

APLICAÇÃO DE MODELO ARIMA NA PREVISÃO DE ÓBITOS POR DOENÇAS DE VEICULAÇÃO HÍDRICA NO ESTADO DO CEARÁ

Douglas Cavalcante Holanda¹

douglas.contato.holanda@gmail.com

Área Temática: Temas Livres em Saúde.

RESUMO

Introdução: Os constantes avanços das políticas públicas de saneamento básico, em conjunto com as técnicas de higiene, vêm contribuindo para uma tendência de diminuição dos óbitos decorrentes de doenças de veiculação hídrica, porém nem todas as regiões do Brasil possuem acesso às infraestruturas de saneamento básico gerando um quadro de necessidade de maiores investimentos na saúde pública, visando o controle epidemiológico e o tratamento referente às doenças de veiculação hídrica. **Objetivo:** Partindo deste pressuposto a presente pesquisa teve como objetivo realizar previsões acerca dos óbitos relacionados a ausência de infraestruturas de saneamento básico e a insegurança hídrica. **Metodologia:** Esta pesquisa teve como aporte metodológico o uso do *software* estatístico RStudio e de seus pacotes, com destaque ao pacote *forecast*, no qual utilizou-se do modelo *Autoregressive Integrated Moving Average* (ARIMA), sendo o modelo ARIMA bastante utilizado em pesquisas que envolvem a previsão baseada em séries temporais univariadas. Os dados utilizados foram extraídos a partir do banco de dados do Sistema IBGE de Recuperação Automática (SIDRA). **Resultados e Discussão:** O modelo ARIMA (0, 1, 1) aplicado a um horizonte de previsão de quatro anos, produziu uma previsão realista acerca dos óbitos atribuídos a doenças de veiculação hídrica no estado do Ceará, com uma visível tendência de redução no número de óbitos. **Conclusão:** Conclui-se que a previsão estatística pode ser bastante útil para o setor público, uma vez que a mesma permite avaliar possíveis melhorias e investimentos na saúde pública, visando uma melhor administração e controle sobre tais eventos. Vale ressaltar, porém, que o modelo utilizado foi aplicado em uma série temporal univariada, ou seja, na prática os diversos fenômenos naturais e antrópicos aleatórios não previstos pelo modelo ARIMA podem induzir ao aumento ou a redução no número de óbitos para o estado do Ceará.

Palavras-chave: Saneamento Básico; ARIMA; Saúde Pública.

1 INTRODUÇÃO

Os constantes avanços nas noções de higiene, bem como a instalação e melhoria das infraestruturas de saneamento básico promoveram reduções significativas sobre o número de óbitos decorrentes de doenças de veiculação hídrica, evidenciando a importância e o impacto que tais investimentos têm sobre a qualidade de vida da população (JUNIOR e MALHEIROS, 2018). Porém ainda existem muitas áreas sem acesso a tais infraestruturas, gerando um quadro de fragilidade e insegurança hídrica uma vez que a falta de acesso ao saneamento básico tende a favorecer uma série de impactos sociais, ambientais e econômicos, afetando negativamente a população e o ecossistema (MOTA, 2018).

O Sistema IBGE de Recuperação Automática (SIDRA), dispõe de uma vasta e ampla coleção de dados dispostos em séries temporais aplicadas a diferentes escalas geográficas. Dentre os dados destaca-se a série temporal para o número de óbitos relacionados a ausência de saneamento básico, insegurança hídrica e a falta de higiene, que são fatores que favorecem as doenças de veiculação hídrica. Porém vale ressaltar que boa parte dos dados do SIDRA, não contemplam os últimos anos (2024, 2025 e 2026), favorecendo o surgimento de lacunas nos registros. Neste sentido, dados importantes relacionados ao gerenciamento do saneamento básico e da saúde pública ficam sem atualizações, gerando um cenário de imprevisibilidade e de incertezas epidemiológicas que afetam sobretudo a população.

