BULLETIN OF THE POLISH ACADEMY OF SCIENCES
TECHNICAL SCIENCES, Vol. 70(2), 2022, Article number: e140694
DOI: 10.24425/bpasts.2022.140694

Arriving air traffic separations generalized model identification

Adrian PAWEŁEK* and **Piotr LICHOTA**

Institute of Aeronautics and Applied Mechanics, Warsaw University of Technology, 00-665 Warsaw, Poland

Abstract. Quick development of computer techniques and increasing computational power allow for building high-fidelity models of various complex objects and processes using historical data. One of the processes of this kind is an air traffic, and there is a growing need for traffic mathematical models as air traffic is increasing and becoming more complex to manage. This study concerned the modelling of a part of the arrival process. The first part of the research was air separation modelling by using continuous probability distributions. Fisher information matrix was used for the best fit selection. The second part of the research consisted of applying regression models that best match the parameters of representative distributions. Over a dozen airports were analyzed in the study and that allowed to build a generalized model for aircraft air separation in function of traffic intensity. Results showed that building a generalized model which comprises traffic from various airports is possible. Moreover, aircraft air separation can be expressed by easy to use mathematical functions. Models of this kind can be used for various applications, e.g.: air separation management between aircraft, airports arrival capacity management, and higher-level air traffic simulation or optimization tasks.

Key words: air traffic; system identification; maximum likelihood; modelling.

1. INTRODUCTION

According to EUROCONTROL's Performance Review Report [1] which assessed European's ATM performance in 2019 and concerned the operational en-route Air Navigation Services (ANS) performance in the EUROCONTROL area, en-route Air Traffic Flow Management (ATFM) delays in 2019 reached 17.2 million minutes causing that 2018 and 2019 were the years with the highest amount of en-route delays since 2010. It was reported that an average en-route delay above one minute per flight occurred on 206 days in 2019, which is equivalent to 56 % of the days. In 2019 en-route delays accounted for 9.9 % of all delayed flights. Air Traffic Control (ATC) capacity accounted for 43.9 %, which is equal to 17.2 million of total delay minutes; ATC staffing accounted for 24.3 %, which is equal to 4.2 million of total delay minutes; ATC disruptions or industrial actions accounted for 7.2 %, which is equal to 1.2 million of total delay minutes; weather accounted for 21.2 %, which is equal to 3.6 million of total delay minutes. Summary is presented in Table 1.

Table 1
En-route ATFM delays in 2019 by attributed delay category

Delay cause	Delayed flights	Flight delay in minutes	Total delay	
			minutes	share
ATC Capacity	5.2 %	13.3	7.6 M	43.9 %
ATC Staffings	2.4 %	16.2	4.2 M	24.3 %
ATC Disruptions	0.4 %	27.3	1.2 M	7.2 %
Weather	1.6 %	20.9	3.6 M	21.2 %
Other	0.4 %	15.0	0.6 M	3.5 %
Total	**9.9 %**	**15.8**	**17.2 M**	**100 %**

In [2] it was stressed that advances in computing and related techniques allow for creating revolutionary breakthroughs in many aspects of everyday life and will data drive many scientific disciplines. One of the scientific disciplines where significant advances are possible is air traffic management (ATM) research, where the application of data science and prediction techniques [3] to large sets of data allows for extracting information by processing sets of historical air traffic data and even making air traffic predictions in real-time, which has been not possible or hard to do in the past. This kind of approach was adopted in [4] where equivalent sound level mathematical model was obtained by processing measurement data using data science techniques.

Given the above factors, the idea of developing a generalized model for a part of the arrival process arose. Generalized model, which comprises traffic from various airports and allows to express aircraft air separation by a set of easy to use mathematical functions, could be useful in many ATM aspects, e.g.: air separation management between aircraft, airport arrival capacity management, higher-level air traffic simulation and optimization tasks, thus, can contribute to reduced ATC-caused delays. Until recently, the creation of ATM models was not an important element of research, as previously noted, due to the lack of motivation in the form of easily manageable air traf-

*e-mail: apawelek@meil.pw.edu.pl

Manuscript submitted 2021-09-06, revised 2021-11-13, initially accepted for publication 2021-12-24, published in April 2022.

fic and computational constraints. However, in recent years, the need for those models started to become higher, e.g., due to ATC Model-Driven Approach proposals [5]. Moreover, studies in which such models based on real air traffic data have begun to appear. In [6] dynamic models to describe some behaviors including aircraft following, holding, and maneuvering were established. In [7] density-speed-based modified cell transmission model was created to simulate the spatio-temporal evolution of traffic flow and airspace congestion in the arrival network. In addition, a number of theoretical approaches to model, assess, and optimize air traffic were described in [8] and [9]. Combination of mathematical modelling, optimization, and real air traffic data can bring significant advances to ATM. In [10] model of airport process and optimization was used for the purpose of efficiency of traffic management processes on the apron evaluation. In [11] mathematical modelling and optimization was used for a purpose of planning and management of aircraft maintenance using a genetic algorithm. Due to many aspects which need to be considered in ATM research and modelling, every research focuses only on selected aspects of air traffic modelling and only collective research efforts may allow to cover a sufficient range of models to allow effective airspace traffic modelling. What impacts ATM research activities is an access to historical air traffic data. Some researchers are using small subsets of real air traffic data or artificial data. Choosing a good source of data and data processing is an important step in the research, but it requires to solve a number of additional issues and requires additional resources [12].

