5-2014

Forecasting Monthly Incidence Rates for Shigellosis in Tennessee

Nancy Murray
nmurray2@utk.edu

FORECASTING MONTHLY INCIDENCE RATES FOR SHIGELLOSIS IN TENNESSEE

Nancy Murray

University of Tennessee, Knoxville
Chancellor's Honors Program

Honors Thesis Project
Faculty Advisor: Dr. William Seaver

May 2014

ABSTRACT

Multiple traditional time-series forecasting models were applied to statewide, monthly data for shigellosis with the goal of forecasting the incidence rate of this disease in Tennessee. The dataset begins in 1995 and ends in 2012 for a total of eighteen years, or 216 months. The year 2012 was used to validate the model results. Forecasting models used include time-series regression, exponential smoothing, decomposition, Box-Jenkins autoregressive integrated moving averages, and dynamic linear regression. The coefficient of determination of the training and validation sets was the main evaluation fit statistic for preliminary results. None of the traditional models fits the data well, and more advanced methods that better respond to cyclical elements will need to be utilized in order to properly forecast the incidence rate of shigellosis. Developing a forecasting model for this infectious disease will aid the public health sector in predicting the severity of an outbreak and will allow for the preparation of an intervention for a disease outbreak.

INTRODUCTION

Shigellosis is an infectious disease caused by a group of bacteria called *Shigella* [1]. Four different subgroups of this species include *S. flexneri*, *S. dysenteriae*, *S. sonnei*, and *S. boydii* [2]. Diarrhea, fever, nausea, and vomiting may accompany this disease [2]. *Shigella* typically lingers on hands that have not been adequately washed and is commonly transmitted fecal-orally [1]. About 14,000 cases of the disease are reported in the United States yearly, and about 72% of those cases are caused by the subgroup *S. sonnei* [3]. However, many cases of shigellosis go unreported [3]. Cases are confirmed through testing of the feces in order to isolate the *Shigella* bacteria [4]. People who have the aforementioned symptoms may forgo a visit to a doctor's office, and then the case cannot be filed; this affects the data and the extent to which researchers can generalize conclusions. Another factor with the shigellosis disease is its high incidence in children. The Centers for Disease Control and Prevention (CDC) reports that children aged two to four years old are more likely to contract shigellosis [1]. This is likely due to poor hand washing and less awareness of personal health in general.

Due to this disease's temporarily debilitating effects and spread in children, we wanted to focus locally on shigellosis in Tennessee for this forecasting project. Overall, we would like to be able to forecast the incidence rate of each month in order to be prepared for an outbreak. An

element of other diseases is monthly seasonality, especially during the summer months when the heat affects the bacteria. Since children in Tennessee typically start school in August, we want to see if some seasonal effect exists. Time-series forecasting models will allow us to determine seasonality and incidence rates for the next year. We look at traditional forecasting methods first, and then we modify our approach to account for both cyclical and seasonal patterns in the data.

MATERIALS AND METHODS

Data source

The Tennessee Department of Health's online data reporting service, Communicable Disease Interactive Data, provides counts of shigellosis for the state of Tennessee by month for years 1995 to 2012.

In order for a case to be reported, an individual with shigellosis must visit a healthcare provider and pass a stool sample that tests positive for *Shigella* bacteria [4]. As a reportable disease, shigellosis must be reported to the Tennessee Department of Health if *Shigella sp.* is isolated from a clinical specimen, if *Shigella sp.* is detected using non-culture based methods, or if a person with diarrhea has been in contact with someone with a confirmed *Shigella* infection [4]. Due to the high level of underreporting in shigellosis cases, the CDC estimates that the actual number of cases may be as much as twenty times greater than reported, on a national level [1].

Data preparation

Data from years 1995 to 2012 provide us with a total of 216 data points. We converted the count data to incidence rates per 100,000 people by using the annual Tennessee population estimate data from the United States Census Bureau [5]. These conversions were performed with Excel. An incidence rate of one can be interpreted as one person having shigellosis per 100,000 people, for example. Incidence rates, as opposed to counts, help the data be more understandable on a population level. The monthly incidence rate per 100,000 people for shigellosis varies in Tennessee from as low as 0.05 to as high as 5.01.

In order to find a forecasting model, we split the data into two sets, the training set and the validation set. The training set has 204 data points for monthly incidence rates from 1995-2011, and the validation set has 12 points for monthly incidence rates in 2012. The validation set may

also be referred to as the holdout set. We want a model that forecasts into the future opposed to only fitting the past data. The holdout sample is essential in forecasting because it validates the effectiveness of the forecasting model constructed with the training data [6].

We performed most statistical analyses in NCSS 8, unless otherwise noted [7]. The forecasting methods used include time-series regression, Holt-Winter's exponential smoothing, decomposition, Box-Jenkins autoregressive integrated moving averages (ARIMA), and dynamic linear regression in varying forms. We used R^2, the coefficient of determination, to determine the adequacy of each model according to the training set and the validation set.

