacional do transporte coletivo urbano.

O sistema GIST é um pacote de *software* com diversos módulos que permitem gerir a informação base relativa à rede de transportes, às linhas e às viagens a realizar. Permite ainda gerar, de uma forma otimizada, os horários das viaturas e dos motoristas, incluindo o escalamento diário destes últimos.

O carácter inovador do sistema GIST passa, essencialmente, pelo seu elevado nível de interatividade com o utilizador, pelas facilidades de parametrização, pela automatização de processos de planeamento e pela gestão integrada de dados que, em qualquer momento, proporciona, à gestão da empresa e ao público, informação atualizada e consistente. A normalização de dados é uma característica fundamental do sistema que permite, de uma forma uniforme e automática, a disponibilização ao público da informação sobre os percursos das linhas e sobre os seus horários.

Atualmente a versão completa do produto GIST (versão 2) é constituída por um conjunto de módulos, descritos pormenorizadamente abaixo, cada um deles com uma função claramente definida, auxiliando o utilizador em cada uma das fases usuais do processo do planeamento.

Em curso a OPT tem em desenvolvimento a nova versão do sistema. A versão GIST 3 integra os módulos de Rede e Megalinhas num só módulo - Rede 3, e os módulos de Viagens e Viaturas e Serviços noutro - Planeamento 3.

O processamento dos horários é realizado no módulo de planeamento e engloba algumas etapas prévias. As etapas principais são descritas de seguida e estão representadas na figura 1.4.

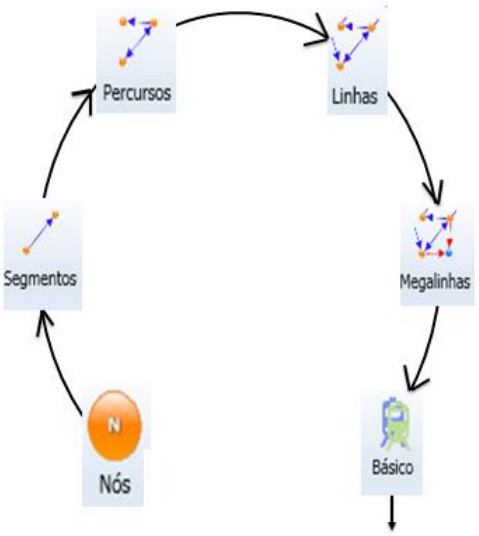


Figura 1.4: Etapas para o Processamento dos Horários

1. Nós: pontos de referência. Nesta etapa do processo, são definidos as coordenadas desses nós;

2. Segmentos: união de dois nós adjacentes. Nesta etapa é ainda definido o tempo de viagem entre nós;

3. Percursos: junção de segmentos;

4. Linhas: junção de percursos adicionando o sentido de ida e volta. No final desta etapa temos definido uma linha;

5. Megalinhas: são adicionados às linhas o tempo de vazio (tempo desde a estação de recolha até à primeira paragem) e ainda definidos alguns atributos de paragem como pontos de rendição (local de troca de motorista), estação de recolha, etc.;

6. Básico: geração automática de horários com alguns parâmetros de entrada, como a hora de inicio, hora do fim, frequência (as frequências são determinadas pela procura e pelo pico do dia), número de linha, tipo de dia, tempo de suporte, etc.;

No final destas etapas, o horário está criado no *software* GIST e o próximo passo é exportá-lo para *Excel*. São produzidos dois tipos de horários que não diferem no conteúdo e informação. Apenas diferem na forma de apresentação: o horário corrido (horário linha público) e o horário de serviço da viatura (horário do motorista).

1.1.2 Problemas e Objetivos

Cada vez mais é necessário prever os tempos de viagem com a maior assertividade possível. A concorrência é forte e se o autocarro não cumprir os tempos que estão nas paragens, corre-se o risco de os passageiros poderem usar outras empresas de transportes.

Desta forma, o objetivo deste trabalho divide-se em duas partes:

I. Construção de um relatório de avaliação da taxa de cumprimento do serviço (horário de passagem real nos pontos de horário definidos vs horário previsto de passagem), incluindo um tratamento estatístico de dados obtidos pelo Sistema de Apoio à Exploração e informação. .

II. Definição e desenvolvimento de um modelo que determine os tempos de percurso nas viagens, para um determinado nível de serviço pré-definido. Definição e construção de uma interface que atualize os tempos de percurso das viagens nas linhas.

As horas de passagem são registadas em tempo real no Sistema de Apoio à Exploração e informação (SAEi), como demonstrado na figura 1.5. No exemplo apresentado, podemos identificar uma discrepância entre os tempos previstos (linhas tracejadas) e os tempos efetuados (linhas a cheio). As cores correspondem a diferentes viagens, e observa-se que a previsão dos tempos de viagem deve ser melhorada. Para o cumprimento desta primeira etapa foi criado um relatório automático devolvendo estatísticas

do nível de cumprimento de serviço, a partir dados recolhidos pelo SAEi.

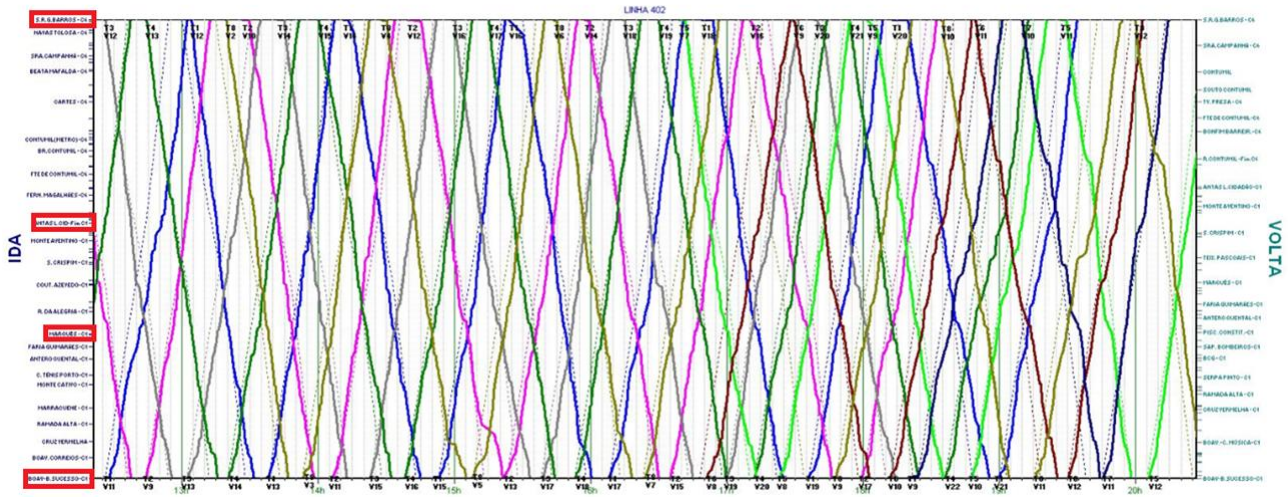


Figura 1.5: Imagem do SAEi

Para a resolução da segunda fase, foi necessário escolher as metodologias mais adequadas ao problema. Começou-se por considerar algumas metodologias empíricas que surgiram naturalmente com o contacto com os dados e a formação de todos os processos antecedentes à criação do horário.

Outra metodologia escolhida surgiu em vários documentos estudados. Segundo Yu et al. (2011), as Máquinas de Suporte Vetorial são o método mais eficaz para prever tempos de viagem. Os autores analisaram dados reais da rede de tráfego de Hong Kong, comparando quatro métodos de Data Mining, nomeadamente Máquinas de Suporte Vetorial, Redes Neuronais Artificiais, Algoritmo dos K-vizinhos mais próximos e Regressão Linear; as máquinas apresentaram uma precisão bastante mais elevada do que os restantes métodos.

A outra metodologia não foi encontrada tão comummente na literatura, estando porém naturalmente associada ao tipo de dados em causa. Pensando em dados longitudinais, o Método dos Mínimos Quadrados Generalizados ocorreu naturalmente como uma das metodologias a considerar. Os dados longitudinais surgem sempre que observações repetidas da variável resposta são obtidas ao longo do tempo, para cada indivíduo ou num ou mais grupos em estudo. Os dados longitudinais podem ser obtidos de uma forma prospetiva ou retrospetiva: na primeira, os indivíduos são seguidos ao longo do tempo e, na segunda, múltiplas medições em cada indivíduo são extraídas do seu historial (Cabral and Gonçalves, 2011).

No que diz respeito aos recursos computacionais, para a primeira parte foi usado *Visual Basic for Applications* (VBA), que é uma linguagem de programação baseada na conhecida linguagem *BASIC Beginner's All-purpose Symbolic Instruction Code*. Está concebida para funcionar em conjunto com diferentes aplicações, de forma a potenciar

a robustez das mesmas. Enquadra-se nos ambientes de programação baseados no processamento de sequência de eventos (*event-driven programming*). Foi inicialmente integrada com o *Excel* 5 em 1994 e, a partir daí, a sua expansão para outras aplicações foi gradual. Foi com a saída do *Office* 97, em 1997, que a *Microsoft* concretizou um dos seus grandes objetivos: ter um ambiente de programação completamente integrado nos seus quatro produtos mais famosos: *Word*, *Excel*, *Access* e *PowerPoint*. Atualmente, o VBA é já por si só um produto independente, que outras companhias podem adotar e incorporar nas suas aplicações (Walkenbach, 2010).

Na segunda parte foi usada a linguagem de programação e *software* de análises estatísticas R (R Core Team, 2014). Trata-se de uma linguagem funcional para computação estatística e que contém também funcionalidades gráficas. Pode ser vista como um dialeto da linguagem S (desenvolvido na AT & T), e foi o motivo pelo qual John Chambers foi premiado em 1998, mencionando que esta linguagem ”vai alterar para sempre a maneira como as pessoas analisam, visualizam e manipulam dados”. (Torgo, 2010)

1.2 Organização da Tese

Esta tese está organizada do seguinte modo: no **capítulo 2** é feita uma análise descritiva e exploratória dos dados.

No **capítulo 3** é proposta uma solução para a questão do relatório de cumprimento de serviço (um relatório automático gerado para qualquer linha e em qualquer período).

Nos **capítulos 4, 5 e 6** são descritas as metodologias estudadas para responder ao problema da previsão dos tempos de viagem. No capítulo 4 são introduzidas três metodologias empíricas; no capítulo 5 é explicado o Método Mínimos Quadrados Generalizados; e no capítulo 6 consideram-se Máquinas de Suporte Vetorial. Nestes três capítulos são descritos primeiramente os métodos e, de seguida, é apresentada a sua aplicação ao problema.

No **capítulo 7** são comparados os cinco métodos descritos nos capítulos 4, 5 e 6, e é selecionado o método que melhor responde ao problema.

Por fim, no **capítulo 8** é apresentado o Trabalho Futuro, que se for desenvolvido, facilitará a implementação prática destas metodologias.

Capítulo 2

Apresentação dos dados

A Base de Dados (BD) da empresa está em *Oracle* e o editor usado para aceder aos dados é o *Toad*. O Toad é um editor bastante intuitivo com boas ferramentas de trabalho e bastante desenvolvido. Os dados recolhidos e usados como exemplo ao longo desta tese estão descritos nos dois subcapítulos seguintes. A escolha desta linha como exemplo prendeu-se com o facto de ser a maior linha da STCP, ao nível dos nós, os nós desta linha são comuns a outras linhas (os 9 nós da linha em estudo são repetidos 15 vezes ao longo das restantes linhas). Como o processamento dos horários é realizado da mesma maneira para todas as linhas, a previsão para qualquer outra linha seria obtida de forma absolutamente análoga.

Neste capítulo é feita a análise descritiva e exploratória dos dados.

2.1 Análise Descritiva e Exploratória dos Dados

A linha selecionada como exemplo foi a linha 205 (Campanhã - Castelo do Queijo) apenas no sentido ida. O mapa da linha encontra-se na figura 2.1.

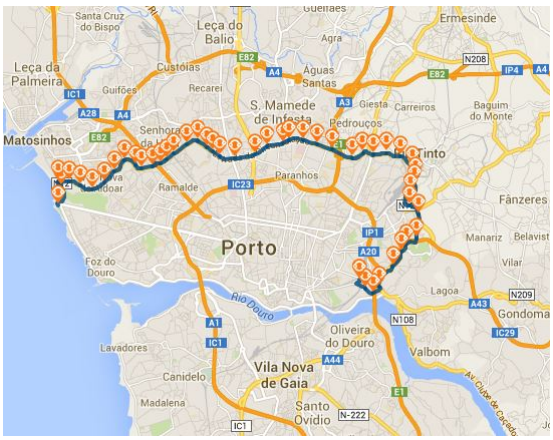


Figura 2.1: Mapa linha 205 (imagem retirada de www.stcp.pt)

Esta linha é composta por 53 paragens, 9 nós (8 segmentos) e uma extensão aproximada de 18,17 Km. A figura 2.2 e a tabela 2.1 descreve em pormenor as paragens e as

características dos segmentos.

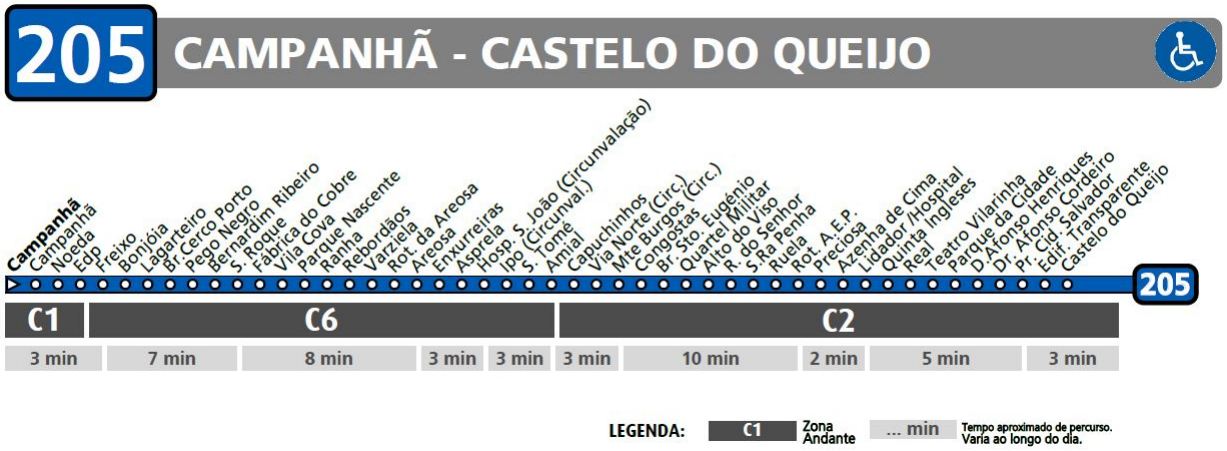


Figura 2.2: Sequência de paragens da linha 205 (imagem retirada de www.stcp.pt)

Devido a uma grande falta de registos na *BD* no que se refere ao primeiro e último nó de todas as linhas, que coincidem com a primeira e última paragem da linha, e pelo facto da última paragem de a ida ser a primeira viagem da volta e os registos nem sempre estarem corretos, considerou-se sempre que o primeiro nó (a primeira paragem) passa a ser a segunda paragem e o último nó (última paragem) a penúltima paragem, para garantir um número significativo de registos fidedignos para a análise.

No	Nº Paragens	Km	Tempo Programado (seg)	Semaforos	Corredor Bus	Rotundas
Campanhã - Freixo	4	1,03	225/135*	Não	Não	Não
Freixo - São Roque	7	3,33	480/420*	Sim	Não	Não
São Roque - Areosa	9	3,11	600/480*	Sim	Não	Sim
Areosa - Hospital São João	4	1,41	300	Sim	Não	Não
Hospital São João - Amial	4	1,42	180/120*	Sim	Não	Não
Amial - Monte Burgos	4	1,46	180	Sim	Não	Não
Monte Burgos - Rotunda AEP	9	2,59	480/600*	Sim	Não	Não
Rotunda AEP - Edifício Transparente	12	3,82	553/547*	Sim	Não	Sim

* Tempos previstos de algumas viagens no período da manhã ao Domingo

Tabela 2.1: Características dos Segmentos da Linha 205

Os dados considerados foram todos obtidos durante o mês de Outubro do ano 2013 (Horário Inverno). A escolha do período teve em conta ser um horário de inverno (maior frequência do que o horário de verão) e ser um mês sem nenhum tipo de operação especial e sem feriados.

Com o objetivo de obter uma previsão com o menor erro possível, foi feita uma partição dos 7 dias da semana e uma divisão de cada dia em vários períodos horários. Mais precisamente, consideraram-se dias úteis (DU), sábados (S) e domingos (D) e acrescentaram-se os períodos: 7-9 horas, 9-16 horas, 16-19 horas e o período das restantes horas. Ficamos assim com um subconjunto de 12 classes.

Os 12 subgrupos conduziram a um total de 17948 observações, não estando distribuídos homogeneamente. O total de motoristas é de 188, sendo em média necessários 36 motoristas para um dia útil e um pouco menos para os fins-de-semana. Os motoristas estão afetos a um grupo de linhas e diariamente é-lhes atribuída uma linha diferente

dentro desse grupo.

As variáveis usadas foram: o número do segmento, o período do dia, o tipo de dia, o tempo entre segmentos e a identificação do motorista. Na previsão não é usada a variável de identificação do motorista.

A tabela 2.2 contém uma breve estatística descritiva do conjunto de dados analisados.

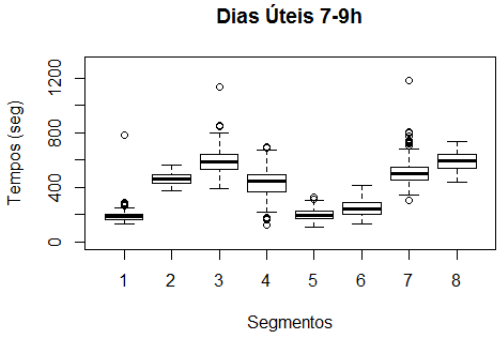
		Dias Úteis					Sábados					Domingos				
Turno	**Seg**	**Nº**	**Estatística Descritiva**				**Nº**	**Estatística Descritiva**				**Nº**	**Estatística Descritiva**			
		DU	Min[s]	Max[s]	Média[s]	DP[s]	S	Min[s]	Max[s]	Média[s]	DP[s]	D	Min[s]	Max[s]	Média[s]	DP[s]
7-9h	1	72	134	783	200.10	77.90	21	132	215	160.76	21.72	32	115	198	149.44	22.37
	2	84	372	562	462.55	40.64	16	391	550	462.94	36.20	32	294	492	399.25	41.04
	3	272	392	1131	593.87	91.58	52	340	664	496.23	69.04	32	374	556	470.91	43.10
	4	222	128	699	435.27	108.77	49	128	320	219.31	43.52	32	148	384	208.81	44.84
	5	269	113	330	203.07	40.97	50	103	247	154.30	31.71	32	102	193	143.16	26.35
	6	252	134	412	248.30	54.15	52	120	277	194.27	35.40	28	98	252	178.18	37.06
	7	238	301	1180	510.98	92.86	53	325	545	419.19	54.93	25	268	453	379.52	52.47
	8	142	434	739	590.72	63.13	47	425	606	504.11	46.63	23	364	573	452.57	55.30
9-16h	1	237	134	722	212.14	54.95	99	133	274	183.31	26.14	111	110	226	164.80	22.71
	2	242	312	584	437.24	37.81	93	345	549	439.14	47.78	106	294	562	412.97	44.97
	3	882	335	1089	561.06	87.30	181	428	865	586.82	85.64	135	348	793	543.96	68.79
	4	884	163	996	328.22	101.08	173	132	591	261.75	74.20	126	149	451	255.41	72.14
	5	933	107	636	206.40	47.13	186	100	421	186.55	38.41	136	99	304	177.66	39.20
	6	939	125	729	232.08	48.00	186	132	332	228.78	39.63	132	139	513	225.05	57.15
	7	943	308	715	477.83	60.82	182	303	680	482.57	65.55	131	287	711	459.41	66.65
	8	431	357	724	561.71	60.81	178	391	677	539.25	59.26	114	352	675	522.48	62.31
16-19h	1	124	137	869	241.64	92.94	53	127	295	175.98	34.36	44	128	231	174.14	24.99
	2	134	337	526	435.45	39.27	44	351	504	420.00	38.76	40	338	487	407.23	31.90
	3	399	53	928	567.23	95.64	86	428	1029	597.40	93.87	77	419	993	589.44	112.12
	4	424	144	633	326.62	80.07	93	158	719	270.18	91.38	81	148	460	294.94	65.97
	5	419	114	1242	282.23	146.88	93	99	259	178.16	31.11	84	120	321	187.46	41.79
	6	415	141	582	282.14	61.94	93	138	380	234.17	40.47	84	142	625	236.70	83.16
	7	416	349	866	503.21	67.57	90	291	956	493.69	92.28	42	307	603	439.71	62.96
	8	234	423	704	576.45	55.10	92	304	695	532.32	66.98	89	348	715	519.93	65.39
outras h	1	260	106	303	162.61	34.45	82	100	225	153.17	24.74	77	94	263	147.90	25.85
	2	258	261	601	372.26	52.06	79	276	451	356.51	44.26	73	277	456	344.75	41.38
	3	373	274	884	472.31	76.56	121	282	741	475.27	75.84	87	312	646	454.60	72.86
	4	381	109	488	238.65	73.06	105	138	284	209.87	35.41	94	128	357	219.00	47.41
	5	403	86	607	183.81	68.46	128	72	236	149.62	33.32	95	92	231	162.47	28.90
	6	392	98	527	223.75	74.42	106	128	361	201.00	39.43	95	114	281	181.54	30.98
	7	375	257	1032	432.54	105.34	107	238	544	391.18	63.46	96	267	498	369.51	57.34
	8	320	309	729	458.68	70.57	106	313	598	445.32	66.01	98	239	652	409.00	67.51

Tabela 2.2: Tamanhos amostrais para cada segmento e estatística descritiva para os dados analisados

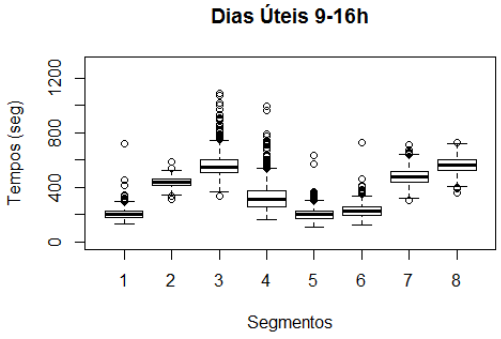
São notórias as diferenças que existem entre os dias e períodos do dia. Por exemplo, o segmento 5 (troço Hospital S. João - Amial) no período 9-16 horas tem um horário previsto de 180 segundos (Apêndice B) e o tempo real é em média de 206 segundos nos dias úteis, 187 segundos aos sábados e 178 segundos aos domingos, sendo que o seu desvio padrão é de 47 segundos, 38 segundos e 39 segundos, respetivamente (tabela 2.1).

Já no que diz respeito a variações entre os diferentes períodos do dia podemos ver, por exemplo, no segmento 3 (São Roque - Areosa) que o tempo previsto de 600 segundos (Apêndice B) é igual para todas as horas do dia, o tempo real é em média aproximadamente de 594 segundos no período das 7-9 horas com um valor máximo de 1131 segundos. Nos outros períodos horários é em média 472 segundos com um valor máximo de 884 segundos (tabela 2.1).