This study aims to create a generalized aircraft air separation model by using continuous probability distributions and nonlinear regression fitting techniques. Models of this kind can be used to generate data samples, compare air traffic data, and validate similar models. Modelling of aircraft air separation by using continuous probability distributions was performed in [13] for a single airport and found efficient. Moreover, in the mentioned case, only a small number of sample days was analyzed. In this study, a variety of airports and a significant amount of data were taken into account, largely expanding previous outcomes. Moreover, by using nonlinear regression, a generalized aircraft air separation model was created using the output data from distribution fitting, which is the main novelty of the following paper.

The structure of the paper is as follows. The paper starts with a brief introduction, which is followed by the methodology overview in Section 2 and air traffic data description in Section 3. The mathematical principles underlying the aircraft air separation identification for a single airport and a selected day are shown in Section 4. The extension of those results to the full data set for various airports (generalized model) are shown in Section 5. The paper finishes with a short summary of the conclusions presented in Secction 6.

2. METHODOLOGY

The aim of this research was to derive a generalized arriving aircraft air separation model in a form of probability density function (PDF), which describes the distribution of air separation between aircraft in a function of arrival number $p(t_s,x)$, where t_s is time separation and x is the number of arrivals. A high-level overview of the whole air separation model identification process developed within this study is presented in Fig. 1.

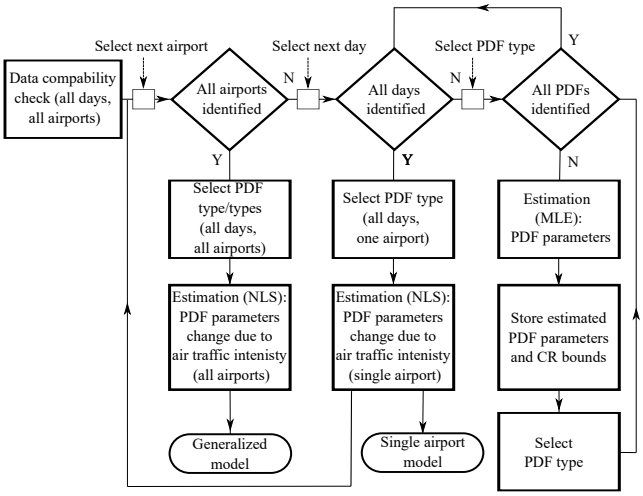

Fig. 1. Overview of the air separations model identification process

Input to the identification process is the separations data for each day from a set of airports. Then for each airport, separations data for each day or arrivals is being fitted to a set of probability density functions using maximum likelihood estimation (MLE). Estimated parameters and correctness of fit values are stored. Once all PDFs for all days and all airports have been identified, PDF type with the best correctness of fit is selected and nonlinear regression (NLR) is performed to obtain the regression of the estimated PDF parameters values in function of air traffic intensity (number of arrivals). Performing NLR for the parameters obtained for a single airport results in a single airport model. Finally, grouping parameters from all airports for the selected PDF and performing NLR results in the identification of the generalized model, which is the goal of this study.

3. AIR TRAFFIC DATA

Data demand repository 2 (DDR2) [14] database was used as a source of the historical air traffic data. DDR2 data has a number of advantages: the data is provided by EUROCONTROL, so it is trustworthy and complete; data is available in easy to store and process AIRAC datasets and can be extracted and partially preprocessed by EUROCONTROL's software *Network Strategy Tool* (NEST) software which also gives confidence to obtained datasets; data contains several fields which allow to easily chose desired airports, waypoints or other data of interest what minimizes introduction of error. NEST was used to convert data from compressed binary *AIRAC* files to separate *so6* text files. M3 trajectories data set was used, which corresponds to the last filled flight plan updated with radar data. For each day, it resulted in a file which contained trajectory data in the form of flight segments for each flight which occurred during that day. Each flight segment entry consisted of the data presented in Table 2.

Table 2
so6 file flight segment data entry structure

Field	Type	Size	Comment
segment identifier	char	8–10	<start> _ <end>
origin of flight	char	4	ICAO code
destination of flight	char	4	ICAO code
aircraft type	char	4	ICAO code
time at segment start	num	6	HHMMSS
time at segment end	num	6	HHMMSS
flight level at segment start	num	1–3	–
flight level at segment end	num	1–3	–
status	char	1	climb/descent/cruise
callsign	char	var	–
date at segment start	num	6	YYMMDD
date at segment end	num	6	YYMMDD
latitude at segment start	float	var	in minute decimals
longitude at segment start	float	var	in minute decimals
latitude at segment end	float	var	in minute decimals
longitude at segment end	float	var	in minute decimals
flight identifier	num	9	unique for every flight
sequence	var		–
segment length	float	var	in nautical miles
segment parity	num	1	–

