Models of predictive modeling on the example of a gas turbine

Echkina E., Lvov V.
CMC, Moscow State University, Moscow, Russia

ejane@cs.msu.ru, valckaaz@gmail.com

Abstract. The ability to accurately predict the operation of a particular mechanism and, on the basis of this, estimate the equipment life is a very important task. The amount of losses of the enterprise can depend on such study, as well as the health of many people whose lives depend on the health of the working installation. As part of this work, the main time series models were considered and the most suitable for the study was selected. The operation of a gas turbine was studied and a forecast was made. On the basis of the study, linear regression, ARIMA and moving average models were built and evaluated.

1. Introduction

Research and analysis of time series as a separate area originated in the second half of the last century. Box and Jenkins published their most famous work on time series back in 1976. This work gave rise to the entire field of predictive analysis.

Today, predictive analysis is used in many areas of modern life and helps to answer the question: “What is more likely to happen?”. Changes in the exchange rate and epidemiological situation, productivity, conversion of online stores, weather, illness or the efficiency of an enterprise - all this is predicted by specialists every day. In manufacturing, huge amounts of data are analyzed, often by AI, to improve business and manufacturing operations. Every year more and more companies are introducing new generation technologies into their production process.[1]

But predictive analysis is used not only, for example, to determine the amount of goods to be produced, but also to answer such a basic, but incredibly important question: “How long will the equipment work and how will it behave tomorrow, the day after tomorrow or in a week?”

Each individual task of predictive analysis requires its own method of solution, and sometimes the most sophisticated and modern methods are not needed to answer certain questions. And often simple but proven and straightforward model gives the most accurate answer. In this article we want to acquaint the reader with classic basic models of predictive modeling using the example of analyzing the operation of a gas turbine. In future we will use more sophisticated methods to analyze the operation of gas turbine pumps.

2. Basic time series models

This chapter will review and describe the basic existing time series models. The choice of a model for each specific time series depends primarily on the applied problem. Also, an equally important factor is the subject area of the series under study. The physical phenomenon, on the basis of the
observations of which the time series was built, has a direct impact on the type, form and method of processing the data of this series.

2.1. Regression model

A regression problem is the problem of predicting the value of a real number \( y \) from the values of the predictor(s) \( x_i \). An example of a regression problem can be the analysis of the value of the price of a house depending on various parameters: area, number of rooms, city district, etc.

In the simplest case (figure 1), the regression model reflects the relationship between the variable \( y \), the value of which you want to predict, and the only predictor variable \( x \):

\[
y_t = \beta_0 + \beta_1 x_{1,t} + \epsilon_t.
\]

![Figure 1. Example of simple linear regression.](image)

2.1.1. Multiple Linear Regression. Multiple Linear Regression is the same linear regression model, but with more predictor numbers added. General view of the multiple linear regression model:

\[
y_t = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \cdots + \beta_k x_{k,t} + \epsilon_t,
\]

where the \( x_1 \) to \( x_k \) values are predictors of the \( y \) variable. One of the ways to write this equation is matrix representation. Let:

\[
y = (y_1, \ldots, y_T)'\quad \epsilon = (\epsilon_1, \ldots, \epsilon_T)', \quad \beta = (\beta_0, \ldots, \beta_k)'.
\]

Then

\[
X = \begin{bmatrix}
x_{1,1} & x_{2,1} & \cdots & x_{k,1} \\
x_{1,2} & x_{2,2} & \cdots & x_{k,2} \\
\vdots & \vdots & \ddots & \vdots \\
x_{1,T} & x_{2,T} & \cdots & x_{k,T}
\end{bmatrix}
\]

And in that case \( y = X\beta + \epsilon \).

When building a linear regression model, be aware of the implicit assumptions we make about variables. First, an assumption is made about a linear relationship with its predictors. Second, those predictors are not just some random numbers. Third, the errors \( (\epsilon_1, \ldots, \epsilon_T) \) errors have a standard normal distribution, do not correlate with each other, and do not have relationship with predictors.

To select the coefficients \( \beta_0, \beta_1, \ldots, \beta_k \), the principle of least squares is used. This approach, as the name suggests, is based on minimizing the squared errors \( \epsilon_t \) to find the coefficients

\[
\sum_{t=1}^{T} \epsilon_t^2 = \sum_{t=1}^{T} (y_t - \beta_0 - \beta_1 x_{1,t} - \beta_2 x_{2,t} - \cdots - \beta_k x_{k,t})^2.
\]

This is called least squares estimation. Finding the linear regression coefficients is called "training" the model. The coefficients \( \hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k \) obtained by this method are called fitted.