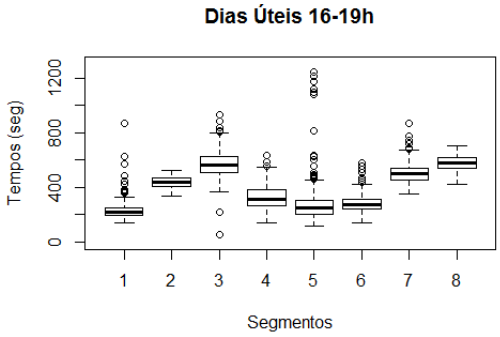
De seguida são apresentados os doze Diagramas de Caixa e Bigodes (*boxplots*) que correspondem às combinações dos períodos e do tipo de dias.



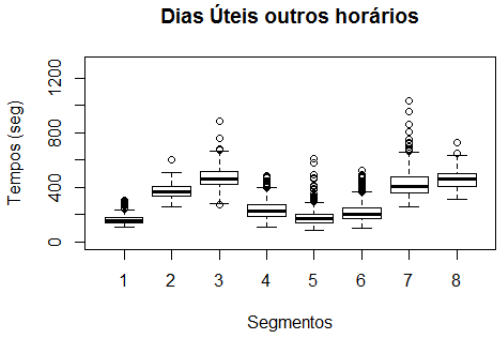
(a)



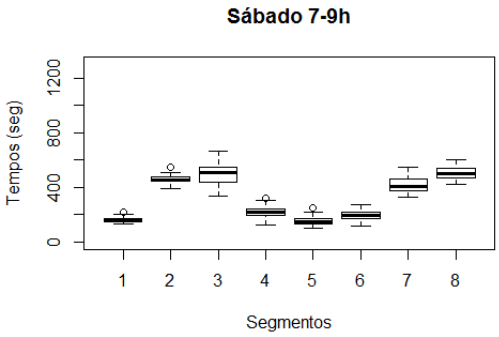
(b)



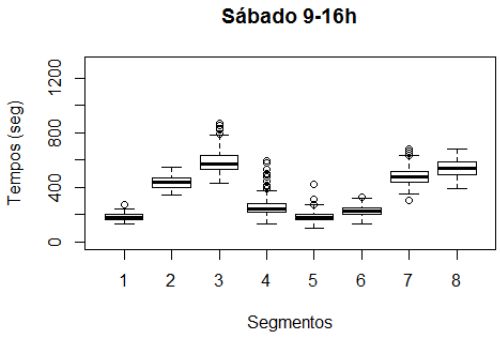
(c)



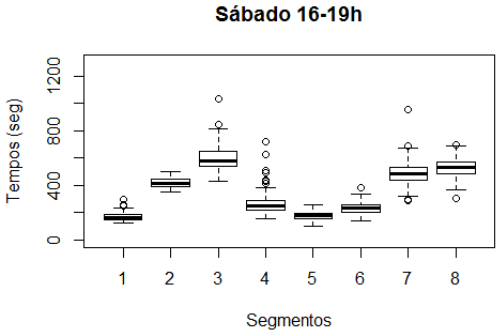
(d)


Sábado 7-9h
Tempos (seg)
0
400
800
1200
1
2
3
4
5
6
7
8
Segmentos

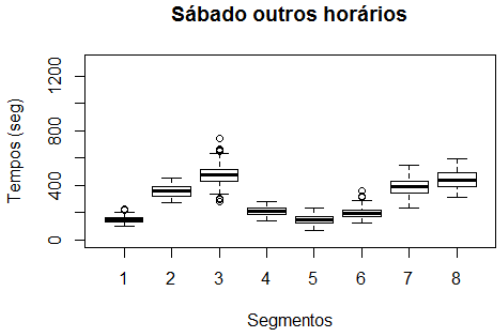
(e)


Sábado 9-16h
Tempos (seg)
0
400
800
1200
1
2
3
4
5
6
7
8
Segmentos

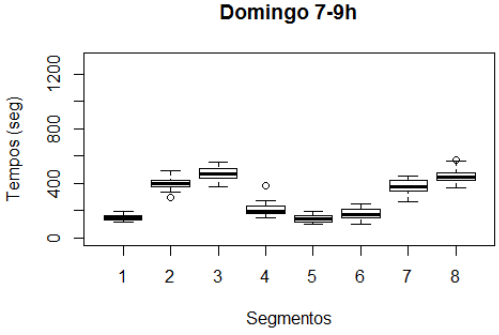
(f)


Sábado 16-19h
Tempos (seg)
0
400
800
1200
1
2
3
4
5
6
7
8
Segmentos

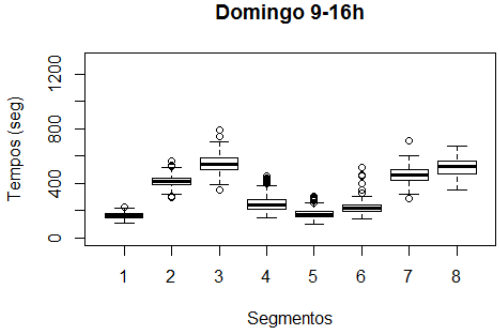
(g)


Sábado outros horários
Tempos (seg)
0
400
800
1200
1
2
3
4
5
6
7
8
Segmentos

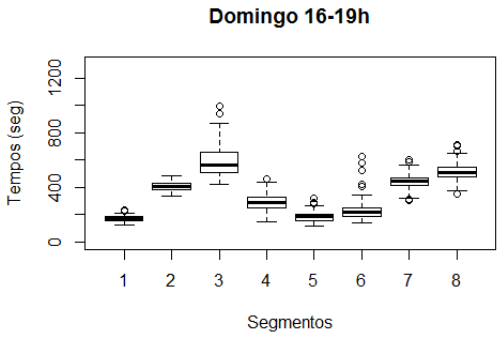
(h)


Domingo 7-9h
Tempos (seg)
0
400
800
1200
1
2
3
4
5
6
7
8
Segmentos

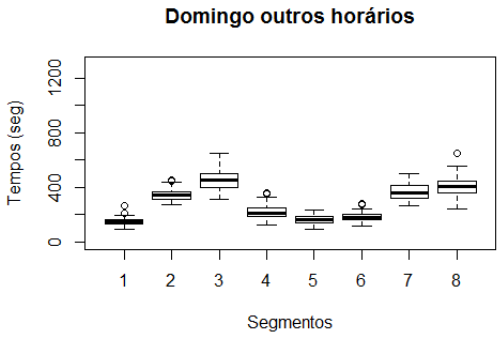
(i)


Domingo 9-16h
Tempos (seg)
0
400
800
1200
1
2
3
4
5
6
7
8
Segmentos

(j)



(k)



(l)

Figura 2.3: Boxplot (a) Dias Úteis 7-9h (b) Dias Úteis 9-16h (c) Dias Úteis 16-19h (d) Dias Úteis outros horários (e) Sábado 7-9h (f) Sábado 9-16h (g) Sábado 16-19h (h) Sábado outros horários (i) Domingo 7-9h (j) Domingo 9-16h (k) Domingo 16-19h (l) Domingo outros horários

Na figura 2.3 é visível a presença de alguns *outliers*, uns moderados e outros severos. Isto pode dever-se à ocorrência de acidentes ou a uma maior afluência de trânsito. Os *outliers* não foram retirados da amostra porque são uma constatação e, não havendo controlo da realidade, é importante que a previsão os tenha em conta porque eles poderão surgir em qualquer dia. A presença de *outliers* severos ocorre apenas nos dias úteis, e com maior frequência nos segmentos centrais, enquanto que aos sábados e domingos apenas encontramos *outliers* moderados, numa escala bastante menos significativa. O tamanho amostral (tabela 2.2) é bastante mais reduzido aos fins-de-semana por haver menos frequência de autocarros.

Quase todas as sub-amostras obtidas das combinações dos dias de semana mais o período horário apresentam simetria, sendo que também têm tamanhos amostrais diferentes (tabela 2.2).

O gráfico (figura 2.4) foi obtido a partir da função *groupedData()* do *package nlme* (Pinheiro et al., 2014) do *software* R (R Core Team, 2014). Um objeto *groupedData* agrupa as observações em grupos distintos de observações através de um fator de agrupamento (Pinheiro and Bates, 2000).

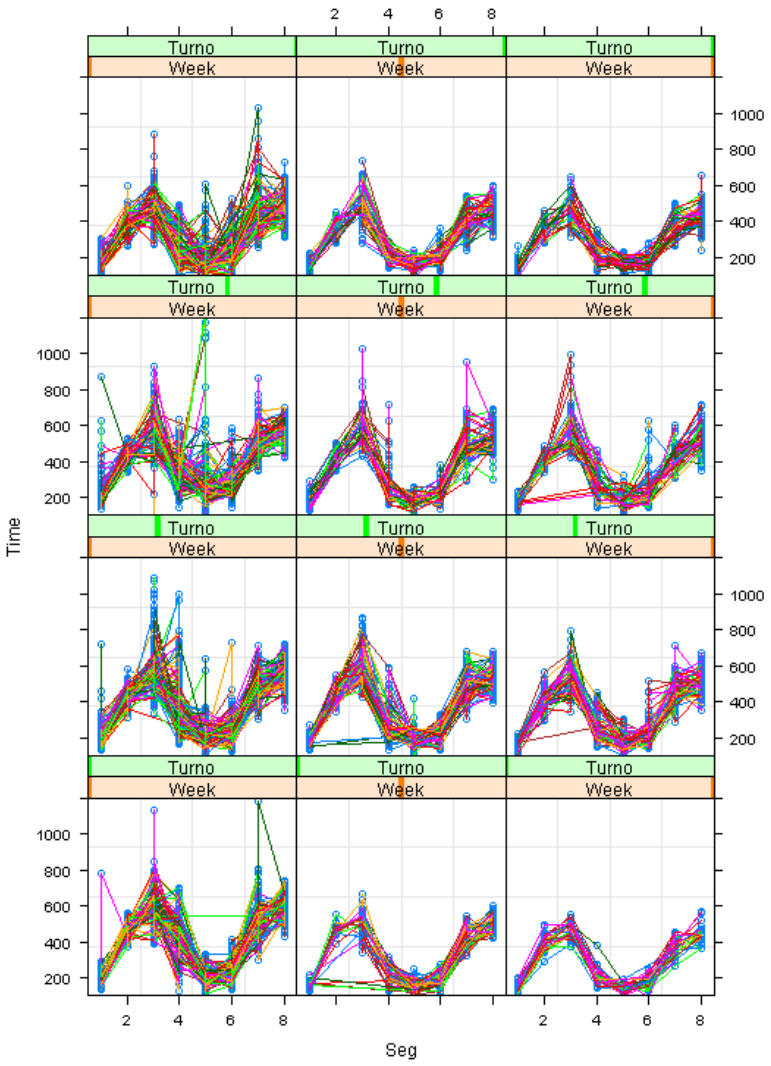


Figura 2.4: Gráfico de perfis (cada perfil corresponde a um motorista por turno, horário e semana)

O gráfico da figura 2.4, os dados foram agrupados por condutor do autocarro (grupo). Novamente estão representados 12 gráficos, o tipo de dia varia da esquerda para a direita na sequência dia útil, sábado e domingo, e o período do dia varia de baixo para cima na sequência 7-9h, 9-16h, 16-19h e outros horários. No eixo horizontal estão representados os segmentos e no vertical o tempo (em segundos).

As linhas representam o fator de agrupamento dos dados, ou seja, as viagens feitas por cada motorista. Algumas viagens não passam por todos os segmentos; isto pode ocorrer por falha no registo na *BD* ou por o motorista ter feito a rendição do motorista anterior nesse momento. Com números diferentes de viagens o padrão apresentado em todos os gráficos é semelhante. Nos dias úteis o desvio entre os tempos de viagem é bastante superior aos restantes dias.

Capítulo 3

Relatório de Avaliação da Taxa de Cumprimento de Serviço

Neste capítulo é apresentada uma ferramenta, que acrescenta grande informação e de fácil utilização e tem um tempo reduzido de execução, que descreve os atrasos e adiantamentos das viagens consideradas na BD.

3.1 Relatório Atual

O relatório atual de cumprimento de serviço é gerado na intranet da empresa (Primavera) onde se pode escolher a linha, data de início e fim, hora de início e fim, turno (um ou mais), tipo de dia (a escolha é feita entre dias úteis, sábados e domingos, só se podendo escolher um) e intervalo (15 minutos em 15 minutos, 30 minutos em 30 minutos ou 60 minutos em 60 minutos). O processo é um pouco lento, e gera um gráfico de viagem com uma média ponderada de viagens, o que acrescenta pouca informação descritiva de atrasos e adiantamentos.

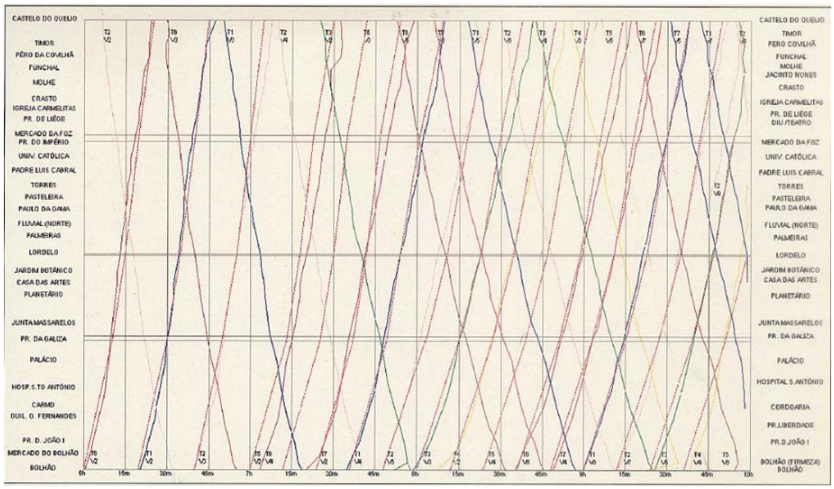


Figura 3.1: Relatório Disponível na S.T.C.P.

A figura 3.1 mostra um exemplo do relatório disponível para a linha 200 (Bolhão – Castelo Queijo) para as escolhas: data início: 31-03-2008 e data fim: 20-04-2008, tipo de dia: útil, turno: 1,2,3,4,5,6,7,8, hora início: 06 e hora fim: 09.

3.2 Solução Proposta

De maneira a acrescentar informação descritiva relativa a atrasos e adiantamentos foi criado em Excel®, com o auxilio de Macros e código VBA (*Visual Basic for Applications*), uma aplicação para tratamento de dados (figura 3.2).

Figura 3.2: Solução Proposta

O modo de funcionamento é simples e intuitivo. A aplicação tem apenas três botões, o INICIAR, LIMPAR e GUARDAR PDF, e está acompanhada de um pequeno manual de instruções no canto inferior esquerdo (figura 3.2).

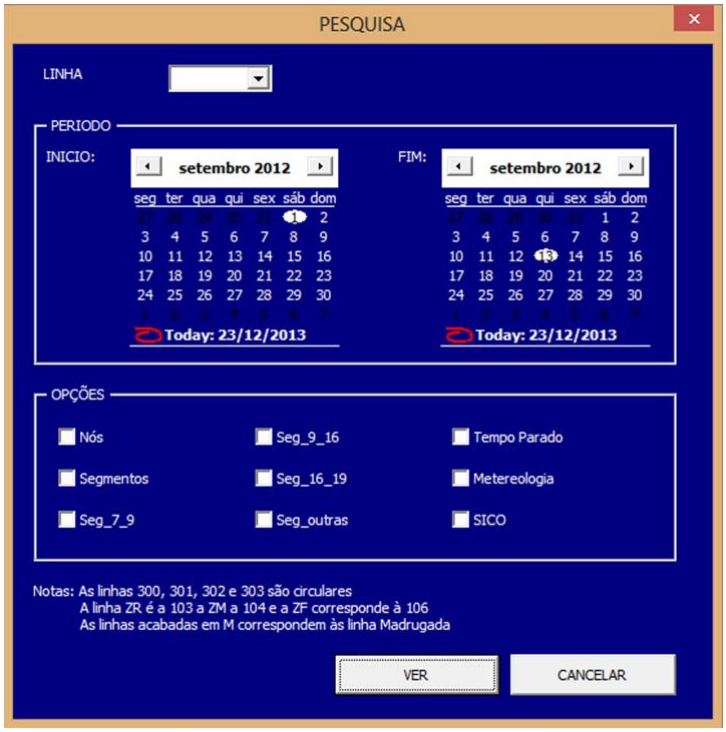


Figura 3.3: Formulário da Solução Proposta

No formulário, visível na figura 3.3, o utilizador encontra uma *Combobox* com todas as linhas da S.T.C.P. e dois calendários, um respeitante à data de início e o outro à data de fim. Tem ainda à escolha 9 opções como *check box*, sendo que deve escolher pelo menos uma. As opções são: Nós, Segmentos, Seg_7_9, Seg_9_16, Seg_16_19, Seg_outras, Tempo Parado, Metereologia e SICO.

Na opção *Nós* é feita uma análise estatística e descritiva de todos os nós da viagem e o resultado apresentado consta de três tabelas e um gráfico. A primeira tabela analisa todos os nós e divide-os em duas classes: atrasados e adiantados. Em cada uma destas duas classes são definidos 7 intervalos: 0-1m, 1-2m, 2-3m, 3-4m, 4-5m, +5m, +10m. O objetivo desta primeira tabela é identificar a percentagem de nós em cada uma das classes de atraso ou adiantamento, o gráfico é referente a esta tabela. Imediatamente abaixo desta tabela aparece um indicador. Este indicador avalia a percentagem de chegadas à hora. Por definição as chegadas à hora são todas aquelas que cheguem com um atraso igual ou inferior a 5 minutos (tabela 3.1, TRB (2000)). Neste trabalho foram ainda considerados todos os adiantamentos inferiores a 1 minuto para o indicador.

Nível de Serviço	Percentagem de chegadas à hora
A	97.5 - 100
B	95.0 - 97.4
C	90.0 - 94.9
D	85.0 - 89.9
E	80.0 - 84.9
F	$<$ 80.0

Tabela 3.1: Indicador da Percentagem de Serviço Cumprido (TRB, 2000)

A letra correspondente ao indicador aparece sublinhada e a cores, com a seguinte correspondência: A,B cor verde; C,D cor amarela; E,F cor vermelha. A segunda tabela da opção *Nós* apresenta uma análise apenas do primeiro e último nós de cada sentido; o objetivo desta tabela é perceber se quando um autocarro parte atrasado/adiantado mantém no final o atraso/adiantamento. A última tabela desta folha contém a legenda das siglas usadas.

A opção *Segmentos* faz uma análise, não por nós como na primeira, mas por segmentos. Para n paragens há n-1 segmentos. A análise é feita usando as mesmas classes e intervalos da primeira tabela da folha anterior. A novidade desta tabela é a classificação dos dias da semana. Na folha encontram-se 4 tabelas: a primeira diz respeito a todos os dias, a segunda corresponde aos dias úteis, a terceira aos sábados e a última aos domingos. A primeira tabela é, naturalmente a soma das três seguintes.

As opções *Seg_7_9*, *Seg_9_16*, *Seg_16_19* e *Seg_Outras* são análogas entre si e parecidas com a folha *Segmentos*; a diferença é que estas estão divididas em períodos horários.

Há ainda um resultado *Gráficos* que está sempre ligado a uma das análises por segmento; não faz parte das opções porque não pode ser escolhida isoladamente. Aqui podemos encontrar 3 diagramas circulares respeitantes a valores gerais e 12 histogramas correspondentes a análises por segmentos em diferentes horas e dias de semana. É de notar que as linhas 300, 301, 302 e 303 são circulares logo só é preenchido um sentido. E ainda que o número de nós das linhas varia entre 2 e 9.

A opção *Tempo_Parado* calcula o tempo mínimo, o tempo máximo, a moda, a média e a variância do tempo parado em cada paragem, por tipo de dia e horário. O objetivo desta tabela é identificar em que paragem o autocarro está parado durante mais tempo.

A opção *Meteo* indica as condições climatéricas no período escolhido, nomeadamente, a temperatura mínima, média e máxima, estação do ano e pluviosidade. O objetivo desta folha é o de transmitir ao utilizador tempos de viagem para diferentes condições climatéricas.

Por fim a folha *SICO* (Sistema de Informação de Controlo Operacional) exporta os dados de uma tabela já existente. São transcritos para esta folha os dados contidos que se re-

ferem a condições de tráfego. O objetivo é o de consolidar toda a informação importante num único sítio e identificar atrasos devido a trânsito congestionado ou acidentes.

3.2.1 Exemplo

A título de exemplo do funcionamento da aplicação desenvolvida escolheu-se a linha número 205, entre o dia 01.10.2013 e o dia 01.11.2013 com as 9 opções selecionadas.

A figura 3.4 ilustra a folha *Nós* preenchida de acordo com as seleções efetuadas.

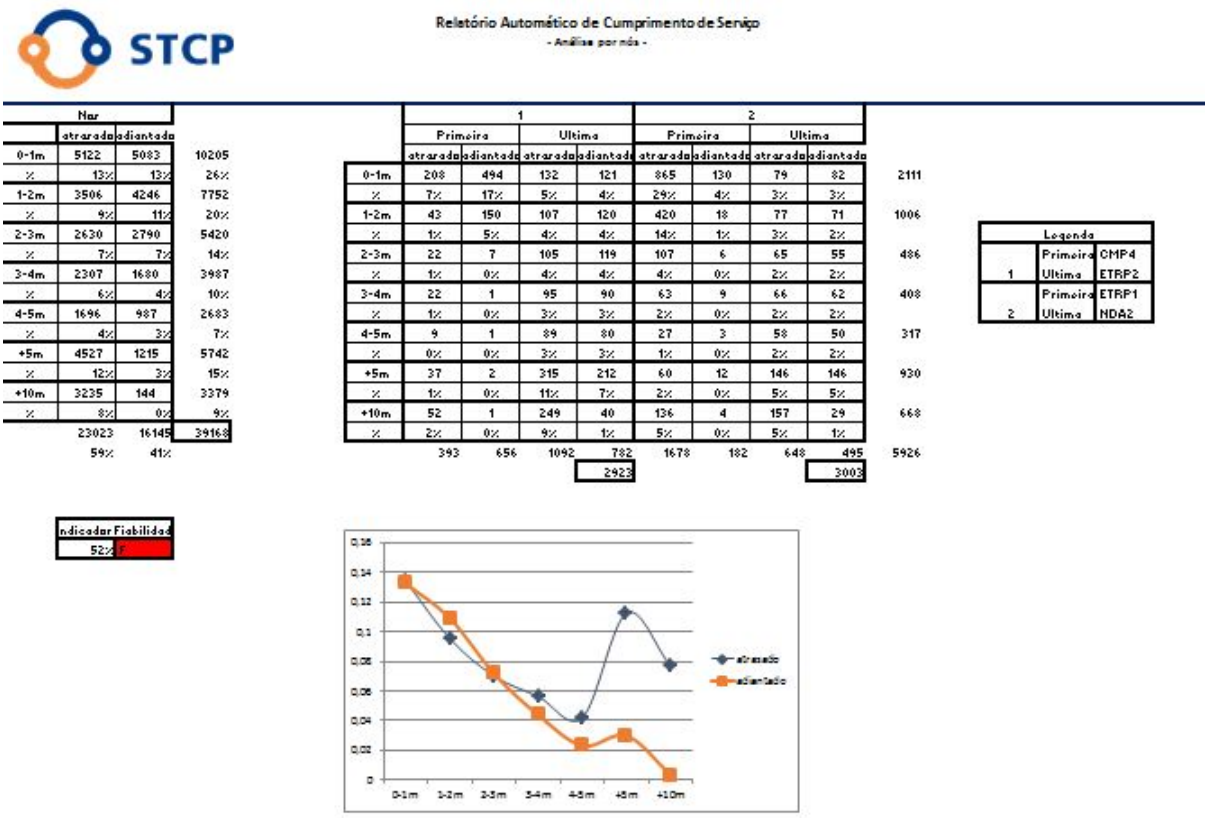


Figura 3.4: Exemplo folha NOS

Podemos verificar que 59% dos autocarros da linha 205 no período indicado chegam atrasados aos nós e que 41% chegam adiantados. Como se considera como cumprimento de serviço os atrasos até 5 minutos e os adiantamentos até 1 minuto, é atribuída uma percentagem de cumprimento de serviço de 52%, classificado com a letra *F*. Ainda se pode constatar que, para o sentido de ida, 13% dos autocarros partem atrasados do primeiro nó, e 39% chegam atrasados último nó. Isto mostra que não são só os autocarros que partem atrasados que chegam atrasados; alguns dos que partem adiantados perdem esse avanço e chegam também atrasados. Este pode ser um indicador que o tempo de viagem é curto para este segmento.

