Revista Tecnologia e Sociedade
ISSN: 1984-3526

https://periodicos.utfpr.edu.br/rts

Um novo método de preenchimento de dados faltantes aplicado a séries temporais de concentração de MP10

Danilo Covaes Nogarotto
nogarotto.danilo@gmail.com
Universidade Estadual de Campinas

Nathalia Morgana Rissi
rissi.nathalia@gmail.com
Universidade Estadual de Campinas

Simone Andréa Pozza
simone.pozza@ft.unicamp.br
Universidade Estadual de Campinas

RESUMO

O estudo da poluição atmosférica, com ênfase em material particulado inalável (MP10), é necessário, devido ao dano causado à saúde da população, além de outros prejuízos. Séries históricas, usadas para previsão de dados, muitas vezes apresentam lacunas devido a vários fatores, que podem prejudicar a qualidade da previsão. O objetivo deste estudo foi propor um novo método de preenchimento de dados faltantes, e após a imputação dos dados, utilizar um modelo de séries temporais para prever a concentração de MP10. Foram obtidas, no Sistema QUALAR da CETESB, dados de concentrações diárias de MP10, entre os anos de 2010 a 2014, referente aos municípios de Campinas, Jundiaí e Paulínia, todos do Estado de São Paulo. O método de preenchimento de dados faltantes, proposto neste trabalho, foi chamado de TDEM (*Time-Dependent Effect Method*). O método TDEM foi comparado com dois outros métodos ("média durante o mês" e "média durante o ano") de preenchimento de dados faltantes, e apresentou os melhores resultados em relação aos Coeficiente de Correlação, Erro Quadrático Médio e Desvio Médio Absoluto. Após o preenchimento da série, os dados foram analisados com o intuito de prever concentrações futuras de MP10. Optou-se por modelos de séries temporais utilizando-se de modelos ARIMA e SARIMA. Os resultados mais satisfatórios foram obtidos pelo modelo SARIMA, cujos dados reais ficaram dentro dos limites de previsão de 95%.

PALAVRAS-CHAVE: Dados faltantes. Metodologia de Box-Jenkins. Material Particulado. Previsão.

INTRODUÇÃO

A caracterização da poluição do ar se dá pela presença de pelo menos uma substância química com concentração suficiente para causar danos aos seres vivos ou em materiais (VALLERO, 2008). Deste modo, é necessário o monitoramento de poluentes atmosféricos como forma de garantir maior qualidade de vida da população.

Um dos principais poluentes do ar é o Material Particulado (MP), que consiste de uma mistura de partículas sólidas e líquidas, capazes de ficar em suspensão, como, por exemplo, a poeira e a fuligem. As principais fontes de emissão de MP são os processos industriais, veículos automotores e a queima da biomassa. O MP10 são as partículas que possuem diâmetro aerodinâmico menor ou igual a 10 µm (ALVES, 2005; VALLERO, 2008; CETESB, 2018) e são conhecidas como partículas inaláveis.

A presença de dados faltantes, em séries temporais de dados, é um problema frequente em vários campos científicos, agravando-se na área de pesquisa ambiental (XIA et al., 1999; GÓMEZ-CARRACEDO et al., 2014; ZAINURI, JEMAIN e MUDA, 2015). O impacto dos dados faltantes no resultado da análise estatística depende do mecanismo que resultou na perda dos mesmos e de como o analista irá lidar com isso. Assim, é de vital importância a busca por critérios adequados para substituir essas lacunas de dados com valores apropriados (PLAIA e BONDI, 2006).

Métodos de preenchimento de dados faltantes

Gómez-Carracedo et al. (2014) analisaram uma estação de monitoramento de 8 poluentes gasosos na cidade de La Coruña, Espanha, durante os anos de 2006, 2009 e 2010. Cinco métodos de imputação de dados foram usados, sendo 4 métodos de preenchimento simples e um método de preenchimento múltiplo. O estudo de Zainuri, Jemain e Muda (2015) abordou formas de diminuir o impacto desse problema, comparando 6 métodos de preenchimento de dados. Neste caso, os métodos do algoritmo EM (*Expectation Maximization*) e o do vizinho mais próximo foram os que apresentaram melhores resultados. Junger e Ponce de Leon (2015) propuseram um novo método de imputação de dados faltantes baseado no algoritmo EM, em que se considera a possibilidade de diversos filtros temporais como ARIMA (*Auto Regressive Integrated Moving Average*), *Splines* (curvas polinomias de interpolação) e GAM (*Generalized Aditive Models* ou Modelos Aditivos Generalizados). O método de preenchimento proposto exibiu boa exatidão e precisão em diferentes configurações em relação aos padrões de observações faltantes.

O método sugerido por Plaia e Bondi (2006) denominado como SDEM (*Site-Dependent Effect Method*), foi aplicado em seus estudos com o objetivo de propor um novo meio de preenchimento de dados faltantes. Seus dados se baseavam na concentração de MP10 medido a cada 2 h, por 8 estações de monitoramento distribuídas pela área metropolitana de Palermo, Itália, durante o ano de 2003. De modo geral, o método SDEM considera um efeito principal (cujas características são similares entre si) e efeitos secundários, estes relacionados ao efeito principal. Em Plaia e Bondi (2006), 8 estações de monitoramento foram consideradas como o efeito principal (fator de espaço), pois suas características são mais próximas

umas das outras. Já os efeitos secundários são a semana, dia e hora, que são considerados em relação às estações (efeitos de tempo, em relação ao espaço).

Após o preenchimento das lacunas dos dados, uma das alternativas de análise são os modelos de séries temporais, que são bastantes utilizados na análise de poluentes atmosféricos (BELL, SAMET e DOMINICE, 2004; POZZA et al., 2010; REISEN et al., 2014), mas, nesses tipos de modelos, é necessário que a série esteja completa (PINTO, REISEN e MONTE, 2018).

Modelos de séries temporais

Para realizar análises e o estudo do comportamento das variáveis poluentes, especificamente do MP, diversos autores (GOYAL, CHAN e JAISWAL, 2006; POZZA et al., 2010; REISEN et al., 2014) têm utilizado modelos de séries temporais. Modelos de Séries Temporais são técnicas que visam explicar a influência do tempo nos dados, tanto nas observações do passado, quanto nas do futuro, que auxiliam na compreensão e na previsão dos dados temporais (MORETTIN e TOLOI, 2006; EHLERS, 2009; PINTO, REISEN e MONTE, 2018). Segundo Bell, Samet e Dominice (2004), estudar as séries temporais da concentração dos poluentes atmosféricos é relevante para os processos de regulamentação e para estabelecer normas para a poluição, em níveis considerados suficientemente seguros para a saúde humana.

Por exemplo, em São José dos Campos (SP), durante os anos de 2000 e 2001, foi realizado um estudo de séries temporais, utilizando-se dados diários do número de internações por pneumonia, da concentração dos poluentes SO_2, O_3, e MP10, além das medições de temperatura e umidade do clima. Para estimar a associação entre as internações por pneumonia e a poluição atmosférica, utilizaram-se Modelos Aditivos Generalizados de regressão de Poisson (NASCIMENTO et al., 2006).

Modelos SARIMA (*Seasonal Auto Regressive Integrated Moving Average*) e ARIMA (*Auto Regressive Integrated Moving Average*) foram usados para prever a concentração do MP em São Carlos, Brasil (LIMA et al., 2009; POZZA et al., 2010) e Daca, Bangladesh (Rahman e Hossain, 2012). Goyal, Chan e Jaiswal (2006) utilizaram de modelos de regressão combinados com modelo ARIMA para a previsão do MP10 nas cidades de Delhi e Hong Kong, conseguindo explicar mais de 70% da variação total. Monte, Albuquerque e Reisen (2015) previram a concentração horária de ozônio na cidade de Vitória (ES) no ano de 2011, usando uma combinação de modelos de regressão, modelos ARMA, e um modelo de Heterocedasticidade condicional autorregressivo generalizado (chamado ARMAX-GARCH). O modelo ARMAX-GARCH se mostrou eficaz para prever episódios de concentração acima de 80 $\mu g/m^3$.