Segundo Sato (2013) e Cardoso (2018), dados provenientes de séries temporais podem ser utilizados para gerar previsões acerca de possíveis aumentos ou reduções nos números de contaminações e de óbitos, sendo úteis para a gestão pública, sobretudo no que diz respeito à eficiência no gerenciamento dos investimentos em saúde pública. Neste sentido Sato (2013), destaca que um dos modelos mais utilizados na área da saúde para gerar previsões baseados em dados de séries temporais é o modelo do tipo *Autoregressive Integrated Moving Average* (ARIMA). O modelo do tipo ARIMA foi proposto por George Box e Gwilym Jenkins (BOX e JENKINS, 1976), sendo uma versão mais robusta do modelo do tipo *Autoregressive Moving Average* (ARMA).

Tendo como principal problemática a ausência de dados recentes no SIDRA para o número de óbitos relacionados a falta de higiene, bem como a ausência de saneamento básico e a insegurança hídrica no estado do Ceará, este trabalho tem como principal objetivo gerar previsões utilizando-se de dados provenientes dos Objetivos de Desenvolvimento Sustentável (ODS), sendo selecionada a tabela (8191) com o indicador (3.9.2), em que selecionou-se a variável “Óbitos atribuídos a fontes de água inseguras, saneamento inseguro e falta de higiene”. O conjunto de dados foi carregado e processado no *software* RStudio, por meio do pacote “*sidrar*” (SIQUEIRA, 2022). Após carregar os dados, elaborou-se previsões baseadas no modelo ARIMA, sendo utilizado para tal o pacote estatístico “*forecast*” (HYNDMAN e KHANDAKAR, 2008).

2 METODOLOGIA

Para a elaboração deste trabalho, primeiramente buscou-se pelos dados no SIDRA, em que selecionou-se a tabela (8191) do indicador (3.9.2) dos ODS, que basicamente trata da taxa de mortalidade atribuída à fonte de águas inseguras, saneamento inseguro e falta de higiene, em que trabalhou-se com a variável número de óbitos para se realizar a previsão. Como área de

estudo selecionou-se o estado do Ceará, buscando assim gerar um modelo específico para tal área, compreendendo portanto as nuances das taxas de óbitos neste estado. Vale ressaltar que tal delimitação na área de estudo pode ser benéfica para a compreensão dos diversos aspectos epidemiológicos das doenças de veiculação hídrica, que geralmente estão fortemente associadas a fatores como a insegurança hídrica, a ausência de saneamento básico, bem como a fatores geográficos e sociais.

Quanto às características demográficas dos dados, não realizou-se distinção por sexo ou faixa etária, neste sentido foi computado o número total de óbitos ocorridos entre os anos (2000 à 2022), sendo os dados da série temporal em questão, anuais. Após a seleção de todas as variáveis a serem consideradas para o estudo, os dados foram carregados no RStudio por meio do pacote “*sidrar*” (SIQUEIRA, 2022). Após carregados, realizou-se os procedimentos padrões, que visam a transformação dos dados para o formato numérico permitindo portanto realizar os testes estatísticos. Após converter os dados para o formato numérico, criou-se um objeto no formato *Temporal Series* (TS), sendo tal formato necessário para a aplicação do modelo de previsão no pacote “*forecast*” (HYNDMAN e KHANDAKAR, 2008).

Pelo fato dos dados serem anuais, não dá para discernir a influência da sazonalidade, portanto tal característica influenciou na seleção do modelo, tendo sido selecionado para tal o modelo do tipo ARIMA. Segundo Sato (2013), o modelo ARIMA pode ser dividido em três termos (p, d, q) sendo o termo (p) relacionado a parte autoregressiva (AR) do modelo, o termo (d) é o termo de diferenciações (I) necessárias para manter a série estacionária, já o termo (q) por sua vez representa a parte de média móvel (MA) do modelo, ressaltando-se porém que nem todo modelo vai possuir as três partes, sendo estas selecionadas em geral por meio da análise dos gráficos de função de autocorrelação e de autocorrelação parcial, denominados respectivamente de (FAC) e (FACP). Segundo Hyndman e Athanasopoulos (2021), o ARIMA completo pode ser expresso pela (Equação 1).

