Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

Previsão, por meio de técnicas de séries temporais e imputação de dados, do volume total de veículos em interseções de Taguatinga para o ano de 2012

Gabriel Gonçalves Mota **08/30054**

Brasília

2012

Gabriel Gonçalves Mota **08/30054**

Previsão, por meio de técnicas de séries temporais e imputação de dados, do volume total de veículos em interseções de Taguatinga para o ano de 2012

Relatório apresentado à disciplina Estágio Supervisionado II do curso de graduação em Estatística, Departamento de Estatística, Instituto de Exatas, Universidade de Brasília, como parte dos requisitos necessários para o grau de Bacharel em Estatística.

Orientador: Prof. Dr. Alan Ricardo da Silva

Brasília

2012

Agradecimentos

Em primeiro lugar à Deus que não cessa de olhar por mim.

Aos queridos pais, à linda irmã e à amada namorada que, apesar das minhas intermináveis murmurações, sempre me apoiaram e acreditaram em mim.

Ao professor Alan e aos demais docentes, os quais, através de suas experiências, foram responsáveis por me impulsionar a enxergar detalhes importantes não abrangidos em livros.

E aos divertidos amigos e companheiros.

"A arte da previsão consiste em antecipar o que acontecerá e depois explicar porque não aconteceu."

Churchill, W.

Resumo

O Volume Médio Diário Anual (VMDA) é um valor muito importante quando discute-se tráfego urbano. A partir dele, os departamentos de trânsito podem adotar diferentes tipos de iniciativa visando sempre a segurança e a fluidez nas vias urbanas. Diretamente ligado ao VMDA, está o volume total de veículos no percorrer do ano, que foi o foco principal de nosso estudo.

Envolvido em um projeto nacional com participação de outras universidades brasileiras, esse trabalho visa prever o volume total de veículos que circulam em certas interseções de estudo. Tal previsão ajudará, entre outros pontos, na decisão de medidas públicas para esses mesmos locais.

As previsões do volume total de veículos no ano serão feitas por séries temporais utilizando o banco de dados fornecido pelo Departamento de Trânsito do Distrito Federal (DETRAN-DF) e observações coletadas em campo por Claude (2012). Entretanto, tal base apresentou número inesperado de *missings* e sub-registros, o que levou à utilização de métodos de imputação a fim de ter observações razoavelmente confiáveis para aplicação das séries temporais.

Palavras-chaves: séries temporais, imputação de dados, interseção, fator de expansão, volume médio diário anual.

Lista de Tabelas

Lista de Figuras

Sumário

Capítulo 1

Introdução

O mercado automobilístico brasileiro vem atraindo muitos consumidores do país. Em grande parte, isso é justificado pela precariedade do transporte público. Uma vez que os brasileiros não podem contar com a pontualidade, conforto e assiduidade de ônibus e metrô, a população procura meios para adquirir um veículo próprio. Tal acontecimento é uma das causas pela qual a frota de carros circulantes no Brasil cresce incessantemente.

Em contrapartida, a largura das vias urbanas, o sistema de segurança viário como semáforos e faixas de pedestres, bem como outros fatores diretamente ligados à movimentação desses veículos permanecem muitas vezes inalterados, sem acompanhar o crescimento do número de automóveis nas ruas.

Um dos motivos para que não haja uma constante modificação na estrutura viária do país é que os órgãos gestores de trânsito atuantes em áreas urbanas não tem, em geral, meios para fazer contagens específicas regulares para a determinação do Volume Médio Diário (VMD) nas vias (trechos e interseções) sob sua jurisdição, pelo menos de um modo abrangente. Segundo o Departamento Nacional de Infraestrutura de Transportes (DNIT, 2006), VMD é o número médio de veículos que, durante um certo período de tempo, percorrem uma seção ou trecho de uma rodovia por dia. Quando não se especifica o período considerado, pressupõe-se que se trata de um ano, assim como será tratado nesse trabalho especificamente.

Atualmente, entretanto, a maioria desses órgãos utiliza equipamentos de fiscalização eletrônica (metrológicos e não-metrológicos) que, além da função específica de fiscalização, armazenam dados de volume agregados em diferentes níveis. No entanto, esses equipamentos apresentam falhas durante a sua operação à longo prazo, caracterizadas por ausência de contagens em alguns períodos e sub-contagens em outros. Normalmente, os maiores problemas ocorrem quando falta energia no local

onde está instalado o aparelho. Por isso, a utilização direta dos dados de volume fornecidos pelos mesmos não é recomendável. Ou seja, é importante que se estabeleçam procedimentos de verificação e ajuste desses dados de modo a permitir sua utilização pelo órgão de trânsito para atividades referentes à:

- revisão e atualização de planos semafóricos de tempo fixo;
- identificação de trechos e interseções críticas por meio do cálculo da taxa de acidentes e taxa de severidade para diferentes locais;
- estimativa de valores de VMD para trechos e interseções não controlados pelos equipamentos de fiscalização por meio da expansão de contagens de curta duração, com o uso de fatores de expansão e fatores de crescimento dos volumes, extraídos de dados coletados por equipamentos localizados nas proximidades do ponto de interesse.

O último tópico apresentado será o mais explorado no decorrer deste trabalho. A ideia é modelar séries temporais a fim de estimar o volume anual de carros para pontos sem equipamento de fiscalização.

1.1 Objetivos

O objetivo geral do trabalho é ajustar modelos de séries temporais nos dados fornecidos por equipamentos eletrônicos do Departamento Estadual de Trânsito do Distrito Federal (DETRAN-DF). A finalidade de tais modelos, além de estimar o volume de carros que circularam em anos anteriores nas interseções onde não existem tais aparelhos, é prever essa mesma informação para o ano em vigência nesses mesmos locais sem observações prévias.

Os objetivos específicos são:

- Executar o método de reamostragem de Jackknife (Rupert G. Miller, 1964), (Efron e Tibshirani, 1994), (Shao e Tu, 1995) e (Cochran, 1977) a fim de identificar sub-registros no banco de dados fornecido pelo DETRAN-DF;
- Realizar imputação de dados nos casos identificados pelo item anterior assim como nos valores faltantes (*missings*);

- Identificar os horários, dias e equipamentos cujos registros são mais bem comportados (segundo critérios a definir) para especificar onde e quando coletar os dados.

Capítulo 2

Contagem volumétrica em interseções

2.1 Introdução

O Código de Trânsito Brasileiro (Brasil, 1997) define interseção como todo cruzamento em nível, entroncamento ou bifurcação, incluindo as áreas formadas por tais cruzamentos, entroncamentos ou bifurcações. Exemplos de interseções são apresentados na Figura 2.1.

Figura 2.1: Exemplos de interseções.
Fonte: *Google Earth*

Para realizar a contagem volumétrica de veículos que circulam nas interseções, pode-se contar com radares fixos, mais conhecidos localmente como "pardais". Como o próprio nome diz, tais radares são afixados em lugares estratégicos das interseções e trabalham continuamente. São alimentados por energia elétrica de corrente alter-

nada e na falta dessa, um *nobreak* integrado possibilita mais um curto período de trabalho. Na seção a seguir, serão introduzidos maiores detalhes sobre tais registradores.

2.2 Equipamentos eletrônicos

O radar fixo mais utilizado no DF que realiza a contagem volumétrica é o Registrador de Infrações de Trânsito 200 (RIT 200) e suas variações (RIT 200F e RIT 200F), fabricados pela ENGEBRAS S/A. Na Figura 2.2, tem-se uma foto do clássico radar. Ele possui uma estrutura resistente a vandalismos e intempéries.

Figura 2.2: Registrador de Infrações de Trânsito 200 (RIT 200).
Fonte: Engebras

Sabe-se que a maior aplicação desse registrador é a fiscalização e controle de tráfego, porém a instalação desse permite também o registro do volume a cada hora de carros nas faixas por ele abrangidas. O controle eletrônico é feito através de um microprocessador. Uma visão do interior do equipamento é ilustrada na Figura 2.3, que pode ser encontrada na Portaria n.° 449 de 19 de novembro de 2009 do Instituto Nacional de Metrologia, Normalização e Qualidade Industrial (INMETRO). O equipamento detecta a passagem de veículos através de sensores instalados na via. O esquema de instalação do equipamento é apresentado nas Figuras 2.4 e 2.5.

Cada uma das três linhas de sensores apresentadas na Figura 2.5 possui bobinas instaladas no interior do asfalto que, conjuntamente, produzem campos eletromagnéticos. Uma vez que os veículos são formados por componentes ferromagnéticos, o campo magnético é desativado no exato momento em que um automóvel passa pelo primeiro sensor e reativado ao acionar o segundo sensor.

Dessa forma, a contagem volumétrica é de fácil obtenção. Basta saber quantas

vezes o campo magnético foi desativado e reativado durante um certo período. Todos os dados são gravados no computador interno do equipamento (Figura 2.3).

A razão de instalar três linhas de sensores é para o cálculo da velocidade do veículo. Suponha que um automóvel esteja na primeira faixa de rolamento. O radar usará então os sensores L1 e L2 para contar o tempo entre as duas linhas. Uma vez que a distância entre esses dois pontos é conhecida, calcula-se a velocidade do veículo, que será recalculada e comparada logo em seguida pelos laços L2 e L3. Comparando as duas velocidades, o computador do equipamento verifica se elas coincidem. Estando corretas e, desde que estejam acima da velocidade permitida para o local, a câmera captura a imagem do veículo infrator.