Desta maneira, este trabalho tem dois objetivos. Primeiramente, o objetivo principal foi propor um novo método simples de preenchimento de dados faltantes da série dos dados de concentração diária do MP10, que considere os efeitos temporais do mês, ano, semana do mês e dia da semana, a fim de obter uma série completa dos dados. Com a série completa dos dados é possível fazer sua análise por qualquer pesquisador, em diferentes situações e usando diversas técnicas. Como exemplo do uso da série completa, este trabalho também teve um segundo objetivo, que foi utilizar modelos de séries temporais para realizar previsões da

concentração de MP10, visando encontrar padrões de tendência, bem como a existência de variação sazonal, baseado nos dados de 3 cidades do interior de São Paulo: Campinas, Jundiaí e Paulínia.

METODOLOGIA

Base de dados

Para a avaliação do método TDEM, foram escolhidos dados de concentração diária do MP10 em 3 cidades do Estado de São Paulo: Campinas, Jundiaí e Paulínia. Tais cidades caracterizam-se pelo clima tropical de altitude, com chuvas no verão e seca no inverno (CEPAGRI, 2018). Campinas destaca-se por ser um polo industrial e tecnológico com mais de 1 milhão de habitantes (D'AMÉLIO, CAMPOS e ALVIM, 2017). Paulínia possui um dos principais complexos petroquímicos do país (D'AMÉLIO, CAMPOS e ALVIM, 2017). Jundiaí apresenta uma população de pouco mais de 410 mil habitantes, e localiza-se entre duas regiões metropolitanas importantes, de São Paulo e Campinas (IBGE, 2018).

O banco de dados utilizado foi obtido no site da CETESB (Companhia Ambiental do Estado de São Paulo), na plataforma Qualidade do Ar (QUALAR, 2016). Os dados de concentração diária de MP10 foram referentes ao período de 5 anos consecutivos (2010 a 2014). Foram utilizados os dados de 2015 para validar o desempenho de previsão do modelo ajustado. Essas concentrações correspondem a medições das estações automáticas da CETESB, existente em cada uma das 3 cidades.

Inicialmente, para realizar o estudo das séries temporais, é necessário que todos os valores diários estejam totalmente preenchidos (MORETTIN e TOLOI, 2006; PINTO, REISEN e MONTE, 2018). Entretanto, existiam de 1,5% até 2,0% de dados faltantes na base de dados das cidades escolhidas. Assim, foi necessário um tratamento prévio, para o preenchimento dos dados faltantes.

Método de preenchimento de dados faltantes

O método SDEM, proposto por Plaia e Bondi (2006), é baseado na equação 1, em que é possível obter uma estimativa do valor do dado faltante ($\hat{x}_{swdh}$).

$$\hat{x}_{swd} = \bar{x}_{.wdh} + \frac{1}{2}\left(\bar{x}_{sw.} - \sum_{s=1}^{S}\frac{\bar{x}_{sw.}}{S}\right) + \frac{1}{2}\left(\bar{x}_{s.d.} - \sum_{s=1}^{S}\frac{\bar{x}_{s.d.}}{S}\right) + \frac{1}{2}\left(\bar{x}_{s..h} - \sum_{s=1}^{S}\frac{\bar{x}_{s..h}}{S}\right) \quad (1)$$

Onde: s = estação de monitoramento (1, 2, 3,...,S), w =semana do ano (1, 2, 3,...,53), d = dia da semana (1, 2, 3,...,7) e h = hora do dia (2, 4, 6,...,24). Os valores

$\bar{x}_{.wdh}$, $\bar{x}_{sw..}$, $\bar{x}_{s.d.}$ e $\bar{x}_{s..h}$ são valores médios considerando os efeitos do tempo. Por exemplo, $\bar{x}_{sw..}$ é a média dos valores observados na semana w e na estação s.

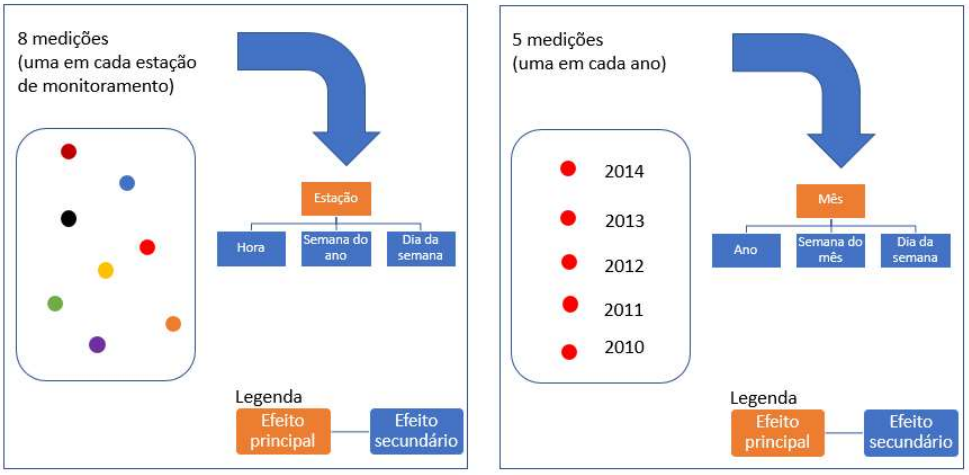
Os dados deste trabalho têm características diferentes do proposto pelo método SDEM original (PLAIA e BONDI, 2006), onde os dados de concentração do MP10 foram medidos em várias estações de monitoramento. Em nosso trabalho, substituímos a variável estação de monitoramento, por 5 anos de coleta numa mesma estação de monitoramento. Por exemplo, suponha a quarta-feira, da terceira semana do mês de janeiro. No banco de dados do Plaia e Bondi (2006), tem-se 8 medições, uma para cada estação. No caso deste estudo, temos 5 medições, uma para cada ano (figura 1).

No método SDEM (PLAIA e BONDI, 2006), para dados de duas estações ("A" e "B") quaisquer entre as 8 disponíveis, supõe-se que medições de uma determinada estação "A" seriam semelhantes com medições numa outra estação "B", cada uma em diferentes locais da cidade, considerando a mesma hora, semana do ano e dia da semana. Ou seja, a estação seria o efeito principal, e hora, semana do ano e dia da semana seriam efeitos secundários, medidos em relação ao efeito principal.

Já com os dados do presente estudo, considerou-se que medições do mesmo mês, ao longo dos anos, possui condições similares para compararmos valores. Por exemplo, dados diários de janeiro de 2010 seriam semelhantes com janeiro de 2011, considerando o mesmo dia da semana e a mesma semana dentro do mês.

Assim, foi proposto um ajuste da equação 1 para adequar a situação aos dados deste estudo. Nesta adaptação, considerou-se os meses do ano como efeito principal, pois suas características são mais próximas umas das outras, devido a condições climáticas (estações do ano, umidade relativa, velocidade do vento, entre outras). Os efeitos secundários (ano, semana do mês e dia da semana) foram considerados em relação ao mês (figura 1).

Figura 1 – Esquema comparativo do efeito principal e secundário para o método de Plaia e Bondi (a) e o método TDEM (b).



(a) (b)

Fonte: autoria própria

Dessa maneira, propôs-se uma adaptação do método SDEM (equação 1). Este novo método de preenchimento de dados faltantes é aqui chamado de TDEM (*Time-Dependence Effect Method*). A estimativa do valor faltante ($\hat{x}_{masd}$) é dada pela equação 2.

$$\hat{x}_{masd} = \bar{x}_{.asd} + \frac{1}{2}\left(\bar{x}_{ma..} - \sum_{m=1}^{M}\frac{\bar{x}_{ma..}}{M}\right) + \frac{1}{2}\left(\bar{x}_{m.s.} - \sum_{m=1}^{M}\frac{\bar{x}_{m.s.}}{M}\right) + \frac{1}{2}\left(\bar{x}_{m..d} - \sum_{m=1}^{M}\frac{\bar{x}_{m..d}}{M}\right) \quad (2)$$

Onde, m = mês (1, 2,..., 12); a = ano (2010, 2011,..., 2014); s = semana do mês (1, 2, 3, 4) e d = dia da semana (1,2,...,7). Os valores $\bar{x}_{.asd}$, $\bar{x}_{ma..}$, $\bar{x}_{m.s.}$ e $\bar{x}_{m..d}$ são valores médios considerando os efeitos do tempo. Por exemplo, $\bar{x}_{ma..}$ é a média dos valores observados no mês m e no ano a. Já o valor de $\bar{x}_{.asd}$, é a média dos valores observados no ano a, na semana do mês s e no dia da semana d.

