SILESIAN UNIVERSITY OF TECHNOLOGY PUBLISHING HOUSE

SCIENTIFIC PAPERS OF SILESIAN UNIVERSITY OF TECHNOLOGY 2019
ORGANIZATION AND MANAGEMENT SERIES NO. 133

INFORMATION TECHNOLOGY OF TOURISM DEMAND MODELING BASED ON COGNITIVE AND STATISTICAL ANALYSIS

Victor M. KRILOV[1*], Hrystyna V. LIPYANINA[2]

[1] Odessa National Polytechnic University, Shevchenko Avenue 1, Odessa 65044, Ukraine;
viktor.kryilov@gmail.com, ORCID: 0000-0003-1950-4690
[2] Ternopil National Economic University, Lvivska Str. 11, Ternopil 46000, Ukraine; xrustya.com@gmail.com,
ORCID: 0000-0002-2441-6292
* Correspondence author

Abstract: The process of tourism demand formation is studied, in which the infrastructure level is estimated on the basis of subjective expertise. There is a strong correlation between the number of collective accommodation facilities and tourism activity subjects, the infrastructure level based on the subjective expertise and the number of recreations. Regression dependencies between the tourism demand and closely correlated factors are determined.

Keywords: information technology, model, tourism demand, correlation analysis, regression dependence.

1. Introduction

The tourism market in Ukraine is at the stage of development, which is determined by socio-economic and political processes of the country.

In 2017, Ukraine was visited by 142 million of foreign tourists. Compared to 2013, foreign tourist arrivals declined by 56%, the number of domestic tourists increased by 0,5%, and the number of excursionists during 2000-2017 also decreased by 29% (Anderson, 1976, p. 756).

The development of the tourism services market and its infrastructure is a topical issue of current economic studies. As infrastructure facilities are crucial for the regional tourism economy, this necessitates the search for ways to develop the infrastructure of the tourism services market, increasing the efficiency of management decision-making process and implementing its development strategy in accordance with the state and trends of the subregional tourism market.

http://dx.doi.org/10.29119/1641-3466.2019.133.7

2. Model characteristics

The demand for tourism is determined by the number of tourists visiting a country or the expenditures of the country. Macroeconomic indicators, such as the income level in different countries, tourism expenditures in Ukraine, transport costs, and the number of collective accommodation facilities are taken into account in the model of tourism demand.

The flow of tourists to Ukraine can be characterized according to the following factors:

$$\mathrm{Y} = F\ (S, V_{TR}, P, R, K, C, I, T),$$

where:
Y – flow of tourists,
S – average salary per person in the tourism industry,
$V_{(TR)}$ – tourism expenditures,
R – the number of collective accommodation facilities,
P – the number of subjects of tourism activity,
K – the number of recreations,
C – the amount of goods and services produced according to the type of economic activity,
I – capital investment by region,
T – transport connection,
N – infrastructure (subjective indicator).

3. The purpose and objectives of the research

The purpose of the work is to develop information technology for modeling tourism demand on the basis of cognitive and statistical analysis.

To achieve the purpose of the research, it is necessary to solve the following scientific tasks:

- to analyze the existing solutions in the field of information technologies and mathematics used for modeling tourism demand on the basis of cognitive and statistical analysis,
- to carry out a correlation analysis of tourism demand, whose distinctive features are qualitative and quantitative parameters,
- to determine the regression dependence of tourism demand on the level of infrastructure development.

4. Methods of modeling tourism demand on the basis of cognitive and statistical analysis

To achieve the stated goals, we need to determine the correlation and regression.

Linear correlation for empirical data, measured according to interval or ratio levels, is estimated using the Pearson correlation coefficient r_{xy}.

$$r_{\mathrm{xy}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where:

x_i and y_i – values of variables X and Y,

$\bar{x}$ and $\bar{y}$ – average values of X and Y,

n – sample volume.

The first stage involves determining the coefficients of the pair correlation (between two numerical arrays – x_i and y_i). The relationship between the values of the coefficients and the nature of the power in the existing relationship is given in Table 1.

Table 1.
Qualitative assessment of the power of relationship

Values of the pair correlation coefficient	Nature of the power in relationship
Up to 0.3	Virtually absent
0.3-0.5	Weak
0.5-0.7	Notable
0.7-0.9	Strong
0.9-0.99	Very strong

In order to consider the indicators as factors suitable for modeling, the relationship between them should be strong or very strong, that is, the resulting variable value of the model should be more than 0.7. If the correlation coefficient values are less than 0.7, the relationship between the factors of this group is considered to be insignificant and these factors should not be considered as having a significant effect on this economic phenomenon. Thus, the procedure for determining the pair correlation coefficients allows us to select those factors among the whole set which really show a close relationship between them and the indicator, selected as a dependent variable of a certain economic process, and not to take into account the factors which have little effect on this process.