$$y_t = c + \phi_1 y_{t-1} + \dots + \phi_p y_{t-p} + \theta_1 \varepsilon_{t-1} + \dots + \theta_q \varepsilon_{t-q} + \varepsilon_t \quad (1)$$

Alguns testes podem ser realizados facilitando a escolha do termo integrado (I), que se refere ao número de diferenciações necessárias para tornar a série temporal estacionária, neste sentido, aplicou-se o teste de Dickey-Fuller Aumentado (DICKEY e FULLER, 1979), que é um teste de raiz unitária, em que a hipótese nula indica uma série não estacionária e a hipótese

alternativa indica uma série estacionária, o teste foi realizado a um nível de significância de 5% ($\alpha = 0.05$). O teste de Dickey-Fuller Aumentado foi aplicado a série temporal via pacote “*tseries*” (TRAPLETTI e HORNIK, 2024), sendo esta uma etapa necessária para descobrir se a série temporal é estacionária ou não estacionária, uma vez que os modelos do tipo ARIMA, tendem a gerar melhores previsões quando a série é estacionária, ou seja, quando a variância residual é constante ao longo do tempo, em outras palavras o modelo torna-se homocedástico.

Para além do teste de Dickey-Fuller Aumentado, optou-se também pela realização do teste de Kwiatkowski-Phillips-Schmidt-Shin (KPSS), que diferente do teste de Dickey-Fuller Aumentado possui como hipótese nula a estacionariedade, ou seja, se o *P-value* for menor que o nível de significância selecionado ($P < \alpha$), significa que a série pode ser considerada como não estacionária, sendo necessário aplicar diferenciação (KWIATKOWSKI et al. 1992). O resultado do teste KPSS foi utilizado neste estudo como complementar ao resultado do teste de Dickey-Fuller Aumentado garantindo assim o número preciso de diferenciações aplicadas evitando portanto um possível *overfitting* do modelo.

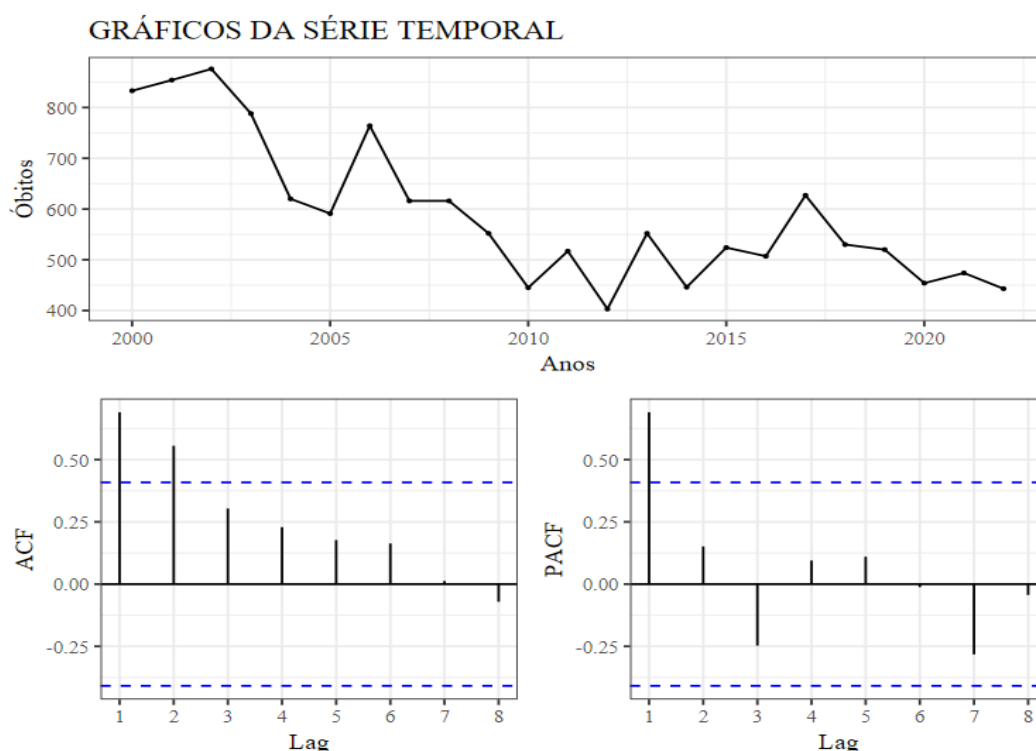
Para a seleção do modelo ARIMA aplicado à previsão, utilizou-se de alguns critérios que segundo Bruce e Bruce (2019), são comumente usados para selecionar o melhor modelo, destacando-se o Critério de Informação de Akaike (AIC), Critério de Informação Bayesiano (BIC), Critério de Akaike corrigido (AICc), para além destes utilizou-se também dos valores da Raiz Quadrada do Erro Quadrático Médio (RMSE) e do Erro Médio Absoluto (MAE). O Critério de Informação de Akaike foi desenvolvido pelo estatístico japonês Hirotugu Akaike (AKAIKE, 1973), tal métrica basicamente penaliza a adição de termos ao modelo, o Critério de Informação Bayesiano é semelhante ao AIC porém tende a ter maiores penalidades sendo portanto uma métrica mais criteriosa na seleção de modelos, o AICc por sua vez é comumente utilizado para modelos com amostras menores.