No método TDEM (equação 2) consideramos os efeitos do ano, da semana do mês e do dia da semana, como efeitos secundários. O efeito principal é do mês m. Ou seja, $\frac{1}{2}\left(\bar{x}_{ma..} - \sum_{m=1}^{M}\frac{\bar{x}_{ma..}}{M}\right)$ é o efeito do ano a, $\frac{1}{2}\left(\bar{x}_{m.s.} - \sum_{m=1}^{M}\frac{\bar{x}_{m.s.}}{M}\right)$ é o efeito da semana do mês s, e $\frac{1}{2}\left(\bar{x}_{m..d} - \sum_{m=1}^{M}\frac{\bar{x}_{m..d}}{M}\right)$ é o efeito dia da semana do d, todos em relação ao mês m.

Deste modo, o método TDEM pode ser usado nos casos em que há apenas uma série de dados disponível em vários anos. Enquanto que o método SDEM (PLAIA e BONDI, 2006) pode ser utilizado em casos em que há várias estações de monitoramento numa mesma cidade.

Para avaliar o método de preenchimento de dados faltantes proposto, o TDEM, foi feita uma comparação com outros dois métodos, descritos pelas equações 3 e 4.

- "Média durante mês": Baseia-se na média do dia da semana, para o mês m, ao longo de todos os anos e semanas (equação 3).

$$\hat{x}_{masd} = \bar{x}_{m..d} \quad (3)$$

- "Média durante o ano": Baseia-se na média de todos os meses do ano a, semana s e do dia d (equação 4).

$$\hat{x}_{masd} = \bar{x}_{.asd} \quad (4)$$

Indicadores de desempenho

Para comparar, e decidir qual método seria utilizado no tratamento dos dados faltantes, utilizou-se os critérios do Erro Quadrático Médio (EQM), do Coeficiente de Correlação (COR) e do Desvio Médio Absoluto (DMA) (PLAIA e BONDI, 2006). Quanto menor os valores de EQM (equação 5) e DMA (equação 7), e quanto maior o valor de COR (equação 6), melhor é o método de preenchimento do dado faltante.

Erro Quadrático Médio: $EQM = \frac{1}{N}\sum_{i=1}^{N}(O_i - P_i)^2$ (5)

Coeficiente de Correlação: $COR = \frac{1}{N}\frac{\sum_{i=1}^{N}[(O_i-\bar{O})(P_i-\bar{P})]}{\sigma_O \sigma_P}$ (6)

Desvio Médio Absoluto: $DMA = \frac{1}{N}\sum_{i=1}^{N}|O_i - P_i|$ (7)

Onde: O_i é o i-ésimo dado observado, com $\bar{O}$ a média dos dados observados, P_i é o i-ésimo dado preenchido, com $\bar{P}$ a média dos dados preenchidos, σ_O é o desvio-padrão dos dados observados, e σ_P é o desvio-padrão dos dados preenchidos, e por fim, N é o total de preenchimento (PLAIA e BONDI, 2006).

Para os cálculos dos critérios (equações 5 a 7) foram considerados apenas os dias em que havia sido monitorado o poluente, ou seja, não se considerou os dados que foram completados. Após completar as lacunas da série, foi ajustado um modelo de séries temporais.

Análise de séries temporais

Com a série completa, a fim de exemplificar uma possível análise de dados, utilizou-se a abordagem de Box Jenkins (MORETTIN e TOLOI, 2006). Considerou-se tanto o ajuste dos modelos ARIMA(p, d, q) quanto o estudo da componente sazonal, por meio dos modelos ARIMA sazonal (SARIMA) de ordem (p, d, q) x $(P, D, Q)\,S$. Assim, p é a ordem do processo autorregressivo, q é a ordem do processo de médias móveis e d é quantidade de vezes que a série original foi diferenciada. Já P é a ordem do processo autorregressivo da parte sazonal, Q é a ordem do processo de médias móveis da parte sazonal, D é o grau de diferenciação da parte sazonal e S é o período da sazonalidade (MORETTIN e TOLOI, 2006). Considerou-se tanto a série diária, quanto a série mensal dos dados de MP10. Ou seja, foram ajustados modelos com os dados diários e com os dados mensais.

Para identificação dos modelos das séries, foram analisadas as funções de auto correlação (fac) e as funções de auto correlação parcial (facp) (MORETTIN e TOLOI, 2006; GOYAL, CHAN e JAISWAL, 2006; GUJARATI e PORTER, 2012). Nos modelos SARIMA, considerou-se a sazonalidade com um período de 7 dias, supondo que as concentrações de MP10 podem ter comportamentos similares no mesmo dia da semana, ou seja, uma sazonalidade semanal. Já para previsão mensal, do ano de 2015, considerou-se a sazonalidade com um período de 12 meses, supondo que a concentração de MP10 pode ter comportamento similar no mesmo mês ao longo dos anos.

Por fim, foram previstas as concentrações diárias do mês de janeiro de 2015, e as concentrações mensais do ano de 2015 das 3 cidades. Para a escolha do melhor modelo, foram realizados os cálculos de três critérios de seleção definidos: AIC (Critério de Informação Akaike), LOG (Log-verossimilhança) e EQMP (erro quadrático médio de previsão) (GUJARATI e PORTER, 2012). As medidas do AIC e EQMP são melhores quando seus valores são menores. Já o LOG, quanto maior for o seu valor, melhor.

Todos os cálculos envolvidos neste trabalho foram realizados com o auxílio dos pacotes da plataforma R (R CORE TEAM, 2017).

DESENVOLVIMENTO (RESULTADOS E DISCUSSÕES)

Imputação dos dados faltantes e comparação dos métodos

Vários estudos apresentaram e compararam diferentes métodos de preenchimento de dados faltantes, como pode ser visto na tabela 1, que apresenta as principais características destes trabalhos. Duas pesquisas apresentaram métodos de preenchimento múltiplo, diferente do apresentado em nosso estudo, de Plaia e Bondi (2006) e de Zainuri, Jermain e Muda (2015), que apenas trabalharam com métodos simples de preenchimento. Os métodos múltiplos ou o algoritmo EM exigem um esforço computacional alto (GÓMEZ-CARRACEDO et al., 2014) maior do que os métodos SDEM e TDEM, que envolvem apenas o cálculo de médias. Como destacado no objetivo, nossa ideia principal é de apresentar um novo método simples de preenchimento de dados faltantes.

Tabela 1: Comparação de métodos de preenchimento de dados faltantes.

Estudo	Método principal	Preenchi-mento	Série preenchida	Período	Local
Este estudo	TDEM	Simples	Dados diários de MP10	2010 a 2014	Campinas, Jundiaí e Paulínia (SP), Brasil
Plaia e Bondi (2006)	SDEM	Simples	Dados a cada 2 horas de MP10	2003	Palermo, Itália
Gómez-Carracedo et al. (2014)	*Multiple Imputa-tion*	Múltiplo	Dados diários de 8 poluentes gasosos	2006, 2009 e 2010	La Coruña, Espanha
Junger e Ponce de Leon (2015)	Algoritmo EM	Múltiplo	Dados diários de MP10	2004	São Paulo, Brasil
Zainuri, Jemain e Muda (2015)	Algoritmo EM	Simples	Dados horários de ozônio	3 meses	Malásia

Fonte: autoria própria

Dentre as cidades do estudo, Campinas foi a que apresentou maior percentual de dados faltantes (2%). Jundiaí teve 1,8% e Paulínia, 1,5% de dados faltantes. Em Paulínia, o maior *gap* foi de 13 dias, enquanto que Jundiaí o maior *gap* foi de 8 dias. Campinas foi a que teve o maior *gap* (18 dias sem dados). O método proposto e os dois outros métodos escolhidos para comparação não exigem nenhum cálculo complexo. A ideia foi comparar apenas métodos simples de preenchimento.

