Technical Note

Open-source Software for Demand Forecasting of Clinical Laboratory Test Volumes Using Time-series Analysis

Emad A. Mohammed[1,2,3,4], Christopher Naugler[2,3,4]

[1]Department of Electrical and Computer Engineering, Schulich School of Engineering, University of Calgary, Departments of [2]Pathology and [3]Laboratory Medicine and [4]Family Medicine, Diagnostic and Scientific Centre, University of Calgary and Calgary Laboratory Services, Calgary, AB, Canada

Received: 08 September 2016 Accepted: 27 January 2017 Published: 28 February 2017

Abstract

Background: Demand forecasting is the area of predictive analytics devoted to predicting future volumes of services or consumables. Fair understanding and estimation of how demand will vary facilitates the optimal utilization of resources. In a medical laboratory, accurate forecasting of future demand, that is, test volumes, can increase efficiency and facilitate long-term laboratory planning. Importantly, in an era of utilization management initiatives, accurately predicted volumes compared to the realized test volumes can form a precise way to evaluate utilization management initiatives. Laboratory test volumes are often highly amenable to forecasting by time-series models; however, the statistical software needed to do this is generally either expensive or highly technical. **Method:** In this paper, we describe an open-source web-based software tool for time-series forecasting and explain how to use it as a demand forecasting tool in clinical laboratories to estimate test volumes. **Results:** This tool has three different models, that is, Holt-Winters multiplicative, Holt-Winters additive, and simple linear regression. Moreover, these models are ranked and the best one is highlighted. **Conclusion:** This tool will allow anyone with historic test volume data to model future demand.

Keywords: Clinical test volume estimation, demand forecasting, forecasting software tool, Holt-Winters model, laboratory utilization

Introduction

Forecasting the future demand for medical services is a key component of health-care planning. This becomes increasingly important in laboratory medicine where unsustainable increases in service requests have occurred in recent years.[1-5] Annual increases in test volumes are the norm in clinical laboratories. However, medical utilization data also often exhibits a strong element of periodicity, meaning that volumes exhibit a repeating temporal pattern, with the baseline tending to increase on a yearly basis. The association of these patterns is a crucial element in predicting future volumes because the traditional method for assessing trends and predicting future volumes (i.e., linear regression[6]) is sensitive only to the baseline change and cannot be used to model short-term variations in volumes.

Time-series forecasting

Time-series forecasting methods have been applied heavily in many fields, for example, economics, bio-medical, meteorology, and electricity consumption.[7,8] Time-series methods are used to analyze historical data and estimate the future values. They have become an essential tool in the modern industrial environment for making decisions.

Time-series methods can be classified as parametric and nonparametric.[9] The parametric approach emphasizes representing the time-series using a statistical model. Modeling a time-series using a statistical approach, for example, Holt-Winters,[10] requires the validation of the model assumptions that describe the structural statistical norms of the process generating the time-series, that is, the residual error is random and normally distributed. If the data can

Address for correspondence: Dr. Christopher Naugler,
Department of Pathology and Laboratory Medicine,
Family Medicine and Community Health Sciences,
Diagnostic and Scientific Centre,
University of Calgary and Calgary Laboratory Services,
C-262, 9, 3535 Research Road NW, Calgary,
AB T2L 2K8, Canada.
E-mail: christopher.naugler@cls.ab.ca

How to cite this article: Mohammed EA, Naugler C. Open-source software for demand forecasting of clinical laboratory test volumes using time-series analysis. J Pathol Inform 2017;8:7.

Available FREE in open access from: http://www.jpathinformatics.org/text.asp?2017/8/1/7/201109

Access this article online

Quick Response Code:

Website:
www.jpathinformatics.org

DOI:
10.4103/jpi.jpi_65_16

comply with the model assumptions, then the model under investigation can be used to detect future values of the data. If the assumptions cannot be validated then nonparametric time-series analysis models, for example, neural networks,[11] can be used to represent the data and predict the future values. A comprehensive classification of various time-series forecasting methods is available.[9]

Material and Methods

Figure 1 illustrates a flow diagram to model a given time-series using the tool described in this paper. The starting point is to understand the underlying characteristics of the time-series under investigation. The time-series characteristics indicate the appropriate selection from among the candidate models. The characteristics may include: (1) the time-series trend, for example, linear, multiplicative, or additive,[12] (2) the seasonality index that describes if the value is above or below the time-series trend, (3) the periodicity of the time-series that describes if a pattern in the data has a specific frequency. These characteristics may indicate candidate parametric or nonparametric models to fit the data. Each model is trained using part of the data and the model's performance parameters are calculated and then the best model is selected and used to forecast the future values of the time-series. If the predicted values are within the 95% prediction interval (PI),[13] the selected model can be used to forecast the future values of the time-series, otherwise, the new recorded actual values are appended to the raw data of the time-series and the whole process restarted. In forecasting, any percentage may be used as a PI, however, it is common to calculate 80% and 95% PI to check for wide ranges of variation around the predicted values.[13]

In this paper, we present a new web-based open-source software based on the R statistical package[14] which is designed to (1) provide user-friendly clinical laboratory volume forecasting, (2) compare different models head-to-head and select the one that best fits the users' data, and (3) provide downloadable predicted test volume data for the time span chosen by the user. It is intended that this publication serves as the citable reference to this software in the published literature.

Time-series models, data characteristics, and model selection

In this section, we describe the models that we use to develop the forecasting tool, the data characterizations that should lead to selection of a certain model, and the selection/ranking criteria of the models.

Holt-Winters model

Model definition and assumption

The Holt-Winters forecasting model includes triple exponential smoothing models. Exponential smoothing model is forecasting model that estimates the predicted values on the history of the time-series data. Exponential smoothing models assume that the historical and predicted data of the time-series data are relatively continuous and have common repeated patterns, and thus, the exponential soothing models are well-matched to short-term predictions. The exponential smoothing models employ smoothing parameters to base the future values on the past ones. Different values of the smoothing parameters will give different exponential decreasing emphasis to the recent values compared to the more distant values in the time-series data.

