Time series analysis of diabetes patients: A case study of Jigme Dorji Wangchuk National Referral Hospital in Bhutan

Thukten Singye and Suntaree Unhapipat
Department of Mathematics, Faculty of Science, Mahidol University, Bangkok 10400, Thailand.
Email: thuktensingye@gmail.com

Abstract: Diabetes has become a concern in Bhutan with its growing number of patients reported to the hospital. Increasing number of patient is likely to result in rising demand for the medical emergencies. Due to having only three referral hospitals in Bhutan, it is important to forecast the future incidences and prepare with proper resource planning. The monthly number of Diabetes patients obtained from Jigme Dorji Wangchuk National Referral Hospital (JDWNRH) is fitted by autoregressive integrated moving average (ARIMA) model. Such dataset starting from January 2006 to December 2016 is divided into two sub datasets; training dataset (January 2006 – December 2015) and validation dataset (January 2016 – December 2016). Using ARIMA, several models were evaluated based on the Bayesian Information Criterion (BIC) and Ljung-Box Q statistics. ARIMA(0, 1, 1) is the best model to describe and predict the future trends of Diabetes incidences on considering the simplest parsimonious lowest order model. Therefore, the proposed model will help to plan appropriately and allocate resources for emergencies.

1. Introduction
Diabetes is a serious life-long disease in which levels of glucose in the blood and urine become too high because the body’s ability to produce or respond to the hormone insulin released by the pancreas gets impaired and cannot move the glucose into the cells. There are two main types of diabetes which are Type 1 and Type 2. These two different types of diabetes have different effects and can be caused by different factors. However, they both can cause serious health complications and need to be treated and managed properly.

Type 1 diabetes is a condition when the pancreas fails to produces insulins because the insulins producing cells are damaged or destroyed, so this causes the elevation of glucose level in the blood. Type 2 diabetes is a condition when the body does not produce enough insulin or there is extremely less insulin to move the glucose to the cells moreover it is also a case when the insulin fails to work properly. Therefore, the sugar level in the blood is risen.

Bhutan is a land lock country situated in the heart of the great Himalaya where its maximum proportion of population depends on the agriculture. However, shift onto the sedintary lifestyle give raise to the diseases like diabetes, heart disease, obesity and some other in the recent years. It is a great concern for country like Bhutan whose population is just about 808,000 as per the United Nations, Department of
Economic and Social Affairs, Population Division (2017)[1]. Out of which 47 percent of population falls under the age group of 25 to 59, so it is in fact alarming to know that this working group of people are the victim of these diseases.

Diabetes Mellitus (DM) is currently a major problem prevailing in the country, which results in the lifelong complication and heavy health expenditure. As per WHO International Diabetes Federation, they estimated 422 million adults were living with diabetes in 2014. Therefore, it showed an increase in associated risk factors such as being overweight or obese.

Unfortunately, the country has only three-referral hospital, few district hospitals and BHU (Basic Health Units). However, they are not that equipped with the infrastructures, equipment and with skilled labor. So ultimately, pressure will come to the referral hospitals because sophisticated laboratory tests are essential to differentiate between type 1 diabetes and type 2 diabetes. To understand whether it requires insulin injections for survival or is it where the body cannot properly use the insulin it produces. Type 2 diabetes are most common of people with diabetes and it had occurred maximum with the adults and few children in our dataset. Therefore, in this paper we will find the best-fit models to forecast the future trends for diabetes patients as a whole and type 2 diabetes patient differently and forecast them subsequently.