Na tabela 2 estão os valores do EQM (equação 5), COR (equação 6) e DMA (equação 7) calculados para os 3 métodos (equações 2, 3 e 4). O método que apresentou menor EQM e DMA, e maiores valores de COR foi o TDEM, proposto neste trabalho. Em seguida, ficou o método “Média durante o mês”, e por último

o método “Média durante o ano”. Os resultados foram similares para as três cidades abordadas.

O método TDEM apresentou correlação maior que 0,60 em todas as cidades. Os outros dois métodos tiveram valores sempre abaixo deste, sendo o maior obtido na cidade de Paulínia, no método “Média durante mês”. Em relação aos valores de EQM, por exemplo, para a cidade de Paulínia, o método TDEM apresentou valor de 167,8, enquanto que no método “Média durante ano”, o valor foi de 257,6. Também houve uma diferença nos valores do DMA. Para o método do TDEM, o valor obtido foi de 9,88, para o método da equação 3 foi de 10,15, e para o método da equação 4 foi de 12,51.

Tabela 2 – Valores do Erro Quadrático Médio (EQM), Coeficiente de Correlação (COR) e Desvio Médio Absoluto (DMA).

Cidade	Método	EQM	COR	DMA
Campinas	TDEM	95,66	0,64	7,43
	Média durante mês	107,03	0,55	7,64
	Média durante ano	139,72	0,30	9,02
Jundiaí	TDEM	165,35	0,60	9,55
	Média durante mês	193,42	0,48	10,09
	Média durante ano	230,76	0,28	11,36
Paulínia	TDEM	167,80	0,66	9,88
	Média durante mês	187,99	0,57	10,15
	Média durante ano	257,59	0,27	12,51

Fonte: autoria própria

De modo geral, o método TDEM (equação 2) mostrou-se como uma alternativa melhor que os outros métodos avaliados, no preenchimento simples de dados faltantes. É um método fácil de utilizar, baseado apenas em médias amostrais. Nele, leva-se em consideração o efeito secundário de 3 fatores do tempo: o ano, semana do mês e dia da semana, além do fator principal que é o mês.

Esta metodologia pode também ser flexibilizada, escolhendo-se outros fatores para serem considerados como efeitos. Por exemplo, poderíamos ter escolhido o fator final de semana. Neste caso, o dia seria ou final de semana ou durante a semana. O fator final de semana poderia ser um substituto para o fator dia da semana. Ou mesmo o fator principal poderia ser substituído. Ao invés de utilizar o mês, poderiam ser as estações do ano.

Por fim, a série temporal dos dados de concentração do MP10 foi preenchida utilizando o método TDEM (equação 2), para as três cidades. Com a série preenchida ajustou-se o modelo para a série diária e para a série mensal.

Identificação dos modelos de séries temporais

Foram identificados diversos modelos ARIMA e SARIMA a fim de comparar qual obteve melhores resultados. Para a escolha dos modelos foram analisados os três critérios em conjunto (AIC, LOG e EQMP), além de uma análise qualitativa dos

gráficos dos ajustes. Para a análise da série diária do MP10 foi decidido realizar as previsões das concentrações diárias do mês de janeiro de 2015. Para a análise da série mensal do MP10, considerou-se a concentrações mensais do ano de 2015 para avaliar o poder de previsão do modelo.

Nas tabelas 3 e 4, constam os resultados dos critérios de seleção obtidos dos diferentes modelos ARIMA (A) e SARIMA (S) das séries temporais utilizando-se dos dados *diários* dos anos de 2010 a 2014. Os valores de p, q, d, P, Q, e D foram escolhidos pela análise dos gráficos de fac e facp.

O modelo S(1,0,2)x(1,0,2)7, que foi o escolhido para a cidade de Campinas está destacado na tabela 3. Para as cidades de Jundiaí e Paulínia, conforme destaque da tabela 4, o modelo escolhido foi o S(2,0,2)x(2,0,2)7.

O modelo S(1,0,2)x(1,0,2)7 foi escolhido para a cidade de Campinas pois apresentou o menor valor de EQMP. Nota-se que, entre os menores valores de EQMP (0,23 ou 0,24), foi o menor AIC obtido (12,43). Analisando todos os modelos, apenas pelo modelo S(1,1,1)x(1,1,1)7 apresentou um AIC inferior, porém este apresentou o maior valor de EQMP (0,29).

Para as cidades de Jundiaí e Paulínia, o modelo S(2,0,2)x(2,0,2)7 apresentou o menor valor de AIC e o maior valor de LOG. Percebe-se ainda que os valores de EQMP ficaram próximos entre os modelos escolhidos nessas duas cidades.

Tabela 3 – Critérios de seleção dos modelos utilizando-se os dados diários de MP10 para a cidade de Campinas.

	Campinas		
	AIC	**LOG**	**EQMP**
A(1,0,2)x(0,0,0)0	192,29	-91,14	0,24
S(1,0,2)x(1,0,2)7	12,43	1,78	0,23
S(1,0,1)x(1,0,1)7	31,88	-9,94	0,24
S(2,0,2)x(2,0,2)7	12,68	3,66	0,24
A(2,0,2)x(0,0,0)0	160,75	-74,37	0,27
S(1,1,1)x(1,1,1)7	-1,04	5,52	0,29

Fonte: autoria própria

Tabela 4 – Critérios de seleção dos modelos utilizando-se os dados diários de MP10 para as cidades de Jundiaí e Paulínia.

	Jundiaí			Paulínia		
	AIC	**LOG**	**EQMP**	**AIC**	**LOG**	**EQMP**
A(1,0,2)x(0,0,0)0	1399,32	-694,66	0,40	1073,89	-531,94	0,34
S(1,0,2)x(1,0,2)7	1355,11	-669,56	0,40	981,40	-482,70	0,32
S(1,0,1)x(1,0,1)7	1352,31	-670,16	0,40	982,94	-485,47	0,32
S(2,0,2)x(2,0,2)7	1321,60	-650,80	0,42	947,20	-463,60	0,33
A(2,0,2)x(0,0,0)0	1369,29	-678,65	0,42	1029,69	-508,85	0,32
S(1,1,1)x(1,1,1)7	1358,34	-674,17	0,43	964,48	-477,24	0,36

Fonte: autoria própria

De modo geral percebe-se que os modelos SARIMA se ajustaram melhor aos dados do que os modelos ARIMA. Isto reflete que a sazonalidade semanal é um fator importante na concentração do MP10.

Já nas tabelas 5 e 6, constam os resultados dos critérios de seleção obtidos por diferentes modelos ARIMA (A) e SARIMA (S) das séries temporais utilizando-se dos dados *mensais* dos anos de 2010 a 2014.

Com as séries mensais, o modelo S(1,0,1)x(1,0,1)12 foi o melhor. Para as 3 cidades, esse modelo apresentou o menor valor de AIC (-53,53, Campinas; -10,81, Jundiaí; e -23,51, Paulínia). Os valores de EQMP em Campinas e Jundiaí foram muito próximos em todos os modelos. Já em Paulínia, o modelo escolhido apresentou o menor valor de EQMP (0,14). Nas 3 cidades, o modelo S(1,0,1)x(1,0,1)12 apresentou o segundo maior valor de LOG, sempre muito próximo do maior valor entre os modelos.

Tabela 5 – Critérios de seleção utilizando-se os dados mensais de MP10 para a cidade de Campinas.

	Campinas		
	AIC	LOG	EQMP
S(1,0,2)x(1,0,2)12	-52,37	34,19	0,12
S(1,0,1)x(1,0,1)12	-53,53	32,76	0,12
S(1,1,2)x(1,0,2)12	-50,11	32,06	0,12
S(1,1,1)x(1,0,1)12	-52,07	31,03	0,12

Fonte: autoria própria

Não foi ajustado nenhum modelo ARIMA para as séries mensais. A sazonalidade é evidente, apresentando os maiores valores de concentração do MP10 sempre nos meses de inverno (julho a setembro).

Após a escolha dos modelos, os ajustes diários e os ajustes mensais foram analisados separadamente, obtendo os resultados descritos nas próximas duas subseções.