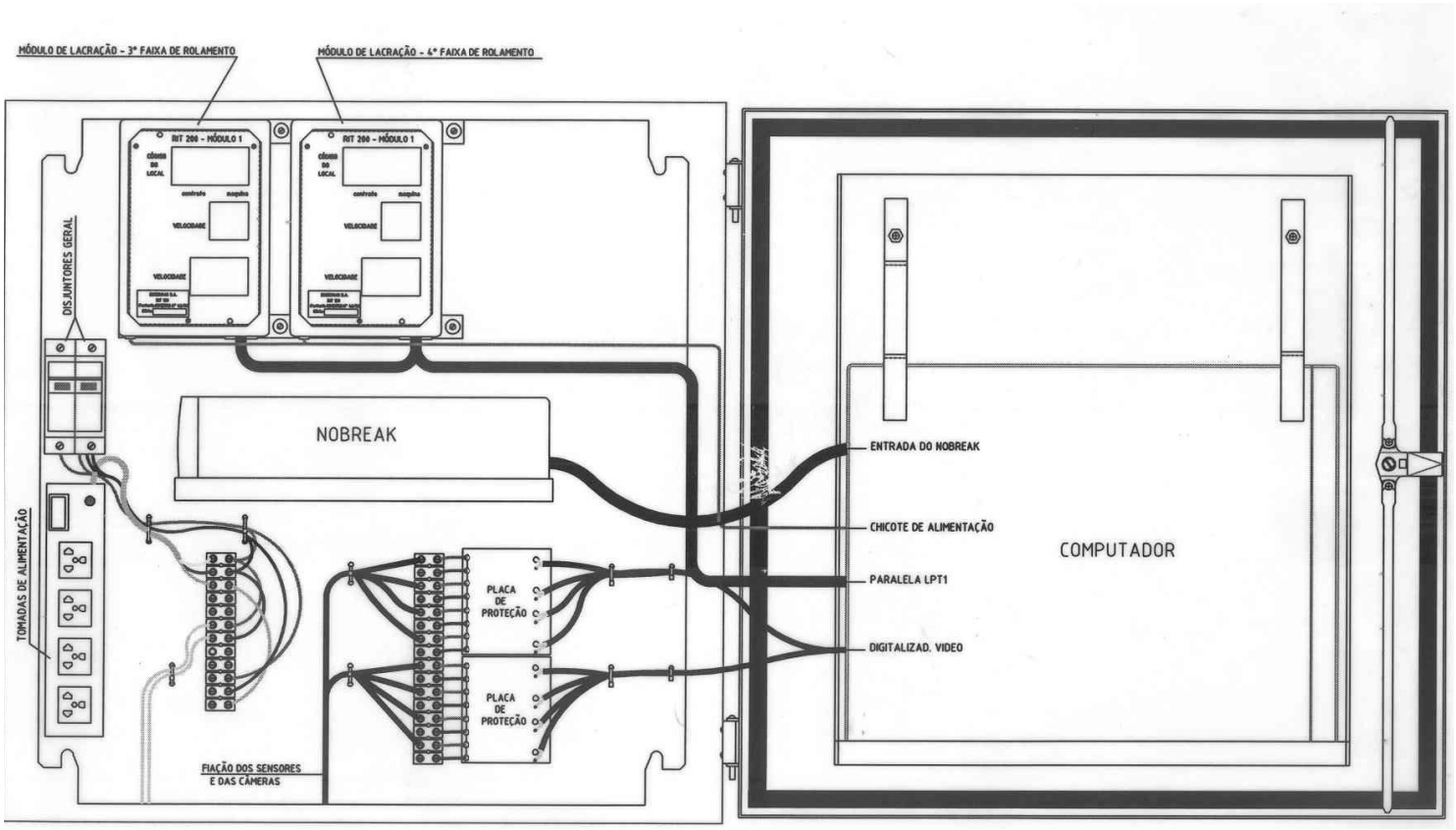


Figura 2.3: Vista do interior do RIT 200.
Fonte: INMETRO

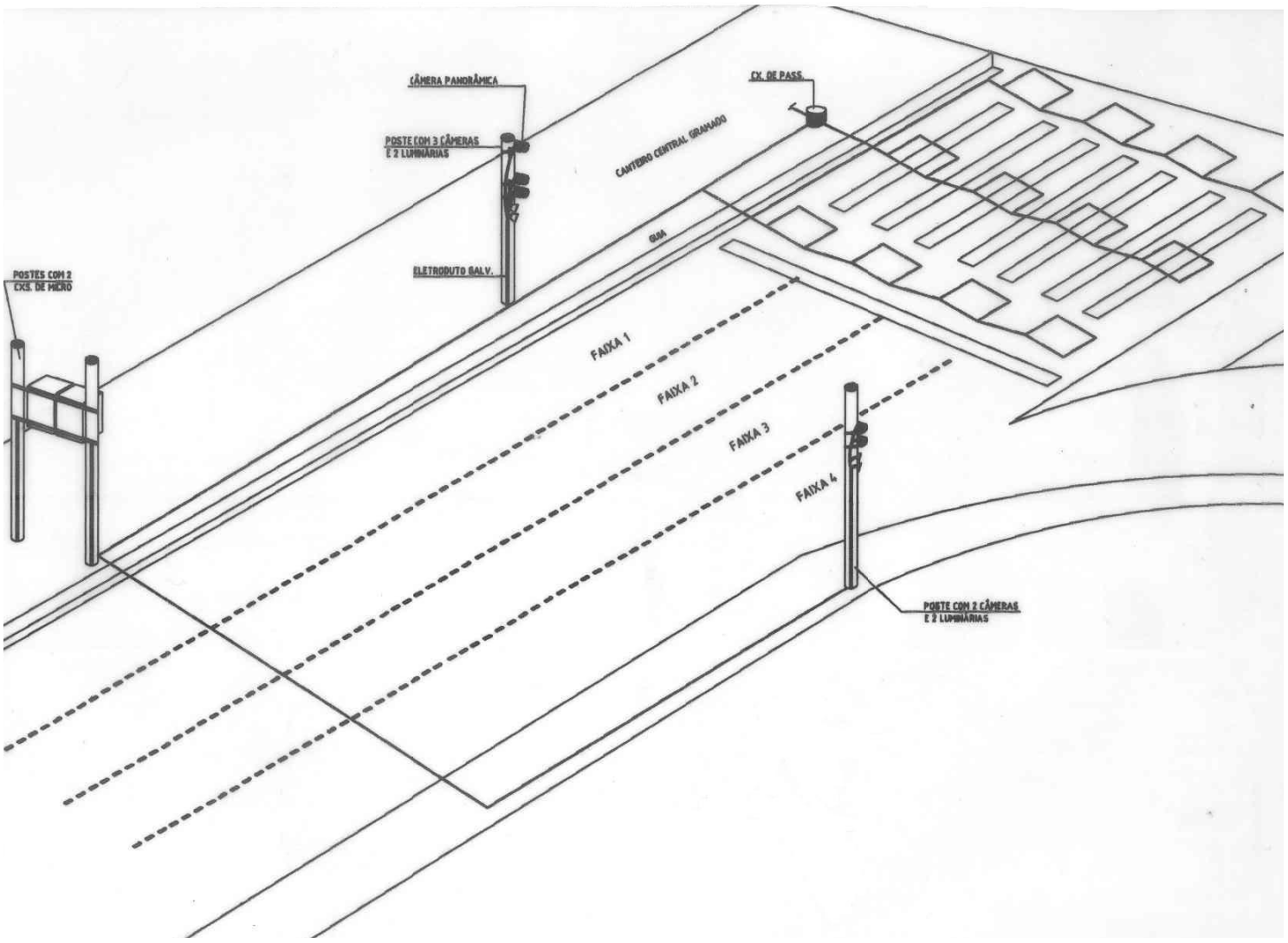


Figura 2.4: Esquema de instalação do RIT 200.
Fonte: INMETRO

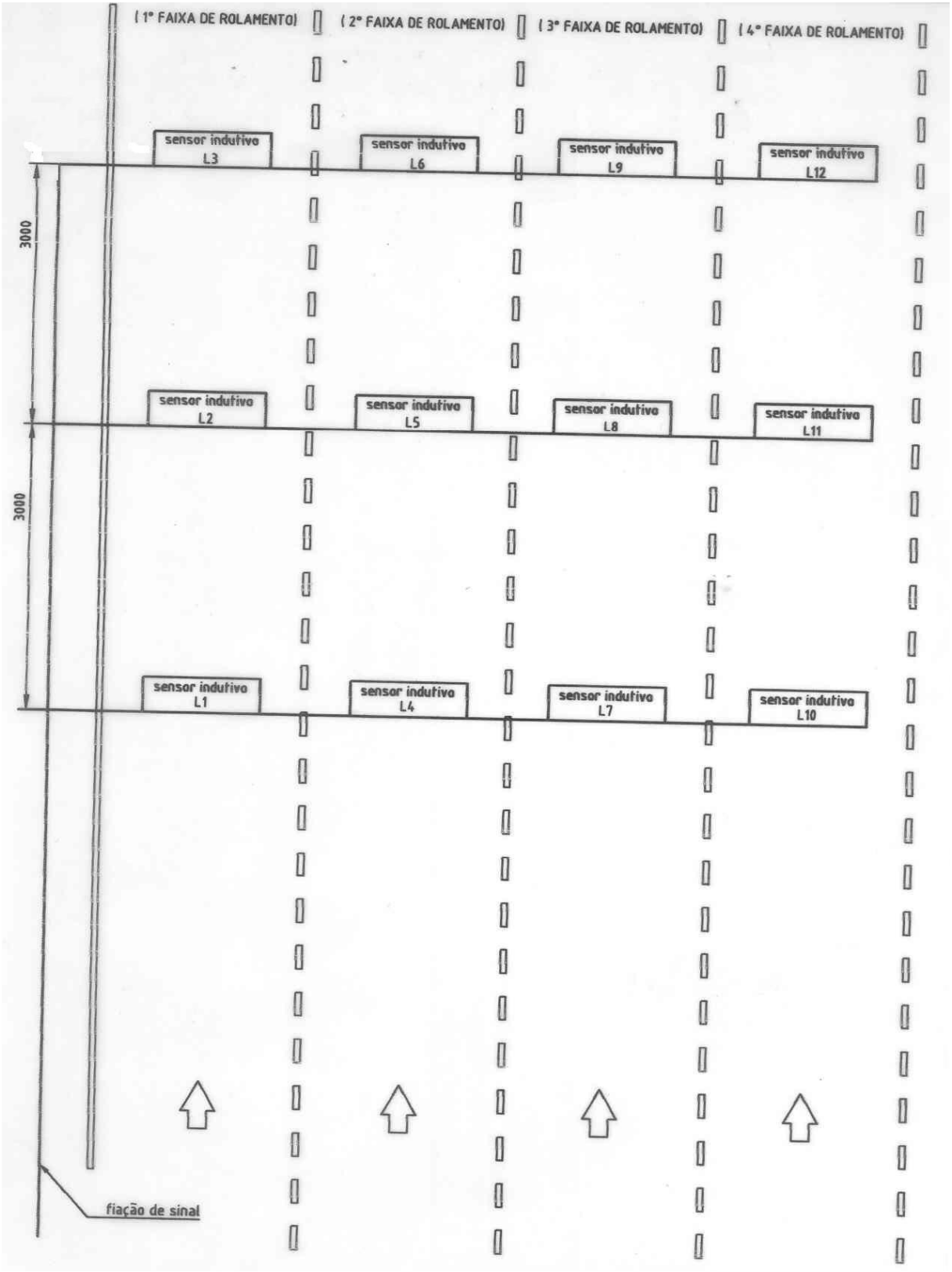


Figura 2.5: Vista superior do esquema de instalação do RIT 200.
Fonte: INMETRO

Capítulo 3

Séries Temporais

3.1 Introdução

Série temporal é todo conjunto de observações que são ordenadas no tempo. Ela pode ser classificada em série *contínua* ou *discreta*. As discretas são aquelas observadas em instantes igualmente espaçados (dias, semanas, meses, entre outros). Por outro lado, as contínuas são séries onde as observações são coletadas sem pausa. O registro de um eletrocardiograma é um exemplo de série contínua. Essas são comumente discretizadas para facilitar a estimação. Pode-se utilizar modelos *paramétricos* e *não-paramétricos* para estimação.

Entre os interesses de estudar uma série temporal X_t, destacam-se:

- fazer previsões de valores futuros;
- analisar o comportamento da série;
- procurar periodicidades interessantes na visão do pesquisador.

A decomposição clássica de uma série temporal X_t é da seguinte forma:

$$X_t = T_t + S_t + \epsilon_t \tag{3.1}$$

sendo que T_t representa a tendência, S_t a sazonalidade e ϵ_t é um *ruído branco*, ou seja, é uma componente aleatória de média zero, variância constante estritamente positiva σ_ϵ^2 e $E(\epsilon_t, \epsilon_h) = 0$. O modelo 3.1 é chamado *aditivo*. Ele é adequado quando uma componente não depende da outra. Caso contrário, usa-se o modelo *multiplicativo*

$$X_t = T_t \times S_t \times \epsilon_t \tag{3.2}$$

que pode ser facilmente transformado em aditivo tomando o logaritmo em ambos os lados da igualdade.

Com os modelos 3.1 e 3.2, o problema que se apresenta é modelar as três componentes. Estima-se então uma mistura de polinômios e funções trigonométricas para representar $f(t) = T_t + S_t$. O inconveniente de tais modelos é que a não correlação entre os erros ϵ_t é raramente verificada nos casos reais.

A metodologia mais usual e prática (intitulada abordagem de Box e Jenkins) consiste em ajustar os modelos paramétricos *auto-regressivos integrados médias móveis* (ARIMA), mais detalhados na seção 3.5. A razão da notoriedade desses modelos é que eles fornecem previsões ótimas no sentido do erro quadrático médio. Segundo Morettin e Toloi (2006), a única desvantagem da utilização da metodologia de Box e Jenkins é que ela requer certa experiência, conhecimento e um *software* onde a técnica esteja bem implementada. Detalhes dessa metodologia são encontrados na própria obra de Box et al. (2008) e no livro de Brockwell e Davis (2002).

3.2 Processos Estocásticos

As séries temporais são, na verdade, observações da trajetória de um processo estocástico, ou seja, observações de uma família de variáveis $(X_t)_{t\in\mathbb{Z}}$. A seguir, é apresentada a definição formal de um processo estocástico.

Definição 1. *Um processo estocástico é uma família $(X_t)_{t\in T}$, tal que, para cada $t \in T$, X_t é uma variável aleatória. T é um conjunto de índices.*

Uma suposição frequente para modelagem de séries temporais é que essa seja *estacionária*.

Definição 2. *Um processo estocástico $(X_t)_{t\in T}$ é fracamente estacionário se e somente se*

(i) $E\{X_t\} = \mu$, *constante para todo* $t \in T$;

(ii) $E\{X_t\} < \infty$, *para todo* $t \in T$;

(iii) $\gamma(t_1, t_2) = Cov\{X_{t_1}, X_{t_2}\}$ *é uma função de* $|t_1 - t_2|$

Caso $(X_t)_{t\in\mathbb{Z}}$ seja um processo estacionário real discreto e de média zero, o fator de autocovariância $\gamma(r) = \text{Cov}\{X_t, X_{t+r}\} = E\{X_t X_{t+r}\}$ satisfaz as propriedades:

(i) $\gamma(0) > 0$;

(ii) $\gamma(-r) = \gamma(r)$;

(iii) $|\gamma(r)| \leq \gamma(0)$;

(iv) $\gamma(r)$ é não negativa definida, ou seja,

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \gamma_{\tau_i - \tau_j} \geq 0 \tag{3.3}$$

para quaisquer números reais $a_1, \ldots, a_n$ e $\tau_1, \ldots, \tau_n$ de $\mathbb{Z}$.

Essas propriedades são de grande uso para realizar o diagnóstico dos modelos ARIMA (Seção 3.5.8).

3.3 Tendência

Pela Definição 2, a estacionariedade de uma série é verificada quando ela tem uma média μ finita e constante, que não dependa do tempo t. Entretanto, grande parte das séries são não-estacionárias nesse aspecto, pois apresentam uma certa *tendência* que pode ou não ser linear. Caso sejam estacionárias, as observações da série flutuam em torno de uma reta constante.

Para exemplificar graficamente essa componente, a Figura 3.1a ilustra uma série sem tendência, já que as observações oscilam em torno de uma média constante próxima ao zero. A Figura 3.1b exemplifica uma série com tendência linear, pois as observações seguem claramente uma reta crescente ao passar do tempo, logo ela é não-estacionária. Essa última figura é o gráfico da venda mensal de passagens aéreas de uma certa companhia entre 1949 e 1960. Logo, era de se esperar que, ao passar dos anos, o número de passageiros aumentasse de uma forma linear, pois tanto a demanda quanto a oferta crescem anualmente.

Caso queira-se eliminar a tendência de uma série, deve-se transformar os dados originais. Para isso, tomam-se *diferenças* sucessivas da série original até eliminar a componente. A diferença de ordem um de $(X_t)_{t\in\mathbb{Z}}$ é definida por:

$$\Delta X_t = X_t - X_{t-1} \tag{3.4}$$

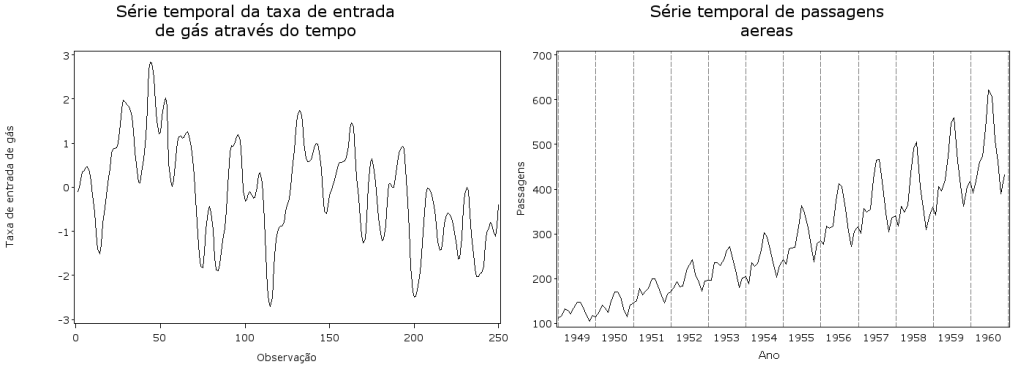
A segunda diferença é

$$\Delta^2 X_t = \Delta\left[\Delta X_t\right] = X_t - 2X_{t-1} + X_{t-2} \tag{3.5}$$

De uma forma geral, a diferença de ordem n:

$$\Delta^n X_t = \Delta \left[\Delta^{n-1} X_t \right] \tag{3.6}$$

Normalmente, com apenas uma ou duas diferenças, consegue-se eliminar a tendência. A Figura 3.2 apresenta o resultado da diferenciação simples da série anteriormente exposta na Figura 3.1b. Percebe-se então que a primeira diferença já foi suficiente para eliminação da tendência.



(a) Série sem tendência (b) Série com tendência

Figura 3.1: Exemplos de Tendência

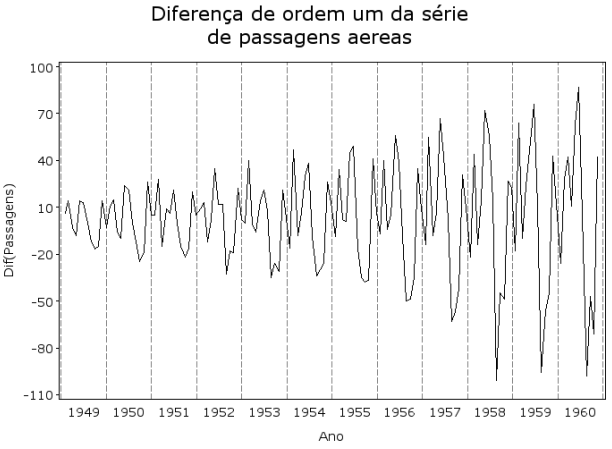
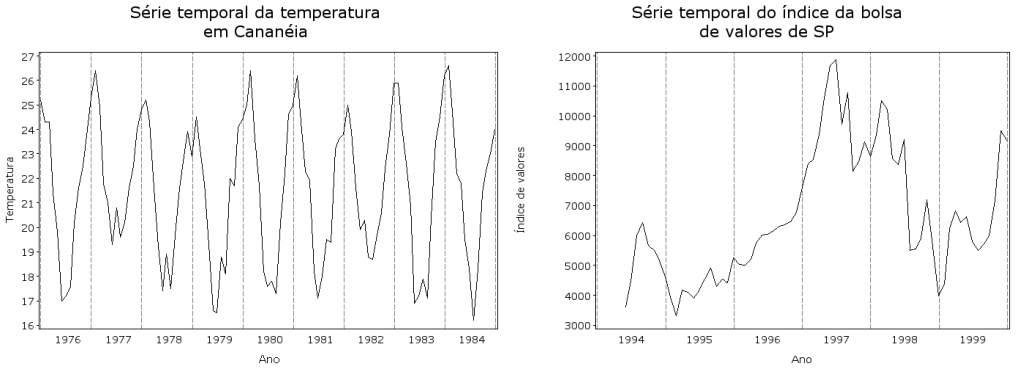


Figura 3.2: Exemplo de Diferenciação

3.4 Sazonalidade

Como o próprio nome diz, *sazonalidade* é um fenômeno que ocorre com certa similaridade em períodos parecidos. Por exemplo, a Figura 3.3a é um caso de série

com sazonalidade, pois nota-se um comportamento parecido a cada ano: a temperatura nos primeiros meses são sempre altas e no meio do ano, baixas. Já o gráfico da bolsa de valores de São Paulo, apresentado na Figura 3.3b, não tem uma sazonalidade clara. A Figura 3.1b é um exemplo de série com tendência e sazonalidade ao mesmo tempo.



(a) Série Com Sazonalidade (b) Série Sem Sazonalidade

Figura 3.3: Exemplos de Sazonalidade

Outro ponto importante de se notar é que a diferenciação torna a série estacionária, mas preserva sua sazonalidade. Tal fato pode ser facilmente verificado com a diferenciação da série de passagens aéreas (Figura 3.2). Observa-se também que tal série é heterocedástica uma vez que sua variância não é constante. Os últimos dados observados variam mais que os primeiros. Como de costume, nesses casos modificam-se os dados utilizando a transformação Box-Cox. Nesse exemplo, ao tomar o logaritmo e diferenciar em seguida, tem-se os dados apresentados na Figura 3.4.

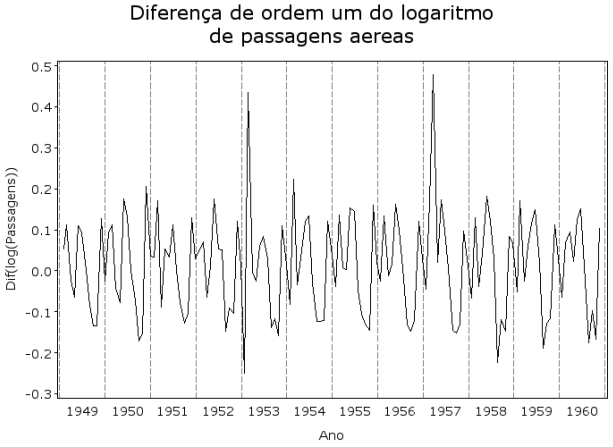


Figura 3.4: Exemplo de transformação logarítmica

3.5 Modelos ARIMA

Como foi dito na Seção 3.1, a metodologia mais usual de análise de séries temporais é ajustar modelos *auto-regressivos integrados médias móveis*, ARIMA(*p,d,q*) (Box et al., 2008). Nessa metodologia, a construção do modelo é feita de forma iterativa. Primeiramente, com base na análise de autocorrelações, identifica-se um modelo, ou seja, os valores de *p, d* e *q* são escolhidos. Em seguida, estima-se os parâmetros para o modelo identificado. Por fim, realiza-se o diagnóstico do modelo ajustado. Essas etapas são repetidas quantas vezes necessárias para comparação de modelos.

3.5.1 Modelos AR

De uma forma geral, o modelo auto-regressivo (AR) de ordem p é apresentado a seguir:

$$\phi_p(B)X_t = \epsilon_t \tag{3.7}$$

Sendo

$$\phi_p(B) = 1 - \varphi_1 B - \varphi_2 B^2 - \ldots - \varphi_p B^p \tag{3.8}$$

ϵ_t um ruído branco e B o operador de translação para o passado:

$$BX_t = X_{t-1},\ B^m X_t = X_{t-m} \tag{3.9}$$

Existe uma infinidade de soluções que resolvem a equação 3.7. Entretanto, uma solução será estacionária e canônica se e somente se todas as raízes do polinômio $\phi_p(B)$ forem, em módulo, maiores que um.

Como exemplo, segue o modelo auto-regressivo de ordem 1, AR(1):

$$X_t - \varphi X_{t-1} = \epsilon_t \tag{3.10}$$

Nesse caso, tem-se $\phi_p(B) = 1 - \varphi B$. A raiz desse polinômio é $B = \varphi^{-1}$. O modelo 3.10 será então estacionário e canônico se e somente se $|\varphi^{-1}| > 1$, ou seja, $|\varphi| < 1$.

3.5.2 Modelos MA

O modelo de médias móveis (MA) de ordem q é da forma:

$$X_t = \Theta_q(B)\epsilon_t \tag{3.11}$$

Novamente, ϵ_t é um ruído branco e B o operador de translação para o passado. Tem-se ainda ainda

$$\Theta_q(B) = 1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q \tag{3.12}$$

Se todas as raízes do polinômio $\Theta_q(B)$ forem, em módulo, maiores que um, verifica-se a condição de invertibilidade de um modelo MA(q).

Considere o modelo MA(1):

$$X_t = \epsilon_t - \theta\epsilon_{t-1} \tag{3.13}$$

A raiz do polinômio $\Theta_q(B) = 1 - \theta B$ é $B = \theta^{-1}$. Assume-se então que o modelo 3.13 está na forma invertível se e somente se $|\theta^{-1}| > 1$, ou seja, $|\theta| < 1$.

