Information systems of passenger transportation forecast in Transbaikal region

L Y Kriklevskaya, O K Kuklina, E A Mihailova and A S Yahina

Chita Institute of Baikal State University, 56, Anokhin str., Chita, 672000, Russia

E-mail: lara.kriklevskaya@mail.ru

Abstract. The paper is concerned with the qualitative assessment of socio-economic development of transport industry of the region made with a help of information systems and mathematical methods. The object of this study is a railway transportation system in Transbaikal region. The research was made on the basis of a priori and posteriori approach with the use of recessionary analysis and correlation. The research of basic trends in socio-economic development of Transbaikal region during a long period of time (10 years) comprising the periods of growth and stagnation made it possible to reveal a set of models which affects the volume of transportation. This will help to forecast a strategy of development of Transbaikal region. The article contains the results of empirical study on interconnection between a set of factors and a volume of passenger transportation in Transbaikal region.

1. Introduction

An automated system of operative management of the transportation process (ASOUP) is used on each road of the public limited company “Russian Railway”. Its creation and adoption became a basis of computer network on the rail roads of the country.

Information about all operations made with the models of transportation process is transmitted directly from the place (a station, a locomotive depot, a wagon depot, a loading area, etc.). This information can be generated in the corresponding AWP (automated working place) of the workplace. Further, the information should be transferred to the ASOUP complex by means of the DTN (data transportation network).

The users can obtain information from the ASOUP-2 complex, the databases of which are stored in the DB2 database system that supports SQL queries, through WEB applications developed by the Research and Design Institute of “Russian Railways”, as well as by the ITC staff.

To develop projects in the Chita Information and Computing Center we adopted the technology of JSP-pages for displaying data and the technology of “data loaders” for downloading information into their databases or application tables. The function of the “loaders” is as follows: at a certain point of time, the Main Frame starts a program written by ITC staff, which selects the information of interest and loads it into the appropriate tables, and also clears outdated information after the expiration date. The ‘loaders’ writing languages are Assembler and C.

In this work, we used real data obtained by a request from ASOUP. Then to the data obtained with the Visual Basic programming language, we applied operations previously recorded with macros. The operations were the utilities of MS Excel “Analysis Package”: “Data Analysis” and “Regression”.

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd
Taking into account the experience of state regulation in different fields of economy and in different types of economic activity it is possible to suppose that the strategy of social and economic perspective development should be based on forecasting and modeling method [1, 2]. The creation of some alternative ways of development based on the analysis of retrospective data, as well as the study of endogenous (with weak state intervention) and exogenous (with strong state regulation) within and inter-sectoral relations, allows one to create scenarios of future development.

2. Results
To make a multifactor forecasting of the total passenger transportation in the Transbaikal region a list of the most significant factors was formed.

A dynamic analysis of the change in the total identified factors for the period of 2000-2016 also was done.

Statistical data provided by the Federal State Statistics Service of Transbaikal region were consolidated into one table (table 1); Pr – time period (years) monitoring (1 – 2000, 2 – 2001, ..., 17 – 2016).

Table 1. Indicators of multifactorial forecasting of the total passenger traffic in the Transbaikal region for 2000 – 2016

| Pr | I_1 | I_2 | I_3 | I_4 | I_5 | I_6 | I_7 | I_8 |
|----|-----|-----|-----|-----|-----|-----|-----|-----|
| 1  | 104 | 457.4 | 74.8 | 1185.8 | 67 | 1405.7 | 173747 | 15179.9 |
| 2  | 108.3 | 423.2 | 92.6 | 1178.8 | 55.3 | 2011.5 | 172515 | 15512.9 |
| 3  | 113.3 | 503.69 | 81.4 | 1165.4 | 44.6 | 2993.1 | 190116 | 15475.9 |
| 4  | 105.2 | 440.3 | 101.8 | 1152.6 | 32.3 | 4017.7 | 192580 | 15174.9 |
| 5  | 97.6 | 453.2 | 147.1 | 1142.7 | 28.8 | 4808.5 | 205126 | 15368.9 |
| 6  | 90.5 | 516.8 | 131.1 | 1132.8 | 26.2 | 5907.6 | 209874 | 15355.9 |
| 7  | 91.7 | 505.9 | 179 | 1123.3 | 23.8 | 7119.2 | 220909 | 29560.9 |
| 8  | 87 | 489.8 | 232.7 | 1115.1 | 23.9 | 8265.4 | 232363 | 27051.7 |
| 9  | 91.1 | 465.8 | 258.6 | 1111.4 | 20.1 | 11047.2 | 264982 | 26935.9 |
| 10 | 87.8 | 479.6 | 269.1 | 1109.2 | 19.9 | 12704.8 | 283759 | 27080.7 |
| 11 | 83.2 | 470.5 | 274.8 | 1109 | 19.3 | 14204.5 | 289329 | 26786.7 |
| 12 | 79.8 | 478 | 276.9 | 1106.2 | 18.9 | 15968.8 | 302968 | 26414.4 |
| 13 | 70.5 | 475.6 | 302.9 | 1099.4 | 17.6 | 17545.6 | 316405 | 26996.4 |
| 14 | 71.6 | 479.4 | 295.1 | 1095.2 | 16.2 | 19886.1 | 344076 | 26750.4 |
| 15 | 78 | 481.5 | 352.5 | 1090.3 | 18 | 20520 | 355349 | 26762.4 |
| 16 | 71.1 | 480 | 235 | 1087.5 | 20.4 | 22980.2 | 359860 | 26444.4 |
| 17 | 92.9 | 477.5 | 291.7 | 1083 | 21.4 | 22846.4 | 368297 | 26519.4 |

where:

- I_1 – the total passenger traffic, mln. people;
- I_2 – the employed population, thousand people;
- I_3 – the building commissioning, thousand sq.m.;
- I_4 – the population size, thou people;
- I_5 – the number of the population with the incomes below the subsistence level, thou people;
- I_6 – the population’s average cash income, ruble;
- I_7 – the number of transport involved in the transportation process, el.;
- I_8 – the extent of railway line, km.

Dynamic analysis of the indicators states the presence of ambiguous trends in their change. For example, the total of passenger traffic for the period from 2000 to 2016 decreased by 12 million people
per year. At the same time, positive trends in this indicator were observed until 2003, from 2004 to 2016 there was a general trend of a decrease in the total value of passenger transportation, however, within the analyzed period, the decline in the level of this indicator is replaced by a slight growth.

The number of vehicles involved in the transportation process during the research period has significantly increased. This trend is possible to explain by sharp reduction in the capacity of transport, i.e. a passenger transportsations are carried out by small transport of Gazel type now.

Dynamic analysis of the selected factors for the research revealed ambiguous trends of change over the analyzed period, what in turn can have a negative impact on the formation of the model [3].

Then a matrix of sample values of the pair correlation coefficients was constructed (table 2).

### Table 2. The values of paired correlation coefficients

|   | I₁ | I₂ | I₃ | I₄ | I₅ | I₆ | I₇ | I₈ |
|---|----|----|----|----|----|----|----|----|
| I₁ | 1.00 | | | | | | | |
| I₂ | -0.30 | 1.00 | | | | | | |
| I₃ | -0.84 | 0.22 | 1.00 | | | | | |
| I₄ | 0.85 | -0.38 | -0.93 | 1.00 | | | | |
| I₅ | 0.77 | -0.41 | -0.82 | 0.91 | 1.00 | | | |
| I₆ | -0.83 | 0.22 | 0.89 | -0.93 | -0.73 | 1.00 | | |
| I₇ | -0.82 | 0.22 | 0.90 | -0.93 | -0.74 | 1.00 | 1.00 | |
| I₈ | -0.75 | 0.34 | 0.87 | -0.85 | -0.75 | 0.75 | 0.75 | 1.00 |

The factor I₂ from the whole set of factors (table 2), based on the analysis of the matrix of sample values of the pair correlation coefficients (table 3), is excluded as it slightly influences the factor I₁ (in table 4 the correlation coefficient is in bold).

Let us estimate the parameters of this model using the method of least squares. The values of the estimates of the model coefficients, their standard errors and the actual values of t-criteria are calculated and presented in table 3.

### Table 3. Values of estimates of the coefficients of the model

| n=17 | Parameter values | Standard parameter errors | Actual values of the t-criteria | P-value |
|------|------------------|---------------------------|---------------------------------|---------|
| b₀   | 199.45102        | 532.61475                 | 0.37448                         | 0.71587 |
| I₃   | -0.11679         | 0.08033                   | -1.45378                        | 0.17666 |
| I₄   | -0.19870         | 0.46193                   | -0.43015                        | 0.67621 |
| I₅   | 0.30169          | 0.50777                   | 0.59415                         | 0.56561 |
| I₆   | -0.00843         | 0.00433                   | -1.94627                        | 0.08024 |
| I₇   | 0.00086          | 0.00049                   | 1.73030                         | 0.11426 |
| I₈   | 0.00001          | 0.00077                   | **0.01000**                     | 0.99222 |

All the factors in model, except I₆, are statistically insignificant (at a level of significance 5%). However, all factors should not be deleted at the same time. It is possible that their insignificance is due to the influence of the “worst” of the insignificant factors, and in the next step of calculations these factors will turn out to be significant. So, the factor I₈, which is characterized by the smallest value of t-criteria should be deleted. In this case we get the following statistics (table 4).

### Table 4. Values of coefficients estimates of the corrected model for the factor “The extent of railway line”
Further, for the same reasons, we successively delete the factors $I_4$ and $I_5$ from the model. So, get the following statistics (table 5).

**Table 5.** Values of coefficients estimates of the corrected model for the factors “The population size” and “The number of the population with the incomes below the subsistence level”

| n=17 | Parameter values | Standard parameter errors | Actual values of the $t$ - criteria | P-value |
|------|------------------|---------------------------|-------------------------------------|---------|
| $b_0$ | 202.39519        | 423.11676                 | 0.47834                             | 0.64177 |
| $I_3$ | -0.11634         | 0.06363                   | -1.82831                            | 0.09472 |
| $I_4$ | -0.20102         | 0.38111                   | **-0.52745**                        | 0.60835 |
| $I_5$ | 0.30391          | 0.43555                   | 0.69775                             | 0.49982 |
| $I_6$ | -0.00842         | 0.00405                   | -2.07811                            | 0.06190 |
| $I_7$ | 0.00085          | 0.00045                   | 1.91402                             | 0.08198 |

The model with the estimated parameters has the form:

$$I_1 = -22.06281 - 0.12204 \cdot I_3 - 0.00820 \cdot I_6 + 0.00088 \cdot I_7.$$  \hspace{1cm} (1)

The graph of the observed values and predicted values from model (10) of the total of passenger traffic is in figure 1.

![Graph](image)

**Figure 1.** The graph of the observed values and predicted values from model (10) of the total of passenger traffic

The multiple linear regression model (1) approximates the statistical data.

The next stage of the research is the verification of the adequacy of the multiple linear regression model.
Using Student $t$-test and Fisher $F$-criteria to check the statistical reliability of the regression coefficients we get the assumptions of residuals $\epsilon_i$ – they are independent random variables, their mean value is 0, they have the same variance and are subject to a normal distribution.

After constructing the regression equation we check the estimates $\epsilon_i$ of those properties that were supposed.

This is due to the fact that estimates of regression parameters should be unbiased and effective.

The adequacy of the models is based on the following criteria described above:
- peaks criteria;
- criteria based on the coefficient of autocorrelation;
- $RS$-criteria.

The homoscedasticity of the variance of residuals due to the small sample size is visually estimated according to the graph of the residues represented in the time sequence. If the residuals on graph are in the form of a horizontal strip then we can say that the time effect does not influence the data (table 6).

### Table 6. Results of the model adequacy check (1):

| Criteria for verification of adequacy | Peaks criteria (in a random series must be $p > 6$) | The homoscedasticity (is estimated by the graph of the residues $\epsilon_i$, picture 2) | Criteria based on the coefficient of autocorrelation (the hypothesis of autocorrelation is rejected if $|r_1| < 0.299$) | $RS$ - criteria (the hypothesis of a normal distribution is accepted if $3.18 < RS < 4.49$) |
|-------------------------------------|-----------------------------------------------|-----------------------------------------------|-----------------------------------------------|-----------------------------------------------|
|                                     | 9                                              | Yes                                           | 0.082                                         | 4.17                                          |

**Figure 2.** The graph of the residues

The adequacy check showed that the multiple linear regression model (1) is adequate and can be used to forecast total passenger traffic. In this regard, we will not check non-linear regression models.

To forecast factors it is proposed to approximate time series by one of the types of functions:
- polynomial function: $\psi(t) = a_0 + a_1 t + a_2 t^2 + \cdots + a_{n-1} t^{n-1}$; \hspace{1cm} (2)
- exponent: $\psi(t) = ab^t$; \hspace{1cm} (3)
- logarithmic parabola $\psi(t) = ab^t c^2$; \hspace{1cm} (4)

Using a variety of functions to approximate time series allows us to improve the quality of forecasting.
The least squares method is used for estimating the parameters of models, when the values of the parameter vector minimize the function [4-7]:

\[ S(\alpha) = \sum_{t=1}^{m} (y_t - (\alpha, \psi(t)))^2. \]  

(5)

The trend model of the time series is considered to be adequate if it correctly reflects the systematic components of the time series. This requirement is equivalent to requiring that the residual component \( e(t) = y_t - \psi(t), t = 1, 2, ..., m \) satisfies the properties of the random component of the time series:

1. randomness of the residual level sequence;
2. the correspondence of the distribution of the random component to the normal distribution;
3. the equality of the mathematical expectation of the random component to zero;
4. independence of the values of the levels of the random component.

The results of experiments of modeling to forecast factors are presented in Table 7 (the values in parentheses are true values). The estimates of model parameters are given in table 8.

A retrospective forecast for 2016 is used according to data from 2000 to 2015 to estimate forecast properties of models.

The best model was chosen as a result (forecast factors for 2016 for this model are shown in bold in table 7).

**Table 7. Forecast of factor values for 2016**

| №  | The name of the model (the method of estimating the parameters) | \( I_3 \)       | \( I_6 \)       | \( I_7 \)     |
|----|-------------------------------------------------------------|-----------------|-----------------|---------------|
|    |                                                             | (291.70)        | 22846.40        | 368297.00     |
| 1  | parabola of the second power                                 | 305.25          | 25492.39        | 390676.91     |
| 2  | parabola of the third power                                  | 225.81          | 24280.86        | **373783.25** |
| 3  | exponent                                                    | 431.67          | 36440.11        | 393557.15     |
| 4  | logarithmic parabola                                         | **281.01**      | **22575.12**    | 396325.41     |

**Table 8. Model parameter estimates**

| №  | The name of the model | Parameter | \( I_3 \)       | \( I_6 \)       | \( I_7 \)     |
|----|-----------------------|-----------|-----------------|-----------------|---------------|
| 1  | parabola of the second power                                 | \( b_0 \)  | 8.61            | 218.70          | 158680.41     |
|    |                                                                    | \( b_1 \)  | 34.00           | 772.05          | 7794.76       |
|    |                                                                    | \( b_2 \)  | -0.97           | 42.04           | 344.24        |
|    |                                                                    | \( b_0 \)  | 88.05           | 1430.22         | 175574.07     |
|    |                                                                    | \( b_1 \)  | -14.83          | 27.30           | -2590.17      |
| 2  | parabola of the third power                                   | \( b_2 \)  | 5.99            | 148.31          | 1826.14       |
|    |                                                                    | \( b_3 \)  | -0.27           | -4.17           | -58.11        |
|    |                                                                    | \( a \)    | 78.88           | 1768.66         | 157911.61     |
| 3  | exponent                                                       | \( b \)    | 1.11            | 1.19            | 1.06          |
|    |                                                                    | \( a \)    | 51.35           | 1095.71         | 159022.35     |
| 4  | logarithmic parabola                                           | \( b \)    | 1.28            | 1.40            | 1.05          |
|    |                                                                    | \( c \)    | 0.99            | 0.99            | 1.00          |

An adequacy check was carried out for the selected factor models. The selected models have been tested for adequacy for all four criteria and will be used to forecast factors.

The forecast value of passenger transportation \( I_1 \) for 2016 is determined by substituting forecast values of factors in the equation of multiple linear regression (1). True value of \( I_1 = 92.90 \) (table 1).
The calculated values of the model (7) are:

\[ I_1 = 89.30, \text{absolute forecast error 3.6}, \text{relative forecast error 3.9\%}. \]

3. Conclusions

The developed model will be used to forecast the volume of passenger traffic in the Transbaikal region in the aspect of updating the “Strategy of social and economic development of the Transbaikal region until 2030” in conditions of ensuring the competitiveness of the region transport system.

4. References

[1] Shiryaev D V, Litvinenko I L, Rubtsova N V, Kolpak E P, Blaginin V A and Zakharova E N 2016 Economic clusters as a form of self-organization of the economic system International J. of Economics and Financial Issues 6(S1) 284–8

[2] Bulatova N N, Korodyuk I S and Sandakova N Y 2015 Model of innovative development of regional transport infrastructure Int. J. of Applied Engineer. Res. 10(20) 40834–40

[3] Tarsitano A and Lombardo R A 2013 Coefficient of correlation based on ratios of ranks and anti-ranks J. of Economics and Statistics 233 (2) 206–24

[4] Uvarov N D, Kuznetsov M P, Malkova A S, Rudakov K V and Strijov V V 2018 Selecting the Superpositioning of Models for Railway Freight Forecasting Moscow University Computational Mathematics and Cybernetics 42(4) 186–93

[5] Gorbatenko E O A methodological approach to the evaluation and prediction of bankruptcy in the region 2018 European Social Sci. J. 4 167–76

[6] Fedotov D Y 2017 Forecasting of regional economic development and budget based on the example of Irkutsk oblast Studies on Russian Economic Development 28(4) 416–22

[7] Krakovsky Y M, Luzgin A N and Mikhailova E A 2018 Interval forecasting of cyberattack intensity on informatization objects of industry using probability cluster model J. of Physics: Conf. Ser. 1015 032088