The Holt-Winters models for time-series analysis have three data components level, trend, and seasonality. The goal of the exponential smoothing model is to estimate the value of the level, trend, and seasonal pattern. These values are then used to construct the Holt-Winters models for future values prediction. The time-series components are time varying components and may have different values at the beginning and end of the time-series. This is in addition to a random noise component that is completely independent of the time-series components.

An exponential smoothing model for a high variation and low noise time-series requires high values for the smoothing parameters. This is mandatory to emphasis more on the most recent values as these values can represent the future values more accurately compared to past values. However,

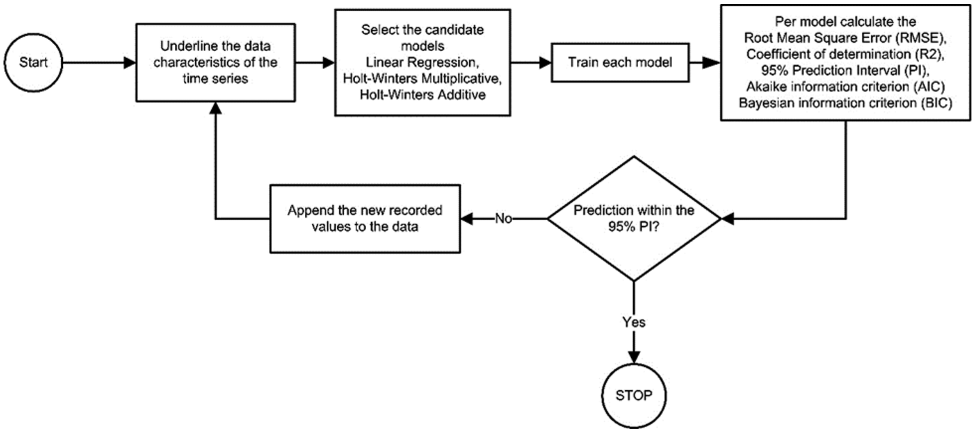

Start
Underline the data characteristics of the time series
Select the candidate models
Linear Regression,
Holt-Winters Multiplicative,
Holt-Winters Additive
Train each model
Per model calculate the
Root Mean Square Error (RMSE),
Coefficient of determination (R2),
95% Prediction Interval (PI),
Akaike information criterion (AIC)
Bayesian information criterion (BIC)
Prediction within the 95% PI?
No
Append the new recorded values to the data
Yes
STOP

Figure 1: Time-series forecasting work flow. The pipeline shows the workflow to accurately forecast the future values of a time-series with certain characteristics, e.g. specific trend. A set of candidates models are trained and the best one is selected based on the performance parameters, i.e., root mean squared error, determination coefficient (R^2), 95% prediction interval, Akakie information criterion, and Bayesian information criterion

exponential smoothing model for a noisy time-series requires more historical data to cancel out the noise to accurately estimate the future values.

There are two types of the Holt-Winters models namely; additive and multiplicative models. The additive models generate constant seasonal variations independent of the time-series trend and multiplicative models generate seasonal patterns that fluctuates as the trend increases/decreases.

Model mathematical characteristic

Holt-Winters is a statistical method of modeling, applied to time-series that exhibit a trend and seasonality, which is founded on the basis of the exponential moving average.[10] The Holt-Winters model has three parts; an equation of the forecasting model characterizes each. The model has two types: (1) additive seasonality (i.e., linear trend) and (2) multiplicative seasonality (i.e., nonlinear trend). In the case of multiplicative models, the seasonality index increases with an increase in the level of the time-series. The additive Holt-Winters model can be used if the seasonal index does not depend on the current level of the time-series.

The following equations represent the multiplicative Holt-Winters model:

$$\text{Level:}\quad L_t = \alpha\left(\frac{Y_t}{S_{t-m}}\right) + (1-\alpha)\times\left(L_{t-1} - b_{t-1}\right) \tag{1}$$

$$\text{Trends:}\quad b_t = \beta\times\left(L_t - L_{t-1}\right) + (1-\beta)\times b_{t-1} \tag{2}$$

$$\text{Seasonal Index:}\quad S_t = \gamma\times\left(\frac{Y_t}{L_t}\right) + (1-\gamma)\times S_{t-m} \tag{3}$$

$$\text{Forecast:}\quad F_{t+k} = \left(L_t + k\times b_t\right)\times S_{t+k-m} \tag{4}$$

The following equations represent the additive Holt-Winters model:

$$\text{Level:}\quad L_t = \alpha\left(Y_t - S_{t-m}\right) + (1-\alpha)\times\left(L_{t-1} - b_{t-1}\right) \tag{5}$$

$$\text{Trends:}\quad b_t = \beta\times\left(L_t - L_{t-1}\right) + (1-\beta)\times b_{t-1} \tag{6}$$

$$\text{Seasonal Index:}\quad S_t = \gamma\times\left(Y_t - L_t\right) + (1-\gamma)\times S_{t-m} \tag{7}$$

$$\text{Forecast:}\quad F_{t+k} = \left(L_t + k\times b_t\right) + S_{t+k-m} \tag{8}$$

Where m is the number of data points of the seasonal cycle, k is an index, t is the time of recording, and Y_t is the recorded data at time t. The smoothing factors are α, β, and γ where $0\leq \alpha \leq 1$, $0\leq \beta \leq 1$ and $0\leq \gamma \leq 1$. The seasonal index represents the differences between the current level and the data at the seasonal cycles.

The root mean square error (RMSE) measure[15] is used to validate the goodness-of-fit and is calculated by the following equation:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{1}^{t}\left(Y_t - F_t\right)^2} \tag{9}$$

Where n is the total number of data points.

