ANALYSIS OF RAIL TRANSPORT ACCIDENTS USING ONE-WAY ANALYSIS OF VARIANCE

Jana Motyková, Libor Švadlenka[1]

Summary: The article deals with the analysis of rail transport accidents in the period of 2008 – 2015 and with the comparison of their causes. The existence of a relationship between causes of accidents and calendar months when the accidents have occurred was determined by using one-way analysis of variance.

Key words: rail transport, accidents, one-way analysis of variance.

INTRODUCTION

The transport is one of the tools to fulfill the needs of society. It belongs to the everyday life of all people. Railway transport is the most traditional mode of transport. Its origins date back to ancient times when the ancient Greeks used the rails to make easier to get into their temples. Former rails did not look like the current rails. There were notches on the road that permitted the passage of vehicles. The first railway in our area was built in 1828 and it was a horse-way Linz – Summerau – Horní Dvořiště – České Budějovice. As that time the road and air transport were on their beginning, railway transport had no rival and it became the fastest way of moving passengers and freights in history. Construction of the railway infrastructure meant the work for thousands of people, encouraged technical development, speed up industrial revolution and allowed expansion of international trade. Places where railways have led registered rapid economic growth. (8, 10)

After World War II, road and air transport have developed and rail transport has distinctly declined. It is mainly due to the fact that road transport can quickly and more flexible respond to the changes in demand for transport (due to a denser network). On the contrary air transport offers very fast transportation but not directly to the city centers. In spite of above mentioned facts relating to the rail transport this mode of transport is irreplaceable. There is transported half a million people and more than quarter of million tons of goods by railway every day in the Czech Republic. And that is why it is necessary to focus attention on safety in the rail transport. (4, 8, 10)

§ 49 of the Railways Act defines an accident in rail transport as *"a serious accident, accident or danger in the rail transport that threatens or impairs safety, regularity and*

[1] Ing. Jana Motyková, University of Pardubice, Jan Perner Transport Faculty, Department of Transport Management, Marketing and Logistics, Studentská 95, 532 10 Pardubice, Czech Republic, Tel.: +420 466 036 383, E-mail: jana.motykova@student.upce.cz
doc. Ing. Libor Švadlenka, Ph.D., University of Pardubice, Jan Perner Transport Faculty, Department of Transport Management, Marketing and Logistics, Studentská 95, 532 10 Pardubice, Czech Republic, Tel.: +420 466 036 375, E-mail: libor.svadlenka@upce.cz

smoothness of rail transport, the safety of persons and the safe function of buildings and equipment or threatens environment." (1)

During the period 2008 – 2015 were happened on average less than 1200 accidents per year in rail transport meaning on average 3 accidents per day. The goal of all transport services providers is to reduce incidents to the lowest possible level. Analysis of the causes of rail transport accidents could help in this way. (9)

Statistics of accidents in rail transport are led by an infrastructure manager since 2008 in the Czech Republic. Infrastructure manager distinguishes accidents depending on the impact that event has, respectively what is the size of human and material losses. This segmentation is specified in Regulation D17 for the reporting and investigation of incidents and detailed specifications are described in the Provision D17-1 related to above mentioned Regulation D17. (11, 12)

For the purpose of the paper, it is necessary to introduce different segmentation compare with that used by infrastructure manager. Accidents won't be divided into the groups by consequences (as it stated in the Regulation D17 and the Provision D17-1) but according to their causes. This adjusted data can be used as the base for the statistical analysis of accidents in rail transport.

1. METHODOLOGY

This chapter provides a theoretical basis for the statistical methods that has been used for the analysis of accidents in rail transport which occurred during the period 2008 – 2015 in the Czech Republic.

