Auto-Regressive Integrated Moving Average Model (ARIMA): conceptual and methodological aspects and applicability in infant mortality

Aline Beatriz dos Santos Silva 1
https://orcid.org/0000-0001-9559-8524

Ana Catarina de Melo Araújo 2
https://orcid.org/0000-0003-1558-9789

Paulo Germano de Frias 3
https://orcid.org/0000-0003-4497-8898

Mirella Bezerra Rodrigues Vilela 4
https://orcid.org/0000-0001-5113-7144

Cristine Vieira do Bonfim 5
https://orcid.org/0000-0002-4495-9673

1 Programa de Pós-Graduação em Saúde Coletiva. Universidade Federal de Pernambuco. Av. Prof. Moraes Rego, 1235. Recife, PE, Brasil. CEP: 50.670-901. E-mail: alinebeatriz92@gmail.com
2 Secretaria Estadual de Saúde de Pernambuco. Recife, PE, Brasil.
3 Instituto de Medicina Integral Prof. Fernando Figueira. Recife, PE, Brasil.
4 Universidade Federal de Pernambuco. Recife, PE, Brasil.
5 Fundação Joaquim Nabuco. Recife, PE, Brasil.

Abstract
This reflective theoretical article, aims to discuss conceptual and methodological aspects about the applications of time series modeling, in particular, the Integrated Auto-regressive Moving Average model and its applicability in infant mortality. This modeling makes it possible to predict future values using past data, outlining and estimating possible scenarios of the health event, highlighting its magnitude. Due to the persistence of infant mortality as a public health problem, the applicability of this method is useful in the timely and systematic management of child health indicators, in addition to being a method with low operating cost, which in contexts of cost reduction in public healthcare services, becomes a potential management tool. However, there are still gaps in the use of statistical methods in the decision-making and policy-making process in public healthcare, such as the modeling in question. These are methodological (robust statistics), institutional (outdated information systems) and cultural obstacles (devaluation of the data produced, mainly at the local level).

Key words Time series studies, Infant mortality, Public health policy

This article is published in Open Access under the Creative Commons Attribution license, which allows use, distribution, and reproduction in any medium, without restrictions, as long as the original work is correctly cited.

http://dx.doi.org/10.1590/1806-93062021000200016
Introduction

The analysis of the situation of health care, the monitoring of priority indicators, and the forecast of scenarios are challenges in all countries, especially those with difficulties to take actions agreed internationally.1 Child mortality is particularly studied based on its expression as a public health problem and on the availability of technology for combating it.1,2 For this event, the use of the time series methodology is beneficial to create future scenarios, monitor and analyze the situation of health care.3,4 This modeling enables the diagnosis and comprehension of the temporal behavior patterns for events that affect a certain population and the assessment of the impacts of interventions in health care.5

Several techniques for analyzing time series can forecast future values, using data from the past, based on statistical conclusions. The purpose is to model the event, building a mathematical function that represents the correlation between the variable and time.6

No proposed model fully assumes the exact forecasts due to occurrences of random variations to which the event observation process is subject. However, they are valuable tools for quickly analyzing the severity of a situation and helping the public health care authorities to define or adjust the control strategies.7 Among the types of models, attention is called to the following: the analysis of a trend (moving average and exponential smoothing); the models of regression, which deal with different patterns occurring in the series, such as inflection points (joinpoint regression analysis); the artificial neural network, which was designed to work mathematically similarly to the human brain and be able to create generalizations based on past non-linear data, and the Autoregressive Integrated Moving Average (ARIMA), or Box-Jenkins Model, widely used in Economics.9,10 In the health care field, in the ’80s, the Center for Control of Diseases (CDC) adopted this modeling as a reference for the analyses in health care, and since then it has been disseminated across studies in the field.7,8

This paper aims to discuss the conceptual and methodological aspects of the time series by using ARIMA modeling and its applicability in child mortality. The purpose of the proposal for creating the manuscript is to contribute with health care practice to reduce persistent inequalities in child health, using strong statistical modeling. By proposing to forecast future scenarios, the systematic incorporation of the method addressed will strengthen the planning of directed actions in the various levels of the health care system.

