PROCESSING DATA FOR TIME SERIES ANALYSIS AND INPUTS OF ALGORITHMS FOR AIRPORT SIMULATIONS

Štefan BEREŽNÝ
Department of Mathematics and Theoretical Informatics, Faculty of Electrical Engineering and Informatics, Technical University of Košice, Němcovej 32, 042 00 Košice, Slovak Republic, tel.: +421 55 602 2447, e-mail: stefan.berezny@tuke.sk

ABSTRACT

This article focuses on the application of given data of flight delays at the airport Košice and its adaptation for further processing. These data were recorded from 2006 till 2009. The airport Košice did not do the data-processing of delays yet. Since the results of this process are useful in the planning or scheduling, we try to establish a methodology for analysing these data. The values of basic statistical parameters for different airlines and different types of flights are shown. We publish their short analysis and commentary. We try to show the problem of prediction of development of these delays at the airport using statistical test.

Keywords: *data processing, statistical hypothesis testing, data analysis*

1. INTRODUCTION

This article provides the processed data from the Košice airport (KEA) for the years 2006–2009. We show processed data representing share of different types of flights and airlines on the total amount of movement on the Košice airport. The primary objective was analysis of the time deviation of arrivals and departures of flights with respect to the scheduled value. This data can help us to detect approaching problems. Based on these results, the scheduling problem for arrivals and departures of flights at the airport can be better solved. In the section where we analyse data of airlines, we use replacement names AIR-01, AIR-02, AIR-03, and AIR-04. The choice of these four airlines is justified by their share on total traffic at the Košice airport (see Table 1). At the end of the article the results of statistical tests that compare the temporal deviations for the years 2006–2009 are presented. These tests tell us, that the prediction based on the given data can be very difficult, even impossible. Despite these negative results, the data were analysed using time series (see articles [4, 5]). To confirm some of the conclusions it would be necessary to conduct a detailed analysis of the problem. Unfortunately, for this we have no space in this article.

2. WHAT DATA WE HAVE AVAILABLE?

The data which is processed in this article has been obtained from the Košice airport. These data cover the period of 48 months from January 2006 until December 2009. We received a rather large amount of data that has undergone initial processing. We have an average of 10 000 rows of data each year, that is about 40 000 lines of data. Each line contains at least 20 data on the performance of an aircraft's arrival or departure. The initial data processing and formal treatment of data for next analysis was done and partially published in articles [1–3, 9–11]. In these works, the process of preparation of data has been described, which leads to obtaining the primary indicators of delays at the Košice airport. Also the basic parameters of realized flights at the Košice airport have been published. The initial analysis forms the basis of our research in this area. We based this research on already processed and provided basic data and also on already published basic parameters of these large data sets.

Based on this information, we had to decide which data will be analysed further and will need to undergo further processing and subsequent analysis. The basic data set provided by the airport also contained data that was incorrect, ambiguous, or required additional modification. All such modifications were made using filtering and sorting. An example of the ambiguity in the data was the value expressing the actual arrival time and departure time of the aircraft at the airport. The air plane of airline AIR-01 has landed with respect to the flight schedule at the airport on 12. 12. 2008 at 23:35, but in fact the flight was delayed 45 minutes and air plane landed at 00:20. In the database of flights is just the time of arrival 00:20. In fact this time belongs already to the day 13. 12. 2008 and not 12. 12. 2008. From these data, it is not clear whether the flight was 45 minutes late or the arrival was 23 hours and 15 minutes sooner than scheduled. This article will focus only on the data collection, processing, treatment, and analysis using basic statistical tests of hypotheses. This analysis reveals more about these data and because of their treatment partial comparison between the years from 2006 to 2009 will be possible.

Our work is built on the articles [9], [10], and [11], describing the basic characteristics and parameters of the recorded data. In these, process of modification of the data and its further processing is described. Given the extensive information, this article is limited only to those parameters that are immediately needed for our analysis.