Recently, there are a number of literatures using time series to model a clinical data. The followings are some time series application for forecasting number of patients in many diseases. However, very few papers are available for diabetes incidence researches using the time series model. In 2005, Earnest et al. under BioMed Central (BMC) health services research uses autoregressive integrated moving average (ARIMA) models to predict and monitor the number of beds occupied during a severe acute respiratory syndrome (SARS) outbreak in tertiary hospital Singapore. They found that the ARIMA (1, 0, 3) model was able to describe and predict the number of beds occupied during the SARS outbreak well. The mean absolute percentage error (MAPE) for the training set and validation set were 5.7% and 8.6% respectively, which they found was reasonable for use in the hospital setting. Furthermore, the model also provided three-day forecasts of the number of beds required. Total number of admissions and probable cases admitted on the previous day were also found to be independent prognostic factors of bed occupancy [2].

In 2011, Liu, et al. studied the forecasting incidence of hemorrhagic fever with renal syndrome in China. By using the ARIMA model, they worked on Chinese hemorrhagic fever with renal syndrome (HFRS) data during 1978 to 2008. The best model is selected based on minimum value of Akaike Information criterion (AIC) and Ljung-Box test were used. They assess their validity of their proposed model using the MAPE between the observed and fitted HFRS incidence. Finally, they used the fitted ARIMA model to forecast the incidence of the years 2009 to 2011 [3]. In 2014, Dan et al. used SARIMA model to forecast malaria mortality rate. In this study, they used Box-Jenkins methodology to build ARIMA for malaria mortality rate for the period 1996 to 2013 with a total of 216 data point. They did the one-year ahead forecast. The best model is selected based on minimum values of AIC and Schwarz Bayesian Information Criterion (SBC). The model ARIMA(1,1,1)×(0,0,1)12 was recommended to be the best model for forecasting the malaria mortality rate for the coming period of January 2014 to December 2014 [4]. In 2015, Appiah et al. has also done time series analysis of Malaria cases in Ejisu Juaben Municipality and found ARIMA (2, 1, 1), autoregressive process of order 2, differencing of order 1 and moving average of order 1 best fit for the secondary data. Using the obtained model, they forecasted for the next two years from 2014 and 2016 [5]. In 2016, Pan et al. used the ARIMA model for forecasting the patient number of epidemic disease. They have used actual data of every day patient number of epidemic disease between January and August 2014, in total 223 days, which they obtained from real life CDC (center of disease control). They identify time series model of ARIMA (7, 1, 0) best fit with the forecast accuracy 92.1% [6]. In 2017, Villani et al. BMC Health Services Research forecasted the prehospital EMS demand for the
emergencies using the time series modelling. Four-one thousand four hundred and fifty-four prehospital diabetic emergencies who attended within the period from January 2009 until December 2015. The SARIMA \((0, 1, 0) \times (0, 1, 0)_{12}\) model provided the best fit, with a MAPE of 4.2\% [7]. Therefore, time series model is an appropriate model dealing with the number of patients. We will use time series to model and predict the diabetes patient dataset obtained from JDWNRH of Bhutan by using Box-Jenkins approach.

2. Data collection and methodology

2.1 Data collection
We are using the monthly dataset of diabetes patient reported to the referral hospital from 2006 to 2016 for this study. The dataset is retrieved from Jigme Dorji Wangchuk National Referral Hospital (JDWNRH), Bhutan. Such data set is divided into two sets; training data (January 2006 – December 2015) and validation data (January 2016 – December 2016). Model generation is done based on the training data set.

2.2 Methodology
We used Box-Jenkins approach to forecast the number of diabetes patients of JDWNRH for next two years using the existing data from 2006 until 2016. The Box-Jenkins forecasting models are based on statistical concepts and principles as follows.

2.2.1 Identification
This is a stage of determining necessity of differencing, to produce stationarity and defining the order of seasonal and non-seasonal Autoregressive (AR) and Moving average (MA) operators for the series. Overall in this step appropriate structure Autoregressive Integrated Moving Average (ARIMA) and the order of the model is specified. Therefore, autocorrelation function (ACF) and partial autocorrelation function (PACF) plays a vital role in the identifying the AR and MA order.