As quatro tabelas seguintes, as tabelas das figuras 3.5, 3.6, 3.7, 3.8, dizem respeito ao preenchimento da folha *Segmentos*. Não são apresentadas tabelas das outras folhas de segmentos, que estão divididas em períodos horários mais restritos, apesar de esta análise ser importante pois os dias têm diferente procura assim como fluxo de trânsito. O processamento dos dados é análogo, naturalmente.

	1															
	A		B		C		D		E		F		G		H	
	atrasado	adiantad	atrasado	adiantad	atrasado	adiantad	atrasado	adiantad	atrasado	adiantad	atrasado	adiantad	atrasado	adiantad	atrasado	adiantado
0-1m	198	485	307	500	417	697	556	566	811	286	962	213	638	386	500	495
%	1%	1%	1%	1%	1%	2%	2%	2%	2%	1%	3%	1%	2%	1%	1%	1%
1-2m	31	258	110	352	177	658	585	395	887	51	817	0	559	232	258	249
%	0%	1%	0%	1%	0%	2%	2%	1%	2%	0%	2%	0%	2%	1%	1%	1%
2-3m	29	4	68	149	57	308	292	57	341	0	344	0	209	226	62	93
%	0%	0%	0%	0%	0%	1%	1%	0%	1%	0%	1%	0%	1%	1%	0%	0%
3-4m	9	3	49	104	29	66	149	1	105	0	140	0	63	113	23	28
%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
4-5m	4	5	47	89	13	10	50	1	33	0	48	0	38	27	15	7
%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
+5m	5	4	132	342	83	13	54	1	31	0	23	2	61	7	92	24
%	0%	0%	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
+10m	32	6	35	218	3	5	7	1	9	1	1	4	4	0	19	9
%	0%	0%	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0,0001	0	0%	0%
	308	765	748	1754	779	1757	1693	1022	2217	338	2335	219	1572	991	969	905

Figura 3.5: Exemplo de uma folha SEGMENTOS I

Dias Uteis	1															
	A		B		C		D		E		F		G		H	
	atrasado	adiantad	atrasado	adiantad	atrasado	adiantad	atrasado	adiantad	atrasado	adiantad	atrasado	adiantad	atrasado	adiantad	atrasado	adiantado
0-1m	150	368	253	401	340	592	494	463	615	224	751	154	631	358	409	366
%	1%	1%	1%	1%	1%	2%	2%	2%	2%	1%	3%	1%	2%	1%	1%	1%
1-2m	28	162	81	262	142	546	497	269	777	35	703	0	556	144	223	169
%	0%	1%	0%	1%	0%	2%	2%	1%	3%	0%	2%	0%	2%	0%	1%	1%
2-3m	23	3	60	124	52	268	263	32	312	0	319	0	209	75	55	54
%	0%	0%	0%	0%	0%	1%	1%	0%	1%	0%	1%	0%	1%	0%	0%	0%
3-4m	9	2	40	90	23	55	139	0	100	0	136	0	63	22	20	16
%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0,0021	0%	0%	0%
4-5m	3	5	39	87	12	7	48	1	33	0	48	0	34	0	10	5
%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
+5m	5	2	108	320	66	13	53	1	31	0	20	2	48	0	74	16
%	0%	0%	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0,0016	0%	0%	0%
+10m	24	6	32	193	3	4	5	1	8	1	0	4	4	0	16	6
%	0%	0%	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0,0001	0	0,0005	0%
	242	548	613	1477	638	1485	1499	767	1876	260	1977	160	1545	599	807	632

Figura 3.6: Exemplo de uma folha SEGMENTOS II

Sábados	1															
	A		B		C		D		E		F		G		H	
	atrasado	adiantad	atrasado	adiantad	atrasado	adiantad	atrasado	adiantad	atrasado	adiantad	atrasado	adiantad	atrasado	adiantad	atrasado	adiantado
0-1m	5	58	27	46	43	63	22	63	105	28	107	26	6	18	57	75
%	0%	2%	1%	1%	1%	2%	1%	2%	3%	1%	3%	1%	0%	1%	2%	2%
1-2m	1	61	10	52	9	60	39	85	67	10	73	0	3	55	20	36
%	0%	2%	0%	2%	0%	2%	1%	3%	2%	0%	2%	0%	0,0009	2%	1%	1%
2-3m	3	0	6	15	2	23	10	13	14	0	14	0	0	73	5	14
%	0%	0%	0%	0%	0%	1%	0%	0%	0%	0%	0%	0%	0	2%	0%	0%
3-4m	0	1	2	8	2	10	6	1	1	0	3	0	0	50	2	6
%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0	1%	0%	0%
4-5m	0	0	5	2	1	0	1	0	0	0	0	0	1	9	2	1
%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0,0003	0%	0%	0%
+5m	0	1	9	20	8	0	1	0	0	0	0	0	7	3	11	1
%	0%	0%	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0,0021	0%	0%	0%
+10m	4	0	1	19	0	1	0	0	1	0	0	0	0	0	0	2
%	0%	0%	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0	0	0	0%
	13	121	60	162	65	157	79	162	188	38	197	26	17	208	97	135

Figura 3.7: Exemplo de uma folha SEGMENTOS III

Domingos	1															
	A		B		C		D		E		F		G		H	
	atrasado	adiantad	atrasado	adiantad	atrasado	adiantad	atrasado	adiantad	atrasado	adiantad	atrasado	adiantad	atrasado	adiantad	atrasado	adiantado
0-1m	43	59	27	53	34	42	40	40	91	34	104	33	1	10	34	54
%	1%	2%	1%	2%	1%	1%	1%	1%	3%	1%	4%	1%	0%	0%	1%	2%
1-2m	2	35	19	38	26	52	49	41	43	6	41	0	0	33	15	44
%	0%	1%	1%	1%	1%	2%	2%	1%	1%	0%	1%	0%	0	1%	1%	1%
2-3m	3	1	2	10	3	17	19	12	15	0	11	0	0	78	2	25
%	0%	0%	0%	0%	0%	1%	1%	0%	1%	0%	0%	0%	0	3%	0%	1%
3-4m	0	0	7	6	4	1	4	0	4	0	1	0	0	41	1	6
%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	0%	0%
4-5m	1	0	3	0	0	3	1	0	0	0	0	0	3	18	3	1
%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0,001	1%	0%	0%
+5m	0	1	15	2	9	0	0	0	0	0	3	0	6	4	7	7
%	0%	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
+10m	4	0	2	6	0	0	2	0	0	0	1	0	0	0	3	1
%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0	0	0%	0%
	53	96	75	115	76	115	115	93	153	40	161	33	10	184	65	138

Figura 3.8: Exemplo de uma folha SEGMENTOS IV

Analisando a figura 3.6 podemos ver que até ao segmento *D* a maioria das viaturas demora menos tempo a percorrer estes segmentos do que o previsto. Em *D* e *E* o cenário altera-se, o que sugere que de *D* para *E* o tempo de viagem deve ser superior ao previsto. É importante relembrar que na análise por segmentos o importante não é se a viatura chega atrasada ou adiantada mas se o tempo de viagem é superior ou inferior ao previsto. A tabela da figura 3.9 é a legenda dos segmentos.

Legenda					
	A	de	CMP4	para	FRX2
	B	de	FRX2	para	SR5
	C	de	SR5	para	ARS5
	D	de	ARS5	para	HSJ12
	E	de	HSJ12	para	AML3
	F	de	AML3	para	MTB2
	G	de	MTB2	para	RAEP3
1	H	de	RAEP3	para	ETRP2

Figura 3.9: Exemplo de uma folha SEGMENTOS V

Os gráficos das figuras 3.10 e 3.11 fornecem dados complementares aos das tabelas, sendo importantes para dar um panorama geral por período horário e tipo de dia devido

à grande dimensão de algumas tabelas.

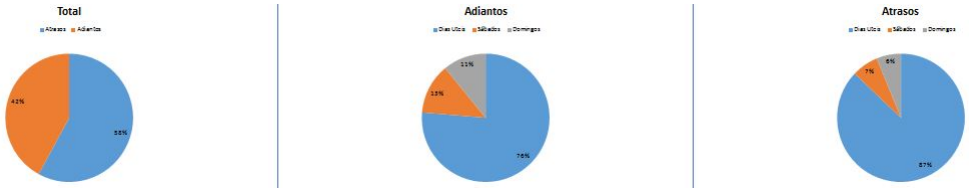


Figura 3.10: Exemplo de uma folha GRAFICOS I

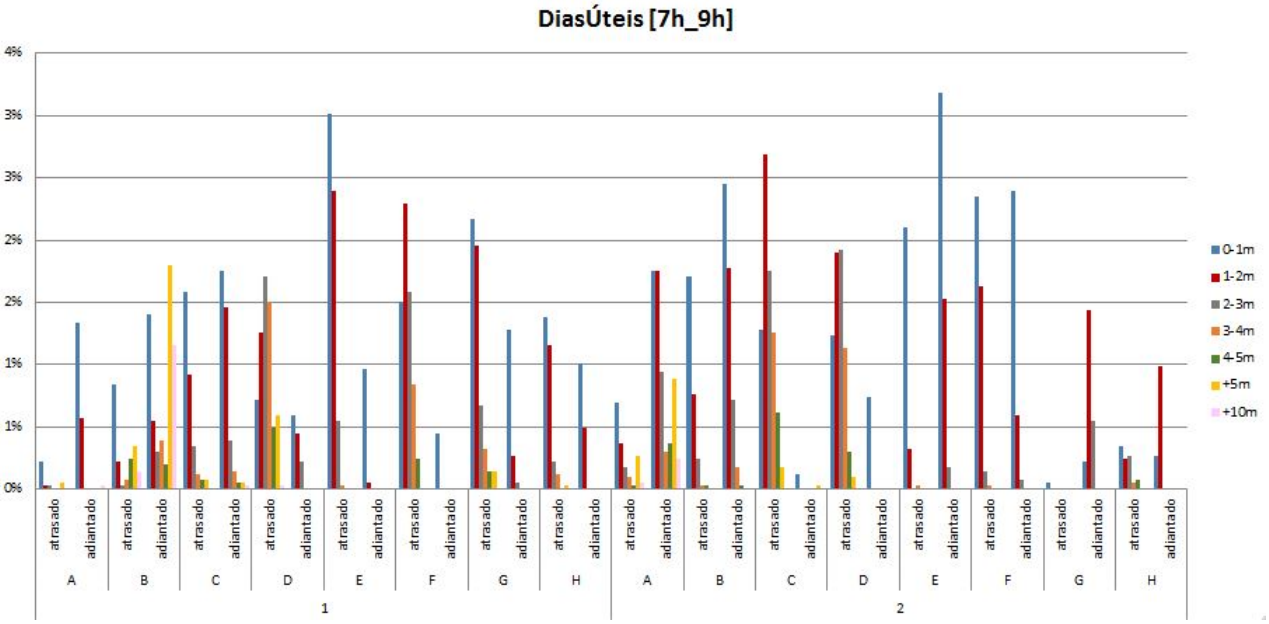


Figura 3.11: Exemplo de uma folha GRAFICOS II

A tabela da figura 3.12 indica o tempo parado na paragem, para a entrada e saída de passageiros e para a cobrança da bilhética.

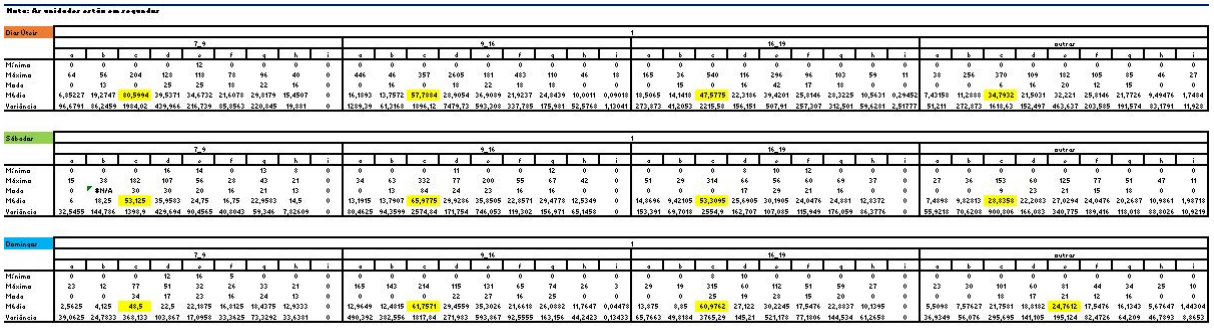
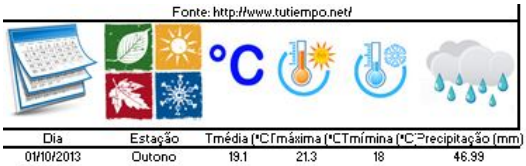


Figura 3.12: Exemplo de uma folha TEMPO PARADO

É de notar que a paragem *C* é a paragem em que o autocarro mais tempo está parado independentemente do dia de semana e divisão horária, o que permite deduzir que é uma paragem com grande afluência de passageiros.

A tabela da figura 3.13 contém a informação climatérica do período escolhido.

Fonte: http://www.tutiempo.net/



Dia	Estação	Tmédia (ºC)	Tmáxima (ºC)	Tmínima (ºC)	Precipitação (mm)
01/10/2013	Outono	19.1	21.3	18	46.99
02/10/2013	Outono	19.2	24	16	-
03/10/2013	Outono	19.3	21.6	18	25.91
04/10/2013	Outono	18.1	21	15	0
05/10/2013	Outono	15.8	20	13	0
06/10/2013	Outono	17.7	26	11	0
07/10/2013	Outono	18.9	26	13	0
08/10/2013	Outono	17.9	26	11	0
09/10/2013	Outono	20.2	27	14	0
10/10/2013	Outono	18.2	25.3	13	0
11/10/2013	Outono	15	19.7	10	0
12/10/2013	Outono	15.8	19.4	12	-
13/10/2013	Outono	16.1	18	12	7.11
14/10/2013	Outono	18	20	16.7	8.89
15/10/2013	Outono	18.8	20	17.6	4.06
16/10/2013	Outono	18.7	19.3	17.9	5.33
17/10/2013	Outono	18.9	21.1	17	4.06
18/10/2013	Outono	18.7	20.9	16.6	7.11
19/10/2013	Outono	18.1	20	16.5	10.92
20/10/2013	Outono	18.2	20.6	16.4	7.87
21/10/2013	Outono	18.8	21.1	16	6.1
22/10/2013	Outono	18.4	20	16.2	41.91
23/10/2013	Outono	18.6	21.8	17	3.3
24/10/2013	Outono	18.5	19.2	17.4	36.07
25/10/2013	Outono	17.5	20	15	1.02
26/10/2013	Outono	15.9	20	13	0.76
27/10/2013	Outono	16.2	20.2	13.5	0
28/10/2013	Outono	15.8	17.4	11	29.97
29/10/2013	Outono	12.7	17.1	8	2.03
30/10/2013	Outono	11.6	18	6	0
31/10/2013	Outono	12.5	19	6	0

Figura 3.13: Exemplo de uma folha METEO

Em relação à tabela da figura 3.13 não há nada de relevante a comentar. Esta tabela acrescenta informações cruciais porque uma viagem feita com chuva ou com sol tem períodos de viagem diferentes e por vezes é necessário saber as condições climatéricas para explicar eventuais diferenças bruscas nos atrasos/adiantamentos.

A tabela da figura 3.14 contém toda a informação disponível de condições de trânsito e sinistros.

ID_REGISTO	LINHA_COD	LINHA_DESCR	TURNO	MG	VIATURA	VIAG_PERDIDAS	HORA	DECISAO_COD	DECISAO_DESCR
215184	205	205 - CAMPANHÃ-CASTELO DO QUEIJC	12			2	01/10/2013	14	Transferido
215155	205	205 - CAMPANHÃ-CASTELO DO QUEIJC	6		1110	0,5	01/10/2013	5	Fora de Serviço até Paragem
215165	205	205 - CAMPANHÃ-CASTELO DO QUEIJC	9	11772		0,5	01/10/2013	5	Fora de Serviço até Paragem
215166	205	205 - CAMPANHÃ-CASTELO DO QUEIJC	10	11383	1125	1	01/10/2013	5	Fora de Serviço até Paragem
215181	205	205 - CAMPANHÃ-CASTELO DO QUEIJC	7	11222		9	01/10/2013	8	Recolheu
215197	205	205 - CAMPANHÃ-CASTELO DO QUEIJC	4			1	01/10/2013	1	Faz Viagem até Paragem
215281	205	205 - CAMPANHÃ-CASTELO DO QUEIJC	13			5	01/10/2013	-1	
215385	205	205 - CAMPANHÃ-CASTELO DO QUEIJC	12	11879		4	02/10/2013	14	Transferido
215432	205	205 - CAMPANHÃ-CASTELO DO QUEIJC	9	11396	1114	0	02/10/2013	-1	
215465	205	205 - CAMPANHÃ-CASTELO DO QUEIJC	4	11828	11269		02/10/2013	-1	
215493	205	205 - CAMPANHÃ-CASTELO DO QUEIJC	6	11415	1116		02/10/2013	-1	
215481	205	205 - CAMPANHÃ-CASTELO DO QUEIJC	13			5	02/10/2013	-1	
215506	205	205 - CAMPANHÃ-CASTELO DO QUEIJC	11			1	02/10/2013	-1	
215512	205	205 - CAMPANHÃ-CASTELO DO QUEIJC	10	12049	1130	1	02/10/2013	5	Fora de Serviço até Paragem
215588	205	205 - CAMPANHÃ-CASTELO DO QUEIJC	10			10	03/10/2013	-1	
215625	205	205 - CAMPANHÃ-CASTELO DO QUEIJC	5	10983	1119	0	03/10/2013	-1	
215582	205	205 - CAMPANHÃ-CASTELO DO QUEIJC	2			0	03/10/2013	-1	
215659	205	205 - CAMPANHÃ-CASTELO DO QUEIJC	5	11117	1119	0	03/10/2013	12	Forma c/Avanço
215674	205	205 - CAMPANHÃ-CASTELO DO QUEIJC	10	12048	1116	0,5	03/10/2013	2	Forma na Paragem
215676	205	205 - CAMPANHÃ-CASTELO DO QUEIJC	1	12013	1124	0	03/10/2013	7	Alterou Destino Para
215684	205	205 - CAMPANHÃ-CASTELO DO QUEIJC	6	11538	1118	0,5	03/10/2013	2	Forma na Paragem
215686	205	205 - CAMPANHÃ-CASTELO DO QUEIJC	10	12048	1116		03/10/2013	-1	
215719	205	205 - CAMPANHÃ-CASTELO DO QUEIJC	12			4	04/10/2013	-1	
215745	205	205 - CAMPANHÃ-CASTELO DO QUEIJC	10	11333	1113		04/10/2013	-1	
215771	205	205 - CAMPANHÃ-CASTELO DO QUEIJC	5	11881	1128		04/10/2013	3	Transbordo na Paragem
215866	205	205 - CAMPANHÃ-CASTELO DO QUEIJC	11	11526	3089	5	04/10/2013	14	Transferido
215850	205	205 - CAMPANHÃ-CASTELO DO QUEIJC	13			5	04/10/2013	-1	
215857	205	205 - CAMPANHÃ-CASTELO DO QUEIJC	9	11415	1114		04/10/2013	2	Forma na Paragem
215876	205	205 - CAMPANHÃ-CASTELO DO QUEIJC	12	11243	1122		04/10/2013	2	Forma na Paragem
215892	205	205 - CAMPANHÃ-CASTELO DO QUEIJC	8	11907	1102	1	04/10/2013	2	Forma na Paragem

Figura 3.14: Exemplo de uma folha SICO

Na tabela da figura 3.14 são apresentados atrasos por trânsito congestionado ou acidentes. Esta tabela pode ser útil para a análise de tempos de viagem muito diferentes dos previstos.

Capítulo 4

Métodos Empíricos

Neste capítulo são apresentados métodos empíricos que surgiram num primeiro contacto com os dados e com o funcionamento de todos os processos.

Os modelos são apresentados pela ordem com que foram criados.

4.1 Modelo 1

Há várias definições de quantil; só no R Core Team (2014) há 9 definições diferentes. A definição aqui usada é a definição fornecida por defeito pelo software. A função quantil genérica produz quantis amostrais correspondentes às probabilidades empíricas observadas. A maior observação corresponde a uma probabilidade de 1, naturalmente. (R Core Team, 2014)

Todos os quantis da amostra são definidos como médias ponderadas das estatísticas de ordem. Tem-se:

$$Q(p) = (1-\gamma)x[j] + \gamma x[j+1] \tag{4.1}$$

onde, $(j)/n \leq p < (j+1)/n$, $x[j]$ é a j^{esima} estatística de ordem, n é o tamanho amostral, $j = \lfloor np \rfloor$ e $g = np - j$.

Para o quantil usado $Q(p)$ é uma função de p descontínua com:

$$\begin{cases} \gamma = 0 \; se \; g = 0 \\ \gamma = 1 \; outros \; casos \end{cases} \tag{4.2}$$

Sabemos que aproximadamente 5% das observações estão à direita do quantil amostral $q_{0.95}$ e que aproximadamente 5% estão à esquerda do quantil amostral $q_{0.05}$. O modelo 1 está esquematizado na figura 4.1

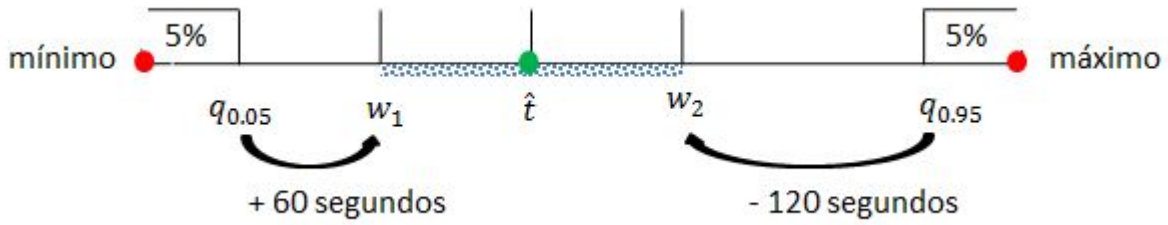


Figura 4.1: Esquema do Modelo 1

Eliminam-se os percursos mais lentos e os mais rápidos. Mais precisamente, considera-se um corte de 5% para cada lado. Gostaríamos que os restantes 90% de tempos amostrados definissem um intervalo de 3 minutos, até 2 minutos atrasados e menos de 1 minuto adiantado. Para se obter esse intervalo ótimo, subtraíram-se 120 segundos ao quantil $q_{0.95}$ e somaram-se 60 segundos ao quantil $q_{0.05}$. A situação ideal seria ter $q_{0.05} + 60 = q_{0.95} - 120$, mas como estes valores não são geralmente coincidentes, definimos o tempo previsto como sendo a média ponderada para estes dois tempos:

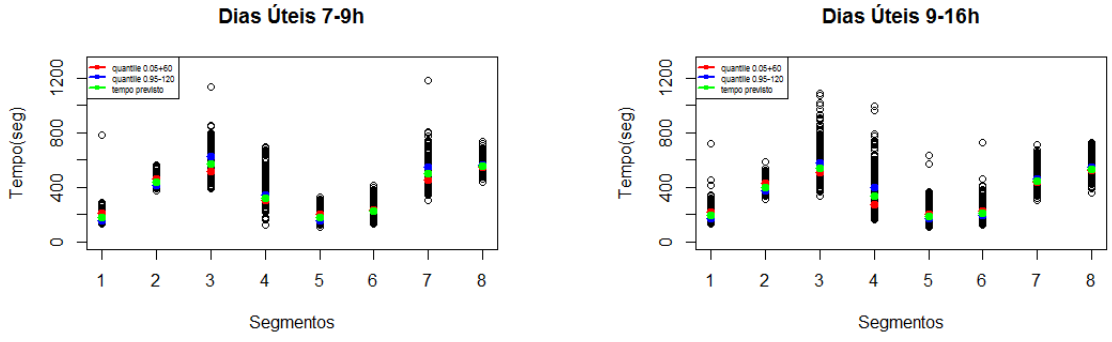
$$\widehat{t}_i = \frac{w_1 + w_2}{2} \tag{4.3}$$

onde $i = segmento * periodo\_dia * tipo\_dia$.