In this study, several hundred days of arrival to each of the airports presented in Table 3 were selected, what results in over a million of flights. Data was collected from before the SARS-CoV-2 coronavirus pandemic. The airports were selected in a way that they cover a variety of runway configurations, i.e., one, two, or three runways and parallel or intersecting runways. They also cover the complete range of arriving traffic intensities, from low through medium to high traffic cases. The majority of the selected airports belong to the list of thirty busiest airports in Europe. For each case, the necessary airport data was obtained from appropriate Aeronautical Information Publication (AIP) providers, depending on the airport's controlling authority. This data included:

- Runway detailed layout and possible landing directions;
- Intermediate fixes (IFs) names and thresholds in nautical miles for every runway direction (IF is a waypoint where usually instrument landing system procedure begins) and are usually located around 10–12 NM from the runway threshold;
- Standard arrival procedures and possible airport landing configurations.

Table 3
Airports used in the study

ICAO Code	Airport
EBBR	Brussels Airport
EGCC	Manchester Airport
EGSS	London Stansted Airport
EKCH	Copenhagen Airport, Kastrup
ENGM	Oslo Airport
EPWA	Warsaw Chopin Airport
LEBL	Barcelona–El Prat Airport
LEPA	Palma de Mallorca Airport
LFPO	Paris Orly Airport
LIRF	Rome–Fiumicino International Airport
LSZH	Zurich Airport

A number of data pre-processing steps were performed to ensure data completeness and correctness as well as a format adequate for further processing. M3 trajectories are updated with radar data if the radar data deviates from the last filled flight plan by any of the thresholds: 5 min, 7 FL or 20 NM. Analysis was performed, which has shown that radar data was present in nearly all cases, which is the real air traffic data. The data was filtered, so only arrivals to selected airports remained. Flight segments were replaced by the trajectory waypoints by merging the segment end point with the next segment start point. The check for duplicated entries was made as in some cases a day with different start and land dates appeared in the files for both days. Subsequently, arrivals were sorted by a criterion such that a given flight belongs to a given day if the arrival date equals this day. Afterwards, to avoid a significant number of outliers, arrivals occurring during night hours were removed, but only for periods where the arrival density was smaller than seven flights per hour. M3 data is provided with insufficient resolution. However, an approximation method has been used to obtain intermediate points between trajectory points in the area of interest. Especially for the case considered within the study, the aircraft should have constant speed, a linear approximation could be applied, what does not impact the accuracy significantly. Finally, separations between arriving aircraft were determined for each day for each airport. Currently distance-based separations are used in arrivals operations. However, many novel traffic synchronization concepts, e.g. [15–17], recognize time-based separations and 4D trajectories as future means of arrival trajectories management, thus, time separations will be considered in this study. Time separation t_s between two consecutive arriving aircrafts belonging to set of arrivals $\mathcal{A}$ at given fix in the airspace was defined as:

$$t_{s_X}^{i,j} = T_X^j - T_X^i, \tag{1}$$

where:

- $i \in \mathcal{A}$ and $i = 1, \ldots, n_A - 1$ is a preceding aircraft index;
- $j \in \mathcal{A}$ and $j = i + 1$ is a following aircraft index;
- n_A is the number of arriving aircraft in given day, i.e. cardinality of set $\mathcal{A}$: $n_A = |\mathcal{A}|$;
- T is a time of a day of arrival at given airspace fix expressed in elapsed seconds from midnight;
- X is selected fix in the airspace for which time separations are calculated.

In the study, longitudinal separations were considered and the intermediate fix was selected as a waypoint for which separations are calculated. For each day included in the study, a separation for each pair of preceding and following aircraft (i, j) was calculated, what resulted in a vectors of length $n_A - 1$ consisting of the time separation values $\mathbf{t_{sx}} = \left[t_{sX}^{1,2}, \ldots, t_{sX}^{n_A-1,n_A}\right]^T$.

4. PROBABILITY DISTRIBUTION IDENTIFICATION

4.1. Maximum likelihood estimation

Fitting probability distributions to separations vector $\mathbf{t_s}$ is a task of unknown parameters identification. Separations vector $\mathbf{t_s}$ can be denoted as observed values vector $\mathbf{z}$. Assuming that aircraft separations have a distribution $p(\mathbf{z}|\boldsymbol{\Theta})$, where $\boldsymbol{\Theta}$ is parameters vector, obtaining optimal estimates requires the maximization of the conditional probability density function [18]:

$$\hat{\boldsymbol{\Theta}} = \operatorname{argmax}(p(\mathbf{z}|\boldsymbol{\Theta})). \tag{2}$$

In practice, maximum likelihood estimation (MLE) [19] is an efficient method for identification problems [20] and maximization task can be reformulated to the minimization of a negative log-likelihood function $L(\boldsymbol{\Theta}|\mathbf{z})$:

$$\hat{\boldsymbol{\Theta}} = \operatorname{argmin}(-\ln L(\boldsymbol{\Theta}|\mathbf{z})). \tag{3}$$

Multidimensional normal distribution is defined as:

$$p(\mathbf{z}_1, \ldots, \mathbf{z}_N) = \frac{1}{\left(\sqrt{(2\pi)^n}\sqrt{|\mathbf{R}|}\right)^N} * \exp\left(-\frac{1}{2}\sum_{k=1}^{N}[\mathbf{z}_k - \mathbf{y}_k]^T \mathbf{R}^{-1}[\mathbf{z}_k - \mathbf{y}_k]\right), \tag{4}$$

where:

- n is a number of dimensions;
- N is a number of observed vectors (measurements);
- $\mathbf{y}$ is a model output vector.

