# Time Series Analysis OF Sexually Transmitted Disease Morbidity: Selected STDs by Age, Gender, and Race/Ethnicity United States and Puerto Rico 1996 - 2014

#### **OROGUN KOFOWOROLA**

**8942692**

Program of Predictive Analytics  
School of Computer Applications  
Conestoga College

STAT8041 (Statistical Forecasting)  
Individual Project  
Date: 2024/06/21

### INTRODUCTION

Data was obtained from [data source](https://data.world/makeovermonday/2019w31/workspace/project-summary?agentid=makeovermonday&datasetid=2019w31). The data are based on case reports of sexually transmitted diseases submitted by state and local health departments to CDC's [Division of STD/HIV Prevention](http://www.cdc.gov/std/dstdp/). The data are reported through [STD\*MIS](http://www.cdc.gov/std/std-mis/), the [National Electronic Telecommunications System for Surveillance (NETSS)](http://www.cdc.gov/ncphi/disss/nndss/netss.htm), or the [National Electronic Disease Surveillance System (NEDSS)](http://www.cdc.gov/nedss/). This dataset contains Sexually Transmitted Disease (STD) morbidity case reports reported to the National Centre for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention ([NCHHSTP](http://www.cdc.gov/nchhstp/)), Centres for Disease Control and Prevention (CDC) for the 50 United States, the District of Columbia and Puerto Rico. The number of cases and disease incidence rates are reported by age group, race/ethnicity and gender of patient, year, type of STD (Chlamydia, Gonorrhoea, Primary and Secondary Syphilis), and area of report.

**The data shape is 42680, 10.**

The dataset contained these columns 'Disease,' 'Disease Code,' 'State,' 'Year,' 'Gender,' 'Age,' 'Age Code,' 'STD Cases,' 'Population,' and 'Rate per 100K'.

The raw data contained some missing values:  
Disease 150  
Disease Code 150  
State 150  
Year 150  
Gender 150  
Age 150  
Age Code 150  
STD Cases 150  
Population 7197  
Rate per 100K 7197.

**DATA DICTIONARY  
disease:** Chlamydia17,496 (41.14%)Gonorrhea16,154 (37.98%)Primary and Secondary Syphilis8,880 (20.88%)Distinct: 3  
Non-empty: 42,530  
Empty: 0 (0%)  
Min Length: 9  
Max Length: 30  
Mean Length: 13.385  
Most Common: Chlamydia  
Next Most Common: Gonorrhea

**Disease code:**Type: integer  
27417,496 (41.14%)28016,154 (37.98%)3108,880 (20.88%)  
Distinct: 3Non-empty: 42,530  
Empty: 0 (0%)  
Mean: 283.796  
Min: 274  
Max: 310  
Std Dev1: 13.723  
Skewness: 1.296  
Kurtosis: -0.115

**State:**Distinct: 51  
Non-empty: 42,530  
Empty: 0 (0%)  
Min Length: 4  
Max Length: 20  
Mean Length: 8.72  
Most Common: California  
Next Most Common: Texas

**Year:**Distinct: 19  
Non-empty: 42,530  
Empty: 0 (0%)  
**gender:** Male19,462 (45.76%)Female18,700 (43.97%)Unknown4,368 (10.27%)  
Distinct: 3  
Non-empty: 42,530  
Empty: 0 (0%)  
Min Length: 4  
Max Length: 7  
Mean Length: 5.187  
Most Common: Male  
Next Most Common: Female

**age**Distinct: 8  
Non-empty: 42,530  
Empty: 0 (0%)  
Min Length: 7  
Max Length: 11  
Mean Length: 10.332  
Most Common: 20-24 years  
Next Most Common: 25-29 years

**age\_code**Distinct: 8  
Non-empty: 42,530  
Empty: 0 (0%)  
Min Length: 2  
Max Length: 5  
Mean Length: 4.405  
Most Common: 20-24  
Next Most Common: 25-29

**Std\_cases**Distinct: > 1,000  
Non-empty: 42,530  
Empty: 0 (0%)  
Mean: 603.134  
Min: 1  
Max: 46,885  
Std Dev: 1,973.091  
Skewness: 9.136  
Kurtosis: 122.447

