Paper 458-2013

Non-Temporal ARIMA Models in Statistical Research

David J. Corliss / Magnify Analytic Solutions / University of Toledo Physics and Astronomy

ABSTRACT

Mathematical models employing an autoregressive integrated moving average (ARIMA) have found very wide applications following work by Box and Jenkins in 1970, especially in time series analysis. ARIMA models have been very successful in financial forecasting, forming the basis of such things as predicting how much gas prices will rise. However, no mathematical requirement exists requiring the data to be a time series: only the use of equally spaced intervals for the independent variable is necessary. This can be done by binning data into standard ranges, such as income by $10,000 intervals. This paper reviews the fundamental statistical concepts of ARIMA models and applications of non-temporal ARIMA models in statistical research. Examples and applications are given in biostatistics, meteorology, and econometrics as well as astrostatistics.

Keywords: ARIMA, Non-Temporal, Time Proxy, Mixed ARIMA Model

INTRODUCTION

Regression is often used in a very general way to identify a relationship between different variables and parameters in order to use several known quantities to predict the value of something that is unknown. The reliance of regression models on the Central Limit Theorem requires that the input data be normally distributed: strongly skewed data renders the technique invalid. Further, the data must consist of Bernoulli Trials: independent observations without autocorrelation and with random errors following a Gaussian distribution. In many cases, however, the values of points in a series may be strongly related to the values of other points. Two common mechanisms result in this lack of independence. In autocorrelation, the pattern between successive points in a series closely matches the pattern seen in other parts of the series. In other cases, the values at each point are driven by the local average to which each point makes a contribution. The significant presence of either on these properties invalidates the use of regression to develop a model. Under these circumstances, ARIMA models may provide a better choice.

ARIMA stands for Autoregressive Integrated Moving Average. These models use a labeling schema designating the number of terms of each type of component – autoregressive, integrated and moving average. This statistical technique examines trends in data by using several successive terms to predict the next value (or next several values) in a series.

These models can have an autoregressive part, predicting successive values in a series using a linear model based on the values of a fixed number of previous terms in the series. Since the success of the model relies on the strength of local trends in the data, it is not invalidated by the presence of autocorrelation like standard regression models and often performs best when autocorrelation is strong.

The relationship between sets of points may be strongest when the two sets are separated by a fixed number of intervening points. An example of this is found in seasonal data, where patterns repeat on an annual basis. Ordinary ARIMA models, with time as the independent variable, often include a term to introduce a time lag to create a *stationary process*, where the probability distribution does not change when displaced by some value of the independent variable. In the labeling schema used to describe ARIMA models, the second number is the “integrated” part of the model and specifies the number of lag

terms. Thus, a model with two autoregressive terms (rarely found but used here for illustration), one lag term and no moving average term is called a (2,1,0) ARIMA model. Introducing the appropriate time lag into a time-based ARIMA model can eliminate periodic variation, resulting in a stationary model, that is, one unchanged under time displacement. However, we find this property can exist for ARIMA models in general, including those for a series of values where the independent variable is not time.

Some data series are marked not by autocorrelation but rather by variation around a local average that changes as the series progresses. In this case, a moving average is used to estimate values for subsequent points in the series. In these Moving Average models, the "error term" - the difference between the prediction and actual value of a point in the series – drives the prediction of subsequent points. Moving average models may also include a lag term to create a stationary data set. In the absence of strong trends found in autoregressive models, moving average models often work best when trying to establish a trend line or prediction in data with known relationships between points while displaying large or chaotic variations.

Autoregressive models are driven by a local *trend*, while moving average models are driven by a local *average*. Both of these are distinct from regression, where the requirement of independence demands the absence of any significant local effects. In practice, ARIMA models will normally have either an autoregressive or a moving average component but not both. This is done to avoid "over-fitting", where a model is overly reactive to errors, describing the noise in the data rather than the underlying behavior. Ordinarily, time is seen as the primary variable driving the model in order to predict values at some point in the future. However, the underlying mathematical architecture of these models does not require the use of time as the input: the only thing that is needed is evenly spaced intervals. This frees these models to be used to study trends and relationships in many different areas.