Similarly, we take basic information from [1–3] where are already published the results of processing, referred to work [9]. These data will serve as input for our analysis and we can also use them for comparison of our obtained results.

At the end of this article we will try to suggest the direction of further work in this field. The scope for research in this field is quite large, because it is possible to examine whether the economic crisis has an impact on the development of parameters and how to effectively predict the trends based on the actual trends of monitored parameters (see [7]). An interesting analysis of these data using time series is presented in the work [6] and articles [4, 5]. The

results of this work are applicable for deciding management of airport for planning of flights and activities at the airport. The data obtained can be used as inputs for the queuing theory or scheduling.

3. USED METHODS

During processing of the data we used the software: MATLAB 2009b, MS Excel 2003, MS Excel 2010, and QtOctave, respectively. MATLAB and QtOctave were used mainly for analysis and testing of processed data and MS Excel for the processing and selection of appropriate data for processing and analysis. Also results published in this article, are consistent by format with these program packages.

Firstly, we try to describe the basic methods that were used in data processing in the next selection. Next, we describe the use of statistical hypothesis testing to assess the quality of data for prediction.

3.1. Data processing and selection

As the basic data set we took the edited file mentioned in [1–3,9–11]. This data has already undergone initial treatment and was checked for the dumb data or data that would lead to errors.

For processing and selection of appropriate data we used filters. Based on these initial filters, we found anomalies in recorded data and we were trying to identify the reason of them. If it was possible to correct them, then they were corrected immediately. If not, the anomalies were marked and later adapted so that they were usable in the calculations.

Similarly, we looked at the data that we showed in different forms in the PivotTable. These tables reveal us what is the structure of records and also show suitable candidates for further processing and analysis. We also used these tables as specific filter, or we used them to detect the frequency of disagreement in each class, we were looking for the causes of these disagreements and we tried to correct them. It could be that we could not fix the recorded value, in which case we deleted it from the list and the calculations were done without the deleted record. There was only minimum of such cases with respect the amount of the records.

3.2. Testing and data analysis

In this section we take as the population already edited file using the methods described above. This file is significantly smaller and more specific for further analysis and testing.

For analysis and testing, the data was downloaded from MS Excel worksheet to the program MATLAB (or QtOctave), and we used statistical functions `ttest` and `vartest2` of MATLAB Statistics toolbox for testing and validation of parameters of retrieved data.

4. BASIC CHARACTERISTICS OF PROCESSED DATA

Data were restricted to airlines listed in articles [1–3, 10, 11]. We used the labels AIR-01 up to AIR-04. Similarly, we marked the types of flights: scheduled domestic flights (SDF), scheduled international flights (SIF), non-scheduled domestic flights (NSDF), non-scheduled international flights (NSIF).

We can see the evolution of the means of the deviations for departures of the airline AIR-02 in the years 2006–2009 in the Fig. 1. The real situation for flight deviations we can see on the Figs. 2 (Departures) and 4 (Arrivals).

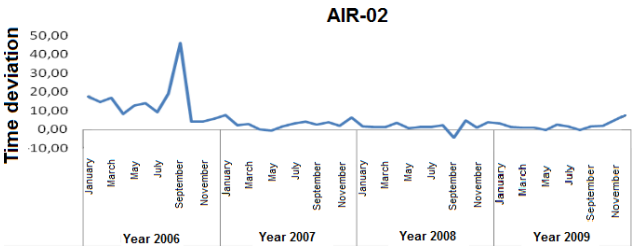

Fig. 1 Mean values of deviations of departures for AIR-02 in the years 2006–2009

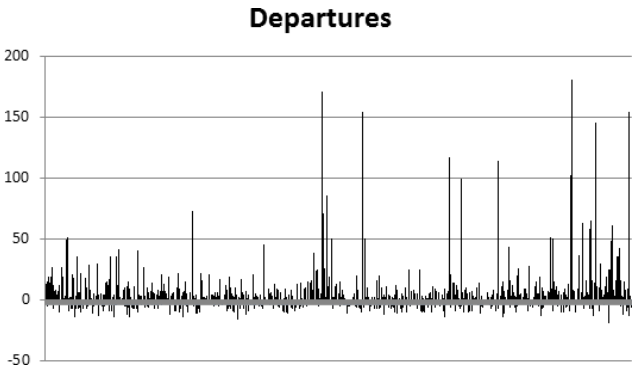

Fig. 2 Values of deviations (in minutes) of departures for AIR-02 in the year 2009

