Predictive modeling of indicators of the tourism sector of the world economy

A.M. Ignatenko^{1,*}, E.V. Onishchenko², and R.A. Simonyan³

¹Sochi State University, 94, Plastunskaya St., Sochi, 354000, Russia

²Federal Research Centre the Subtropical Scientific Centre of the Russian Academy of Sciences, 2/28 Jan Fabricius St., Sochi, Russia

³Kuban State University, 149, Stavropol St., Krasnodar, 350040, Russia

Abstract. The article studies the statistical indicators of the tourism sector of the world economy. For the study, predictive modeling technology is used, which includes methods of correlation and regression data analysis. A modern R programming language is used as a modeling tool, which has wide functionality. The authors analyze the main indicators of tourism, as well as predictors of scientific, technical and innovative development. As a result of the analysis, a close relationship between indicators is proved, as well as a linear relationship is identified.

1 Introduction

Tourism is one of the main sectors of the world economy, which contributes to the economic growth of countries, an increase in the number of additional jobs, the formation of a positive image. This industry also contributes a significant share of cash to the formation of countries' GDP.

In order to understand how effectively a particular industry is functioning, statistical information about its state and the use of modern methods of data analysis are needed, which will help to identify trends in the development of the industry, determine dependencies and determine the predicted values of indicators. One of these methods is predictive modeling.

To ensure comparability of international indicators for the development of the tourism industry, UNWTO has developed International Recommendations for Tourism Statistics. These recommendations distinguish 2 main groups of statistical data for accounting for inbound tourism: accounting for inbound flows and accounting for expenses.

2 Materials and methods

To analyze the data, we will use the following indicators of international tourism for 2018, designating these predictors with the corresponding variables for the convenience of calculations [1, 2]:

The number of arrivals - the number of tourists who visited the country within a period not exceeding 12 months (thousand people) – A1.

* Corresponding author: allrededor@mail.ru

Number of departures - the number of people who left the country of permanent residence for any other country for any reason (thousand people) – A2.

International Tourism Expenses - Visitor spending in other countries, including payments to foreign carriers for international transport (million US dollars) – A3.

Travel expenses for passenger transport - travel expenses for passenger transport are the expenses of tourists traveling abroad from all over the world in other countries, for services provided by non-resident carriers in the process of international transportation (million US dollars) – A4.

Travel and shopping expenses - goods and services purchased by or on behalf of a tourist for his personal use or use as gifts, goods (million US dollars) – A5.

International tourism revenues are the expenses of foreign inbound visitors, including payments by national carriers for international transport. This income includes all prepayments for goods and services purchased in the destination country. They can also include income from tourists visiting the country for one day, unless they deserve a separate classification. (million US dollars) – A6.

Total contribution to GDP - the share of travel or employment spending across the economy in published national income accounts or in labor market statistics (million US dollars) - A7.

Direct contribution to employment is employment in the hospitality industry, travel agencies, transportation, as well as catering and leisure when providing services directly to tourists (million US dollars) – A8.

Domestic international tourist spending - domestic international tourist spending (million US dollars) – A9.

Residents' expenses within the country (million US dollars) – A10.

Government spending - the share of travel or employment spending across the economy in published national income accounts or in labor market statistics (million US dollars) – A11.

Business travel expenses (million US dollars) – A12.

For modeling, 20 countries were taken, which are leaders in the "Number of arrivals" indicator, presented in Figure 1.

Fig. 1. Number of arrivals - the number of tourists visiting the country within a period not exceeding 12 months.

In addition, to determine the presence of a correlation and relationship between indicators of tourism and scientific and technological development, we add the following predictors, designating them with the corresponding variables [4, 5]:

1. ICT publications in Scopus-indexed publications (pcs.) - B1.
2. R&D personnel (thousand people) – B2.
3. Number of researchers (thousand people) – B3.
4. Internal R&D costs (million US dollars) – B4.
5. State budget allocations for research and development by country (million US dollars) – B5.
6. Articles in scientific and technical journals (pcs.) – B6.