AUTOREGRESSIVE MODELS

(1,1,0) ARIMA models are sometimes referred to as Box–Jenkins models, following the seminal work by George Box and Gwilym Jenkins in the 1970's. These models are very widely used in economic forecasting today. It is important to remember that these models work best in the presence of strong trends, such as organic growth. The example below is a classic application of a (1,1,0) ARIMA model: used in econometrics to forecast the future price of a commodity, in this case, the US average retail price per gallon of regular gasoline.

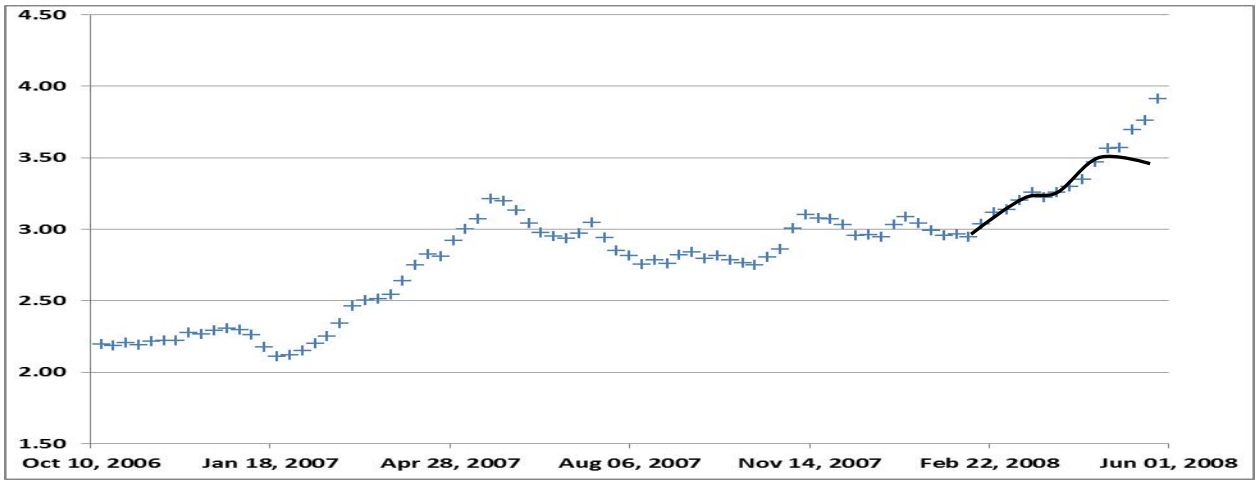

Figure 1. Regular gasoline, retail price per gallon, U.S. weekly average

Because (1,1,0) ARIMA models rely on the strength of local trends to make a forecast, they tend to work best when there is a smooth change over time. Chaotic changes introduce a large random element, decreasing the relative amount of autocorrelation and weakening the model. The success and failure of this gas price model demonstrates the strengths and weaknesses of auto-regressive models: the model performed very well so long as the changes were smooth and failed with the advent of chaotic fluctuations in gas prices in late spring of 2008.

A simple autoregressive ARIMA model can be created in Base SAS using a macro to perform a PROC REG recursively, stepping forward at each iteration a fix time interval between successive data points. This example uses the Cryer milk data, a well-known data set often used to illustrate techniques in Time Series Analysis. These data are a compilation of dairy farm's monthly milk per cow, tracked over a fourteen year period from 1962 to 1975.