The error covariance matrix $\boldsymbol{R}$ can be estimated as:

$$\boldsymbol{R} = \frac{1}{N}\sum_{k=1}^{N}[\mathbf{z}_k - \mathbf{y}_k]\,[\mathbf{z}_k - \mathbf{y}_k]^T; \tag{5}$$

where $\mathbf{z}$ has length $n_Z = n_A - 1$. After assuming residuals independence, the following cost function is obtained:

$$J(\boldsymbol{\Theta}) = \sum_{k=1}^{n_Z}(z_k - y_k)^2, \tag{6}$$

where y_k is model output value corresponding to observed value z_k.

Fitting probability distributions using MLE was performed as described in [13] where several dozens of various continuous probability distributions were used to find the best fit distribution, and for each of them probability density function parameters were estimated.

4.2. Relative standard errors

Relative standard error (RSE) [21] smallest mean value was used as a criterion for selecting the distribution type which fits best to the arrival traffic data most often. One of the advantages of this criterion is being resistant to the number of distribution parameters: e.g., the sum of squared errors criteria can be advantageous for distributions with larger number of parameters as they usually better fit the data (due to greater flexibility), but sometimes it may lead to overfitting and not be a good candidate for generalized model. RSE were derived from observed Fisher information matrix. Observed Fisher information matrix [18] was evaluated using maximum likelihood estimates ($\boldsymbol{\Theta} = \hat{\boldsymbol{\Theta}}$):

$$\boldsymbol{F} = \frac{\partial^2 J}{\partial \boldsymbol{\Theta}^2} \approx \sum_{k=1}^{n_Z}\left[\frac{\partial y_k}{\partial \boldsymbol{\Theta}}\right]^T \boldsymbol{R}^{-1}\left[\frac{\partial y_k}{\partial \boldsymbol{\Theta}}\right], \tag{7}$$

where:

- J is a cost function;
- n_Z is a number of measurements;
- y is a model output;
- $\boldsymbol{R}$ is estimated as below:

$$\boldsymbol{R} = \frac{1}{n_Z}\sum_{k=1}^{n_Z}[z_k - y_k]\,[z_k - y_k]^T. \tag{8}$$

Model output partial derivatives can be calculated using the forward difference scheme:

$$\left[\frac{\partial y_k}{\partial \boldsymbol{\Theta}}\right]_j \approx \frac{y_k(\boldsymbol{\Theta} + \Delta\Theta_j \boldsymbol{e}^j) - y_k}{\Delta\Theta_j}, \tag{9}$$

where:

- $j = 1, \ldots, n_\Theta$ and n_Θ is a number of estimated parameters;
- $\boldsymbol{e}^j$ is a column vector with 1 in the j$^{\text{th}}$ row and 0 in remaining rows;
- $\Delta\Theta_j$ is perturbation in j$^{\text{th}}$ component of $\boldsymbol{\Theta}$ and $\Delta\Theta_j = 10^{-4}\Theta_j$ in this study.

After obtaining observed Fisher information matrix, it can be inverted to obtain the asymptotic covariance matrix estimator $\sigma^2(\hat{\boldsymbol{\Theta}}_{ML})$:

$$\sigma^2(\hat{\boldsymbol{\Theta}}_{ML}) = \left[\boldsymbol{F}(\hat{\boldsymbol{\Theta}}_{ML})\right]^{-1}. \tag{10}$$

Relative standard errors are finally calculated:

$$\boldsymbol{\varepsilon}(\hat{\boldsymbol{\Theta}}_{ML}) = \sqrt{diag\left(\sigma^2(\hat{\boldsymbol{\Theta}}_{ML})\right)} \oslash \boldsymbol{\Theta} * 100\%, \tag{11}$$

where $\oslash$ means for element-wise division.

4.3. Results

After solving MLE problems for each day and applying RSE criteria to obtained results, *Cauchy* distribution turned out to be the best fit distribution, constituting the best fit for 77–90% of days for each airport being the subject of this study.

For *Cauchy* distribution $\boldsymbol{\Theta} = [x_0, \gamma]$ and the probability density function (PDF) is defined as follows [22]:

$$p(x, x_0, \gamma) = \frac{1}{\pi\gamma\left(1 + \frac{(x - x_0)^2}{\gamma^2}\right)}, \tag{12}$$

where:

- x_0 is the location parameter and specifies peak location, where maximum value of probability density function is located;
- γ is the scale parameter and specifies half-width at half-maximum, that determines the slope inclination of the probability density function shape.

That characteristic makes *Cauchy* distribution even more suitable for expressing aircraft air separation distribution, as it contains a high peak and a long tail that is typical for separation distribution.