It should be noted that there are often problems of the so-called multicollinearity while creating the multi-factor regression equation: when there is a close relationship not only between the factor features and the functional (dependent) feature, but also between the factor features themselves. In such a case, the functional relationship may be distorted and the simulated process is not shown properly. That is why, in addition to determining the pair correlation coefficients between the independent variables and dependent variable, it is expedient to determine the pair correlation coefficients between the factor features themselves and remove those factors which can show the presence of multicollinearity and a less close relationship between them and the functional variable. The second stage in building the regression models is selecting the formula of the regression equation itself. A linear multifactor regression which describes a linear relationship between the data under investigation is the simplest one:

$$y = a_0 + a_1x_1 + \cdots + a_nx_n$$

where:

y – dependent variable, function,

$a_0, a_1, \dots, a_n$– regression coefficients,

$\mathrm{x}_1, \dots, \mathrm{x}_n$ – dependant variables.

Empirical formulas can be varied because when choosing an analytic dependence one should not follow some strict theories (physical or economic). Only one condition is necessary, that is, close possible correspondence of the values, calculated by the formula, to the experimental data. The corresponding regressions can be constructed "manually", but such calculations are rather cumbersome and time-consuming.

While building multi-factor regression models, the following stages can be distinguished:

1. Selecting and analyzing all possible factors that affect the process or indicator which is being studied.
2. Measuring and analyzing the factors found. If some factors cannot be quantitatively or qualitatively determined or the model parameters statistics are not available for them, then they are removed from further consideration.
3. Mathematical and statistical analysis of the factors. At this stage, when there is a lack of information in dynamic rows, the information may be restored by means of specified methods and the basic assumptions of the classical regression analysis are verified.
4. Selecting the regression multifactor model.
5. Estimating unknown parameters of the regression model.
6. Verifying the significance of the found parameters of the model and its correspondence to the reality by means of the Fisher's F-criterion and Student's t-criterion. Fisher's F-statistics is calculated with m and (n-m-1) degrees of freedom:

$$F = \frac{\frac{\sum_{i=1}^{n}(y_{ip}-\bar{y})^2}{m}}{\frac{\sum_{i=1}^{n}(y_i-y_{ip})^2}{n-m-1}}$$

where:

m – the number of factors included into the model,

n – the total number of observations,

y_{ip} – the estimated value of the dependent variable with the i-th observation,

$\bar{y}$ – the average value of the dependent variable,

y_i – the value of the dependent variable with the i-th observation.

According to Fischer's F-tables, the critical value of F_{kr} with m and (n-m-1) degrees of freedom is determined due to the previously set confidence level (1- α) · 100%. F > F_{kr}, shows the adequacy of the model constructed.

If the model is inadequate, it is necessary to return to the stage of its construction and additional factors can be introduced, or the nonlinear model is used. T-statistics for multi-factor regression parameters are as follows:

$$t = \frac{a_i}{\sigma_{a_i}^2}$$

where:
a_i – estimation of the i-th parameter,
$\sigma_{a_i}^2$ – mean square deviation of the i-th parameter estimation.

If the t-value exceeds the critical value, which is not included into the t-criterion table, then the corresponding parameter is considered to be statistically significant and it greatly influences the aggregate indicator.

7. Determining the main characteristics (multiple correlation coefficient), analyzing the results received, drawing conclusions.

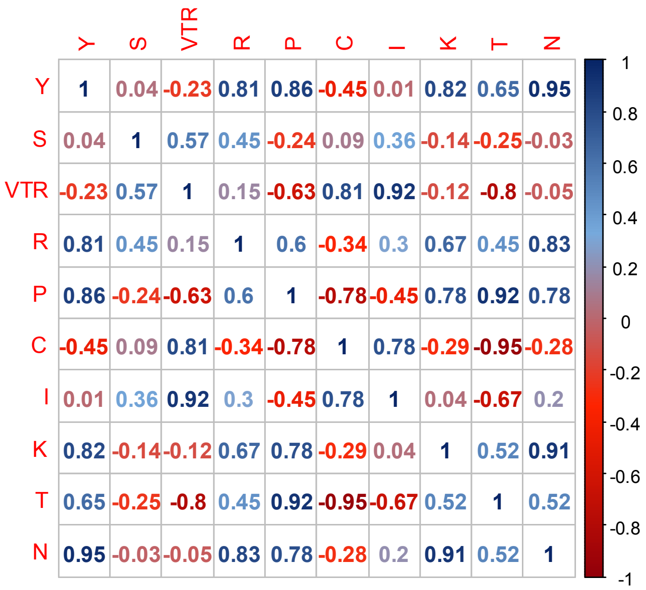
Multiple correlation coefficient is the main indicator of the correlation density of the aggregate indicator and factors. If its value, calculated by the formula, is close to 1, the relationship between the indicator and the factors is considered to be dense.

$$R = \sqrt{1 - \frac{\sum_{i=1}^{n}(y_i - y_{ip})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

5. Results of the study of tourism demand modeling based on cognitive and statistical analysis

To analyze the factors affecting the tourism demand, data of the Ukrainian Statistics Service is used (Dubrova, 2003, p. 205). The infrastructure indicator is an average assessment made by the experts in the field of infrastructure, that is, this indicator is qualitative.