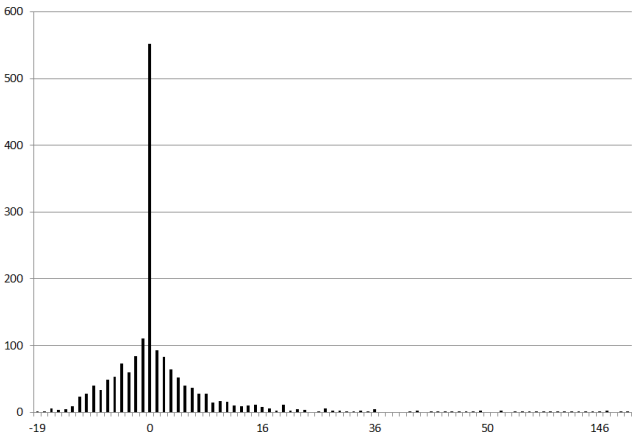

Fig. 3 Deviations frequencies of departures for AIR-02 in the year 2009

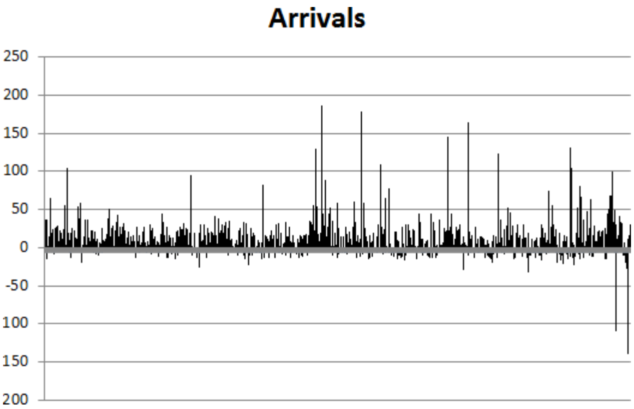

Fig. 4 Values of deviations (in minutes) of arrivals for AIR-02 in the year 2009

For a better understanding of the actual structure of the files we created Tables 1 and 2. These tables show the share of selected companies on the overall movements of the Košice airport and their mutual share in arrivals and departures during the years 2006–2009. In 2006 they made 7 166 arrivals and departures, in 2007 9 329 arrivals and departures, in 2008 10 340 arrivals and departures, and in 2009 9 636 arrivals and departures. The first column in each year is the relative percentage within the groups AIR-01 up to AIR-04 (together 100%). The second column in each year represents a percentage of movements of the above airlines with respect to all movements at the airport in given year. The values in the last row (under the table) show why we have chosen these four airlines. They have conducted majority of movements at the airport.

Basic characteristics are related to deviations from the planned arrival and departure times, respectively. Later departure or arrival time (called delay) is represented by positive values and the earlier arrival or departure is represented by negative values. The ideal situation is zero expected value with minimal dispersion. Actual mean values in minutes and corresponding variances are shown in Tables 3 and 4. Table 3 describes the situation for the arrivals and Table 4 for departures.

In Fig. 3 is shown graph of frequencies for each occurrence time deviations (departures) of airline Air-02 in the year 2009. It is a nice example of the structure of data and it confirms that the expected mean value is zero. It can also be seen that most of the values are within the range from −10 to 16 minutes. Special value of zero represents nearly a 50 percent share of all deviations recorded for airline AIR-02 for departures during the year 2009. The nice shape of histogram is confirmed by the numerical values of statistical parameters, and this are the mean value and variance, which are listed in the Table 4.