1.1 Adjustment for calendar variations

Adjustment of time series of calendar variations is particularly important for the reason that the data of rail transport accidents are related to different time intervals. Some months last 31 days, some 30 days and February has only 28 or 29 days. So we have to recalculate all seasons to the unitary time interval. Then the results of statistical analysis won't be misleading. This operation is called an adjustment for calendar variations and it is performed by the following formula (1). (3, 6)

$$y_t^{(\emptyset)} = y_t \frac{\overline{k}_t}{k_t} \tag{1}$$

When the value y_t is the value of adjusting indicator in the relevant period of the year t (in our case it is the number of rail transport accidents in the relevant month and in the relevant year). The value $\overline{k}_t$ indicates the average number of calendar days in the relevant time interval of the year. And the value k_t indicates the number of calendar days in the relevant time interval of the year. (3, 6)

1.2 One-way ANOVA

One-way ANOVA is the simplest sort of the analysis of variance when we analyze the influence of one factor x (in our case month) on the dependent variable y (in our case it is accidents that were founded by specific causes). One-way ANOVA detects differences between the averages of several groups that represent individual level (category) of the reference factor by calculation of testing criterion F_n. We find whether the groups created by classification factor are similar or whether the individual averages generate some identifiable groups (homogeneous subgroups with similar values). (2, 3, 5, 6, 7, 13)

One-way ANOVA tests the null hypothesis saying that mean values of the groups are equal (it counts with the fact that the groups are influenced by natural variability). Testing is performed on the basis of the analysis of relationships between the variances in each group using the *F*-test. In other words, the measurable variables *y* don't depend on the variable *x*. The alternative hypothesis says that not all mean values are the same, i.e. at least one is different from others. Thus there is a dependence between the variables *x* and *y*. Hypotheses are given by equations (2) and (3). (2, 3, 5, 6, 7, 13)

$$H_0: \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_n \quad (2)$$

$$H_1: \mu_1 \neq \mu_2 \neq \mu_3 \neq \cdots \neq \mu_n \quad (3)$$

Basic statistics calculated in the one-way ANOVA model is a generally testing criterion F_n, which is calculated by formula (4), where *n* is the total number of observations, *k* is the number of groups, *t* is the monitored period, $SSR_n(t)$ is given by equation (5) and $SSE_n(t)$ is given by equation (6). (2, 3, 5, 6, 7, 13)

$$F_n(t) = \frac{SSR_n(t)}{k-1} \Big/ \frac{SSE_n(t)}{n-k} \quad (4)$$

$$SSR_n(t) = \sum_{i=1}^{k} n_i \left(\bar{Y}_i(t) - \bar{Y}(t)\right)^2 \quad (5)$$

$$SSE_n(t) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(Y_{ij}(t) - \bar{Y}_i(t)\right)^2 \quad (6)$$

The result of a generally testing criterion F_n is compared with critical value $F_{1-\alpha;(k-1,\, n-k)}$ where α is the significance level, $k-1$ and $n-k$ are degrees of freedom. If the equation (7) is valid i.e. the value of testing criterion falls into the critical area, we should reject the null hypothesis H_0 and we should accept the alternative hypothesis H_1. Alternative hypothesis H_1 says that the dependence between the researched variables exists. (2, 3, 5, 6, 7, 13)

$$F > F_{1-\alpha;(k-1,n-k)} \quad (7)$$

Otherwise, if the value of testing criterion F_n falls outside the critical area (defined by the critical value $F_{1-\alpha;(k-1,\, n-k)}$) , we should accept the null hypothesis H_0. The dependence between the researched variables doesn't exist. Results of the partial calculations are usually written into tabular form (see Table 1). (2, 3, 5, 6, 7, 13)

Table 1 – Table for one-way ANOVA

Source of variability	*Sums of Squares (SS)*	*Degree of Freedom (DF)*	*Mean Square (MS)*	*Testing Criterion (F_n)*	*Critical value (F)*
Factor	$SSR_n(t)$	k – 1	$SSR_n(t)$ / (k – 1)	$F_n(t)$	$F_{1-\alpha;(k-1,\, n-k)}$
Error	$SSE_n(t)$	n – k	$SSE_n(t)$ / (n – k)	–	–
Total	$SSR_n(t)$+ $SSE_n(t)$	n – 1	–	–	–

Source: (2, 3, 5, 6, 7, 13), authors

2. MEASUREMENT

Infrastructure manager keeps detailed statistics about the occurrence of rail transport accidents in the Czech Republic. The statistical data about accidents in rail transport related to the period 2008 – 2015 are the base for this paper. Authors of the paper don't divide the incidents according to the extent of human and material losses (as the infrastructure manager does) but they divide incidents by the causes. These factors have been found and filtered in statistics:

1. unavoidable events (natural causes),
2. railway operators,
3. operators of rail transport,
4. suicides,
5. third persons (e.g. drivers of motor vehicles, bicyclists, etc.).