3 Correlation modeling

Correlation analysis is the analysis of the relationship of quantitative data. Indicators for this relationship are in the range from -1 to 1. The closer this indicator is to 1, the stronger the relationship between the predictors. The simulation was carried out using the programming language R. In R, there is a huge number of correlation tests that allow you to calculate both partial correlation, and polychoric coefficients, and multi-row. For the analysis, the function "corr.test ()" was used, which allows you to calculate 3 correlation coefficients: Spearman's rank correlation coefficient, which shows the relationship between the ranked variables, Kendall's correlation coefficient, a nonparametric rank correlation indicator, and Pearson's correlation coefficient, showing the degree of linear relationship between quantitative variables. We will use the latter to determine the presence of a relationship between the predictors of the world economy tourism sector under study. Function syntax: corr.test (x, use =, method =), where "x" is the data table; "use" is a parameter that allows you to initiate gaps in the data and makes it possible to either ignore them, or give the user an error; "method" - specifies the type of coefficient used (Pearson, Spearman or Kendall). Figure 2 shows the result of using the "corr.test ()" function between variables A1-A12 and B1, as well as the result of checking the statistical significance of the calculated correlation coefficients.

```
> library(psych)
> corr.test(Ran,use="complete")
Call:corr.test(x = Ran, use = "complete")
Correlation matrix
      A1  A2  A3  A5  A6  A7  A8  A9  A10 A11 A12 B1
A1  1.00 0.27 0.46 0.42 0.64 0.55 0.32 0.74 0.52 0.44 0.50 0.47
A2  0.27 1.00 0.85 0.87 0.34 0.72 0.71 0.58 0.77 0.71 0.64 0.75
A3  0.46 0.85 1.00 0.99 0.48 0.93 0.88 0.77 0.94 0.79 0.82 0.96
A5  0.42 0.87 0.99 1.00 0.38 0.90 0.92 0.71 0.90 0.73 0.76 0.95
A6  0.64 0.34 0.48 0.38 1.00 0.70 0.13 0.88 0.70 0.74 0.82 0.55
A7  0.55 0.72 0.93 0.90 0.70 1.00 0.77 0.91 0.99 0.84 0.95 0.98
A8  0.32 0.71 0.88 0.92 0.13 0.77 1.00 0.56 0.74 0.50 0.55 0.86
A9  0.74 0.58 0.77 0.71 0.88 0.91 0.56 1.00 0.88 0.80 0.89 0.82
A10 0.52 0.77 0.94 0.90 0.70 0.99 0.74 0.88 1.00 0.90 0.96 0.96
A11 0.44 0.71 0.79 0.73 0.74 0.84 0.50 0.80 0.90 1.00 0.88 0.79
A12 0.50 0.64 0.82 0.76 0.82 0.95 0.55 0.89 0.96 0.88 1.00 0.89
B1  0.47 0.75 0.96 0.95 0.55 0.98 0.86 0.82 0.96 0.79 0.89 1.00
Sample Size
[1] 18
Probability values (Entries above the diagonal are adjusted for multiple tests.)
      A1  A2  A3  A5  A6  A7  A8  A9  A10 A11 A12 B1
A1  0.00 0.66 0.45 0.49 0.08 0.27 0.66 0.01 0.33 0.46 0.41 0.44
A2  0.28 0.00 0.00 0.00 0.66 0.02 0.02 0.20 0.01 0.02 0.08 0.01
A3  0.06 0.00 0.00 0.00 0.44 0.00 0.00 0.01 0.00 0.00 0.00 0.00
A5  0.08 0.00 0.00 0.00 0.60 0.00 0.00 0.02 0.00 0.01 0.01 0.00
A6  0.00 0.17 0.04 0.12 0.00 0.03 0.66 0.00 0.03 0.01 0.00 0.27
A7  0.02 0.00 0.00 0.00 0.00 0.00 0.01 0.00 0.00 0.00 0.00 0.00
A8  0.20 0.00 0.00 0.00 0.62 0.00 0.00 0.26 0.01 0.41 0.27 0.00
A9  0.00 0.01 0.00 0.00 0.00 0.00 0.02 0.00 0.00 0.00 0.00 0.00
A10 0.03 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
A11 0.07 0.00 0.00 0.00 0.00 0.00 0.04 0.00 0.00 0.00 0.00 0.00
A12 0.03 0.00 0.00 0.00 0.00 0.00 0.02 0.00 0.00 0.00 0.00 0.00
B1  0.05 0.00 0.00 0.00 0.02 0.00 0.00 0.00 0.00 0.00 0.00 0.00
```