2.2.2 Estimation
The second step is to estimate the coefficients of the model. Coefficient of AR models can be estimated by least-square regression. The parameter estimation usually requires more complicated iteration procedure but using the computer programing automatically generates it. In this study, we use the Statistical Package Social Sciences (SPSS 18) software to estimate the coefficient.

2.2.3 Diagnostic checking
In this step, we verify whether the estimates of the parameters are significant. Then we ensured the residual plots is a white noise. Once the model is fitted, we need to do diagnostic checking by studying the ACF of residual. If all the autocorrelation and partial correlation are small then the model is working fine but if some of the autocorrelations are large, the values of autoregressive order \(p\) or moving average order \(q\) are adjusted and the model is re-estimated. The checking of residuals and adjusting the values \(p\) and \(q\) continues until the resulting residuals contain no additional structure. Once appropriate model is selected, the program may be used to generate forecasts and associated probability limits.

2.2.4 Model selection
Finally, the best ARIMA model is selected by using the Bayesian Information Criterion (BIC) and the selected model is used for forecasting the future diabetes patients of the JDWNRH.
2.3 Box-Jenkins method

2.3.1 The Autoregressive AR \((p)\) process
The AR model is a value of time series after regression on previous values from that same time series. It expresses time series as a linear function of past values. The AR model of order \(p\), denoted by AR \((p)\), is

\[ y_t = \phi_0 + \sum_{i=1}^{p} \phi_i Y_{t-i} + \epsilon_t \]  \hspace{1cm} (1)

where \(\phi_0\) is a constant, \(\phi_1, \phi_2, \phi_3, \ldots, \phi_p\) are the model parameters and \(\epsilon_t \sim N(0, \sigma^2)\) is a white noise error.

2.3.2 The Moving-Average MA\((q)\) process
The MA model of order \(q\), denoted by MA \((q)\), is given as

\[ y_t = \epsilon_t + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} \]  \hspace{1cm} (2)

where \(\epsilon_t, \epsilon_{t-1}, \epsilon_{t-2}, \ldots, \epsilon_{t-q}\) are white noise error and \(\theta_1, \theta_2, \theta_3, \ldots, \theta_q\) are model parameters

2.3.3 The Autoregressive Moving Average (ARMA) process
If the data perfectly stationary, we can simply model in the format in the model below. By using the linear combination of AR and MA models, given in the equation (1) and (2), the ARMA model of order \((p, q)\) can be written as

\[ y_t - \phi_1 y_{t-1} - \cdots - \phi_p y_{t-p} = 1 - \theta_1 B - \cdots - \theta_q B^{q} \epsilon_t \]  \hspace{1cm} (3)

Further can be abbreviated as:

\[ \phi_p(B) Y_t = \theta_q(B) \epsilon_t \]  \hspace{1cm} (5)

where \(\phi_p(B) = 1 - \phi_1 B - \cdots - \phi_p B^{p}\) and \(\theta_q(B) = 1 - \theta_1 B - \cdots - \theta_q B^{q}\).

2.3.4 The ARIMA \((p, d, q)\) forecasting equation
However, it is impossible to have the data always stationary. Therefore, to make the process we have to do differencing in non-stationary series so the differenced series \((1-B)^d Y_t\) must be add in the process ARMA \((p, q)\).

Where the \(a_i\) is a sequence of identically distributed uncorrelated deviates, referred to as "white noise".
In many situations where differencing is employed, a non-zero constant term \( \theta_0 \) will not be required. For brevity, the equation is generally written as:
\[
\Phi_p(B)(1 - B)^dY_t = \theta_0 + \theta_q(B)\epsilon_t, \tag{6}
\]

Where
\[
y_t = a_t + \sum_{i=1}^{p} \phi_i Y_{t-1} + \epsilon_t + \sum_{j=1}^{q} \theta_t \epsilon_{t-1}
\]

and \((1 - B)^dY_t\) is a differenced series and \(\epsilon_t\)'s is a sequence of identically distributed uncorrelated deviates, referred to as "white noise"[8].