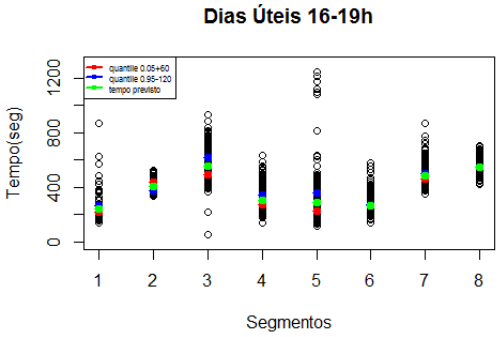
$$\begin{cases} w_1 = q_{0.05} + 60 \\ w_2 = q_{0.95} - 120 \end{cases} \tag{4.4}$$

4.1.1 Resultados da Aplicação do Modelo 1

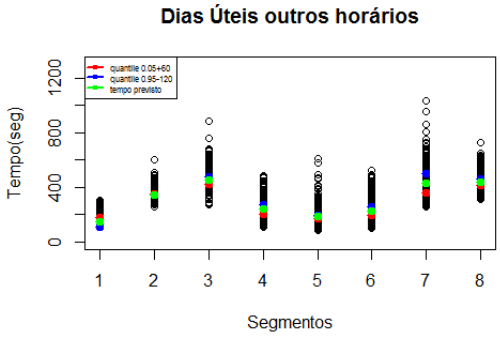
Para todas as sub-amostras foram calculados os valores de w_1 e w_2 e o respetivo tempo previsto (figura 4.2).



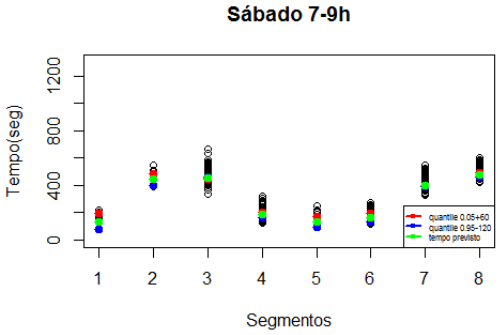
(a) (b)


Dias Úteis 16-19h
quantile 0.05+60
quantile 0.95-120
tempo previsto
Tempo(seg)
0
400
800
1200
1
2
3
4
5
6
7
8
Segmentos

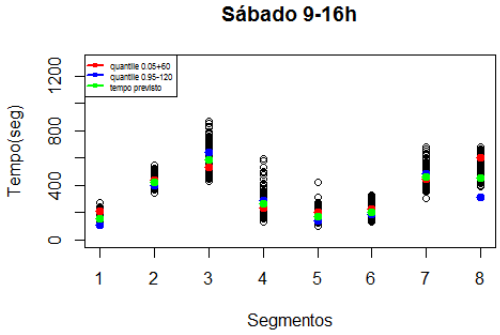
(c)


Dias Úteis outros horários
quantile 0.05+60
quantile 0.95-120
tempo previsto
Tempo(seg)
0
400
800
1200
1
2
3
4
5
6
7
8
Segmentos

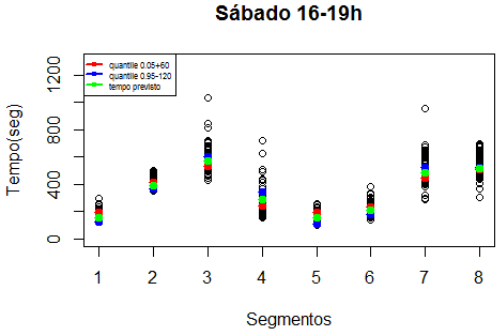
(d)


Sábado 7-9h
quantile 0.05+60
quantile 0.95-120
tempo previsto
Tempo(seg)
0
400
800
1200
1
2
3
4
5
6
7
8
Segmentos

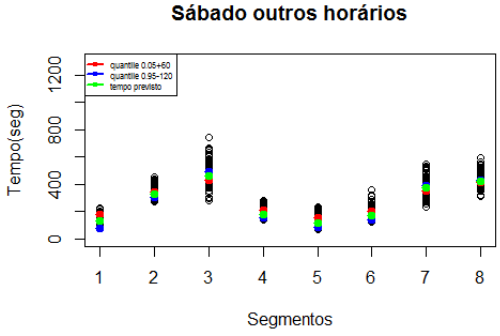
(e)


Sábado 9-16h
quantile 0.05+60
quantile 0.95-120
tempo previsto
Tempo(seg)
0
400
800
1200
1
2
3
4
5
6
7
8
Segmentos

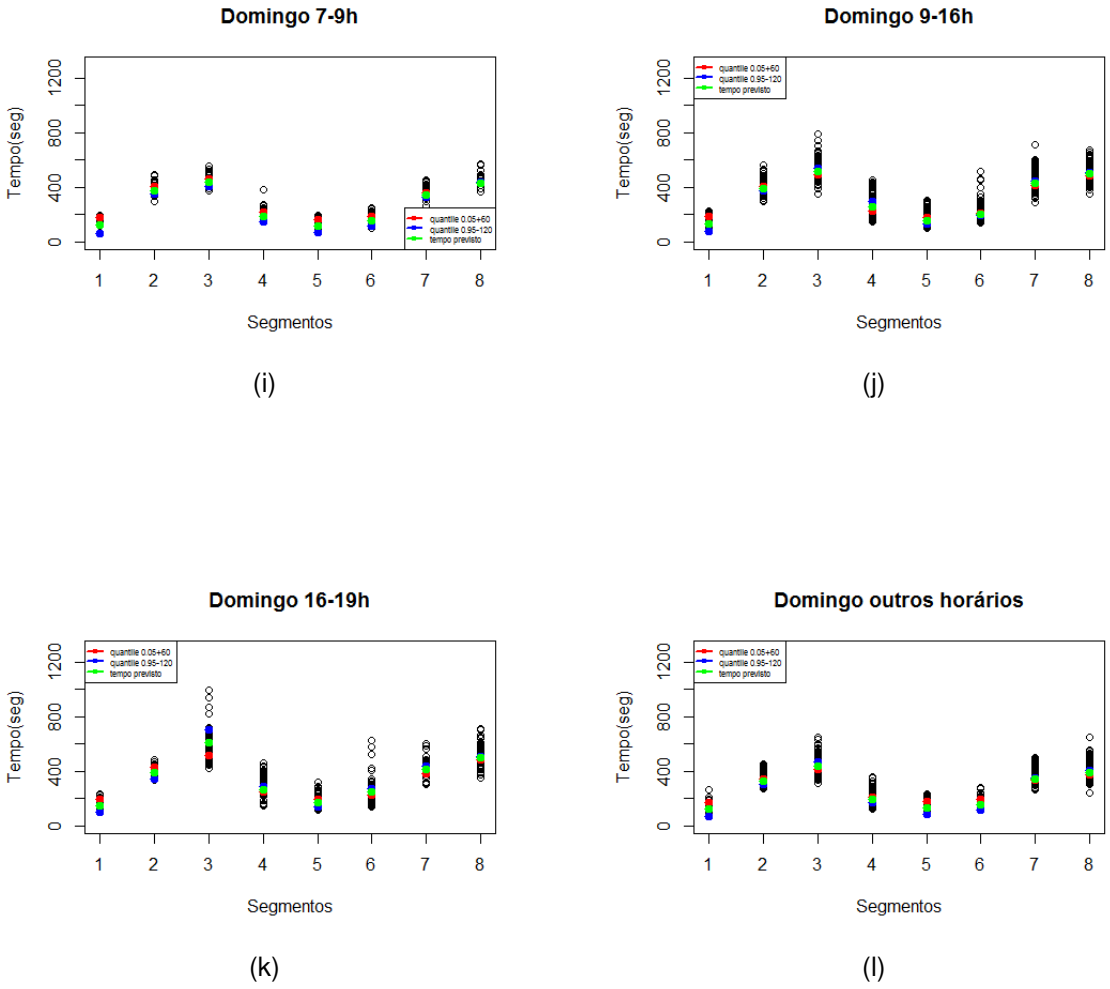
(f)


Sábado 16-19h
quantile 0.05+60
quantile 0.95-120
tempo previsto
Tempo(seg)
0
400
800
1200
1
2
3
4
5
6
7
8
Segmentos

(g)


Sábado outros horários
quantile 0.05+60
quantile 0.95-120
tempo previsto
Tempo(seg)
0
400
800
1200
1
2
3
4
5
6
7
8
Segmentos

(h)



(i) (j)

(k) (l)

Figura 4.2: Gráficos com a previsão e quantis para o modelo empírico 1 (a) Dias Úteis 7-9h (b) Dias Úteis 9-16h (c) Dias Úteis 16-19h (d) Dias Úteis outros horários (e) Sábado 7-9h (f) Sábado 9-16h (g) Sábado 16-19h (h) Sábado outros horários (i) Domingo 7-9h (j) Domingo 9-16h (k) Domingo 16-19h (l) Domingo outros horários

Em quase todos os gráficos é possível distinguir os pontos w_1, w_2 e $\widehat{t}$. Nos casos em que parecem estar sobrepostos temos $w_1 \approx w_2$.

4.2 Modelo 2

O modelo 2 surge como uma melhoria do modelo 1. Ao contrário do modelo 1, o tempo previsto é calculado tendo em conta o intervalo entre w_1 e w_2. A figura 4.3 e as equações 4.5 e 4.6 descrevem o modelo 2.

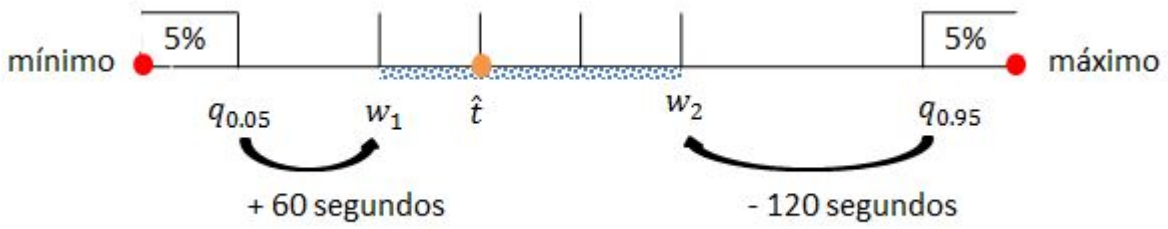


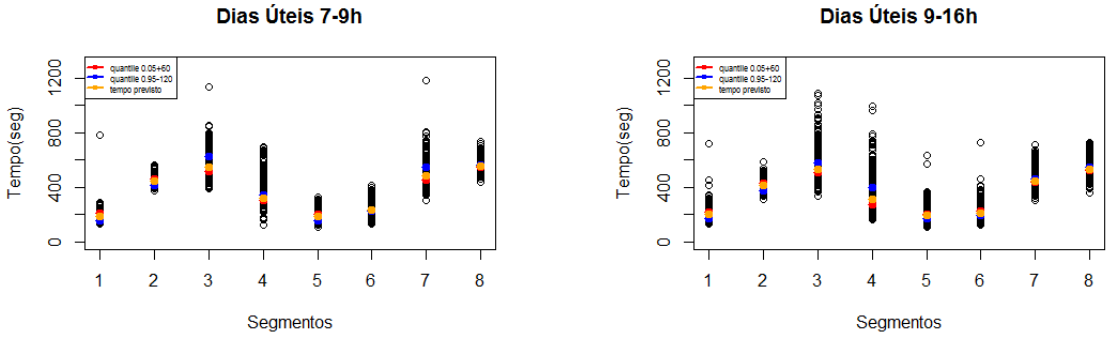
Figura 4.3: Esquema do Modelo 2

$$\widehat{t_i} = w_1 + \frac{w_2 - w_1}{3} \tag{4.5}$$

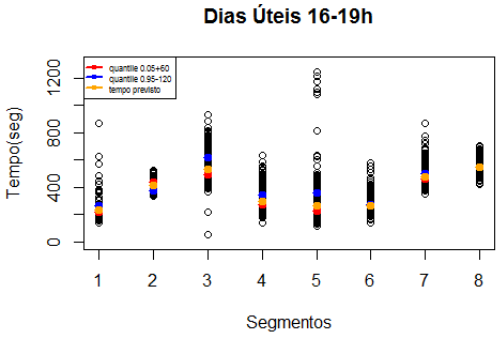
$$\begin{cases} w_1 = q_{0.05} + 60 \\ w_2 = q_{0.95} - 120 \end{cases} \tag{4.6}$$

4.2.1 Resultados da Aplicação do Modelo 2

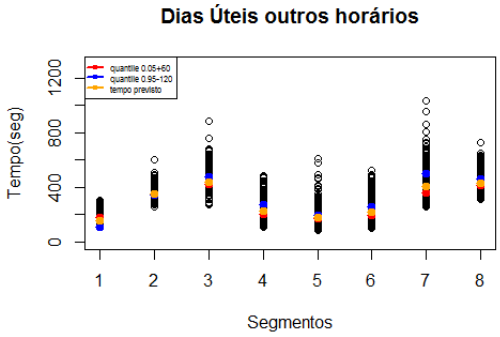
Do mesmo modo que para o modelo 1, para todas as sub-amostras foram calculados os valores para w_1 e w_2 e $\widehat{t}$.



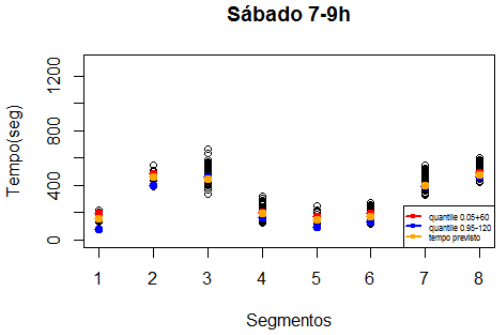
(a) (b)


Dias Úteis 16-19h
quantile 0.05+60
quantile 0.95-120
tempo previsto
Tempo(seg)
0
400
800
1200
1
2
3
4
5
6
7
8
Segmentos

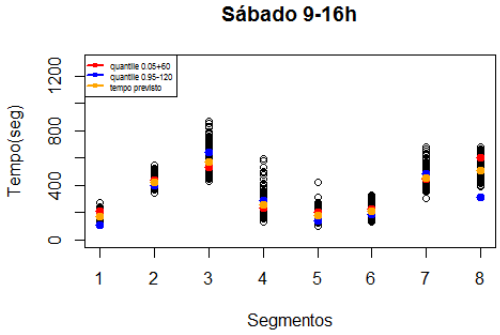
(c)


Dias Úteis outros horários
quantile 0.05+60
quantile 0.95-120
tempo previsto
Tempo(seg)
0
400
800
1200
1
2
3
4
5
6
7
8
Segmentos

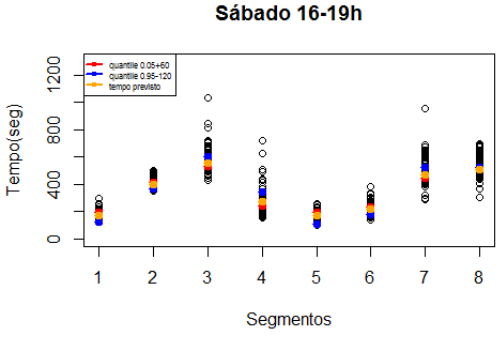
(d)


Sábado 7-9h
quantile 0.05+60
quantile 0.95-120
tempo previsto
Tempo(seg)
0
400
800
1200
1
2
3
4
5
6
7
8
Segmentos

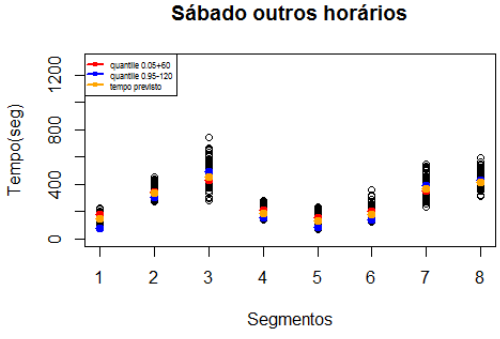
(e)


Sábado 9-16h
quantile 0.05+60
quantile 0.95-120
tempo previsto
Tempo(seg)
0
400
800
1200
1
2
3
4
5
6
7
8
Segmentos

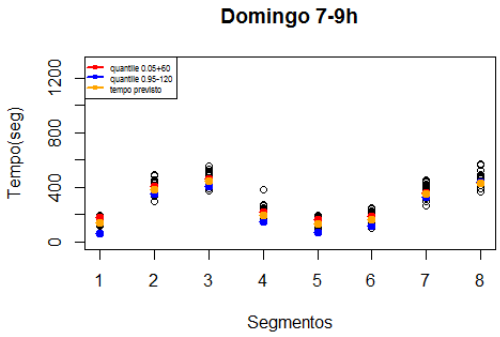
(f)


Sábado 16-19h
quantile 0.05+60
quantile 0.95-120
tempo previsto
Tempo(seg)
0
400
800
1200
1
2
3
4
5
6
7
8
Segmentos

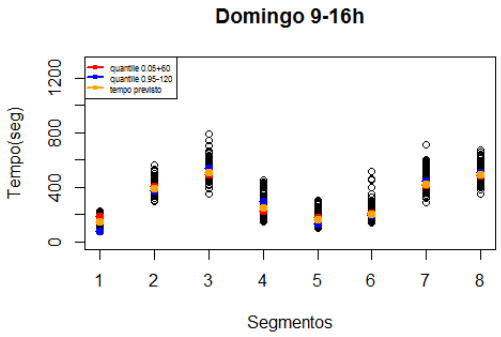
(g)


Sábado outros horários
quantile 0.05+60
quantile 0.95-120
tempo previsto
Tempo(seg)
0
400
800
1200
1
2
3
4
5
6
7
8
Segmentos

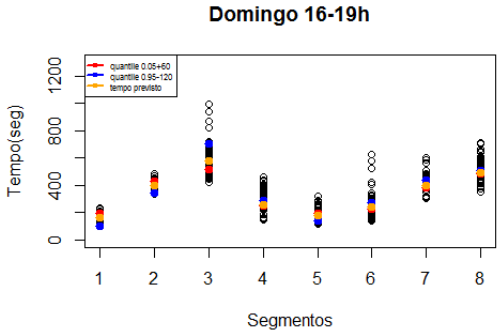
(h)



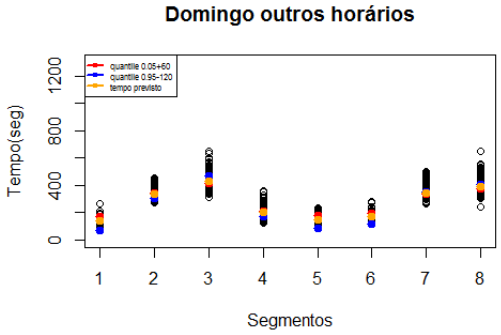
(i)



(j)



(k)



(l)

Figura 4.4: Gráficos com a previsão e quantis para o modelo empírico 2 (a) Dias Úteis 7-9h (b) Dias Úteis 9-16h (c) Dias Úteis 16-19h (d) Dias Úteis outros horários (e) Sábado 7-9h (f) Sábado 9-16h (g) Sábado 16-19h (h) Sábado outros horários (i) Domingo 7-9h (j) Domingo 9-16h (k) Domingo 16-19h (l) Domingo outros horários

4.3 Modelo 3

O modelo 3 segue os mesmos princípios do modelo 2, mas tem em conta que a situação $w_1 > w_2$. Para controlar esta situação, procedeu-se a uma redução das caudas de modo a garantir que $w_1 < w_2$ pode ocorrer. O fluxograma apresentado na figura 4.5 e o esquema da figura 4.6 explicam o processo correspondente ao modelo 3.

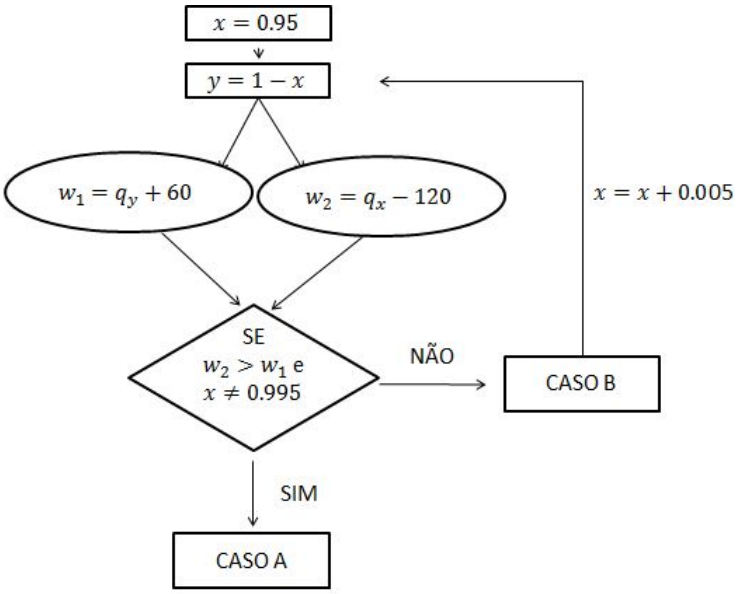
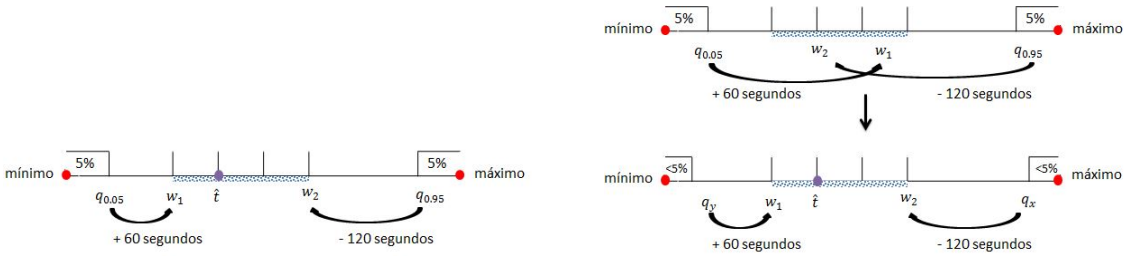


Figura 4.5: Fluxograma do Modelo 3



(a) (b)

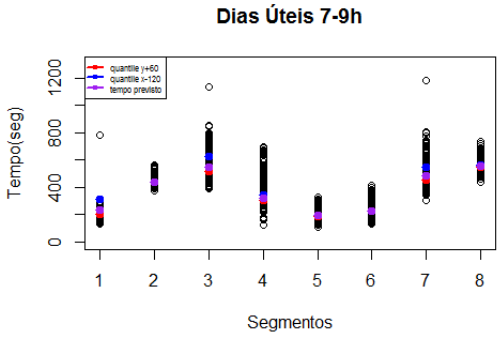
Figura 4.6: Esquema do Modelo 3: (a) Caso A, (b) Caso B

Quando $w_1 < w_2$ estamos na presença do caso A (igual ao modelo 2); quando a condição não é verificada estamos na presença do caso B. Nesta situação, procede-se a um incremento de 0.005 no quantil q_x e à atualização do quantil q_{1-x}. Este processo é iterado até que se recaia no caso A. Em alguns casos, verificou-se, que a desigualdade $w_1 > w_2$ nunca foi verdadeira. Para esses casos foi considerado $x = 0.995$. A equação 4.7 descreve o cálculo do valor para o tempo previsto.