It should be noted that Cauchy distribution does not have a theoretical definition of mathematical expectation (also known as expected value). Mathematical expectation defines a central (or average) value. For separations distributions it is not a good metric, due to the usually high peak and long tail of the distribution. Instead, Cauchy's distribution location and scale parameters are good metrics, as they determine the maximum value and slope inclination, respectively. Mathematical expectation has certain applications, which are not the case in the research done within the article. Lack of definition of mathematical expectation for Cauchy distribution does not impact creating and using of the model.

Example results of fitting distribution to arrival separation data are presented in Fig. 2. Horizontal axis presents the time separation in seconds, the vertical axis presents the normalized number of aircraft, blue histogram bars (bin width was set to 40 s) present the distribution of real data, and the red line presents *Cauchy* distribution fit to the data. On this day 417 arrivals took place, and $\hat{\boldsymbol{\Theta}} = [113.13, 27.05]$. In Fig. 2 it can be also observed that some aircrafts have very small separation. This is usually related to the data measurement noise, data inaccuracies, and approximations described in Section 3. Additional

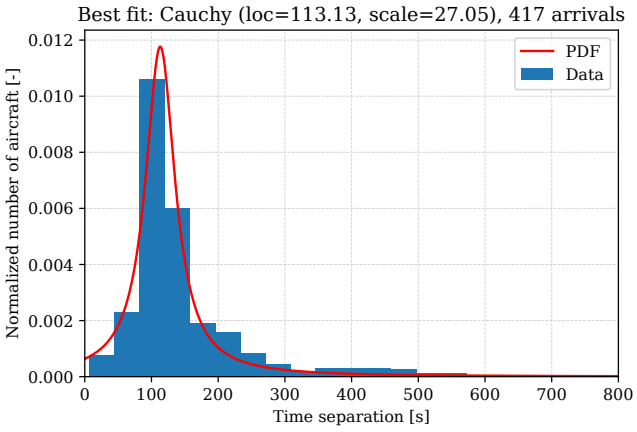

Fig. 2. Example results of fitting distribution to separations data, *Cauchy* distribution

factor is the histogram bin width set to 40 s, where the data is usually closer to 40 s than 0 s. Violation of ICAO separations is a rare case, however it may be present within the data as well.

For each airport, A results of fitting $\hat{\boldsymbol{\Theta}} = [x_0, \gamma]$ were grouped into observed distribution parameters values $\mathbf{z}^{\mathbf{A,M}}$ vectors ($M = \{1, 2\}$) which contained all n_P values of the given parameter for the given airport:

$$\mathbf{z}^{A,M} = \left[z_1^{A,M}, z_2^{A,M}, \ldots, z_{n_P}^{A,M}\right]. \tag{13}$$

Also vectors $\mathbf{x}^A$ can be defined:

$$\mathbf{x}^A = \left[x_1^A, x_2^A, \ldots, x_{n_P}^A\right]. \tag{14}$$

which contain the number of flights corresponding to each entry in $\mathbf{z}^{A,M}$ vectors.

In Fig. 3 and Fig. 4 the grouped probability density function parameters for LEPA airport are shown. Horizontal axis presents the number of arrivals in a given day and the vertical axis presents *location* and *scale* parameters values, respectively. It can be observed that for both parameters their values decrease with increasing number of arrivals, also the slope becomes smaller with the increasing number of arrivals. It means

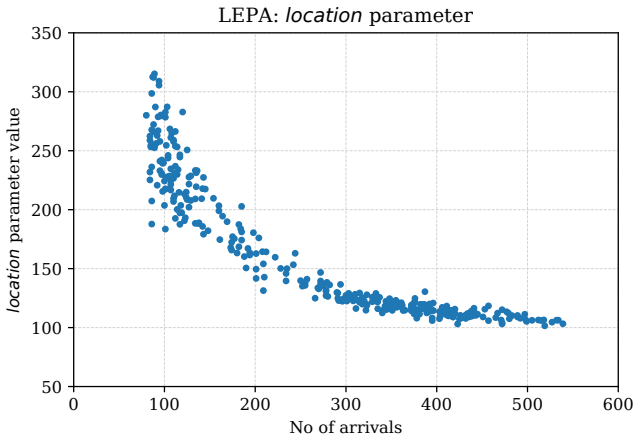

Fig. 3. *Location* parameter values for LEPA airport

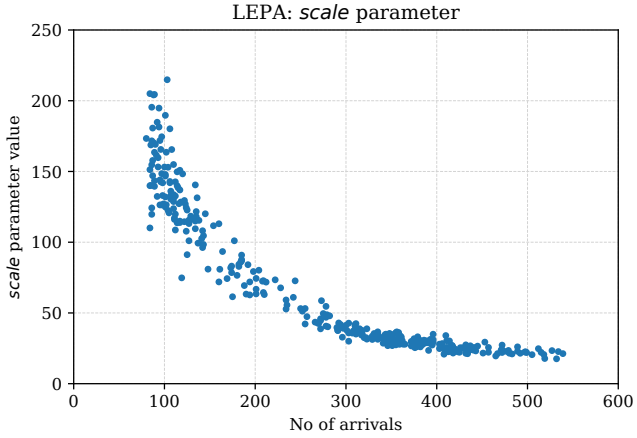

Fig. 4. *Scale* parameter values for LEPA airport

that *Cauchy*'s distribution peak moves left (towards 0 s) and becomes higher and tighter. Another observation is that in both cases the parameters values variation becomes smaller with increasing number of arrivals. This is a reasonable behaviour, as with increasing number of arriving aircraft the airspace becomes more constrained and there is less space for variation.