3.5.3 Modelos ARMA

Muitas vezes, os modelos AR e MA encontrados como mais adequados para um conjunto de dados tem vários parâmetros. Isso, além de demandar certo tempo de estimação, não é uma solução parcimoniosa. Uma alternativa seria então de combinar esses dois modelos de forma que haja uma redução considerável no número de parâmetros a estimar. Sendo assim, tem-se então os modelos auto-regressivos e de médias móveis (ARMA).

O modelo ARMA (p, q) é da forma:

$$\phi_p(B)X_t = \Theta_q(B)\epsilon_t \tag{3.14}$$

$\phi_p(B)$ e $\Theta_q(B)$ foram definidos nas Equações 3.8 e 3.12. As condições para estacionariedade, canonicidade e invertibilidade do modelo ARMA continuam sendo aquelas apresentadas anteriormente, ou seja, todas as raízes dos polinômios $\phi_p(B)$ e $\Theta_q(B)$ devem estar fora do disco unitário.

O modelo $X_t = 0,8X_{t-1} + \epsilon_t - 0,3\epsilon_{t-1}$ é um exemplo de processo ARMA(1,1) estacionário, canônico e invertível, pois $|\varphi| = 0,8 < 1$ e $|\theta| = 0,3 < 1$.

3.5.4 Modelos ARIMA

Os modelos AR, MA e ARMA são apropriados para descrever séries sem tendência. Entretanto, como citado na seção 3.3, muitas séries possuem tendências, sejam elas lineares ou não. Caso a série X_t apresente essa componente, utiliza-se o operador da diferença, definido na Equação 3.6. Supondo que seja necessária a utilização da

diferença de ordem d para que X_t perca sua tendência, tomar-se-á $W_t = \Delta^d X_t$, onde W_t é uma série sem tendência. Dessa forma, a aplicação dos modelos ARMA em W_t, que é uma diferença de X_t, torna-se viável. Por definição, tem-se então que X_t é uma *integral* de W_t. Por essa razão, o modelo a seguir é chamado de auto-regressivo, *integrado* e de médias móveis, ARIMA(p, d, q):

$$\phi_p(B)\Delta^d X_t = \Theta_q(B)\epsilon_t \tag{3.15}$$

Logo abaixo, tem-se um processo ARIMA(1,1,1):

$$(1 - \varphi B)\Delta X_t = (1 - \theta B)\epsilon_t \tag{3.16}$$

$$(1 - \varphi B)(1 - B)X_t = (1 - \theta B)\epsilon_t \tag{3.17}$$

$$[1 - (1 + \varphi)B + \varphi B^2]X_t = (1 - \theta B)\epsilon_t \tag{3.18}$$

$$X_t - (1 + \varphi)X_{t-1} + \varphi X_{t-2} = \epsilon_t - \theta\epsilon_{t-1} \tag{3.19}$$

Se $|\varphi| > 1$ e $|\theta| > 1$, o processo 3.19 é estacionário, canônico e invertível.

Vale notar que é possível obter os modelos AR, MA e ARMA controlando p, d e q do modelo ARIMA. Por exemplo, se $d = q = 0$, o modelo ARIMA(p, d, q) resultará em um AR(p).

3.5.5 Modelos SARIMA

Os modelos ARIMA foram desenvolvidos para séries com a componente tendência. De forma similar, o modelo ARIMA *sazonal multiplicativo* (SARIMA) lida com a componente sazonalidade. Supondo que X_t tenha uma sazonalidade de periodicidade S, o modelo SARIMA$\{(P, D, Q)_S, (p, d, q)\}$ é da forma:

$$\phi_P(B)\phi_p(B)\Delta_S^D\Delta^d X_t = \Theta_Q(B)\Theta_q(B)\epsilon_t \tag{3.20}$$

Sendo

$$\phi_P(B) = 1 - \varphi_1^* B^S - \varphi_2^* B^{2S} - \ldots - \varphi_p^* B^{pS} \tag{3.21}$$

$$\Theta_Q(B) = 1 - \theta_1^* B^S - \theta_2^* B^{2S} - \ldots - \theta_q^* B^{qS} \tag{3.22}$$

$$\Delta_S^D = \left(1 - B^S\right)^D \tag{3.23}$$

Dados observados mensalmente, como aqueles ilustrados nas Figuras 3.1b e 3.3a, tem periodicidade $S = 12$.

Novamente, não é difícil perceber que pode-se obter os modelos apresentados anteriormente a partir de um modelo SARIMA. Um processo MA(q), por exemplo, é um SARIMA$\{(0, 0, 0)_S, (0, 0, q)\}$.

3.5.6 Estimação

A segunda etapa da metodologia desenvolvida por Box e Jenkis é a estimação dos parâmetros do modelo identificado. Dois dos métodos empregados para estimar os parâmetros são: Método dos Momentos e Método de Máxima Verossimilhança. Em certos casos, a utilização de procedimentos iterativos não-lineares será necessária. Para maiores detalhes de estimação, veja Box et al. (2008), Morettin e Toloi (2006) ou Brockwell e Davis (2002).

3.5.7 Escolha do modelo

Tendo em mãos alguns modelos ARMA já estimados, deve-se então compará-los para escolher aquele que parece ser o melhor. Dois métodos de comparação entre modelos mais utilizados são o Critério de Informação de Akaike (AIC) (Akaike, 1974) e o Critério de Informação Bayesiano (BIC) (Akaike, 1977), (Rissanen, 1978) e (Schwartz, 1978). Sugere-se escolher o modelo ARIMA(p, d, q) cujas ordens p e q minimizam o valor de AIC ou de BIC.

$$\text{AIC}(p,d,q) = N\,\ln(\hat{\sigma}_\epsilon^2) + \frac{N}{N-d}\,2(p+q+1+\delta_{d0}) + N\,\ln(2\pi) + N \tag{3.24}$$

$$\text{BIC} = \ln(\hat{\sigma}_\epsilon^2) + (p+q)\frac{\ln N}{N} \tag{3.25}$$

Sendo

$$\delta_{d0} = \begin{cases} 1 & \text{se } d = 0 \\ 0 & \text{se } d \neq 0 \end{cases} \tag{3.26}$$

Um terceiro método de comparação é citado por Pindyck (2004). Ele trata do coeficiente de desigualdade de Theil. Definindo y_t^s, y_t^a e T como valor estimado, valor efetivo e o número de períodos respectivamente, tem-se U:

$$U = \frac{\sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(y_t^s - y_t^a\right)^2}}{\sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(y_t^s\right)^2} + \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(y_t^a\right)^2}} \tag{3.27}$$

Sendo que U pode assumir valores entre 0 e 1. É fácil verificar que quanto mais próximo de 0, melhor é o modelo estimado, pois neste caso, $y_t^s = y_t^a, \forall t$.

3.5.8 Diagnóstico do modelo

Uma das formas de diagnóstico é analisar os resíduos do modelo. Suponha que o modelo escolhido seja um ARIMA(p, d, q) estacionário, canônico e invertível apre-

sentado na equação 3.15. Se este for um modelo que satisfaça as condições de Box e Jenkins, pode-se escrever $\epsilon_t = \Theta_q^{-1}(B)\phi_p(B)\Delta^d X_t$ de forma que ϵ_t seja um ruído branco. Ou seja, ao analisar os gráficos de autocorrelações dos resíduos, não deve-se achar correlações significantemente diferentes de zero.

Capítulo 4

Material e Métodos

4.1 Introdução

Como foi dito no capítulo 1, a utilização direta do banco de dados fornecido pelo DETRAN-DF não é recomendada devido a problemas comuns como sub-registros na falta de energia. Com a finalidade de obter dados mais confiáveis, serão imputados valores no banco. Para identificação dos valores que devem ser imputados, a técnica de reamostragem de Jackknife será utilizada. Além dessa imputação melhorar a estimação da série temporal, esse estudo servirá como auxílio na escolha de interseções onde a coleta de dados será mais útil, pois ter-se-á conhecimento de quais equipamentos funcionaram melhor, ou seja, aqueles que estavam ligados a maior parte do tempo e não computaram muitos sub-registros.

4.2 Estrutura

O banco consta de dados disponíveis por hora, por dia, por mês e por ano, para 29 equipamentos diferentes de contagem volumétrica veicular.

Deficiências observadas:

- Inexistência de dados registrados para determinados horários de um dia;
- Inexistência de dados registrados para determinados dias da semana (todos os horários sem dados registrados);
- Valores observados em determinados horários bastante discrepantes dos registrados para o mesmo dia da semana de um mesmo mês/ano.

Devido a esses problemas citados, um tratamento de dados será feito. Na seção a seguir será explicado como manipular o banco de dados para a futura análise.

4.3 Tratamento dos dados de cada equipamento

Esse tratamento é aplicável para cada mês/ano. No decorrer desse texto, será usado o termo **conjunto** para se referir a cada um dos 7 conjuntos formados pelos diferentes dias da semana (domingo à sábado) e **dia** para se referir a cada elemento de um determinado conjunto. As etapas do tratamento serão as seguintes:

4.3.1 Agrupamento dos dados de volumes horários de cada dia em função do dia da semana

Serão formados 7 conjuntos de dias, referentes aos seguintes tipos de dias da semana: 1- Domingo e Feriado; 2- Segunda; 3- Terça; 4- Quarta; 5- Quinta; 6- Sexta; 7- Sábado. O número de elementos de cada conjunto dependerá do mês e ano pertinentes.

4.3.2 Construção da base de dados imputada

Aplicar o método de imputação de dados (Seção 4.4) para:

- atribuir valores de volumes horários para os casos em que estes volumes forem faltantes (*missings*);

- atribuir valores de volumes horários para o caso em que estes volumes estiverem fora do que seria esperado para o horário, levando em conta o respectivo dia da semana (o conjunto a que pertencem).

Após a imputação, haverá dois bancos de dados que serão comparados mais adiante para um estudo de consistência: Banco Original (antes da imputação) e Banco Imputado.

4.3.3 Métodos de limpeza dos bancos

Tendo em pose os dois bancos (Original e Imputado), serão adotados dois métodos de limpeza de dados a fim de calcular fatores de expansão diferentes. Dessa forma, será possível comparar fatores obtidos a partir de quatro bases de dados distintas: aplicando o Método 1 no Banco Original e em seguida no Banco Imputado, assim como o Método 2 em ambos os bancos. Tais métodos de limpeza são descritos a seguir:

Métodos de limpeza 1

(a) para cada equipamento, ano e mês, os conjuntos que não tiverem pelo menos um dia com contagem volumétrica em todos os horários serão retirados da base de dados;

(b) para os dias de cada conjunto com informações para todos os horários serão calculados os volumes diários;

(c) os volumes diários calculados em (b) serão usados para calcular o volume médio diário do conjunto correspondente;

(d) o volume médio diário calculado em (c) será adotado como o valor do volume diário representativo do conjunto;

(e) o volume mensal será calculado pela soma dos volumes diários representativos de cada conjunto, multiplicados pelo número de dias do conjunto a que se referem;

(f) considerando os volumes horários dos dias usados para calcular o volume médio diário de cada conjunto (ver letra *b*), calcular os volumes médios horários correspondentes.

Métodos de limpeza 2

(a) para cada equipamento, ano e mês, os conjuntos que não tiverem todos os dias com contagem volumétrica em todos os horários serão retirados da base de dados;

(b) para os dias de cada conjunto com informações para todos os horários serão calculados os volumes diários;

(c) os volumes diários calculados em (b) serão usados para calcular o volume médio diário do conjunto correspondente;

(d) o volume médio diário calculado em (c) será adotado como o valor do volume diário representativo do conjunto;

(e) o volume mensal será calculado pela soma dos volumes diários representativos de cada conjunto, multiplicados pelo número de dias do conjunto a que se referem;

(f) considerando os volumes horários dos dias usados para calcular o volume médio diário de cada conjunto (ver letra *b*), calcular os volumes médios horários correspondentes.

A diferença entre os dois métodos de limpeza reside no item (a). Enquanto no Método de Limpeza 1 basta que um conjunto tenha um dia com informações completas para que todos os horários sejam mantido na base de dados, no Método de Limpeza 2, a permanência do conjunto requer que todos os dias nele contidos estejam com os volumes horários completos.

Nos dois métodos de limpeza, no caso de um ou mais conjuntos serem eliminados, o volume mensal calculado em (e) será um volume subestimado para o mês. Os meses que estiverem nessa situação não serão considerados para efeito da determinação dos fatores de expansão diário (referente a cada conjunto) e mensal (referente ao mês).

A aplicação dos dois métodos de limpeza produzirá quatro bases de dados, a partir das quais serão calculados os fatores de expansão horário, diário e mensal. São elas: (i) Banco Original - Método 1; (ii) Banco Imputado - Método 1; (iii) Banco Original - Método 2 e (iv) Banco Imputado - Método 2.

Esperava-se que a base obtida pelo método de limpeza 2 aplicado no banco imputado (Banco Imputado - Método 2) resultasse na base mais consistente porque somente foram incluídos dias onde todos os volumes horários foram revisados e ajustados pelo método de imputação, isto é, tanto os valores nulos como os sub-registros foram eliminados.

No entanto, o uso desta base reduziu bastante o número de fatores de expansão diário e mensal, dificultando assim a utilização destes para a previsão dos fatores nos anos seguintes. Por tal razão, a base utilizada na determinação das previsões foi aquela obtida pelo Método de limpeza 1 aplicado no banco imputado (Banco Imputado - Método 1).

A seguir, o método usado para atribuição de valores faltantes ou sub-registros é especificado.

4.4 Método de imputação

O procedimento de imputação será aplicado com os seguintes objetivos:

- estimar valores de fluxos horários para o caso da não ocorrência de registros em determinados horários (*missing values* na base original);
- estimar valores de fluxos horários para substituir valores registrados que forem detectados pela análise estatística como abaixo do esperado para o horário (valores discrepantes na base original).

A imputação é sempre feita para volumes horários de um determinado dia da semana de um mês específico.

No máximo poderão ser feitas três imputações para um mesmo horário/dia/mês/ano/equipamento. Se for necessário mais do que três imputações, então não se faz nenhum tratamento nos dados e os volumes horários existentes não serão alterados. No caso de algum desses dados representarem sub-registros e pertencerem a um dia com volumes presentes para todos os horários, ter-se-á volumes diários calculados sem a devida qualidade.

4.4.1 Identificação dos valores a serem imputados pelo método de Jackknife

Ao fixar um equipamento e certo dia da semana de um mês em algum ano, tem-se então n observações para cada horário desse dia. Por exemplo, o mês de agosto de 2011 teve cinco segundas-feiras. Nesse caso $n = 5$, pois são cinco observações em cada horário da segunda-feira desse mês naquele ano. Definindo x_i como o *i-ésimo* valor observado para o horário em estudo daquele dia/mês/ano/equipamento (com $i = 1, ..., n$), pode-se obter a média e o desvio-padrão desse horário:

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{4.1}$$

$$dp = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n-1}} \tag{4.2}$$

***Observação**: valores faltantes para um determinado dia referentes ao horário estudado não são considerados. Ou seja, n é o número de observações disponíveis para o horário considerado.*

Em seguida, aplica-se a técnica de reamostragem de Jackknife (Rupert G. Miller, 1964), (Efron e Tibshirani, 1994), (Shao e Tu, 1995) e (Cochran, 1977) com os seguintes passos:

1. Gera-se a amostra Jackknife 1 e calcula-se sua respectiva média

$$X_{(1)} = \{x_2, x_3, \ldots, x_{n-1}, x_n\}$$
$$\bar{X}_{(1)} = \frac{\sum_{i=2}^{n} x_i}{n-1} \tag{4.3}$$

2. Gera-se a amostra Jackknife 2 e calcula-se sua respectiva média

$$\begin{aligned} X_{(2)} &= \{x_1, x_3, \ldots, x_{n-1}, x_n\} \\ \bar{X}_{(2)} &= \frac{x_1 + \sum_{i=3}^{n} x_i}{n-1} \end{aligned} \tag{4.4}$$

3. Gera-se a amostra Jackknife p e calcula-se sua respectiva média

$$\begin{aligned} X_{(p)} &= \{x_1, x_2, \ldots, x_{p-1}, x_{p+1}, \ldots, x_{n-1}, x_n\} \\ \bar{X}_{(p)} &= \frac{\sum_{i=1}^{p-1} x_i + \sum_{i=p+1}^{n} x_i}{n-1} \end{aligned} \tag{4.5}$$

4. Gera-se a amostra Jackknife n e calcula-se sua respectiva média

$$\begin{aligned} X_{(n)} &= \{x_1, x_2, \ldots, x_{n-1}\} \\ \bar{X}_{(n)} &= \frac{\sum_{i=1}^{n-1} x_i}{n-1} \end{aligned} \tag{4.6}$$

Para cada amostra $X_{(i)}$ de Jackknife, a seguinte função indicadora é aplicada:

$$I_1(X_{(i)}) = \begin{cases} 1 & \text{se} \quad \bar{X}_{(i)} > \bar{X} \\ 0 & \text{se} \quad \bar{X}_{(i)} \leq \bar{X} \end{cases} \tag{4.7}$$

Dessa forma, será possível separar os conjuntos de dados que contem as observações com valores menores daquelas com maiores valores.

Seja C o conjunto formado pela união das amostras de Jackknife onde $I_1(X_{(i)}) = 1$. Essa união deverá ser feita de forma que valores presentes em duas ou mais amostras deverão constar em C quantas vezes aparecerem em amostras diferentes.