The following macro uses PROC REG to predict the next value in a series. The macro requires a new record for each successive point in the series. The macro parameter N gives the number of points to be used by PROC REG. The macro selects the last N in the data set, corresponding to the last N points in the series; only these points are used to predict the next value. The predicted value is then added to the end of the time series. Repeated application of the macro applies the PROC REG recursively to create an ARIMA model.

```
%MACRO AC(N);

PROC SORT DATA=WORK.TSERIES;
   BY DUMMY;
RUN;

DATA WORK.LAST;
   SET WORK.TSERIES;
   BY DUMMY;
   IF LAST.DUMMY;
   RECENT = _N_ - &N. + 1;
   KEEP DUMMY RECENT;
RUN;

DATA  WORK.RECENT;
   MERGE WORK.TSERIES WORK.LAST;
   BY DUMMY;
   IF _N_ GE RECENT;
   DROP RECENT;
RUN;


PROC REG DATA=WORK.RECENT NOPRINT;
   MODEL Y=T;
   OUTPUT OUT=WORK.TREND PREDICTED=AUTOREG_TERM RESIDUAL=RESIDUAL;
RUN;


DATA WORK.TREND;
   SET WORK.TREND;
   OUTPUT;
   T_PREVIOUS = T;
   LAG_TERM = ((-1)**RAND(BERNOULLI,))* RAND('NORMAL',LAMBDA,SIGMA);
   Y_PREVIOUS = AUTOREG_TERM + LAG_TERM;
   RETAIN T_PREVIOUS Y_PREVIOUS;
RUN;

DATA WORK.NEW;
```

```
      SET WORK.TREND;
      BY DUMMY;
      IF LAST.DUMMY;
      DELTA_T = T - T_PREVIOUS;
      T = T + DELTA_T;
      DELTA_Y = Y - Y_PREVIOUS + 1;
      Y = Y + DELTA_Y;
      KEEP T Y DUMMY;
   RUN;

   DATA WORK.TSERIES;
      SET WORK.TSERIES WORK.NEW;
   RUN;

%MEND AC;
```

The value-added SAS Economics and Time Series software (SAS / ETS) has much more extensive capabilities, featuring the ARIMA procedure. This procedure includes tools to identify a lag term needed for stationarity, output datasets, dealing with missing data, periodicity, output and other features.

MOVING AVERGE MODELS

In the standard classification system for ARIMA models, the third digit is the number of Moving Average terms. Thus, a (0,1,1) ARIMA model combines Moving Average and a lag term without the use of an autoregressive component. While (1,1,0) ARIMA models tend to work best when there is a smooth change over time, moving average models often are better suited to data with chaotic fluctuations. Following the failure of the (1,1,0) gas price model in the earlier example, a (0,1,1) model to forecast gas prices was successfully implemented in the summer of 2008. At present, both the (1,1,0) and (0,1,1) ARIMA models are in use, each making their own forecast for gas prices.

A simple moving average ARIMA model can be created in Base SAS. The first step is to create a moving average with an appropriate number of points. This code develops a moving average routine using code from the SAS institute to create a moving average using eleven points.

```
**** MOVING AVERAGE MACRO - USES MOVING AVERAGE CODE FROM THE SAS INSTITUTE ****;

DATA TEMP;
    DO X=1 TO 4 BY 0.1;  SUMX+X;  OUTPUT;  END;
 RUN;


DATA NEW;
   IF _N_=1 THEN DO;
      DO N=1 TO 11;
         SET TEMP;  AVERAGE=SUMX/N;  OUTPUT;  END;  END;
   ELSE DO;
      MERGE TEMP TEMP(FIRSTOBS=12 RENAME=(SUMX=SUMX2));
      IF SUMX2 ^= . ;
      AVERAGE=(SUMX2-SUMX)/11;  OUTPUT;  END;
RUN;
```