It seems to be logical to focus on the largest group of accidents. But sometimes it is not possible and easy. Methods of mathematical statistics can detect various dependencies and tightness of dependencies between accidents and monitored indicators. In this case we have investigated the correlation between the frequency of accidents caused by a specific cause and month of the year.

After we had divided accidents into the groups according to cause, we adjusted the time series from calendar variations. It is due to the fact that all months of the year have not the same number of days. In the monitored period 2008 – 2015 were two leap years. Thus adjusted time series can be analyzed by one-way ANOVA.

2.1 Application of adjustment for calendar variations

The first step of the analysis of rail transport accidents is the time series adjustment for calendar variations. Individual monitored time interval of the year (months) are converted to the same time interval according to equation (1). Input data are the number of accidents in the years 2008 – 2015 so we have to introduce two formulas for calculating the adjusted values.

The first formula is applied for classical, respectively non-leap year, and it is defined by equation (8). Where y_t is the original unadjusted number of accidents which occurred in a particular month of a particular year (e.g. 90 accidents occurred in February 2011 – value $y_t = 90$). The value of indicator $\overline{k}_t$ is the average duration of the months in the monitored year (e.g. this value is 365/12 for the year 2013). The k_t value is the number of days in the monitored month in the year (e.g. the value is 28 days in February 2011).

$$y_t^{(ON)} = y_t \frac{365/12}{k_t} \qquad (8)$$

The second formula is applied for leap years (for the years 2008 and 2012) and it is defined by equation (9). The difference in the calculation is the value of indicator $\overline{k}_t$ which is equal to 366/12 and the value of indicator k_t which is equal to 29 in the leap year's February.

$$y_t^{(OL)} = y_t \frac{366/12}{k_t} \qquad (9)$$

2.2 Application of one-way ANOVA

The number of rail transport accidents which occurred in each of months during the period 2008 – 2015 is known and simply divided into the groups according to cause (see Chapter 1) and adjusted for calendar variations (see Chapter 2.1). The task of the one-way ANOVA model is to determine the dependence between the months of the year and the number of accidents in rail transport, e.g. in winter could be more accidents than in summer because of the icing, snow drifts, etc.

So we test the null hypothesis saying that accidents caused by a specific cause (i.e. measurable variable *y*) are not depending on the month (i.e. variable *x*). The null hypothesis is defined by the equal (2). On the opposite, the alternative hypothesis says that dependence between monitored variables exists. The alternative hypothesis is defined by the equal (3).

Table 2 is the default table for the calculation of the one-way ANOVA model for individual causes of accidents in the monitored period 2008 – 2015. The one-way ANOVA model was testing on the significance level α = 0,05 (5 %). A detailed interpretation of the results is given in Chapter 3.

Table 2 – One-way ANOVA – Rail transport accidents

Source of variability	*Sums of Squares (SS)*	*Degree of Freedom (DF)*	*Mean Square (MS)*	*Testing Criterion (F_{96})*	*Critical value (F)*
Factor	$SSR_{96}(t)$	12 – 1	$SSR_{96}(t)$ / (12 – 1)	$F_{96}(t)$	$F_{1-0,05;(12-1,\,96-12)}$
Error	$SSE_{96}(t)$	96 – 12	$SSE_{96}(t)$ / (96 – 12)	–	–
Total	$SSR_{96}(t)$+ $SSE_{96}(t)$	96 – 1	–	–	–

Source: authors

3. RESULTS

As we describe above, the goal of the analysis of rail transport accidents is to determine the dependence between the months of the year and the number of accidents during the period 2008 – 2015. Dependence has been assessed by using one-way ANOVA model in 5 levels according to the cause of accidents. These are:

1. unavoidable events (natural causes),
2. railway operators,
3. operators of rail transport,
4. suicides,
5. third persons (e.g. drivers of motor vehicles, bicyclists, etc.).