The references to support the concepts and methods were characterized by books and scientific papers published in journals indexed in the databases provided by LILACS and the US National Library of Medicine (PubMed), in addition to SciELO’s virtual library. The explanations and considerations were organized into the following topics: Time series; theoretical and methodological aspects and ARIMA modeling; Applications of time series analysis in child mortality; methodological possibilities and limitations. The child mortality rates in Brazil were distributed into a time series to exemplify the applicability of the ARIMA modeling using a linear chart and the autocorrelation charts. The data for calculating the child mortality rate were extracted from the Information Systems about Mortality (SIM – Portuguese acronym) and from the Information System about Live Births (Sinasc – Portuguese acronym) and obtained from the DATASUS website available for the public at: https://datasus.saude.gov.br/.

Time series: theoretical and methodological aspects and ARIMA modeling

The time series is a sequence of observed values of a certain phenomenon distributed into a time basis. The mathematical expression that describes a time series is \( \{Y(t), t \in T\} \), where \( Y \) represents the variable of interest and \( T \) represents the index set related to the measurement times. The time series analysis aims at building explanatory or deterministic models for the phenomenon studied and, with that, making the respective forecasts.6

The type of values observed characterizes the time series as continuous or discrete. Based on the general function, a discrete series \( T = t_1, t_2, ..., t_n \), the observations are made in fixed time intervals that can be enumerated, which can be equispaced or not. In the continuous series, the pieces of data regarding a certain variable are evidenced sequentially in a time interval \( T = t : t_1 < t < t_2 \), for example, the continuous measurement of a certain biological sign, such as arterial pressure.8

In the analysis of the series, some basic aspects must be taken into account to enable a better understanding of the behavior of the variable observed. The first one refers to the frequency, which depends on the type of the variable under analysis, namely, continuous or discrete, and on the objective of the study proposed. In the case of the child mortality rate, its frequency is related to the availability of data in official information systems and health inves-
tigators who perform constant critical analyses to measure the magnitude of the event. Another important aspect refers to the comparability between time series of different contexts; in this situation, it is necessary to respect certain circumstances that allow contrasting the series.

Non-stationarity is a characteristic considered important for time series and should be avoided because it hinders forecasts. A non-stationary series results from the frequency of the data studied and may be unstable in its average, variance, and autocovariance. Regarding the child mortality rate, very-high or very-low values of deaths within a period cause great variations in the series. Accordingly, the processing of the series, regarding its stationarity, is a necessary condition in time series modeling. A non-stationary series implies that the model is determined by the randomness of the observations, by chance.

To estimate the model, first, it is necessary to make sure that it is not a purely random sequence, also referred to as random variation, white noise, or random residual. If it is not, the function of the model may be formed by components that represent partial regularities or patterns of the series studied, namely: trend, seasonality, and the estimate variability of the random residual to create the confidence intervals for the forecasts arising from the model.

The first component to be identified is the trend through the time chart and some methods that smooth the fluctuations of the original series, such as moving averages or linear transformations. The moving averages, or linear filters, are used whenever a series proves to be markedly irregular, with many

*Stationarity test (Augmented Dickey-Fuller): p-value 0.38. This indicates that the series is not stationary.
Source: Mortality Information System and Live Birth Information System – DATASUS, Ministry of Health
fluctuations, hindering the visualization of the trend. The more sinuous the trend, the more smoothing is necessary for identifying the real trend of the series.

The trend is removed with a technique to differentiate the series itself in order to make it become stationary (Figure 3). A difference eliminates a linear trend and two differences eliminate an exponential trend. The number of differences is related to the degree of the polynomial estimated for the trend.

The seasonality component is a phenomenon that occurs regularly in the time. The relations of the observations in series with seasonality occur frequently in yearly and monthly time series, and may also occur in series measured in other time dimensions. Seasonal patterns are identified in the original chart (Figure 1) and in the correlogram based on oscillations in the same frequency (Figure 3). The analysis is performed based on the autocorrelation coefficients in the seasonal periods.

