Features of forecasting the demand for tourism services based on a linear model

Victor Ziborov,
ziborov@ukr.net,
http://orcid.org/0000-0003-4397-1782,
Igor Patrakeyev,
patr@ukr.net,
http://orcid.org/0000-0002-0448-8790,
Yelyzaveta Ziborova,
lizaziborova@gmail.com,
https://orcid.org/0000-0002-7976-6687

KEY WORDS: economic indicator forecasting, demand for tourism services, time series, trend, flat, linear forecasting model, linear regression model, generalized inverse, condition number, method of least squares.

ABSTRACT. An economic indicator forecasting by a quantitative method implies choosing a mathematical model. Since the tourism industry is characterized by a constant market growth from year to year (being a phenomenon of the 20th and 21st centuries), it is advisable to use a linear model for forecasting. In this article, we consider the features of constructing such a model using the least squares method. Using the condition number of the matrix of residual equations, this article provides a concrete example of the most effective solution to this problem.

INTRODUCTION

There are two categories of forecasting methods. The first category is qualitative methods. They are based on the opinions, value judgments of various experts, as well as other consumers of services. This is the brain attack method, the Delphi method [1], the goal tree method, the matrix method, etc.

The second category is quantitative forecasting methods [2]. They are based on the construction of a function (model), depending on the historical data of the time series. At the same time, it is assumed that the dynamics of the time series will continue in the future. This article describes this particular approach to forecasting the demand for tourism services.

The first step of the quantitative method is to assess the state of the time series: the current period is characterized by either an uptrend, or a downtrend, or lack of a trend (that is, a flat) [3]. We will not explain these concepts as they are known. An example of the trend dynamics of a time series is shown in the Figure below.

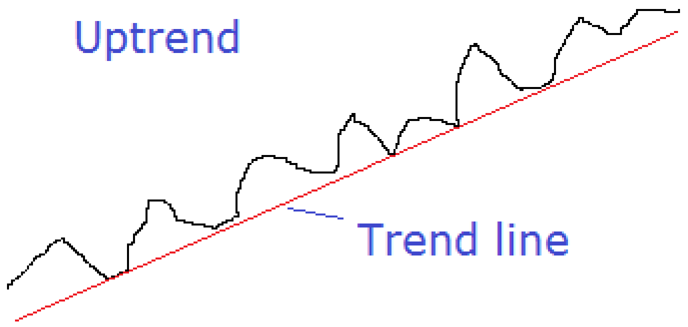

Figure 1. An example of an uptrend

A side trend (flat) is shown in Figure 2. As you can see, the time series does not have a specific direction up or down, it forms a lateral movement. This time dependence manages to exit the flat and form a downward movement.

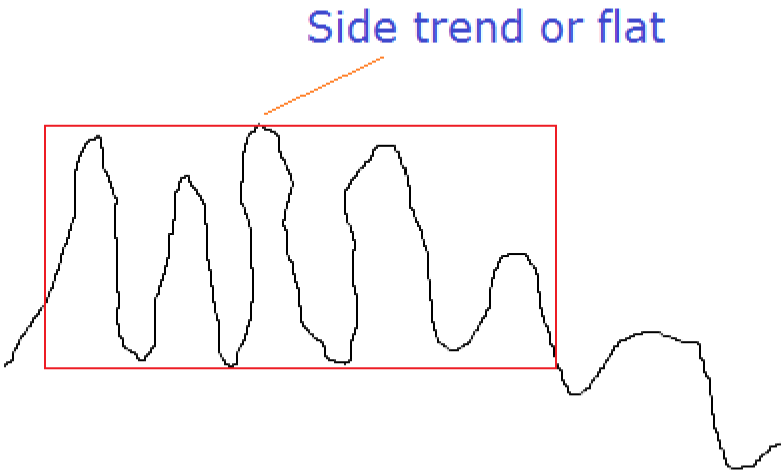

Figure 2. An example of side trend

ECONOMIC INDICATOR FORECASTING METHOD

The second stage of the quantitative method involves the choice of model. When conducting speculative operations in the forex market, with the securities and raw materials market, moving averages are most often used to predict price behavior over time.