Results of the analysis are presented in the tables below (see Table 3 – Table 7). The most important values are always indicators F_{96} and F whose mutual comparison according to equation (7) shows that the null hypothesis is confirmed (see equation (2)) or the alternative hypothesis is confirmed (see equation (3)).

Results of one-way ANOVA are given in the Table 3. In this case we have examined the dependence between the number of accidents in rail transport caused by unavoidable events and the month when it had occurred. Results confirm the alternative hypothesis. The equal $F_{96} > F$ is valid. The dependence exists in here.

Table 3 – One-way ANOVA – Unavoidable events

Source of variability	*SS*	*DF*	*MS*	***F₉₆***	***F***
Factor	587,7128	11	53,42844	**2,241463**	**1,904539**
Error	2002,259	84	23,83642		
Total	2589,972	95			

Source: authors

Table 4 shows the results of the analysis of accidents that were caused by the railway operators. From the values of indicators F_{96} and F is evident that the dependence between accidents and months doesn't exist. We have accepted the null hypothesis.

Table 4 – One-way ANOVA – Railway operators

Source of variability	*SS*	*DF*	*MS*	$\boldsymbol{F_{96}}$	$\boldsymbol{F}$
Factor	1419,234	11	129,0213	**0,746622**	**1,904539**
Error	14515,76	84	172,8067		
Total	15934,99	95			

Source: authors

The analysis of rail transport accidents that were caused by the operators of rail transport using the one-way ANOVA (see Table 5) have similar result like the analysis of accidents that were caused by the railway operators (see Table 4). The dependence between accidents and month doesn't exist.

Table 5 – One-way ANOVA – Operators of rail transport

Source of variability	*SS*	*DF*	*MS*	$\boldsymbol{F_{96}}$	$\boldsymbol{F}$
Factor	592,6621	11	53,87837	**0,633767**	**1,904539**
Error	7141,083	84	85,0129		
Total	7733,745	95			

Source: authors

The analysis of rail transport accidents that were caused by the suicides can't be predictable or impressible by the infrastructure manager. Evidence of this is the result of the analysis described in the Table 6. The comparison of the values F_{96} and F shows that the dependence between accidents that were caused by the suicides and the month doesn't exist.

Table 6 – One-way ANOVA – Suicides

Source of variability	*SS*	*DF*	*MS*	$\boldsymbol{F_{96}}$	$\boldsymbol{F}$
Factor	178,7518	11	16,25016	**0,700247**	**1,904539**
Error	1949,331	84	23,20632		
Total	2128,083	95			

Source: authors

The last accidents that have been analyzed by one-way ANOVA are the accidents that were caused by the third persons (see Table 7) e.g. drivers of motor vehicles, bicyclists, etc. As in the previous three cases, the comparison of the values F_{96} and F confirms the validity of the null hypothesis. It means that there is no dependence between accidents that were caused by the third persons and month of the year.

Table 7 – One-way ANOVA – Third persons

Source of variability	*SS*	*DF*	*MS*	F_{96}	*F*
Factor	557,3414	11	50,6674	**0,456043**	**1,904539**
Error	9332,594	84	111,1023		
Total	9889,936	95			

Source: authors

CONCLUSION

The goal of the paper was the analysis of rail transport accidents in the period 2008 – 2015 in the Czech Republic. We have examined the dependence between the causes of the accidents and the month when the accidents occurred by the one-way ANOVA model. Statistical data were obtained from the infrastructure manager. The results of the analysis written in Chapter 3 declared that the only reason that is dependent on the month of the year is unavoidable event. This means that the month, respectively the weather that belongs to month, influenced the number of accidents.