Tabela 6 – Critérios de seleção utilizando-se os dados mensais de MP10 para as cidades de Jundiaí e Paulínia.

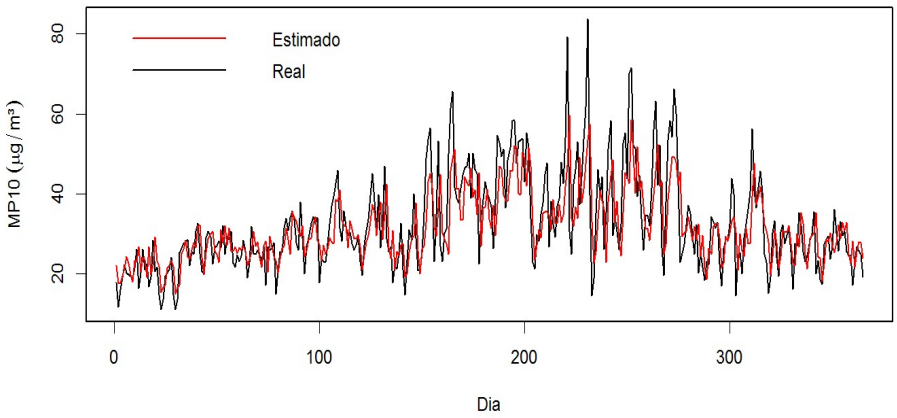
	Jundiaí			Paulínia		
	AIC	LOG	EQMP	AIC	LOG	EQMP
S(1,0,2)x(1,0,2)12	-6,86	11,43	0,18	-20,83	18,44	0,15
S(1,0,1)x(1,0,1)12	-10,81	11,41	0,18	-23,51	17,75	0,14
S(1,1,2)x(1,0,2)12	-5,54	9,77	0,18	-19,43	16,71	0,15
S(1,1,1)x(1,0,1)12	-9,15	9,57	0,18	-22,41	16,20	0,14

Fonte: autoria própria

Análise das séries temporais diárias

O gráfico apresentado na figura 2 representa, a série temporal (vermelho) utilizando o modelo S(1,0,2)x(1,0,2)7, e os dados diários (preto), obtidos pela estação de monitoramento de Campinas. Nas figuras 3 (Jundiaí) e 4 (Paulínia) representam a série temporal (vermelho) utilizando o modelo S(2,0,2)x(2,0,2)7, e os dados diários obtidos (preto) pela estação de monitoramento localizada em cada cidade. Foi apresentado apenas a série ajustada para o ano de 2011, para exemplificar como ficou o ajuste. Nos demais anos, o comportamento da série foi similar.

Figura 2 – Série ajustada do ano de 2011, modelo S(1,0,2)x(1,0,2)7, utilizando os dados diários da cidade de Campinas.



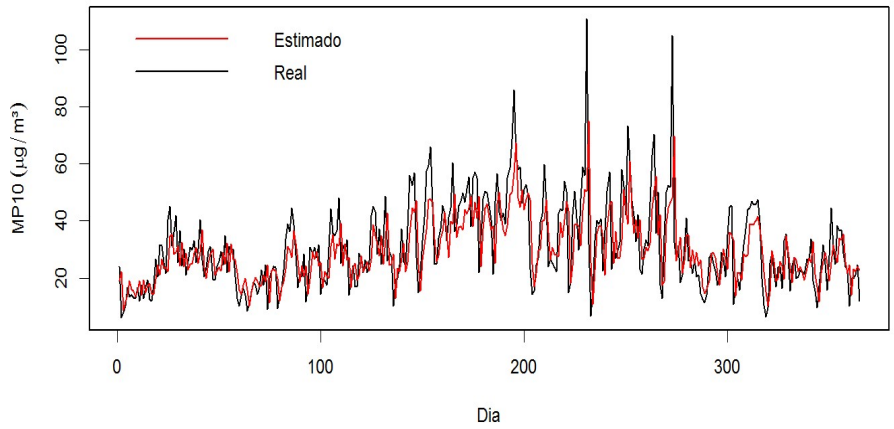
Fonte: autoria própria

De modo geral, a série estimada consegue acompanhar a série dos dados reais. Nos dias de maiores picos de concentração, os dados estimados ajustam o pico, porém não atingem o valor. A mesma análise pode ser feita para os dias com baixa concentração de MP.

Outro fator que pode ser observado é em relação à defasagem do ajuste. Nota-se que a série estimada parece estar ajustando o valor do dia presente, apenas no dia posterior. Este é um indício que a série pode ter o que é chamado de processo de memória longa, sendo indicado o ajuste do modelo ARFIMA (Auto Regressivo Fracionário Integrado de Médias Móveis), ou o ARFIMA sazonal (SARFIMA). Estas séries com memória longa, apresentam dependência significativa entre observações separadas por um longo intervalo de tempo (MORETTIN e TOLOI, 2006; REISEN et al. 2014).

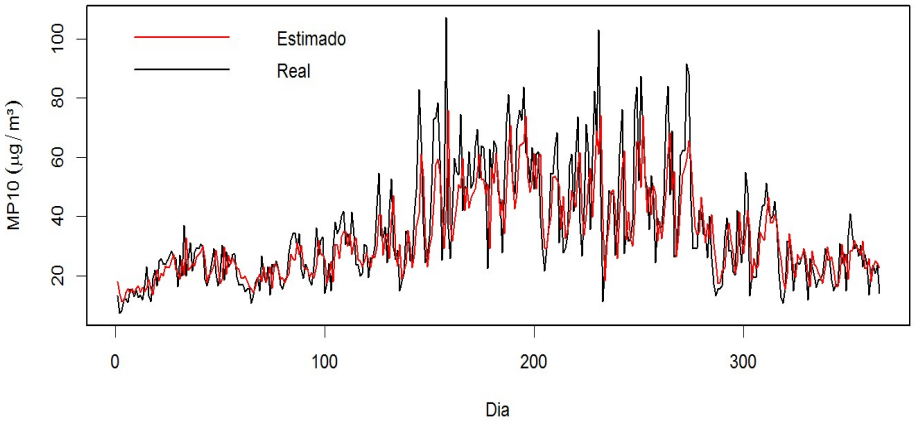
O período escolhido para previsão foi o mês de janeiro de 2015 (figuras 5, 6 e 7). Nestas figuras, os valores reais estão na linha preta, os valores ajustados na linha vermelha e os limites (inferior e superior) de previsão estão em azul. Percebe-se que os valores reais ficaram na sua maioria todos dentro dos limites de previsão de 95% obtidos pelo modelo ajustado, indicando um ajuste satisfatório.

Figura 3 – Série ajustada do ano de 2011, modelo S(2,0,2)x(2,0,2)7, utilizando os dados diários para a cidade de Jundiaí.



Fonte: autoria própria

Figura 4 – Série ajustada do ano de 2011, modelo S(2,0,2)x(2,0,2)7, utilizando os dados diários para a cidade de Paulínia.

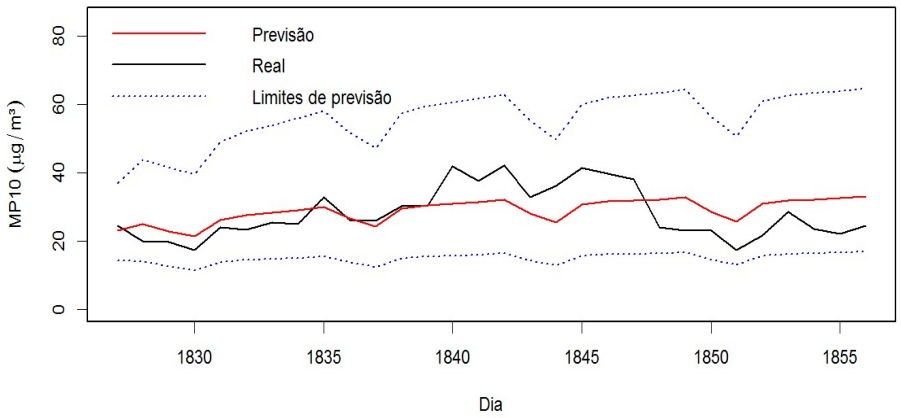


Fonte: autoria própria

Nas figuras 6 e 7, observa-se que a série prevista para o mês de janeiro de 2015, obtida do modelo SARIMA não conseguiu acompanhar o comportamento dos dados da estação de monitoramento em Jundiaí e Paulínia, respectivamente. Percebeu-se que a série real teve um aumento, ficando acima de 30, entretanto a previsão apontou concentração sempre próxima de 20 µg/m³, ao longo do mês.