The RMSE the goodness-of-fit of the model, which describes the magnitude of the error in terms that would be relatively more useful to decision makers compared to other error measures.[15]

The coefficient of determination[12] (R^2) is used to measure the relative enhancement in the forecasting of the future values of the regression model compared to the mean model (i.e., the average value of the observations). R^2 can have values from 0 to 1, where zero indicates the failure of the model to improve the forecasting over the mean model and one indicates perfect forecasting. R^2 can be calculated as:

$$R^2 = \frac{1}{n}\sum_{1}^{i}\frac{\left(F_i - \bar{Y}\right)^2}{\left(Y_i - \bar{Y}\right)^2} \tag{10}$$

where $\bar{Y}$ is the average value of the observations.

Linear regression model

Model definition and assumption

Linear regression is a method for modeling the linear relationship between a scalar dependent variable (response variable) denoted as Y and one or more independent variables (explanatory variables) denoted as X. The case of one explanatory variable is known as simple linear regression.

The simple linear regression model assumes a linear relationship between the independent and dependent variables. The linearity assumption can be visually tested with a scatter plot between the independent variable on the X-axis and the dependent variable on the Y-axis. The simple linear regression analysis requires the independent variable to be normally distributed. If the independent is not normally distributed a nonlinear transformation, e.g., log-transformation, may be used to transform the independent variable to normally distributed variable. This is in addition to the assumption of independence of the residual error that must be independent from the explanatory variable. Moreover, simple linear regression analysis requires that there is little or no autocorrelation in the data.

Model mathematical characteristics

A linear regression model[6] represents the relationship between two variables (X and Y) by fitting a line to the recorded data. The X variable is the explanatory/independent variable, and the Y variable is the predicted/dependent variable. A linear regression line can be described as:

$$Y = \mathrm{a} + \mathrm{b}\times X \tag{11}$$

Where X is the explanatory/independent variable and Y is the predicted/dependent variable. The intercept of the line is a and the slope of the line is b.

The least-squares method[6] is used to calculate the model parameters by finding the best line that can fit the recorded data by minimizing the sum of the squares of the error from each data point to the line.

Model selection criteria

In the development of the time-series forecasting model, we train three different models (i.e., Holt-Winters multiplicative, Holt-Winters additive, and linear models). Too use these models for forecasting, it is required to select the optimal model, the initial values, and the values of the parameters α, β, and γ.

Akaike information criterion (AIC)[16,17] is a method used to calculate the likelihood/probability of the model to predict the future values. We calculate the AIC per model and select the one that minimizes the AIC value.

Bayesian information criterion (BIC)[17] is another method for model selection. BIC measures the trade-off between model fit and complexity. A lower AIC or BIC value indicates a better fit.

The following formulas are used to calculate the AIC and BIC of a model:

$$\text{AIC} = -2 \times \ln(L) + 2 * k \quad (12)$$

$$\text{BIC} = -2 \times \ln(L) + 2 \times \ln(N) \times k \quad (13)$$

Where L is the value of the likelihood function calculated at the parameter estimates, N is the number of observations, and is the number of estimated parameters.

Model validation

Forecasting model validation is the process of testing a model against unseen samples and recording of the prediction error. The prediction error can be used as a criterion to select among different models. The validation process is a method of measuring the predictive performance of a statistical model. Model goodness-of-fit statistics, that is, RMSE, is not an ultimate indicator on how well a model will predict the future values as it is easy to over-fit the training dataset to minimize the goodness-of-fit error. However, the predictions from the model on unseen dataset will generally get worse.

To construct a predictive model, the dataset is first divided into training and validation datasets. The training dataset is used to estimate the model parameters and decide upon the models complexity to mitigate the effect of overfitting. The validation dataset is then used to test the model against unseen dataset and record the generalization error (prediction accuracy) of the predictions. The predictive accuracy of a model can be measured by RMSE on the validation dataset (testing dataset).

There are many method for predictive models validation, among them are: k-fold, leave-one-out, and hold-out validation methods.[18-20] These methods assume that the observations in the input dataset are independent of each other. However, the observations in time-series are not, and thus, the validation process becomes more difficult as leaving out random observations do not remove all the associated information because of the time-dependency between observations.

In this paper, the time-series forecasting models are trained and validated as follows:

1. The time-series dataset must have at least 2-cycles of observations (24 observations for monthly and 14 observations for weekly cycles) to compute the Holt-Winters models
2. If the time-series data have more than 2-cycles of observations, then at least the first 2-cycles must be used as the training dataset. The remaining observations should be used as the validation dataset. Increasing the size of the training data is mandatory in case of noisy data to better estimates the different components in the time-series. The software can predict up to fifty estimates in the future, and thus, if there are enough observations, a common practice is to keep the last fifty observations as the validation dataset and train the model of the remaining data
3. Save the training and validation datasets into two separate CSV files. Load the training dataset file, set the parameters, and download the predicted values. (see section "Using the software" for more details on how to use the software)
4. Use the predicted values to compute the RMSE per model using equation (9).

Data preparation

In this paper, we used three different datasets to illustrate the usage of the forecasting software tool with real-life use cases (see the Result and Discussion section for model training and testing results per dataset)

Clinical laboratory test volumes

A dataset of the test volumes of all different clinical tests are recorded monthly for the period of April 2011–March 2015 from all medical facilities located at the Province of Alberta, Canada. This dataset was collected by the Alberta Health Services Laboratory Utilization Office in Alberta, Canada. The dataset consists of forty observations and the first 24 observations are used as training while the remaining 16 observations are used for validation. This dataset can be downloaded from the software (see section using the software). There are many parameters that influence clinical laboratory test orders, amongst them are: Patient severity, patient assurance, number of patient visits, etc., that should be used to normalize the clinical laboratory test volumes. However, these parameters are not possible to collect in the scope of this paper as there are concerns for patient privacy.