Por exemplo, para $n = 4$, correspondente a um conjunto sem nenhum valor faltante, suponha que: $I_1(X_{(1)}) = 1, I_1(X_{(2)}) = 0, I_1(X_{(3)}) = 0$ e $I_1(X_{(4)}) = 1$. Sabe-se que:

$$X_{(1)} = \{x_2, x_3, x_4\}$$

$$X_{(4)} = \{x_1, x_2, x_3\}$$

Logo, $C = \{x_1, x_2, x_2, x_3, x_3, x_4\}$

Seja n_i a frequência de x_i presente no conjunto C. Para x_i não pertencente ao conjunto C, atribui-se $n_i = 10$.

Para o exemplo anterior: $n_1 = 1, n_2 = 2, n_3 = 2$ e $n_4 = 1$.

Define-se então uma segunda variável indicadora:

$$I_2(n_i) = \begin{cases} 1 & \text{se} \quad n_i > min(n_i) \quad \text{e} \quad n_i \neq 10 \\ 0 & \text{se} \quad n_i = min(n_i) \quad \text{ou} \quad n_i = 10 \end{cases} \tag{4.8}$$

No exemplo anterior, $I_2(n_1) = 0, I_2(n_2) = 1, I_2(n_3) = 1$ e $I_2(n_4) = 0$, pois $min(n_i) = 1$.

Sendo assim, é possível escrever:

$$\bar{X}^* = \frac{\sum_{i=1}^{n} [x_i \times I_2(n_i)]}{\sum_{i=1}^{n} I_2(n_i)} \tag{4.9}$$

$$dp^* = \sqrt{\frac{\sum_{i=1}^{n} \left[\left(x_i - \bar{X} \right)^2 \times I_2(n_i) \right]}{\left[\sum_{i=1}^{n} I_2(n_i) \right] - 1}} \tag{4.10}$$

Se $min(n_i) = max(n_i)$, então $\bar{X}^*$ e dp^* não são calculáveis, porque $I_2(n_i) = 0$ para todo n_i.

A intenção de gerar $\bar{X}^*$ e dp^* é obter uma média e um desvio-padrão do horário em estudo onde os valores discrepantes (sub-registros) não influenciem.

Seja LI um limite inferior, tal que:

$$LI = \begin{cases} \bar{X}^* - 2dp^* & \text{se } \bar{X}^* \text{ e } dp^* \text{ existem} \\ \bar{X} - dp & \text{caso contrário} \end{cases} \tag{4.11}$$

Tal limite é calculado com a finalidade de auxiliar na identificação dos valores que devem ser imputados. Se $x_i < LI$ ou se x_i for uma informação faltante (*missing*), então deve-se imputar um valor em seu lugar. Caso contrário, mantem-se x_i.

Lembre-se que a imputação é realizada para no máximo três valores de um certo horário/dia/mês/ano/equipamento. Isso se deve ao fato de que, na maioria das vezes, há apenas quatro ou cinco observações para um horário.

4.4.2 Imputação de valores

Uma vez identificados os valores a serem imputados, se faz necessário ordená-los de forma crescente (caso haja valores faltantes, esses serão os primeiros da ordem). Cria-se então outra variável indicadora:

$$I_3(x_i) = \begin{cases} 1 & \text{se } x_i \geq LI \\ 0 & \text{se } x_i < LI \text{ ou } x_i \text{ é um valor faltante} \end{cases} \tag{4.12}$$

Ou seja, serão imputados valores somente para substituir os x_i que apresentarem $I_3(x_i) = 0$.

No primeiro valor a ser imputado (o menor ou um *missing*), imputa-se $\bar{X}$ calculado em 4.1, que é a média de todos os dados observados para aquele horário. No segundo, imputa-se a média dos valores que são iguais ou estão acima do limite inferior:

$$\hat{X}^{**} = \frac{\sum_{i=1}^{n} [x_i \times I_3(x_i)]}{\sum_{i=1}^{n} I_3(x_i)} \tag{4.13}$$

O terceiro e último valor será substituído por uma média geral novamente:

$$\bar{X}^{***} = \sum_{i=1}^{n} \frac{(x_i)_{atualizado}}{n} \tag{4.14}$$

Como os dois menores valores já foram atualizados anteriormente, $\bar{X}$ será diferente de $\bar{X}^{***}$.

4.5 Fatores de expansão

A coleta de dados foi feita em algumas interseções durante uma ou três horas de um certo mês e ano. Uma vez que o objetivo é expandir o fluxo de carros coletado nesse curto período para um ano, cria-se fatores de expansão. Serão obtidos três desses fatores, detalhados a seguir.

4.5.1 Fator de expansão horária para um determinado dia/mês/ano/equipamento

Permite a estimativa do volume médio diário para um tipo de dia da semana d =[domingo/feriado, segunda, terça, ... , sábado] a partir de contagens realizadas no período de uma hora h =[0h-1h, 1h-2h, ... , 23h-24h].

$$F_{h(d,mês,ano)} = \frac{\text{Volume médio diário do dia da semana } d}{\text{Volume médio da hora } h \text{ no dia } d} \tag{4.15}$$

Para um determinado mês/ano, ter-se-ão 168 fatores de expansão horária (24×7). Para um ano, 2016 fatores horários ($24 \times 7 \times 12$).

Analogamente, pode-se calcular tal fator de expansão a partir de contagens em intervalos de tempos maiores. Um exemplo seria agrupar três horas, de forma a obter

h =[0h-3h, 1h-4h, ... , 21h-24h]. A vantagem de utilizar intervalos de hora maiores é que a observação é uma fração mais impactante do dia. Por exemplo, espera-se que o fator de expansão criado a partir de 8 horas observadas em um dia (um terço do dia) se aproxime mais da realidade do que aquele calculado com apenas 1 hora observada no dia inteiro.

4.5.2 Fator de expansão diário para um determinado mês/ano/equipamento

Permite a estimativa do volume mensal para um mês Y =[Jan, Fev, ... , Dez] a partir dos valores obtidos para o dia da semana d.

$$F_{d(mês,ano)} = \frac{\text{Volume total do mês } Y}{\text{Volume médio diário do dia da semana } d} \tag{4.16}$$

Para um determinado mês/ano, ter-se-ão 7 fatores de expansão diária. Para um ano, 84 fatores diários (7×12)

4.5.3 Fator de expansão mensal para um determinado ano/equipamento

Permite a estimativa do volume anual para um ano Z =[2005, 2006, ... ,2010] a partir dos valores obtidos para o mês Y.

$$F_{mês(ano)} = \frac{\text{Volume total do ano } Z}{\text{Volume total do mês } Y} \tag{4.17}$$

Os três fatores de expansão serão calculados para cada uma das quatro bases de dados (Banco Original - Método 1, Banco Original - Método 2, Banco Imputado - Método 1 e Banco Imputado - Método 2).

Será possível então realizar a comparação entre as quatro estimativas de volumes mensais e anuais obtidas pela aplicação dos diferentes fatores de expansão.

4.6 Comparação entre métodos

Uma vez que os dados são assimétricos (não respeitando o critério da normalidade), foram feitos testes de Wilcoxon para os quatro bancos resultantes das limpezas a fim de verificar se os métodos de limpeza e imputação implicam em alguma mudança no cálculo dos fatores. É de se esperar que tais diferenças sejam nítidas para conjunto de dias da semana que originalmente estavam com vários valores faltantes

ou sub-registros. Como um auxílio ao teste formal não-paramétrico, analisa-se também gráficos com o objetivo de averiguar visualmente os casos onde os métodos se distinguem.

4.7 Previsão dos fatores de expansão

Essa seção é apresentada com o intuito de introduzir a ideia utilizada no desenvolvimento da programação com aproximadamente três mil linhas. Para a estimação dos modelos de séries temporais, bem como suas previsões, foram utilizadas rotinas macros no *software* SAS 9.2.

Para estimar um modelo ARIMA(p, d, q) da variável x, basta utilizar as linhas de comando apresentadas na Figura 4.1.

```
proc arima data=banco;
      identify var=x(d);run;
      ods output optsummary=msg;
      estimate plot p=p q=q outstat=estatisticas
            outmodel=parametros;run;
quit;
```

Figura 4.1: Estimação de um modelo ARIMA(p, d, q) no SAS 9.2

Esse comando estimará θ_i $(i = 1, \ldots, p)$ e φ_j $(j = 1, \ldots, q)$ do modelo $\phi_p(B)\Delta^d X_t = \Theta_q(B)\epsilon_t$ especificado na seção 3.5.4 e os salvará em uma tabela intitulada "parametros". As estatísticas de interesse para comparação entre modelos como o AIC serão gravadas em um banco de dados chamado "estatisticas". Caso o modelo estimado por esse comando forneça parâmetros que não satisfaçam as condições de estacionariedade, canonicidade e invertibilidade citadas na seção 3.5.3, um banco de dados chamado "msg" será criado trazendo uma mensagem de erro. Sendo assim, durante a rotina é fácil excluir tais modelos insatisfatórios e guardar apenas os que satisfazem essa condição para futura comparação e escolha do melhor modelo.

Um modelo ARIMA com restrições de parâmetros também é utilizado nas comparações entre modelos. Um modelo restrito é aquele onde apenas as *lags* de interesse são estimadas enquanto as outras são fixadas com o valor zero. No modelo ARIMA$(p, d, q)_r$ (o índice r marca a restrição), fixam-se $\varphi_i = 0$ para $i < p$ e $\theta_j = 0$ para $j < q$, resultando então no modelo $\left(1 - \varphi_p B^p\right) \Delta^d X_t = \left(1 - \theta_q B^q\right) \epsilon_t$. Tal modelo pode ser estimado no SAS 9.2 com o comando apresentado na Figura 4.2.

Analogamente, a Figura 4.3 traz como estimar um modelo SARIMA$\{(P, 1, Q)_S, (p, d, q)\}_r$.

```
proc arima data=banco;
      identify var=x(d);run;
      ods output optsummary=msg;
      estimate plot p=(p) q=(q) outstat=estatisticas
            outmodel=parametros;run;
quit;
```

Figura 4.2: Estimação de um modelo ARIMA$(p, d, q)_r$ no SAS 9.2

```
proc arima data=banco;
      identify var=x(d,S);run;
      ods output optsummary=msg;
      estimate plot p=(p)(P) q=(q)(Q) outstat=estatisticas
      outmodel=parametros;run;
quit;
```

Figura 4.3: Estimação de um modelo SARIMA$\{(P, 1, Q)_S, (p, d, q)\}_r$ no SAS 9.2

A programação desenvolvida nesse trabalho utiliza-se basicamente desses comandos para estimação e comparação entre modelos. De forma iterativa, a macro estima todos os modelos possíveis com p e q variando de 0 à 12, enquanto d toma os valores 0, 1, 6 e 12 (as únicas diferenças interpretáveis para as variáveis em estudo). Essa iteração foi feita tanto com comandos baseados na Figura 4.1 quanto na 4.2. Vale frisar que, como a primeira estima mais de um parâmetro nas médias móveis e nos auto-regressivos, testes foram realizados para saber quais deles são significativos ao nível 5% ($\alpha = 0,05$). Após esses testes, o modelo é reestimado apenas com os parâmetros significativos.

Capítulo 5

Análise dos Resultados

5.1 Introdução

O banco contem 29 equipamentos eletrônicos. Observa-se dados de janeiro de 2005 à junho de 2011. É importante salientar que nem todos os registradores tem observações a partir de 2005, pois alguns foram instalados após essa data. Executado o procedimento de imputação, verificou-se que aproximadamente 30,5% dos dados foram substituídos, evidenciando a inconsistência do banco de dados. Na seção a seguir são apresentados gráficos para visualização do resultado do método de imputação.

5.2 Imputação de dados

As Figuras 5.1 e 5.2 ilustram a funcionalidade do método de imputação de dados. Os pontos interligados por uma linha tracejada e vermelha são os originais. Já os que estão ligados por uma linha contínua e azul, representam o conjunto de dados após a execução do método de imputação. Para o equipamento ASV012, nas proximidades do primeiro dia de setembro de 2008 (Figura 5.1), nota-se facilmente como o processo conseguiu capturar e reproduzir a variabilidade do processo ao redor desses dados que, à princípio, eram *missings*. Ao longo do tempo, percebe-se também a eficácia do método ao tratar sub-registros. Entretanto, como comentado no Capítulo 4, o processo não realiza mais que três imputações para um conjunto de dias em um certo mês/ano/equipamento. Tal fato pode ser notado nas proximidades do mês de novembro de 2008 para o equipamento ASV012 (Figura 5.1) ou próximo ao mês de fevereiro de 2009 do ASV063 (Figura 5.2).

Com o método de imputação realizado, gerou-se uma lista com os equipamentos

eletrônicos que eram menos inconsistentes durante o período de funcionamento. Para gerar tal lista, foi levado em consideração a quantidade de *missings* e sub-registros antes da imputação. Por questões operacionais, o DETRAN-DF escolheu os locais próximos de onde estão instalados os equipamentos ASV013, ASV014, ASV032, ASV033, ASV131 e ASV132 para uma coleta de dados em 2011 e 2012. Essa amostra foi utilizada no cálculo da previsão do volume anual de veículos (Seção 5.5) nesses locais da coleta.

5.3 Fatores de expansão

Após realizada a imputação de dados e a limpeza descrita na Seção 4.3.3, foram calculados os fatores de expansão (seguindo as instruções da Seção 4.5). Geraram-se então gráficos para comparação visual entre os métodos de limpeza.

As Figuras 5.3 e 5.4 apresentam essa comparação para o fator de expansão dia-mês aplicado no equipamento ASV063, ano de 2008 e para os meses de julho e dezembro, respectivamente. Na primeira figura, percebe-se que não há uma diferença tão notável entre os métodos quanto na figura seguinte. Isso evidencia o fato de que, para esse registrador e nesse ano, os dados coletados em julho foram mais consistentes que em dezembro.

Ainda com relação às duas figuras, perceba que o fator de expansão para os domingos e feriados é muito maior que os outros. Tal resultado era esperado, pois o contingente de carros nesses dias é bem menor. De forma que, caso queira-se expandir o volume observado em um domingo para o mês inteiro, seria necessário um fator mais alto do que o de uma terça-feira, por exemplo, que é um dia útil.

A diferença entre os métodos pode também ser observada na Figura 5.5, que ilustra o fator de expansão mês-ano do registrador ASV063 no ano de 2008. Entretanto, não se notam grandes diferenças no fator de expansão hora-dia, principalmente após as 5 horas, da Figura 5.6 para o mesmo pardal, mesmo ano, no mês de dezembro para as sextas-feiras.

Com o objetivo de apresentar quão próximas as estimativas por fatores de expansão estão do volume real de carros em um ano, foi realizado um simples exemplo. Para cada um dos seis equipamentos escolhidos pelo DETRAN-DF citados na seção 5.2, selecionou-se aleatoriamente um dia e, em seguida, um horário. Os fatores de expansão referentes àquele local, dia e horário foram aplicados à observação a fim de estimar o volume do ano inteiro. Segue na Tabela 5.1 os valores encontrados.

Percebe-se então que as estimativas estão relativamente próximas do real.

Tabela 5.1: Cálculo da estimativa do volume de veículos no ano

Equipamento	Dia Selecionado	Hora Selecionada	Volume observado	FE Hora-Dia	FE Dia-Mês	FE Mês-Ano	Volume ano (real)	Volume estimado do ano
ASV013	12/10/2009	15 às 16	126,56	22,87	53,02	10,93	1677397	1388994,88
ASV014	05/01/2009	13 às 14	389,25	14,45	29,39	11,70	1933942	1967499,46
ASV032	12/09/2008	15 às 16	1291,25	15,76	27,54	12,59	7056610	7361279,09
ASV033	09/05/2010	19 às 20	1091,08	16,64	41,60	11,34	8567729	9092237,33
ASV131	04/09/2009	14 às 15	415,00	20,06	28,21	12,20	2866050	2813219,47
ASV132	07/06/2010	9 às 10	426,50	20,63	29,91	12,01	3161152	3187033,65

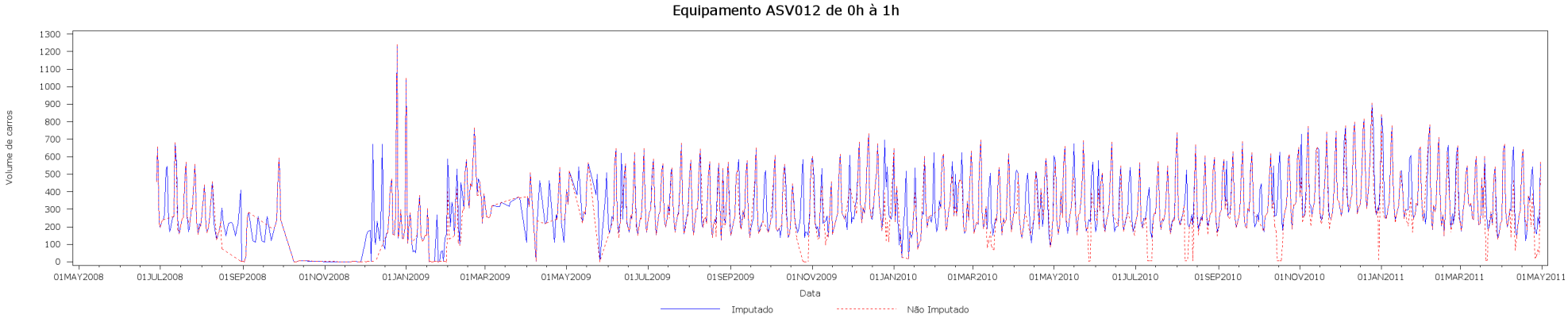


Figura 5.1: Comparação Banco Original e Banco Imputado (Pardal ASV012)

