Bits & Bytes: Time Series and Its Analysis

|  |  |  |
| --- | --- | --- |
|  | | |
| What a nice drawing?  If you see this pattern while studying data analytics, you might catch that there is a data flow over time. (Something changes over time.)  We call this kind of data TIME SERIES. | | |
| https://upload.wikimedia.org/wikipedia/commons/thumb/8/86/First_flight2.jpg/220px-First_flight2.jpg | https://upload.wikimedia.org/wikipedia/commons/thumb/a/ab/DeHavilland_DH50_biplane.jpg/220px-DeHavilland_DH50_biplane.jpg | https://upload.wikimedia.org/wikipedia/commons/thumb/4/44/DeHavilland_Comet.jpg/220px-DeHavilland_Comet.jpg |
| After world war II, one of big changes was the popularity in international air travel, driven by rapidly advancing technology, modernised industry structure and growth of market demand. Statisticians Box & Jenkins collected monthly data about how many people travelled using international flight, their first recording was 112000 in January 1949 . This number is much smaller than the daily number of international travellers nowadays (i.e. 11M in 2017). Their data collection finished in December 1961 with 432000 international travellers. This means that the data were collected for 12 years, and so there are 144 data points (12 years \* 12 months) as the data were collected monthly. The Data looks like 112, ..., …, …, …, …, 432 (1000) – very simple and with no other information provided. The actual time series plot from the international flight data is the plot below, in which we can see the time on the horizontal axis and changes in the number of passengers on the vertical axis. | | |
|  | | |
| Imagine that a chef has a piece of nice meat without any other ingredients or spices. It might be very challenging to make a delicious dish – they might need salt and pepper at least.  It is same in analysing TIME SERIES…  In time series analysis, there is no other information that we can add to the consideration. Thus you might assume that it is very easy to analyse, as the data structure is very simple. The reality is opposite to this assumption however. Time series are one of the hardest analytical data forms as they is too simple. To mitigate this challenge it is very important to analyse different time frames such as long-term to short-term diversities of data. We normally consider four different time frames: long-term, mid-term, short-term and each data point.  First, we consider LEVEL - the overall value of the whole data, which is the average value of all 144 data points (approximately, 280300). The shape of the LEVEL is not exciting at all- just a horizontal line that has the height of 280300 as shown in figure below. | | |
|  | | |
| Second, we consider TREND. This is the mid-term analysis to help us to figure out the change of the data over time. The easy way of capturing TREND is that Ignore the fluctuations and just draw a straight line from the fist value towards the last value, though the trend line does not necessarily to be straight. We can easily capture that the trend line for this data is close to a straight line as shown below. | | |
|  | | |
| Next, we count SEASONALITY. This is a short-term analysis to find any seasonal effects in any regular time period. When the SEASONALITY has been separated from the whole data as shown below, it is really clear that there is a repetitive seasonal effect with the same strength for each year. It is a bit of trick in time series that we can decompose the seasonality like that – change the height (strength) of the seasonal effects in respect to the yearly data values. | | |
|  | | |
| Last, we look at RESIDUAL - the individual data points as the difference from the counts of LEVEL, TREND and SEASONALITY all put together. Look at the plot below. There seem to be some sort of seasonal effects in the residual values, though they are not very clear like the previous plot of SEASONALITY. | | |
|  | | |
| To talk about time series model, I could not avoid the use of a technical term. Namely ARIMA – this is the composite of AR, I and MA. The reason is that almost all the time series models are based on ARIMA:   * AR (Autoregression): Arrange the data in time order and copy it, then shift the copy one-step, two-step, and so on (namely lagged) to find the dependency relationship between time-shifted values. * I (Integrated): Calculate the difference between the current and previous values to catch the level of stability of data. * MA (Moving Average): Calculate the average and error for a framed window (with a fixed time-frame) and shift the window to continue the computation of average and error – this activity is to smooth the values.   Though there are more technical stories in the depth of the ARIMA, I suppose that this is enough to introduce its concept today.  The core value of any TIME SERIES is FORECASTING. To do this, we need to find the best formula for predicting the future values. When I learnt about time series a few years ago, I needed to apply a number of different models manually and spent a lot of time to select an appropriate formula (model). These days, with the advanced development in tools (packages) it is easy to automate this process. I used one of the auto-functions to fit the model. From the model fitting, I found the prediction value equals the sum of these four components:   * 0.6 times of lag-2 autoregression * 0.2 times of 12-month seasonality * -0.99 times of lag-1 errors * Random errors (residuals)   Exciting time… P..R..E..D..I..C..T..I..O..N…  To forecast some future values, say for the next three years (i.e. 36 months) using the formula above, just compute for the time stamps 145 to 180 (remember that we already have 144 historical data points).  The blue line is the prediction of the number of international travellers using air travel for three years after actual observation (i.e. 1962, 1963 and 1964) with a variation error of 95% confidence level (shaded). From this model, the total number passengers in 1964 is about predicted to be 6.9M. | | |
|  | | |
| I was curious about how accurate the prediction from the model is. So I searched the Internet to get the actual statistics of the total number of international flight passengers. Unfortunately, I could not find the information for 1964. The earliest year of statistics from the International Civil Aviation Organization is for 1970, and the number for that year is over 310M (as shown below).  Can we predict from our model as far as 1970 to compare the actual statistic? Not in actuality…  This is because the forecasting for a long future time is normally nonsense.  However if you are still curious about this, the prediction for 1970 is 9.4M. That is far behind of the actual figure of 310M that we expected. Interestingly, the number of international air travellers is continuously increasing not only linearly but also in a gentle curve, though this is out of topic for today, haha…  Hope this is a bit of information on TIME SERIES for you. | | |
| <https://data.worldbank.org/indicator/IS.AIR.PSGR> | | |