The Autoregressive Integrated Moving Average (ARIMA), or Box-Jenkins Model, is widely used for forecasting. This technique uses data from the past to estimate future values, and the instability with its fluctuations is an obstacle to the applicability of the technique. The identification of the stability of the series, or stationarity, is the first stage of the modeling. Given the possibility that there are seasonal periods in the series, the ARIMA modeling has an extension, the Seasonal Autoregressive Integrated Moving Average (SARIMA), or seasonal ARIMA. In series without stationarity due to the fluctuations (trend, seasonality) inherent in themselves, the ARIMA model enables obtaining satisfactory results.

The inducement to stationarity respects an order based on an assumption about the preservation of the fluctuations of the original series. First, the non-stationarity of the variance is assessed, followed by the autocorrelations and, finally, the average. Depending on the degree of the trend present in the series, by simply correcting the non-stationarity of the variance, the average and autocorrelations consequently stabilize. That is why the abovementioned order for adjustment of the series must be respected so that it is not necessary to induce stationarity for the autocorrelations and average.

Generally, the ARIMA method proposed by George Box and Gwilym Jenkins can diagnose the series regarding the stationarity condition and identify the proper model based on an iterative cycle with the data of the series itself, with the following stages: identification, specification, estimation, and diagnosis.

By assessing the elements of the acronym ARIMA during the stage of identification of the potential models, we note that the AR, or $p$ order, represents the autoregressive process, that is, the influence of the previous value of the variable on the value under consideration; the $I$, or $d$ order, is related to the number of differenciations to induce to stationarity, and the MA, or $q$ order, is related to the influence of the noise generated in the previous value. The orders of the ARIMA and SARIMA models are estimated by the chart of the autocorrelation function (ACF) and by the correlogram of the partial autocorrelation (PACF); the ACF is sugges-

![Figure 2](image)

**Figure 2** Autocorrelation function (ACF) and partial autocorrelation function (PACF) of the monthly child mortality rate in Brazil, 2000-2018.

ARIMA models and applicability in infant mortality

Observing the ACF and the PACF is useful to get an idea of the model to be tested, in order to choose the one with the smallest amount of parameters (the most parsimonious one).

The SARIMA models have a non-seasonal \((p, d, q)\) and a seasonal order \((P, D, Q)\), and, for this reason, they are referred to as multiplicative models. The estimation of the parameters (or coefficients) uses the following methods: least squares, maximum likelihood, or method of moments.

The time series modeling process may be systematically outlined based on the following flowchart (Figure 4).

**Applications of time series analysis in child mortality: methodological possibilities and limitations**

Acting in the main determinants and conditions of health hazards requires studies with strong modeling, such as the time series models, which allow generating scientific evidence to support the decision-making process for health care.

The incorporation of time-series studies in the health care field arises from the need for planning where to allocate the investments, in order to impact the main epidemiological indicators, such as child and maternal mortality. Additionally, it is a practical model, as it allows the use of demographic, epidemiological environmental, and socio-economic data provided by official sources.

Based on this type of study, it is possible to

*Figure 3*

Monthly child mortality rate and autocorrelation function (ACF) and partial autocorrelation function (PACF) after the differentiation process \((d=1)\), Brazil, 2000-2018.

*Stationarity test (Augmented Dickey-Fuller): \(p\)-value 0.01. This indicates that the null hypothesis of non-stationarity was rejected. Source Mortality Information System and Live Birth Information System: – DATASUS, Ministry of Health.*
To induct stationarity by means of differentiation techniques (order)

The temporal series is stationary?

No

Yes

To identify possible ARIMA models (p, q order)

ACF and PACF

Temporal chart ACF and PACF stationarity tests?

ITo identify possible ARIMA models (p, q order)

ACF and PACF

Estimate model parameters

Least Squares Method

Maximum likelihood

Method of moments

The estimate model is adequate? (Diagnosis)

No

Yes

Forecast

ACF = Autocorrelation Function; PACF = Partial Autocorrelation Function; AIC = Akaike Information Criterion; BIC = Schwarz’s Bayesian Criterion.