Next, this moving average routine is used as part of a macro for creating a (0,1,1) ARIMA model in base SAS. The number of successive points to be used in the moving average is a parameter in the macro.

```
**** (0,1,1) ARIMA MODEL IN BASE SAS ****;

**** DAVID J CORLISS, UNIVERSITY OF TOLEDO DEPT. OF PHYSICS AND ASTRONOMY, 2009 ****;

DATA WORK.TSERIES;
   DO X=1 TO 15 BY 0.1;  OUTPUT;  END;
RUN;



   %MACRO MA(N);

   DATA WORK.TEMP;
      SET WORK.TSERIES;
      SUMX+X;
   RUN;

   DATA WORK.TREND;
   %LET M = %EVAL(&N+1);
      IF _N_=1 THEN DO;
         DO N=1 TO &N.;
            SET WORK.TEMP;  AVERAGE = SUMX / N;
         OUTPUT;  END;  END;
      ELSE DO;
         MERGE WORK.TEMP WORK.TEMP(FIRSTOBS=&M. RENAME=(SUMX=SUMX2));
         IF SUMX2 ^= . ;
         AVERAGE = (SUMX2 - SUMX) / &N.;
      OUTPUT;  END;
      MA_TERM = AVERAGE;
      RETAIN MA_TERM;
   RUN;

   DATA WORK.TREND;
      SET WORK.TREND;
      DUMMY = 1;
      RESIDUAL = MA_TERM - X;
      FORECAST = MA_TERM - RESIDUAL;
   RUN;

   DATA WORK.NEW;
      SET WORK.TREND;
      BY DUMMY;
      IF LAST.DUMMY;
      X = X - (RESIDUAL / ((&N. + 1) / 2));
      KEEP X;
   RUN;

   DATA WORK.TSERIES;
      SET WORK.TSERIES WORK.NEW;
      KEEP X;
   RUN;

   %MEND MA;
```

The next example applies this routine to the Cryer milk data. The data are highly periodic, varying over a period of 12 months. By selecting 12 as the number of points to be included in the moving average, seasonality is eliminated. This reveals variations in the long-term trend previously obscured by seasonal variation (e.g., the above-average performance in 1972).

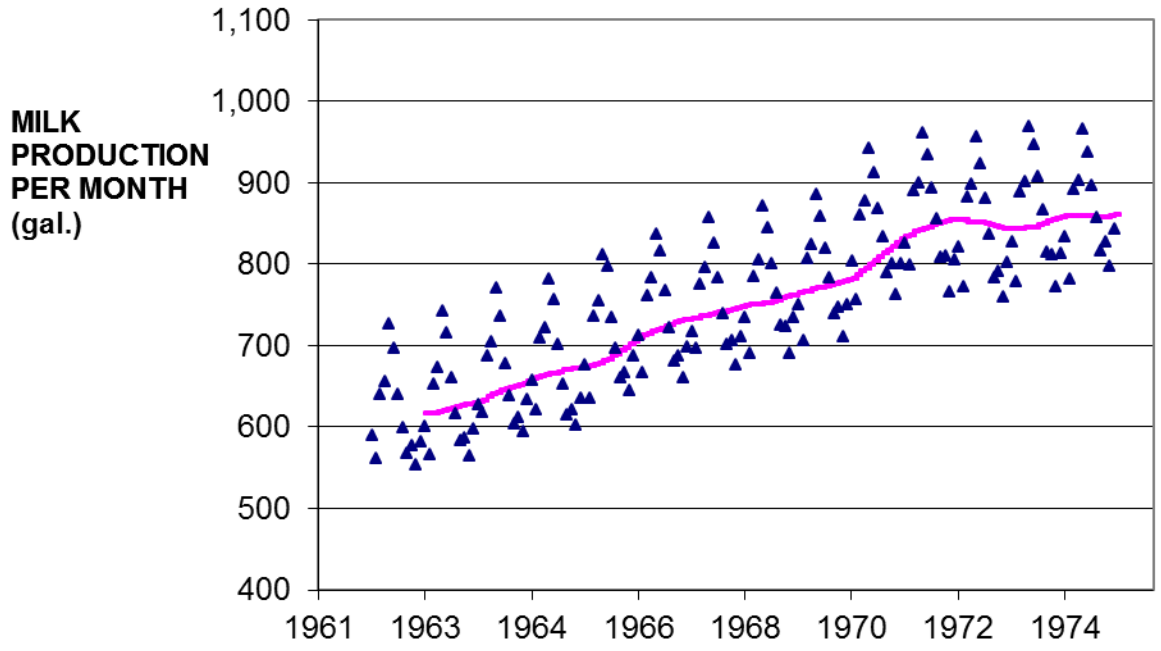

Figure 2. Cryer Milk Data (Cryer, J.D., Time Series Analysis, Duxbury Press, Belmont, 1986, p. 269). The observed values from the Cryer data are marked with blue triangles; a (0,1,1) ARIMA model (pink line) plots the long-term trend corrected for seasonal variation

NON-TEMPORAL ARIMA MODELS

In these traditional examples of ARIMA models, time is used as the independent variable. However, the mathematical structure of these models makes no such requirement: only even-spaced intervals in the independent variable is necessary. This allows for application of ARIMA models in many situations with evenly spaced intervals and a known relationship between the data points. An application in astrostatistics is found in the investigation of a population anomaly in White Dwarf stars known as the DB Gap.