In tables we can observe the temporal evolution of the average deviations for arrivals and departures at the Košice airport from 2006 until 2009. These are values for years preceding the global economic crisis. Interesting are figures for regular and irregular lines and comparison of them with the total number of lines at the Košice airport. Table 3 shows only three negative values in the third row in 2006 and 2008 and first row in 2009. This indicates that most airlines were landing (or departing) on time, and if not, they produced rather delay on arrival (or departure) than they arrived (departed) earlier. We can notice that between 2006 and 2007 there was a significant decrease in the variation and the mean value of delay, and this trend continued in following years. We can notice that between 2006 and 2007 there was a significant decrease in the dispersion and the delay, and this trend continued in following years. This means that since 2007 has significantly improved discipline of airlines in complying with flight plan.

In Table 4, showing the values for the departure, is not even a negative value except for the first row in year 2009, which corresponds to the fact that the aircraft would not leave earlier than planned to have a plane in flight. It is an interesting development of variations in the average time each year, together with the evolution of variances. A more detailed analysis of individual rows of table requires much more space, so it is not shown here. Nevertheless, it is easy to see that the values in the table of departures are better than values in the table of arrivals (variance values are significantly lower). It means that compliance with the schedule of departures for the airlines is simpler and more easily achievable.

Even more interesting are the values of those airlines that are divided into the national and the international and also scheduled and non-scheduled. The results are shown in Tables 5 and 6. From these tables we can see quite a big difference between scheduled and non-scheduled flights. Similarly, we observed differences between the arrivals and departures. The question is how to interpret these differences, and where to find the causes of these deviations. Scheduled flights are significantly different from non-scheduled lines because these flights are mostly charter flights during the holiday period. They often do not comply with the flight plan. Also there are private flights whose complying with the flight plan is more formal than real.

Table 1 Share of airlines AIR-01 to AIR-04 on mutual and total air traffic at Košice airport

	2006		2007		2008		2009	
AIR-01	22.26%	14.88%	16.37%	12.23%	12.62%	9,44%	15.32%	9.17%
AIR-02	45.21%	30.21%	48.81%	36.48%	41.14%	30.76%	60.43%	36.16%
AIR-03	31.74%	21.21%	28.08%	20.99%	41.02%	30.68%	23.91%	14.31%
AIR-04	0.79%	0.53%	6.74%	5.04%	5.21%	3.90%	0.35%	0.21%
		66.83%		74.73%		74.78%		59.84%

Table 2 The share of different types of flights and the total air traffic at the Košice airport

	2006	2007	2008	2009
	share in %	share in %	share in %	share in %
scheduled domestic flights	8.97%	7.84%	5.37%	31.00%
scheduled international flights	20.00%	27.47%	26.47%	40.00%
non-scheduled domestic flights	27.46%	22.79%	24.80%	8.00%
non-scheduled international flights	43.57%	41.90%	43.37%	21.00%

Table 3 Basic characteristics for time deviation of arrivals at Košice airport (in minutes)

	2006		2007		2008		2009	
	mean value	variance	mean value	variance	mean value	variance	mean value	variance
AIR-01	2.75	7656.13	3.95	843.69	0.49	1892.62	−4.27	253.36
AIR-02	12.92	6521.10	3.77	333.23	5.07	1133.18	7.05	1496.69
AIR-03	−5.36	28557.22	10.51	1880.00	−1.96	1140.61	24.40	3685.25
AIR-04	71.16	74034.87	16.17	1678.28	27.06	2784.36	9.70	717.61
scheduled f.	3.45	13753.95	5.84	891.48	2.27	891.32	8.77	922.13
non-scheduled f.	35.99	18858.89	13.21	1811.21	15.81	1232.23	17.72	1485.27
Košice Airport	15.34	15864.97	8.10	1185.18	6.36	1032.84	11.91	2882.35

Table 4 Basic characteristics for time deviation of departures at Košice airport (in minutes)