Para além das métricas anteriormente apresentadas, adotou-se o teste de Ljung-Box para a análise residual (LJUNG e BOX, 1978). O teste de Ljung-Box avalia a predominância de autocorrelação serial residual, ou seja, o teste avalia se os resíduos são independentes e normalmente distribuídos, sendo portanto denominados de ruídos brancos. Para o cálculo dos intervalos de previsão gerados pelo modelo, é necessário que os resíduos satisfaçam a duas propriedades, sendo elas a variância constante (homocedasticidade) e a distribuição normal residual, quando tais propriedades não são satisfeitas o modelo tende a gerar intervalos de previsão imprecisos, portanto segundo Hyndman e Athanasopoulos (2021), a análise residual se faz indispensável na construção do modelo de previsão.

3 RESULTADOS E DISCUSSÃO

Após carregar os dados sobre os óbitos relacionados à doenças de veiculação hídrica no estado do Ceará no RStudio via pacote “*sidrar*”, converteu-se as variáveis de estudo para o formato numérico permitindo assim transformar os dados do *dataframe* para o formato de *Temporal Series* (TS). Ao transformar os dados para o formato de série temporal extraiu-se os correlogramas com as funções de autocorrelação (FAC) e de autocorrelação parcial (FACP). A (Figura 1), contém a série temporal original com os correlogramas da FAC e da FACP.

Figura 1 – Plotagem da Série temporal com os correlogramas da FAC e da FACP.



Fonte: Autor (2026).