Pattern analysis and outlier identification

In order to get an initial idea of the data, a few analyses must be performed. First, it is critical to look at the monthly incidence rate over time. Figure 1, displayed below, illustrates the cyclical pattern of the data but no definite seasonal aspects. Figure 2 and Figure 3 display the training set and the validation set over time. Figure 4 allows us to see if any months are consistently having higher incidence rates over time; we can see that no months have the highest incidence rates on a year-to-year basis.

Figure 5 allows us to examine the seasonal effect more closely. Many months have outliers, but June, September, October, and November all have what we consider to be extreme outliers. We also perform a Kruskal-Wallis One Way ANOVA on the monthly incidence rates to identify equal medians, and we accept the null hypothesis that the medians are equal ($p > 0.05$). Therefore, our monthly seasonal impact hypothesis cannot be proven by this data; however, since we are aware of such rampant underreporting of this disease, we would still like to incorporate some sort of seasonal element into our forecasting models.

Outliers may affect the data and its analysis. It is important to identify these outliers at an early stage. Figure 6 contains the same time plot as Figure 2 but also contains the mean of the training set as the red line and one, two, and three standard deviations above the mean as the blue lines. The two data points in October and November of 1998 are well above three standard deviations from the mean. We were cognizant of these points throughout our analysis, but we could not remove them since they do provide insight for the cycles and the seasons of the data.

Now that we have determined the initial observations of the data, we can proceed to time-series modeling.

Time-series forecasting models

All time-series forecasting models tested had to meet certain criteria to be considered a good model. The initial statistic evaluated for each model was the coefficient of determination, or R^2 statistic, on the training data. The R^2 for the validation set was evaluated second. R^2 can be presented as either a proportion or percentage; we will use the former. R^2 ranges from 0 to 1, with 1 being the most variation explained by the model. Equation (1) and equation (2) provide computation of R^2 for both the training and validation sets, respectively:

$$R^2_{\text{Prediction}} = 1 - \frac{\sum_{i=1}^{n}[y_i - \hat{y}_{(-i)}]^2}{\sum_{i=1}^{n}[y_i - \bar{y}]^2} \quad (1)$$

$$R^2_{\text{Holdout}} = 1 - \frac{\sum_{i=1}^{n2}[y_i - \hat{y}_{(i)}]^2}{\sum_{i=1}^{n2}[y_i - \bar{y}]^2} \quad (2)$$

Typically, if the R^2 prediction and R^2 holdout statistics are close in number, then the model is considered adequate, and we can then perform residual analysis [8]. Various forecasting techniques were utilized throughout the duration of this project including time-series regression, Winters Exponential Smoothing, decomposition, ARIMA, and dynamic linear regression, and the R^2 values were the initial fit statistics for these models.

We examined residual plots and normality using Shapiro-Wilk's goodness-of-fit for normality. We also examined the existence of white noise in the residuals using the Q-statistic [9]. When no pattern exists in the residuals, the model does not need further improvement.

In addition to the traditional models, we attempted combination models with the top performing models. We used the simple combination method that averages the forecasts from each model. The three simple combination models with averages of the forecasts are Winters exponential smoothing and ARIMA(1,0,1), Winters exponential smoothing and decomposition, and decomposition and ARIMA(1,0,1).

After evaluating traditional forecasting models and combination models, we smoothed the two high outliers. We perform decomposition, dynamic linear regression, and moving origin, fixed horizon decomposition models on the smoothed data.

Time-series regression

Regression models allow us to use trend and seasonal elements. We utilized ordinary least squares methods in the NCSS macro for multiple regression [7]. After investigating trend only models and trend plus seasonal models, the results were better with the seasonal element. Due to interpretation issues, we excluded any interaction variables from the model. The base month for the seasonal variables is October. We used additive time-series regression and multiplicative (log-transformed) time-series regression. The formulas for the additive and multiplicative time-series regression models are outlined in equation (3) and (4), respectively.

$$Y_t = \beta_0 + \beta_1 x_{trend} + \sum_{i=2}^{p=12} \beta_i d_i + \varepsilon_t \qquad (3)$$

$$Y_t = \beta_0 + \beta_1 x_{trend} + \prod_{i=2}^{p=12} \beta_i d_i + \varepsilon_t \qquad (4)$$

Holt-Winters exponential smoothing

Holt-Winters exponential smoothing, or simply Winters exponential smoothing, utilizes smoothing constants along with seasonal parameters. NCSS has a Winters exponential smoothing macro [7]. The model that best fit the shigellosis data had an additive trend and multiplicative seasonality. Winters exponential smoothing uses the equations (5) through (8) below:

$$S_T = \alpha \frac{Y_t}{I_{t-L}} + (1-\alpha)(S_{t-1} + b_{t-1}) \qquad (5)$$

$$b_t = \beta(S_t - S_{t-1}) + (1-\beta)b_{t-1} \qquad (6)$$

$$I_t = \gamma \frac{Y_t}{S_t} + (1-\gamma)I_{t-L} \qquad (7)$$

$$\hat{Y}_{t+m} = (S_t + b_t m)I_{t-L+m} \qquad (8)$$

where α, β, γ are smoothing constants, and we account for trend, slope, and seasonality. S_t is the smoothed value at end of *t* after adjusting for seasonality, b_t is the smoothed value of trend through period *t*, and I_t is the smoothed seasonal index at end of period *t* [10].

