Abstract
Autocorrelation-integrated Moving Average (ARIMA) approaches to time series analysis are increasingly used in education. The purpose of this paper is to illustrate the use this approach using real high school attendance data from one urban high school. The basic concepts and assumptions of ARIMA are discussed, as well as the data diagnostics, residual analysis, model selection and interpretation of parameter estimates. The analysis of this particular school reveals a strong weekly cyclical pattern and a heavy skewness in the series due to occasional instances of very low attendance. Most of these trends would not have been captured by traditional models (central tendency, variability, OLS). Implications of the findings and the potential for ARIMA for the education and related disciplines are further discussed.

Rationale for the Study
• Write common in other disciplines, e.g., econometrics (interest rates), medicine (heart beats) and meteorology (weather prediction), ARIMA is rarely used in education.
• Urban school districts that publicize daily attendance rates for their schools have created an opportunity to conduct these types of analyses and explore ARIMA’s utility for educational research.

Steps in the ARIMA Analysis
1. Initial Exploration of the Data
2. Testing for Stationarity of the Time Series
3. Examination of Correlated Error Patterns
4. Identification of Seasonal Cycles
5. Estimating the Impact of Outlying Values on the Series
6. Selection of Best-Fitting Model and Interpretation of its Parameters
7. Analysis of the Residuals

Basic ARIMA Concepts
The Autoregressive (AR) Process
\[ Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + \epsilon_t \]
This model estimates \( Y_t \) using \( p \) lags, with a parameter \( \phi \) associated with each past value included in the model.

The Moving Average (MA) Process
\[ Y_t = \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \ldots + \theta_q \epsilon_{t-q} \]
A moving average process describes time dependencies as a pattern of disturbances, where the estimation \( \theta \) is based on the weighted sum of previous and current errors in the series. A parameter \( \theta \) is associated with each innovation included in the model.

Seasonal AR and MA Processes
A cyclical pattern at lag = 1 is estimated, i.e., five days in the school week:
\[ Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + \epsilon_t \]
The Autocorrelation Function (ACF)
The autocorrelation function for lag k is defined as:
\[ r_k = \frac{\sum (Y_t - \bar{Y})(Y_{t+k} - \bar{Y})}{\sum (Y_t - \bar{Y})^2} \]
The Partial Autocorrelation Function
The partial correlation function for lag k estimates the association between \( Y_t \) and \( Y_{t+k} \) after removing the effects of the effects of the lags in-between.
\[ \hat{\rho}_k = r_k / \sqrt{r_0 \cdot r_k} \]

Outliers
Outliers are additive if:
\[ Y_t = Y' + \omega_t \]
Outliers are innovative if:
\[ Y_t = Y' + \omega_t | Y_{t-1}, \ldots, Y_{t-q} \]

Table 1. Summary Statistics: School Characteristics and Daily Attendance

<table>
<thead>
<tr>
<th>School Characteristics</th>
<th>Daily Attendance</th>
<th></th>
</tr>
</thead>
</table>
| % Male                  | 52%             | 52%
| % Urban                | 80%             | 80%
| % Public               | 50%             | 50%
| % Attendance Rate      | 80%             | 80%
| % Participation Rate   | 90%             | 90%
| % Graduation Rate      | 70%             | 70%
| % Transfer             | 10%             | 10%
| % Non-English Speakers | 20%             | 20%

Table 2. Model Selection Process: Coefficients and Goodness of Fit Statistics

<table>
<thead>
<tr>
<th>Model</th>
<th>Parameter</th>
<th>Pulse</th>
<th>Residuals</th>
<th>AIC (1-L)</th>
<th>AIC (1-L, X)</th>
<th>BIC (1-L)</th>
<th>BIC (1-L, X)</th>
<th>AIC (p-L)</th>
<th>AIC (p-L, X)</th>
<th>BIC (p-L)</th>
<th>BIC (p-L, X)</th>
</tr>
</thead>
<tbody>
<tr>
<td>V</td>
<td>n</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>V</td>
<td>n</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>V</td>
<td>n</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>V</td>
<td>n</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>V</td>
<td>n</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
</tbody>
</table>

Summary of the Results
• Attendance rates are correlated with immediately preceding rates, but we need to model the extreme values in order to be able to detect this;
• There is a strong cyclical pattern at lag = 5 (weekly);
• Outlying values produce heavy skewness due to incidental occurrences of low attendance;
• Traditional summary statistics (mean, variance, OLS) would not reveal the time dependency of these trends;
• The time factor is therefore important and needs to be modeled.

Discussion
• To learn more about the causal effects underlying attendance rates, cross-sectional and longitudinal approaches provide complementary information.
• Understanding the time dependencies in daily high school attendance is helpful to practitioners and policy makers in education.
• More research is needed in this area so that we can:
  – estimate long-term dependencies (e.g., self-organized criticality);
  – compare a larger number of schools on the dynamical features of their attendance rates;
  – Link school-wide daily attendance patterns to the trauma and dropout behavior of individual students.

References