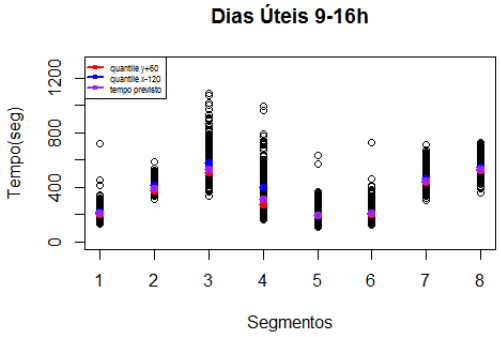
$$\widehat{t_i} = w_1 + \frac{w_2 - w_1}{3} \tag{4.7}$$

4.3.1 Resultados da Aplicação do Modelo 3

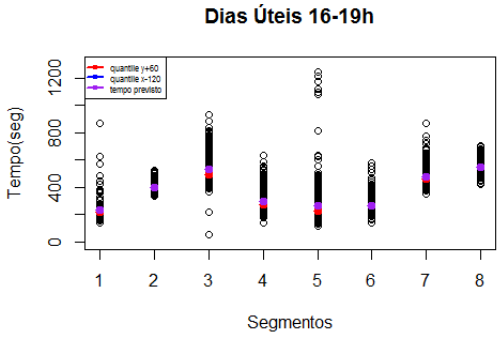
Os resultados da aplicação deste modelo estão representados graficamente na figura 4.7.


Dias Úteis 7-9h
quantile y+60
quantile x-120
tempo previsto
Tempo(seg)
0
400
800
1200
1
2
3
4
5
6
7
8
Segmentos

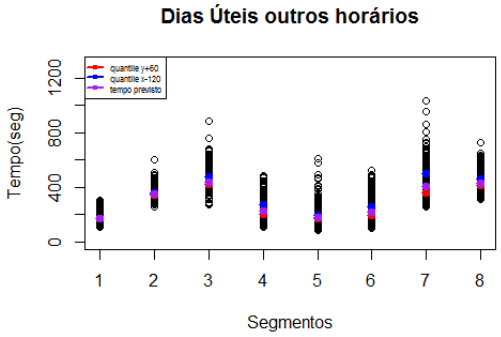
(a)


Dias Úteis 9-16h
quantile y+60
quantile x-120
tempo previsto
Tempo(seg)
0
400
800
1200
1
2
3
4
5
6
7
8
Segmentos

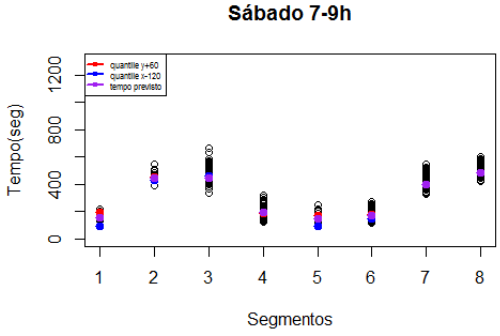
(b)


Dias Úteis 16-19h
quantile y+60
quantile x-120
tempo previsto
Tempo(seg)
0
400
800
1200
1
2
3
4
5
6
7
8
Segmentos

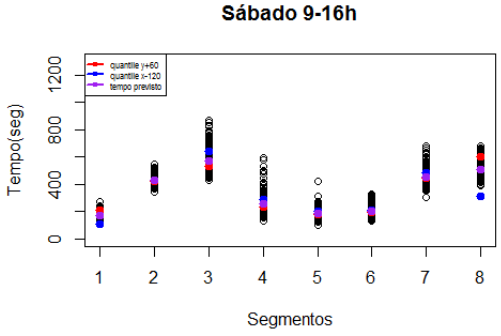
(c)


Dias Úteis outros horários
quantile y+60
quantile x-120
tempo previsto
Tempo(seg)
0
400
800
1200
1
2
3
4
5
6
7
8
Segmentos

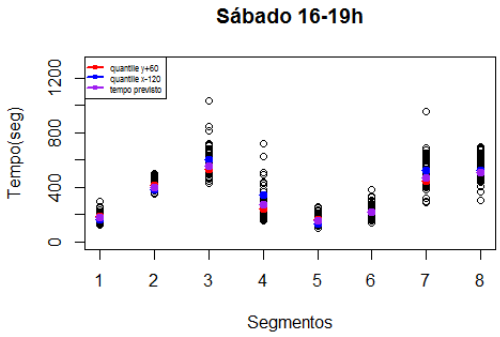
(d)


Sábado 7-9h
quantile y+60
quantile x-120
tempo previsto
Tempo(seg)
0
400
800
1200
1
2
3
4
5
6
7
8
Segmentos

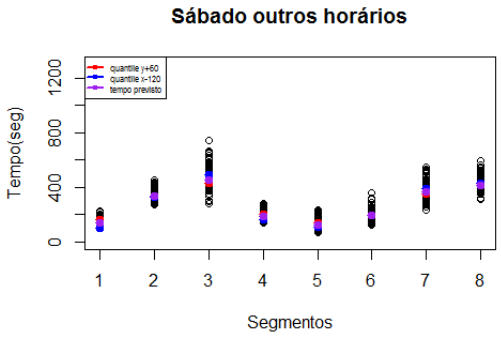
(e)


Sábado 9-16h
quantile y+60
quantile x-120
tempo previsto
Tempo(seg)
0
400
800
1200
1
2
3
4
5
6
7
8
Segmentos

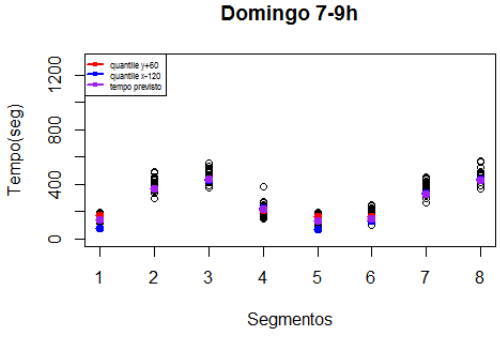
(f)



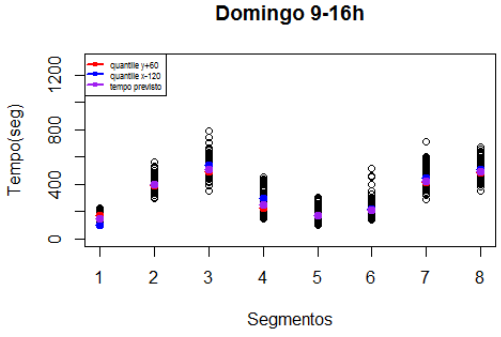
(g)



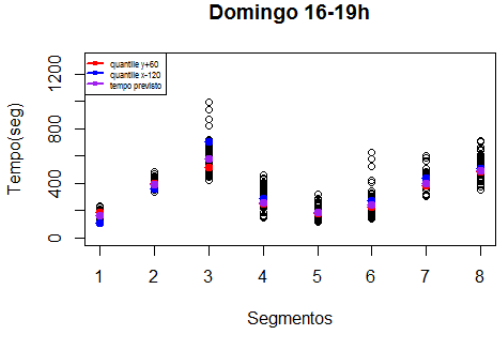
(h)



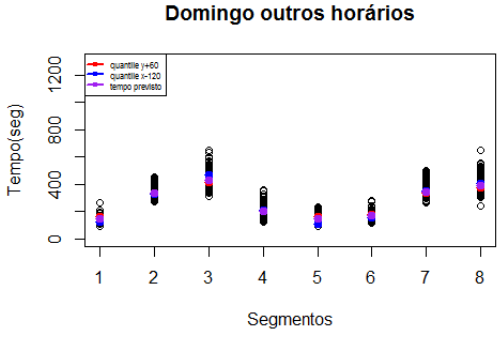
(i)



(j)



(k)



(l)

Figura 4.7: Gráficos com a previsão e quantis para o modelo empírico 3 (a) Dias Úteis 7-9h (b) Dias Úteis 9-16h (c) Dias Úteis 16-19h (d) Dias Úteis outros horários (e) Sábado 7-9h (f) Sábado 9-16h (g) Sábado 16-19h (h) Sábado outros horários (i) Domingo 7-9h (j) Domingo 9-16h (k) Domingo 16-19h (l) Domingo outros horários

Neste capítulo não é averiguado qual a melhor metodologia empírica, essa comparação é realizada no capítulo de Discussão de Resultados.

Capítulo 5

Método dos Mínimos Quadrados Generalizado

Neste capítulo é apresentado o Método dos Mínimos Quadrados Generalizado (MMQG[1]). O capítulo está dividido em duas partes: na primeira são apresentados os pressupostos teóricos do método e na segunda parte é feita a aplicação do método ao caso de estudo. O MMQG é uma técnica que serve para estimar os parâmetros desconhecidos de um modelo de regressão linear que considere erros com variâncias diferentes (heterocedasticidade) e/ou com correlações não nulas.

5.1 Pressupostos Teóricos

5.1.1 O Modelo

Considere-se o modelo de regressão linear dado pela equação: (Pinheiro and Bates, 2000, Kariya and Kurata, 2004)

$$\begin{aligned} y_i &= X_i\beta + \epsilon_i \\ \epsilon_i &\sim N(0, \sigma^2\Lambda_i) \\ i &= 1, ..., n \end{aligned} \tag{5.1}$$

onde n é o número de observações, $y_i = (y_{i1}, ..., y_{iT})$ é o vetor resposta para a observação i, X_i é a matriz de covariáveis, β é o vetor dos coeficientes estimados pela regressão e ϵ_i é o vetor dos erros. A matriz de variância-covariância dos erros semi-definida positiva é dada por $\sigma^2\Lambda_i$, onde a matriz Λ_i é parametrizada por parâmetros λ.

A matriz Λ_i admite uma raiz quadrada $\Lambda_i^{1/2}$ invertível, com inversa $\Lambda_i^{-1/2}$, de modo que:

$$\Lambda_i = (\Lambda_i^{1/2})^\top \Lambda_i^{1/2}$$

[1]GLSM - Generalized Least Squares Method

e

$$\Lambda_i^{-1} = \Lambda_i^{-1/2}(\Lambda_i^{-1/2})^\top.$$

Considere-se a seguinte reparametrização do modelo:

$$\begin{aligned} y_i^* &= (\Lambda_i^{-1/2})^\top y_i \\ X_i^* &= (\Lambda_i^{-1/2})^\top X_i \\ \epsilon_i^* &= (\Lambda_i^{-1/2})^\top \epsilon_i. \end{aligned} \tag{5.2}$$

Tendo em atenção que:

$$\epsilon_i^* \sim N\left[(\Lambda_i^{-1/2})^\top 0, \sigma^2(\Lambda_i^{-1/2})^\top \Lambda_i \Lambda_i^{-1/2}\right] = N(0, \sigma^2 I), \tag{5.3}$$

podemos reescrever a equação (5.1) na forma:

$$y_i^* = X_i^* \beta + \epsilon_i^*, \tag{5.4}$$

onde $\epsilon_i^* \sim N(0, \sigma^2 I)$, $i = 1, ..., M$. (Pinheiro and Bates, 2000).

5.1.2 Estimação dos Parâmetros do Modelo

Atendendo à expressão da função de máxima verosimilhança, o logaritmo da mesma é uma função dependente apenas de λ dada por: (Kariya and Kurata, 2004, Pinheiro and Bates, 2000)

$$l(\lambda|y) = \ constante \ - N log\|y^* X^* \widehat{\beta}(\lambda)\| - \frac{1}{2}\sum_{i=1}^{n} log|\Lambda_i| \tag{5.5}$$

Fixando λ, os estimadores de máxima verosimilhança de β e σ^2 são obtidos pela resolução de um problema de mínimos quadrados. Representando por X^* a matriz que agrega todas as matrizes X correspondentes às diferentes observações, os estimadores de máxima verosimilhança dos mínimos quadrados generalizados são dados por: (Kariya and Kurata, 2004, Pinheiro and Bates, 2000)

$$\widehat{\beta}(\lambda) = [(X^*)^T X^*]^{-1}(X^*)^T y^* \tag{5.6}$$

$$\widehat{\sigma^2}(\lambda) = \frac{\|y^* X^* \widehat{\beta}(\lambda)\|^2}{N} \tag{5.7}$$

5.1.3 Decomposição da Matriz de Variância - Covariância

As matrizes Λ_i podem ser decompostas no produto de matrizes mais simples dado por: (Pinheiro and Bates, 2000)

$$\Lambda_i = V_i C_i V_i \tag{5.8}$$

onde V_i é uma matriz diagonal e C_i é uma matriz de correlação, ou seja, uma matriz semi-definida positiva com todos os elementos da diagonal iguais a 1. A matriz V_i pode não ser única, pois podemos multiplicar cada uma das linhas por -1 e obter a mesma decomposição. Para garantir a sua unicidade impõe-se que V_i tenha todos os elementos da diagonal principal positivos. Por outro lado,

$$\begin{aligned} var(\epsilon_{ij}) &= \sigma^2 \left[V_i\right]^2_{jj} \\ corr(\epsilon_{ij}, \epsilon_{jk}) &= \left[C_i\right]_{jk}, \end{aligned}$$

portanto V_i descreve a variância dos erros ϵ_i dentro do grupo e C_i descreve as suas correlações. Esta decomposição de Λ_i numa estrutura de variância e numa estrutura de correlação, permite a modelação destas estruturas separadamente. As funções de variância para representar a componente V_i e as estruturas de correlação para a componente C_i serão descritas em § 5.1.3.1 e § 5.1.3.2, respetivamente. (Pinheiro and Bates, 2000)

5.1.3.1 Funções de Variância para Modelar a Heterocedasticidade

As funções de variância são usadas para modelar a estrutura da variância dos erros dentro do grupo, usando eventualmente covariáveis. (Pinheiro and Bates, 2000)

De acordo com a parametrização proposta por Davidian & Giltinian (1995) a função de variância dos erros dentro do grupo associada ao modelo (5.1) pode ser definida da seguinte forma: (Cabral and Gonçalves, 2011)

$$var(\epsilon_{ij}) = \sigma^2 g^2(\mu_{ij}, \nu_{ij}, \delta),\ i = 1, ..., n;\ j = 1, ..., n_i, \tag{5.9}$$

onde $\mu_{ij} = E[y_{ij}] = x_{ij}^T \beta$, ν_{ij} é um vetor de covariáveis da variância, δ é o vetor dos parâmetros da variância e $g(.)$ é a função de variância, contínua em δ. (Cabral and Gonçalves, 2011, Pinheiro and Bates, 2000) Esta função é escolhida de modo a refletir a variabilidade intra-indivíduo e pode ser, por exemplo, a função exponencial, logarítmica, potência ou uma combinação destas funções. (Cabral and Gonçalves, 2011) Na tabela 5.1 são dadas algumas funções de variância: (Cabral and Gonçalves, 2011, Pinheiro and Bates, 2000).

Descrição da classe	Variância ($var(\epsilon_{it})$)
Variância Fixa	$\sigma^2\nu_{it}$
Variâncias diferentes por classe	$\sigma^2\delta^2_{s_{it}}$
Potência de uma covariável	$\sigma^2\|\nu_{it}\|^{2\delta}$
Exponencial de uma covariável	$\sigma^2 exp(2\delta\nu_{it})$
Constante + Potência de uma covariável	$\sigma^2(\delta_1 + \|\nu_{it}\|^{\delta_2})^2$

ν_{it} - covariável; s_{it} - variável de estratificação; $\delta_1 > 0$

Tabela 5.1: Funções de variância

A formulação da função de variância descrita em (5.9) é muito flexível e intuitiva na medida em que permite que a variância dentro do indivíduo dependa dos efeitos fixos β através dos valores esperados μ_{it}. (Pinheiro and Bates, 2000)

5.1.3.2 Estruturas de Correlação para Modelar a Dependência

As estruturas de correlação são usadas para modelar a dependência entre as observações. No contexto dos modelos lineares, são usadas para modelar a dependência dos erros aleatórios dentro do grupo. Nesta secção é assumido que as estruturas de correlação são isotrópicas, isto é, que a correlação entre dois erros ϵ_{it} e $\epsilon_{it'}$, depende da posição dos vetores p_{it} e $p_{it'}$ apenas através da distância entre eles $d(p_{it}, p_{it'})$, e não através dos valores particulares que tomam (Pinheiro and Bates, 2000).

A expressão geral para a estrutura de correlação dentro do grupo é expressa na forma, para $i = 1, ..., M$ e $j, j^{'} = 1, ..., T_i$,

$$corr(\epsilon_{it}, \epsilon_{it'}) = h[d(p_{it}, p_{it'}), \rho], \tag{5.10}$$

onde ρ é um vetor de parâmetros de correlação e $h(.)$ é uma função de correlação que assume valores entre -1 e 1, contínua em ρ e tal que $h(0, \rho) = 1$ (Pinheiro and Bates, 2000). Quanto mais próximos, no espaço ou no tempo, estiverem dois erros aleatórios, maior a sua dependência.

As estruturas de correlação foram desenvolvidas para duas classes de dados: séries temporais e dados espaciais. A primeira está associada a observações indexadas por um tempo que tome valores inteiros (unidimensional), enquanto que a segunda está associada a observações indexadas a um vetor espacial tomando valores no plano real (bidimensional). Neste trabalho apenas foram estudadas as estruturas de correlação serial.

Estruturas de Correlação Serial
As estruturas de correlação serial são usadas para modelar a dependência em dados de séries temporais, ou seja, em observações feitas sequencialmente ao longo do tempo. Simplificando o pressuposto de isotropia, o modelo geral de correlação serial é dado por:

$$corr(\epsilon_{it}, \epsilon_{it'}) = h[|p_{it} - p_{it'}|, \rho]$$

No contexto das séries temporais, a função de correlação $h(.)$ é designada por função de autocorrelação. A função de autocorrelação empírica, que consiste numa estimativa não paramétrica da função de autocorrelação, é bastante útil para verificar a correlação serial nos dados. Sejam

$$r_{it} = (y_{it} - \hat{y}_{it})/\hat{\sigma}_{it} \tag{5.11}$$

onde $\hat{\sigma}_{it}$ é o estimador da variância de ϵ_{it}, os resíduos padronizados do modelo linear. A função de autocorrelação no espaçamento (lag) l é dada por:

$$\hat{\rho}(l) = \frac{\sum_{i=1}^{n}\sum_{t=1}^{T_i-l} r_{it}r_{i(t+l)}/N(l)}{\sum_{i=1}^{n}\sum_{t=1}^{T_i} r_{it}^2/N(0)} \tag{5.12}$$

onde $N(l)$ representa o número de pares de resíduos utilizados no somatório do numerador da função.

Se as observações forem igualmente espaçadas, a função de autocorrelação empírica pode ser usada (graficamente) para identificar o processo. Analisando o gráfico desta função podemos deduzir que:

- se os valores se aproximam de zero gradualmente então o processo pode ser autoregressivo,
- caso a função de autocorrelação seja consistente dentro de

  $$\pm z_{(1-\alpha/2)}/\sqrt{N(l)}z_{(1-\alpha/2)}$$

  após o lag 2 ou 3 então o modelo pode ser identificado como um processo de médias móveis de ordem 1 ou 2.

As estruturas de correlação serial mais usadas na prática são:

Simetria Composta Assume-se uma correlação igual entre todos os erros aleatórios dentro do mesmo grupo; isto é, para o mesmo indivíduo os erros correspondentes a diferentes tempos estão todos igualmente correlacionados. O modelo de correlação é dado por:

$$corr(\epsilon_{it}, \epsilon_{it'}) = \rho, \forall t \neq t',\ h(k,\rho) = \rho,\ k = 1, 2, ...$$

onde o único parâmetro de correlação ρ é designado por coeficiente de correlação intraclasse.

Esta estrutura é útil para aplicações que envolvem séries com um curto período de tempo dentro do grupo, ou quando todas as observações dentro do grupo são recolhidas ao mesmo tempo (Pinheiro and Bates, 2000).

A estrutura de correlação de simetria composta tende a ser muito simples nas aplicações práticas que envolvem séries temporais. Assim, em geral, é mais realista assumir uma estrutura em que a correlação entre duas observações diminui em valor absoluto com a sua distância (Cabral and Gonçalves, 2011, Pinheiro and Bates, 2000).

Geral Na estrutura de correlação geral, cada correlação entre as observações é dada por um parâmetro diferente, e a função de correlação é dada por:

$$h(k, \rho) = \rho_k, \ k = 1, 2, ... \tag{5.13}$$

Pelo facto do número de parâmetros em (5.13) aumentar quadraticamente com o número máximo de observações dentro do grupo, esta estrutura leva a modelos sobre-parametrizados, sendo útil apenas quando existem poucas observações por grupo (Cabral and Gonçalves, 2011, Pinheiro and Bates, 2000).

Auto-regressivo - Médias Móveis Esta família de estruturas de correlação inclui diferentes classes de modelos lineares estacionários: modelos auto regressivos (AR), modelos de médias móveis (MA) e modelos auto-regressivos de médias móveis (ARMA).

Estes modelos assumem que as observações são feitas em intervalos de tempo inteiros e, para simplificar vamos omitir o índice referente ao individuo, pelo que, ϵ_t designa a observação que ocorreu no instante de tempo t. A distância (lag), ou período de tempo, entre duas observações ϵ_t e ϵ_s é dada por $|t - s|$, pelo que $lag1$ refere-se a observações feitas com uma unidade de distância, $lag2$ a observações feitas com duas unidades de distância, e assim sucessivamente (Pinheiro and Bates, 2000).

1. **Modelo Auto-regressivo - AR** Estes modelos exprimem uma observação como combinação linear das observações anteriores acrescentada de um ruído homocedástico, a_t, centrado em zero $E[a_t] = 0$ e independente das observações anteriores

   $$\epsilon_t = \phi_1 \epsilon_{t-1} + ... + \phi_p \epsilon_{t-p} + a_t,$$

   onde p é o número de observações anteriores e designado por ordem do modelo auto-regressivo e escreve-se $AR(p)$. Assim, existem p parâmetros de autocorrelação num modelo $AR(p)$ dados por $\Phi = (\phi_1, ..., \phi_p)$.
   O único parâmetro da correlação, ϕ, tem de ser não negativo.

   Para modelos auto-regressivos de ordem superior a 1, a função de correlação é definida recursivamente através da equação às diferenças (Pinheiro and Bates, 2000)

   $$h(k, \phi) = \phi_1 h(|k - 1|, \phi) + ... + \phi_p h(|k - p|, \phi), \ k = 1, 2, ...$$

2. **Modelo de Média Móvel - MA** Nestes modelos assume-se que qualquer observação é uma combinação linear de termos de ruído, isto é,

   $$\epsilon_t = \theta_1 a_{t-1} + ... + \theta_q a_{t-q} + a_t, \tag{5.14}$$

onde q é número de termos de ruído incluídos. O valor q é designado por ordem do modelo de médias móveis, $MA(q)$. Existem q parâmetros de correlação dados por $\theta = (\theta_1, ..., \theta_q)$ e a função de correlação para um modelo $MA(q)$ tem a forma:

$$h(k,\theta) = \begin{cases} \dfrac{\theta_k + \theta_1\theta_{k-1} + ... + \theta_{k-q}\theta_q}{1 + \theta_1^2 + ... + \theta_q^2}, & k = 1, ..., q \\ 0, & k = q+1, q+2, ... \end{cases} \tag{5.15}$$

As observações separadas por mais do que q unidades de tempo não estão correlacionadas, uma vez que não partilham qualquer termo de ruído.