Decomposition

Various decomposition methods were utilized throughout this project. NCSS has a macro for decomposition that performs automatic multiplicative decomposition with stable seasonal variation [7]. Due to the nature of decomposition, one is also able to perform decomposition methods within the spreadsheet feature of NCSS with either stable seasonal variation or changing seasonal variation. The basic formula for decomposition is outlined below in equation (9):

$$Y_t = T_t C_t S_t E_t \qquad (9)$$

Equation (9) displays the trend, cyclical, seasonal, and error components, respectively. The nature of this model allows us to isolate these elements for forecasts when necessary.

Box-Jenkins autoregressive integrated moving averages (ARIMA)

Various ARIMA models were tested in this project. ARIMA models are set-up as follows.

$$ARIMA(p, d, q)(P, D, Q) \qquad (10)$$

Equation (10) encompasses seasonal ARIMA. The elements include p as the number of autoregressive terms, d as the number of nonseasonal differences, q as the number of lagged forecast errors in the prediction equation (moving average), P as the number of seasonal autoregressive terms, D as the number of seasonal differences, and Q as the number of seasonal moving average terms [11].

We ran a myriad of ARIMA models in NCSS with autoregressive and moving averages terms up to two and differencing as zero or one. The best model, ARIMA(1,0,1), is represented by the backshift operator equation (11):

$$(1 - \phi_1 B) y_t = (1 - \theta_1 B) e_t \qquad (11)$$

Dynamic linear regression

SAS software allows users to perform dynamic linear regression through the Time Series Forecasting System macro [12]. Dynamic linear regression is a type of transfer function, another model type attributed to Box and Jenkins [13]. Transfer functions are represented by equation (12):

$$Y_t = a + \sum_{i=0}^{k} v_i x_{t-i} + N_t \qquad (12)$$

Dynamic regression utilizes trend elements along with other predictor elements like ARIMA and cyclical components, making it a possible model for this time-series data.

RESULTS

Pattern analysis

Another preliminary step before analyzing the data involves examining autocorrelation (ACF) and partial autocorrelation (PACF) plots [14]. Figure 7 shows the autocorrelation plot for the Tennessee shigellosis incidence rate data; the slightly exponential pattern here demonstrates a need for a model with an autoregressive process of order 2. Figure 8, the partial autocorrelation plot, has one large, significant spike at the beginning of the series, indicating a need for a model with an autoregressive process of order 1. Both of these plots confirm that some sort of time dependency exists in the data that can be accounted for by lags.

Outlier identification and smoothing

Since the outliers in 1998 are large, we smooth them down to better reflect the peak of the typical cycle. The highest data point is now at three standard deviations away from the training dataset's mean with the second highest data point slightly below three standard deviations, as seen in Figure 9. We perform decomposition, dynamic linear regression, and moving origin, fixed horizon decomposition models on the smoothed data. All of these models allow for cyclical and seasonal elements, but we are still unable to achieve preferred model performance since the R^2 training and R^2 holdout are not close, as seen in Table 1. Therefore, the smoothing of the outliers is unnecessary.

Time-series model results

Table 2 provides the summary of the models examined including time-series regression, Winters Exponential Smoothing, decomposition, ARIMA, and dynamic linear regression. Most of the models do not perform well because they are not catching the cyclical nature of the series. None of the R^2 training values is close to its corresponding R^2 holdout value. Additionally, the only model to result in a positive R^2 holdout value is Holt-Winters Exponential Smoothing with an additive trend and multiplicative seasonality.

Three simple combination models with averages of the forecasts for Winters exponential smoothing and ARIMA(1,0,1), Winters exponential smoothing and decomposition, and decomposition and ARIMA(1,0,1) did not result in much better R^2 values for the training set, and all R^2 validation values were negative. The exact R^2 values are in Table 3. This method shows that combining the already poor performing models in the case of this data does not improve the fit.

Holt-Winters Exponential Smoothing

Winters exponential smoothing model with additive trend and multiplicative seasonality is the only forecasting model that has positive R^2 for both the training set and the validation set. With the smoothed data, the holdout fit actually slightly decreases with this model. Therefore, it was in our best interest to keep the original, unmodified data set. Even though the R^2 values are both positive, we ideally would like the two R^2 values to be similar, and 0.597 and 0.258 are not close; the model does not adequately forecast the incidence rates.

Figure 10 and Figure 11 visually show how the Winters exponential smoothing model performs on this Tennessee shigellosis incidence rate data. The Winters exponential smoothing model captures the cycles' peaks and valleys as well as some of the seasonality, but the data and the forecasts do not perfectly align, causing the R^2 values to be lower.