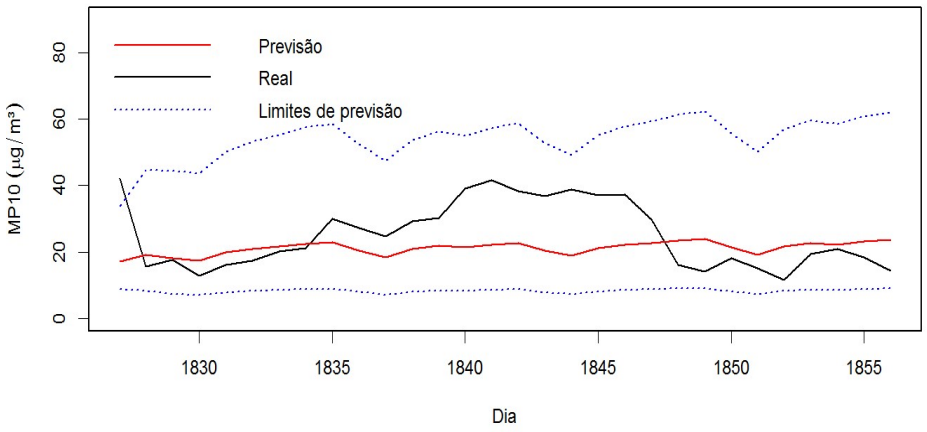
Para a cidade de Jundiaí, perceba que o primeiro dia ficou fora dos limites de previsão de 95% (figura 6). Entretanto, os demais pontos ficaram dentro dos limites.

Figura 5 –Previsão da concentração diária no mês de Janeiro de 2015 para a cidade de Campinas, utilizando o modelo S(1,0,2)x(1,0,2)7.



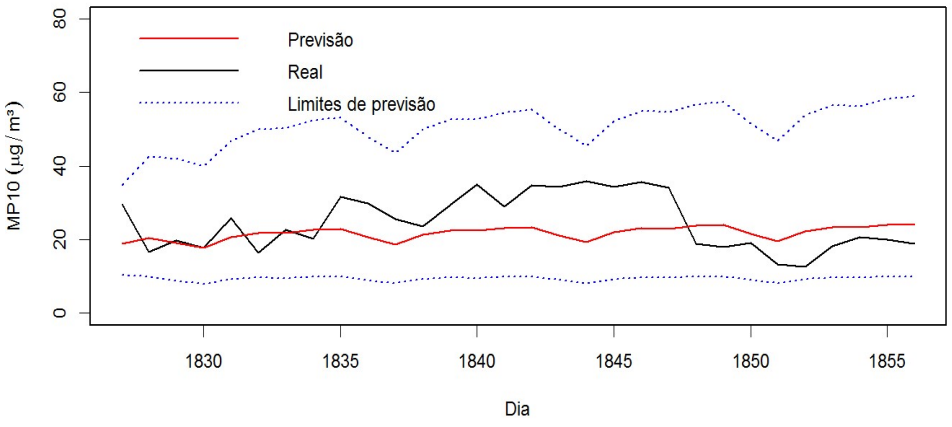
Fonte: autoria própria

Figura 6 – Previsão da concentração diária no mês de Janeiro de 2015 para a cidade de Jundiaí, utilizando o modelo S(2,0,2)x(2,0,2)7.



Fonte: autoria própria

Figura 7 – Previsão da concentração diária no mês de Janeiro de 2015 para a cidade de Paulínia, utilizando o modelo S(2,0,2)x(2,0,2)7.



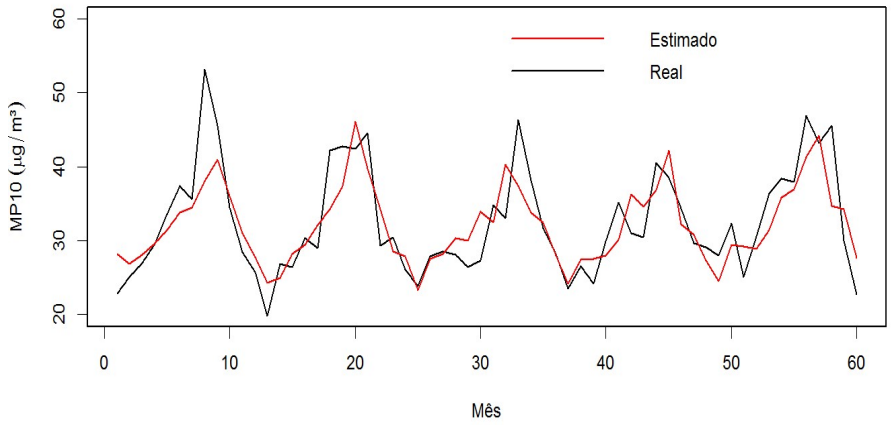
Fonte: autoria própria

Análise das séries temporais mensais

A figura 8 representa a série temporal (vermelho) utilizando o modelo S(1,0,1)x(1,0,1)12, e os dados mensais reais (preto), obtidos pela estação de monitoramento de Campinas.

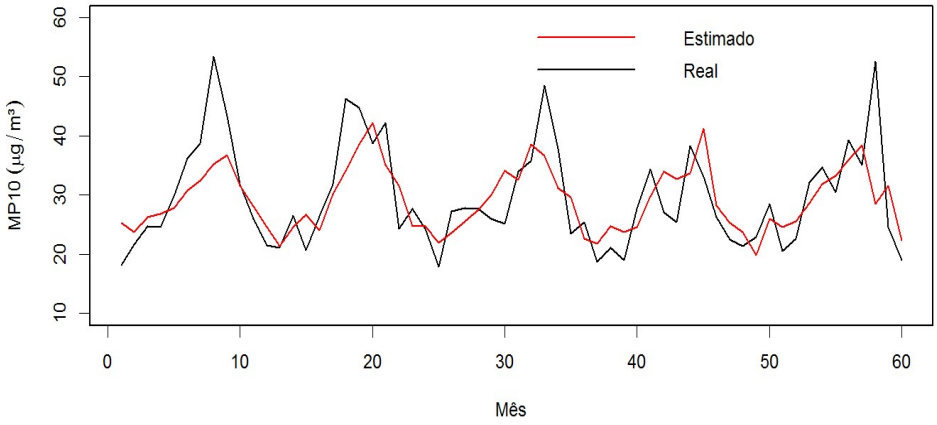
Os gráficos apresentados nas figuras 9 e 10 representam a série temporal ajustada pelo modelo S(1,0,1)x(1,0,1)12, nas cidades de Jundiaí e Paulínia, respectivamente.

Figura 8 – Série ajustada no período de 2010 a 2014, no modelo S(1,0,1)x(1,0,1)12, dados mensais para a cidade de Campinas.



Fonte: autoria própria

Figura 9 – Série ajustada no período de 2010 a 2014, no modelo S(1,0,1)x(1,0,1)12, dados mensais para a cidade de Jundiaí.



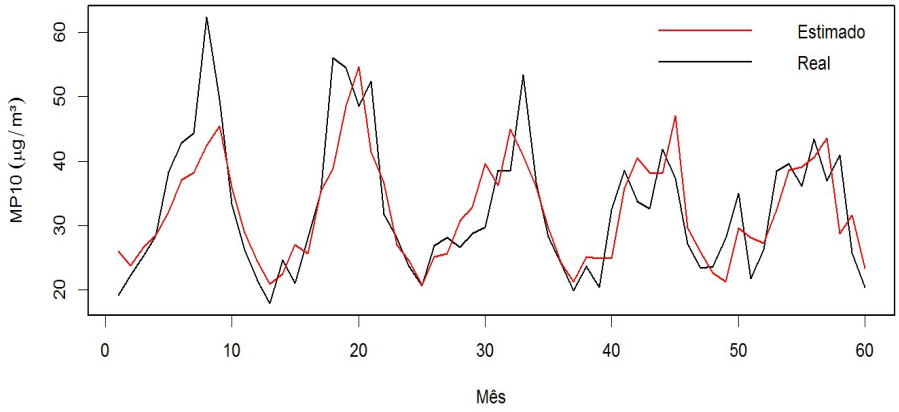
Fonte: autoria própria

Nas três cidades, os picos de concentração foram sempre nos meses de inverno (julho a setembro), caracterizando muito bem a sazonalidade mensal. A série estimada dos dados consegue capturar essa sazonalidade, apontando os

picos de concentração do MP10 nessa estação do ano. Entretanto, a série ajustada não conseguiu prever o valor da concentração no pico.