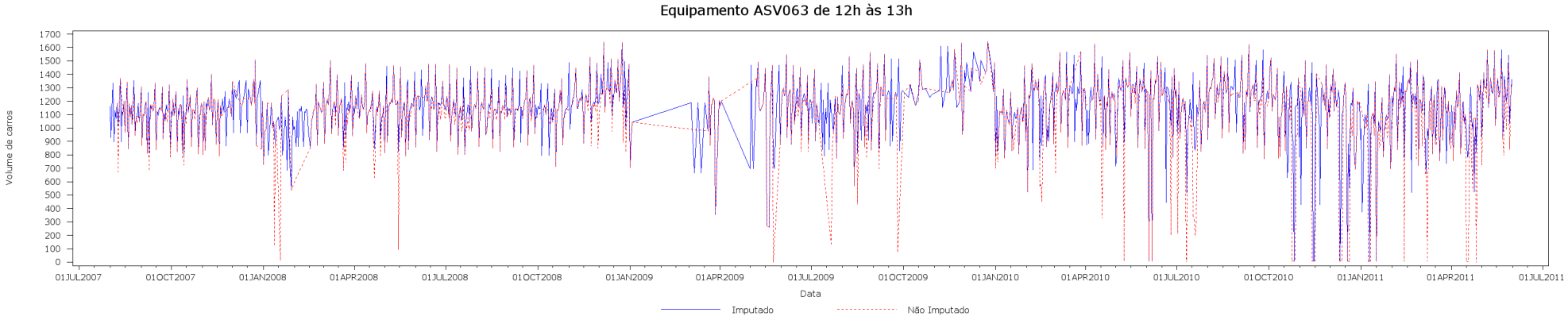


Figura 5.2: Comparação Banco Original e Banco Imputado (Pardal ASV063)

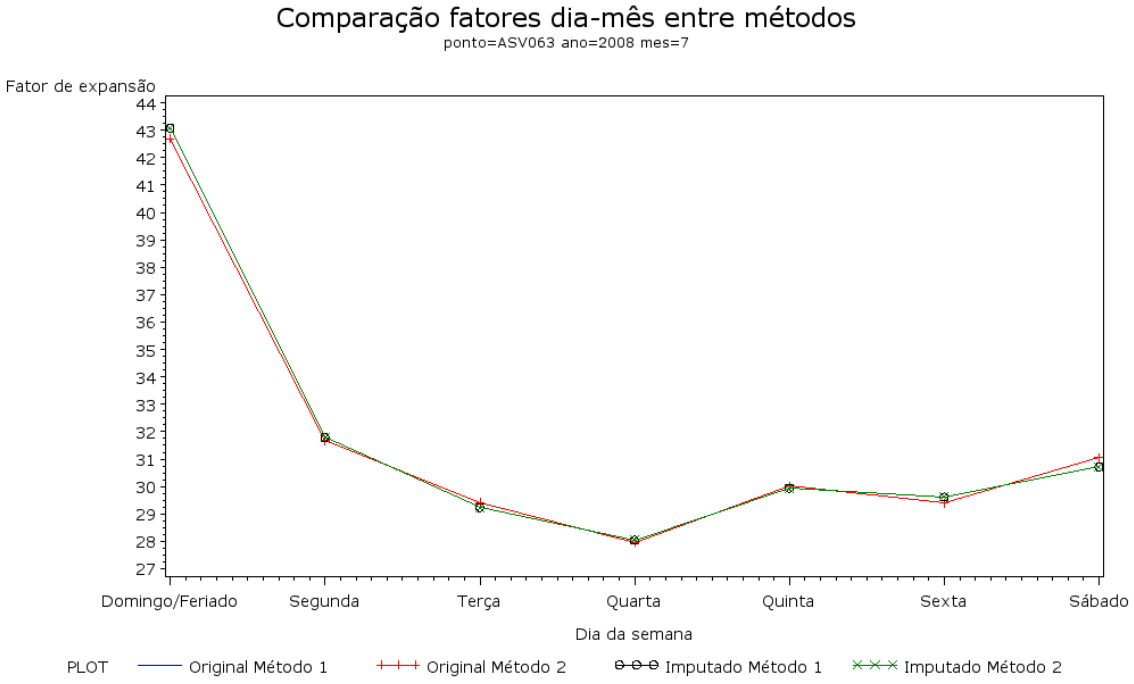


Figura 5.3: Comparação entre métodos de limpeza

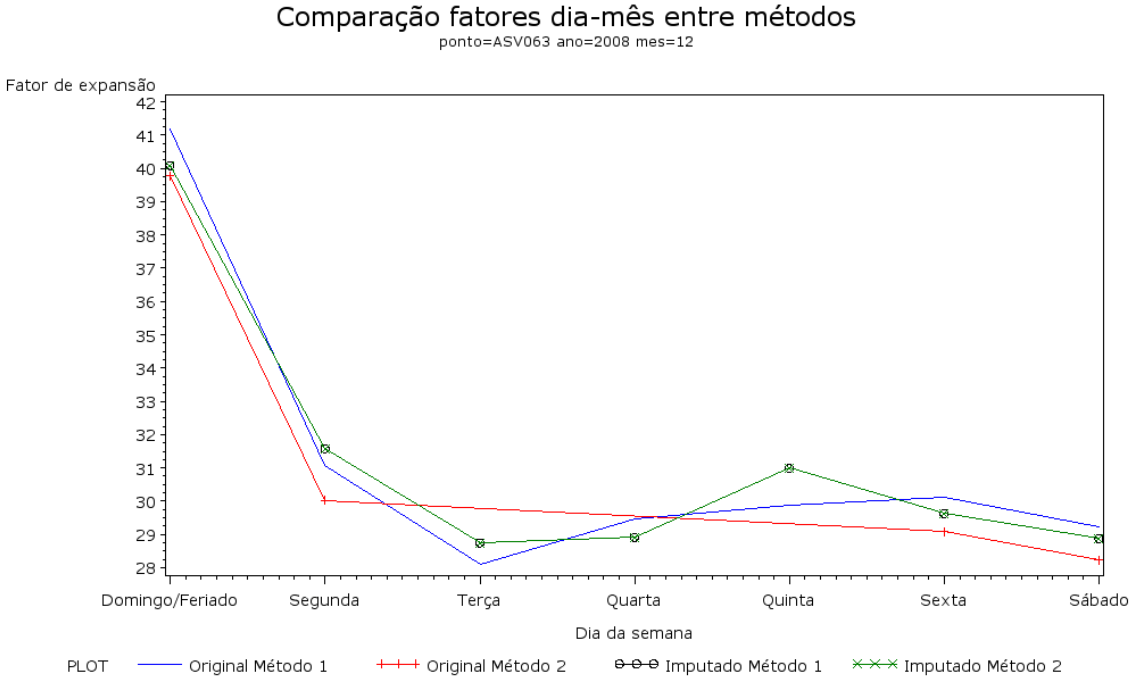


Figura 5.4: Comparação entre métodos de limpeza

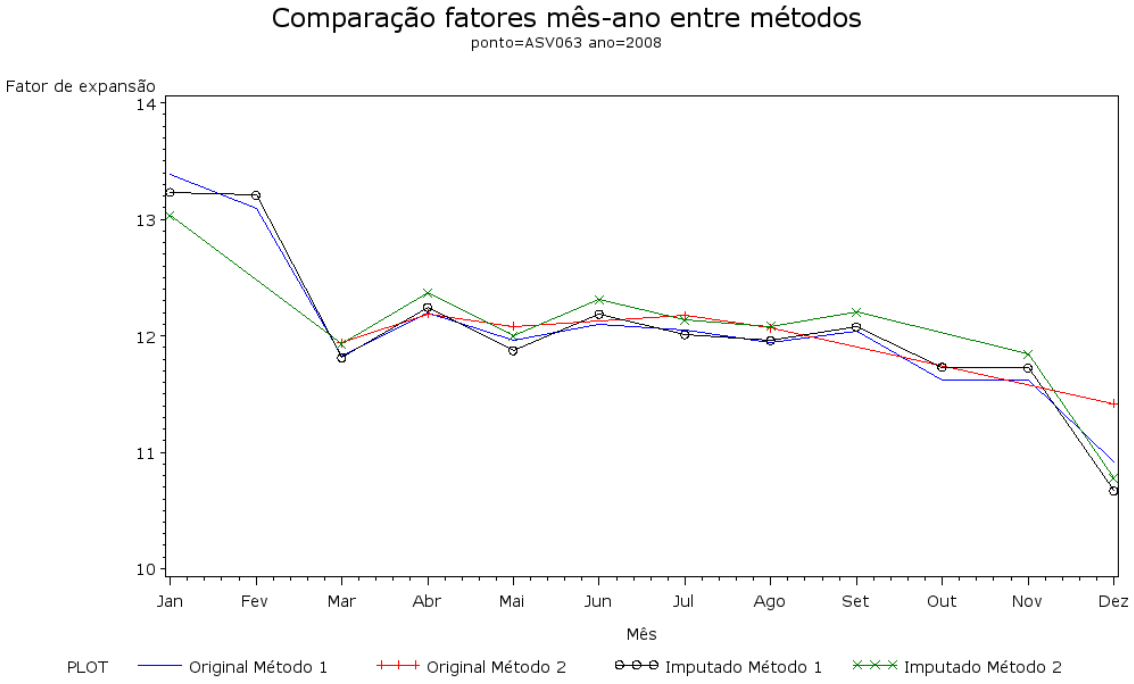


Figura 5.5: Comparação entre métodos de limpeza

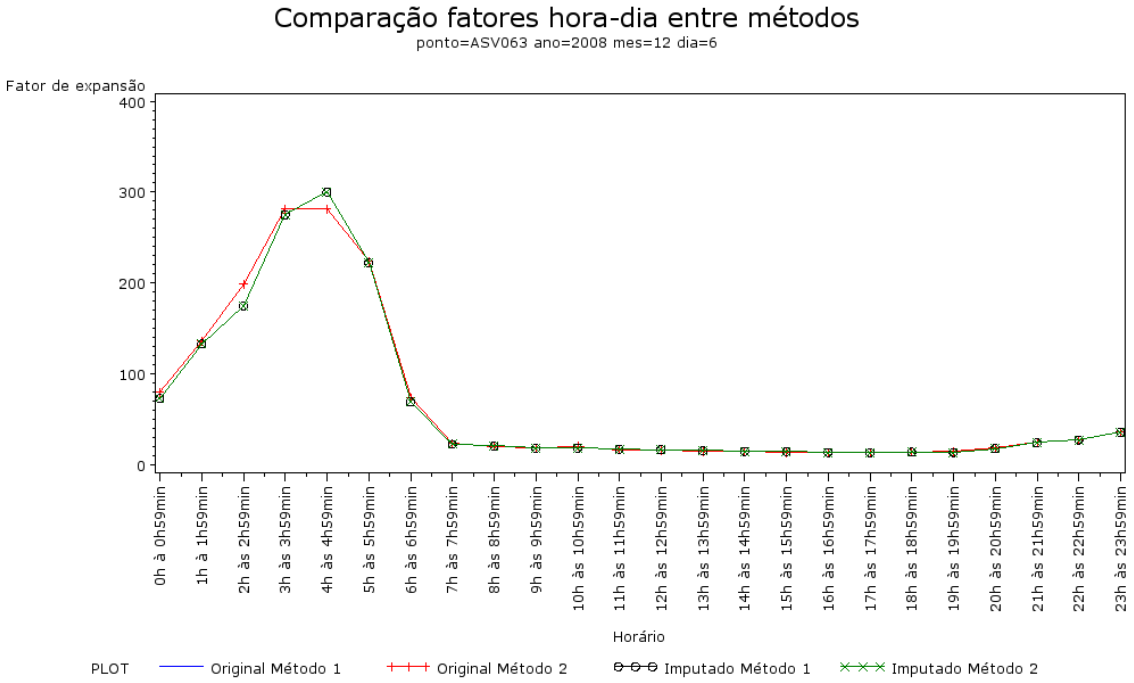


Figura 5.6: Comparação entre métodos de limpeza

5.4 Previsão dos fatores de expansão

Todo o trabalho de previsão dos fatores de expansão foi desenvolvido de acordo com a necessidade de uma dissertação de mestrado realizado em paralelo (Claude,

2012). Por isso, foram estimados 97 modelos para o fator de expansão hora/dia, 70 para o fator dia/mês e 52 para o fator mês/ano. Visto a quantidade de modelos a estimar, percebe-se a importância de ter preparado a macro citada na seção 4.7.

Essa presente seção traz os resultados de alguns modelos estimados pela programação desenvolvida. Primeiramente apresentar-se-á um modelo que se ajustou bem aos dados. Em seguida, será exposto um modelo de ajustamento razoável e um ruim. Nesse trabalho, o modelo é considerado bem ajustado se, além de atender aos pressupostos de Box et al. (2008), as previsões estiverem próximas das observações. Na Seção 5.4.4 é apresentado um *Box-Plot* para se ter uma visão resumida da qualidade (segundo o coeficiente de Theil) de todos os 219 modelos estimados.

5.4.1 Exemplo de um modelo bem ajustado

O modelo apresentado aqui tem como objetivo prever os fatores que expandem o volume de uma sexta-feira para o mês do equipamento ASV090.

A Figura 5.7 mostra que, mesmo realizado o processo de limpeza e de imputação de dados, alguns valores extremos permaneceram no banco. Infelizmente, devido ao pico no mês de fevereiro de 2008, nenhum modelo SARIMA conseguiu uma boa adaptação. Por isso, em casos de picos persistentes na base de dados, adotou-se a seguinte decisão: os valores extremos são substituídos pela média entre o valor anterior e posterior. Ou seja, se X_t for um valor aberrante, ele será substituído por $(X_{t-1} + X_{t+1})/2$. Ressalta-se que o critério de escolha de valores aberrantes foi feita de forma subjetiva já que as metodologias de imputação e limpeza apresentadas anteriormente não foram capazes de detectá-los. Entretanto, as únicas mudanças desses valores foram feitas ao longo do trabalho se nenhum modelo SARIMA se adequasse bem. No caso de haver pelo menos um modelo (mesmo sendo ruim) que se ajustasse às observações, os valores aberrantes não seriam alterados.

Após a substituição do valor aberrante, pode-se observar a série em estudo na Figura 5.8 e o seu respectivo correlograma e correlograma parcial na Figura 5.9. Para tais observações, o modelo escolhido foi um SARIMA$\{(1,0,0)_{12},(0,0,0)\}$. As Figuras 5.10 e 5.11 mostram os correlogramas, o histograma e o *QQ-Plot* dos erros do modelo estimado. Conclui-se então que ele satisfaz os pressupostos de Box et al. (2008). Vale lembrar que a macro desenvolvida descarta os modelos cujas raízes dos polinômios pertinentes aos processos AR e MA são menores do que um em módulo. Logo, o modelo escolhido também atende ao critério de ser estacionário, canônico e

invertível. A Figura 5.12 traz as previsões obtidas através do modelo. Nota-se então que as previsões estão bem próximas das observações, o que nos leva a caracterizar o modelo como bem ajustado.

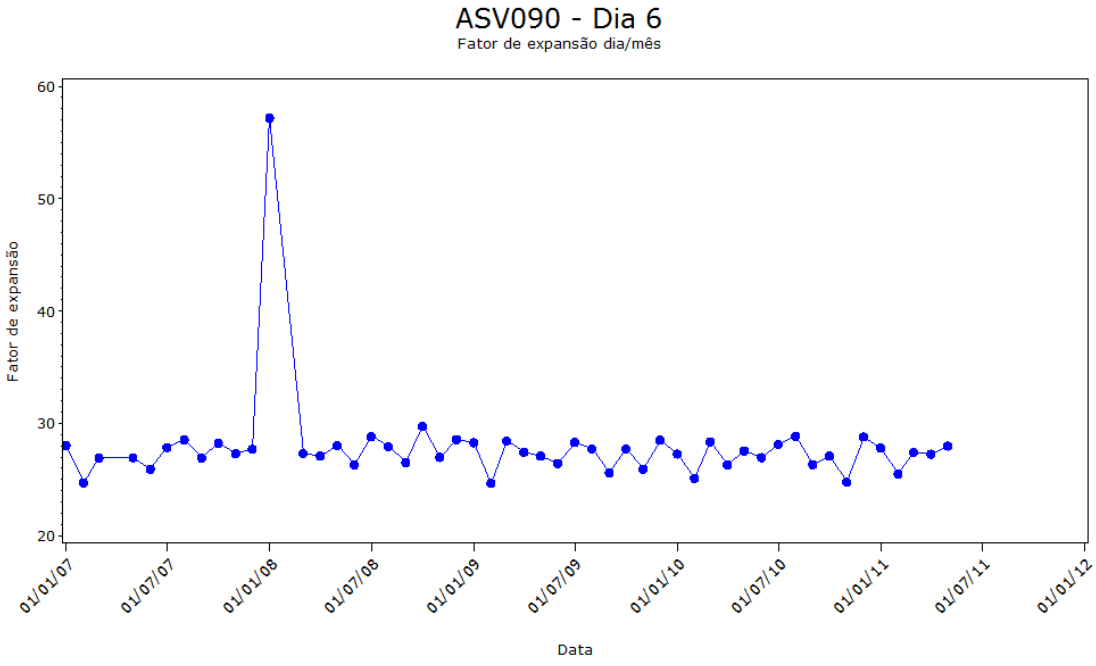


Figura 5.7: Observação dos fatores de expansão dia/mês do ponto ASV090

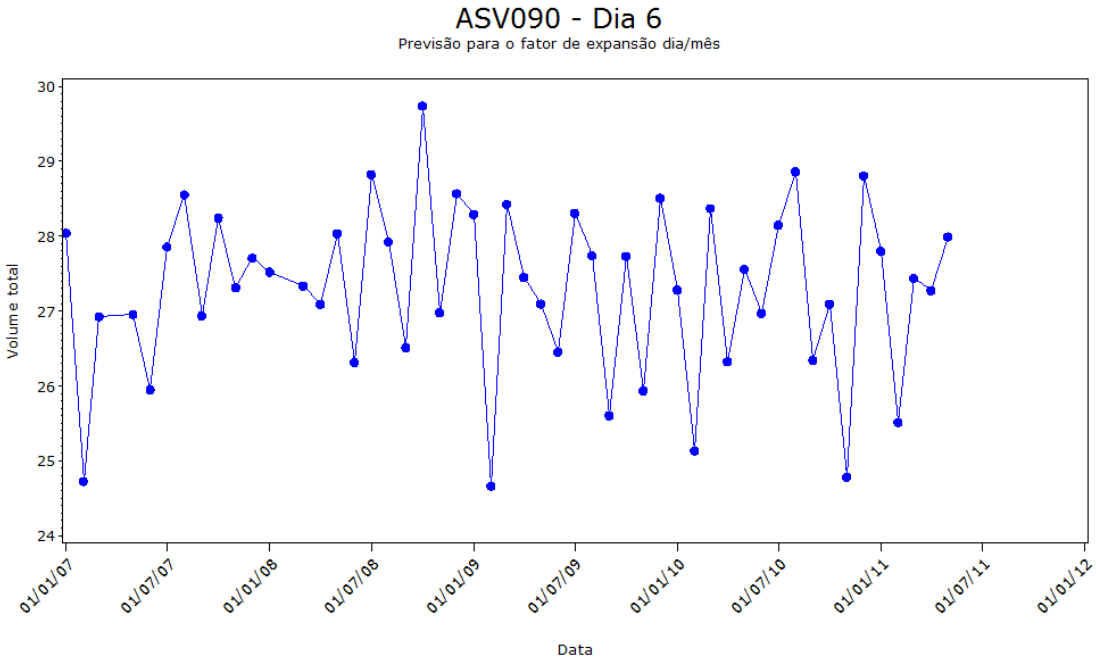


Figura 5.8: Observação dos fatores de expansão dia/mês do ponto ASV090 (corrigido)