5. AIR TRAFFIC IDENTIFICATION

5.1. Nonlinear regression

Nonlinear regression was performed by using the nonlinear least squares (NLS) method. It allowed to fit $\boldsymbol{\beta}=[\beta_1,\beta_2,\ldots,\beta_{n_B}]$ coefficients of nonlinear functions which were chosen as candidates for aircraft separations model. Estimated functions coefficients $\hat{\boldsymbol{\beta}}$ were obtained by solving the following minimization problem:

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \operatorname{argmin} \sum_{k=1}^{n} \left[z_k - f\left(x_k, \boldsymbol{\beta}\right)\right], \\ \boldsymbol{\beta} &= \left(\beta_1, \beta_2, \ldots, \beta_n\right).\end{aligned} \tag{15}$$

Levenberg-Marquardt algorithm [23] was used to solve the optimization task.

In theory, infinite numbers of function linear combinations can be utilized for the task of finding the best match to data. However, in practice, a set of common simple functions is usually used, what makes the model convenient for further use. The following functions were considered as the candidate regression model $f\left(x, \boldsymbol{\beta}\right)$:

- Exponential: $\beta_1 e^{\beta_2 x}$;
- First order polynomial: $\beta_1 x + \beta_2$;
- Logarithmic: $\beta_1 \ln x + \beta_2$;
- Power: $\beta_1 x^{\beta_2}$;
- Second order polynomial: $\beta_1 x^2 + \beta_2 x + \beta_3$;
- Reciprocal: $\dfrac{\beta_1}{(x+\beta_2)} + \beta_3$.

Basing on the characteristics of the data, whose sample is presented in Fig. 3 and Fig. 4, first-order polynomial was rejected as the parameter value change becomes smaller with increasing arrival number, which means that the first derivative should monotonically increase. Moreover, the second-order polynomial was rejected due to the fact that the function changes monotonicity after reaching its minimum. Basing on the knowledge of physical phenomena, only monotonically decreasing functions can properly represent the distribution parameters regression in this case. Remaining of the listed functions remained good candidates for the regression model.

Before proceeding with the regression fit, days with less than 80 arrivals were removed as they usually stood for unusual events, like holidays or closed airport and contained outliers. Afterwards, to obtain the generalized model, the results for all airports were concatenated into $\mathbf{z}^M$ and $\mathbf{x}$ vectors of length n_G:

$$\mathbf{z}^M = \left[z^{1,M}, z^{2,M}, \ldots, z^{n_{\text{Airports}},M}\right], \tag{16}$$

$$\mathbf{x} = \left[x^1, x^2, \ldots, x^{n_{\text{Airports}}}\right]. \tag{17}$$

Predicted model distribution parameters $\boldsymbol{y}^M$ were obtained by passing $\boldsymbol{x}$ and $\hat{\boldsymbol{\beta}}$ to candidate functions:

$$\boldsymbol{y}^M = [y_1^M, y_2^M, \ldots, y_{n_G}^M]. \tag{18}$$

$$y_k^M = f\left(x_k, \hat{\boldsymbol{\beta}}\right), \quad k = 1, \ldots, n_G. \tag{19}$$

5.2. Coefficient of determination

Coefficient of determination R^2 [24], which measures how well model replicates observation, was used as a first factor to assess correctness of fit for each parameter M of the distribution:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}, \tag{20}$$

where:

- SS_{tot} is the total sum of squares:

$$SS_{tot} = \sum_{k=1}^{n_G} \left[z_k - \bar{z}\right]^2; \tag{21}$$

- SS_{res} is the sum of squares of residuals:

$$SS_{res} = \sum_{k=1}^{n_G} \left[z_k - y_k\right]^2; \tag{22}$$

- $\bar{z}$ is mean of observed parameter values $\boldsymbol{z}$:

$$\bar{z} = \frac{1}{n_G} \sum_{k=1}^{n_G} z_k. \tag{23}$$

5.3. Theil's inequality coefficient

Theil's inequality coefficient (TIC) U [25] was used as a second factor to assess correctness of fit for each parameter M of distribution:

$$U = \frac{\sqrt{\dfrac{1}{n_G} \sum_{k=1}^{n_G} \left[z_k - y_k\right]^2}}{\sqrt{\dfrac{1}{n_G} \sum_{k=1}^{n_G} \left[z_k\right]^2} + \sqrt{\dfrac{1}{n_G} \sum_{k=1}^{n_G} \left[y_k\right]^2}}. \tag{24}$$