There are several methods for determining the correlation coefficient level. The method of least squares is the most well-known one. However, instead of this rather laborious calculation, functional and statistical dependencies in the programming language R can be successfully used. This allows us to find the correlation coefficients within a few minutes.



	Y	S	VTR	R	P	C	I	K	T	N
Y	1	0.04	-0.23	0.81	0.86	-0.45	0.01	0.82	0.65	0.95
S	0.04	1	0.57	0.45	-0.24	0.09	0.36	-0.14	-0.25	-0.03
VTR	-0.23	0.57	1	0.15	-0.63	0.81	0.92	-0.12	-0.8	-0.05
R	0.81	0.45	0.15	1	0.6	-0.34	0.3	0.67	0.45	0.83
P	0.86	-0.24	-0.63	0.6	1	-0.78	-0.45	0.78	0.92	0.78
C	-0.45	0.09	0.81	-0.34	-0.78	1	0.78	-0.29	-0.95	-0.28
I	0.01	0.36	0.92	0.3	-0.45	0.78	1	0.04	-0.67	0.2
K	0.82	-0.14	-0.12	0.67	0.78	-0.29	0.04	1	0.52	0.91
T	0.65	-0.25	-0.8	0.45	0.92	-0.95	-0.67	0.52	1	0.52
N	0.95	-0.03	-0.05	0.83	0.78	-0.28	0.2	0.91	0.52	1

Figure 1. Matrix of correlations.

Table 2.
Estimation of the relationship density between the dependent variable Y and independent variables

Effective indicator (dependent variable y)	Factor (independent variable x)	Value of the pair correlation coefficient (r)	Nature of the linear relationship (density)	Characteristics of the relationship
Y – tourist flow	*R* – number of collective accommodation facilities	0.8137670	strong	direct
	S – average salary per person in the tourism industry	0.04219225	virtually absent	direct
	V_{TR} – tourism expenditures	-0.2271771	virtually absent	reverse
	P – number of subjects of tourism activity	0.8613155	strong	direct
	Z – ecology	0.8080230	strong	direct
	N – level of infrastructure based on subjective expertise	0.94793356	very strong	direct
	K – number of recreations	0.8180624	strong	direct
	C – the amount of goods and services produced according to the type of economic activity	-0.44915574	weak	reverse
	I – capital investment by region	0.006036328	virtually absent	direct
	T – transport connection	0.6452539	notable	direct

According to the results of the correlation matrix, the tourist flow to Ukraine depends on the factors, which show a level of dependence of (Y) ≥0.7, that is, the nature of the relationship is strong or very strong.

Therefore, the model of the tourism demand formation looks as follows:

$$\mathrm{Y} = F\ (P, R, K, N),$$

where:

Y – tourist flow,

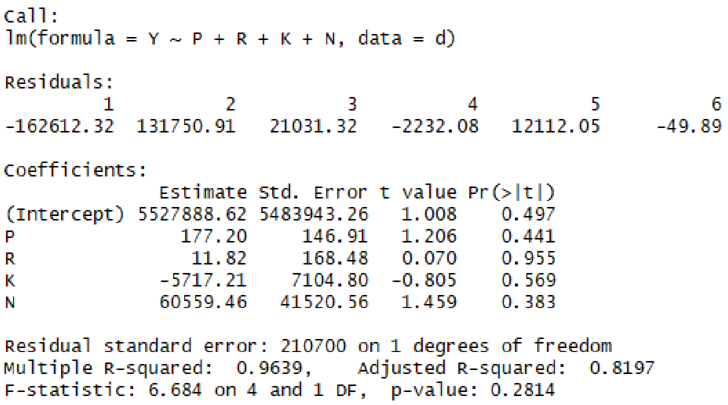
R – the number of collective accommodation facilities,

P – the number of subjects of tourist activity,

K – the number of recreations,

N – infrastructure level based on subjective expertise.