Writing this equation in matrix representation, we get \( \epsilon' \epsilon = (y - X\beta)'(y - X\beta) \). This shows that this equation takes its minimum value when \( \hat{\beta} = (X'X)^{-1}X'y \).
The following parameters can be used as predictor variables for time series linear regression models:

- Time series lags.
- Statistics related to the sliding window - a set of previous values of the series:
  - The maximum or minimum value in the window.
  - Average / median value of the series in the window.
  - Standard deviation.

Date and time parameters:

- Day of the week, specific hour of the day, etc.
- A binary value representing whether the day is a holiday or some other special day.
- Values predicted by other models.

2.2. Classic Box-Jenkins models.

Autoregressive models invented by Box and Jenkins are one of the most popular forecasting methods [1,10]. The most frequently used model of autoregressive moving average (ARMA) is based on two models: autoregressive (AR, autoregressive) and moving average (MA, moving average) [2].

When considering the Box-Jenkins models, it is customary to divide the series into stationary and non-stationary. The stationary series is characterized by the average constant unchanged value of the series (the level of the series), as well as the presence of constant variance. This means that the series that have a seasonal pattern or trend are not stationary. It should be understood that the series in which non-seasonal cyclicity can be observed can be stationary.

In the paper by Box and Jenkins, it is said that autoregressive models are extremely useful when examining time series encountered in practice. The model describes each value of the series as a linear combination of the previous values of the series and the error, also known as "white noise".

\[ y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \varepsilon_t \]

where \( \varepsilon_t \) is a “white noise”. Such a model is called an autoregressive model and is denoted AR (p), where p is the order of autoregression.

If we introduce the offset operator \( L \): \( Ly_t = y_{t-1} \), then the autoregressive model can be written as follows:

\[ y_t = c + \sum_{i=1}^{p} \phi_i L^i y_t + \varepsilon_t \]

The stationarity of the process is achieved when the roots of the characteristic polynomial \( \Phi(z) = 1 - \sum_{i=1}^{p} \phi_i z^{-i} \) lie outside the unit circle in the complex plane \( |z| > 1 \) [3 pp. 88.90].

The moving average model takes a slightly different approach to determine the value of a series. Rather than using a weighted value of a specified number of prior values in a series, the moving average model combines prior errors to build a forecast.

\[ y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q} \]

where \( \varepsilon_t \) stands for white noise. Such a model is called the q-order moving average model and is denoted as MA (q), where q is the moving average order. It should be mentioned that in addition to this model, there are models of weighted, exponential, cumulative moving average.

By combining the autoregressive and moving average models, we get an autoregressive moving average (ARMA) model.

\[ y_t = c + \varepsilon_t + \sum_{i=1}^{p} (\phi_i y_{t-i}) + \sum_{i=1}^{q} (\theta_i \varepsilon_{t-i}) \]

This model is intended for the analysis of stationary series.

To study non-stationary series, an autoregressive integrated moving average (ARIMA) model was developed. The difference between the ARIMA model and ARMA is that the ARIMA model takes the differences of the original series of order d as input values.
\[ y_t' = c + \phi_1 y_{t-1}' + \cdots + \phi_p y_{t-p}' + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t, \]

Where \( y_t' \) are the differences of the series. This model is denoted as ARIMA (p, d, q), where p is the autoregressive order, d is the order of integration, q is the moving average order. The confidence interval can be calculated using the following formula:

\[ y_{t+1}' \pm 1.96 \hat{\sigma} \]

where \( \hat{\sigma} \) is the standard error variance.

When constructing this model, it is assumed that the series will be stationary after subtracting adjacent values. However, in practice, this does not always happen. If the time series has seasonal patterns, the integration of the series will not bring it to a stationary state. Therefore, the development of the ARIMA model was the seasonal ARIMA model - SARIMA (Seasonal autoregressive integrated moving average).

The SARIMA model is obtained by adding seasonal parameters to the ARIMA model. This model is designated as SARIMA (p, d, q) (P, D, Q, s), where the parameters p, d and q are the same parameters as in the ARIMA and the parameters P, D, Q and s are responsible for the seasonal component of the series. P is the autoregressive order of the seasonal component. Q is the moving average order of the seasonal component. D is the order of seasonal integration.

The advantages of autoregressive models are their clear mathematical interpretation. However, the data requirements for using the ARIMA and SARIMA models make them rather difficult to apply in practice. To use the ARIMA model, at least 40 observations are recommended, and for the SARIMA model, about 6-10 seasons [4].

3. Research

After considering the most basic time series models, a study of the provided data obtained from the sensor of the gas turbine (figure 2) was carried out.