O modelo mais simples é o modelo de ordem 1, $MA(1)$ e tem-se:

$$h(1,\theta) = \rho_1 = \frac{\theta_1}{1+\theta_1^2},\ |\rho_1| < 0.5.$$

3. **Modelo Auto-regressivo de Média Móvel - ARMA** Estes modelos são obtidos combinando os modelos auto-regressivos e os modelos de média móvel. Um modelo $ARMA(p,q)$ é dado por:

$$\epsilon_t = \sum_{i=1}^{p} \phi_i \epsilon_{t-i} + \sum_{j=1}^{q} \theta_j a_{t-j} + a_t.$$

Este modelo tem $p+q$ parâmetros de correlação, que correspondem à combinação dos p parâmetros auto-regressivos $\Phi = (\phi_1, ..., \phi_p)$ e dos q parâmetros de média móvel $\theta = (\theta_1, ..., \theta_q)$.

A função de correlação de um modelo $ARMA(p,q)$ é dada por:

$$h(k,\rho) = \begin{cases} \phi_1 h(|k-1|,\rho) + ... + \phi_p h(|k-p|,\rho) + \theta_1\psi(k-1,\rho) + ... + \theta_q\psi(k-q,\rho), & k = 1, ..., q \\ \phi_1 h(|k-1|,\rho) + ... + \phi_p h(|k-p|,\rho), & k = q+1, q+2, ... \end{cases} \tag{5.16}$$

onde $\psi(k,\phi,\theta) = \dfrac{E[\epsilon_{t-k}a_t]}{var(\epsilon_t)}$. Note-se que $\psi(k,\phi,\theta) = 0$ para $k = 1, 2, ...$ dado que, neste caso ϵ_{t-k} e a_t são independentes e $E[a_t] = 0$.

5.2 Resultados

A aplicação do Método dos Mínimos Quadrados Generalizado foi feita com recurso à instrução *gls()* da biblioteca *nlme* (Pinheiro et al., 2014) do *software* R (R Core Team, 2014).

Foram testados vários modelos, descritos na tabela 5.2. A escolha do melhor modelo foi feita tendo em conta o Critério de Informação Bayesiano (BIC[2]). Nestes critérios, quanto menor o valor do critério melhor o modelo. Assim, quando usado o critério BIC para comparação de dois ou mais modelos, escolhemos o modelo com menor valor de BIC. (Pinheiro and Bates, 2000).

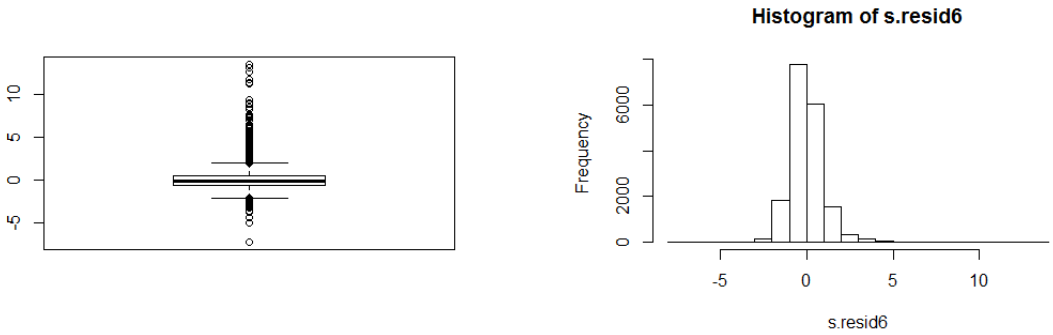
	Correlação	Variância	BIC
Modelo1	corAR1(form=∼1| MG/Turno/Week/id.turno)	varIdent(form= ∼1 | Week)	203748.5
Modelo2	corSymm(form=∼1| MG/Turno/Week/id.turno))	varIdent(form= ∼1 | Week)	*
Modelo3	corCompSymm(form=∼1| MG/Turno/Week/id.turno))	varIdent(form= ∼1 | Week)	203603.1
Modelo4	corCompSymm(form=∼1| MG/Turno/Week/id.turno))	varIdent(form=∼1|Turno))	203748.2
Modelo5	corCompSymm(form=∼1| MG/Turno/Week/id.turno))	varIdent(form=∼1|Week*Turno))	203125.9
Modelo6	corAR1(form=∼1| MG/Turno/Week/id.turno)	varIdent(form=∼1|Turno)	203892.6
Modelo7	corAR1(form=∼1| MG/Turno/Week/id.turno)	varIdent(form=∼1|Week*Turno)	203283

* Este modelo não foi executado, atingiu o limite de memória interna

Tabela 5.2: Estruturas de correlação e covariância dos Modelos analisados

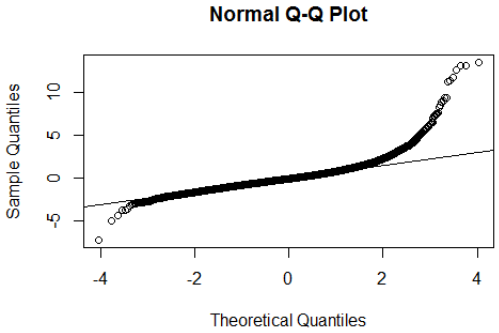
O significado das estruturas de correlação da tabela 5.2 é o seguinte: corAR(1) corresponde ao modelo auto-regressivo de ordem 1, corSymm à correlação geral e corCompSymm à correlação de simetria composta. No que se refere às funções de variância a usada foi variância constante por grupo, que difere de acordo com um fator de agrupamento. Esta decisão foi baseada no gráfico 2.4 do capítulo § 2.1. O significado das variáveis é o seguinte: o *MG* refere-se à identificação do motorista que conduzia o autocarro (Matrícula Geral), o *Turno* representa os diferentes períodos do dia, a variável *Week* representa o tipo do dia da semana e a variável *id.turno* é uma variável que diferencia o número de viagens que o motorista efetua no mesmo segmento, turno e tipo de dia. O melhor modelo, modelo 5 da tabela 5.2, indica que a correlação é constante entre quaisquer dois segmentos. A estrutura da variância é fixa uma vez escolhidos o período horário e tipo de dia.

[2]Bayesian Information Criterion



(a)

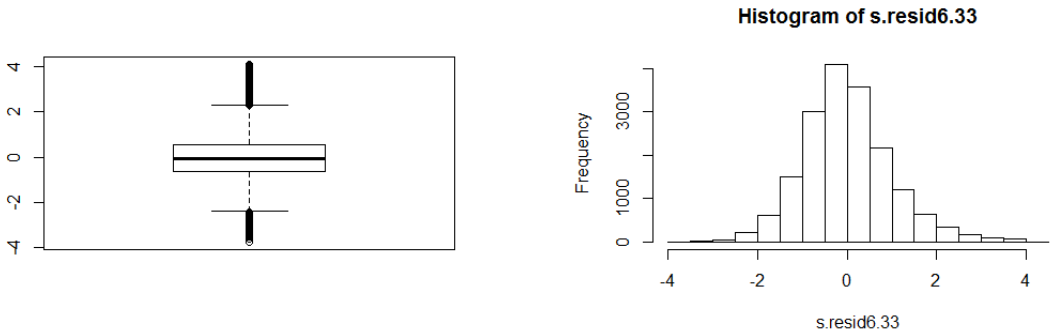
(b)



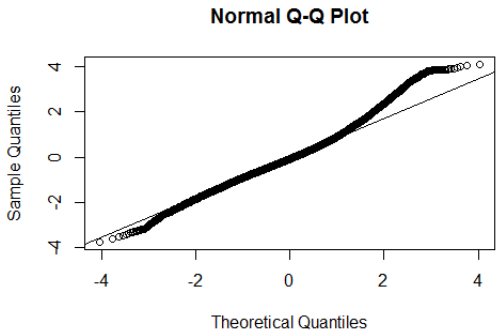
(c)

Figura 5.1: Resíduos estandarizados do modelo final estimado pelo MMQG (a) Diagrama de Caixa e Bigodes (b) Histograma (c) Gráfico dos Quantis

Os gráficos da figura 5.1 apresentam os resíduos para o modelo 5. É possível observar que os resíduos não apresentam exatamente simetria. Para tentar melhorar a qualidade de ajustamento do modelo retiraram-se os resíduos superiores em módulo a 3.3 (163 observações).



(a) (b)



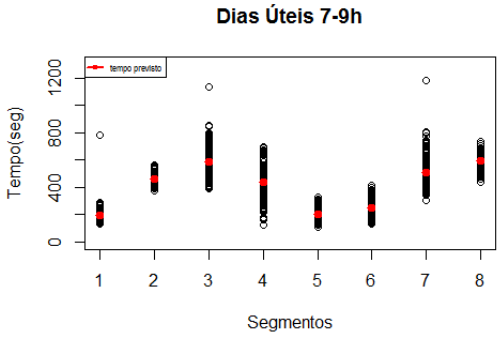
(c)

Figura 5.2: Resíduos estandarizados do modelo final após remoção de outliers (a) Diagrama de Caixa e Bigodes (b) Histograma (c) Gráfico de Quantis

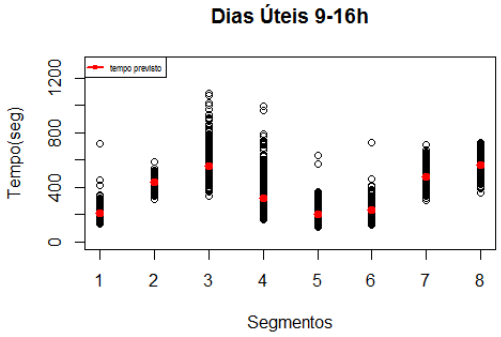
A figura 5.2 apresenta os gráficos dos resíduos estandarizados do novo modelo. A simetria dos resíduos não parece estar comprometida.

Os resultados da estimação estão listados em anexo (Apêndice A). A grande maioria das variáveis explicativas é estatisticamente significativa ($p<0.001$) e as variáveis com $p>0.001$ apresentam uma relativa semelhança com a variável de referência nos tempos de viagem.

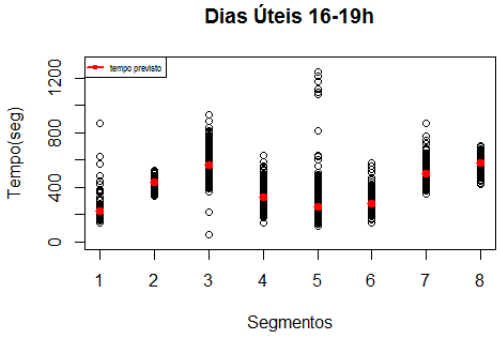
Os gráficos da figura 5.3 apresentam a previsão conseguida através do métodos dos mínimos quadrados generalizados.


Dias Úteis 7-9h
tempo previsto
Tempo(seg)
0
400
800
1200
1
2
3
4
5
6
7
8
Segmentos

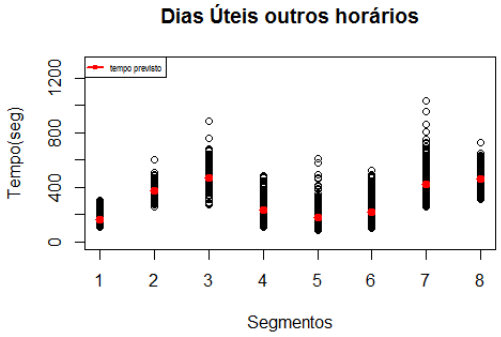
(a)


Dias Úteis 9-16h
tempo previsto
Tempo(seg)
0
400
800
1200
1
2
3
4
5
6
7
8
Segmentos

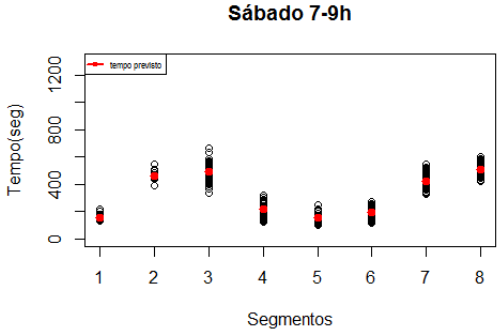
(b)


Dias Úteis 16-19h
tempo previsto
Tempo(seg)
0
400
800
1200
1
2
3
4
5
6
7
8
Segmentos

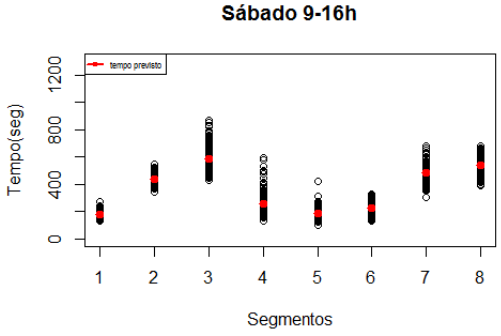
(c)


Dias Úteis outros horários
tempo previsto
Tempo(seg)
0
400
800
1200
1
2
3
4
5
6
7
8
Segmentos

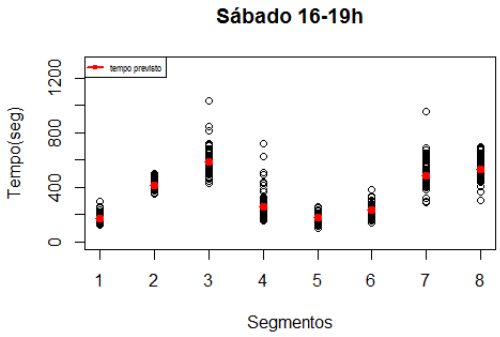
(d)


Sábado 7-9h
tempo previsto
Tempo(seg)
0
400
800
1200
1
2
3
4
5
6
7
8
Segmentos

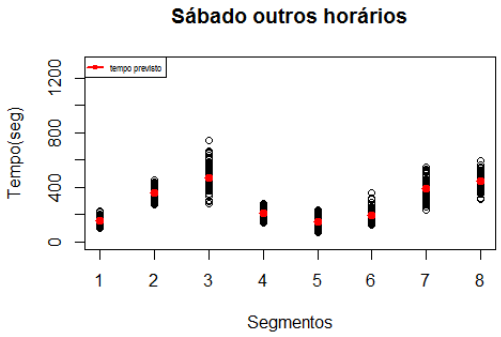
(e)


Sábado 9-16h
tempo previsto
Tempo(seg)
0
400
800
1200
1
2
3
4
5
6
7
8
Segmentos

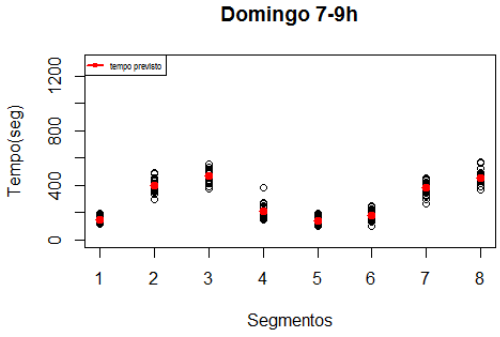
(f)



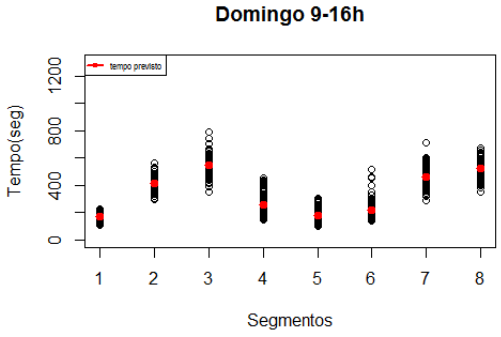
(g)



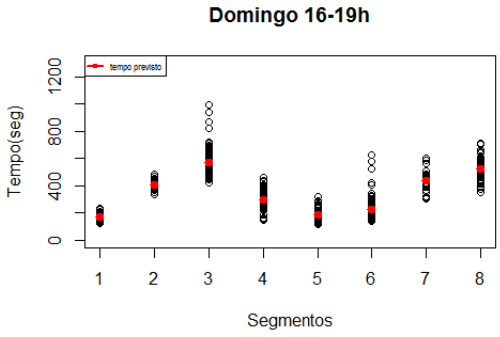
(h)



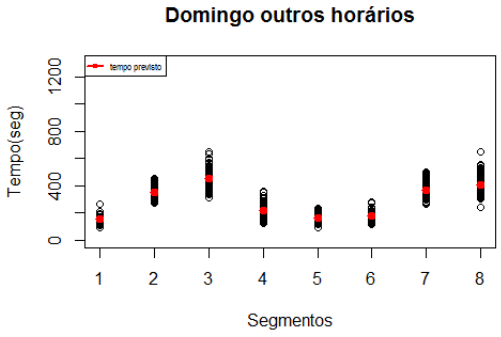
(i)



(j)



(k)



(l)

Figura 5.3: Previsão dos Tempos usando o MMQG a) Dias Úteis 7-9h (b) Dias Úteis 9-16h (c) Dias Úteis 16-19h (d) Dias Úteis outros horários (e) Sábado 7-9h (f) Sábado 9-16h (g) Sábado 16-19h (h) Sábado outros horários (i) Domingo 7-9h (j) Domingo 9-16h (k) Domingo 16-19h (l) Domingo outros horários

Capítulo 6

Máquinas de Suporte Vectorial

Neste capítulo é apresentada alguma teoria sobre Máquinas de Suporte Vetorial (MSV[1]) e descrita a sua aplicação ao caso de estudo. A forma mais simples de descrever as Máquinas de Suporte Vectorial é considerar um problema de classificação com duas classes, onde o objetivo é encontrar um hiper-plano que separe as observações em duas classes. Contudo, existem inúmeros hiper-planos que verificam esta condição. Com este modelo pretende-se escolher o hiper-plano que maximiza a distância à observação mais próxima, ou seja, aquele que tem uma maior margem (Torgo, 2010).

6.1 Pressupostos Teóricos

A ideia básica por detrás das MSV é a de enviar os dados originais para um novo espaço, de alta dimensão, onde seja possível a aplicação de modelos lineares para obter um plano de separação, por exemplo, separando em classes, no caso de um problema de classificação. O envio dos dados originais para este novo espaço é obtido com a ajuda das funções núcleo[2]. As MSV são máquinas lineares que operam em representações duais induzidas por funções do núcleo. (Torgo, 2010)

A separação feita pelo hiperplano na nova representação dual é feita com frequência, maximizando a margem de separação entre os casos pertencentes a diferentes classes. Este é um problema de otimização frequentemente resolvido através de métodos de programação quadrática. Métodos com margens suaves permitem que uma pequena proporção de casos estejam do lado ”errado”da margem, cada um deles associado a um certo ”custo”. (Torgo, 2010)

As MSV podem ser adaptadas para a regressão de uma resposta quantitativa com a introdução de uma função de penalização (exemplo na figura 6.1), herdando algumas das propriedades do classificador das MSV. (Hastie et al., 2001). O objetivo é encontrar uma função linear por pedaços $f(x)$ que tenha no máximo um erro de ε para todas as

[1] SVM - Support Vector Machines
[2] Kernel functions

observações do conjunto de treino. (Smola and Schölkopf, 2004)

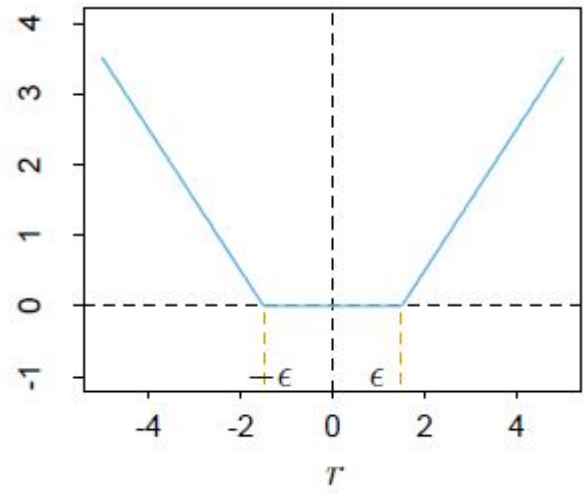


Figura 6.1: Exemplo de uma função de penalização linear por pedaços (Fonte: Hastie et al. (2001))

Considere-se o conjunto de dados de treino $\{(x_1, y_1), (x_2, y_2), ..., (x_l, y_l)\} \subset X \times \mathbb{R}$, onde X é o espaço das variáveis de entrada, x_i é um vetor de entrada de dimensão n, como por exemplo o tempo de viagem do segmento atual, y_i é o valor pretendido, tal como o tempo de viagem observado do próximo segmento. Considerando o caso em que *f* é uma função linear, ela pode ser descrita da seguinte forma: (Bin et al., 2006, Smola and Schölkopf, 2004)

$$f(x) = \langle w, x \rangle + b \tag{6.1}$$

com $w \in X$, $b \in \mathbb{R}$, w representa os vetores de suporte e $\langle ., . \rangle$ denota o produto interno. O objetivo é que w tome valores pequenos. Uma maneira de assegurar tal facto é minimizar a norma, ou seja, $\|w\|^2 = \langle w, w \rangle$, vamos ter uma função o mais plana possível (Bin et al., 2006). O problema pode ser escrito como um problema de otimização: (Smola and Schölkopf, 2004)

$$\begin{aligned} & minimizar\ \frac{1}{2}\|w^2\| \\ s.a.\ & \begin{cases} y_i - \langle w, x_i \rangle - b \leq \varepsilon \\ \langle w, x_i \rangle + b - y_i \leq \varepsilon \end{cases} \end{aligned} \tag{6.2}$$

A formulação feita pressupõe que a função *f(x)* realmente existe e que aproxima todos os pares (x_i, y_i) com uma precisão ε. Quando o pressuposto não é verificado é necessário introduzir variáveis de afrouxamento ξ_i, ξ_i^* de modo a permitir que alguns pontos estejam do lado errado. Geometricamente, pode-se visualizar o problema (figura 6.2) como um tubo de tamanho 2ε em torno da função $f(x)$ e cujos pontos que estão fora do mesmo são erros.(Smola and Schölkopf, 2004, Bin et al., 2006)

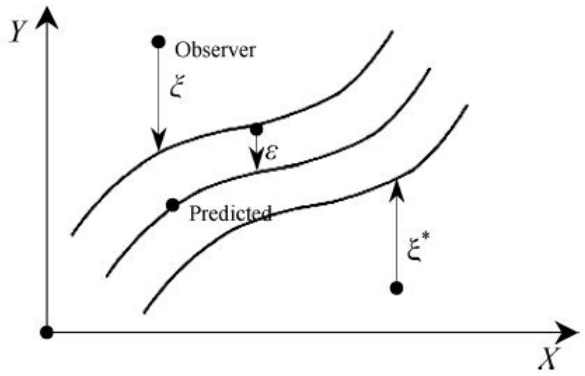


Figura 6.2: Valores previstos e valores observados (Fonte: Bin et al. (2006))

Com a introdução das variáveis de afrouxamento, obtém-se a seguinte formulação do problema: (Smola and Schölkopf, 2004)

$$minimizar\ \frac{1}{2}\|\omega\|^2 + C\frac{1}{l}\sum_{i=1}^{l} L_\varepsilon(y_i, f(x_i)) \qquad (6.3)$$
$$s.a.\ \begin{cases} y_i - \langle w, x_i\rangle - b \leq \varepsilon + \xi_i \\ \langle w, x_i\rangle + b - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases}$$