Precipitation in millimeters Eastport, USA, 1887–1950:

This dataset[21] represents a monthly time-series (January 1887–December 1950) with high level of noise. This dataset has 768 observations and is divided into training dataset (first 718 observations) and validation dataset (last fifty observations).

Airlines passenger dataset:

This dataset[22] represents the number of international passengers per month on an airline in the United States and were obtained from the Federal Aviation Administration for the period 1946–1960. This dataset has exponential raising trend. This dataset has 135 observations and is divided into training dataset (first 85 observations) and validation dataset (last fifty observations).

Implementation

The forecasting software tool is implemented using the R statistical packages and the Web interface is built using the

Shiny R package. In the following section the layout of the Web interface, functionalities, and the tool usage are described.

Availability

The forecasting tool software is freely available from the authors. The software can be accessed online through the following link: https://github.com/ClinicalLaboratory/Clinical-Laboratory

Results and Discussion

Figure 2 shows a linear regression model fitted to the monthly clinical laboratory test volumes for the period of April 2011–March 2015 from all medical facilities located in the Province of Alberta, Canada. The vertical dotted line represents the starting point to forecast future values. Figure 3 shows a Holt-Winters multiplicative model fitted to the same data; however, the fitted values are closer to the actual values compared to the data illustrated in Figure 2. Moreover, Figure 2 shows that the predicted values have a wider 95% PI compared to the predicted values shown in Figure 3. It is obvious that for predictions at the level of monthly test volumes, linear regression is inadequate, whereas the Holt-Winters multiplicative model can provide more accurate results. This is due to the fact that the linear regression model fitted/predicted value at time t is completely independent of the fitted/predicted value at time *t* – 1, however, the Holt-Winter models, that is, multiplicative and additive, provide this dependency using the smoothing parameters, that is, α, β, and γ. Moreover, the independent variable (X) in the linear regression model is represented as a numerical time index, which does not reflect the seasonality measure that exists in the dependent variable (Y). However, if the X-axis is restructured to reflect the seasonality index, for example, using repeated categorical values such as name of the month instead of the numerical time index, the model can capture only the seasonality variation and miss the variation in the year over year trend represented by the data. By contrast, the Holt-Winters models include separate representation for the level, trend, and seasonality of the data, which makes it a better model to represent clinical laboratory test volume data.

Time-series analysis has been employed by a number of authors to model epidemiology,[23-28] physiology,[29-31] and resource utilization.[32] Although its usage in modeling laboratory test volumes was suggested over 35 years ago[33] it is rarely used for clinical laboratory test volume prediction.

Indeed, the choice of the best statistical model to use in a given solution is often difficult. In addition to linear regression, there are several variations of time-series model from which to choose.[16,34] Moreover, the use of these models generally involves advanced programming knowledge for the open-source versions or the purchase of proprietary software packages.[35]

This software is primarily designed to be used in medical laboratory settings to estimate clinical test volumes. However, in this section the results of applying the forecasting models are illustrated using 3 different datasets representing different data characteristics case studies.

The models used in the forecasting software tool (Holt-Winters multiplicative, Holt-Winters multiplicative, and linear regression) are trained with the three different datasets (see data preparation section for details). Figures 3-5 show the fitting and prediction results of the best model per dataset. The predictions are then used to compute the RMSE per model per dataset and the results are illustrated in Table 1.

Table 1 shows the RMSE per model per dataset. The clinical laboratory dataset is best modeled by the Holt-Winters multiplicative model as the dataset shows multiplicative trend. This is the same case for the airlines passenger dataset where the dataset shows multiplicative and exponentially rising trend.

The Holt-Winters models are best fit the dataset when it has a continuity property as illustrated by the clinical laboratory test volumes and airlines passenger datasets. This continuity is not achieved in the precipitation dataset and the sudden changes

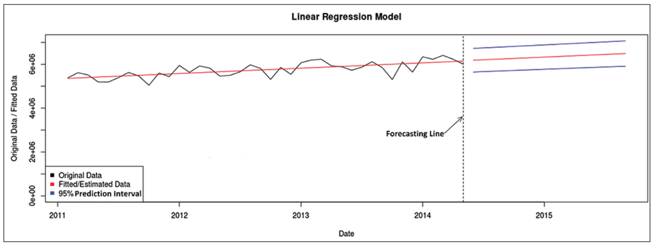


Figure 2: Linear regression model of the monthly test volumes for all clinical laboratory tests in the Province of Alberta, Canada

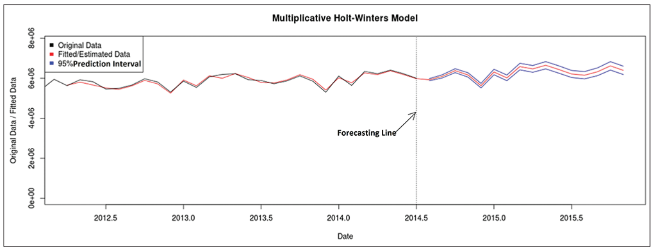


Figure 3: Multiplicative Holt-Winters model of the monthly test volumes for all clinical laboratory tests in the Province of Alberta, Canada

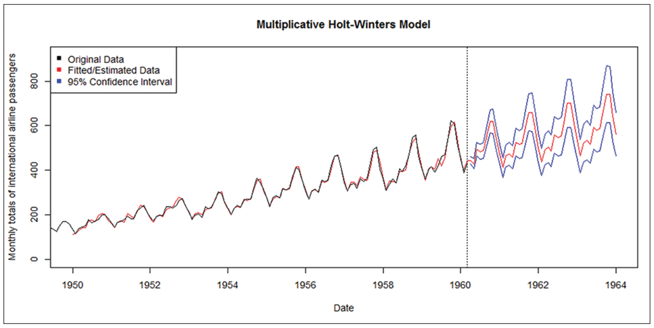


Figure 4: Multiplicative Holt-Winters model of the monthly volume of airline passengers

in the trend and seasonal patterns cannot be captured correctly by Holt-Winters models and the linear regression model is the best model to fit the data in this case with minimum RMSE.