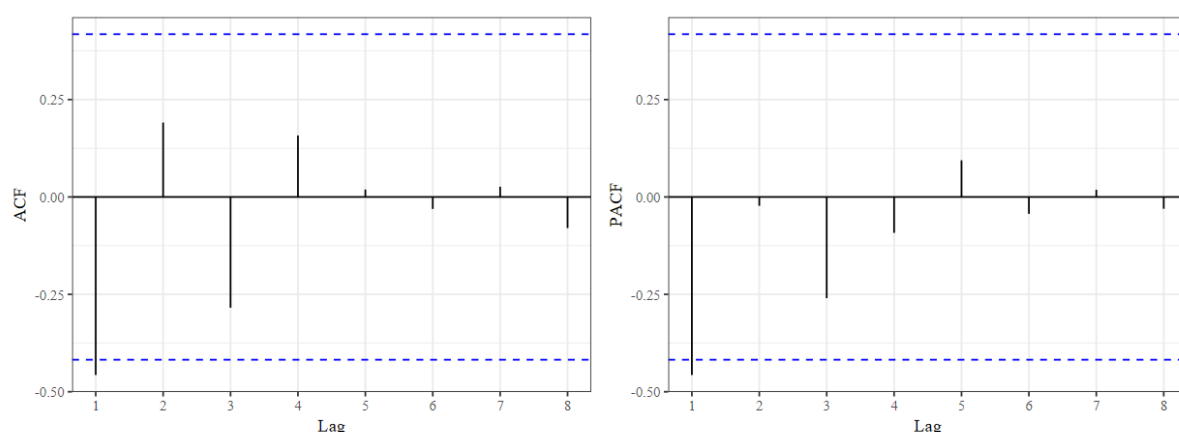
A série temporal abrange registros do período de (2000 à 2022), em que percebe-se uma tendência de redução ao longo dos anos, a série temporal em questão é anual, ou seja não apresenta o componente de sazonalidade. Analisando os correlogramas da FAC e da FACP, é possível notar alguns *lags* fora do intervalo de significância, sobretudo na FAC, indicando a necessidade de ajustes aos termos (p, d, q) do modelo ARIMA.

Iniciando pelo termo (d), analisou-se por meio dos testes de Dickey-Fuller Aumentado e do teste KPSS a necessidade de diferenciação para tornar a série estacionária. O resultado do teste de Dickey-Fuller Aumentado evidenciou a necessidade da diferenciação de primeira ordem ($d = 1$), com *P-value* igual a (0.39). Reaplicando o teste de Dickey-Fuller Aumentado à série já diferenciada obteve-se um *P-value* igual a (0.04), descartando a hipótese nula ao nível

de significância de 5% ($\alpha = 0.05$), indicando portanto que a série se tornou estacionária após a diferenciação de primeira ordem ($d = 1$).

Aplicando o teste KPSS evidenciou-se ao nível de significância de 5% ($\alpha = 0.05$) que a série temporal não é estacionária, aplicando-se portanto uma diferenciação de primeira ordem ($d = 1$). Reaplicando o teste KPSS na série já diferenciada, obteve-se *P-value* superior a (0.1) não rejeitando a hipótese nula de variância constante (homocedasticidade), neste sentido ficou evidente por meio dos testes aplicados, que é necessário uma diferenciação de primeira ordem ($d = 1$) para tornar a série temporal estacionária. Após tornar a série estacionária, analisou-se novamente os gráficos da FAC e FACP, estando os mesmos apresentados na (Figura 2).

Figura 2 – Correlogramas da FAC e da FACP após diferenciação de primeira ordem.



Fonte: Autor (2026).

Após a aplicação da diferenciação a série apresentou somente um *lag* fora do intervalo de confiança, a partir disso foi possível elaborar alguns modelos. Os modelos foram testados tanto com *drift* quanto sem *drift*, para o critério de seleção dos modelos utilizou-se do AIC, BIC, AICc, RMSE e MAE. Aplicando o *drift* percebeu-se uma redução geral da RMSE e do MAE para todas as combinações, além de melhores resultados no teste de Ljung-Box, neste sentido aplicou-se o *drift* à todas as combinações testadas. O *drift* pode ser definido segundo Hyndman e Athanasopoulos (2021), como a variação da média ao longo da série temporal.

A partir dos gráficos de autocorrelação sugeriu-se cinco modelos, destacando-se entre eles o modelo ARIMA com os termos (p, d, q) igual a (0, 1, 1), em outras palavras um modelo com uma diferenciação de primeira ordem ($d = 1$) e um termo de média móvel ($q = 1$), sendo tal modelo o que apresentou os menores valores dos critérios de seleção estabelecidos: AIC, BIC e AICc. A (Tabela 1), apresenta os modelos testados bem como os respectivos resultados dos critérios de seleção.