Fig. 2. Correlation coefficient and their statistical significance between predictors A1-A12 and B1.

Let us note the most highly correlated values between the predictors. We note right away that the relationship between all the variables is direct. In a very close relationship (the correlation coefficient exceeds 0.9 and is statistically significant) there are tourism indicators such as «Expenditure on international tourism» and «Expenditure on travel and shopping» (0.99), «Expenditure on international tourism» and «Total contribution to GDP» (0.93), «Expenditure on international tourism» and «Expenditure by residents within the country» (0.94), «Expenditure on travel and shopping» and «Total contribution to GDP» (0.9), «Travel and Shopping Expenses» and «Domestic Residents Expenses» (0.9), «Travel and Shopping Expenses» and «Direct Contribution to Employment» (0.92), «Business Travel Expenses» and «Total contribution to GDP» (0.95).

In addition, we can highlight the relationship between the predictors with a correlation coefficient higher than 0.8: «Expenditure on international tourism» and «Travel expenses on passenger transport» (0.85), «Number of departures» and «Expenses on travel and shopping» (0.87), «International Tourism Expenditure» and «Direct Contribution to Employment» (0.88), «International Tourism Income» and «Domestic International Tourist Expenditure» (0.88), «Government Expenditure» and «Total Contribution to GDP» (0.84), «Business Travel Expenditure» and «Domestic International Tourist Expenditure» (0.89).

In a separate group, it is worth highlighting a direct and very close relationship between predictors and indicators of scientific and technological progress, the correlation coefficients of which are presented in Table 1.

Table 1. The relationship of key predictors of tourism with indicators of scientific and technological progress.

	Internation al tourism expenses	Travel and shopping expenses	Total contribution to GDP	Residents' expenses within the country	Direct contribution to employment	Domestic travel expenses of international tourists
Articles in scientific and technical journals	0.96	0.95	0.98	0.96	0.86	0.82
ICT publications in Scopus-indexed publications	0.98	0.97	0.96	-	0.91	0.84
Research and development personnel	0.9	0.93	0.78	-	0.98	-
Number of researchers	0.92	0.89	0.95	0.92	0.83	0.82
Internal R&D costs	0.9	0.83	0.99	0.98	0.72	0.93
State budget allocations for research and development by country	0.85	0.77	0.96	0.96	0.62	0.96

It should be noted that the indicators of scientific and technological progress and innovations have very high and statistically significant correlation coefficients with almost all the main indicators of the tourism sector of the world economy.

4 Regression modeling

Regression analysis is a statistical method based on identifying the mathematical relationship between indicators. In addition to the usual methods of regression analysis, R has many functions for the selection of regression models. Regression types used in R: simple linear, polynomial, multiple, multivariate, logistic, Poisson, Cox proportional hazards, time series, nonlinear, nonparametric, robust. The calculations also take into account the functions of detecting multicollinearity, heteroscedasticity and other problems arising in regression modeling.