Software architecture

In this section the software architecture is explained. It is designed in a multi-tier architecture[36] and is comprised of two tiers. These tiers are illustrated in Figure 6 and are explained in the following:

Client tier

The client tier interacts with the users to obtain the prediction results. Since the software application conforms to a two layered services application it hosts the presentation layer components, that is, web interface/browser. For the forecasting web application, the client tier comprises the user workstations/computers, and other devices that host a web browser, e.g., tablets. The data are stored on the local file system of the client tier.

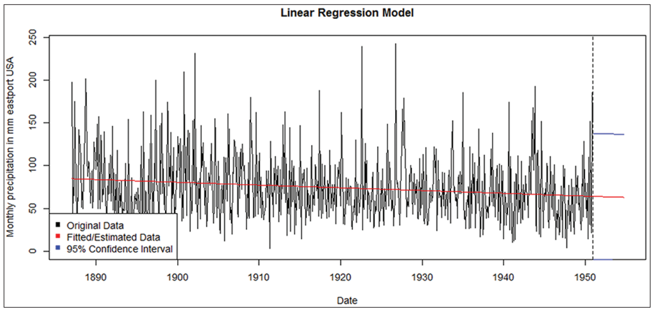


Figure 5: Linear regression model of monthly precipitation

Application tier

The servers used in the application tier are responsible for hosting all the application's libraries and the Web servers are provided by the RShiny server.[37] In this case a user does not have to install the RStudio or any forecasting packages, e.g., the CARET package.[14] Moreover, the RShiny server is responsible for instantiating the application per user and running the user commands.

Separating the client computer from the application logic supports the development and distribution of thin-client applications that require minimum software at the client tier, for example, a web browser.

The initial version of the forecasting tool was deployed on the RShiny server; however for privacy concern of the data, we chose to upload code of the GitHub repository as described below.

Laboratory demand forecasting software functionalities

R and RStudio must be installed on the user machine before the tool can be used. The next step is to download the project

Table 1: Root mean square error for different models and validation datasets

Model/dataset	Holt-Winters multiplicative	Holt-Winters additive	Linear regression
Clinical laboratory test volumes	176,062	174,431	276,074
Airlines passenger	30.801	49.78	66.23
Precipitation	39.73	39.61	38.77

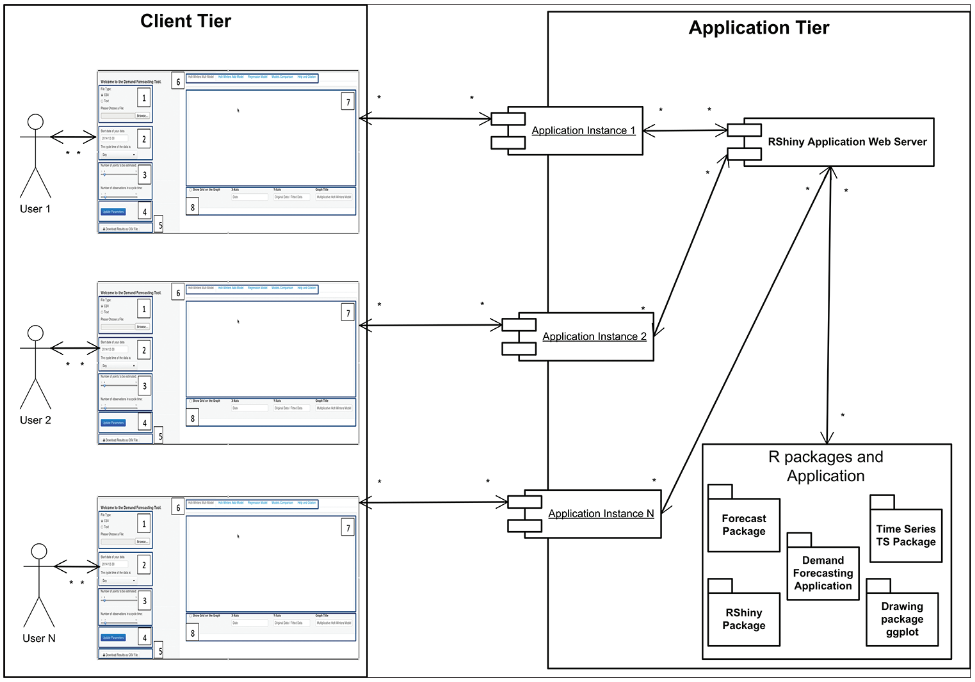


Figure 6: The demand forecasting software architecture. The system is designed in multi-tier architecture style. The client tier represents multiple users that can interact with the web application simultaneously and independently. The users' requests are handed by the application tier, where the R packages are hosted along with the web server application

files from the following GitHub repository: https://github.com/ClinicalLaboratory/Clinical-Laboratory.

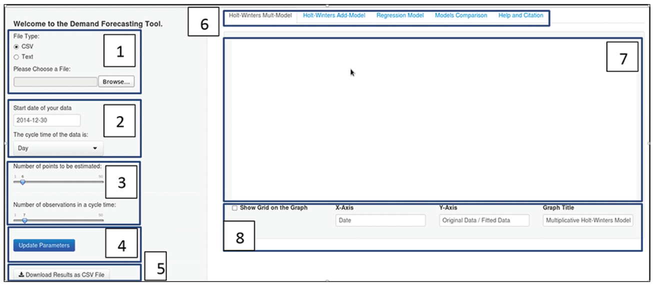
After downloading all the files, run the "ClinicalLaboratory.Rproj" file and press the "Run App" button in the RStudio interface and finally click on "Open in Browser" to use all the functionalities of the tool as described below.