To understand the behavior of variables during a period, discover atypical patterns in morbimortality, and understand the determination of the causes; and it is useful to assess the impact caused by health care interventions.\(^{18}\)

To forecast scenarios, it is necessary to use diversified techniques, including those that assess the characteristics of the data studied. The ARIMA modeling has this versatility regarding epidemiological variables, which are dynamic by nature; that is why it is useful in stationary or non-stationary series.\(^{12,19}\) Another methodological advantage, compared to other forecasting techniques, is the ability to build parsimonious models with a reduced amount of parameters (iterative cycle estimation stage) and the fact that the forecasts obtained can be quite accurate in several contexts.\(^{6}\)

Despite the potentials presented, there are some limitations, which may arise when using time series modeling. One of them, inherent in this type of study, considers the existence of a linear relation between the pieces of data observed and the pieces of data from the past, which is not the case for real pieces of data, which have a complex, non-linear relation.\(^{7}\) Child mortality, for example, has a multifactorial determination resulting from the interaction of biological, social, economic, and care-related variables.\(^{20}\)

Another limitation is related to the quality of the data available, as it interferes directly with the modeling.\(^{15}\) Available and reliable data are essential for providing the necessary information for determining policies and delimiting the vulnerable groups. Therefore, it is essential to provide the health care information systems with proper data.\(^{21}\)

Even though the time-series studies are classic and essential for understanding the situation of a certain health care issue, sometimes they are
ARIMA models and applicability in infant mortality

This document discusses the application of ARIMA models in the field of infant mortality, highlighting the challenges and potential of using these models. ARIMA models are regarded as complex and hard to implement, but they are not used in all of their potential. There are still difficulties in the handling and application of the ARIMA modeling in particular in the health care field, due to its mathematical nature arising from the economic field. It is a sophisticated model that requires familiarity with the theoretical aspects and training in statistical analysis to complete the entire iterative cycle and the forecast.

Its persistence as a global public health problem makes child mortality a consensual agenda in the health care field, and the ARIMA modeling is a methodological possibility for its management. By using it, it is possible to delimit and analyze the most likely health-related trends, characterizing them as important devices for planning interventions.

The use of ARIMA models in child health care would work as a tool that precedes and supports the practice of assistance. Its incorporation in the planning of strategic impact actions may contribute to the reduction in unfavorable outcomes in the child and maternal health-disease process. Recently, some studies used the ARIMA modeling to express different dimensions of the issue. Table 1 shows a few examples.

In spite of that, there is a significant gap in the use of statistical methods in the health-related decision-making and policy-formulating process. An example of that is the incorporation of time-series studies as health-management tools. The forecasts require the overcoming of an already mentioned limitation: data collected systematically (proper and quality data) and access to epidemiological data, not only by the decision makers. This is a reality in most of the poorest countries, especially those with difficulties in consolidating the quality information systems and creating a data usage culture.

Building and strengthening sturdy health-related information systems is still a challenge. In 2015, as a result of the discussions about the objectives and goals of the 2030 Agenda, key principles (focus, relevance, innovation, equity, leadership, and national ownership) were proposed to achieve strong models for global monitoring of child and maternal health indicators, including the ARIMA modeling, and bring them into the national and local realities.

The focus refers to the definition of child and maternal health indicators standardized globally so it is possible to measure, monitor, and estimate them in all national health sublevels. The purpose is to value the production of national and local data, delegating joint decision-making responsibility to the actors in these spaces. Therefore, it is necessary to provide them with the technical capacity to understand the data handling complexity. Additionally, subjective incentives are required, such as encouragement and motivations for using the data, by identifying the potential of the modeling. In addition to these principles, the perspective of innovation is included. The ARIMA model, among others addressed, is a promising tool for data interpretation. The forecasts enabled by the ARIMA modeling may precede strategic actions and mitigate the morbidity-mortality load; the forecasts are used to increase the number of alternatives in decision-making processes. Even considering the limitations inherent and foreign to the method discussed, its potentials, applications, and versatility stand out, characterizing it as a feasible method in health care practice.