DISCUSSION

Conclusion and Suggestions

Winters exponential smoothing resulted in the best forecasting model for Tennessee's shigellosis data. About 60% of the variation in the shigellosis incidence rates between 1995 and 2011 can be explained by this model, whereas only about 25% of the variation in shigellosis

incidence rates in the holdout sample can be explained by Winters exponential smoothing. Although this forecast provides us some sense of knowledge about the patterns in the data, there is no normality or white noise of residuals, so there is room for improvement in the model. For this reason, we do not recommend using Winters exponential smoothing with additive trend and multiplicative seasonality on Tennessee's shigellosis incidence rates. Instead, more advanced models that better incorporate the cyclical and seasonal elements should be utilized. For instance, Winters exponential smoothing with ARIMA for the cyclical element in the decomposition mode might be an option.

Other disease data are better modeled by more advanced models. Tuberculosis, as well as other diseases, in China, for example, is best modeled by a hybrid model that combines ARIMA models and generalized regression neural network models [15]. Therefore, more advanced models may be necessary for Tennessee's shigellosis data.

Also, a different holdout period, forecast range, or cycle prediction may improve our time-series forecasting models. Other suggestions for improvement include the inclusion of a geographical breakdown by region or county and a breakdown by age since shigellosis is common in young children.

Limitations

All of the data used for this forecasting project had to be laboratory confirmed cases of shigellosis. The Centers for Disease Control and Prevention estimate the actual number of cases to be much higher so our modeling may not reflect the actual incidence rates. Without confirmed cases, records of outbreaks are difficult to pinpoint.

Additionally, we limited ourselves to using NCSS and SAS software packages. More complicated models as well as other packages may have provided more accurate models for Tennessee's disease data.

ACKNOWLEDGEMENTS

Thanks to the Tennessee Department of Health for providing the data from its website.
Also, thanks to Dr. William Seaver for all of his support and advice throughout the analysis and writing of this thesis project.

REFERENCES

1. **Centers for Disease Control and Prevention (CDC).** (http://www.cdc.gov/nczved/divisions/dfbmd/diseases/shigellosis/). Accessed 4 May 2014.

2. **Niyogi SK.** Shigellosis. *The Journal of Microbiology.* 2005; 43(2): 133-143.

3. **Centers for Disease Control and Prevention (CDC).** (http://www.cdc.gov/nczved/divisions/dfbmd/diseases/shigellosis/technical.html). Accessed 4 May 2014.

4. **Tennessee Department of Health.** (http://health.state.tn.us/ReportableDiseases/ReportableDisease.aspx/). Accessed 4 May 2014.

5. **U.S. Census Bureau.** (Intercensal Estimates of the Resident Population for the United States). Accessed 7 Dec 2013.

6. **DeLurgio SA.** *Forecasting Principles and Applications*, 1st edn. Boston, MA: Irwin McGraw-Hill, 1998, pp. 26.

7. **Hintze J.** NCSS, PASS and GESS. In: *NCSS*. Kaysville, Utah, USA, 2012.

8. **Seaver WL.** Multiple Linear Regression (Unpublished Chapter). In: *The Bottom-Line in Business Statistics*. Stamford, CT: Thomson Publishing, 2007, pp. 329.

9. **DeLurgio SA.** *Forecasting Principles and Applications*, 1st edn. Boston, MA: Irwin McGraw-Hill, 1998, pp. 88.

10. **DeLurgio SA.** *Forecasting Principles and Applications*, 1st edn. Boston, MA: Irwin McGraw-Hill, 1998, pp. 224.

11. **Duke University.** (http://people.duke.edu/~rnau/seasarim.htm). Accessed 4 May 2014.

12**. SAS Institute.** Statistical analysis systems (SAS), version 9.3. Cary, North Carolina, USA: SAS Institute Inc., 2010.

13. **SAS Institute Inc**. (http://support.sas.com/documentation/cdl/en/etsug/63348/HTML/default/viewer.htm#etsug_tfprdvar_sect005.htm). Accessed 4 May 2014.

14. **DeLurgio SA.** *Forecasting Principles and Applications*, 1st edn. Boston, MA: Irwin McGraw-Hill, 1998, pp. 67.

15. **Zhang G, *et al*.** Application of a Hybrid Model for Predicting the Incidence of Tuberculosis in Hubei, China. *PLoS ONE*. 2013; 8(11): e80969.

TABLES AND FIGURES

Table 1. Time-series forecasting models and their performance in forecasting smoothed shigellosis incidence rates in Tennessee.

	R^2	R^2 Holdout
Stable Seasonal Automatic Decomposition	0.6754825	-9.0469167
Changing Seasonal Decomposition	0.445713932	-5.3311764
Dynamic Linear Regression with cyclical element, cubic trend, and seasonal dummies	0.632	-2.85

R^2 is the training R^2 ,or the prediction R^2.
R^2 Holdout is the holdout R^2, or the validation R^2 .
Models ideally have similar, positive R^2 values.

Table 2. Time-series forecasting models and their performance in forecasting shigellosis incidence rates in Tennessee.