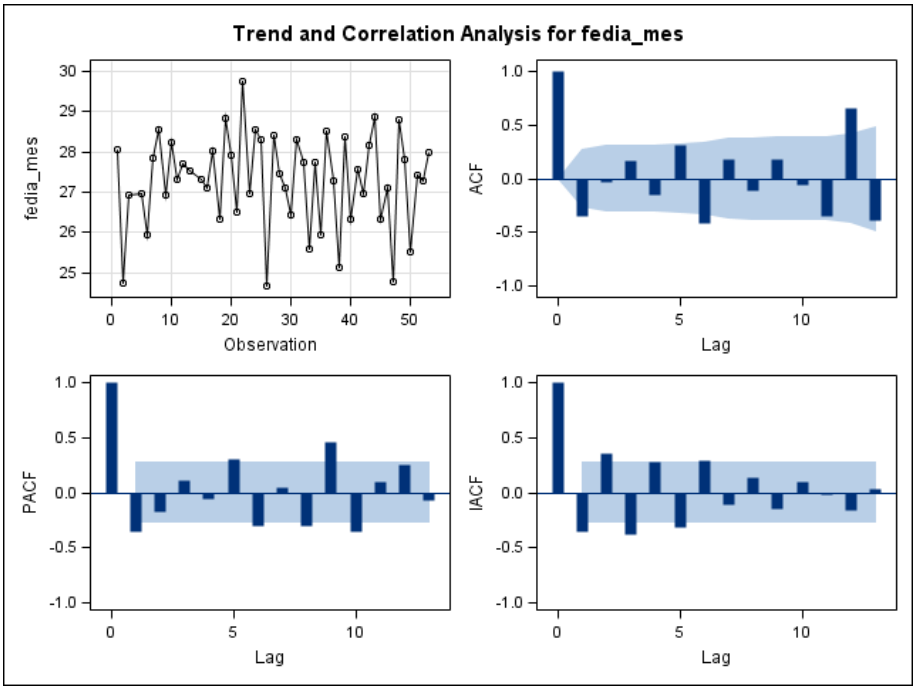


Figura 5.9: Correlograma dos fatores de expansão dia/mês do ponto ASV090

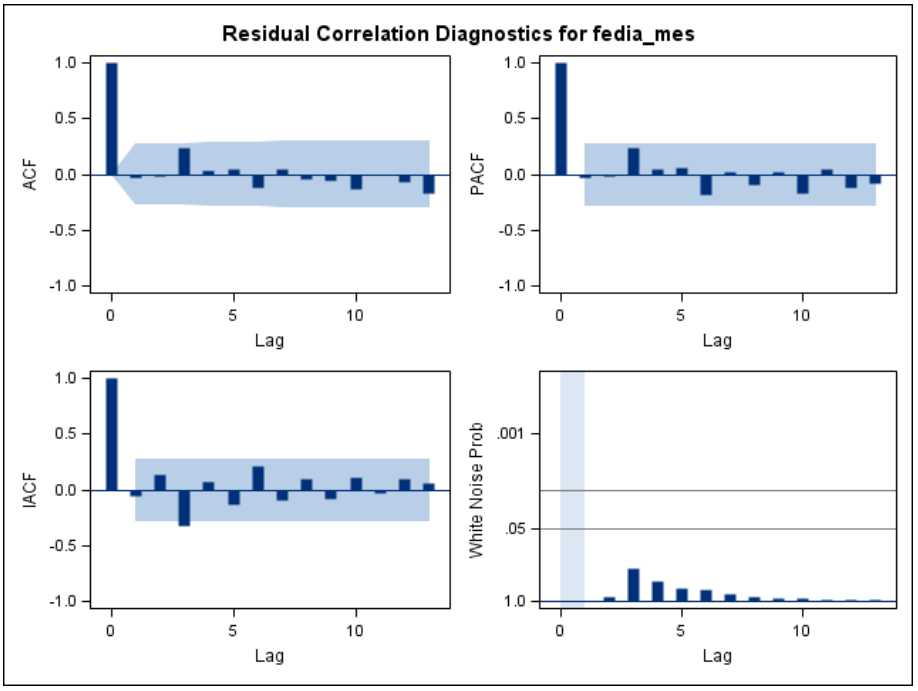


Figura 5.10: Correlograma dos resíduos do modelo SARIMA$\{(1,0,0)_{12},(0,0,0)\}$

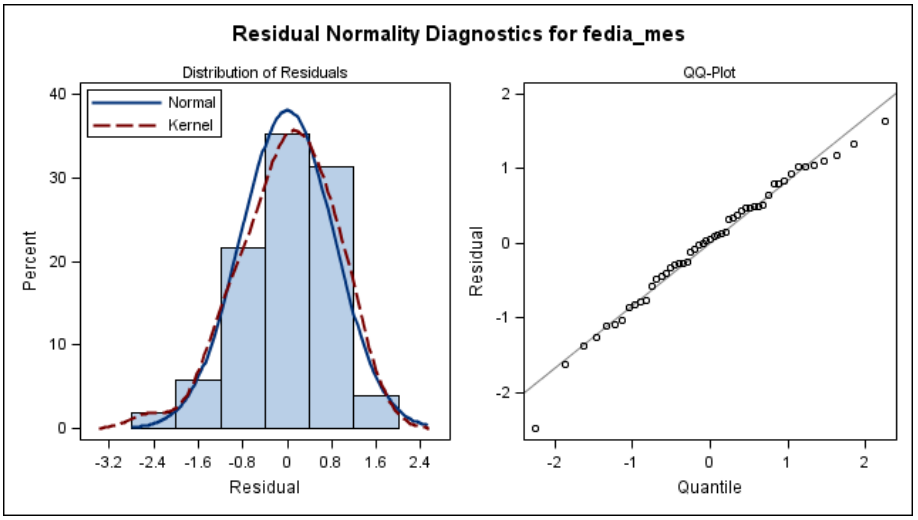


Figura 5.11: Histograma e *QQ-Plot* dos resíduos do modelo SARIMA$\{(1,0,0)_{12},(0,0,0)\}$

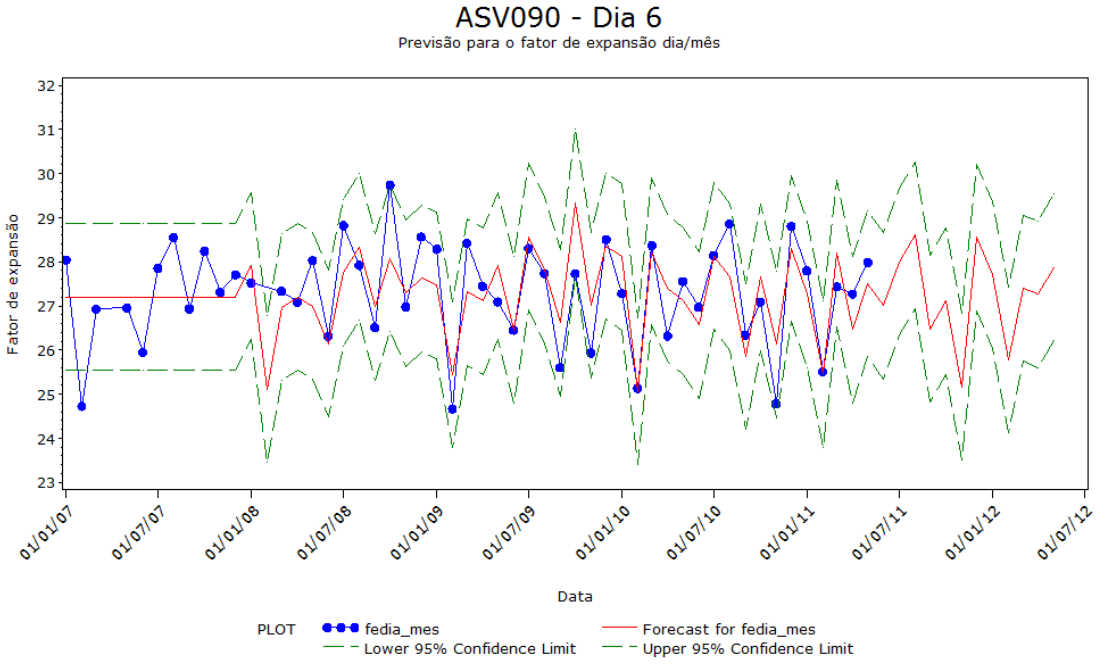


Figura 5.12: Previsão dos fatores de expansão dia/mês do ponto ASV090

5.4.2 Exemplo de um modelo razoavelmente bem ajustado

Agora o interesse é modelar o fator que expande o volume de três horas observadas em uma terça-feira para esse dia inteiro do equipamento ASV131. Análogo ao tópico anterior, a Figura 5.13 apresenta as observações e a Figura 5.14 os seus respectivos correlogramas. Nesse caso, o modelo escolhido foi um ARIMA$(0, 1, 4)_r$ com θ_4 à estimar e $\theta_i = 0$ para $i < 4$. O histograma na Figura 5.16 não se parece com uma distribuição normal, mas como há poucas observações, o foco foi feito no *QQ-Plot*, que sugere uma normalidade dos resíduos. Mesmo os pressupostos sendo atendidos, as previsões na Figura 5.17 não ficaram tão bons quanto na subseção anterior. Nota-se que a partir de um certo momento, as previsões ficam constantes. Ainda assim, o modelo foi capaz de gerar previsões que seguiram a variação das observações. Por essas razões, esse modelo foi classificado como razoavelmente bem ajustado.

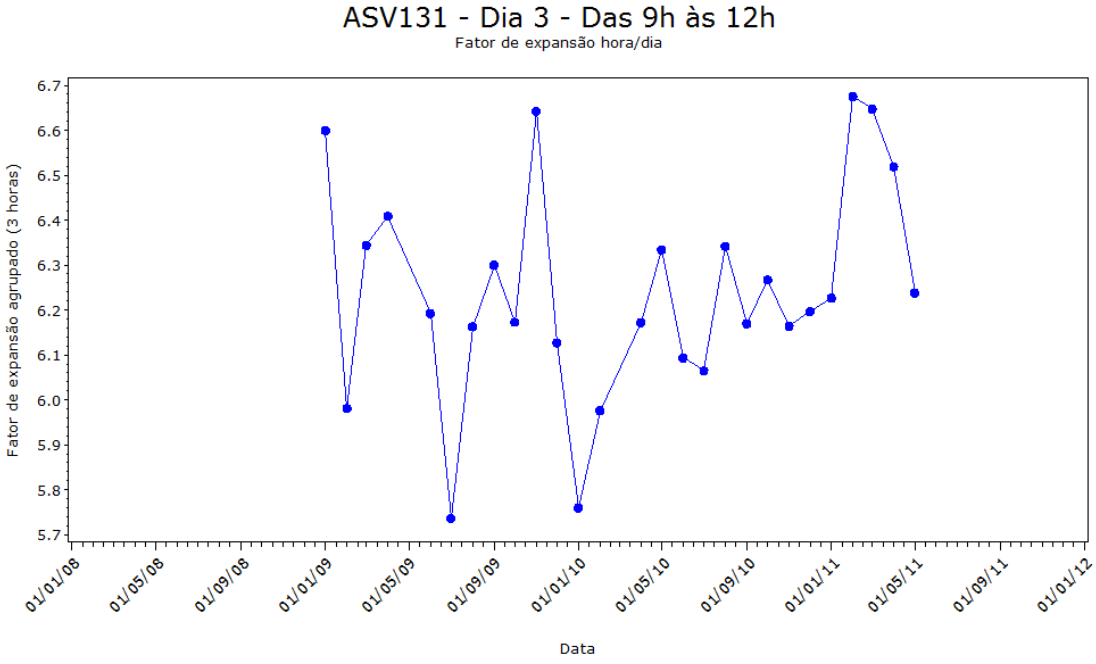


Figura 5.13: Observação dos fatores de expansão hora/dia do ponto ASV131

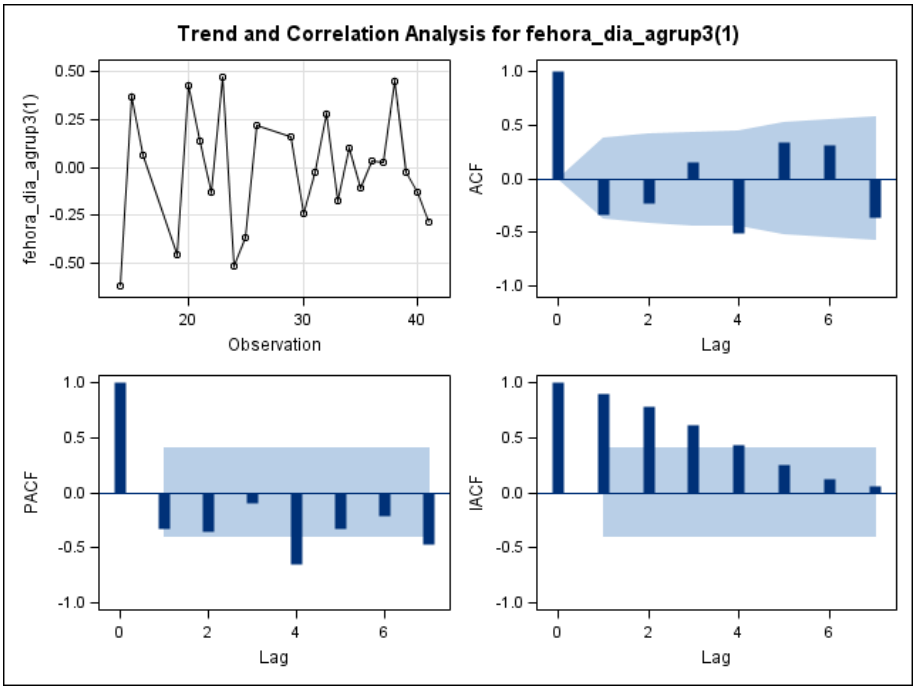


Figura 5.14: Correlograma dos fatores de expansão hora/dia do ponto ASV131 (d=1)

