SESUG Paper 245-2018

# Predictive Modeling for Healthcare Professionals: The use of time-series analysis for health-related data and the application of ARIMA modeling techniques in SAS for Public Health Practice

Camillia R. Comeaux, MSPH, Florida A&M University

## ABSTRACT

The use of time-series analysis in public health practice is an under-utilized tool that can aid in effective activities such as: program and health planning; appropriate health service and provider delivery; improved emergency preparedness action; and much more.

The ARIMA technique is a type of trend recognition tool in time-series analysis, that can sort through large amounts of data and create a statistical model for forecasting. Time-series models, such as ARIMA, were historically used in financial industries to assess risks and market changes overtime to predict future economic outcomes. The ARIMA technique is a process in which the stages of model identification, parameter estimation, and diagnostic checking are repeated to find the most appropriate fitting model for prediction (Chen, 2008). This modeling technique is operational when data is assumed to have stationarity, or without a trend, and uses longitudinal data with at least forty-five data points to increase the accuracy of forecasting (Chen, 2008).

Engle (2001) suggests that the utility of modeling techniques in time-series analysis, is their ability to factor in major shocks and volatility shifts over time; therefore, in the public health field, the social, the economic, and the ecological factors correlated to these shocks over time can be analyzed for health outcome forecasting and future population health planning.

This product gives a generalized overview of time-series analysis and its application to public health practice. Using the SAS statistical package, health professionals will be empowered to use time-series analysis, specifically ARIMA modeling techniques, to analyze and interpret large health data sets to predict or forecast factors that impact health outcomes on populations.

## INTRODUCTION

The use of time-series analysis in public health practice is an under-utilized tool that can aid in effective activities such as: program and health planning; appropriate health service and provider delivery; improved emergency preparedness action; and much more. The importance of time-series analysis lies in the researcher's ability to take a collection of past observations, estimate their parameters, choose the most suitable model, and ultimately predict future outcomes based on that past data (Balasubramanian & Ravindran, 1979). Therefore, an understanding of statistical tools in time-series analysis can be used to investigate historical data, visualize trends, and demonstrate correlations between hypothesized risk factors and volatile health outcomes.

The ARIMA technique is a type of trend recognition tool in time-series analysis that can sort through large amounts of data and create a statistical model for forecasting. Time-series models, such as ARIMA, were historically used in financial industries to assess risks and market changes overtime to predict future economic outcomes. The ARIMA technique is a process in which the stages of model identification, parameter estimation, and diagnostic checking are repeated to find the most appropriate fitting model for prediction (Chen, 2008). This modeling technique is operational when data is assumed to have stationarity, or without a trend, and uses longitudinal data with at least forty-five data points to increase the accuracy of forecasting (Chen, 2008).

Engle (2001) suggests that the utility of modeling techniques in time-series analysis, is their ability to factor in major shocks and volatility shifts over time; therefore, in the public health field, the social, the economic, and the ecological factors correlated to these shocks over time can be analyzed for health outcome forecasting and future population health planning.

This product gives a generalized overview of time-series analysis for health professionals and its application to public health practice. The goals of this product are to: (1) give a basic overview of time-series analysis and its key components; (2) describe the three stages of the ARIMA technique, a statistical modeling tool for time-series analysis; and (3) discuss the implications and how to apply time-series analysis to public health practice and large health-related data sets.

Using the SAS statistical package and this product, health professionals will be empowered to use time-series analysis, specifically ARIMA modeling techniques, to analyze and interpret large health data sets to predict or forecast factors that impact health outcomes on populations.

## TIME-SERIES ANALYSIS OVERVIEW AND KEY COMPONENTS

Time-series is when observations or measurements are in a sequence and equally spaced through time (Zeger, Irizarry & Peng, 2005). Examples include values such as: the daily number of live births or deaths, hospital admissions, the monthly average temperature, the U.S. per capita income, etc. Unlike typical regression analysis, health professionals use time-series analysis to understand how predictor variables that are close together in time tend to be correlated with one another; in which, those correlated variables influence a particular health outcome (Zeger, Irizarry & Peng, 2005).

Time-series analysis can often be used concurrently with regression analysis to explain the dependence of the response on the correlated predictor variables at each specified time period (Bolleslev, Engle & Nelson, 1994). The internal correlation of between observations in time-series analysis is known as autocorrelation ("Time-series Introduction", n.d.) This dependency from each observed time-period makes forecast values reliable. The model accuracy step in various time-series analysis techniques confirms the forecasting ability of the model and signifies the level of agreement between the actual and forecasted values (Bolleslev, Engle & Nelson, 1994).