	R^2	Normality of residuals?	White noise in residuals ?	R^2 Holdout
Multiplicative Regression with Trend and Seasonal Components without Interaction terms	-0.01715	No	No	-2.17487
Additive Regression with Trend and Seasonal Components without Interaction terms	0.0626	No	No	-9.28059
Additive Trend, Multiplicative Seasonality Winter's Exponential Smoothing	0.59733	No	No	0.25794
Stable Seasonal Automatic Decomposition	0.60575	No	No	-7.9688
Changing Seasonal Decomposition	0.38925	No	No	-10.5252
ARIMA(1,0,1)	0.54208	No	Yes	-10.0994
Dynamic Linear Regression with Exponential Trend + ARIMA(1,0,1)s No Intercept	0.63283	No	No	-0.165

R^2 is the training R^2, or the prediction R^2.
Normality of residuals refers to the distribution of the residuals for each model.
White noise in the residuals describes the adequacy of the model fit to the data.
R^2 Holdout is the holdout R^2, or the validation R^2 .
The ideal model will have similar, positive R^2 values, normality of residuals, and white noise.

Table 3. Time-series combination forecasting models and their performance in forecasting shigellosis incidence rates in Tennessee.

	R^2	R^2 Holdout
Winters Exponential Smoothing and ARIMA(1,0,1)	0.593556502	-2.394410276
Winters Exponential Smoothing and Decomposition	0.773730285	-0.446069372
Decomposition and ARIMA(1,0,1)	0.539510984	-5.711060638

R^2 is the training R^2 ,or the prediction R^2.
R^2 Holdout is the holdout R^2, or the validation R^2 .
Models ideally have similar, positive R^2 values.

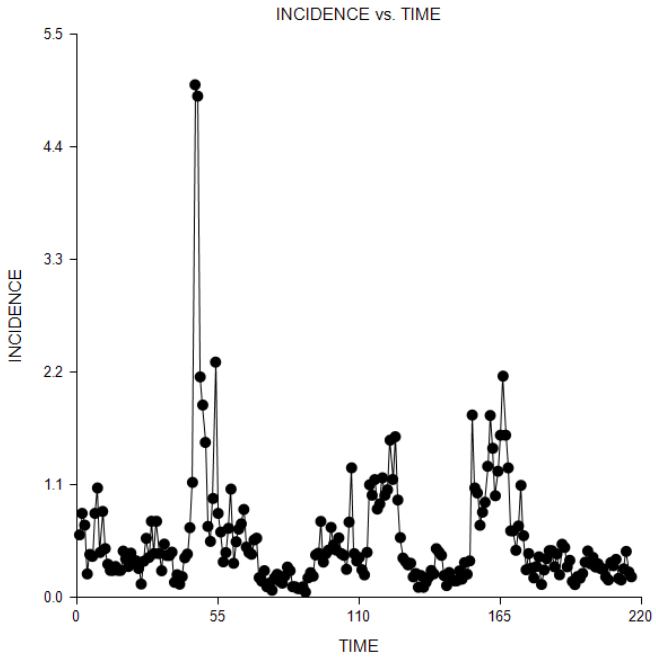

Figure 1. Time-series plot of the monthly incidence rate of shigellosis in Tennessee from 1995 to 2012.

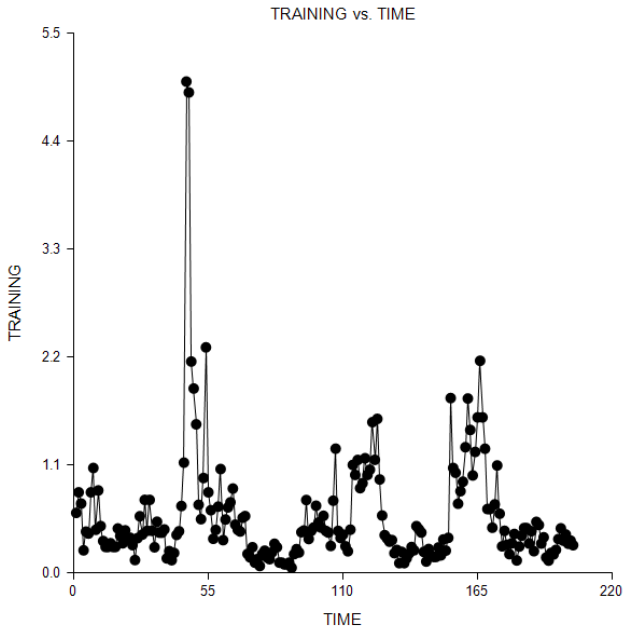

Figure 2. Time-series plot of the monthly incidence rate of shigellosis in Tennessee from 1995 to 2011.

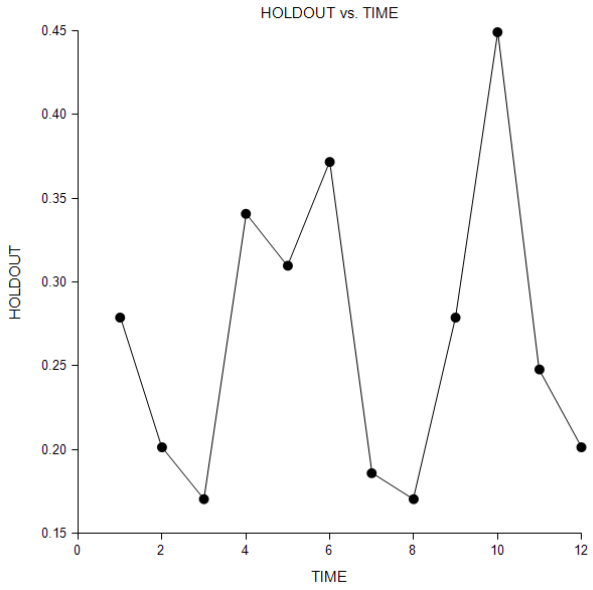

Figure 3. Time-series plot of the monthly incidence rate of shigellosis in Tennessee in 2012.