C é a chamada constante de regularização. O primeiro termo, $\|\omega\|^2$, é chamado de termo regularizado. O segundo termo $\frac{1}{l}\sum_{i=1}^{l} L_\varepsilon(y_i, f(x_i))$ é o erro empírico medido pela função de perda ε-insensível, o qual é definido a seguir: (Hastie et al., 2001)

$$L_\varepsilon(y_i, f(x_i)) = \begin{cases} |y_i - f(x_i)| - \varepsilon,\ |y_i - f(x_i)| \geq \varepsilon \\ 0,\ \text{outros casos} \end{cases} \qquad (6.4)$$

Na figura 6.3 está esquematizado o processo, apenas os pontos fora da região sombreada contribuem para o custo. (Smola and Schölkopf, 2004)

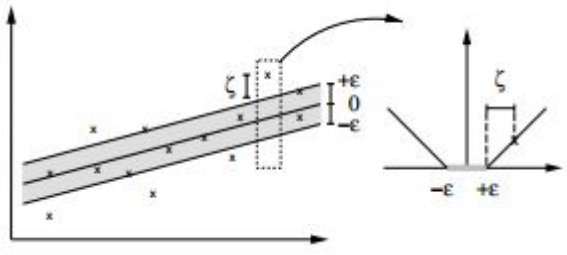


Figura 6.3: Introdução de variáveis de afrouxamento em MSV linear (Fonte: Smola and Schölkopf (2004))

O problema de otimização (6.3) pode ser resolvido mais facilmente na sua forma dual, para isso é necessário introduzir os multiplicadores de Lagrange, procedendo da se-

guinte forma:

$$L = \frac{1}{2}\|w^2\| + C\sum_{i=1}^{l} \xi_i + \xi_i^* - \sum_{i=1}^{l} \eta_i\xi_i + \eta_i^*\xi_i^* \tag{6.5}$$
$$- \sum_{i=1}^{l} \alpha_i(\varepsilon + \xi_i - y_i + \langle w, x_i\rangle + b)$$
$$- \sum_{i=1}^{l} \alpha_i^*(\varepsilon + \xi_i^* + y_i - \langle w, x_i\rangle - b)$$

onde L é o Lagrangeano e $\eta_i, \eta_i^*, \alpha_i, \alpha_i^*$ são multiplicadores de Lagrange. Assim, as variáveis duais em (6.5) têm de satisfazer as restrições:
(Smola and Schölkopf, 2004, Bin et al., 2006)

$$\eta_i, \eta_i^*, \alpha_i, \alpha_i^* \geq 0 \tag{6.6}$$

Decorre da condição de ponto de sela que as derivadas parciais de L em relação às variáveis primais (w, b, ξ_i, ξ_i^*) são:

$$\partial_b L = \sum_{i=1}^{l}(\alpha_i + \alpha_i^*) = 0 \tag{6.7}$$

$$\partial_\omega L = \omega - \sum_{i=1}^{l}(\alpha_i + \alpha_i^*)(x_i) = 0 \tag{6.8}$$

$$\partial_{\xi_i} L = C - \alpha_i - \eta_i = 0 \tag{6.9}$$

$$\partial_{\xi_i^*} L = C - \alpha_i^* - \eta_i^* = 0 \tag{6.10}$$

Substituindo (6.7), (6.8), (6.9) e (6.10) em (6.6) obtemos o problema de otimização dual:

$$maximizar\ \{-\frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i + \alpha_i^*)(\alpha_j - \alpha_j^*)\langle x_i, x_j\rangle - \varepsilon\sum_{i=1}^{l}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{l} y_i(\alpha_i + \alpha_i^*)\} \tag{6.11}$$

sujeito a:

$$\sum_{i=1}^{l}(\alpha_i + \alpha_i^*) = 0 \text{ e } \alpha_i, \alpha_i^* \in [0, C] \tag{6.12}$$

De (6.8) podemos obter:

$$\omega = \sum_{i=1}^{l}(\alpha_i + \alpha_i^*)x_i \tag{6.13}$$

assim,

$$f(x) = \sum_{i=1}^{l} (\alpha_i + \alpha_i^*)\langle x_i, x\rangle + b \tag{6.14}$$

Para os problemas não lineares é necessário enviar os dados para um novo espaço de características Φ. Para resolver o problema do mapeamento dos dados é introduzida a função núcleo $K(x_i, x_j)$. O valor de $K(x_i, x_j)$ é igual ao produto interno de dois vetores, $K(x_i, x_j) = \langle \phi(x_i), \phi(x_j)\rangle$. Através do uso de núcleos, todos os cálculos necessários podem ser efetuados diretamente no espaço de entrada, sem ter de calcular a função $\phi(x)$. Algumas funções núcleo mais conhecidas estão na Tabela 6.1. Usando diferentes funções núcleo, pode-se construir diferentes máquinas. (Bin et al., 2006)

Linear	$K(x_i, x_j) = \langle x_i, x_j\rangle$	
Polinomial	$K(x_i, x_j) = \langle x_i, x_j + 1\rangle^d$	d
Radial	$K(x_i, x_j) = exp(-\gamma\|x_i - x_j\|^2)$	$\gamma > 0$
Sigmoide	$K(x_i, x_j) = tanh(b\langle x_i, x_j\rangle + c)$	b, c

Tabela 6.1: Várias funções núcleo

Do desenvolvimento das equações (6.13) e (6.14) obtemos:

$$\omega = \sum_{i=1}^{l} (\alpha_i + \alpha_i^*)\Phi(x_i) \tag{6.15}$$

e

$$f(x) = \sum_{i=1}^{l} (\alpha_i + \alpha_i^*)K(x_i, x_j) + b \tag{6.16}$$

6.2 Resultados

Para a aplicação das Máquinas de Suporte Vetorial foi usada a função *svm()* da biblioteca *e1071* (Meyer et al., 2014) do *software* R (R Core Team, 2014).

O conjunto de dados foi dividido em dois subconjuntos: subconjunto de treino e subconjunto de teste, em 70% e 30% das observações, respetivamente. O conjunto de treino foi usado para a construção do modelo de previsão, já o conjunto de teste foi utilizado para medir a performance do modelo obtido. Os dois conjuntos são disjuntos de forma a assegurar a fiabilidade do modelo. Esta divisão serve também para evitar problemas de sobreajustamento dos dados.

O primeiro passo consiste da escolha dos melhores parâmetros para os vários núcleos; para isso recorreu-se à função *tune.svm()* da biblioteca *e1071* (Meyer et al., 2014) do *software* R (R Core Team, 2014).

Esta função escolhe os melhores parâmetros para cada núcleo, no sentido em que identifica os parâmetros que produzem o menor erro quadrático médio utilizando validação cruzada, ou seja, obter K subconjuntos de igual tamanho e aleatórios dos dados de treino. Para cada um desses subconjuntos K, construir um modelo usando os restantes K-1 conjuntos e avaliar este modelo no sub-conjunto K^{th}. Guarda-se o desempenho do modelo e repete-se este processo para todos os restantes sub-conjuntos. No final, temos K medidas de desempenho, todas obtidas de dados não utilizados na construção do modelo em causa. A estimativa de validação cruzada é a média destas K medidas (Torgo (2010)). O valor de K usado foi de K=10. Na tabela 6.2 estão os melhores parâmetros para cada núcleo e o intervalo em que os valores variaram.

	Grau	Grau Opt	Gama	Gama Opt	Custo	Custo Opt
Núcleo Polinomial	$2:5$	3	$10^{(-6:-2)}$	0.01	$10^{(1:4)}$	10000
Núcleo Linear	-	-	-	-	$10^{(1:4)}$	10000
Núcleo Radial	-	-	$10^{(-6:-2)}$	0.01	$10^{(1:4)}$	10000
Núcleo Sigmoid	-	-	$10^{(-6:-2)}$	0.001	$10^{(1:4)}$	10000

Tabela 6.2: Parâmetros ótimos para cada núcleo e o respetivo intervalo de pesquisa

Usando os parâmetros ótimos da tabela 6.2 foi executado o modelo para cada núcleo. A tabela 6.3 e o gráfico da figura 6.4 descrevem os erros para cada segmento usando os vários núcleos. A medida de desempenho usada foi o erro absoluto médio (EAM [3]) (Yu et al. (2011))

$$EAM_l = \frac{\sum |t_{l,n}^{running} - \widehat{t}_{l,n}^{running}|}{N} \tag{6.17}$$

em que l designa o segmento, $n = tipo\_dia * periodo\_dia$, N o número total de observações. $t^{running}$ o tempo de viagem observado e $\widehat{t}^{running}$ o tempo de viagem previsto.

	1	2	3	4	5	6	7	8
Núcleo Polinomial	27,5	34,3	63,9	62	39,9	39,9	54,9	50
Núcleo Linear	27,3	34,3	63,9	62	39,9	39,9	54,9	50
Núcleo Radial	27,3	34,3	63,9	62	39,9	39,9	54,9	50
Núcleo Sigmoid	40	39,8	76,8	65,1	44,9	41,7	60,7	62,8

Tabela 6.3: Valores dos EAM para os diferentes núcleos usados MSV

[3]MAE - Mean Absolute Error

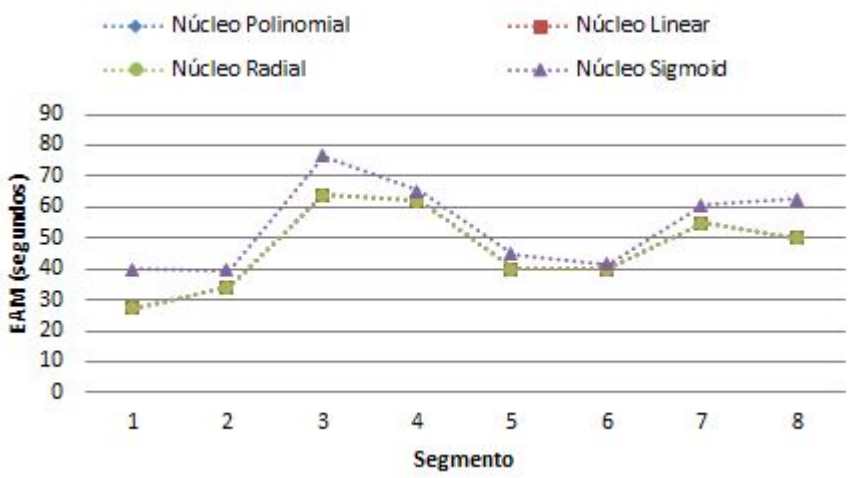
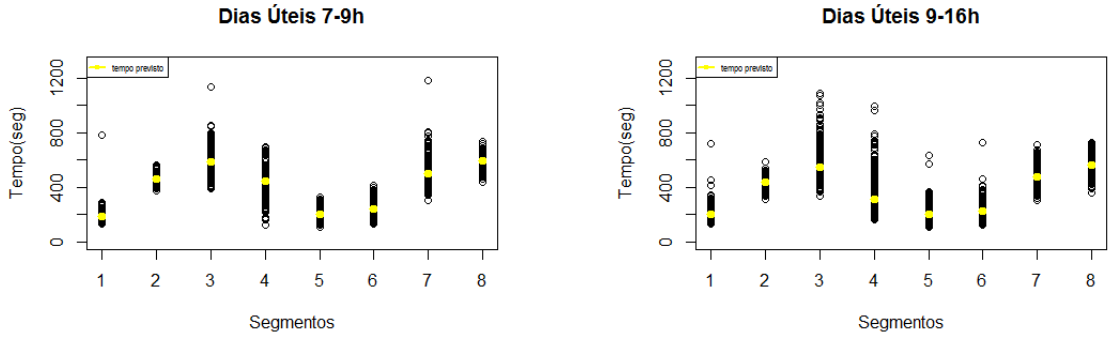


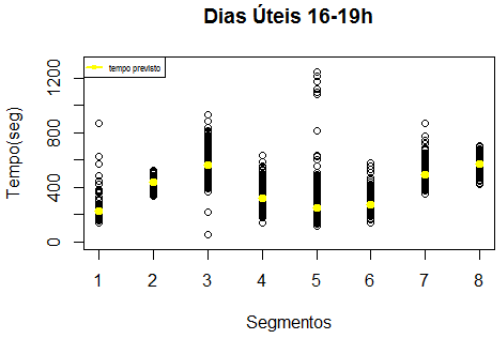
Figura 6.4: EAM para a previsão usando máquinas de suporte vetorial para diferentes núcleos

Os modelos que obtiveram menor erro (EAM), foram aqueles em que se usou o núcleo linear e radial. De entre estes dois modelos, o da função linear precisou de 9177 vetores de suporte, enquanto que o que utilizou a função radial necessitou de 9168 vetores de suporte. Assim sendo, dado o menor número de vetores de suporte, o núcleo radial é o modelo que apresenta melhores resultados.

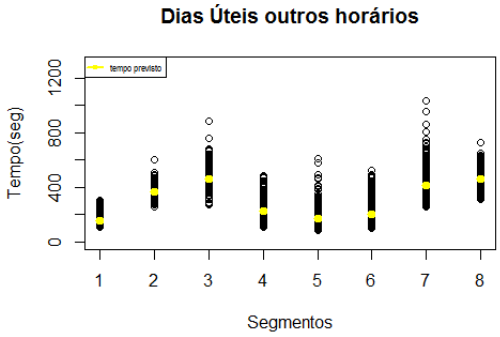
Foram calculados os valores previstos usando o núcleo radial no conjunto de teste; os gráficos da figura 6.5 apresentam o resultado:



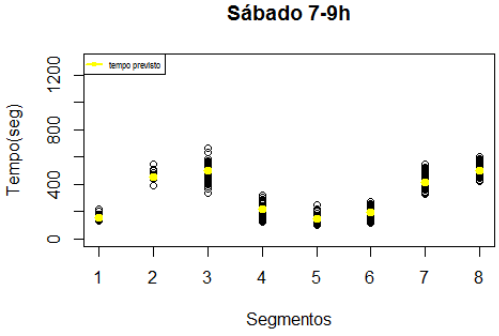
(a) (b)


Dias Úteis 16-19h
tempo previsto
Tempo(seg)
0
400
800
1200
1
2
3
4
5
6
7
8
Segmentos

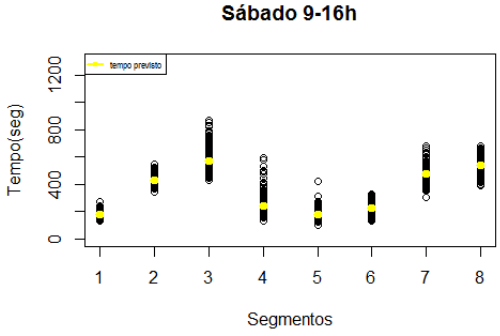
(c)


Dias Úteis outros horários
tempo previsto
Tempo(seg)
0
400
800
1200
1
2
3
4
5
6
7
8
Segmentos

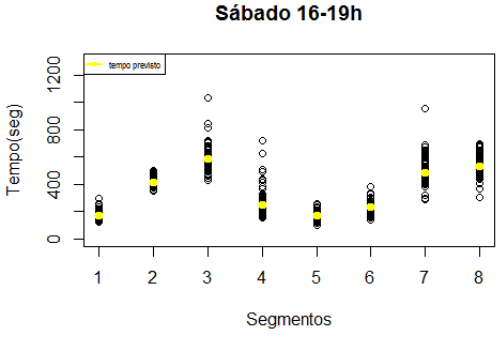
(d)


Sábado 7-9h
tempo previsto
Tempo(seg)
0
400
800
1200
1
2
3
4
5
6
7
8
Segmentos

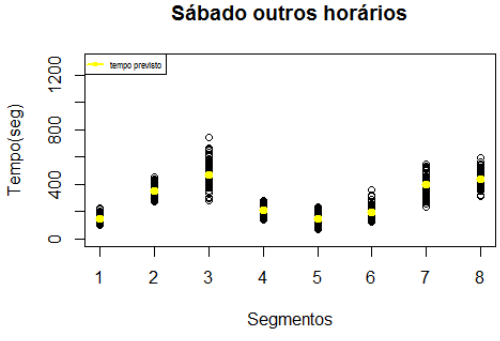
(e)


Sábado 9-16h
tempo previsto
Tempo(seg)
0
400
800
1200
1
2
3
4
5
6
7
8
Segmentos

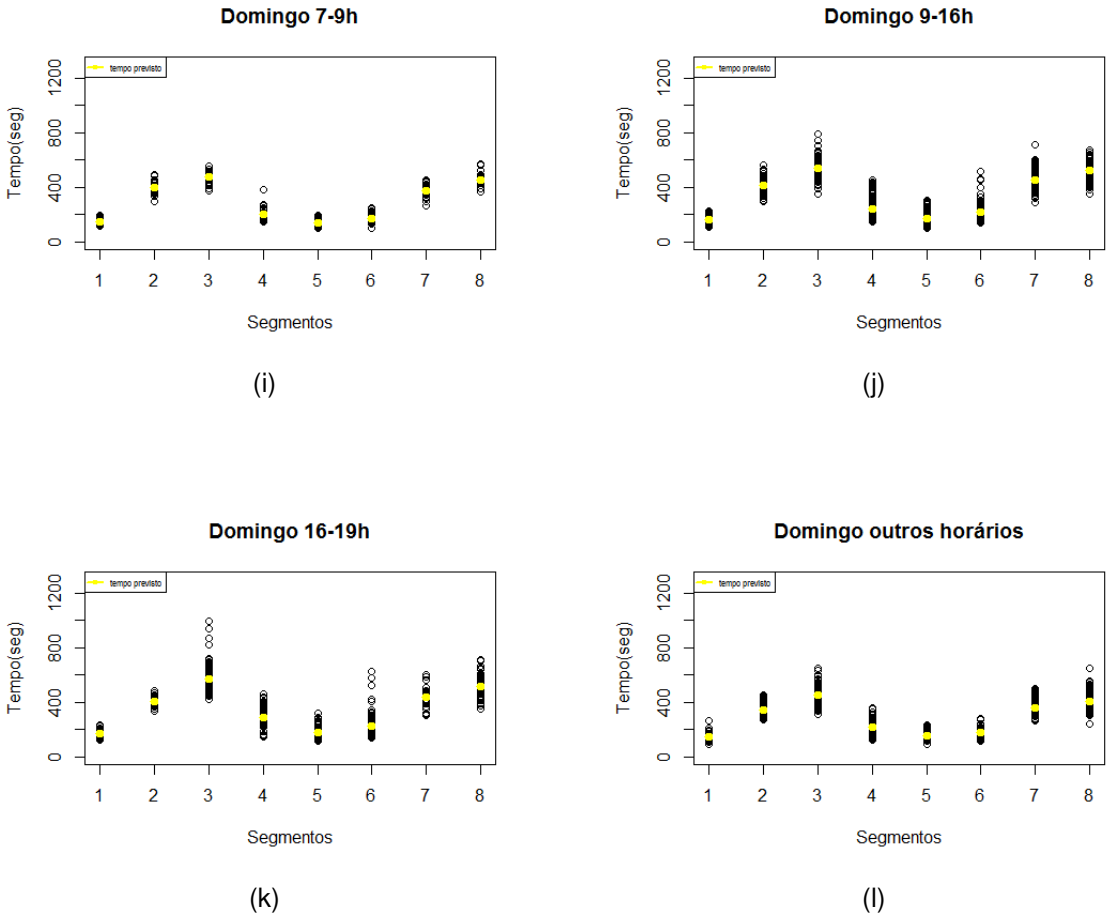
(f)


Sábado 16-19h
tempo previsto
Tempo(seg)
0
400
800
1200
1
2
3
4
5
6
7
8
Segmentos

(g)


Sábado outros horários
tempo previsto
Tempo(seg)
0
400
800
1200
1
2
3
4
5
6
7
8
Segmentos

(h)



(i) (j)

(k) (l)

Figura 6.5: Previsão dos tempos usando MSV a) Dias Úteis 7-9h (b) Dias Úteis 9-16h (c) Dias Úteis 16-19h (d) Dias Úteis outros horários (e) Sábado 7-9h (f) Sábado 9-16h (g) Sábado 16-19h (h) Sábado outros horários (i) Domingo 7-9h (j) Domingo 9-16h (k) Domingo 16-19h (l) Domingo outros horários

Capítulo 7

Discussão de Resultados

Neste capítulo comparam-se os resultados obtidos pelas 5 metodologias utilizadas: 3 métodos empíricos, 1 modelo linear com parâmetros estimados pelo método dos Mínimos Quadrados Generalizados e as máquinas de suporte vetorial. É também feita sempre a comparação com o horário em vigor, designado por horário atual.
As medidas de desempenho usadas para comparar os modelos foram: o erro absoluto médio (EAM[1]) dado pela equação (7.1), erro percentual absoluto médio (EPAM[2]) dado pela equação 7.2 e a raiz do erro quadrático médio (REQM[3]) dado pela equação 7.3

$$EAM_l = \frac{\sum |t_{l,n}^{running} - \widehat{t}_{l,n}^{running}|}{N} \tag{7.1}$$

$$EPAM_l = \frac{1}{N} \sum \frac{|t_{l,n}^{running} - \widehat{t}_{l,n}^{running}|}{t_{l,n}^{running}} \times 100\% \tag{7.2}$$

$$REQM_l = \sqrt{\frac{\sum (t_{l,n}^{running} - \widehat{t}_{l,n}^{running})^2}{N-1}} \tag{7.3}$$

em que l designa o segmento, $n = tipo\_dia * periodo\_dia$, N o número total de observações. $t^{running}$ o tempo de viagem e $\widehat{t}^{running}$ o tempo previsto. (Yu et al., 2011)

O gráfico da figura 7.1 e a tabela 7.1 apresentam o erro absoluto médio de cada um dos 6 modelos em cada um dos 8 segmentos.

[1]MAE - Mean Absolute Error
[2]MAPE - Mean Absolute Percentage Error
[3]RMSE - Root Mean Square Error

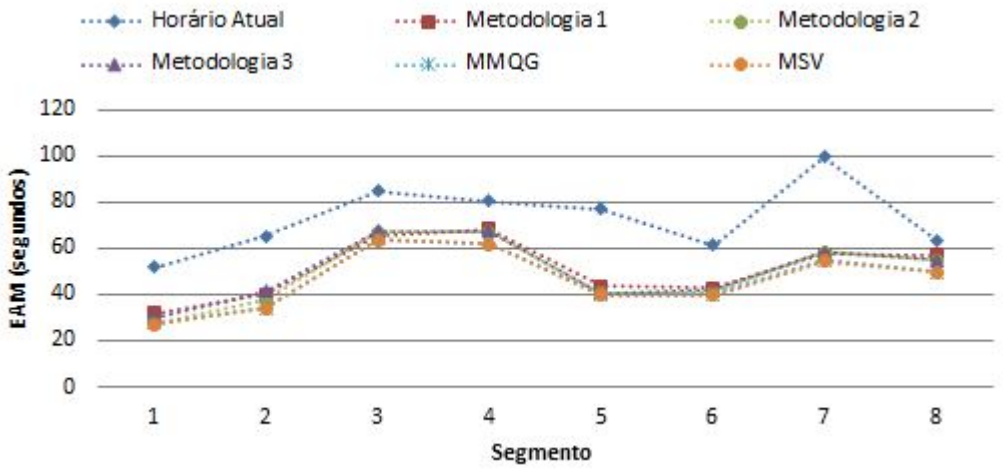


Figura 7.1: EAM de cada uma das metodologias aplicadas em cada um dos 8 segmentos

	Segmentos							
	1	2	3	4	5	6	7	8
Horário Atual	51,9	65,3	84,9	80,5	77	61,5	99,7	63,7
Metodologia 1	32,2	40,4	65,3	68,6	43,8	43,2	57,6	57,1
Metodologia 2	28	37,7	67,4	67,4	41	41,4	58,6	54,9
Metodologia 3	30,3	41,8	67,4	67,4	41	42	58,7	55,2
MMQG	27,8	34,5	64	62,4	40,2	40,3	55	50
Máquinas Suporte Vetorial	27,3	34,3	63,9	62	39,9	39,9	54,9	50

Tabela 7.1: EAM de cada uma das metodologias aplicadas em cada um dos 8 segmentos

Os cinco modelos criados têm um erro absoluto médio inferior ao horário em vigor. Entre os métodos empíricos o modelo 2 é o que apresenta melhores resultados. O Método dos Mínimos Quadrados Generalizado e as Máquinas de Suporte Vetorial apresentam valores bastante semelhantes, mas as Máquinas de Suporte Vetorial apresentam um erro absoluto médio igual ou ligeiramente inferior em todos os nós.