The start-up screen illustrated in Figure 7 shows the following areas:

1. Area #1 contains the file types that can be processed by the software, namely comma separated value "CSV" and text files
2. Area #2 is where the user records the start date of the data. It is used to set the time stamp of the recorded time-series; if you click inside this area a calendar will open and the start date can be chosen. Another button named "The cycle time of the data is" is used to select the time interval between two successive recordings. In this version of the software, the possible intervals are day of the week and month of the year
3. Area #3 is used to select the forecasting horizon (the number of future points to be estimated). The slider can set the forecasting horizon from 1 to 50 increments (day or month) in the future
4. Area #4 is an "update" button. Whenever there is a change in Area 1, 2, 3, or 8, the update parameter button must be clicked for the changes to take effect
5. Area #5 is the button used to save the estimation results in a single CSV file
6. Area #6 contains tabs to choose individual models, compare models, view help files and view the suggested citation
7. Area #7 is the time-series plot area, which illustrates the original, fitted, and estimated models
8. Area #8 is the plot attributes control.

Using the software

1. Getting started - To start using the software, a time-series in CSV or text format must be loaded first (make sure that you select the right format of your file), if you have a stored time-series on your local hard disk, click "Browse" and click on the file contains your time-series. The next step is to adjust the time-series parameters located in Area #2, and 3. Finally click the "Update Parameters" button. If you do not have a time-series file, you can click the "Help and Citation" tab and click "Download Sample Text data" or "Download Sample CSV data". This will also show you the proper format for your own data. When the estimation cycle calculations are completed, a table containing the estimated points will appear under the Area #8. This is illustrated in Figure 8. At least 2 cycles of data (i.e. 2 years of data are required, if the time interval is monthly) are needed to perform time-series analyses
2. Adjust the plot attributes - you can add grid lines to your plot by checking the option "Show Grid on the Graph." A new X-axis, Y-axis, and title can be displayed on the plot by writing the appropriate labels in the corresponding fields in Area #8 and then clicking "Update Parameters"
3. Model Comparison - the software contains three different models (i.e., Holt-Winters multiplicative, Holt-Winters additive and linear regression models) that are used to estimate the future values of the loaded time-series. These models are examined for the ability of fitting and estimating the future values of the time-series and the best model is selected based on this metric. This is demonstrated in the "Models Comparison" tab that is illustrated in Figure 9
4. In the model comparison tab, you can view the stationary nature of the processed time-series "Area #1" (i.e., if the process generating the time-series is stable or not), the residual error form fitting the time-series by every model is shown in "Area #2, #3, and #4." All models are ranked by prediction power and the rank is displayed in a table showing the model name and rank in "Area #5"
5. Saving the results - to save the estimation results for the entire models click on "Download Results as CSV File." The file will be automatically named, although you may wish to rename it at this point.



File Type:
● CSV
○ Text
Please Choose a File:

Start date of your data
2014-12-30
The cycle time of the data is:
Day
Number of points to be estimated:
1 6 50
Number of observations in a cycle time:
1 7 50

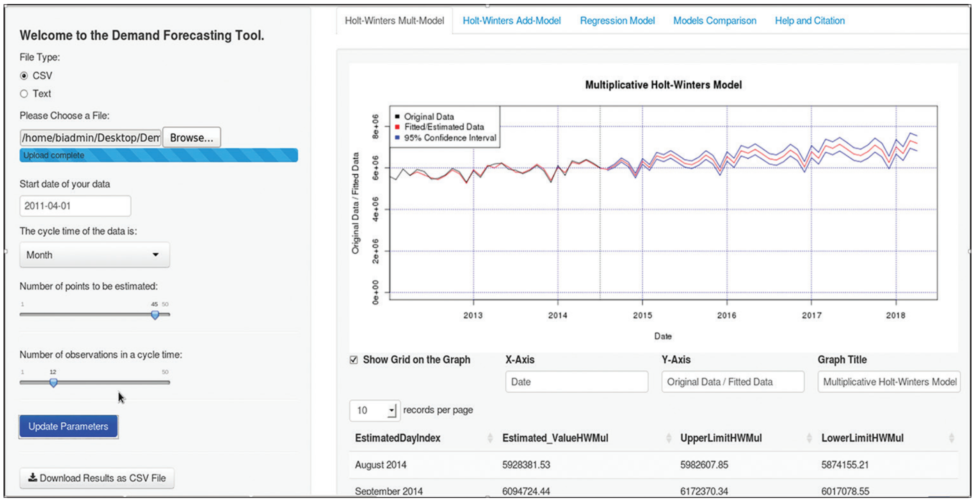
□ Show Grid on the Graph
X-Axis
Date
Y-Axis
Original Data / Fitted Data
Graph Title
Multiplicative Holt-Winters Model

Figure 7: Start-up screen of the demand forecast tool

Conclusion

Simple models are easier to build, implement, interpret and update. Increasing model complexity leads to complex implementation and interpretation. In most cases, the ability to understand the model and it's parameters is preferred over a complex model that may not be easier to interpret. Linearity and continuity are common assumptions for time-series modeling, which are considered as weak assumptions. Weak assumptions that are coupled with complex algorithms are more inefficient than using more data with simpler algorithms. This is because a training dataset is a subset of relevant data and with more data, the estimates of the future values can be more accurate under the weak assumptions. With much more data, the sample variation accurately represents the underlying population and the future estimates tend to be more accurate.

Readymade algorithms are used as a "black box" that is impossible to understand or modify, and therefore, leads to very complex training phase and model validation that may not be user-friendly for many users. R and RStudio provide a programming environment to design and implement different time-series prediction algorithms. However, it requires trained



Welcome to the Demand Forecasting Tool.