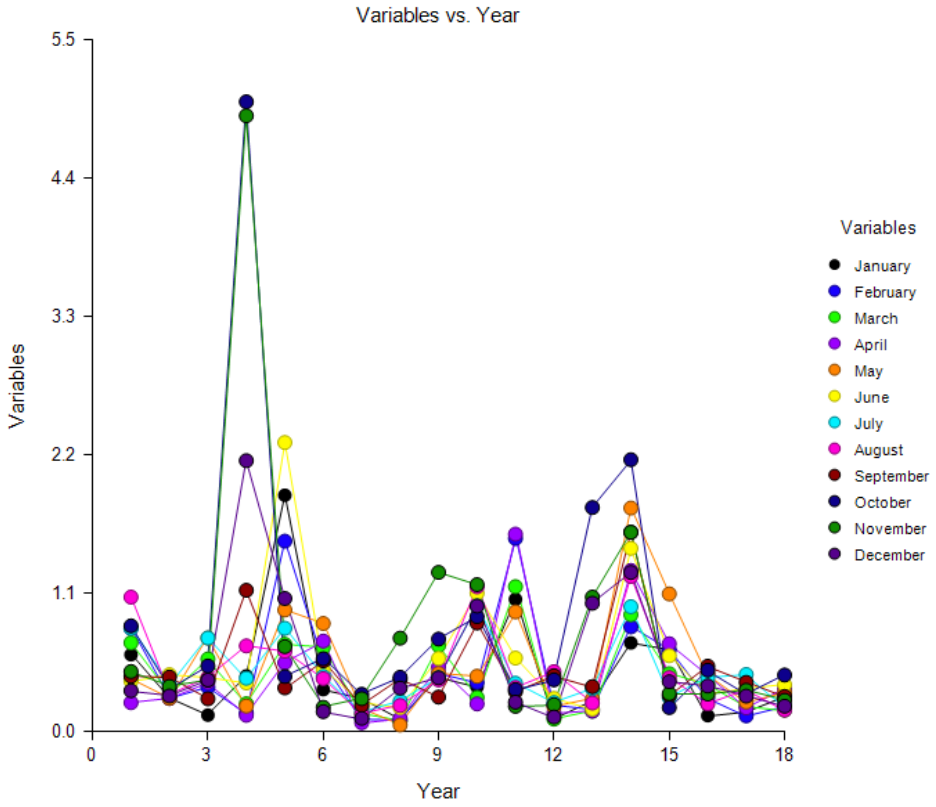

Figure 4. Monthly incidence rate of shigellosis in Tennessee by year from 1995 to 2012.

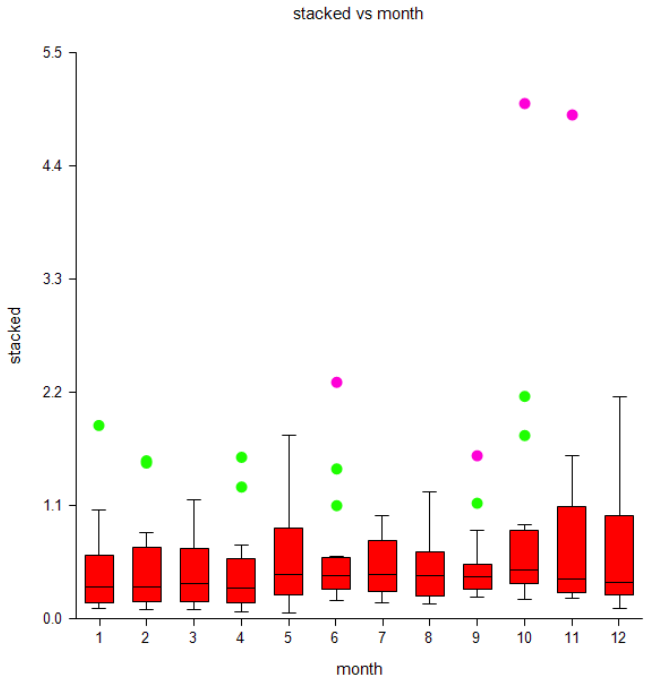

Figure 5. Box plots of monthly incidence rate of shigellosis in Tennessee from 1995 to 2012.