However, in the field of tourism we are dealing with an ever-growing market for tourism services [4]. Therefore, we believe a priori that the demand for tourism services is characterized by an uptrend. We will determine the rate of this growth by the closest known values. Choosing a model which we will make a forecast with, we will keep in mind that non-linearity in the general case is poorly predicted. Therefore, this growth relative to the nearest known values of the indicator will be considered linear.

To demonstrate our approach to forecasting, we will solve a specific problem. The number of tourists arriving in Ukraine from 2016 to 2019 is known, see Table 1. These data are obtained from cross-border statistics.

Year, X	Number of arrivals in million tourists, Y
2016	21.73
2017	21.57
2018	22.11
2019	23.49
2020	?

Table 1. The source data for our task

It is required to make a forecast of how many tourists will come to Ukraine in 2020.

As can be seen from the Table, the number of tourists arriving in Ukraine from year to year is constantly growing. We conclude that in the market of travel services there is an upward trend movement. Let's take a linear growth model: $Y = a \cdot X + b$. The time (years) in Table 1 is denoted by the independent variable X, and the number of tourists' arrivals (in millions) is denoted by the dependent variable Y, see Fig. 3.

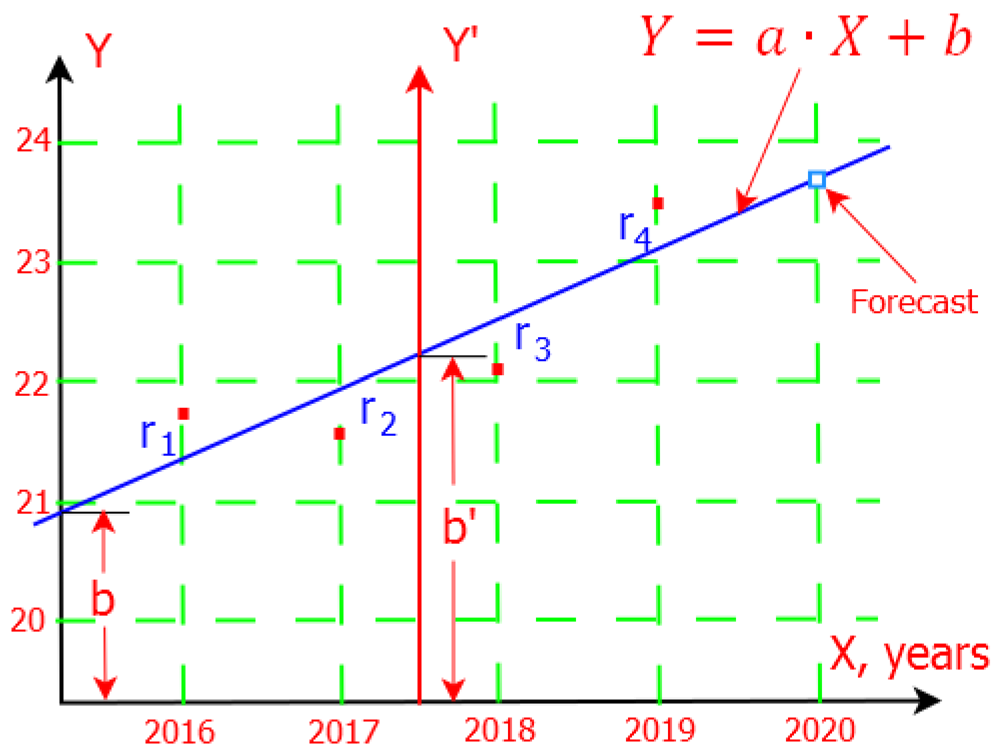

Fig. 3. Graphical representation of a linear model

In this model, the coefficient a shows the growth rate of tourists' arrivals. Geometrically, the coefficient a represents the tangent of the slope of the straight line (see Fig. 3) to the X axis. The coefficient b is the segment that cuts off the line $Y = a \cdot X + b$ on the Y axis.

Thus, we solve the problem of approximating points by the analytic function $Y = a \cdot X + b$ using the least squares method. The points are presented in Table 1.