3. Results and discussion

3.1 Identification
The monthly Diabetes patients reported to the hospital from 2006 to 2016 is plotted in the Figure 1. The data plot display trends, which indicates that the mean is not stationary. However, it is necessary to check variance stationarity first. To investigate the stationarity of variance, Box-Cox transformation is used. Figure 1 shows the time series plot of the monthly number of diabetes patients from JDWNRH, Bhutan.

![Figure 1. Monthly number of diabetes patients from JDWNRH (January 2006- December 2016)](image)
3.1.1 Tests for stationarity
It is necessary to check the stationarity of the variance. We can use Shapiro-Wilk and Kolmogorov-Smirnov\textsuperscript{a} for the normality test. According to the result providing in Table 1, we can conclude that the variance was not constant (\(p\)-value < 0.05). Hence, root transformation which is Box-Cox power transformation is used to stabilize the variance as shown in the Table 2. The transformed series is plotted in Figure 2.

### Table 1. Tests of Normality of the empirical data.

| Diabetes Patients | Kolmogorov-Smirnov\textsuperscript{a} | Shapiro-Wilk |
|-------------------|--------------------------------------|--------------|
| Statistic          | df   | P-value | Statistic | df | P-value |
| .094              | 120 | .011    | .972      | 120 | .014 |

\(a\). Lilliefors Significance Correction

### Table 2. Tests of Normality of the transformed data.

| Root transformation | Kolmogorov-Smirnov\textsuperscript{a} | Shapiro-Wilk |
|---------------------|--------------------------------------|--------------|
| Statistic           | df   | P-value | Statistic | df | P-value |
| .078               | 120 | .070    | .972      | 120 | .496 |

\(a\). Lilliefors Significance Correction

The power transformation analysis shows that the need of root transformation for the data. Further, to make the mean stationary, first order differencing is used. The monthly root transformed diabetes data with first order differencing is plotted in the Figure 2. Notice that the root transformed with differenced series is mean stationary and constant variance.

![Figure 2. Time series plot of Diabetes patients after the first differencing.](image-url)
3.1.2 Graphical representation of autocorrelations

From the ACF plots below (Figure 3), the value of its function is less than one. The parameter of any given function is less the one, it means that there is stationarity.

![Sample ACF plot of monthly square root transformed Diabetes data after first differencing](image)

**Figure 3.** Sample ACF plot of monthly square root transformed Diabetes data after first differencing

The plot shows a negative spike at lag 1 and there is no other spikes showing that it is over differenced. Further the plot decays to zero on the both sides of the mean. From the Figure 3 it can be concluded that an appropriate model is moving average process of order 1, i.e., MA(1).

![Sample PACF of monthly square root transformed Diabetes data after first differencing](image)

**Figure 4.** Sample PACF of monthly square root transformed Diabetes data after first differencing
Observing the plots of ACF and PACF plot above, it could be noticed that the ACF displays the sharper
cuts off compared to the PACF. According to the PACF plot in Figure 4, we use the autoregressive
process of order 2, i.e., AR(2).

Based on the Ljung-Box Q statistics, the models with significant p-values are selected and shown in
Table 3 below. The resultant BIC providing in Table 4 shows that the most appropriate model to fit
diabetes patients’ data is ARIMA(0, 1, 1). We are going to estimate parameters of the model in the next
subsection.

**Table 3.** Competitive models for the monthly diabetes patients based on the Ljung-Box Q statistics test

| Model         | Q statistic | P-value |
|---------------|-------------|---------|
| 1. ARIMA (0, 1, 0) | 79.624      | 0.000   |
| 2. ARIMA (1, 1, 0) | 50.858      | 0.000   |
| 3. ARIMA (0, 1, 1) | 21.669      | 0.198   |
| 4. ARIMA (1, 1, 1) | 21.069      | 0.176   |
| 5. ARIMA (2, 1, 0) | 23.806      | 0.094   |
| 6. ARIMA (2, 1, 1) | 19.379      | 0.197   |