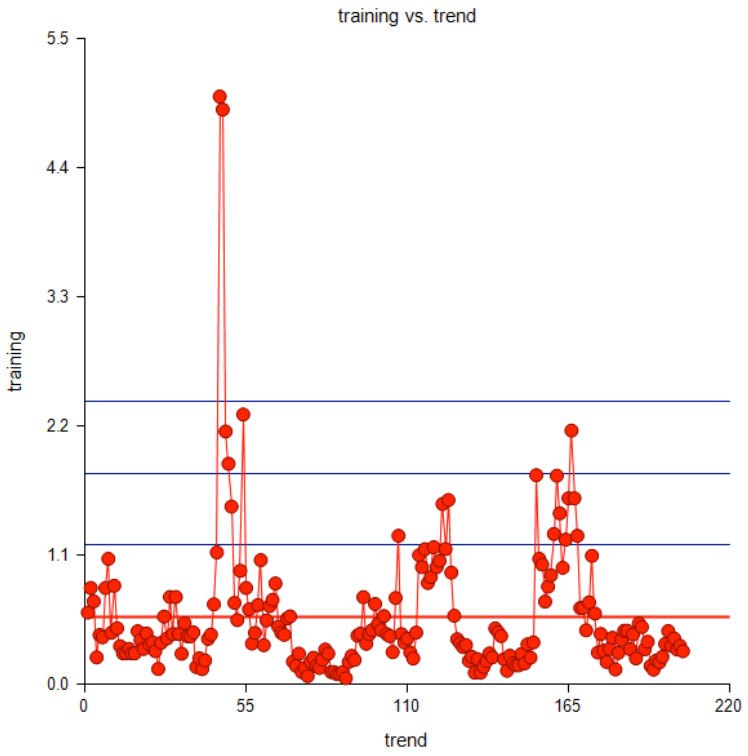

Figure 6. Monthly incidence rate of shigellosis in Tennessee from 1995 to 2011, emphasizing the outliers above 3 standard deviations.

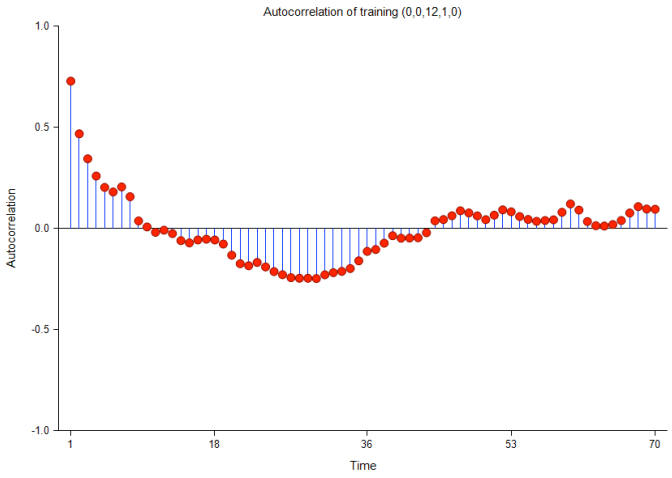

Figure 7. Autocorrelation plot for Tennessee's shigellosis incidence rate per 100,000 people.

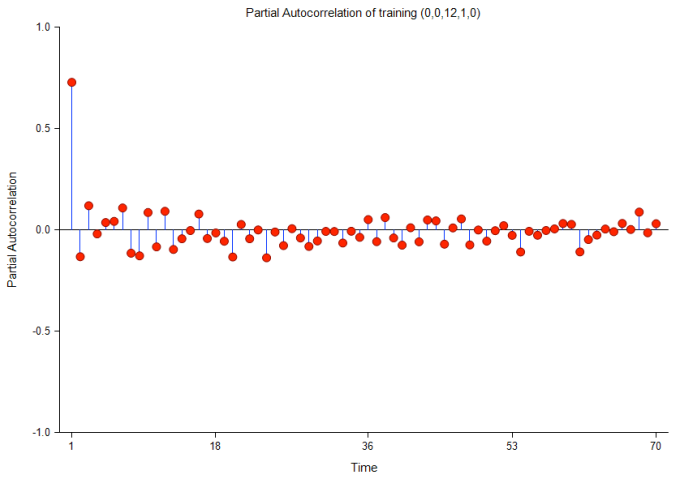

Figure 8. Partial autocorrelation plot for Tennessee's shigellosis incidence rate per 100,000 people.