Currently, the weather changes can be predicted with relatively high accuracy. So prediction of the accidents caused by the weather changes should be easy in the same way. The question is, if the measures realized against it by infrastructure manager are cost-effective. We have to consider the financial and material consequences that these incidents make. Accidents caused by the unavoidable events make up only 5 % of all the accidents in rail transport that were occurred during the period 2008 – 2015. The material losses are insignificant in comparison with other caused accidents. It is necessary to focus attention on other causes whose effect could have for the passengers and providers of transport services far more serious consequences.

ACKNOWLEDGEMENT

The work was created in connection with the scientific research project of the University of Pardubice SGS_2016_008. The authors are grateful for their support.

REFERENCES

(1) ČESKO. Zákon č. 266 ze dne 14. prosince 1994 o dráhách. In: *Sbírka zákonů České republiky*. 1994, p. 3041 – 3054. Available from: <http://aplikace.mvcr.cz/sbirka-zakonu/SearchResult.aspx?q=266/1994&typeLaw=zakon&what=Cislo_zakona_smlouvy> ISSN 1211 – 1244.

(2) GÓRECKI, T., SMAGA, Ł. A Comparison of Tests for the One-way ANOVA Problem for Functional Data. *Computational Statistics,* 2015, vol. 30, i. 4, p. 987-1010, ISSN 1613-9658.

(3) HINDLS, R., HRONOVÁ, S., NOVÁK, I. *Metody statistické analýzy pro ekonomy.* Praha: Management press, 2000. 259 p. ISBN 80-7261-013-9

(4) *Informace pro dopravní analýzy* [online]. © 2015 [cit. 2016-05-11]. Available from: <http://www.vyrocenky.cz/drazni>.

(5) KLIESTIK, T., LYAKIN, A., N., VLASAKOVA, K. Stochastic Calculus and Modelling in Economics and Finance. In: *2nd International Conference on Economics and Social Science (ICESS) 2014.* China, Shenzhen, 2014, vol. 61, p. 161-167. ISBN 978-1-61275-068-2.

(6) KUBANOVÁ, J., *Statistické metody pro ekonomickou a technickou praxi.* Bratislava: STATIS, 2008. 247 p. ISBN 978-80-85659-47-4

(7) NUNES, G. et al. One-way Fixed Effects ANOVA with Missing Observations. In: *Proceedings of the International Conference on Numerical Analysis and Applied Mathematics 2014.* Greece, Rhodes, 2015, vol. 1648. ISBN 978-07-3541-287-3

(8) PEŘINA, L., ZELENÝ, L., *Doprava, dopravní infrastruktura.* Praha: Vysoká škola ekonomická v Praze, 2000. 106 p. ISBN 80-245-0110-4

(9) *Statistics of rail transport incidents in the period 2008 – 2015.*

(10) SŽDC. Historie železnice. *SŽDC.cz* [online]. © 2009 – 2012 [cit. 2016-05-02]. Available from: http://www.szdc.cz/o-nas/zeleznice-cr/historie-zeleznice.html

(11) SŽDC. D17 Předpis pro hlášení a šetření mimořádných událostí. *SŽDC.cz* [online]. © 2009 – 2012 [cit. 2016-05-02]. Available from: http://www.szdc.cz/dalsi-informace/dokumenty-a-predpisy/provozne-technicke.html?page=detail&docid=1%3B%2375ec707a-7b2a-4af4-8fe2-2c05aaecc6cb

(12) SŽDC. D17 – 1 Prováděcí opatření k předpisu pro hlášení a šetření mimořádných událostí. *SŽDC.cz* [online]. © 2009 – 2012 [cit. 2016-05-02]. Available from: http://www.szdc.cz/dalsi-informace/dokumenty-a-predpisy/provozne-technicke.html?page=detail&docid=1%3B%237f8d5d59-4bea-4a70-86e8-a0717e00937b

(13) TABACHNICK, B., G., FIDELL, L., S. *Using Multivariate Statistics.* United States of America: Allyn & Bacon, 2001. 966 p. ISBN 0-321-05677-9