**Table 4.** RMSE and BIC of all possible models resulting in Table 5.

| Model         | RMSE | BIC   |
|---------------|------|-------|
| ARIMA (0, 1, 1) | 0.678| -0.736|
| ARIMA (1, 1, 1) | 0.670| -0.639|
| ARIMA (2, 1, 0) | 0.701| -0.551|
| ARIMA (2, 1, 1) | 0.671| -0.598|

### 3.2 Diagnostic checking and parameter estimation

Diagnosis checking is also known as verification. It is to do with the testing of the goodness of fit test
statistics of a model. We study the ACF and PACF of the residual plot to see if it is white noise. If all the
autocorrelation and partial correlation are small then the model is working fine but if some of the
autocorrelations are large, the values of $p$ or $q$ are adjusted and the model is re-estimated. The checking of
residuals and adjusting the values of $p$ and $q$ continues until the resulting residuals contain no additional
structure.

Form Figure 5 the sample ACF and PACF of the model shows that the autocorrelations of the residual
are all close to zero which mean they are uncorrelated, hence the residual assume mean of zero and
constant variance. Finally, the p-value (0.198) for the Ljung-Box statistic clearly exceeds 5% for all lag
orders. Thus, the selected model ARIMA (0, 1, 1) satisfies all the model assumptions.

Table 5 shows the estimated parameter of ARIMA(0, 1, 1). Therefore, the estimated model is $y_t = \varepsilon_t - 0.825\varepsilon_{t-1}$.

**Table 5.** Parameter estimates of model ARIMA(0, 1, 1)

| Model         | Estimate | SE  | t    | P-value |
|---------------|----------|-----|------|---------|
| ARIMA(0, 1, 1)| Square Root Transformation | Difference | 1     |         |
|               |          | MA Lag 1 | .825  | .054    | 15.297  | .000    |
3.3 Forecasting using ARIMA (0, 1, 1)

The forecast values with 95 percent forecast limit of the ARIMA (0, 1, 1) of model for monthly diabetes patients are shown in Table 6 with standard error, lower and upper limit and its actual forecasted values.

Table 6. Forecast updates for 1 year- monthly Diabetes patient reported to the hospital

| Date    | Diabetes patient | Forecasted value | 95% confidence limit | Error |
|---------|------------------|------------------|----------------------|-------|
|         |                  |                  | Lower                | Upper |
| Jan 2016| 9                | 10.89            | 3.92                 | 21.44 |
| Feb 2016| 4                | 10.89            | 3.84                 | 21.62 |
| Mar 2016| 13               | 10.89            | 3.76                 | 21.81 |
| Apr 2016| 16               | 10.89            | 3.69                 | 22.00 |
| May 2016| 8                | 10.89            | 3.61                 | 22.18 |
| Jun 2016| 10               | 10.89            | 3.53                 | 22.37 |
| Jul 2016| 8                | 10.89            | 3.46                 | 22.56 |
| Aug 2016| 10               | 10.89            | 3.39                 | 22.66 |
| Sep 2016| 10               | 10.89            | 3.35                 | 22.85 |
| Oct 2016| 9                | 10.89            | 3.28                 | 23.04 |
| Nov 2016| 7                | 10.89            | 3.20                 | 23.23 |
| Dec 2016| 10               | 10.89            | 3.13                 | 23.43 |
4 Conclusion
In this paper, we studied the Diabetes Patients using the ARIMA modelling and forecasting method. Box-Jenkins was used to study the monthly Diabetes patients reported to the JDWRNH in Bhutan for the period of January 2006 to December 2016. Our main intension for this study is to forecast the monthly diabetes case for the coming period of January 2017 to December 2018. Table 7 below shows the forecast for next two years and Figure 6 is the scatter plot, it determine potential relationships among Actual and Forecast value. List of tentative ARIMA models are developed based on the Ljung-Box Q statistic test, BIC and after testing the significant of the estimated parameters, ARIMA(0, 1, 1) model was seen best fit for forecasting the Diabetes patient.