	2006		2007		2008		2009	
	mean value	variance	mean value	variance	mean value	variance	mean value	variance
AIR-01	5.47	159.28	4.47	162.35	3.81	334.87	−4.57	4780.54
AIR-02	16.96	4252.16	2.95	247.21	1.30	811.16	2.11	177.28
AIR-03	17.58	1651.09	16.06	1397.24	9.38	1777.21	26.78	6579.45
AIR-04	26.05	984.47	17.57	2056.78	38.09	3938.90	18.60	1007.84
scheduled f.	14.21	2491.02	7.19	614.78	4.88	1179.89	7.54	708.09
non-scheduled. f.	26.59	10524.71	15.35	1851.24	18.77	1666.23	15.96	1837.41
Košice Airport	18.71	5445.18	9.69	1006.94	9.07	1367.16	10.47	2402.11

Table 5 Basic characteristics for time deviation of arrivals at Košice airport (in minutes)

	2006		2007		2008		2009	
	mean value	variance	mean value	variance	mean value	variance	mean value	variance
SDF	−5.09	31340.53	7.85	1407.12	−4.29	804.99	9.36	675.22
SIF	7.33	5710.85	4.52	549.66	6.28	901.69	8.18	886.44
NSDF	86.23	51060.28	10.75	1618.40	10.12	645.08	13.04	1456.68
NSIF	19.49	7175.79	14.08	1876.69	17.00	1346.74	22.39	2038.64

Table 6 Basic characteristics for time deviation of departures at Košice airport (in minutes)

	2006		2007		2008		2009	
	mean value	variance	mean value	variance	mean value	variance	mean value	variance
SDF	25.38	6926.48	10.38	1046.62	3.29	1352.17	11.62	887.61
SIF	9.03	350.91	5.09	320.01	5.84	1072.21	3.45	678.45
N-SDF	70.82	34231.75	11.25	1067.83	16.72	1375.21	9.29	1466.58
N-SIF	12.23	1981.77	16.72	2104.65	19.23	1730.27	22.62	1964.21

5. PROCESSED DATA TESTING

This section describes how to check behaviour of different sets of data files for different analysed years. As first, it is necessary to test files whether they meet the conditions for using statistical tests of hypotheses and then describe the tests and their results. Given the limited space of this article, we shortly describe the verification of individual results and focus mainly on the description and interpretation of the tests carried out for individual files.

From all possible tests we carried out one sample test for mean values and variances. Then we conducted series of two-sample t-tests of the mean value. At the end we did the multi-dimensional test of equality of mean values. All these tests of statistical hypotheses were conducted on a 5% significance level (i.e., $\alpha = 0.05$). All results are presented in overview tables.

Table 7 shows the results of the two-sample tests for mean values between sets of time deviation of arrivals and departures for each year. We can see that all of the tests have given a negative answer on question, whether we can at a significance level of 0.05 consider average time deviation of arrivals and departures to be comparable across years 2006–2009. Due to fluctuations in the mean values in each year and due to the results of tests, it can be assumed that the prediction of time variations will be very difficult, or even impossible.

Table 7 Two-sample tests of equality of mean value of time deviation of sets of arrivals and departures

	2006	2007	2008	2009
2006	–	No	No	No
2007	No	–	No	No
2008	No	No	–	No
2009	No	No	No	–

Table 8 shows the results of two-sample tests for mean values between sets of time deviation of arrivals for each year. In this case we have the same results as in Table 7. For the same reason you can not expect that prediction of temporal deviation of arrivals will be reliable.

Table 8 Two-sample tests of equality of mean value of time deviation of sets of arrivals

	2006	2007	2008	2009
2006	–	No	No	No
2007	No	–	No	No
2008	No	No	–	No
2009	No	No	No	–

Table 9 Two-sample tests of equality of mean for time deviation of sets of departures

	2006	2007	2008	2009
2006	–	No	No	No
2007	No	–	Yes	No
2008	No	Yes	–	No
2009	No	No	No	–

Comparison between 2008 and 2007 became negative, but on significance level of 1% would have been equality. Table 9 shows the results of the two-sample tests of equality of mean for time deviation of sets of departures for all years.