To solve this problem, following the traditional least-squares method, we compose the residual equations in the general form:

$$r_i = a \cdot X_i + b - Y_i.$$

Next, we compose the residual equations in numerical form:

$$\begin{cases} r_1 = a \cdot 2016 + b - 21.73 \\ r_2 = a \cdot 2017 + b - 21.57 \\ r_3 = a \cdot 2018 + b - 22.11 \\ r_4 = a \cdot 2019 + b - 23.49 \end{cases}$$

We write the last system in matrix-vector form:

$$\begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \end{bmatrix} = \begin{bmatrix} 2016 & 1 \\ 2017 & 1 \\ 2018 & 1 \\ 2019 & 1 \end{bmatrix} \cdot \begin{bmatrix} a \\ b \end{bmatrix} - \begin{bmatrix} 21.73 \\ 21.57 \\ 22.11 \\ 23.49 \end{bmatrix}$$

or:

$$r = A \cdot z - l.$$

Here r is the residuals vector,
A – residual equations matrix,
$z = \begin{bmatrix} a \\ b \end{bmatrix}$ – vector of unknown values,
l – free member vector.

Further, it would be possible to solve the system of residual equations using, for example, a pseudoinverse matrix:

$$z = A^+ \cdot l, \qquad (1)$$

where $A^+ = (A^T \cdot A)^{-1} \cdot A^T$ – generalized inverse of matrix A.

However, first we calculate the condition number of the matrix A. This number is equal to

$$\kappa(A) = \|A^+\| \cdot \|A\| = 3.6 \cdot 10^6.$$

Here $\kappa(A)$ is the condition number of the matrix A,
$\|A^+\|$ – the norm of a pseudoinverse matrix A,
$\|A\|$ – the norm of matrix A.

The condition number $\kappa(A)$ characterizes the problem (1). It shows how many times the output value $\frac{\|\Delta z\|}{\|z\|}$ can change with a small change in the input data $\frac{\|\Delta l\|}{\|l\|}$. The large value of the condition number indicates the poor stability of the matrix A^+. It is clear that we cannot consider the result of solving the problem $z = A^+ \cdot l$ reliable if the condition number $\kappa(A)$ is 3.6 million. Therefore, we will try to reduce the condition number of the residual equations matrix.

For this, it is possible to make the elements of the residual equations matrix A, if possible, equal the same order of magnitude. For example, you can drop the total value of the first column of 2010. And in subsequent calculations, we can take into account this total value. This will significantly reduce the condition number $\kappa(A)$.

However, let's go in a more convenient way. First let's find the average value of the elements from the first column. It turns out to be equal to 2017.5. Then we subtract this average from all elements from the first column. These new received abscissas are called centered abscissas. Geometrically, we moved the Y axis to a point $(0,\ 2017.5)$, see Fig. 3. Our line intersects with the new Y' axis at point $(b',\ 2017.5)$.

Then the residual equations will have the following form:

$$\begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \end{bmatrix} = \begin{bmatrix} -1.5 & 1 \\ -0.5 & 1 \\ 0.5 & 1 \\ 1.5 & 1 \end{bmatrix} \cdot \begin{bmatrix} a \\ b' \end{bmatrix} - \begin{bmatrix} 21.73 \\ 21.57 \\ 22.11 \\ 23.49 \end{bmatrix}$$

In this case, the condition number of the matrix A will be equal to a small value:

$$\kappa(A) = 1.118$$

Bear in mind that the "best" conditional number is one. In this case, with a small change in the input data, the output values change by about the same amount.

We make normal equations $A^T \cdot A \cdot z = A^T \cdot l$:

$$\begin{bmatrix} 5 & 0 \\ 0 & 4 \end{bmatrix} \cdot \begin{bmatrix} a \\ b' \end{bmatrix} = \begin{bmatrix} 2.91 \\ 88.9 \end{bmatrix}$$

As you can see, the system of normal equations was transformed into two independent equations. This transformation is a consequence of our introduction of centered abscissas. From the first equation we find the tangent of the slope of the model a:

$$a = \frac{2.91}{5} = 0.582 \text{ million per year}$$

Model parameter a characterizes the growth rate of visitors to the country per year. In fact, we calculated a with the following formula:

$$a = \frac{\sum x \cdot Y}{\sum x^2}. \qquad (2)$$

Here x are the centered coordinates obtained by subtracting the average value from the vector of the original coordinates X,

Y are source coordinates from the Table 1.