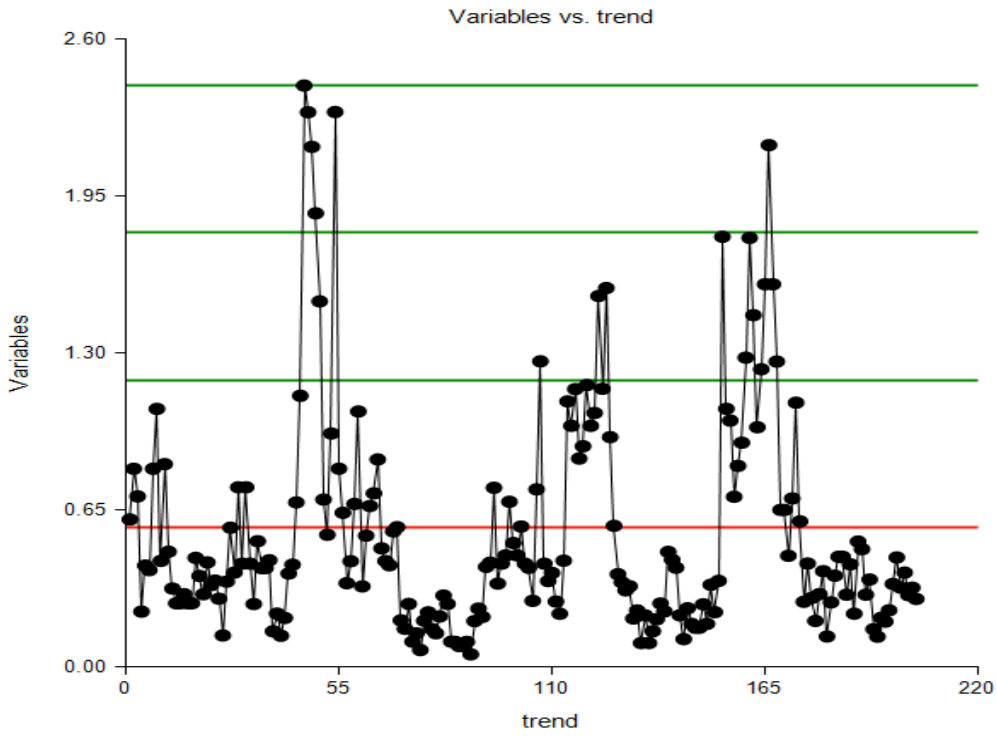

Figure 9. Smoothed time-series plot of monthly incidence rate of shigellosis in Tennessee from 1995 to 2011.

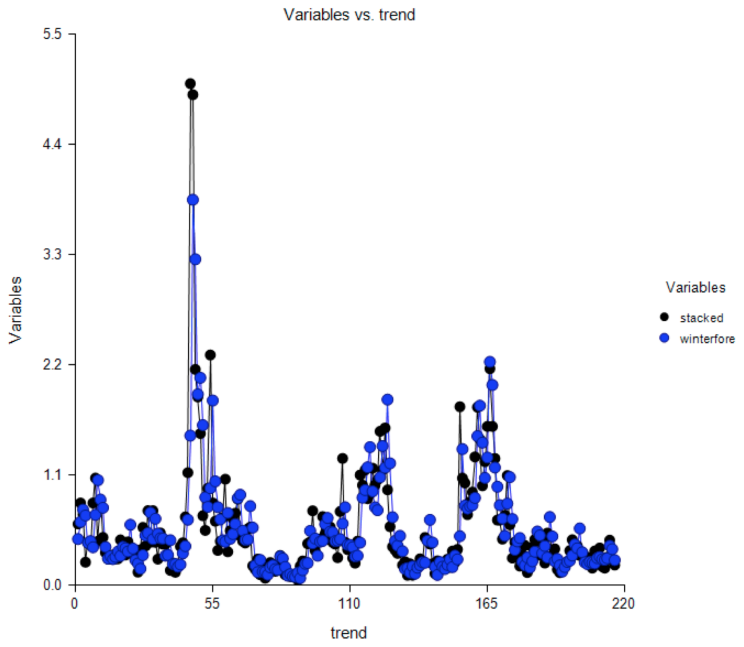

Figure 10. Overlay plot of Winters exponential smoothing model monthly forecasts and the training data for shigellosis incidence rates in Tennessee.

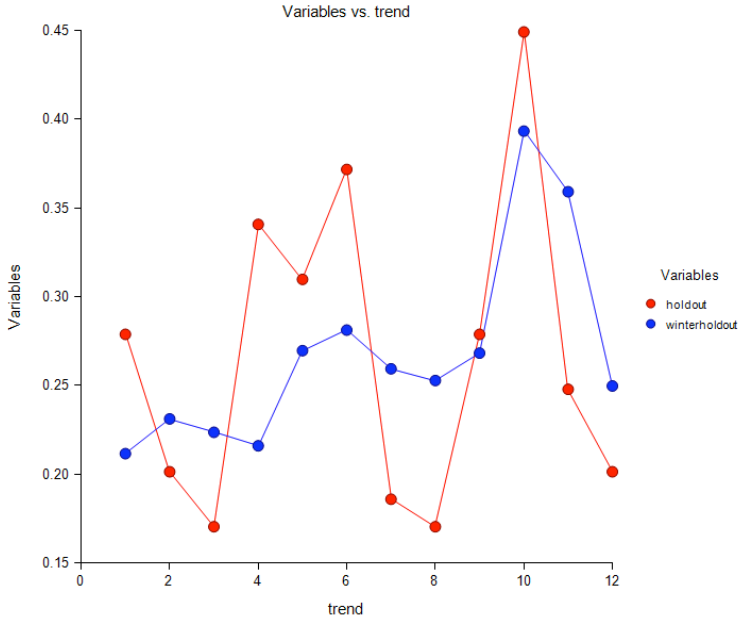

Figure 11. Overlay plot of Winters exponential smoothing model monthly forecasts and the holdout data for shigellosis incidence rates in Tennessee.