ASTROSTATISTICS – AN ANOMALY IN A POPULATION DISTRIBUTION

Type DA white dwarfs are found with surface temperatures over a wide range of values up to 50,000 K. By contrast, type DB white dwarfs are not found at surface temperatures above 30,000 K, even though DA stars are common in this range. A (1,1,0) non-temporal ARIMA model was developed using temperature as the independent variable. The requirement of evenly-spaced intervals was met by binning of stars in 1,000 degree ranges. Two models were built: one for the DA type stars, which are not subject to the population anomaly, and a second model for the DB type stars. Counts were expressed as the % of the total population found in each 1,000 degree bin. The difference between the two model lines indicates the deficiency in the DB population in the “Gap”. This research concluded that the DB Gap includes a reduction in the population of stars beginning near 25,000 K, some 5,000 K lower than previously observed. This lowered limit can have a significant impact on our understanding of the physical process underlying the existence of this population anomaly.

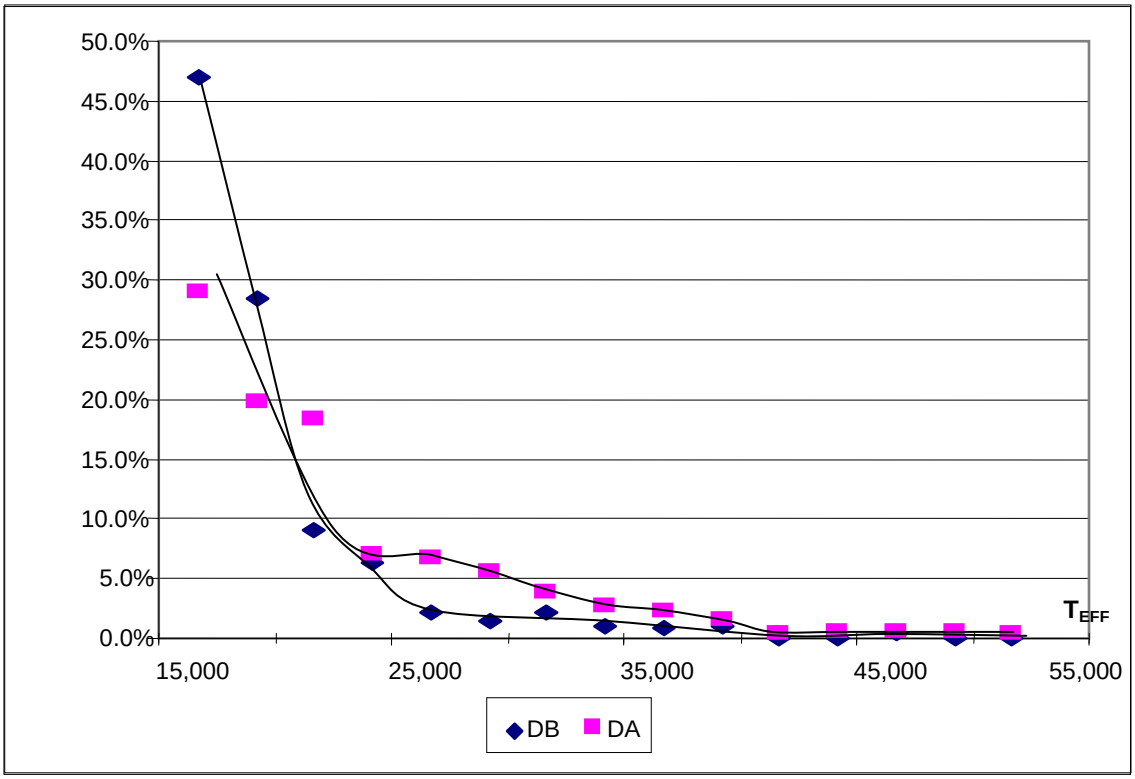

Figure 3. ARIMA models of the population distributions of DA and DB white dwarfs in the Sloan Digital Sky Survey. Data from Eisenstein et al. 2006. The "DB Gap" anomaly is indicated by the difference between the two distributions. As population by temperature is largely a function of time for these stars, temperature is found to be a time proxy.