Then, the tourism demand regression model based on cognitive and statistical analysis is built (Figure 2).



```
Call:
lm(formula = Y ~ P + R + K + N, data = d)

Residuals:
         1          2          3          4          5          6
-162612.32  131750.91   21031.32   -2232.08   12112.05     -49.89

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 5527888.62 5483943.26   1.008    0.497
P               177.20     146.91   1.206    0.441
R                11.82     168.48   0.070    0.955
K             -5717.21    7104.80  -0.805    0.569
N             60559.46   41520.56   1.459    0.383

Residual standard error: 210700 on 1 degrees of freedom
Multiple R-squared:  0.9639,    Adjusted R-squared:  0.8197
F-statistic: 6.684 on 4 and 1 DF,  p-value: 0.2814
```

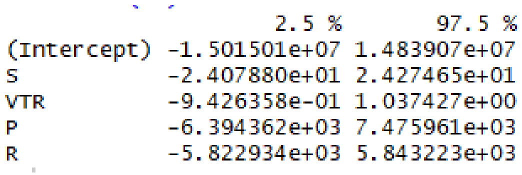
Figure 2. Results of the tourism demand regression model based on the statistical data for the years 2012-2017.

According to the statistical data for 2012-2017, the determination coefficient $R^2 = 0.99639$ and the value of the F observation statistics = 6.684. These values show the model adequacy, since the determination coefficient is close to 1, its observed F-statistics value is 6.684, which is more than the critical F-statistics value at the level of significance of 4.5.

The availability of these regularities of the tourism sector in Ukraine is confirmed by correlation and regression analysis on the basis of statistical data for 2012-2017. By means of this analysis, the dependence of the size of tourist flow (Y) on the change in the number of collective accommodation facilities (R), the number of subjects of tourism activity (P), the number of recreations (K), and infrastructure level based on the subjective expertise (N) are determined and can be introduced as linear multiple regression:

$$Y_t = 177.20P_t + 11.82R_t - 5717.21K_t + 60559.46N_t + 5527888.62 \tag{1}$$

Thus, the received multiple regression shows the dependence of the volume of tourism demand on the main indicators of tourism activity. Therefore, based on the predictive values of the number of collective accommodation facilities, the number of subjects of tourism activity, the number of recreations and the infrastructure level resulting from the subjective expertise, it is possible to determine the predictive values of tourist flows.



```
                    2.5 %       97.5 %
(Intercept) -1.501501e+07 1.483907e+07
S           -2.407880e+01 2.427465e+01
VTR         -9.426358e-01 1.037427e+00
P           -6.394362e+03 7.475961e+03
R           -5.822934e+03 5.843223e+03
```

Figure 3. Confidence intervals for model indicators.

Using the obtained values of confidence intervals for the parameters of the empirical model (Fig. 3), it is possible to write the functions of the upper and lower limits of the confidence interval, within which Y*dependent variable values can be determined with a given reliability:

Lower limit

$$Y_t = -1689.5 \cdot P_t - 2128.9 \cdot R_t - 95992.3 \cdot K_t - 467009.3 \cdot N_t - 64152217.2$$

Upper limit

$$Y_t = 2043.9 \cdot P_t + 2152.5 \cdot R_t + 84557.9 \cdot K_t + 588128.2 \cdot N_t + 75207994.4$$

One of the tasks of regression analysis is to predict the future value of a dependent variable on the basis of the obtained multiple regression model.

6. Conclusions

The process of tourism demand formation is studied in this paper. Its infrastructure level is assessed on the basis of subjective expertise. A correlation analysis is carried out and a strong correlation between the number of collective accommodation facilities and the number of subjects of tourism activity, the infrastructure level based on subjective expertise and the number of recreations is found. According to the results received, the tourism demand model is built on the basis of cognitive and statistical analysis.

References

1. Anderson, T. (1976). *Statistical analysis of time series*. Moscow: Mir, 756.
2. Dubrova, T.A. (2003). Statistical methods of forecasting. *Finance and Statistics,* 205.
3. Holovko, O.M. (2013). Promising directions of tourism development in small cities. *Scientific Bulletin of NLTU of Ukraine, 23.14*, 67-73.
4. Hunk, D.E. (2003). *Business Forecasting*. Moscow: Publishing house “Wiliams”, 656.
5. Official statistical information of the Ukraine State Statistics Committee, http://www.ukrstat.gov.ua.
6. The Law of Ukraine "On Tourism" (18.11.2003). №1282 – IV, http://zakon4.rada.gov.ua/ laws/show /324/95-%D0%B2%D1%80.
7. The World Tourism Organization (UNWTO), http://www2. nwto.org.
8. Tsibulskiy, V.O. (2015). Investigation of the essence of demand and supply in the market of tourist services and factors of influence on them. *Economy. Management. Innovations, 2(14)*, 14-24.