| Months   | 2017 Lower limit | forecast | Upper limit | 2018 Lower limit | forecast | Upper limit |
|----------|------------------|----------|-------------|------------------|----------|-------------|
| January  | 3.06             | 10.89    | 23.52       | 2.43             | 10.89    | 25.50       |
| February | 3.03             | 10.89    | 23.72       | 2.37             | 10.89    | 25.70       |
| March    | 2.96             | 10.89    | 23.91       | 2.34             | 10.89    | 25.81       |
| April    | 2.89             | 10.89    | 24.11       | 2.28             | 10.89    | 26.01       |
| May      | 2.86             | 10.89    | 24.21       | 2.34             | 10.89    | 26.11       |
| June     | 2.79             | 10.89    | 24.40       | 2.19             | 10.89    | 26.11       |
| July     | 2.722            | 10.89    | 24.60       | 2.16             | 10.89    | 26.42       |
| August   | 2.69             | 10.89    | 24.70       | 2.10             | 10.89    | 26.63       |
| September| 2.62             | 10.89    | 24.90       | 2.07             | 10.89    | 26.73       |
| October  | 2.56             | 10.89    | 25.00       | 2.02             | 10.89    | 26.83       |
| November | 2.53             | 10.89    | 25.20       | 1.99             | 10.89    | 27.04       |
| December | 2.46             | 10.89    | 25.40       | 1.96             | 10.89    | 27.14       |

Figure 6. Scatter plot for Actual and Forecast value
References
[1] United Nation: New York. (2017). Department of Economic and Social Affairs. Retrieved from World Population Prospects: https://esa.un.org/unpd/wpp/Publications/Files/WPP2017_KeyFindings.pdf.
[2] Earnest, A., Chen, M.I., Ng, D., & Sin, L.Y. (2005, May 11). Using autoregressive integrated moving average (ARIMA) models. Retrieved from BMC Health Services Research: https://bmchealthservres.biomedcentral.com/track/pdf/10.1186/1472-6963-5-36?
site=bmchealthservres.biomedcentral.com.
[3] Liu Q., Liu X., Jiang B. and Yang W. (2011). Forecasting incidence of hemorrhagic fever with renal syndrome in China using ARIMA model.
[4] Dan D. E., Jude O and Idochi O. (2014). Modelling and Forecasting Malaria Mortality Rate using SARIMA Models (A Case Study of Aboh Mbaize General Hospital, Imo State Nigeria). Science Publishing Group.
[5] Sampson, A. T., Henry, O., & Nabubie, I. B. (2015, June 06). Time Series Analysis of Malaria cases in Ejsu-Juaben Municipality. Retrieved from International journal of scientific and technology research volume 4.: http://www.ijstr.org/final-print/june2015/Times-Series-Analysis-Of-Malaria-Cases-In-Ejsu-juaben-Municipality.pdf
[6] Pan, Y., Zhang, M., Chen, Z., Zhou, M., & Zhang, Z. (2016). An ARIMA Based Model for Forecasting the Patient. Retrieved from Shenzhen University: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7538560
[7] Villani, M., Earnest, A., Nanayakkara, N., Smith, K., Courten, B., & Zoungas, S. (2017). Time Series Modelling to forecast Prehospital EMS demand for diabetic emergencies. Retrieved from BMC Health Services Research: https://bmchealthservres.biomedcentral.com/track/pdf/10.1186/s12913-017-2280-6?site=bmchealthservres.biomedcentral.com.
[8] Wei W.S. (1990). Time series analysis, Addison-Wesley, USA.

Acknowledgements
We would like to thank Thai International Cooperation Agency (TICA) for providing the financial support. We also would like to thank Jigme Dorji Wangchuk National Referral Hospital (JDWNRH), Bhutan for providing and letting us use latest available data.