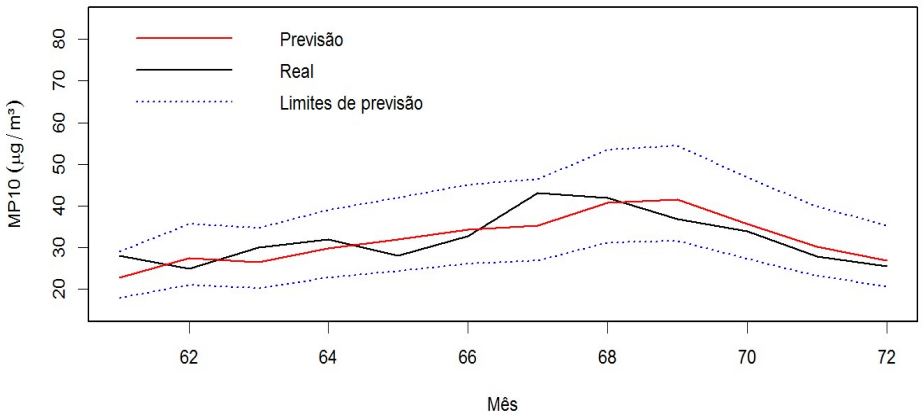
Em relação a previsão mensal do ano 2015 (figuras 11, 12 e 13), o comportamento da série obtida ficou semelhante aos dados reais. Todos os valores reais ficaram dentro dos limites de previsão, e o ajuste previsto conseguiu acompanhar a série real. Percebe-se que os valores reais se encontram dentro dos limites de previsão de 95%.

Figura 10 – Série ajustada no período de 2010 a 2014, no modelo S(1,0,1)x(1,0,1)12, dados mensais para a cidade de Paulínia.



Fonte: autoria própria

Figura 11 – Previsão da concentração mensal do ano de 2015 para a cidade de Campinas, no modelo S(1,0,1)x(1,0,1)12.

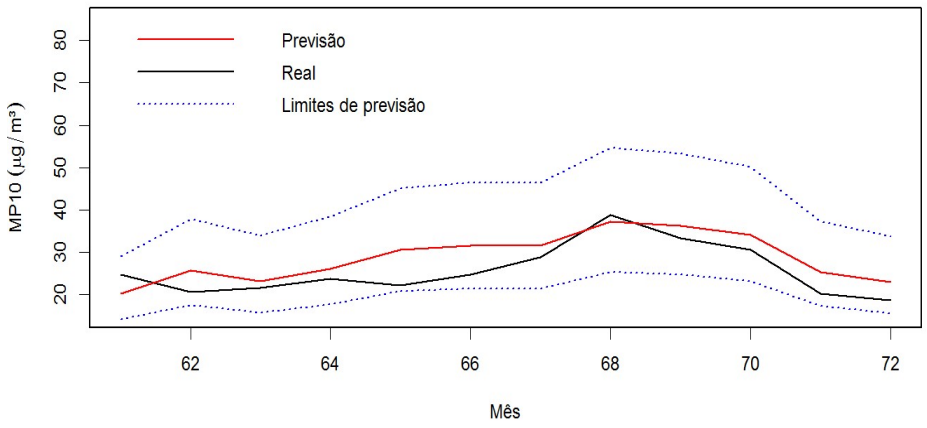


Fonte: autoria própria

Por fim, os modelos SARIMA conseguiram obter resultados com um apropriado desempenho de previsão para as séries mensais dos dados, ou seja, a série dos dados reais ficaram dentro dos limites de previsão de 95%. Os resultados de previsão desses modelos foram melhores que os ARIMA, como notado nos critérios de seleção (tabelas 3, 4, 5 e 6). Aparentemente, as séries diárias de

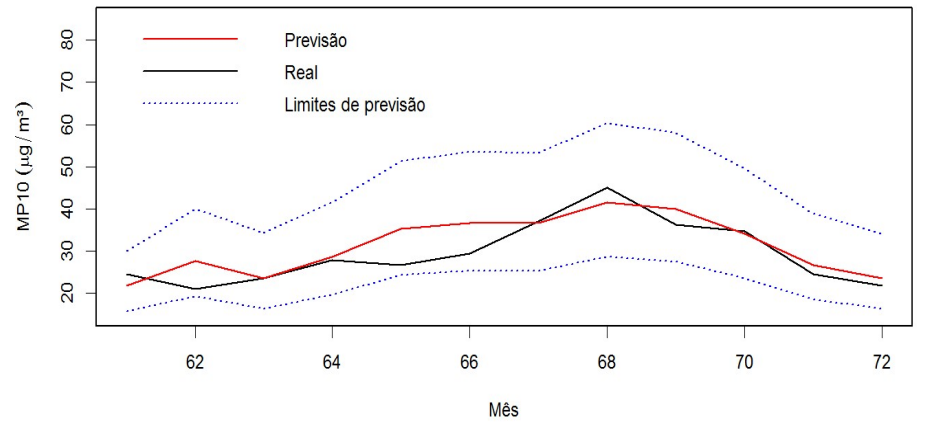
concentração do MP10 apresentam a característica de processo de memória longa. Neste caso, modelos SARFIMA poderiam melhorar os resultados de previsão da série diária.

Figura 12 – Previsão da concentração mensal do ano de 2015para a cidade de Jundiaí, no modelo S(1,0,1)x(1,0,1)12



Fonte: autoria própria

Figura 13 – Previsão da concentração mensal do ano de 2015para a cidade de Paulínia, no modelo S(1,0,1)x(1,0,1)12.



Fonte: autoria própria

CONSIDERAÇÕES FINAIS

O método TDEM foi o que apresentou melhores resultados para preencher os dados faltantes das séries temporais da concentração diária do MP10. Os resultados foram similares nas 3 cidades analisadas, de acordo com os critérios analisados.

Entre suas propriedades, o método consegue captar o comportamento da série, além de apresentar flexibilidade na opção dos fatores, ou seja, os fatores

temporais podem ser escolhidos de acordo com o estudo do pesquisador. No presente estudo, foram considerados como fatores do tempo, o mês (efeito principal), ano, semana do mês e dia da semana (efeitos secundários).

O método TDEM é uma alternativa para séries de dados coletadas em uma única estação de monitoramento, porém em diversos anos. Este método utiliza apenas valores da própria série para completar as lacunas. Diferentemente do método SDEM, que utiliza medições feitas em outras estações de monitoramento, nem sempre disponíveis na cidade.

Destaca-se também, que o método SDEM foi testado por Plaia e Bondi (2006) com altos percentuais de dados faltantes (chegando a 13% numa das estações de monitoramento) e obteve ótimos resultados. Em nosso estudo, não tivemos esse percentual em nenhuma das cidades (máximo de 2% em Campinas), mas também obtivemos resultados satisfatórios para o método TDEM. Isso nos leva a concluir que temos um método robusto para preenchimento dos dados ausentes. Além disso, devido a flexibilidade do método TDEM, é possível inclusive optar por outros fatores, sejam relacionados ao tempo, ou mesmo ao espaço. Pesquisas futuras podem ser conduzidas para avaliar esses outros fatores.

A série completa dos dados permite realizar diversas outras análises. Algumas, como a metodologia de Box-Jenkins usada aqui, exigem que a série esteja completa para sua realização. Então, obter um método de preenchimento de dados faltantes foi muito importante, e nos permitiu fazer os ajustes das séries temporais.