The physics of white dwarf stars requires that they cool slowly and monotonically, with no additions and few eliminations of individuals over the range of temperatures where the DB Gap is found. As a result, the individuals found in one temperature range constitute a cohort that, with few exceptions, will be found at other temperature ranges at later points in time. The existence of a monotonic relationship between time and temperature in white dwarf stars identifies temperature as a *time proxy*. Time proxy ARIMA models can be developed whenever time can be expressed as a function of the proxy. Examples of time proxies can include cumulative bond yields and outstanding debt on loans, population counts in unrestricted growth such as in the early stages of an infection and the percent demographic composition of a large population where the rates of change are well known. While technically non-temporal, the relationship between time and the proxy can be used to transform a non-temporal ARIMA into a temporal one.

A GEOGRAPHIC EXAMPLE

The plot below shows the results of a non-temporal ARIMA model of the number of attorneys per capita by longitude. As substantial variability is seen between successive points, a moving average model has been used. The trend clearly rises near the coasts, especially the east coast, possibly reflecting a greater degree of litigiousness.

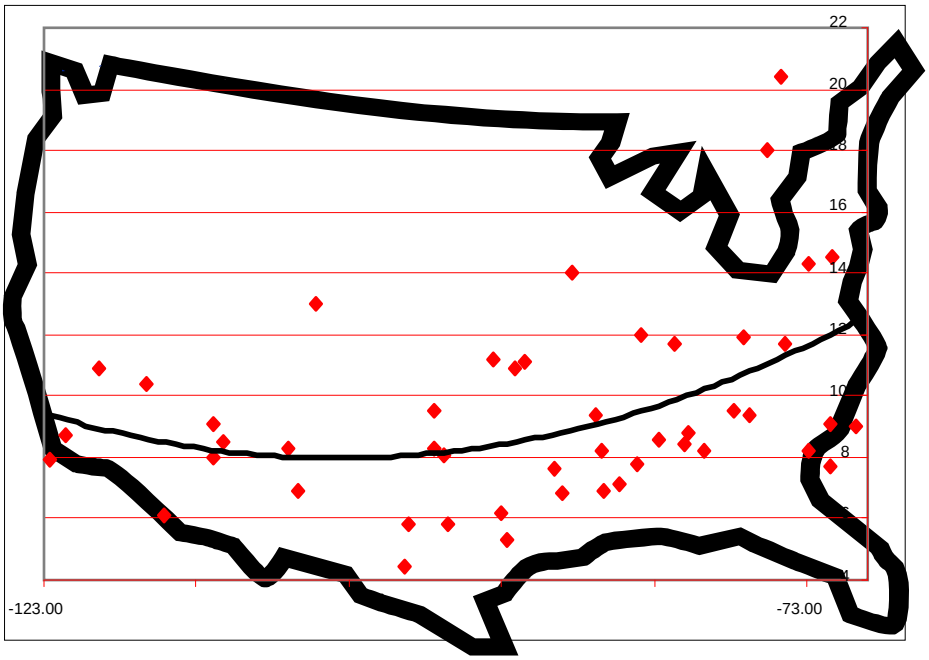

Figure 4. ARIMA model of the distribution of attorneys per capita by longitude.

A DEMOGRAPHIC EXAMPLE

Policies implemented by governmental agencies can display some interesting mathematical properties. As governments will often seek answers to policy questions by examining the successful practices of other governments. While practices are likely to be copied from places with similar socioeconomic profiles, geography will often play a role. A large eastern US city such as Philadelphia, for example, is more likely to use Baltimore as a success story than, say, Prague despite its greater economic similarity. The tendency for governments to model policies and practices after nearby locations with similar economics results in the effects of these polices showing economic and geographic correlation.