**Population**Distinct: > 1,000  
Non-empty: 35,483  
Empty: 7,047 (16.57%)  
Mean: 438,979.404  
Min: 12,937  
Max: 8,880,836  
Std Dev: 760,558.75  
Skewness: 4.827  
Kurtosis: 32.109

**rate\_per\_100k:**how many people in 100k people are infected by a particular disease  
Distinct: > 1,000  
Non-empty: 35,483  
Empty: 7,047 (16.57%)  
Mean: 325.662  
Min: 0.02  
Max: 9,078.95  
Std Dev: 697.545  
Skewness: 3.96  
Kurtosis: 20.22

**CLEANING**For the sake of the project, the dataset had to be divided using only rows with disease as Chlamydia. The new dataset had ZERO NULL values.  
Duplicates were checked for, and there were no duplicate values

**Why I Chose This Dataset:**

1. **Public Health Significance**:
   * Sexually Transmitted Diseases (STDs) pose a significant public health challenge, affecting millions of people annually across various demographics. Understanding the trends and patterns in STD morbidity is crucial for effective public health planning and intervention.
   * This dataset offers detailed insights into the incidence and prevalence of various STDs across different age groups, genders, and racial/ethnic backgrounds, providing a comprehensive view of how these diseases impact diverse populations.
2. **Rich Temporal Data**:
   * The dataset spans multiple years and includes time-stamped records of STD cases, making it ideal for time series analysis. Analyzing such temporal data can reveal trends, seasonal patterns, and other temporal dynamics that are vital for understanding the spread and control of STDs.
   * The longitudinal aspect of the data allows for the examination of how STD morbidity evolves over time, which is critical for developing effective forecasting models.
3. **Diverse Demographic Breakdown**:
   * The detailed breakdown by age, gender, and race/ethnicity enables a nuanced analysis of how STDs affect different segments of the population. This can help in identifying vulnerable groups and tailoring public health interventions to address their specific needs.
   * By incorporating demographic variables, we can enhance the precision of our models and gain a deeper understanding of the sociocultural factors influencing STD spread.
4. **Data Availability and Relevance**:
   * This dataset is publicly accessible, allowing for transparency and reproducibility in research. Its relevance to ongoing public health efforts makes it a valuable resource for informing policies and strategies to combat STDs.

**Aims of Forecasting the Data:**

1. **Predicting Future Trends**:
   * By forecasting STD cases, we can anticipate future outbreaks and prepare timely responses. Predictive models can help allocate resources efficiently, such as ensuring adequate supply of medical treatments and planning public health campaigns.
   * Forecasting allows us to project the burden of STDs in different demographic groups, aiding in the development of targeted intervention strategies.
2. **Identifying Seasonal Patterns**:
   * Time series analysis can uncover seasonal variations in STD incidence, which are critical for planning prevention efforts. For example, if certain STDs exhibit higher cases during specific times of the year, public health officials can proactively increase awareness and preventive measures during those periods.
   * Understanding these patterns can also help in scheduling educational programs and outreach efforts to coincide with periods of higher risk.
3. **Assessing the Impact of Interventions**:
   * By comparing forecasts with actual data over time, we can evaluate the effectiveness of public health interventions aimed at reducing STD rates. This feedback loop is essential for continuous improvement of health strategies and policies.
   * The models can also simulate the potential outcomes of various intervention scenarios, providing a data-driven basis for decision-making.
4. **Supporting Resource Allocation**:
   * Accurate forecasts enable health authorities to allocate resources more effectively. Knowing where and when to expect surges in STD cases helps in deploying healthcare personnel, supplies, and educational resources to the areas of greatest need.
   * This is particularly important in resource-constrained settings where efficient use of available resources can significantly impact public health outcomes.