### KEY COMPONENTS OF TIME-SERIES ANALYSIS

1. *Seasonality*. This means the data experiences regular and predictable changes that recur in regular intervals through time.
2. *Cyclic or Periodic*. This means the data is observed in a non-fixed pattern over a span of time.
3. *Trend*. This means the data is either increasing or decreasing over a span of time.
4. Irregular. This means the data points observed are unpredictable over a span of time.
5. *Ordering of the time points matter*. This is because observations at one time point might influence observations in the future.
6. *The average, or mean, of observation points must hover around $\mu = 0$ to produce a reliable forecasting model*. This means the model produced from time-series analysis is an estimation of all past observations with consideration of potential error calculations.
7. *Stationarity*. This means statistical values such as mean value, the variance, and the autocorrelation must be made constant over time to do the analysis (i.e removing the trend).

## ARIMA TECHNIQUE OVERVIEW

The Box-Jenkins Autoregressive Integrated Moving Average (ARIMA) model is a type of trend recognition tool in time-series analysis, that can sort through large amounts of data and create a statistical model for forecasting. Within time-series analysis, there are several techniques that develop statistical models used for forecasting or prediction. The ARIMA technique is a process in which the stages of model identification, parameter estimation, and diagnostic checking are repeated to find the most appropriate fitting model for prediction (Chen, 2008). This modeling technique is operational when data is assumed to have stationarity, or without a trend, and uses longitudinal data with at least forty-five data points to increase the accuracy of forecasting (Chen, 2008).

The ARIMA modeling technique assumes stationarity (Chen, 2008). This means that the random observed data points that are equally spaced over time, now have a constant mean and variance which does not change when shifted in space or time. In the analysis of health-related data, the ARIMA modeling technique is used to determine the best fitting model and suggest estimated parameters for the risk factors or explanatory variables in forecasting.

As a result of collecting a large number of volatile data points over a span time; during the phases of analysis, forecast errors or residuals are likely to occur. Residuals, or forecast errors, are the differences in the actual and forecasted values. Therefore, when there are a small number of errors then the model is considered to have high accuracy (Chen, 2008). Mean error (ME), mean absolute error (MAE), sum of the squared errors ($\varepsilon^2$), mean percentage error (MPE), and mean absolute percentage error (MAPE) are calculated values in time-series analysis that statistically demonstrate the accuracy of the developed ARIMA models. Additionally, those calculated values are used to determine the best estimated parameter values of the risk factors or explanatory variables in forecasting (Bollerslev, Engle & Nelson, 1994).

## ARIMA MODEL FITTING PROCEDURE

The ARIMA model fitting procedure follows a three-step process:

### (1) Model Identification

In the model identification step, a generalized model is identified based on the 3 graphs: the sequence plot of the health outcome rates; the plot of the autocorrelation function (**ACF**) of the residual series; and the plot of the partial autocorrelation function (**PACF**) of the residual series. [*Reminder, the residual series, or the series of forecast errors are the differences between the actual and forecasted values.*]

The general ARIMA model is known as ARIMA (*p, d, q*). The *p, d, q* parameter values are denoted by the *p*th order of the autoregressive effect, the *d*th order of the differencing, and the *q*th order of the moving average. The order of autoregressive effect (*p*) is number of lags of past values in the regression equation of the series. Lags are the amount of time or space between observed events (Soyiri & Reidpath, 2013). The order of differencing (*d*) is the number of terms in a statistical process that removes a trend and establishes stationarity. The order of moving average is the number of terms to be included in the model that reflects a combination of past residual values. Different ARIMA (*p, d, q*) models are developed to determine the best fitting model for forecasting.

#### *(A) Sequence Plot*

The first step in the model identification process is the analysis of the sequence, or series, plot. The plot is a graphical representation of the actual past values of the data. If the sequence plot graph presents a trend, the method of differencing must be used to remove the trend. Remember, ARIMA is only operational under the assumption of stationarity of the data. The removal of trends clarifies the analysis of the proper causal relationship in the data series and achieves stationarity requirement (Chen, 2008). First order of differencing (*d*=1) is designed to remove the trend.

#### *(B) ACF and PACF plots*

The ACF and PACF residual plots are used to identify the model structure. The model structure will be found to be either an AR(*p*) model or a MA(*q*) model (or mixed model). The Autocorrelation Function (ACF) is the correlation of error terms over a span of time. It measures the association between actual and forecasted variables by lag over time. The order of the ACF is the value for the number of lags between the data points over a span of time. The PACF is the partial correlation of error terms controlling for the lags. The PACF plot evaluates the strength of the association between the actual and forecasted values.