In this case, it would be appropriate to carry out more tests in addition to scheduled flights and non-scheduled flights. It should be also tested separately for selected four airlines AIR-01 up to AIR-04. Given the values in Tables 5 and 6, it can be estimated, where equality occurs and where equality does not occur. Due to limited space we will not publish the complete results of the tests of time deviations of mean values for each of the selected data sets. Even so narrowed result we can say that in 2007 and 2008 flights departed from Košice International Airport with almost the same delay. Similarly, we might note that in those years flights landed with approximately the same delay. Other cases have varied significantly, which is essential information for exploring the causes of these disagreements.

We can conduct also one-sample tests with regard to the expected mean value (zero). Based on the results, we can create several groups of airlines. These would be arranged with respect to the distance from zero. It would also create some overview of who is trying to follow schedule of flights and who is not.

Table 10 Two-sample tests (F-test) of equality of variations of time deviation of sets of arrivals and departures

	2006	2007	2008	2009
2006	–	No	No	No
2007	No	–	No	No
2008	No	No	–	No
2009	No	No	No	–

Comparison of variances is shown in the Table 10. The results are even more negative than in the case of mean values. Also, this test confirms the low predictability of this parameter for the future.

6. CONCLUSIONS

From these results, we have an idea how the temporal variation was proceeding at the Košice airport from 2006 till 2009. These tests have confirmed that for most of these cases the equality of the average delays was not confirmed. It is true that we have done detailed analysis, but general tests create space for next analysis of this issue. It would be appropriate to examine separately the behaviour of the time deviations of individual flights. Then it would be appropriate to analyse these results.

These results can be used in the analysis described in [4], [5], and [12]. Results from articles [4] and [5] confirm our conclusions that the prediction of the time deviations in these cases is very difficult, despite the fact that they used time series analysis. Similarly, we could divide these tests for scheduled and non-scheduled flights between domestic and international. With this finer partitioning results could be more specific.

This analysis is essential for development and testing of time deviations of flights at the airport and analysis by time series which can be represented by months, weeks or days of the week. Such data processing and results are ideal for time analysis, which is described in [6], [7], and [8]. Time series analysis should be the target of all of these partial results of research described in this article well as in the aforementioned articles. The results of analysis of the time deviations are very important source of information for the management of Košice International Airport. This model of analysis is of course also applicable to other airports and the results could be compared between different airports.

Even more interesting is the analysis of extended data starting from the global economic crisis in 2009. It is interesting to follow developments of the time deviations in 2009, compared with the years before the crisis (from 2006 to 2008) and to try to describe the reasons for any differences based on the results obtained (see [6, 7]). On the other hand, the time evolution of these deviations should theoretically not be affected by the crisis, but by weather or unforeseen event. We need to suggest that this conclusion is non-standard behaviour. More research in this area will be necessary to clarify the hidden patterns of temporal anomalies in arrivals and departures at the airport. This could be later effectively used in planning and management of the airport.

Further direction of research in this area can be therefore divided into two main lines. One line should address expanding the basic set of statistics for more years, or extend the set of data from other airports. The second line should analyse these data in more detail and depth.

ACKNOWLEDGEMENT

I want to thank to Mrs. RNDr. Zuzana Petrášová, Mrs. Ing. Marta Horváthová, and Mr. Ing. Ján Palko, for their willingness, time and information. Operational data for this research was provided by Košice International Airport.

REFERENCES

[1] BEREŽNÝ, Š. – ANDREJKOVIČ, M.: *Statistical processing of flights on Košice airport*, Acta Avionica, Vol. 11, No. 17 (2009), 9–14, ISSN 1335-9479.

[2] BEREŽNÝ, Š.: *Statistical processing of arrivals and departures on Košice airport in 2007*, Acta Avionica, Vol. 12, No. 19 (2010), 42–48, ISSN 1335-9479.