**Practical Problems Solved with the Forecast:**

1. **Public Health Preparedness**:
   * Forecasting STD morbidity helps in preparing healthcare systems for potential surges, thereby reducing the risk of overwhelming facilities and ensuring better patient care.
   * It also aids in the strategic planning of vaccination campaigns, testing initiatives, and other preventive measures tailored to predicted outbreak patterns.
2. **Targeted Interventions**:
   * With detailed forecasts, public health officials can design and implement interventions targeted at the most at-risk populations. This could include community-specific education programs, increased access to testing and treatment in high-risk areas, and focused outreach to vulnerable demographic groups.
   * For instance, if the data reveals that certain age groups or ethnic communities are at higher risk during specific times, resources can be directed towards these groups with culturally and temporally appropriate interventions.
3. **Policy Development**:
   * The insights gained from forecasting can inform policymakers in creating data-driven policies that address the root causes of STD spread and enhance preventive measures. Policies can be tailored based on predicted trends and the demographic breakdown of cases.
   * This can include funding allocations, the introduction of new health programs, or changes in regulations regarding sexual health education and services.
4. **Educational Campaigns**:
   * Forecasting helps in timing public awareness campaigns more effectively. Knowing when STD cases are likely to rise allows for preemptive educational efforts, potentially curbing the increase before it becomes more widespread.
   * These campaigns can be designed to reach specific demographics identified as at risk through the forecast models, thereby increasing their effectiveness.

### Basic Visualizations

* **Time Plot**: Plot the data over time to visualize any trends or patterns.  
  The time plot reveals a clear upward trend in sales over the years, with noticeable seasonal spikes around the holiday seasons. This suggests a strong seasonal component in the data.

A graph showing the growth of cases

Description automatically generated

* **ACF Plot**: Generate an Autocorrelation Function plot to identify any correlations over time lags.  
  The ACF plot shows significant autocorrelation at lag 1, indicating that sales figures are correlated with those of the previous day. The gradual decline in autocorrelation suggests a trend component.

A graph of a plot of cases

Description automatically generated

## TRANSFORMATIONS

### Decomposition and Box-Cox Transformation

To perform a decomposition and Box-Cox transformation of your time series data on sexually transmitted disease (STD) morbidity, we’ll follow these steps:

1. **Load and Prepare the Data**: Load the dataset and ensure it is in a format suitable for time series analysis.
2. **Decompose the Time Series**: Break down the time series into its components: trend, seasonality, and residuals.
3. **Apply the Box-Cox Transformation**: Transform the data to stabilize variance and make it more suitable for analysis.

A graph of different cases

Description automatically generated with medium confidence

**Decomposition Analysis**

1. Observed Component: This is the original time series data. Visualizing this helps in understanding the overall patterns in the data, including any visible trends and seasonal cycles.
2. Trend Component: This component represents the long-term progression of the series. For STD morbidity data, the trend component can show whether the overall number of cases is increasing, decreasing, or stable over time. Identifying the trend is crucial for understanding the general direction of the disease spread.
3. Seasonal Component: This represents repeating short-term cycles in the data. For STD data, the seasonal component might reveal patterns such as higher incidence rates during certain times of the year, possibly related to behavioral changes or reporting practices. Understanding seasonality can help in planning targeted interventions during peak periods.
4. Residual Component: These are the remaining fluctuations in the data after removing the trend and seasonal effects. Ideally, the residuals should resemble white noise, indicating that all structured information has been captured by the trend and seasonal components. Analyzing the residuals helps in assessing the adequacy of the decomposition.

Applying a Box-Cox transformation was necessary to stabilize the variance, as indicated by the heteroscedasticity in the raw sales data. The transformed data appears more stable and suitable for modeling.

A graph showing a line

Description automatically generated

**Box-Cox Transformation Analysis**

1. **Stabilizing Variance**: The Box-Cox transformation is particularly useful when the time series data exhibits heteroscedasticity (i.e., the variance changes over time). Stabilizing the variance makes the series more suitable for statistical modeling and improves the accuracy of forecast models.
2. **Lambda Value**: The lambda value in the Box-Cox transformation is a parameter that determines the power to which all data points are raised. A lambda value close to zero suggests a logarithmic transformation, while other values suggest different levels of power transformation. The chosen lambda value is critical in achieving the desired stabilization effect.
3. **Transformed Data**: Visualizing the transformed data helps to confirm whether the variance has been stabilized and if the data is now more symmetric. This is a crucial step before applying any further statistical models, as many models assume homoscedasticity (constant variance).

## FORECASTING AND ANALYSIS

### Model Selection and Forecasting

The ARIMA and Holt-Winters (HW) Models were used to forecast STD cases. The ARIMA model showed a closer fit to the actual STD cases data, as indicated by the lower RMSE.