Author’s contribution

Silva ABS, Frias PG, Vilela MBR, Bonfim CV, and Araújo ACM contributed to all stages of the preparation of the paper: conception, outlining, writing, and approval of the final version of the paper.
<table>
<thead>
<tr>
<th>Title</th>
<th>Authors (Year of publication)</th>
<th>Applicability of the method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Change of Outcomes in Pediatric Intestinal Failure: Use of Time-Series Analysis to assess the evolution of an Intestinal Rehabilitation Program</td>
<td>Oliveira et al. (2016)</td>
<td>The study evidenced the versatility of the ARIMA method, expanding its scope of application in the clinic. The efficacy of the treatment options in the improvement of the clinical status of neonatal or infantile intestinal failure was assessed, showing the downward trend in the primary outcome (mortality).</td>
</tr>
<tr>
<td>Post Millennium Development Goals Prospect on Child Mortality in India: An Analysis Using Autoregressive Integrated Moving Averages (ARIMA) Model</td>
<td>De et al. (2016)</td>
<td>The ARIMA modeling was used to support the monitoring of the Sustainable Development Goals in India and forecast the health care actions to be given priority in order to achieve the respective goals, as well as an essential tool for formulating an overall national health planning.</td>
</tr>
<tr>
<td>Predicting infant mortality in India using time series models</td>
<td>Singh and Singh (2018)</td>
<td>The method was applied to forecast the child mortality rate in some states of India. It allowed the comparison between the respective historical series and the forecasts.</td>
</tr>
<tr>
<td>Trend of Neonatal Mortality in Nigeria from 1990 to 2017 using Time Series Analysis</td>
<td>Usman et al. (2019)</td>
<td>The ARIMA modeling was used to investigate the trend of incidence of neonatal mortality, making the 20-year forecast. A consistent decline in the neonatal mortality rate is evidenced, in addition to the fact that the local neonatal health policy is operational and committed to reducing the incidence of neonatal deaths.</td>
</tr>
<tr>
<td>Trends and future of maternal and child health in Bangladesh</td>
<td>Rajia et al. (2019)</td>
<td>The method enabled the analysis of the trend of the maternal and child health indicators. With the scenario forecasted for the rates, it is characterized as a potential tool for supporting the decision-making of local management in monitoring and the implications of public policies to achieve the goals proposed in the 2030 Agenda.</td>
</tr>
<tr>
<td>A mortalidade infantil no estado de São Paulo: uma previsão da taxa por meio da modelagem SARIMA</td>
<td>Chaib (2019)</td>
<td>The historical series was used to describe the evolution of the child mortality rate over time (1996-2016). The ARIMA method applied enabled evidencing the seasonal behavior of child mortality, as it has drops in specific periods, hence identifying a SARIMA model. The forecast evidences a downward trend for the rate.</td>
</tr>
<tr>
<td>Forecasting Indian infant mortality rate: An application of autoregressive integrated moving average model</td>
<td>Mishra et al. (2019)</td>
<td>The ARIMA modeling allowed defining an overview for a future scenario of child mortality. The forecast was for the nine-year period (2017-2025), with a downward trend. The study highlights one potential of the method: the reliability of the forecast when it estimates the period of the sample using available data. It evidences that it is a statistical tool of great value in health care, supporting the planning of interventions.</td>
</tr>
<tr>
<td>Modeling and Forecasting Infant Deaths in Zimbabwe using ARIMA Models</td>
<td>Nyoni and Nyoni (2020)</td>
<td>The model was applied to forecast the number of child deaths in a country under peripheral capitalism with a high mortality rate. Having such information would bring gains more effectively to the local health intervention programs.</td>
</tr>
</tbody>
</table>
References


Received on September 1, 2020
Final version presented on January 18, 2021
Approved on March 24, 2021