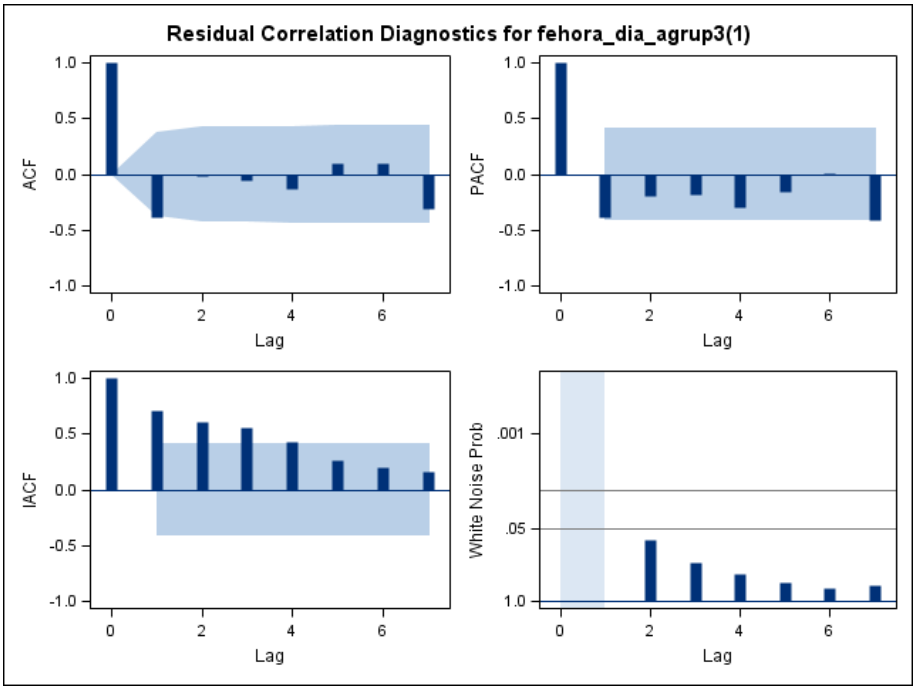


Figura 5.15: Correlograma dos resíduos do modelo ARIMA$(0, 1, 4)_r$

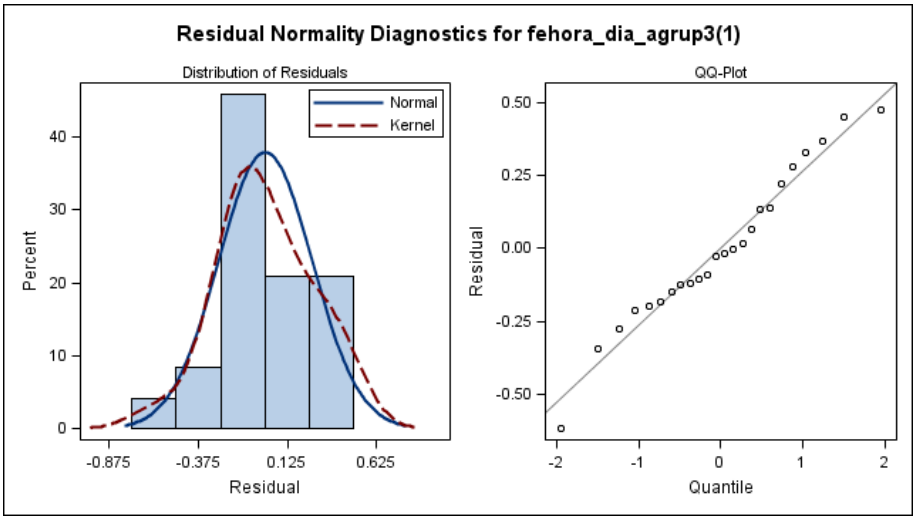


Figura 5.16: Histograma e *QQ-Plot* dos resíduos do modelo ARIMA$(0, 1, 4)_r$

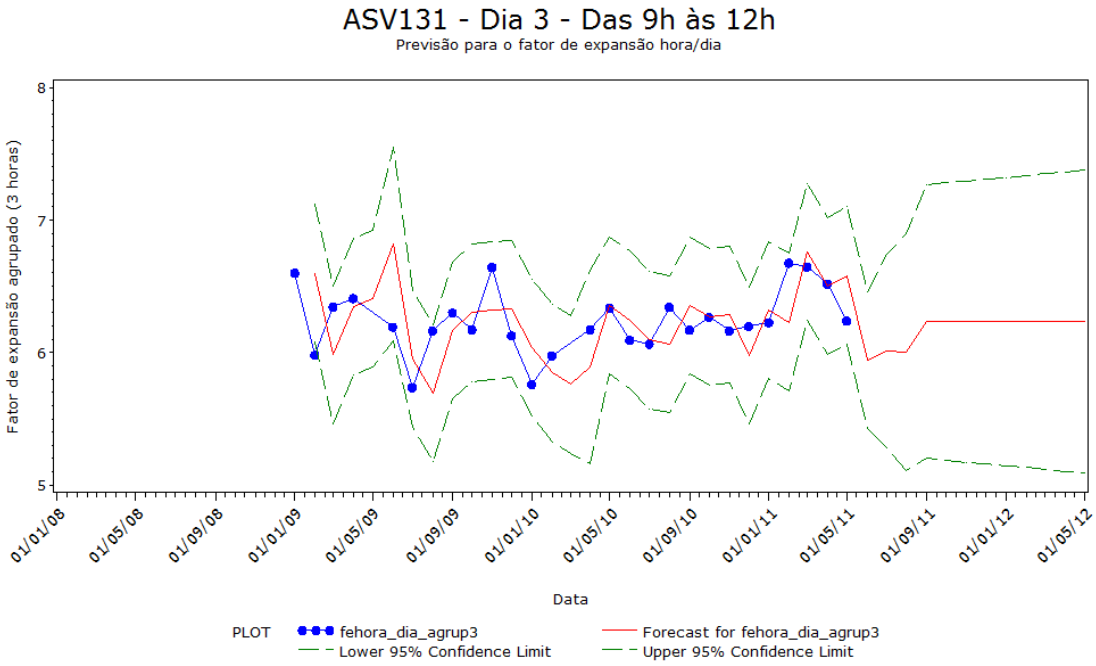


Figura 5.17: Previsão dos fatores de expansão hora/dia do ponto ASV131

5.4.3 Exemplo de um modelo mal ajustado

Nesse tópico é apresentada a tentativa de modelar o fator que expande o volume total dos meses para o ano inteiro do equipamento ASV012. A série é apresentada na Figura 5.18. Nessa série, nenhum valor foi interpretado como aberrante. Os seus correlogramas estão na Figura 5.19. Esses não sugerem que haja alguma correlação entre as observações e o tempo. Entretanto, uma vez que nossa programação não capta essa informação do gráfico, modelos SARIMA são estimados e comparados. Aqui, foi escolhido o modelo ARIMA$(1, 1, 0)$. A Figura 5.21 revela que o modelo não atende os pressupostos de normalidade dos resíduos. Da mesma forma, as previsões expostas na Figura 5.22 estão "atrasadas", ou seja, não estão próximas às observações. Tudo isso confirma o que já havia sido constatado pelo correlograma inicial: as observações não caracterizam uma série temporal. Logo, o modelo é considerado como mal ajustado.

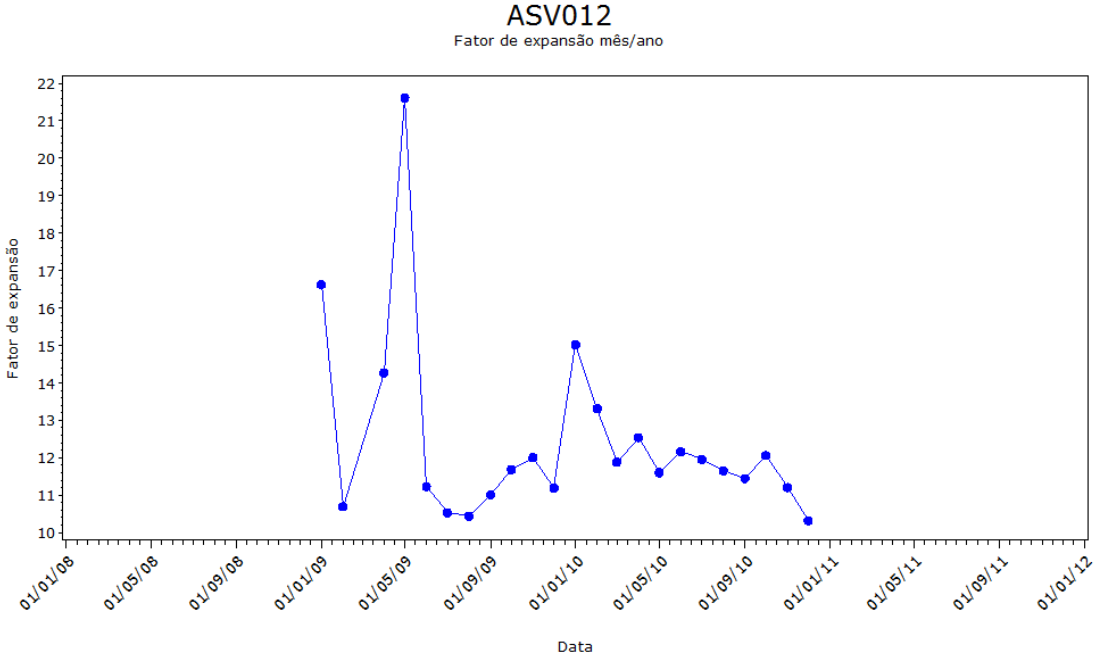


Figura 5.18: Observação dos fatores de expansão mês/ano do ponto ASV012

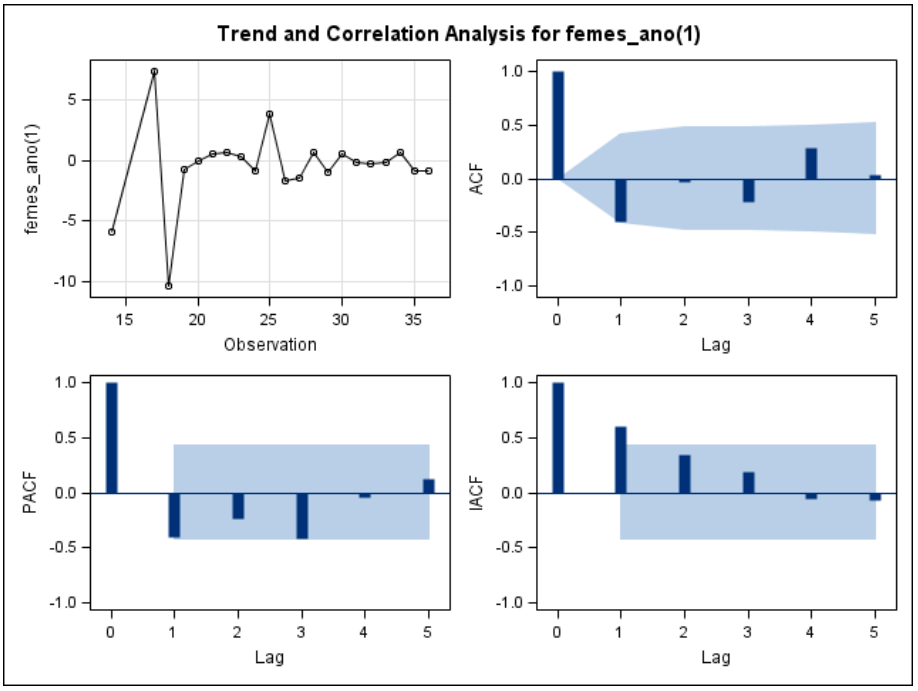


Figura 5.19: Correlograma dos fatores de expansão mês/ano do ponto ASV012

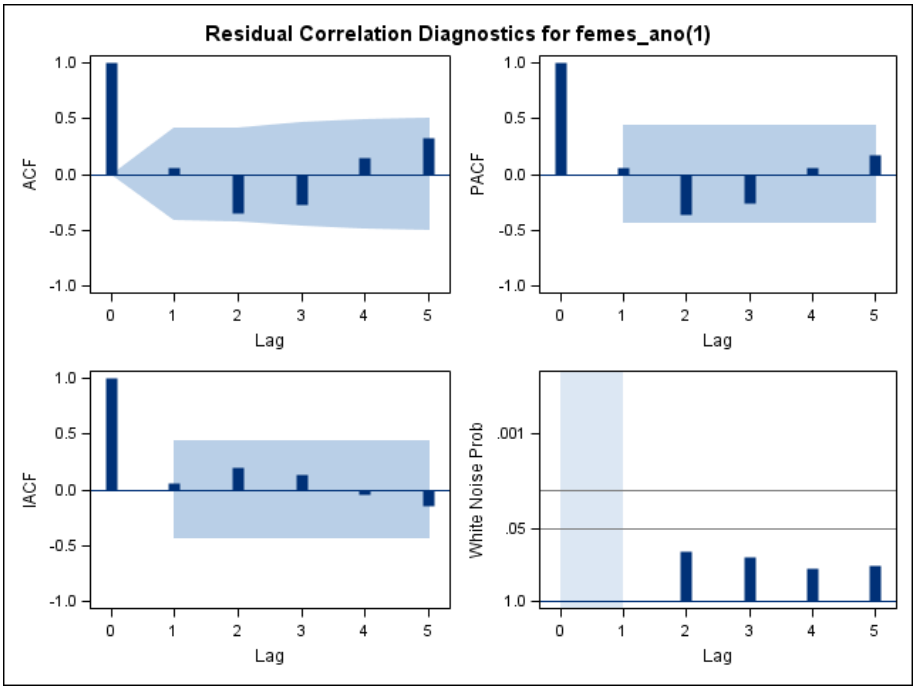


Figura 5.20: Correlograma dos resíduos do modelo ARIMA$(1, 1, 0)$

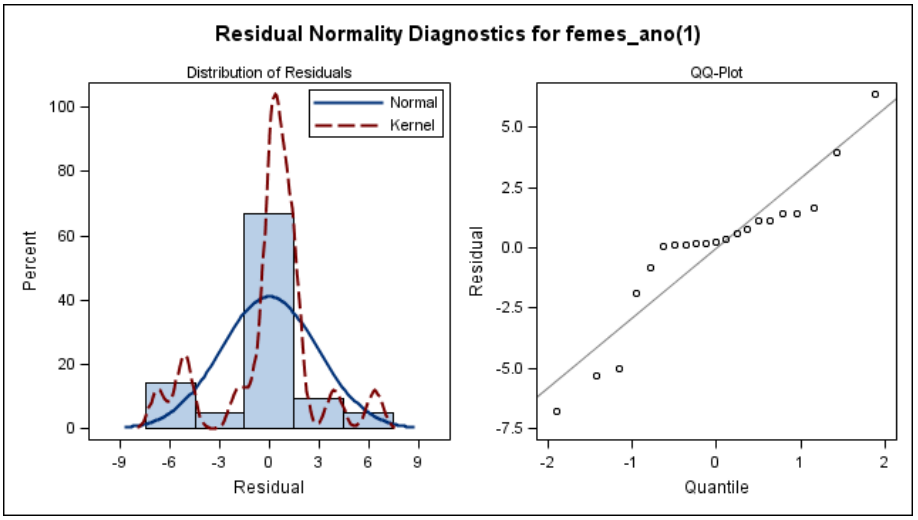


Figura 5.21: Histograma e *QQ-Plot* dos resíduos do modelo ARIMA$(1, 1, 0)$

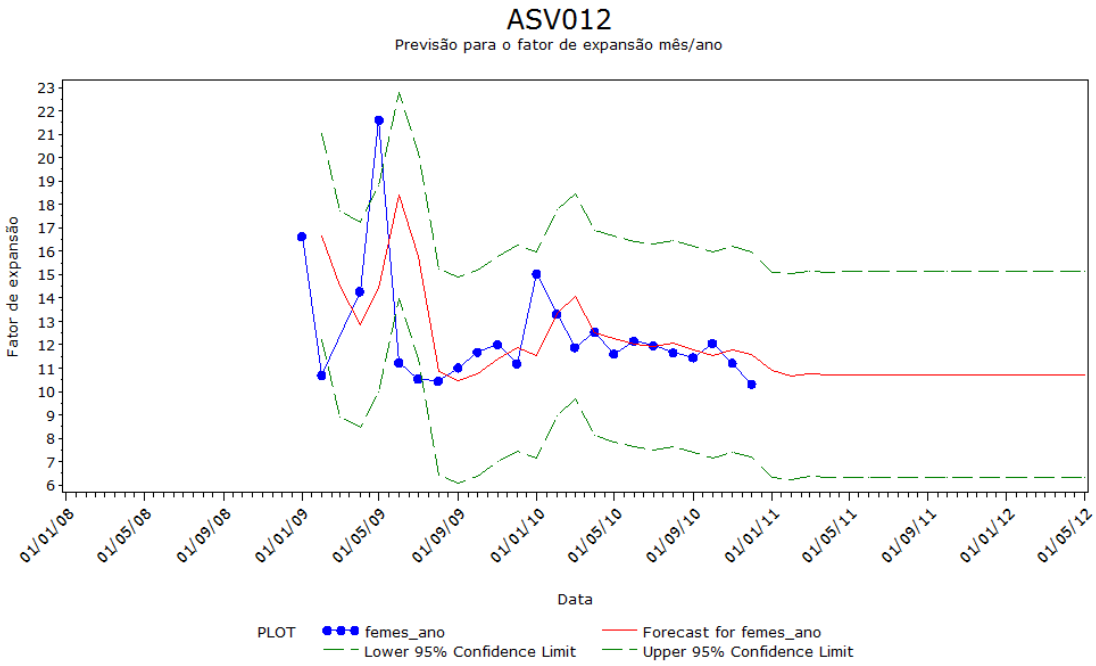


Figura 5.22: Previsão dos fatores de expansão mês/ano do ponto ASV012

5.4.4 Coeficiente de Theil

Com o intuito de apresentar um resumo sobre todos os modelos, um *Box-Plot* dos coeficientes de Theil (seção 3.5.7) foi montado e é exposto na Figura 5.23. Segundo esses valores, verifica-se então que os modelos estão bem ajustados, pois os coeficientes estão próximos de zero. Nota-se que esses valores nos modelos para o fator de expansão mês/ano tem uma variabilidade maior do que aqueles para os outros dois fatores. Isso aconteceu porque em vários equipamentos, não era possível calcular uma quantidade razoável do fator em questão. Logo, a modelagem realizada nesses casos, contava com poucas observações, o que gera modelos poucos consistentes.