A graph showing a number of cases

Description automatically generated

Residual analysis revealed that both models had random residuals, suggesting a good fit. However, ARIMA had slightly less autocorrelation in the residuals compared to the Holt-Winters (HW) Model model.

Based on the RMSE and residual analysis, the ARIMA model provided more accurate forecasts for the sales data.

A graph with blue and orange lines

Description automatically generated

Interpreting the results of forecasting models, such as ARIMA (AutoRegressive Integrated Moving Average) and Holt-Winters (HW), based on Root Mean Squared Error (RMSE) involves understanding how well each model performs in predicting STD (sexually transmitted disease) cases. Here’s how you can interpret these RMSE values:

In time series forecasting, the Root Mean Square Error (RMSE) is a commonly used metric to measure the accuracy of a forecasting model. RMSE quantifies the difference between the actual values and the values predicted by the model. Lower RMSE values indicate better model performance as they reflect smaller differences between predicted and observed values.

Given the RMSE values for your models:

* **ARIMA Model RMSE**: 594,840.46
* **Holt-Winters (HW) Model RMSE**: 981,159.76

Here’s how to interpret these results:

**1. Comparative Accuracy**

The ARIMA model has a significantly lower RMSE (594,840.46) compared to the Holt-Winters model (981,159.76). This suggests that the ARIMA model provides more accurate forecasts for STD cases than the Holt-Winters model, given the data and the forecasting horizon.

* **ARIMA Model**: The Autoregressive Integrated Moving Average (ARIMA) model combines autoregressive terms, moving average terms, and differencing to account for non-stationarity. This model is flexible and can be tuned to handle a wide range of time series data, including those with trends and seasonality.
* **Holt-Winters Model**: This model, also known as the Triple Exponential Smoothing model, is typically used for data with both trends and seasonality. It involves smoothing levels, trends, and seasonal components to make forecasts.

The substantially lower RMSE for the ARIMA model indicates that it is better suited to capture the patterns in the data related to STD cases for this particular dataset and forecasting period.

**2. Forecasting Implications**

Given the lower RMSE, the ARIMA model's forecasts are closer to the actual observed values than those of the Holt-Winters model. This implies that:

* **Forecast Reliability**: Decision-makers can rely more on the ARIMA model's forecasts for planning and resource allocation regarding STD cases.
* **Model Selection**: For future forecasting efforts or similar datasets, the ARIMA model might be preferred over the Holt-Winters model unless new data or changes in the dataset characteristics suggest otherwise.

**3. Model Characteristics and Data Suitability**

* **ARIMA Model**: Its better performance could be due to its capability to model both the autoregressive and moving average components effectively while addressing non-stationarity through differencing. This indicates that the STD case data may have patterns that are better captured by ARIMA’s flexible structure.
* **Holt-Winters Model**: The higher RMSE might suggest that the trend and seasonal patterns in the data are not as well captured by the smoothing techniques of the Holt-Winters model. It could be that the data requires more intricate modeling of autocorrelations and non-stationarity, which ARIMA handles more robustly.

**4. Practical Use and Decision-Making**

* **Resource Allocation**: Accurate forecasting of STD cases allows public health authorities to allocate resources efficiently, plan interventions, and respond proactively to changes in case numbers.
* **Policy Development**: Understanding the performance of different forecasting models can guide the development of public health policies and strategies tailored to the dynamics of STD spread.

**Summary**

The ARIMA model’s lower RMSE suggests it is more accurate for forecasting STD cases in this context compared to the Holt-Winters model. This could be due to ARIMA's greater flexibility in handling complex patterns and trends in the data. For future forecasting and analysis, the ARIMA model should be the preferred choice, providing more reliable predictions that can better support public health planning and interventions.

REFERENCES  
US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for HIV, STD and TB Prevention (NCHSTP), Division of STD/HIV Prevention, Sexually Transmitted Disease Morbidity for selected STDs by age, race/ethnicity and gender 1996-2014, CDC WONDER Online Database. Accessed at http://wonder.cdc.gov/std-race-age.html on Jun 23, 2024 10:23:53 PM