TIC was also decomposed into separate factors which account for bias (U^M), variance (U^S) and covariance (U^C) [18]:

$$U^M = \frac{\left(\bar{z} - \bar{y}\right)^2}{\dfrac{1}{n_G} \sum_{k=1}^{n_G} \left[z_k - y_k\right]^2}, \tag{25}$$

$$U^S = \frac{\left(\sigma_z - \sigma_y\right)^2}{\dfrac{1}{n_G} \sum_{k=1}^{n_G} \left[z_k - y_k\right]^2}, \tag{26}$$

$$U^C = \frac{2\left(1-\rho\right)\sigma_z \sigma_y}{\dfrac{1}{n_G} \sum_{k=1}^{n_G} \left[z_k - y_k\right]^2}, \tag{27}$$

where:

- $\bar{z}$ is defined in equation (23);
- $\bar{y}$ is mean of model predicted parameter values $\boldsymbol{y}$:

$$\bar{y} = \frac{1}{n_G} \sum_{k=1}^{n_G} y_k \,. \tag{28}$$

- σ_z is standard deviation of observed parameter values $\boldsymbol{z}$:

$$\sigma_z = \sqrt{\frac{1}{n_G} \sum_{k=1}^{n_G} [z_k - \bar{z}]^2} \,; \tag{29}$$

- σ_y is standard deviation of model predicted parameter values $\boldsymbol{y}$:

$$\sigma_y = \sqrt{\frac{1}{n_G} \sum_{k=1}^{n_G} [y_k - \bar{y}]^2} \,; \tag{30}$$

- ρ is correlation coefficient of $\boldsymbol{z}$ and $\boldsymbol{y}$:

$$\rho = \frac{1}{\sigma_z \sigma_y} \frac{1}{n_G} \sum_{k=1}^{n_G} [z_k - \bar{z}]\,[y_k - \bar{y}] \,. \tag{31}$$

$U^M + U^S + U^C = 1$ and in ideal scenario $U^M = 0$, $U^S = 0$ and $U^C = 1$ as the first two are measures of systematic error and ability to duplicate variability by the model respectively, whereas the latter is a measure of nonsystematic error.

5.4. Regression fitting results

Regression fit results are presented in Fig. 5, in Fig. 6 and in Table 4. Power function $\beta_1 x^{\beta_2}$ turned out to be the best fit regression function and the function coefficients and correctness of fit factor values for both parameters are presented in Table 4. Values R^2, U, U_M, U_S, U_C testify that the fit is satisfactory, as $R^2 \geq 0.75$ and $U \leq 0.2$ show high level of correlation, while U_M, U_S, U_C show that nonsystematic error is dominant part of fit error.

Location parameter values for all airports are presented in Fig. 5 by blue scatter points. Regression representing the generalized model is presented in Fig. 5 by scattered black points.

Scale parameter values for all airports are presented in Fig. 6 by blue scatter points. Regression representing the generalized model is presented in Fig. 6 by black scattered points.

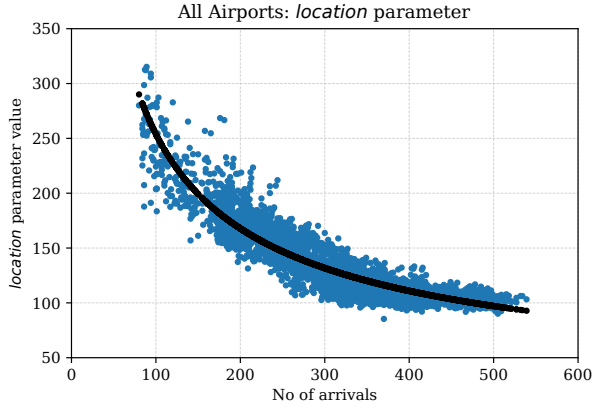

Fig. 5. *Location* parameter values for all airports and regression line (generalized model)

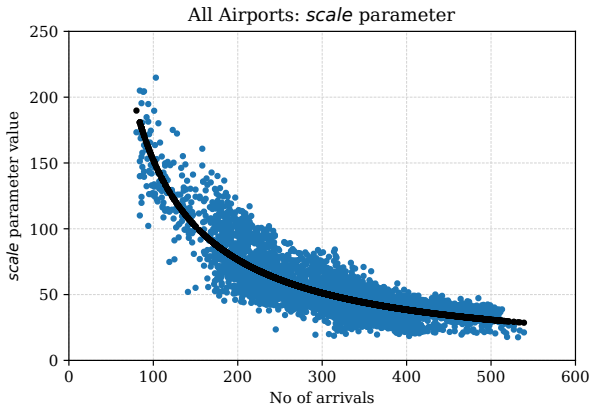

Fig. 6. *Scale* parameter values for all airports and regression line (generalized model)

Table 4
Regression fit for power function $\beta_1 x^{\beta_2}$: coefficients values and correctness of fit factors for generalized model

	location parameter	*scale* parameter
β_1	3972.89	14631.41
β_2	–0.597245	–0.991530
R^2	0.845397	0.756742
U	0.047211	0.107445
U_M	0.000031	0.000126
U_S	0.052675	0.084699
U_C	0.947294	0.915175