In the example, below, the success rates on a state-wide standardized test for a number of public school districts in metropolitan Detroit are compared by per capita income. While it is always possible to draw a regression line through a group of points, it is not always desirable to do so. In this case, the school districts in question often borrow teaching methodologies, text book recommendations and test-taking strategies from nearby districts with similar per capita incomes. Given the significant relationships between many of the factors driving test outcomes between successive data points, a non-temporal ARIMA model is recommended. In this model, the requirement of evenly spaced intervals in the independent variable is met by binning school districts together in $1,000 ranges in per capita income. As the data are fairly smooth and trends are clear, a (1,1,0) ARIMA model is used. In this analysis, the values level off above a certain value: given the maximum possible success rate of 100%, the upward trend is eventually broken. Accordingly, the model only includes points where the success rate increases monotonically with per capita income.

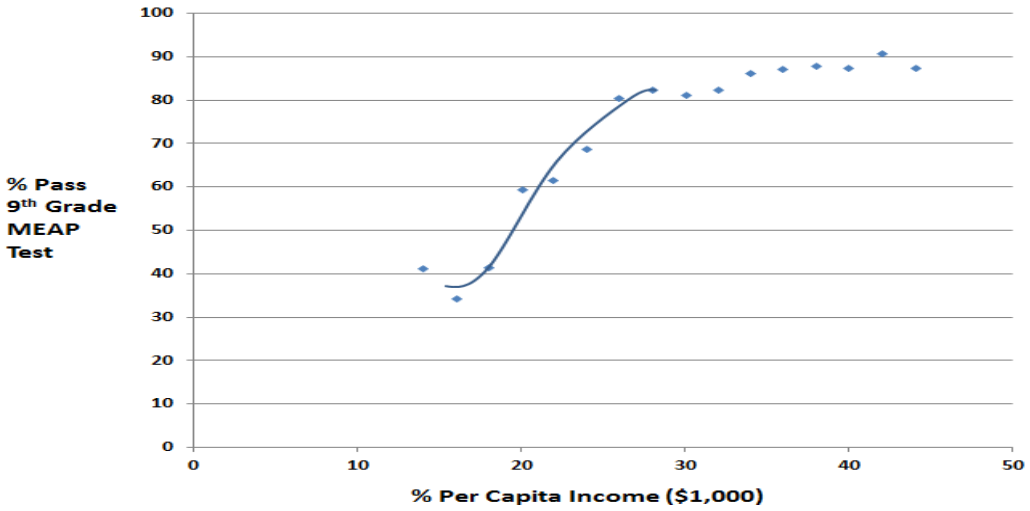

Figure 6. ARIMA model of the relationship between per capita income and educational performance on a state standardized test in metropolitan Detroit school districts.

MIXED ARIMA MODELS

Once it is understood that the independent variable only needs to be evenly spaced – that it doesn't have to be time – it becomes clear that it does not need to be a single thing. Mixed ARIMA models contain multiple components in which at least one is an ARIMA effect. An example of this can be found in an enhancement of the school test score / per capita income model given above. A second, smaller effect is observed in which test scores vary by population density. In this instance, an overall trend is clearly seen but the effect is small in comparison to the residuals. Accordingly, a moving average model is used.

The next example below combines two different types of ARIMA models to establish a trend line: one autoregressive and one a moving average model. In this mixed model, the test scores from the city of Detroit was not used in the development of