Os modelos ARIMA e SARIMA apresentaram um desempenho adequado no ajuste dos dados da concentração do MP10 nas cidades de Campinas, Jundiaí e Paulínia, para o período estudado. De modo geral, os modelos SARIMA apresentaram resultados mais satisfatórios quando comparados aos modelos ARIMA. Há evidências que as séries temporais diárias do MP10 têm comportamento de memória longa (REISEN et al., 2014), conforme observado. Neste caso é indicado um ajuste pelo modelo SARFIMA, sendo esta uma das possibilidades para trabalhos futuros.

A new method of missing data imputation applied to time series of PM10 concentration

ABSTRACT

The study of atmospheric pollution, with the emphasis of inhalable particulate matter (PM10), is necessary; given the damage was done for population health, besides other losses. Historical series, used for forecasting data, often have gaps due to several factors, which can detract from the quality of the forecast. The aim of this study was purposed a new method of missing data imputation, and after this, to use a time series model to forecast PM10 concentration. It was obtained the PM10 daily concentration in QUALAR system of CETESB, related to cities of Campinas, Jundiaí and Paulínia, all in São Paulo State. The method of imputation of missing data, purpose in this study, was called TDEM (Time-Dependent Effect Method). The TDEM method was compared to others two methods ("Mean during month" and "Mean during year") of imputation of missing data, and it presented better results related to correlation coefficient, mean square error and mean absolute deviation. After imputation data of series, the data were analysed in order to forecast future PM10 concentrations. It was used ARIMA and SARIMA for time series models. The more satisfactory results were obtained for SARIMA models, which real data remained within the 95% forecast limits.

KEYWORDS: Missing data. Box-Jenkins Methodology. Particulate Matter. Forecast.

AGRADECIMENTOS

Os autores gostariam de agradecer a CAPES e o CNPq pela concessão de bolsas de doutorado e de iniciação científica.

REFERÊNCIAS

ALVES, C. Aerossóis atmosféricos: perspectiva histórica, fontes, processos químicos de formação e composição orgânica. **Química Nova**, v. 28, p. 859-870, 2005.

BELL, M. L.; SAMET, J. M.; DOMINICE, F. Time-Series studies of particulate matter. **Annual Review of Public Health**, v.25, p. 247–80, 2004.

CEPAGRI (Centro de Pesquisas Meteorológicas e Climáticas Aplicadas a Agricultura). UNICAMP - Campinas, São Paulo. Disponível em: <http://www.cpa.unicamp.br/outras-informacoes/clima-dos-municipios-paulistas.html>. Acesso em: 26/01/2018.

CETESB (Companhia Ambiental do Estado de São Paulo). Qualidade do ar - Poluentes. São Paulo. Disponível em:<http://cetesb.sp.gov.br/ar/poluentes/>. Acesso em: 12/01/2018.

D'AMÉLIO, M. T. S., CAMPOS, L. C. L., ALVIM, D. S. Estudo da variabilidade do monóxido de carbono atmosférico na região metropolitana de Campinas-SP e comparação com São Paulo-SP. Ensaios USF, v. 1, n. 1, p. 80-90, 2017.

EHLERS, R.S. Análise de Séries Temporais, 2009. Disponível em: <http://www.icmc.usp.br/ ehlers/stemp/stemp.pdf>. Acesso em: 12/01/2018.

GÓMEZ-CARRACEDO, M. P.; ANDRADE. J. M.; LÓPEZ-MAHÍA, P.; MUNIATEGUI, S.; PRADA, D. A practical comparison of single and multiple imputation to handle complex missing data in air quality datasets. **Chemometrics and Intelligent Laboratory Systems**, v. 134, p. 23-33, 2014.

GOYAL, P., CHAN, A. T., JAISWAL, N. Statistical models for the prediction of respirable suspended particulate matter in urban areas. **Atmospheric Environment**, v. 40, p. 2068-2077, 2006.

GUJARATI, D. N.; PORTER, D. C. **Econometria básica**. 5. ed. Porto Alegre: AMGH Ed., 2012. 924p.

IBGE, Cidades – Panorama. Disponível em: <https://cidades.ibge.gov.br/brasil/sp/jundiai/panorama> Acesso em: 29/10/2018.

JUNGER, W. L.; PONCE DE LEON, A. Imputation of missing date in series for air pollution. **Atmospheric Environmental**, v. 102, p. 96-104, 2015.

LIMA, E. P., POZZA, S. A., GIMENES, M. L., COURY, J. R. Uso de modelos ARIMA Sazonais no estudo da série temporal de MP10 da cidade de São Carlos. In: **XXXIV Congresso Brasileiro de Sistemas Particulados (ENEMP)**. Anais... São Paulo: Campinas, Out. 2009.

MONTE, E. Z.; ALBUQUERQUE, T. T. A.; REISEN, V. A. Previsão da concentração de ozônio na região da grande Vitória, Espírito Santo, Brasil, utilizando o modelo ARMAX-GARCH. **Revista Brasileira de Meteorologia**, v. 30, n. 3, p. 285-294, 2015.

MORETTIN, P. A., TOLOI, C. M. C. **Análise de séries temporais**. 2ª. ed. Rio de Janeiro: Blucher. 2006, 538p.

NASCIMENTO, L. F. C.; PEREIRA, L. A. A.; BRAGA, A. L. F.; MÓDOLO, M. C. C.; CARVALHO, J. A. J. Efeitos da poluição atmosférica na saúde infantil em São José dos Campos, SP. **Revista de Saúde Pública**, v. 40, p. 77-82, 2006.

PINTO, W. P., REISEN, V. A., MONTE, E. Z. Previsão da concentração de material particulado inalável, na Região da Grande VItória, ES, Brasil, utilizando o modelo SARIMAX. **Engenharia Sanitária e Ambiental**, v. 23, n. 2, p. 307-318, 2018

PLAIA, A.; BONDI A. L. Single imputation method of missing values in environmental pollution data sets. **Atmospheric Environment**, v. 40, p. 7316-7330, 2006.

POZZA, S. A., LIMA, E. P., COMIN, T. T., GIMENES, M. L., COURY, J. R. Time series analysis of PM2.5 and PM10-2.5 mass concentration in the city of São Carlos, Brazil. **International Journal of Environment and Pollution**, v. 41, p. 90-108, 2010.

QUALAR - CETESB: banco de dados. Disponível em: <http://ar.cetesb.sp.gov.br/qualar/>. Acesso em: 15/04/2016.

RAHMAN, A., HOSSAIN, A. Time Series Analysis Model for Particulate Matter of Air Pollution Data of Dakha City. **Asian Journal of Water, Environment and Pollution**, v. 9, p. 63-69, 2012.

R CORE TEAM (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Disponível em:<https://www.R-project.org/>.

REISEN, V. A., SARNAGLIA, A. J. Q., REIS Jr, N. C., LÉVY-LEDUC, C., SANTOS, J. M. Modeling and forecasting daily average PM10 concentrations by a seasonal long-memory model with volatility. **Environmental Modelling & Software**, v. 51, p. 286-295, 2014.

VALLERO, D. **Fundamentals of Air Pollution**. 4ª. ed. Amsterdan: Academic Press/Elsevier. 2006, 942p.

XIA, Y.; FABIAN, P.; STOHL, A.; WINTERHALTER, M. Forest climatology: estimation of missing values for Bavaria. **Agricultural and Forest Meteorology**, v. 96, p. 131-144, 1999.

ZAINURI, N. A.; JEMAIN, A. A.; MUDA, N. A Comparison of Various Imputation Methods for Missing Values in Air Quality Data. **Sains Malaysiana**, v. 43, p. 449-456, 2015.

Recebido: 23 jul. 2018.
Aprovado: 21 nov. 2018.
DOI: 10.3895/rts.v15n37.8594
Como citar: NOGAROTTO, D. C. et al. Um novo método de preenchimento de dados faltantes aplicado a séries temporais de concentração de MP10. **R. Tecnol. Soc.,** Curitiba, v. 15, n. 37, p. 275-296, jul./set. 2019. Disponível em: <https://periodicos.utfpr.edu.br/rts/article/view/8594>. Acesso em: XXX.
Correspondência: