https://doi.org/10.26552/com.C.2015.2.30-34

Ludmila Janosikova - Martin Slavik *

MODELLING PASSENGERS' ARRIVALS AT PUBLIC TRANSPORT STOPS

The paper presents statistical exploration of passengers' arrivals at bus stops in urban public transport. First, it describes the methodology which is applied on the urban public transport system where the following conditions are met: (i) passengers are familiar with the timetables, (ii) the vehicles run on time, and (iii) the capacity of the vehicles is sufficient. The methodology is demonstrated on the urban public transport in Zilina, Slovak Republic. The correlation analysis reveals that there is a correlation between the waiting time and headway. The relationship between these variables can be described by a linear function or better by a logarithmic function. The Kolmogorov-Smirnov and the chi-square tests accept the hypothesis that the Gumbel minimum distribution fits for modelling passengers' arrivals rates. The proposed models are helpful in public transport planning.

Keywords: *Public transport, passengers' waiting time, correlation analysis, regression analysis.*

1. Introduction

The paper describes the results of our research aimed at the behaviour of passengers in urban public transport, particularly at their arrivals at bus stops. The research goal was:

1. To determine whether there exists a relationship between passengers' waiting time and the line frequency (headway between the successive buses);
2. If the waiting time depends on the headway, to determine the mathematical model of this dependence;
3. To propose a suitable mathematical model of passengers' arrivals rate at a stop.

The time passengers spend at the stops waiting for a bus has been regarded as an important criterion of the public transport system quality (see for example [1, 2, 3, 4 and 5]. Therefore, it is important to know the aspects affecting passengers' arrivals at the stops in order to design an effective public transport system or to improve its quality. Moreover, description of arrival patterns by mathematical models is necessary if one wants to use sophisticated methods for public transport planning, such as operations research methods [4, 6 and 7] or computer simulation [8].

There is a widespread belief that the waiting time depends on the line frequency. Published mathematical models of waiting time consider both the headway and the arrival time (or waiting time) as random variables. The most simple model is based on the assumption that passengers do not know the timetables, they arrive at the original stop randomly, and the mean waiting time is proportional to the headway (inversely proportional to the line frequency, respectively). Under the assumption that passengers arrive at a constant rate, the waiting time is a function of the mean headway and its variance:

$$E(W) = \frac{E(H)}{2}\left(1 + \frac{D(H)}{E^2(H)}\right), \tag{1}$$

where $E(W)$ denotes the mean waiting time, $E(H)$ the mean headway and $D(H)$ the headway variance.

When a transportation service operates with long headways, passengers do not arrive at stops randomly but they tend to arrive few minutes before the planned vehicle departure. Previous studies performed in Europe in the 1970s [9] and in the U.S.A. [4] were aimed at the determination of the minimum headway with non-random arrival pattern and the model for the relationship between the waiting time and the headway. The headway threshold varied from 5 to 12 minutes and the models were linear or quadratic. The recent European study was carried out in Zürich by Luethi et al. [9]. Passengers were supposed to belong to one of two groups: those who were familiar with the schedule and those who did not know the schedule. As a consequence, the authors suggest an arrival rate model that combines the uniform distribution for informed passengers with the shifted Johnson S_B distribution for uninformed passengers. The Johnson distribution is shifted with a small value due to the observation that some passengers arrive short time after the vehicle departure. The reported share of these

* **Ludmila Janosikova, Martin Slavik**
Department of Mathematical Methods and Operations Research, Faculty of Management Science and Informatics, University of Zilina, Slovakia
E-mail: Ludmila.Janosikova@fri.uniza.sk

passengers is quite high (from 5 to 16 %). The authors explain passengers' early arrivals by the fact that passengers do not trust the service reliability and rely on a regular delay but then they do not catch the bus. Passengers waiting the whole period would be observed also if they failed to board due to the insufficient vehicle capacity. Regarding the dependence of the mean waiting time on the headway, Luethi et al. propose the logarithmic model.

2. Methodology

The first goal of our study was to find out whether the waiting time depends on the headway even in the case of reliable service with the sufficient capacity of vehicles and with passengers being familiar with the schedules. We proceeded from the situation in the Slovak Republic, which is similar to most European countries, where the public transport users are well-informed. The timetables are available at the stops as well as on the Internet, so passengers are able to obtain schedule information almost everywhere by using new information technologies.

As it was said before, both the waiting time (or passengers' arrival time, respectively) and the headway are random variables. The dependence of random variables is the matter of the correlation analysis. There are several correlation coefficients measuring the dependence between two random variables X and Y. The most common of these are the Pearson correlation coefficient $R_{X,Y}$ and the Spearman correlation coefficient R_S. They are calculated using a series of n measurements of X and Y denoted as (x_i, y_i) for i = 1, ..., n. The estimated correlation coefficients $R_{X,Y}$ and R_S are almost always different from zero. Therefore a statistical test should be preformed to verify whether their value is statistically significant. The null hypothesis
H_0: $\rho = 0$ (correlation is insignificant)
is tested against the alternative hypothesis
H: $\rho \neq 0$.

Several tests with different test criteria are available. The test criteria are functions of the estimated correlation coefficient.

The first test is based on the assumption that the sample of the pairs (x_i, y_i) for i = 1, ..., n comes from the two-dimensional normal distribution with the correlation coefficient ρ. The test criterion

$$T = \frac{R_{X,Y}\sqrt{n-2}}{\sqrt{1-R_{X,Y}^2}} \tag{2}$$

has the Student's t distribution with n – 2 degrees of freedom under the null hypothesis.

The second test is based on the same assumption as for the normal distribution. The test uses the Fisher z transformation that converts the Pearson correlation coefficient to the variable Z. The formula for the transformation is:

$$Z = \frac{1}{2}\ln\frac{1+R_{X,Y}}{1-R_{X,Y}} \tag{3}$$

Z is approximately normally distributed with the mean $\frac{1}{2}\ln\frac{1+\rho}{1-\rho}$ and the standard deviation $\frac{1}{\sqrt{n-3}}$.

The test criterion

$$Z^* = \frac{\sqrt{n-3}}{2}\ln\frac{1+R_{X,Y}}{1-R_{X,Y}} \tag{4}$$

has the standard normal distribution $N(0,1)$ under the null hypothesis.

If the assumption of the above mentioned tests is not met, the nonparametric test for the Spearman correlation coefficient can be used. The value R_s is compared with the tabulated critical value r_α. If $|R_S| \leq r_\alpha$, then the null hypothesis is accepted.

In the case that random variables X and Y are dependent, one can describe their relationship by a regression function. The most simple form of the regression function is the linear function $y = ax + b$. The logarithmic function $y = a\ln(x) + b$ is often used as well. Unknown coefficients a and b need to be estimated using a measured sample of X and Y. The most common method of estimation is the least squares method.

3. Case study

As a case study for passengers' arrivals and waiting time modelling we chose the urban public transport in the city of Žilina. The transportation service in the city is provided by the transportation operator Dopravny podnik mesta Ziliny, s.r.o. (DPMZ). During the day, 8 trolleybus lines and 10 bus lines operate. At night, the city area is served by 1 bus line.

The data for the analysis were collected at 6 stops in Zilina on weekdays during the morning peak and off-peak periods (from 6:00 to 11:00). The stops were selected according to the following criteria:

- Passengers are not supposed to change lines at the stop.
- The stop must be busy enough to enable collecting sufficient data.

The data were collected "by hand", i.e. by observing passengers' arrivals at stops and recording the passenger's arrival time, the number of the line taken by the passenger, and the vehicle departure time.

3.1 The results of the correlation analysis

Using the measured data we want to determine whether there exists a relationship between passengers' waiting time (random variable Y) and the line headway (random variable X). The size of the sample used in the following calculations is n = 467.

The relationship is measured by the Pearson and Spearman correlation coefficients that take the values $R_{X,Y}$ = 0.134 and

R_S = 0.759. Both coefficients are different from zero, which means that the waiting time and the headway are related according to a monotonic function. This finding can be verified by the test of significance of the theoretical correlation coefficient ρ. The null hypothesis

H_0: ρ = 0 (correlation is insignificant)

is tested against the alternative hypothesis

H: $\rho \neq 0$.

at the level of significance α = 0.05.

In the first two tests with the Pearson correlation coefficient it is assumed that the sample of pairs (X,Y) comes from the normal distribution. As regards the variable X (the line headway), it takes only several values, most often 10, 15, 20, and 30 minutes that are common in public transport operation and therefore it is impossible to make a test on its probability distribution. The sample of Y was tested on the probability distribution for particular headways x_i (see Section 3.3). Although the chi-square test failed to reject the hypothesis about the normal distribution for some x_i, for all data together the hypothesis was rejected at the significance level α = 0.05. Although the assumption of the first two tests was not proved, all three tests mentioned in Section 2 were performed.

The value of the test criterion according to (2) is T = 2.919. It is greater than the critical value of the Student distribution for the significance level 0.05 and n-2 degrees of freedom (1.965), therefore we reject the null hypothesis H_0 and accept the alternative hypothesis H that X and Y are dependent random variables.

The value of the test criterion according to (4) is Z^* = 2.907. It is greater than the critical value of the standard normal distribution for the significance level 0.05 (1.96), therefore we reject the null hypothesis H_0 and accept the alternative hypothesis H that X and Y are dependent random variables.

The same outcome is obtained using the third nonparametric test with the Spearman correlation coefficient. The value R_s = 0.759 is greater than the tabulated critical value r_α = 0.091, so the null hypothesis is rejected.

Further, the confidence interval of the Pearson correlation coefficient can be calculated. The 95% confidence interval is $\langle 0.044, 0.222 \rangle$. This interval includes the Pearson correlation coefficient $R_{X,Y}$ = 0.134. This fact confirms the correlation between the waiting time and the headway.

3.2 The results of the regression analysis

To specify the dependence mathematically, a regression function can be derived, which describes the dependence of the pair of random variables (X,Y). Using the least square method, linear and logarithmic functions were proposed, further the significance of coefficients was investigated and the quality of both models was examined using the F-test.

The linear function was estimated as y = 0.088x + 3.932. The 95% confidence intervals of the coefficients are: $a \in \langle 0.029, 0.147 \rangle$, $b \in \langle 2.770, 5.095 \rangle$. None of the intervals contains 0, therefore both coefficients a and b are significant.

The validation of the model can be done through the F-test on the statistical significance of the regression model. Let us denote $SS_{reg} = \frac{S_{reg}}{p-1}$ and $SS_{err} = \frac{S_{err}}{n-p}$, where p is the number of regression parameters, S_{reg} is the regression sum of squares, also called the explained fraction of variance, and S_{err} is the residual sum of squares, also called the unexplained fraction of variance.. The regression model is considered to be statistically significant if SS_{reg} is significantly greater than SS_{err}. The F-test states the null hypothesis

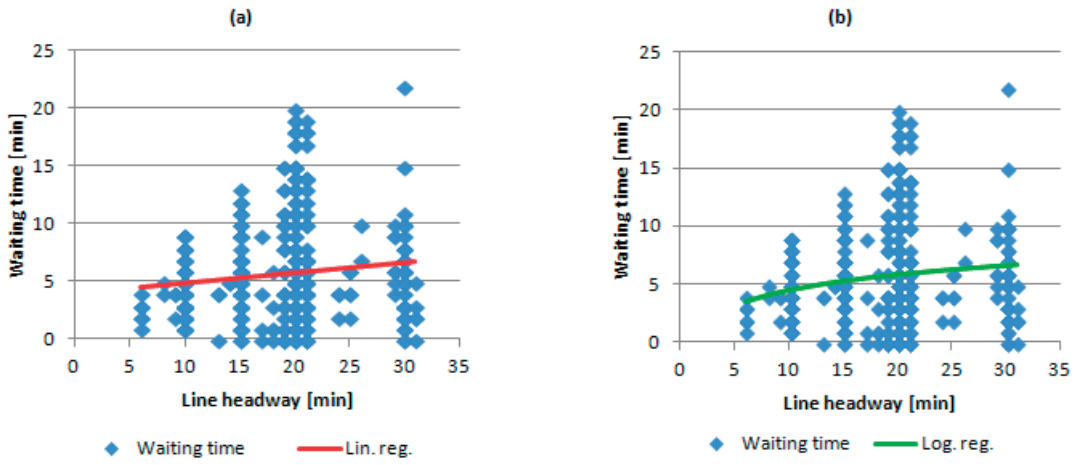

Fig. 1 (a) Linear and (b) Logarithmic regression

H_0: $SS_{reg} = SS_{err}$ (the regression model is insignificant) against the alternative hypothesis H: $SS_{reg} > SS_{err}$.

The value of the test criterion is $F = 8.521$. It is greater than the critical value of the Fisher-Snedecor distribution with the parameters $v = p - 1 = 1$ and $w = n - p = 465$ for the significance level 0.05 ($F_{0.95}(1,465) = 3.861$), therefore we reject the null hypothesis H_0 and accept the alternative hypothesis H that the explained fraction of variance is significantly greater than the unexplained fraction of variance. The coefficient of determination for this linear model is $R^2 = 0.018$.

The logarithmic function was estimated as $y = 1.930\ln(x) + 0.047$. The 95% confidence intervals of the coefficients are: $a \in \langle 0.923, 2.937 \rangle$, $b \in \langle -2.857, 2.950 \rangle$. The confidence interval for coefficient b contains 0, but this only means that the absolute part of the function may be zero, however, the type of the function is still logarithmic.

The value of the F-test criterion is $F = 14.185$. It is greater than the critical value of the Fisher-Snedecor distribution ($F_{0.95}(1,465) = 3.861$), therefore we reject the null hypothesis H_0 and accept the alternative hypothesis H that the logarithmic model is statistically significant. The coefficient of determination is $R^2 = 0.030$.

The coefficients of determination for both regression models are quite small. It means that the values $\hat{y}_i$ on the regression curve are far away from the observed values y_i. The reason is that for each headway x_i there were a lot of different waiting times y_i observed, which can also be seen in Figs. 1a - 1b. In accordance with [9], the logarithmic dependence seems to be a better approximation of the relationship between the examined random variables.

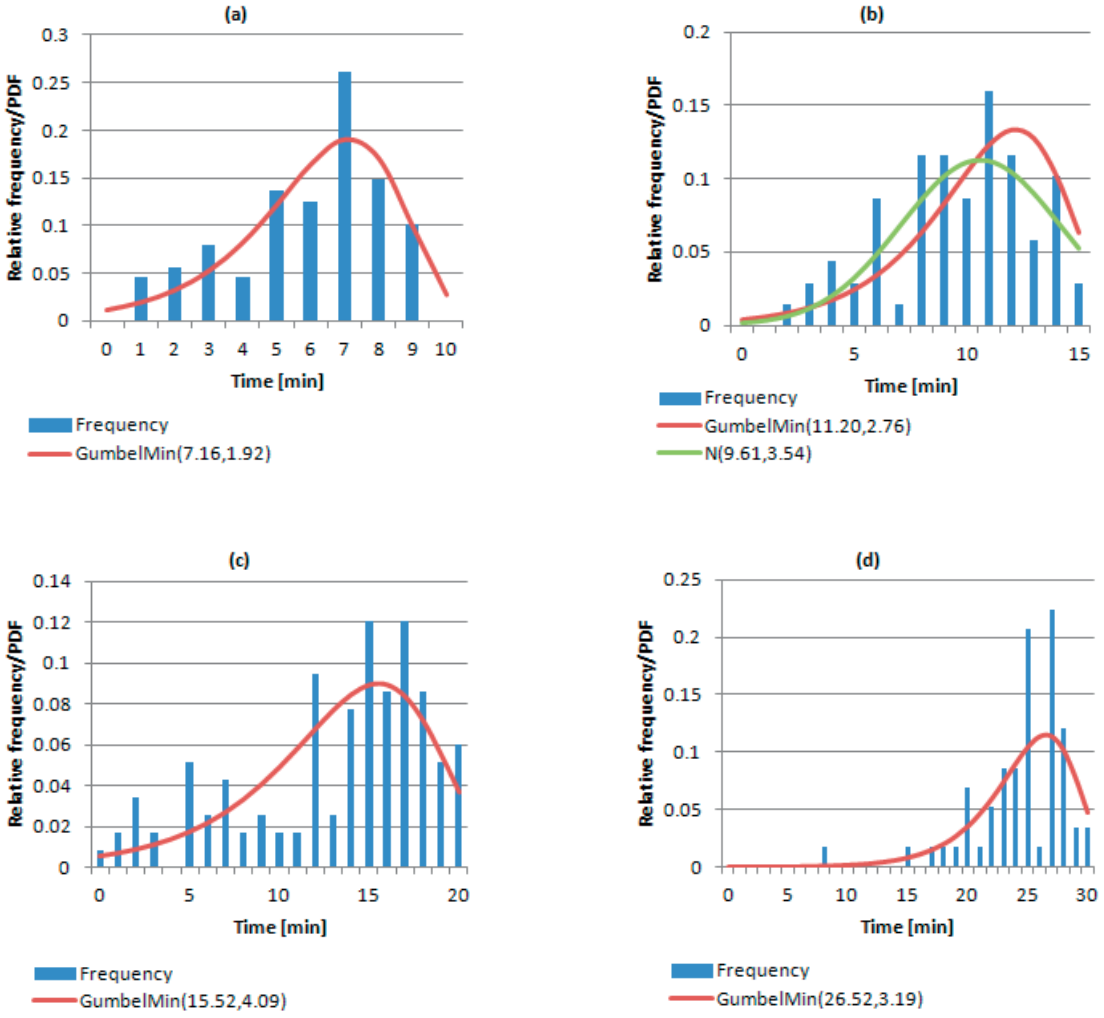

Fig. 2 Passengers' arrivals in (a) 10-minute headway, (b) 15-minute headway, (c) 20-minute headway, (d) 30-minute headway