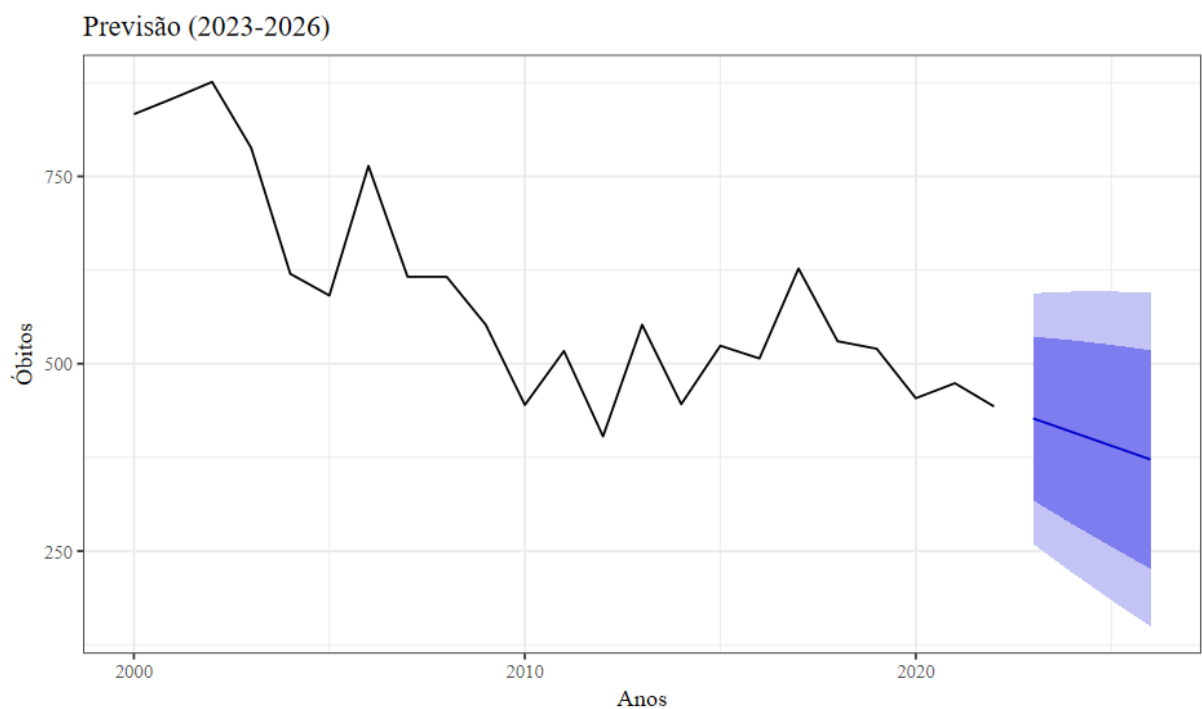
Tabela 1 – Modelos testados e seus respectivos resultados de AIC, BIC, AICc, RMSE e MAE.

Modelo	AIC	BIC	AICc	RMSE	MAE
(0, 1, 1)	262.20*	265.48*	263.54*	79.47	61.04
(1, 1, 0)	262.37	265.64	263.70	79.86	61.59
(1, 1, 1)	263.16	268.52	266.51	79.40*	60.15*
(0, 1, 2)	264.18	268.54	266.53	79.43	60.51
(2, 1, 0)	264.35	268.71	266.70	79.83	61.45

Fonte: Autor (2026).

Após selecionar o modelo ARIMA (0, 1, 1), realizou-se o teste de Ljung-Box, com o intuito de avaliar se os resíduos possuem características de um ruído branco, ao aplicar o teste no modelo selecionado, obteve-se como resultado ($Q^* = 1.957$ e $P\text{-value} = 0.7437$), indicando portanto que os resíduos possuem características de ruído branco, ou seja os resíduos tendem a seguir uma distribuição normal sem autocorrelação, tais características dos resíduos de um modelo são necessários para se obter boas previsões, com intervalos de confiança adequados, a (Figura 3), apresenta o gráfico com as previsões geradas pelo modelo selecionado.