O gráfico da figura 7.2 e a tabela 7.2 apresentam as três medidas de desempenho para os 6 modelos.

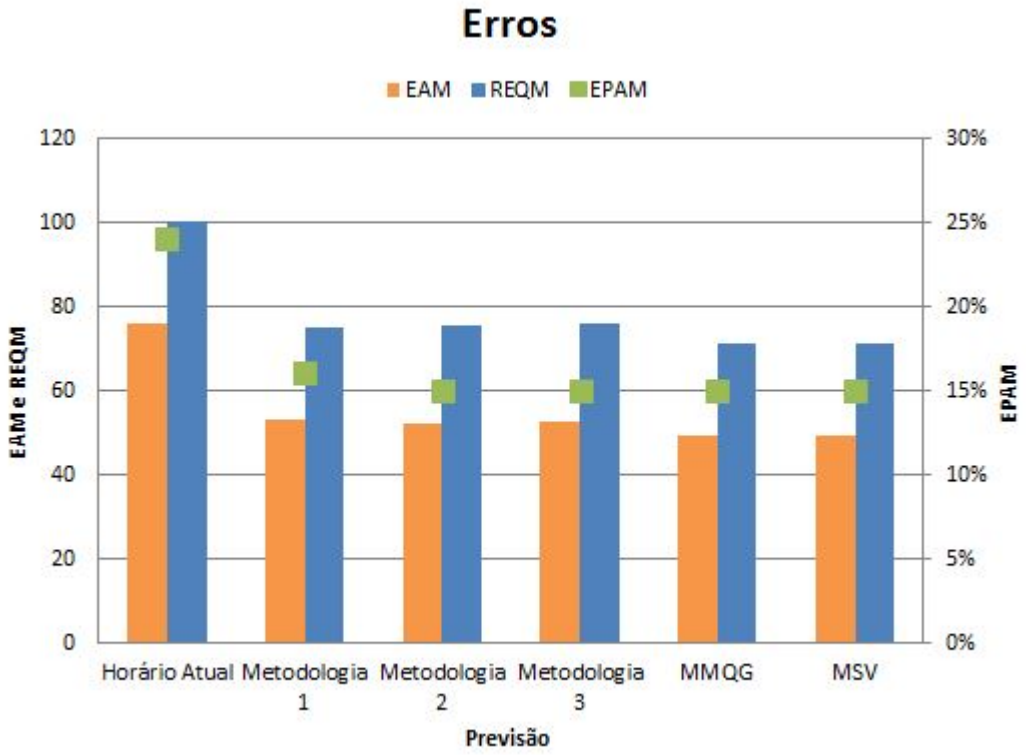


Figura 7.2: Medidas de Desempenho para todas as metodologias

	EAM	EPAM	REQM
Horário Atual	75,86	24%	100,19
Metodologia 1	53,06	16%	75,09
Metodologia 2	51,93	15%	75,33
Metodologia 3	52,51	15%	75,67
Método Mínimos Quadrados Generalizado	49,11	15%	70,9
Máquinas Suporte Vetorial	49	15%	71,3

Tabela 7.2: Medidas de Desempenho para todas as Metodologias

A leitura do gráfico da figura 7.2 é feita da seguinte forma: o eixo vertical esquerdo corresponde às medidas EAM e REQM e o eixo vertical direito corresponde à medida EPAM.

Como já tinha sido verificado por segmento no gráfico 7.1 e na tabela 7.1 o método que apresenta o menor erro absoluto médio são as Máquinas de Suporte Vetorial, seguido do Método dos Mínimos Quadrados Generalizado.

No que respeita ao erro percentual absoluto médio, o método com piores resultados é o do horário atual, sendo a diferença entre o horário atual e o menor erro encontrado de 9%.

A raiz do erro quadrático médio é mais baixo no Método dos Mínimos Quadrados Generalizados.

As medidas de desempenho do gráfico 7.2 (ou da tabela 7.2) sugerem que os dois melhores métodos de previsão são as Máquinas de Suporte Vectorial e o modelo linear estimado pelo Método dos Mínimos Quadrados Generalizado.

A tabela 7.3 contém os erros máximos em valor absoluto, cometidos na previsão por segmento e para cada um dos 6 modelos.

	Segmentos							
	1	2	3	4	5	6	7	8
Horário Atual	644	219	547	696	1122	549	700	265
Metodologia 1	626	253	562	659	950	520	679	289
Metodologia 2	633	251	580	680	973	515	694	297
Metodologia 3	633	250	580	680	973	520	694	297
Mínimos Quadrados Generalizados	639	228	541	679	982	497	675	270
Máquinas Suporte Vetorial	869	601	1131	996	1242	729	1180	739

Tabela 7.3: Erros Máximos (segundos) para as várias previsões, em valor absoluto

O maior erro encontrado, em qualquer um dos segmentos, pertence sempre ao método das Máquinas de Suporte Vetorial. Apesar de ser um método robusto parece não estar a conseguir lidar bem com os *outliers* das observações.

A tabela 7.4 apresenta os coeficientes de correlação de pearson.

	Atual	**Met1**	**Met2**	**Met3**	**MQG**	**MSV**
r	0,8234	0,89	0,8932	0,8924	0,8961	0,8956

Tabela 7.4: Coeficientes de Correlação de pearson (r) entre cada uma das metodologias e o horário observado

O coeficiente de correlação entre cada uma das metodologias e o horário observado reflete a precisão da previsão dos métodos. Todos os métodos estudados, apresentam um coeficiente de correlação entre o valor previsto pela metodologia e o horário real aproximado de 0.9, o que implica que a proporção entre o tempo previsto de viagem para cada modelo é bem ajustada com o tempo de viagem real.

Tendo em conta todos os resultados analisados, o modelo melhor ajustado e com menor erro é o obtido pelo Método dos Mínimos Quadrados Generalizado.

A tabela 7.5 compara as diferenças entre os tempos de viagem do horário atual e o tempo previsto pelo Método dos Mínimos Quadrados Generalizado com os tempos

observados.

	Segmentos			
	Atual		MQG	
	atrasado	adiantado	atrasado	adiantado
0-1m	4602	4192	7140	5681
%	26%	23%	40%	32%
1-2m	3094	2711	2153	1825
%	17%	15%	12%	10%
2-3m	1049	1175	271	524
%	6%	7%	2%	3%
3-4m	348	399	26	165
%	2%	2%	0%	1%
4-5m	112	117	3	73
%	1%	1%	0%	0%
+5m	107	27	3	69
%	1%	0%	0%	0%
+10m	12	0	0	12
%	0%	0%	0%	0%
	9324	8621	9596	8349
	52%	48%	53%	47%

Tabela 7.5: Análise do cumprimento de serviço Atual *vs* MQG

Pela tabela 7.5 é possível ver que a taxa de cumprimento de serviço sobe de 75% para 85% se compararmos o horário atual com o horário obtido usando o Método dos Mínimos Quadrados Generalizados. Usando a escala do indicador de fiabilidade passaria da classificação F para a classificação D.

Com a previsão obtida pelo Método dos Mínimos Quadrados Generalizado a percentagem presente na classe $0 - 1m$ passa de 49% para 62%. A previsão revela-se mais precisa do que o horário em vigor.

Capítulo 8

Trabalho Futuro

Para por em prática o melhor modelo é necessário otimizar o processo, desde a extração da BD, à ordenação dos registos, à previsão usando o método dos MQG, e ao cálculo dos erros. Para agilizar todo o processo, seria interessante implementar uma interface gráfica no *software* GIST que permitisse aceder a qualquer previsão, e que introduzisse os valores previstos diretamente no processamento dos horários. Nas figuras 8.1 e 8.2 estão exemplos da interface a implementar. Os exemplos foram produzidos em Excel. Na figura 8.1 está representado o formulário de seleção e na figura 8.2 um exemplo da seleção.

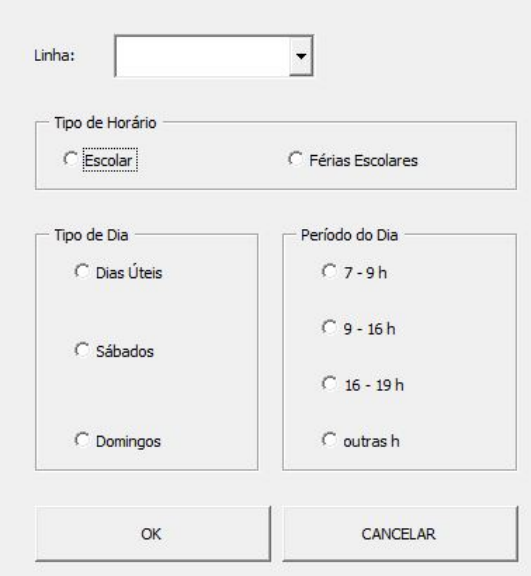


Figura 8.1: Exemplo da interface gráfica a implementar no *software* GIST

Com base nos resultados obtidos pela aplicação de diferentes metodologias de modelação estatística, espera-se que esta nova ferramenta posso contribuir de forma significativa para um melhor desempenho da S.T.C.P. ao serviço prestado aos utentes.

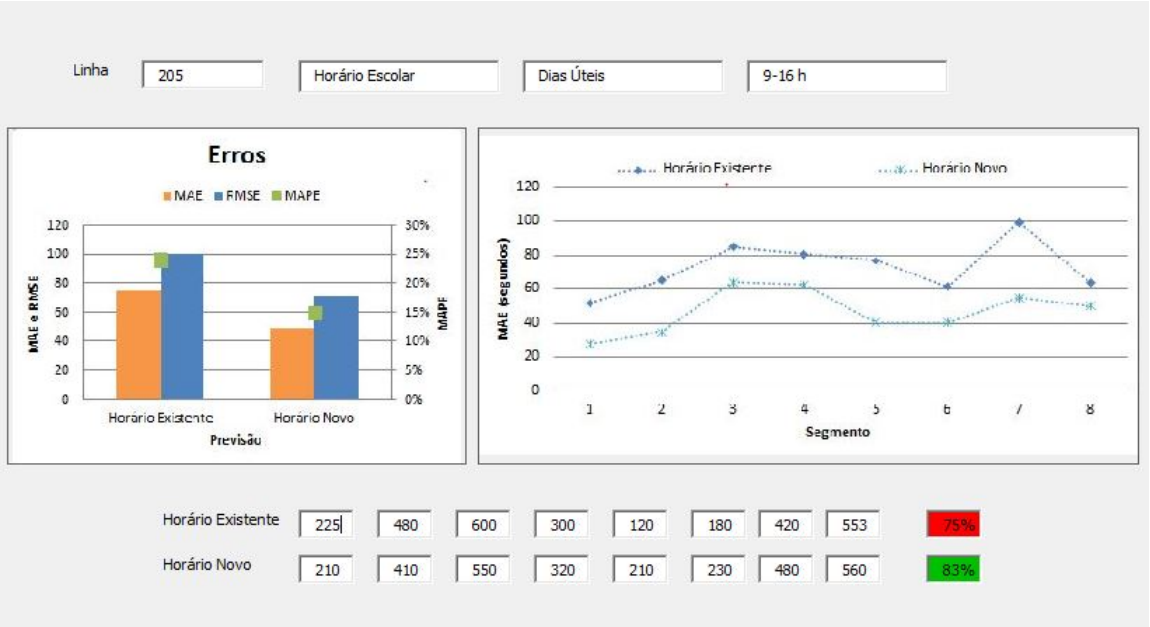


Figura 8.2: Exemplo de pesquisa na interface gráfica

Referências

Bin, Y., Zhongzhen, Y., and Baozhen, Y. (2006). Bus arrival time prediction using support vector machines. *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*, 10(4):151 – 158.

Cabral, M. S. and Gonçalves, M. H. (2011). *Análise de Dados Longitudinais.* Sociedade Portuguesa de Estatística.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA.

Kariya, T. and Kurata, H. (2004). *Generalized Least Squares*. Wiley series in probability and statistics. John Wiley & Sons Ltd, England.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F. (2014). *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*. R package version 1.6-3.

Pinheiro, J. and Bates, D. (2000). *Mixed-Effects Models in S and S-PLUS*. Springer-Verlag, New York.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., and R Core Team (2014). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-117.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Smola, A. J. and Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222.

STCP (2012). *Relatório de Contas*.

Torgo, L. (2010). *Data Mining with R: Learning with Case Studies*. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series. Taylor & Francis.

TRB (2000). *Highway Capacity Manual. Special Report Board. National Research Council*. National Academy of Science, United States of America. ISBN 0-309-06681-6.

Walkenbach, J. (2010). *Excel® 2010 Power Programming with VBA*. Wiley Publishing, Inc, Indianopolis, Indiana. ISBN 978-0-470-47535-5.

Yu, B., Lam, W. H., and Tam, M. L. (2011). Bus arrival time prediction at bus stop with multiple routes. *Transportation Research Part C: Emerging Technologies*, 19(6):1157 – 1170.

Apêndice A

Resultados do Método dos Mínimos Quadrados Generalizado

	Coeficiente	Valor p
	192.966	<0.001
Segmento 2	271.053	<0.001
Segmento 3	397.413	<0.001
Segmento 4	241.653	<0.001
Segmento 5	10.346	0.238
Segmento 6	55.346	<0.001
Segmento 7	312.194	<0.001
Segmento 8	398.652	<0.001
Período (9-16h)	14.077	0.129
Período (16-19h)	37.141	<0.001
Período (outros horários)	-28.653	<0.001
Sábados	-33.106	<0.001
Domingos	-43.529	<0.001
Segmento 2 : Período (9-16h)	-42.637	<0.001
Segmento 3 : Período (9-16h)	-49.978	<0.001
Segmento 4 : Período (9-16h)	-131.578	<0.001
Segmento 5 : Período (9-16h)	-11.876	0.221
Segmento 6 : Período (9-16h)	-30.802	<0.001
Segmento 7 : Período (9-16h)	-41.585	<0.001
Segmento 8 : Período (9-16h)	-45.456	<0.001
Segmento 2 : Período (16-19h)	-64.917	<0.001
Segmento 3 : Período (16-19h)	-61.836	<0.001
Segmento 4 : Período (16-19h)	-146.936	<0.001
Segmento 5 : Período (16-19h)	19.827	0.075
Segmento 6 : Período (16-19h)	-4.758	0.669
Segmento 7 : Período (16-19h)	-40.679	<0.001
Segmento 8 : Período (16-19h)	-51.211	<0.001
Segmento 2 : Período (outros horários)	-62.501	<0.001
Segmento 3 : Período (outros horários)	-90.057	<0.001
Segmento 4 : Período (outros horários)	-168.44	<0.001
Segmento 5 : Período (outros horários)	4.837	0.627
Segmento 6 : Período (outros horários)	-0.049	0.996
Segmento 7 : Período (outros horários)	-56.153	<0.001
Segmento 8 : Período (outros horários)	-104.018	<0.001
Segmento 2 : Sábado	26.698	0.138
Segmento 3 : Sábado	-61.442	<0.001
Segmento 4 : Sábado	-182.197	<0.001
Segmento 5 : Sábado	-16.213	0.266
Segmento 6 : Sábado	-20.985	0.149
Segmento 7 : Sábado	-52.753	<0.001
Segmento 8 : Sábado	-53.289	<0.001

Tabela A.1

(Continuação)		
	Coeficiente	Valor p
Segmento 2 : Domingo	-21.24	0.124
Segmento 3 : Domingo	-75.944	$<$0.001
Segmento 4 : Domingo	-182.278	$<$0.001
Segmento 5 : Domingo	-16.627	0.188
Segmento 6 : Domingo	-26.261	0.042
Segmento 7 : Domingo	-80.825	$<$0.001
Segmento 8 : Domingo	-94.491	$<$0.001
Período (9-16h) : Sábado	9.050	0.534
Período (16-19h) : Sábado	-23.15	0.158
Período (outros horários) : Sábado	23.174	0.111
Período (9-16h) : Domingo	4.965	0.689
Período (16-19h) : Domingo	-14.388	0.345
Período (outros horários) : Domingo	33.097	$<$0.001
Segmento 2 : Período (9-16h) : Sábado	1.973	0.922
Segmento 3 : Período (9-16h) : Sábado	114.133	$<$0.001
Segmento 4 : Período (9-16h) : Sábado	144.501	$<$0.001
Segmento 5 : Período (9-16h) : Sábado	19.859	0.229
Segmento 6 : Período (9-16h) : Sábado	42.093	0.011
Segmento 7 : Período (9-16h) : Sábado	81.646	$<$0.001
Segmento 8 : Período (9-16h) : Sábado	56.381	$<$0.001
Segmento 2 : Período (16-19h) : Sábado	11.555	0.610
Segmento 3 : Período (16-19h) : Sábado	141.489	$<$0.001
Segmento 4 : Período (16-19h) : Sábado	172.926	$<$0.001
Segmento 5 : Período (16-19h) : Sábado	-9.202	0.623
Segmento 6 : Período (16-19h) : Sábado	31.369	0.093
Segmento 7 : Período (16-19h) : Sábado	95.609	$<$0.001
Segmento 8 : Período (16-19h) : Sábado	63.747	$<$0.001
Segmento 2 : Período (outros horários) : Sábado	-31.773	0.114
Segmento 3 : Período (outros horários) : Sábado	72.461	$<$0.001
Segmento 4 : Período (outros horários) : Sábado	162.464	$<$0.001
Segmento 5 : Período (outros horários) : Sábado	-4.551	0.785
Segmento 6 : Período (outros horários) : Sábado	10.175	0.543
Segmento 7 : Período (outros horários) : Sábado	31.654	0.059
Segmento 8 : Período (outros horários) : Sábado	47.155	0.006

Tabela A.2

(Continuação)		
	Coeficiente	Valor p
Segmento 2 : Período (9-16h) : Domingo	41.155	0.011
Segmento 3 : Período (9-16h) : Domingo	104.151	<0.001
Segmento 4 : Período (9-16h) : Domingo	160.846	<0.001
Segmento 5 : Período (9-16h) : Domingo	26.999	0.068
Segmento 6 : Período (9-16h) : Domingo	54.382	<0.001
Segmento 7 : Período (9-16h) : Domingo	99.715	<0.001
Segmento 8 : Período (9-16h) : Domingo	96.752	<0.001
Segmento 2 : Período (16-19h) : Domingo	48.189	0.017
Segmento 3 : Período (16-19h) : Domingo	139.479	<0.001
Segmento 4 : Período (16-19h) : Domingo	209.026	<0.001
Segmento 5 : Período (16-19h) : Domingo	0.813	0.963
Segmento 6 : Período (16-19h) : Domingo	27.456	0.129
Segmento 7 : Período (16-19h) : Domingo	76.847	<0.001
Segmento 8 : Período (16-19h) : Domingo	95.259	<0.001
Segmento 2 : Período (outros horários) : Domingo	9.400	0.566
Segmento 3 : Período (outros horários) : Domingo	71.625	<0.001
Segmento 4 : Período (outros horários) : Domingo	175.079	<0.001
Segmento 5 : Período (outros horários) : Domingo	10.782	0.475
Segmento 6 : Período (outros horários) : Domingo	-0.634	0.967
Segmento 7 : Período (outros horários) : Domingo	40.820	0.009
Segmento 8 : Período (outros horários) : Domingo	51.715	0.001

Tabela A.3: Resultados da estimação usando o método dos MQG

Apêndice B

Horário Atual vs Horário Proposto

	Atual	MQG	Diferença
(1) DU 7h9	225	190	-35
(1) DU 9h16	225	210	-15
(1) DU 16h19	225	230	5
(1) DU outros horários	225	160	-65
(1) S 7h9	225	160	-65
(1) S 9h16	225	180	-45
(1) S 16h19	225	170	-55
(1) S outros horários	225	150	-75
(1) D 7h9	135	150	15
(1) D 9h16	135	160	25
(1) D 16h19	225	170	-55
(1) D outros horários	225	150	-75
(2) DU 7h9	480	460	-20
(2) DU 9h16	480	440	-40
(2) DU 16h19	480	440	-40
(2) DU outros horários	480	370	-110
(2) S 7h9	480	460	-20
(2) S 9h16	480	440	-40
(2) S 16h19	480	420	-60
(2) S outros horários	480	360	-120
(2) D 7h9	420	400	-20
(2) D 9h16	420	410	-10
(2) D 16h19	480	410	-70
(2) D outros horários	300	350	50
(3) DU 7h9	600	590	-10
(3) DU 9h16	600	550	-50
(3) DU 16h19	600	570	-30
(3) DU outros horários	600	470	-130
(3) S 7h9	600	500	-100
(3) S 9h16	600	580	-20
(3) S 16h19	600	590	-10
(3) S outros horários	480	470	-10
(3) D 7h9	480	470	-10
(3) D 9h16	480	540	60
(3) D 16h19	600	570	-30
(3) D outros horários	600	460	-140
(4) DU 7h9	300	430	130
(4) DU 9h16	300	320	20
(4) DU 16h19	300	320	20
(4) DU outros horários	300	240	-60

Tabela B.1

Continuação			
	Atual	MQG	Diferença
(4) S 7h9	300	220	-80
(4) S 9h16	300	250	-50
(4) S 16h19	300	260	-40
(4) S outros horários	300	210	-90
(4) D 7h9	300	210	-90
(4) D 9h16	300	260	-40
(4) D 16h19	300	290	-10
(4) D outros horários	300	220	-80
(5) DU 7h9	180	200	20
(5) DU 9h16	120	210	90
(5) DU 16h19	120	260	140
(5) DU outros horários	180	180	0
(5) S 7h9	120	150	30
(5) S 9h16	120	190	70
(5) S 16h19	120	180	60
(5) S outros horários	180	150	-30
(5) D 7h9	120	140	20
(5) D 9h16	120	180	60
(5) D 16h19	120	190	70
(5) D outros horários	180	160	-20
(6) DU 7h9	180	250	70
(6) DU 9h16	180	230	50
(6) DU 16h19	180	280	100
(6) DU outros horários	180	220	40
(6) S 7h9	180	190	10
(6) S 9h16	180	230	50
(6) S 16h19	180	230	50
(6) S outros horários	180	200	20
(6) D 7h9	180	180	0
(6) D 9h16	180	220	40
(6) D 16h19	180	220	40
(6) D outros horários	180	180	0
(7) DU 7h9	480	500	20
(7) DU 9h16	420	480	60
(7) DU 16h19	420	500	80
(7) DU outros horários	480	420	-60
(7) S 7h9	600	420	-180
(7) S 9h16	600	480	-120
(7) S 16h19	600	490	-110
(7) S outros horários	600	390	-210

Tabela B.2

Continuação			
	Atual	MQG	Diferença
(7) D 7h9	600	380	-220
(7) D 9h16	600	460	-140
(7) D 16h19	600	440	-160
(7) D outros horários	600	370	-230
(8) DU 7h9	553	590	37
(8) DU 9h16	553	560	7
(8) DU 16h19	553	580	27
(8) DU outros horários	553	460	-93
(8) S 7h9	547	510	-37
(8) S 9h16	547	540	-7
(8) S 16h19	547	530	-17
(8) S outros horários	547	440	-107
(8) D 7h9	547	450	-97
(8) D 9h16	547	520	-27
(8) D 16h19	547	520	-27
(8) D outros horários	387	410	23

Tabela B.3: Hórario atual vs Horário proposto pelo método MQG (segundos)