In formula (2), we only centered on the abscissas to find a. Note that some authors also center the vector Y in a similar formula. In this case, they get the same value, that is:

$$\frac{\sum x \cdot Y}{\sum x^2} = \frac{\sum x \cdot y}{\sum x^2}$$

Therefore, it is not necessary to center the vector Y as well.

To find the second parameter of line b, we use the fact that the line constructed by the least squares method passes exactly through the midpoint $(\bar{X}, \bar{Y} = b')$, see Fig. 3. The equation of the line passing through the midpoint $(\bar{X}, \bar{Y})$ has the following form:

$$\bar{Y} = a \cdot \bar{X} + b,$$

where:

$$\bar{X} = \frac{\sum X}{4} = 2017.5,$$

$$\bar{Y} = \frac{\sum Y}{4} = 22.225.$$

From here we find the second unknown parameter of the line:

$$b = \bar{Y} - a \cdot \bar{X} = -1151.96$$

Now we check the correctness of the calculation of the parameters of the line a and b. This control follows from our following reasoning. When solving the least squares problem, we look for the minimum of the function:

$$S = \sum r^2 = r^T \cdot r = r^T \cdot (A \cdot z - l) = r^T \cdot A \cdot z - r^T \cdot l$$

But

$$r^T \cdot A \cdot z = z^T \cdot A^T r = z^T \cdot A^T (A \cdot z - l) = z^T \cdot (A^T A \cdot z - A^T \cdot l) = z^T \cdot \begin{bmatrix} 0 \\ 0 \end{bmatrix} = 0.$$

Therefore, the equality $r^T \cdot r = -r^T \cdot l$ can serve as a control of the correctness of the solution of the least squares problem.

For our problem, we find both of these sums:

$$\sum r^2 = 0.59388$$

$$\sum r \cdot l = -\,0.59388$$

As you can see, both sums coincide in absolute value and opposite in sign. This indicates the correctness of the solution.

Now we can calculate the forecast value for 2020:

$$Y = a \cdot X + b = 0.582 \cdot 2020 - 1151.96 = 23.68 \text{ million.}$$

Thus, according to our forecast, the number of tourists visiting Ukraine in 2020 will be 23.68 million people. Perhaps our forecast would have been correct, providing the coronavirus COVID-19, being a force majeure, had not occurred.

CONCLUSIONS. In the quantitative method of an economic indicator forecasting, an effective approach is to initially assess the state of the time series. The time series in question may have an uptrend, downtrend or sideways trend, i.e. flat. Next, if we have an uptrend or downtrend, then we build a linear trend model. In this case, using the least squares method, one should calculate the condition number of the residual equations matrix. If the condition number is too large (for example, millions, as in the above-mentioned example), then the residual equations matrix should be simplified, for example, by entering centered coordinates. When calculating the tangent of the slope of the model, it is advisable to center only the abscissas and to use the original ordinates. The solution should be checked with the formula: $r^T \cdot r = -r^T \cdot l$.

REFERENCES

1.Tobias Prokescha, Heiko A.von der Gracht, Holger Wohlenberg. Integrating prediction market and Delphi methodology into a foresight support system — Insights from an online game // Technological Forecasting and Social Change, Volume 97, August 2015, Pages 47-64.
https://doi.org/10.1016/j.techfore.2014.02.021

2. FORECASTING: PRINCIPLES AND PRACTICE By Rob J Hyndman and George // Athanasopoulos 2nd edition, May 2018. https://otexts.com/

3. Ramon Lawrence. Using Neural Networks to Forecast Stock Market Prices // Department of Computer Science, University of Manitoba, December 12, 1997. https://people.ok.ubc.ca/rlawrenc/research/Papers/nn.pdf

4.Ghalehkhondabi, I., Ardjmand, E., Young, W. and Weckman, G. (2019). A review of demand forecasting models and methodological developments within tourism and passenger transportation industry // Journal of Tourism Futures, Vol. 5 No. 1, pp. 75-93. https://doi.org/10.1108/JTF-10-2018-0061