**IF**… the higher order ACF takes zero value + the PACF has a spike
**THEN**… the autoregressive (AR) model may be most appropriate.
*In this case the PACF plot is the most helpful tool for identifying the order of an AR model.*

**IF**… the PACF rapidly approaches a zero value without a spike
**THEN**… the MA model may be appropriate.
*In this case the ACF plot is the most helpful tool for identifying the order of a MA model.*

**(2) Parameter Estimation**

In the estimation stage, each of the tentative models derived are estimated and various coefficients are examined (Ďurka & Silvia, 2012). The regression coefficients are estimated based on the least-squares method. Using least squares method in a nonlinear capacity requires that the regression coefficients are derived in a way that the estimates that the health outcome rate series come as close as possible to the actual health outcome rate series.

**(3) Diagnostic Checking**

The diagnostic checking step aims to determine if the derived model is the most suitable and adequate model for forecasting. The chi-square test is used to evaluate if the ACF of the residual series exhibits any systematic pattern. This step closely examined because large forecasting errors could be the result of a model identified as not suitable. If a model is not suitable, the model fitting procedure is repeated until the most suitable model is found.

## IMPLICATIONS AND THE APPLICATION OF TIME-SERIES ANALYSIS TO PUBLIC HEALTH PRACTICE

The complexity of time-series statistical analysis typically deters health professionals not well-versed in application and usefulness of tool. But, utilizing time-series analysis proves valuable to the health care field for predicting future health-related events. The application of time-series analysis in public health practice can aid in effective activities such as: program and health planning; appropriate health service and provider delivery; improved emergency preparedness action; and much more.

### KEY IMPLICATIONS AND APPLICATIONS TO PUBLIC HEALTH PRACTICE

1. *Large data sets*. The analysis is designed to support large data sets with multiple collection points. The greater the number of equally spaced values over the span of time increases the reliability and accuracy of the forecasting model for future related events.
2. *Ability to factor in major shocks in volatility shifts over time*. This means in the application to public health practice, health professionals can assess the social, the economic, and the ecological factors correlated to these shocks over time associated with health outcomes for forecasting and future population health planning.
3. *Rationalization for preventive programming*. This means that predictive risk factors associated with health outcomes can be forecasted to reduce incidence rates of future events. Therefore, the operationalization of this tool can be used for policy development, emergency preparedness, strategic budget planning, etc.
4. *Important for health service delivery analysis and planning*. According to Soyiri & Reidpath (2013), time-series analysis can enhance preventive health care services; create alerts for the management of patient overflows; and significantly reduce the associated costs in supplies and staff redundancy.

## REFERENCES


Balasubramanian, P. & Ravidran, A. (1979). A time series aggregation model for predicting the incidence of syphilis. *Sexually Transmitted Diseases, 6*(1), 14-18.

Bollerslev, T., Engle, R. F., & Nelson, D. B. (1994). ARCH models. In D. M. RF Engle (Ed.), *Handbook of Econometrics* (pp. 2961-3030; 49)

Boyle, J.; Jessup, M.; Crilly, J.; Green, D.; Lind, J.; Wallis, M.; et al. (2011). Predicting emergency department admissions. *Emergency Medicine Journal*. Doi: 10.1136/emj.2010.103531

Chen, C. (2008). An integrated enrollment forecast model. *IR Applications: Using Advanced Tools, Techniques, and Methodologies, 15*, 1-18.

Ďurka, P. & Silvia, P. 2012. "ARIMA vs. ARIMAX – which approach is better to analyze and forecast macroeconomic time series?" Proceedings of 30th International Conference Mathematical Methods in Economics Section, 136-140. Karviná, Czech Republic. Silesian University in Opava, School of Business Administration in Karviná.

Engle, R. (2001). GARCH 101: The use of ARCH/GARCH models in applied econometrics. *Journal of Economic Perspectives, 15*(4), 157-168.

Soyiri, I. N, & Reidpath, D. D. (2012). An overview of health forecasting. *Environmental Health and Preventative Medicine, 18*(1), 1-9. doi: 10.1007/s12199-012-0294-6

Time-Series Introduction. (n.d.). Retrieved from https://people.maths.bris.ac.uk/~magpn/Research/LSTS/STSIntro.html

Zeger, S. L.; Irizarry, R.; & Peng, R. D. (2006). On time series analysis of public health and biomedical data. *Annual Review of Public Health, 27*(1), 57-79. doi: 10.1146/annurev.publhealth.26.021304.144517


## CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:


Camillia R. Comeaux, MSPH

Florida A&M University, Doctor of Public Health Student

816-868-8069

camilla1.comeaux@famu.edu