Figura 3 – Gráfico da série temporal contendo a previsão para um horizonte de quatro anos (2023-2026).



Fonte: Autor (2026).

O modelo gerado utilizando o *drift* evidenciou uma tendência de redução linear dos óbitos com o passar dos anos, o modelo em questão apresentou bons intervalos de confiança

gerando margem para diferentes cenários. Para o ano de 2023 o modelo registrou uma média de 427 óbitos, para 2024 foram previstos 409 óbitos, para 2025 o modelo previu 391 óbitos e para o último ano do horizonte de previsão o modelo previu 372 óbitos. A (Tabela 2), contém os valores previstos com seus respectivos intervalos de confiança.

Tabela 2 – Valores com os seus respectivos intervalos de confiança..

Período (Anos)	Valor Médio (Previsão)	Intervalo 80% (Inferior)	Intervalo 95% (Inferior)	Intervalo 80% (Superior)	Intervalo 95% (Superior)
2023	427	318	260	536	594
2024	409	286	221	531	596
2025	391	256	185	525	596
2026	372	227	149	518	595

Fonte: Autor (2026).

Os intervalos de confiança abrangeram probabilidades para diferentes cenários, sendo estes otimistas ou pessimistas, uma vez que tais previsões são passíveis de oscilações, ou seja, dependem de fatores políticos, sociais, econômicos e climáticos. O modelo ARIMA foi obtido a partir de uma série temporal univariada, neste sentido toda a previsão do modelo é baseado no que já ocorreu no passado, outros modelos como os de série temporal multivariados podem ser utilizados agregando outras variáveis em conjunto a variável principal.

Vale ressaltar que tais previsões são benéficas ao setor público, sobretudo na área da saúde, uma vez que tendem a ser relativamente simples de se obter, além de possuírem uma alta eficiência na previsão (SATO, 2013). De forma geral os resultados indicam uma queda nos números de óbitos, que pode estar relacionada às constantes melhorias na infraestrutura e no acesso ao saneamento básico, bem como aos investimentos na saúde pública, sobretudo na educação aos cuidados e prevenção das doenças de veiculação hídrica.

4 CONCLUSÃO

O presente trabalho utilizou-se da metodologia de Box-Jenkins na elaboração de um modelo de previsão aplicado ao número de óbitos por doenças de veiculação hídrica no estado do Ceará, tal metodologia já consagrada na análise de séries temporais, possibilitou prever o número de óbitos com intervalos de confiança realistas, sendo amplamente aplicada a área da saúde, uma vez que possui alto poder preditivo possibilitando a tomada de decisões por parte da administração pública, promovendo portanto melhorias na saúde pública, no saneamento básico e no bem estar da população.

O pacote *forecast* disponível no *software* RStudio demonstrou possuir alto nível para análises e previsões baseadas em séries temporais, vale ressaltar as inúmeras funções que vão desde a plotagem, à sumarização do modelo e aplicação de testes de resíduos, possuindo uma gama de possibilidades para se avaliar os modelos gerados. O modelo ARIMA se mostrou útil nesta pesquisa ante a ausência de dados dos últimos anos (2023-2026), fornecendo possíveis valores que auxiliassem na análise de tais dados, sendo portanto considerado uma boa opção para pesquisas que envolvam a análise preditiva de variáveis provenientes de séries temporais univariadas, ou que envolvam dados de série temporais desatualizados.

De forma geral o modelo ARIMA (0, 1, 1) utilizado nesta pesquisa, apresentou bons resultados em três dos cinco critérios utilizados para a escolha do modelo, sendo portanto o melhor modelo para análise preditiva de tais dados. A modelagem apresentou de forma geral uma tendência de redução no número de óbitos relacionados à doenças de veiculação hídrica, evidenciando um quadro de melhoria na qualidade de vida da população quanto a segurança hídrica, o saneamento básico e as noções de higiene.