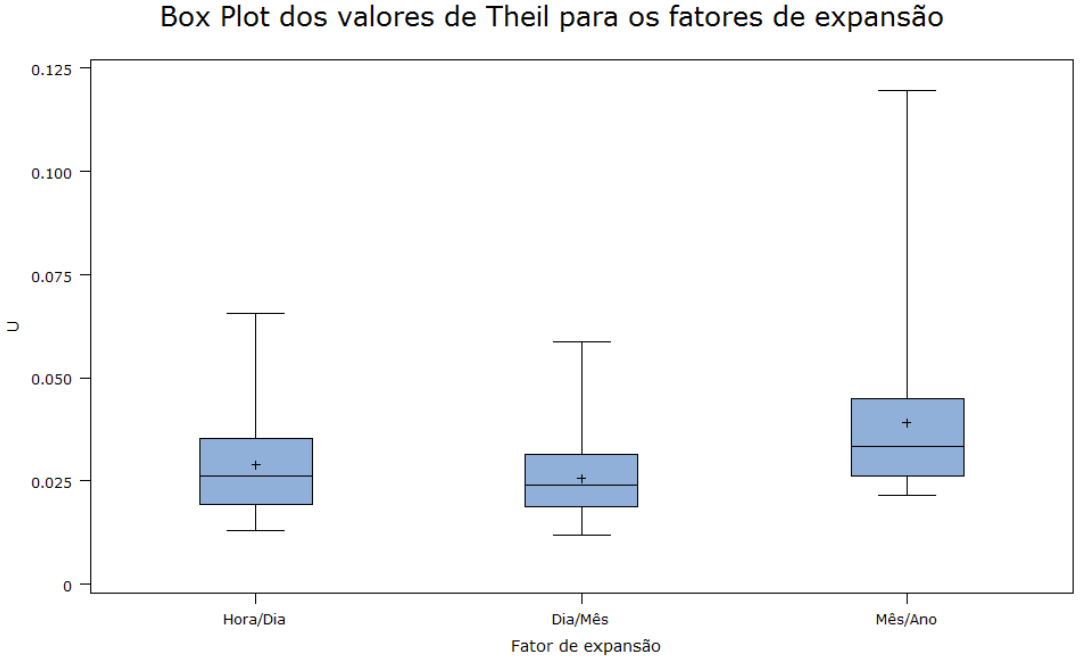


Figura 5.23: *Box-Plot* do coeficiente de Theil para os fatores de expansão

5.5 Previsão do volume médio diário anual

Com as previsões dos fatores de expansão já calculadas, basta observar a quantidade de veículos que circularam em uma certa interseção para aplicar os fatores e obter previsões do volume total no ano inteiro. Sendo assim, Claude (2012) foi à campo em 2011 e em 2012 para coletar dados com a ajuda de uma equipe do DETRAN-DF. Uma vez que o trabalho dela diz respeito à interseções onde não existem equipamentos eletrônicos, aplicaram-se às observações coletadas os fatores de expansão do equipamento mais próximo. Nas Tabelas 5.3 e 5.2, as colunas *inter-*

seção, via, ano, mês, dia da semana e *horário* são referentes ao local onde os dados foram coletados. Já a coluna *equipamento*, informa qual foi o ponto fiscalizador do qual foram utilizados os fatores de expansão. A última coluna é o resultado do produto entre o observado em campo e os fatores de expansão. Por exemplo, a primeira linha da Tabela 5.3 explicita que houve coleta de dados em uma segunda-feira de outubro de 2011 das 9h às 12h na via principal da interseção 5. Foi relatado que essa interseção está relacionada com aquela onde se encontra o equipamento ASV032. Seja X_c o valor coletado nesse local, $\hat{F}_{hd}$ a previsão do fator de expansão hora/dia do equipamento ASV032 no mês de outubro de 2011 na hora e dia especificados, $\hat{F}_{dm}$ a previsão do fator de expansão dia/mês de segunda para outubro de 2011 e $\hat{F}_{ma}$ a previsão do fator de expansão mês/ano. Portanto, a previsão do volume de veículos na via principal da interseção 5 ($\hat{V}_{2011}$) será igual à $X_c \times \hat{F}_{hd} \times \hat{F}_{dm} \times \hat{F}_{ma}$.

Tabela 5.2: Previsão do volume total de veículos no ano de 2012

Interseção	Equipamento	Via	Ano	Mês	Dia da semana	Horário	Previsão do volume em 2012
1	ASV131	Principal	2012	Janeiro	Terça	9h - 12h	2.781.677
1	ASV131	Secundária	2012	Janeiro	Terça	9h - 12h	395.401
1	ASV132	Principal	2012	Janeiro	Terça	9h - 12h	3.725.915
1	ASV132	Secundária	2012	Janeiro	Terça	9h - 12h	488.685
2	ASV131	Principal	2012	Janeiro	Terça	9h - 12h	4.432.648
2	ASV131	Secundária	2012	Janeiro	Terça	9h - 12h	92.491
3	ASV131	Principal	2012	Fevereiro	Quarta	9h - 12h	2.342.793
3	ASV131	Retorno	2012	Fevereiro	Quarta	9h - 12h	186.359
3	ASV132	Principal	2012	Fevereiro	Quarta	9h - 12h	3.457.435
3	ASV132	Secundária	2012	Fevereiro	Quarta	9h - 12h	276.079
4	ASV131	Principal	2012	Janeiro	Terça	13h - 16h	3.183.435
4	ASV131	Secundária	2012	Janeiro	Terça	13h - 16h	608.046
4	ASV132	Principal	2012	Janeiro	Terça	13h - 16h	3.248.162
7	ASV016	Principal	2012	Fevereiro	Quinta	14h - 17h	14.805.385
7	ASV016	Secundária	2012	Fevereiro	Quinta	14h - 17h	875.136
17	ASV090	Principal	2012	Janeiro	Sexta	14h - 17h	2.643.612
17	ASV094	Principal	2012	Janeiro	Sexta	14h - 17h	5.832.956
17	ASV094	Secundária	2012	Janeiro	Sexta	14h - 17h	1.009.774
18	ASV014	Principal	2012	Janeiro	Sexta	14h - 17h	5.743.610
18	ASV014	Secundária	2012	Janeiro	Sexta	14h - 17h	5.569.314
18	ASV094	Principal	2012	Janeiro	Sexta	14h - 17h	6.696.149
18	ASV094	Secundária	2012	Janeiro	Sexta	14h - 17h	5.248.962
22	ASV079	Principal	2012	Fevereiro	Quinta	9h - 12h	2.216.273
22	ASV079	Secundária	2012	Fevereiro	Quinta	9h - 12h	1.204.743
22	ASV135	Principal	2012	Fevereiro	Quinta	9h - 12h	1.689.406
22	ASV135	Secundária	2012	Fevereiro	Quinta	9h - 12h	754.814
26	ASV091	Principal	2012	Fevereiro	Sexta	14h - 17h	4.585.683
26	ASV091	Secundária	2012	Fevereiro	Sexta	14h - 17h	316.417
26	ASV139	Principal	2012	Fevereiro	Sexta	14h - 17h	4.541.464
26	ASV139	Secundária	2012	Fevereiro	Sexta	14h - 17h	530.390
27	ASV091	Principal	2012	Fevereiro	Sexta	14h - 17h	4.335.383
27	ASV091	Secundária	2012	Fevereiro	Sexta	14h - 17h	281.784
27	ASV139	Principal	2012	Fevereiro	Sexta	14h - 17h	4.379.400
27	ASV139	Secundária	2012	Fevereiro	Sexta	14h - 17h	405.159
29	ASV138	Principal	2012	Março	Quinta	14h - 17h	5.128.614
29	ASV138	Secundária	2012	Março	Quinta	14h - 17h	170.222
29	ASV139	Principal	2012	Março	Quinta	14h - 17h	5.714.958
29	ASV139	Secundária	2012	Março	Quinta	14h - 17h	190.859
30	ASV011	Principal	2012	Março	Quinta	14h - 17h	20.920.773
30	ASV011	Secundária	2012	Março	Quinta	14h - 17h	500.891

Tabela 5.3: Previsão do volume total de veículos no ano de 2011

Interseção	Equipamento	Via	Ano	Mês	Dia da semana	Horário	Previsão do volume em 2011
5	ASV032	Principal	2011	Outubro	Segunda	9h - 12h	12.528.787
5	ASV033	Principal	2011	Outubro	Segunda	9h - 12h	12.850.607
5	ASV131	Secundária	2011	Outubro	Segunda	9h - 12h	6.581.540
5	ASV132	Secundária	2011	Outubro	Segunda	9h - 12h	5.830.555
6	ASV016	Principal	2011	Dezembro	Terça	14h - 17h	12.169.943
6	ASV016	Secundária	2011	Dezembro	Terça	14h - 17h	3.196.918
6	ASV032	Principal	2011	Dezembro	Terça	14h - 17h	3.592.837
8	ASV034	Principal	2011	Dezembro	Segunda	14h - 17h	9.645.428
8	ASV034	Secundária	2011	Dezembro	Segunda	14h - 17h	2.840.403
8	ASV035	Principal	2011	Dezembro	Segunda	14h - 17h	19.345.636
8	ASV035	Secundária	2011	Dezembro	Segunda	14h - 17h	2.643.210
9	ASV035	Principal	2011	Dezembro	Sexta	14h - 17h	13.378.659
9	ASV035	Secundária	2011	Dezembro	Sexta	14h - 17h	1.602.438
9	ASV063	Principal	2011	Dezembro	Sexta	14h - 17h	9.897.133
10	ASV015	Principal	2011	Dezembro	Quinta	14h - 17h	5.418.763
10	ASV015	Secundária	2011	Dezembro	Quinta	14h - 17h	2.330.118
10	ASV035	Principal	2011	Dezembro	Quinta	14h - 17h	11.321.279
10	ASV141	Secundária	2011	Dezembro	Quinta	14h - 17h	4.513.358
11	ASV015	Principal	2011	Dezembro	Quinta	13h - 16h	4.347.275
11	ASV015	Secundária	2011	Dezembro	Quinta	13h - 16h	633.303
11	ASV035	Principal	2011	Dezembro	Quinta	13h - 16h	393.861
12	ASV035	Principal	2011	Dezembro	Quinta	14h - 17h	7.764.805
12	ASV035	Secundária	2011	Dezembro	Quinta	14h - 17h	65.678
13	ASV093	Principal	2011	Dezembro	Quinta	8h - 11h	2.471.728
13	ASV093	Secundária	2011	Dezembro	Quinta	8h - 11h	208.146
13	ASV141	Principal	2011	Dezembro	Quinta	8h - 11h	2.485.734
14	ASV093	Principal	2011	Dezembro	Quinta	8h - 11h	3.314.317
14	ASV141	Principal	2011	Dezembro	Quinta	8h - 11h	3.347.431
14	ASV141	Secundária	2011	Dezembro	Quinta	8h - 11h	157.165
15	ASV092	Principal	2011	Dezembro	Quinta	14h - 17h	4.237.341
15	ASV092	Secundária	2011	Dezembro	Quinta	14h - 17h	1.885.060
15	ASV093	Principal	2011	Dezembro	Quinta	14h - 17h	3.470.106
15	ASV093	Secundária	2011	Dezembro	Quinta	14h - 17h	2.338.650
16	ASV090	Principal	2011	Dezembro	Terça	9h - 12h	4.373.086
16	ASV090	Secundária	2011	Dezembro	Terça	9h - 12h	4.126.352
16	ASV093	Principal	2011	Dezembro	Terça	9h - 12h	3.757.085
16	ASV093	Secundária	2011	Dezembro	Terça	9h - 12h	3.657.571
19	ASV013	Secundária	2011	Outubro	Terça	9h - 12h	2.688.630
19	ASV014	Principal	2011	Outubro	Terça	9h - 12h	2.918.795
19	ASV094	Principal	2011	Outubro	Terça	9h - 12h	5.588.700
19	ASV094	Secundária	2011	Outubro	Terça	9h - 12h	4.529.848
20	ASV011	Principal	2011	Dezembro	Sexta	14h - 17h	14.386.062
20	ASV012	Principal	2011	Dezembro	Sexta	14h - 17h	12.679.438
20	ASV117	Secundária	2011	Dezembro	Sexta	14h - 17h	2.561.511
20	ASV135	Secundária	2011	Dezembro	Sexta	14h - 17h	2.028.948
21	ASV079	Principal	2011	Dezembro	Sexta	9h - 12h	3.396.897
21	ASV135	Principal	2011	Dezembro	Sexta	9h - 12h	3.199.784
21	ASV135	Secundária	2011	Dezembro	Sexta	9h - 12h	224.850
23	ASV011	Principal	2011	Dezembro	Quinta	14h - 17h	14.997.690
23	ASV011	Secundária	2011	Dezembro	Quinta	14h - 17h	560.760
24	ASV091	Principal	2011	Dezembro	Segunda	14h - 17h	2.901.349
24	ASV091	Secundária	2011	Dezembro	Segunda	14h - 17h	727.518
24	ASV139	Principal	2011	Dezembro	Segunda	14h - 17h	3.409.471
24	ASV139	Secundária	2011	Dezembro	Segunda	14h - 17h	785.317
25	ASV078	Secundária	2011	Dezembro	Quarta	14h - 17h	4.177.481
25	ASV091	Principal	2011	Dezembro	Quarta	14h - 17h	6.451.679
25	ASV091	Secundária	2011	Dezembro	Quarta	14h - 17h	4.302.939
25	ASV139	Principal	2011	Dezembro	Quarta	14h - 17h	3.678.428
28	ASV138	Principal	2011	Dezembro	Quarta	9h - 12h	6.754.551
28	ASV138	Secundária	2011	Dezembro	Quarta	9h - 12h	757.385
28	ASV139	Principal	2011	Dezembro	Quarta	9h - 12h	7.085.447
28	ASV139	Secundária	2011	Dezembro	Quarta	9h - 12h	817.846
31	ASV011	Principal	2011	Dezembro	Sexta	14h - 17h	18.873.086
31	ASV011	Secundária	2011	Dezembro	Sexta	14h - 17h	2.835.831
32	ASV011	Principal	2011	Dezembro	Segunda	9h - 12h	20.041.975
32	ASV011	Secundária	2011	Dezembro	Segunda	9h - 12h	1.268.820

Capítulo 6

Conclusão

O trabalho realizado é apenas o início de um projeto maior, que conta com o apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPQ), a fim de desenvolver modelos de previsão de acidentes em interseções, e que pode ser vastamente discutido para implementação de métodos complementares.

Um tratamento inicial do banco de dados mostrou a falta de precisão na contagem volumétrica. Problemas técnicos nos equipamentos de fiscalização impediram a obtenção de uma base 100% confiável. Tal transtorno pôde ser parcialmente contornado com o método de imputação de dados. A reamostragem de Jackknife se mostrou capaz de detectar grande parte das sub-contagens e dos *missings*. Inicialmente a metodologia pareceu ser robusta, mas notou-se em casos particulares que alguns sub-registros não foram detectados. Uma revisão desse procedimento é sugerida para trabalhos futuros. O primeiro passo seria averiguar se o intervalo para identificação de sub-registro poderia ser aumentado a fim captar mais observações com valores baixos.

O procedimento de imputação teve grande utilidade e substituiu da forma desejada os valores interpretados como erros do aparelho. Observou-se que a metodologia de imputação criada conseguiu captar a variabilidade dos dados e reproduzi-la na substituição de valores acusados como errados. Após a imputação dos valores, foi possível averiguar as interseções que menos apresentavam problemas, o que pôde facilitar a escolha de Claude (2012) quanto aos locais onde ela deveria coletar os dados para obtenção de resultados mais fidedignos e desenvolver sua dissertação de mestrado com mais credibilidade.

A rotina criada para escolha do melhor modelo SARIMA mostrou-se eficiente segundo o coeficiente de Theil, pois esses tiveram valores próximos de zero. Infelizmente, alguns modelos, principalmente quando a variável de interesse era o fator de

expansão mês/ano, não foram bem ajustados. Nesses casos, como foi apresentado no capítulo anterior, não havia razões para acreditar que as observações fossem correlacionadas com o tempo. Sugere-se então que seja implementada à rotina outro método de previsão de valores em situações parecidas.

De uma forma geral, o trabalho atendeu às expectativas iniciais e os resultados puderam ser aproveitados na dissertação de mestrado desenvolvida por Claude (2012) sob o título "Previsão da ocorrência de acidentes de trânsito em interseções de vias arteriais urbanas - O caso de Taguatinga-DF". Em tal pode-se notar a importância do cálculo do volume anual de veículos para uma previsão de acidentes em locais específicos.

Referências Bibliográficas

Akaike, H. (1974). *A new look at the statistical model identification. IEEE Transactions on Automatic Control*, AC-19:p. 716–723.

Akaike, H. (1977). *On entropy maximization principle. Applications of Statistics*, (P.R. Krishnaiah, Ed.)):p. 27–41.

Anderson, T. W. (1994). *The Statistical Analysis of Time Series*. Wiley.

Box, G.; Jenkins, G. M. e Reinsel, G. C. (2008). *Time Series Analysis: Forecasting and Control*. Prentice Hall.

Brasil (1997). *Código de trânsito brasileiro. Lei Nº 9.503*.

Brockwell, P. J. e Davis, R. A. (2002). *Introduction to Time Series Analysis and Forecasting,*. Respringer.

Chatfield, C. (1996). *The Analysis of Time Series: An Introduction*. Chapman and Hall.

Claude, G. F. M. (2012). *Previsão da ocorrência de acidentes de trânsito em interseções de vias arteriais urbanas - O caso de Taguatinga-DF*. Dissertação de mestrado, ENC - FT - UnB.

Cochran, W. G. (1977). *Sampling Techniques*. Wiley.

DNIT (2006). *Manual de Estudos de Tráfego. Publicação IPR No. 723.* Departamento Nacional de Infra-estrutura de Transportes.

Efron, B. e Tibshirani, R. (1994). *An Introduction to the Bootstrap*. Chapman and Hall.

Morettin, P. A. e Toloi, C. M. C. (2006). *Análise de Séries Temporais*. Edgard Blucher.

Pindyck, R. S. (2004). *Econometria: Modelos e Previsões. Elsevier, Rio de Janeiro.*

Rissanen, J. (1978). *Modelling by shortest data description. Automatica*, v. 14:p. 465–471.

Rupert G. Miller, J. (1964). *A Trustworthy Jackknife. The Annals of Mathematical Statistics*, v. 35:p. 1594–1605.

Schwartz, G. (1978). *Estimating the dimension of a model. Annals of Statistics*, v. 6:p. 461–464.

Shao, J. e Tu, D. (1995). *The jackknife and bootstrap*. Springer.

Silva, A. R.; Araújo, C. E. F. e Rocha, C. H. (2006). *Previsão da demanda de passageiros no eixo de oportunidades Taguatinga-Ceilândia. XX Congresso de Pesquisa e Ensino em Transportes*, v. 1:p. 467–478.