The lm() function was used to calculate the regression parameters. Its syntax is as follows: lm (formula, data), where formula - describes the mathematical form of the model that needs to be calculated; data - a table with data containing values for creating a model. During the analysis, linear and multiple regression models were built. The results of regression modeling (containing only those regression equations in which the coefficients were statistically significant) are presented in Table 2. It presents the main relationships between the predictors. The indicator "Total contribution to GDP" was considered as an endogenous variable.

Table 2. Results of regression modeling for key indicators of the tourism sector of the world economy.

No.	Regression equation	Multiple R-squared	F-statistic
1	«Total contribution to GDP» = -34912.078+6.188· «International tourism expenses»	0.87	109.3
2	«Total contribution to GDP» = 137445.27+54.591· «Direct contribution to employment »	0.82	72.81
3	«Total contribution to GDP» = 52774.61+72.1· «Domestic travel expenses of international tourists»	0.71	38.36
4	«Total contribution to GDP» = 9226.13+1.82· «Residents' expenses within the country»	0.9	145.3
5	«Total contribution to GDP» = -34377.38+13.195· «Internal R&D costs»	0.99	82.76
6	«Total contribution to GDP» = -123360.75+6.31· «Domestic travel expenses of international tourists»+ 28.56· «Government spending»	0.92	49.01

Thus, the following conclusions can be drawn. Among the constructed models, the highest coefficient of determination (Multiple R-squared, equal to 0.99) has a linear regression equation describing the relationship between the predictors "Total contribution to GDP" and "Internal expenditures on research and development." It can be argued that with an increase in research and development costs by an average of \$ 1 million. the total contribution to GDP will increase by \$ 13.195 million. In addition, with an average increase in international tourism spending by \$ 1 million. GDP will grow by \$ 6.188 million. The results of regression modeling show that the indicators of the tourism sector of the world economy directly affect the increase in GDP of the studied countries.

5 Cluster analysis

Cluster analysis is one of the most popular and modern methods of data analysis. This tool is also actively used in the tourism sector. The task of cluster analysis is to break down the population (objects), in our case, 20 leading countries in the field of tourism by the number of tourists visiting the country, into groups that would be similar in a certain way. Clustering algorithms for the implemented methods are divided into 2 groups: hierarchical and non-hierarchical.

With a large number of observations, hierarchical methods of cluster analysis are not suitable. In such cases, non-hierarchical methods are used. The work analyzed data for 20 countries and 11 indicators.

To analyze the data, we will use a non-hierarchical k-means squared error algorithm. The algorithm is the simplest to implement, but it assumes that the number of clusters is determined in advance. This problem can be overcome by performing a hierarchical analysis with a random sample of observations. Applying the k-means algorithm in the R programming language, we get the dendrogram shown in Figure 3. The dendrogram shows the presence of 5 clusters.

Fig. 3. Dendrogram.

The first cluster includes China and the United States, exceeding all other countries in all respects. Russia was included in the largest cluster along with Austria, Greece, Malaysia, UAE, Canada, Poland and Holland. France, Italy, Spain and Mexico form a separate group.

6 Conclusion

After the analysis, it should be noted that due to the huge data set and the dynamic development of world tourism, it will also be relevant to consider the complex patterns of the influence of indicators, identifying not only linear, but also nonlinear dependencies between them using modern means and data processing techniques. Also highlight and analyze the main indicators of scientific and technological progress, which most strongly affect the indicators of tourism development.

References

1. <https://knoema.ru/atlas/topics/> Tourism
2. <https://knoema.ru/atlas/topics/> Research and development
3. Kuhn Max, Johnson Kjell 2019 *Predictive modeling in practice* p 640
4. Abdrakhmanova G I, Vishnevsky K O, Gokhberg L M et al 2020 *Digital economy: 2020: a short statistical collection* (Moscow: HSE)
5. G. Abdrakhmanova, K. Vishnevskiy, L. Gokhberg et al 2019 *Digital Economy Indicators in the Russian Federation: 2019: Data Book* (Moscow: HSE)