[3] BEREŽNÝ, Š.: *Statistical processing of arrivals and departures on Košice airport in 2006*, Acta Avionica, Vol. 12, No. 19 (2010), 35–41, ISSN 1335-9479.

[4] BEREŽNÝ, Š. – ANDREJIOVÁ, M. – BUŠA Jr., J.: *Analysis of Time Variations of Flights at Košice Airport in The Years 2006-2009 for Selected Airline Companies*, Acta Avionica, Vol. 13, No. 21 (2011), 104–110, ISSN 1335-9479.

[5] BEREŽNÝ, Š. – BUŠA Jr., J.: *Analysis of time variations of flights at Košice airport in the years 2006–2009*, Acta Avionica, Vol. 13, No. 22 (2011), 7–14, ISSN 1335-9479.

[6] BALÁŽOVÁ, A.: *Časová analýza meškaní letov na Košickom letisku v období 2006–2009*, [Temporal analysis of delays at the airport in Košice, between 2006–2009], Diploma thesis. Košice: Technical University of Košice, Faculty of Aeronautics, 2011, 74 pp.

[7] HRUŠKOVÁ, M.: *Analýza časových odchýlok príletov a odletov na Košickom letisku a vplyv hospodárskej krízy na ich vývoj v roku 2009*, [Analysis of time deviation, arrivals and departures at the airport Košice and the impact of economic crisis on their developement in 2009], Diploma thesis. Košice: Technical University of Košice, Faculty of Aeronautics, 2011, 61 pp.

[8] KIMÁKOVÁ, Z. – ANDREJIOVÁ, M.: *Analýza časového radu emisií sox pomocou programu R*, [SOX Emisions Time Series Analysis Using Program R], In: Forum statisticum slovacum, Nitra, Vol. 5, No. 6 (2009), 63–68, ISSN 1336-7420.

[9] LACZKÓ, R.: *Štatistické spracovanie meškaní letov na Košickom letisku*, [Statistical processing of flies delays at Košice airport], Bachelor thesis. Košice: Technical University of Košice, Faculty of Aeronautics, 2009, 48 pp.

[10] RONDOŠOVÁ, Ľ.: *Tvorba časových radov odchýlok letov na Košickom letisku pre vybrané letecké spoločnosti*, [Creation of time series deviations of flights at Kosice airport for selected airlines], Bachelor thesis. Košice: Technical University of Košice, Faculty of Aeronautics, 2011, 50 pp.

[11] RONDOŠ, R.: *Tvorba časových radov odchýlok letov na Košickom letisku vzhľadom na typ letu*, [Creation of time series deviations of flights at Kosice airport for given type of flight], Bachelor thesis. Košice: Technical University of Košice, Faculty of Aeronautics, 2011, 49 pp.

[12] TKÁČ, M. – ANDREJKOVIČ, M. – HAJDUOVÁ, Z.: *Analysis of Sky Europe airlines flights on Košice international airport*, Acta Avionica, Vol. 11, No. 17 (2009), 174–179, ISSN 1335-9479.

Received July 27, 2012, accepted September 28, 2012

BIOGRAPHY

Štefan BEREŽNÝ was born on 31. 10. 1974. In 1998 he graduated (MSc.) at the Faculty of Science at P. J. Šafárik University in Košice. He defended his Ph.D. in the field of Discrete Mathematics in 2005. He was working as assistant professor at the Department of Mathematics and Physics at the Air Force Academy of gen. M. R. Štefánik from 1998 till 2004. Since 2003 he is working as assistant professor at the Department of Mathematics and Theoretical Informatics at Faculty of Electrical Engineering and Informatics at the Technical University of Košice. His scientific research is focusing on Discrete Mathematics, Graph Theory, Discrete Optimization, Algorithms, and Complexity. In addition, he also investigates questions related to the Mathematical Statistics and Applied Mathematics.