REFERÊNCIAS BIBLIOGRÁFICAS

AKAIKE, Hirotugu. Maximum likelihood identification of Gaussian autoregressive moving average models. **Biometrika**, v. 60, n. 2, p. 255-265, 1973. Disponível em: [biometrika1979.pdf](#). Acesso em: 15 de fevereiro de 2026.

BOX, G.; JENKINS, G. M. Analysis: Forecasting and Control. **San Francisco**, 1976.

BRUCE, P; BRUCE, A. **Estatística Prática para Cientistas de Dados: 50 Conceitos Essenciais**. Rio de Janeiro: Alta Books, 2019.

CARDOSO, Maria Regina Alves. Epidemiologia Ambiental. In: PHILIPPI, A. J. (Org). **Saneamento, saúde e ambiente: fundamentos para um desenvolvimento sustentável**. 2. ed. Barueri: Manole, 2018. p. 92-117.

DICKEY, David A.; FULLER, Wayne A. Distribution of the estimators for autoregressive time series with a unit root. **Journal of the American statistical association**, v. 74, n. 366a, p. 427-431, 1979. Acesso em: 08 de fevereiro de 2026.

HYNDMAN, Rob J.; KHANDAKAR, Yeasmin. Automatic time series forecasting: the forecast package for R. **Journal of Statistical Software**, v. 27, p. 1-22, 2008. Disponível em: <https://doi.org/10.18637/jss.v027.i03>. Acesso em: 07 de fevereiro de 2026.

HYNDMAN, Rob J.; ATHANASOPOULOS, George. **Forecasting: principles and practice**. 3. ed. OTexts, 2021. Disponível em: <https://otexts.com/fpppg>. Acesso em: 07 de fevereiro de 2026.

JUNIOR, Arlindo Philippi; MALHEIROS, Tadeu Fabrício. Saneamento e Saúde Pública: Integrando Homem e Ambiente. In: PHILIPPI, A. J. (Org). **Saneamento, saúde e ambiente: fundamentos para um desenvolvimento sustentável**. 2. ed. Barueri: Manole, 2018. p. 3-35.

KWIATKOWSKI, D., PHILLIPS, P. C., SCHMIDT, P., & SHIN, Y. Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?. **Journal of econometrics**, v. 54, n. 1-3, p. 159-178, 1992. Disponível em: [https://doi.org/10.1016/0304-4076\(92\)90104-Y](https://doi.org/10.1016/0304-4076(92)90104-Y). Acesso em: 10 de fevereiro de 2026.

LJUNG, Greta M.; BOX, George EP. On a measure of lack of fit in time series models. **Biometrika**, v. 65, n. 2, p. 297-303, 1978. Disponível em: <https://www.jstor.org/stable/2335207>. Acesso em: 10 de fevereiro de 2026.

MOTA, Suetônio. Conhecimentos para Promoção em Saúde. In: PHILIPPI, A. J. (Org). **Saneamento, saúde e ambiente: fundamentos para um desenvolvimento sustentável**. 2. ed. Barueri: Manole, 2018. p. 118-144.

SATO, Renato Cesar. Gerenciamento de doenças utilizando séries temporais com o modelo ARIMA. **Einstein (São Paulo)**, v. 11, p. 128-131, 2013. Disponível em: <https://doi.org/10.1590/S1679-45082013000100024>. Acesso em: 05 de fevereiro de 2026.

SIQUEIRA, Renato Prado. sidrar: An Interface to IBGE's SIDRA API. **R package**. version 0.2.9, 2022. Disponível em: <https://CRAN.R-project.org/package=sidrar>. Acesso em: 07 de fevereiro de 2026.

TRAPLETTI A; HORNIK, K. tseries: Time Series Analysis and Computational Finance. **R package**. version 0.10-58. 2024. Disponível em: <https://CRAN.R-project.org/package=tseries>. Acesso em 08 de fevereiro de 2026.