File Type:

● CSV

○ Text

Please Choose a File: /home/biadmin/Desktop/Dem

Start date of your data 2011-04-01

The cycle time of the data is: Month

Number of points to be estimated: 1 45 50

Number of observations in a cycle time: 1 12 50

☑ Show Grid on the Graph

X-Axis

Date

Y-Axis

Original Data / Fitted Data

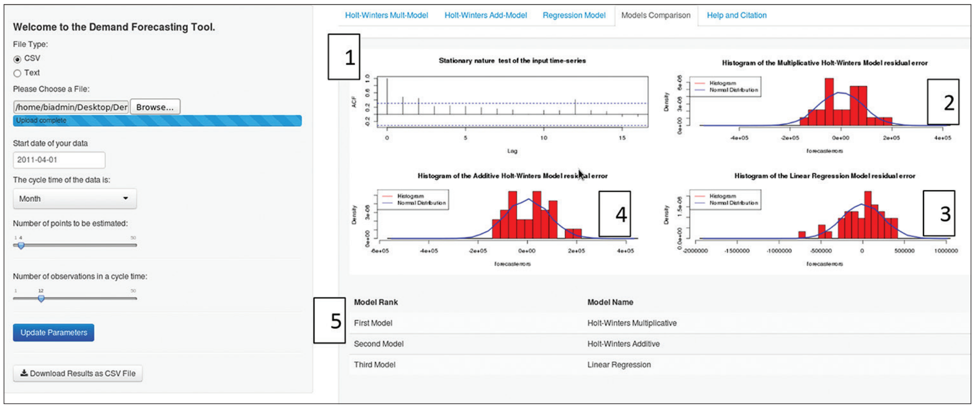
Graph Title

Multiplicative Holt-Winters Model

10 records per page

EstimatedDayIndex	Estimated_ValueHWMul	UpperLimitHWMul	LowerLimitHWMul
August 2014	5928381.53	5982607.85	5874155.21
September 2014	6094724.44	6172370.34	6017078.55

Figure 8: Sample time-series estimation for 45 months in the future



Welcome to the Demand Forecasting Tool.

File Type:

● CSV

○ Text

Please Choose a File: /home/biadmin/Desktop/Der

Start date of your data 2011-04-01

The cycle time of the data is: Month

Number of points to be estimated: 1 4 50

Number of observations in a cycle time: 1 12 50

Model Rank	Model Name
First Model	Holt-Winters Multiplicative
Second Model	Holt-Winters Additive
Third Model	Linear Regression

Figure 9: Model comparison and best model selection

personnel to design, implement, validate, select the best model, and interpret the model parameters. The open-source software described in this paper provide a user-friendly interface and make it easier to load a time-series dataset, build three different models to predict the future values of the time-series data and choose the best model.

In this paper, we present a new open-source program for future demands prediction based on a comparison of linear regression and two forms of time-series analysis, that is, Holt-Winters multiplicative and additive models. This software fills an important gap in the available open-source software and greatly simplifies the process of demand forecasting. Although the software was developed with the clinical laboratory in mind, the software could be equally useful in other areas of medicine or business.

In clinical laboratories the authors foresee two main applications. First the tool can be used to predict future test volumes for the purpose of reagent, staffing, and analyzer needs. This may help to reduce waste, staff overtime, and testing delays due to inadequate resources.

A second and more innovative use involves the evaluation of utilization management initiatives. Measures designed to promote the cost-effective use of medical laboratory tests are widespread in regions of Europe and North America.[2,4] These "utilization management" initiatives often result in changes in overall test volumes in the range of 5%–10%. However, as seen in Figure 2, actual observed test volumes may vary by up to 20% from month to month, potentially completely masking any effect of a utilization management initiative. The use of the new demand forecasting tool can detect utilization management effects as small as 1%–2% in some instances. To do this, the user would need at least 24 months of historical data to establish the pattern of predicted future volumes. Forecasting is simplified if the planned intervention begins on the first of a month. The period of the historic forecasting would then include the month immediately prior to the start of the intervention and the predicted demand would begin on the 1st day of the intervention. As the software generates 95% PI, it is a simple matter to compare the observed intervention volumes with the predicted volumes. If the observed volumes fall outside of the 95% PI, it could be concluded that the intervention had a significant effect. The percentage change attributable to the intervention could then be determined by comparing the observed and predicted values. This method may detect intervention effects as small as a few percentage points as soon as 1 month after the start of a utilization management intervention.

The forecasting software tool has the following advantages compared to the popular tool WEKA:[38]

1. The initial parameters of the models are calculated by the software and do not require any knowledge from the user
2. The residual error of the fitted data and the stationary nature of the data are displayed for the user as a visual validation of the model assumptions
3. The models are ranked according to their forecasting performance and complexity.

We examine the software tool using two other use-cases of real-life data and show how to validate the models performance.

Limitations

The time-series methods described in this article are of the parametric type. The model assumptions must be verified to consider a model to be valid. Another limitation of these models are the sensitivity to outliers, which may cause significant errors in the predicted values. The parameters of the Holt-Winters models required by the forecasting tool must be entered manually, for example, "The cycle time of the data." This mandates that the user is aware of the characteristics of the time-series data.

Future work

The future enhancement of this tool is to fully automate the data characterization process, i.e. the software should be able to identify the periodicity and handle the outliers.