![Image of a gas turbine](image.png)

**Figure 2.** Schematic representation of a gas turbine.

On the schematic representation of a gas turbine you can see a number of points highlighted. There are set of vibration sensors, part of them piezo-crystals accelerometers and others are current-vortex proximeters. Sensors measure vibration in three directions: vertically, horizontally and axial. After receiving raw data, transforming force or induction to voltage, digitizing and additional calculations we get a familiar acceleration m/s² or movement in μm. This movement characteristic is often calculated as a squared mean and called Overall level.
Depending on the values of these frequency and time characteristics we can determine the state of the turbine, but also understand what kind of malfunctions should we be weary of. The original series was as follows (figure 3):

![Figure 3. Original time series. X axis is date and y axis is general level.](image)

Since the outliers in the data were associated with intentional turbine shutdown, these values were discarded. And also, the data received before October 2020 was not considered, because, as can be seen in the graph (figure 4), the installation was reconfigured this month and the data up to that point was no longer relevant.

![Figure 4. Series after transformation.](image)

3.1. ARIMA model.
As described above, when constructing the ARIMA model, the assumption is made that the time series will be stationary after subtracting adjacent values. The series was checked for stationarity using the Dickey-Fuller test (figure 5).

![Figure 5. Above we can see the original series and the p-value for Dickey-Fuller test. On the bottom left is the autocorrelation plot and on the bottom right is a partial autocorrelation plot. Their y axis’s correspond to p-values of correlation and x axis’s- to the number of lags.](image)
On autocorrelation plot we can see the correlation between every value and the next. Partial autocorrelation plot shows us the correlation between last value of the series and previous 100 values. It's clear that the series are not stationary. An explicit autocorrelation of data is visible. From partial autocorrelation plot we can see that the value of the series and it’s first lag heavily correlate. And when subtracting adjacent values, the series becomes stationary (figure 6).

![Figure 6](image.png)

**Figure 6.** On top is the series after subtracting first lag from every value. Dickey-Fuller p-value is 0.0. Strong autocorrelation is gone.

As we can see from autocorrelation and partial autocorrelation plots – every value of the series correlates only with previous value. It’s safe to assume that ARIMA (1,1,1) (figure 7) model should do quite well.

![Figure 7](image.png)

**Figure 7.** Result of making ARIMA (1,1,1). The blue and green graphs are the true values of the series, the red ones are the ARIMA forecast, and the gray zone is the 95% confidence interval.

The model was trained on all values of the series, except for the last 20: it was tested on these values. As you can see from the graph, ARIMA was able to give a good forecast and all values of the series fell into the confidence interval.

3.2. Linear Regression Model.
24 lags of the time series will be used as variable predictors in the linear regression model (figure 8).
After building a linear regression model with L1 regularization, the following result was obtained (figure 9):

![Figure 9](image)

**Figure 9.** Result of building linear regression. On the x axis are dates, on y axis are value of general level. Here, the blue graph is the true values of the series, the green one is the forecast, and the red lines limit the 95% confidence interval.

L1 regularization discarded all predictors except the first lag, since its value makes the largest contribution to the current value of the series. Thus, the linear regression model gave a mean absolute percentage error of 0.84%. For such a model, this is a pretty good result. At the same time, no preliminary data verification was required to build this model, as was the case when building the ARIMA model.

3.3. Moving average model.

Since the previous value was found to play the most important role in predicting the current value of the series when building the linear regression model, it was decided to build a simple moving average model with a sliding window of size 2 (figure 10).
Figure 10. Moving average model. Here, the blue graph is the true values of the series, the green one is the forecast, and the red lines limit the 95% confidence interval. Red dots show where series values cross this interval.

The model gave an average absolute percentage error of 0.56%, which is the best result. But due to its simplicity, the forecast cannot be built far into the future.

4. Conclusion.
In this paper, the existing time series models were studied. ARIMA, linear regression and moving average models were built and compared based on the data obtained from the gas turbine sensor. When comparing the constructed models, the moving average gave the best result. But at the same time, due to its simplicity, this model cannot make a forecast far into the future. The ARIMA model gave a good result and all predicted values were within the 95% confidence interval. The linear regression model gave a good result, and at the same time, its construction was simpler than the ARIMA model.

Acknowledgments
We are grateful to BMSsoft d.o.o. for provided data and cooperation.

References
[1] Leurent, H., Boer, E. D., “Fourth Industrial Revolution Beacons of Technology and Innovation in Manufacturing” White Paper, January 10, 2019. https://www.weforum.org/whitepapers/fourth-industrial-revolution-beacons-of-technology-andinnovation-in-manufacturing
[2] Box, George; Jenkins, Gwilym Time series analysis: forecasting and control, rev. ed. // Oakland, California: Holden-Day. — 1976.
[3] Shumway, Robert; Stoffer, David (2010). Time series analysis and its applications : with R examples (3rd ed.). Springer. ISBN 144197864X.
[4] Ханк Д. Э, Уичери Д. У., Райте А. Дж. Бизнес-прогнозирование: пер. с англ. 7-е изд. М.: Вильямс, 2003. С. 506