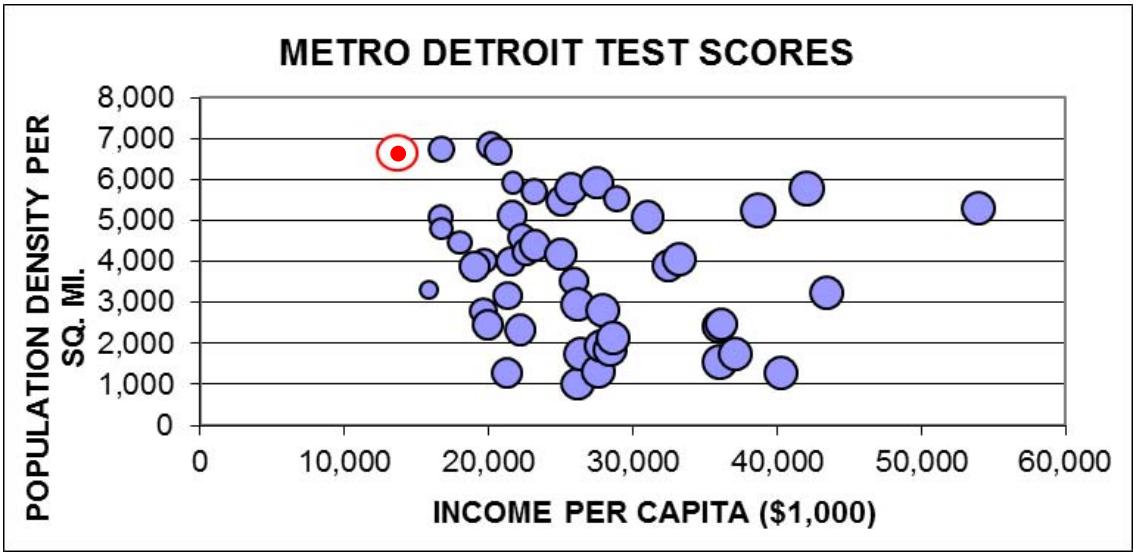

Figure 7. ARIMA model of the relationship between poverty factors and basic educational outcomes in metropolitan Detroit school districts. Detroit, in red, is found to perform above the level predicted by its socio-economic context.

either the autoregressive component, describing the effect of per capita income, or the moving average component for population density. Comparison of the % pass rate predicted by the mixed model for the Detroit Public Schools to the actual values for Detroit provides a measure of the district's performance within the context of surrounding districts with differing socio-economic and demographic characteristics. In the figure, the size of each circle indicates the % pass rate. It will be noted that the circles become larger with increasing annual per capita income up to about $28,000. The slight trend toward lower test scores with increasing population density is also visible. In this plot, the predicted value for Detroit is given by the size of the red dot, while the (much larger) actual value is given by red circle. This analysis yields the surprising result that the performance of the Detroit Public Schools on this particular test, while poor at only 41.2%, is still much better than can be expected (30.1%) given context of the socio-economic challenges the district faces.

CONCLUSION

While ARIMA models are typically used only in time series analysis, the only structural requirement is for evenly-spaced intervals in the independent variable. This allows the development of non-temporal applications of both autogressive and moving average ARIMA models. The requirement of evenly-spaced intervals can be met by binning data into bins of fixed size, supporting ARIMA methodology for many situations where independence of individual data points is not present due to a relationship between successive values in a series. In cases where the independent variable can be described as a function of time, the independent variable may be described as a time proxy. However, the use of ARIMA models is not restricted to cases where such a proxy exists. Non-Temporal ARIMA models can be applied to diverse areas a research including the natural sciences, social sciences, economic and financial applications, among others.

REFERENCES

Box, George and Jenkins, Gwilym, 1970, *Time series analysis: Forecasting and Control*, San Francisco, Holden-Day

Corliss, David J, 2009, Proceedings Midwest SAS User Group Conference

Cryer, J.D., Time Series Analysis, Duxbury Press, Belmont, 1986, p. 269

Eisenstein, D.J., et al., 2006, ApJS, 167, 40

Michigan Educational Assessment Program, Grade 9, Michigan Department of Education, 2008, www.michigan.gov/meap.

United States Census Bureau, 2008, http://factfinder.census.gov

ACKNOWLEDGMENTS

CONTACT INFORMATION

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

Magnify Analytic Solutions
1 Kennedy Square, Suite 500
Detroit, MI 48224
Phone: 313.202.6323
Email: dcorliss@magnifyas.com