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

References

1. Bossuyt X, Verweire K, Blanckaert N. Laboratory medicine: Challenges and opportunities. Clin Chem 2007;53:1730-3.
2. Naugler C. A perspective on laboratory utilization management from Canada. Clin Chim Acta 2014;427:142-4.
3. Alonso-Cerezo MC, Martín JS, García Montes MA, de la Iglesia VM. Appropriate utilization of clinical laboratory tests. Clin Chem Lab Med 2009;47:1461-5.
4. Plebani M, Zaninotto M, Faggian D. Utilization management: A European perspective. Clin Chim Acta 2014;427:137-41.
5. Huck A, Lewandrowski K. Utilization management in the clinical laboratory: An introduction and overview of the literature. Clin Chim Acta 2014;427:111-7.
6. Preacher KJ, Curran PJ, Bauer DJ. Computational tools for probing interactions in multiple linear regression, multilevel modeling, and latent curve analysis. J Educ Behav Stat 2006;31:437-48.
7. Yule GU. On a method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers. Philos Trans R Soc Lond A 1927;226:267-98.
8. Rosner B. Fundamentals of Biostatistics. Paducah, KY: Cengage Learning, Kentucky Publishing Inc. 1540 McCracken Blvd; 2010.
9. Fan J, Yao Q, editors. Nonlinear Time Series: Nonparametric and Parametric Methods. Spring Street, NY 10013-1578, USA: Springer Science+Business Media; 2003.
10. Chatfield C, Yar M. Holt-Winters forecasting: Some practical issues. Statistician 1988;37:129-40.
11. Yegnanarayana B. Artificial Neural Networks. 1st ed. India: PHI Learning Pvt. Ltd., India Institute of Technology; 2006. p. 476.
12. De Gooijer JG, Hyndman RJ. 25 years of time series forecasting. Int J Forecast 2006;22:443-73.
13. Chatfield C, editor. Prediction intervals for time-series forecasting. In: Principles of Forecasting. Spring Street, NY 10036, USA: Springer; 2001. p. 475-94.
14. The R Project for Statistical Computing. Available from: http://www.r-project.org/. [Last accessed on 2015 Feb 20].
15. Armstrong JS, Collopy F. Error measures for generalizing about forecasting methods: Empirical comparisons. Int J Forecast 1992;8:69-80.
16. Akaike H. A new look at the statistical model identification. IEEE Trans Automat Contr 1974;19:716-23.
17. Stone M. Comments on model selection criteria of Akaike and Schwarz. J R Stat Soc Ser B (Methodol) 1979;41:276-8.
18. Box GE, Jenkins GM, Reinsel GC, Ljung GM. editors. Time Series Analysis: Forecasting and Control. Hoboken, NJ 07030-5774, USA: John Wiley and Sons; 2015.
19. Chatfield C, editor. The Analysis of Time Series: An Introduction. Boca Raton, Florida, USA: CRC Press; 2016.
20. Ripley B. Statistical Data Mining. New York: Springer-Verlag; 2002.
21. Available from: https://www.datamarket.com/data/set/22y6/precipitation-in-mm-eastport-usa-1887-1950#!ds = 22y6 and display = line. [Last accessed on 2016 Dec 11].
22. Airlines Passenger Dataset. Available from: https://www.rpubs.com/nohaelprince/47545. [Last accessed on 2016 Dec 12].
23. Yang L, Bi ZW, Kou ZQ, Li XJ, Zhang M, Wang M, *et al.* Time-series analysis on human brucellosis during 2004-2013 in Shandong Province, China. Zoonoses Public Health 2015;62:228-35.
24. Dangor Z, Izu A, Hillier K, Solomon F, Beylis N, Moore DP, *et al.* Impact of the antiretroviral treatment program on the burden of hospitalization for culture-confirmed tuberculosis in South African children: A time-series analysis. Pediatr Infect Dis J 2013;32:972-7.
25. Huang Y, Deng T, Yu S, Gu J, Huang C, Xiao G, *et al.* Effect of meteorological variables on the incidence of hand, foot, and mouth disease in children: A time-series analysis in Guangzhou, China. BMC Infect Dis 2013;13:134.
26. Spaeder MC, Fackler JC. Time series model to predict burden of viral respiratory illness on a pediatric Intensive Care Unit. Med Decis Making 2011;31:494-9.
27. Lopman B, Armstrong B, Atchison C, Gray JJ. Host, weather and virological factors drive norovirus epidemiology: Time-series analysis of laboratory surveillance data in England and Wales. PLoS One 2009;4:e6671.
28. López-Lozano JM, Monnet DL, Yagüe A, Burgos A, Gonzalo N, Campillos P, *et al.* Modelling and forecasting antimicrobial resistance and its dynamic relationship to antimicrobial use: A time series analysis. Int J Antimicrob Agents 2000;14:21-31.
29. Gupta AK, Udrea A. Beyond linear methods of data analysis: Time series analysis and its applications in renal research. Nephron Physiol 2013;124:14-27.
30. Miyake K, Miyake N, Kondo S, Tabe Y, Ohsaka A, Miida T. Seasonal variation in liver function tests: A time-series analysis of outpatient data. Ann Clin Biochem 2009;46(Pt 5):377-84.
31. Matthews DR. Time series analysis in endocrinology. Acta Paediatr Scand Suppl 1988;347:55-62.
32. Abdel-Aal RE, Mangoud AM. Modeling and forecasting monthly patient volume at a primary health care clinic using univariate time-series analysis. Comput Methods Programs Biomed 1998;56:235-47.
33. Elwell GR. Forecasting: Which type is for you? Am J Med Technol 1979;45:131-5.
34. Engle RF, Granger CW. Co-integration and error correction: Representation, estimation, and testing. Econometrica 1987;55:251-76.
35. SAP Business One Software. Available from: http://www.softwareadvice.com/ca/accounting/sap-business-one-profile/. [Last accessed on 2015 Feb 11].

36. Yang J, Pang J, Qi N, Qi T. On-demand self-adaptivity of service availability for cloud multi-tier applications. In: Cluster, Cloud and Grid Computing (CCGrid), 2015 15th IEEE/ACM International Symposium on. IEEE, 2015. p. 1237-40.
37. RShiny Application Hosting Server: Available from: https://www.shinyapps.io/. [Last Access on 2016 Sep 01].
38. Machine Learning Algorithms: Weka 3.7 ‘s Description with this Set of Tools You Can Extract Useful Information From Large Databases. You Can Work with Filters, Clusters, Classify Data, Perform Regressions, Make Associations, etc., Weka Includes Two Executable Options: Command Line or Graphical User Interface (GUI). Available from: http://www.weka.software.informer.com/3.7/. [Last Access on 2016 Sep 01].