3.3 The distribution of the arrival rates

The next step in our research was to specify the distribution of passengers' arrivals for the most common headways (10, 15, 20, and 30 minutes). The random variable is now the time elapsed between the departure of the previous vehicle and the arrival of the passenger. The frequency diagrams of the passengers' arrival times were constructed for each headway. The diagrams suggested that the Gumbel minimum distribution would be a suitable model. The formula for the probability density function (PDF) of the Gumbel minimum distribution is

$$f(x) = \frac{1}{b}\exp\left(\frac{x-a}{b} - \exp\left(\frac{x-a}{b}\right)\right) \qquad (5)$$

for $x \in (-\alpha, \alpha)$; $a \in (-\alpha, \alpha)$ is the location parameter and $b > 0$ is the scale parameter.

Using the Kolmogorov-Smirnov and the chi-square tests we can accept the hypothesis that the arrivals follow the Gumbel minimum distribution. Moreover, for the 15-minute headway also the normal distribution was accepted by the tests.

For illustration, Figs. 2a - 2d display the relative frequencies of passengers' arrivals and the PDF of the Gumbel minimum distribution for four most common headways. As it can be seen, the location parameter of the PDF strongly depends on the headway since passengers tend to arrive at the boarding stop few minutes before the planned bus departure. Only a couple of passengers arrive at the beginning of the period. So we can conclude that most passengers are familiar with the timetables and adjust their arrivals to the schedule.

4. Conclusions

We proposed a methodology where statistical methods are used for the exploration of passengers' arrivals at bus stops in urban public transport. The main results of our research are as follows:

1. Passengers' waiting time and line headway are correlated random variables.
2. The relationship between these variables can be modelled by a linear function or better by a logarithmic function.
3. The Gumbel minimum distribution is the suitable mathematical model of passengers' arrivals; the location parameter of the PDF strongly depends on the line headway.

The results can be generalized for every public transport system in which the users are well-informed. The proposed models can be used to generate demand input in the operations research and simulation methods that are helpful tools in the effort for improving public transport quality.

Acknowledgements

This research was supported by the Scientific Grant Agency of the Ministry of Education of the Slovak Republic and the Slovak Academy of Sciences under project VEGA 1/0339/13 "Advanced microscopic modelling and complex data sources for designing spatially large public service systems" and by the Slovak Research and Development Agency under project APVV-0760-11 "Designing of Fair Service Systems on Transportation Networks". The authors wish to thank students who helped with data collection: Jozef Piecka, Ivana Urbanicova, and Alzbeta Janosikova.

References

[1] ALVAREZ, A., CASADO, S., GONZALEZ VELARDE, J. L., PACHECO, J.: A Computational Tool for Optimizing the Urban Public Transport: A Real Application. *J. of Computer and Systems Sciences International* 49(2), 2010, pp. 244-252.

[2] AVINERI, E.: A Cumulative Prospect Theory Approach to Passengers Modeling: Waiting Time Paradox Revisited. *J. of Intelligent Transportation Systems* 8(4), 2004, pp. 195-204.

[3] DAGANZO, C. F.: *Fundamentals of Transportation and Traffic Operations*. New York: Elsevier Science Inc., 1997.

[4] FAN, W., MACHEMEHL, R. B.: *Optimal Transit Route Network Design Problem: Algorithms, Implementations, and Numerical Results.* Report No. SWUTC/04/167244-1, Center for Transportation Research: University of Texas at Austin, 2004.

[5] OSUNA, E. E., NEWELL, G. F.: Control Strategies for an Idealized Public Transportation System. *Transportation Science* 6(1), 1972, pp. 52-72.

[6] JANACEK, J., KOHANI, M.: Waiting Time Optimisation with IP-solver. *Communications - Scientific Letters of the University of Zilina*, 12(3A), 2010, pp. 36-41.

[7] JANOSIKOVA, L., KOHANI, M., BLATON, M., TEICHMANN, D.: Optimization of the Urban Line Network Using a Mathematical Programming Approach. *Intern. J. of Sustainable Development and Planning* 7(3), 2012, pp. 288-301.

[8] ERATH, A., VAN EGGERMOND, M. A. B., FOURIE, P. J., CHAKIROV, A.: *Decision Support Tools in Transport Planning: From Research to Practice.* Paper presented at the 13th Swiss Transport Research Conference, Ascona, April 2013.

[9] LUETHI, M., WEIDMANN, U., NASH, A.: *Passenger Arrival Rates at Public Transport Stations.* Working paper doi:10.3929/ethz-a-005704674, Zurich : Institute for Transport Planning and Systems: ETH Zurich, 2006.