Finally, the generalized model can be expressed in the form of the probability density function as follows:

$$p\left(t_s^{i,j}, x\right) = \frac{1}{14631.41\pi x^{-0.99153}\left(1 + \frac{(t_s^{i,j} - 3972.89x^{-0.597245})^2}{14631.41x^{-1.98306}}\right)}, \tag{32}$$

where x stands for a number of arrivals and the following constraints apply due to the mechanism of the separation phenomenon, *Cauchy* distribution properties and assumptions in the study:

- $x \in \langle 80; 550 \rangle$;
- $p(t_s^{i,j}, x) \leq 0$ shall be rejected as they are not physical;
- as *Cauchy* distribution is defined on $(-\infty, +\infty)$ interval, it may sometimes yield big nonphysical values, so reasonable upper threshold should be also defined.

Example of sampling from the generalized model was performed for comparison with the real data. Daily fit presented in Fig. 2 was compared with the output from equation (32). Comparison is presented in Fig. 7. Red solid line presents the daily fit, the black solid line presents the model fit, and the blue histogram bars present results of the sampling model presented by the black line.

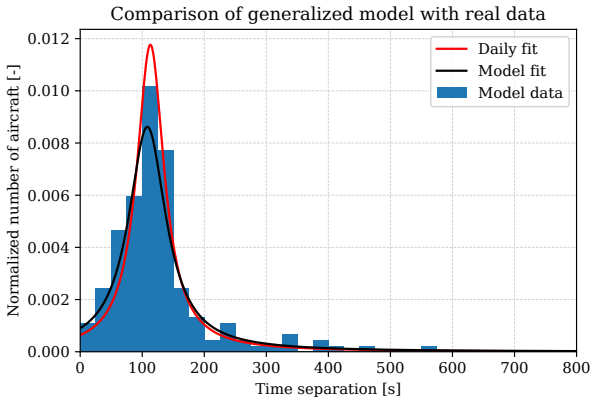

Fig. 7. Example comparison of daily fit and fit from generalized model (Model fit)

Results show that building a generalized model which comprises traffic from various airports is possible and aircraft air separation can be expressed by a single expression in the form of probability density function as shown in equation (32). As shown by the correctness of fit, and visible in Fig. 7, *location* parameter estimated value is usually more precise than *scale* parameter estimated value. However, peak value location looks to be a more important parameter for a task of modelling of the aircraft separations. As can be seen in Fig. 7, despite visible difference in *scale* parameter, thanks to good estimate of *location* parameter model was able to provide an output which is very close to the true data.

5.5. Generalized model application

Regarding practical applications, the major advantages of the generalized model of aircraft separations in the form of probability distribution are its simplicity, versatility, and flexibility of use. Relying on one simple explicit formula allows to apply the model quickly and reliable in various simulation or computational environments. A number of applications are proposed, and the applications are not limited with those listed. Examples of model applications:

- generation of samples for air traffic management algorithms testing and evaluation. Obtaining and processing historical air traffic data is time-consuming and requires additional resources. Moreover, the obtained data sets are usually limited to a certain number of days. Arriving aircraft separations model allows to quickly generate an unlimited number of samples which can be used for generating sequences of arriving aircraft, e.g., in the development of arrivals sequencing [15] or taxing algorithms [10];
- comparison of arriving traffic; Two metrics, location and scale parameters of Cauchy distribution, allow to quickly assess arrival traffic at different time intervals. Parameters values and parameters values differences (e.g., deviations from desired values) can be used for the assessment of arriving traffic with a goal of improving arriving traffic;
- validation of other separation models; Some models, e.g., neural network models, do not provide explicit formulas and function as black-box models. Even well-trained models of this kind might sometimes yield unrealistic data. The model provided within this study provides an explicit formula and can be used for quick validation with the use of e.g., TIC or R^2 factors provided in Section 5.

Sample application analysis was also performed. One thousand days of arrival sequences were generated assuming arriving traffic intensities of 300, 400 and 500 aircraft. Generation of samples was performed on a computer with Intel® Core™ i7 processor and Windows 10 operating system.

Results presenting the time consumed to generate samples are presented in Table 5. Total sampling times were around one second and increased with the number of samples. Median times for generating one sample were less than 0.001 seconds in all cases. Median times were provided, as due to nondeterministic threads executions times on Windows, some samplings took a few times longer than usual, so the mean value was not a good metric in this case. Sampling times prove time efficiency of the derived model.

5.6. Single airports models

In addition to the generalized model, single airport models were derived using Cauchy distribution and power function regression. Results are presented in Table 6, Fig. 8 and Fig. 9.

Table 5
Generalized model sampling timing analysis

Number of arrivals	Total sampling time [s]	Median sampling time [s]
300	0.702968	0.0006642
400	0.875851	0.0008324
500	1.138790	0.0009730

Location parameter values for all airports are presented in Fig. 8 by scatter points, where each colour represents a different airport, and each dot represents an individual day. Regression representing the generalized model is presented in Fig. 8 in black colour. The same scheme was used for presenting the *scale* parameters as can be seen in Fig. 9 in black colour.

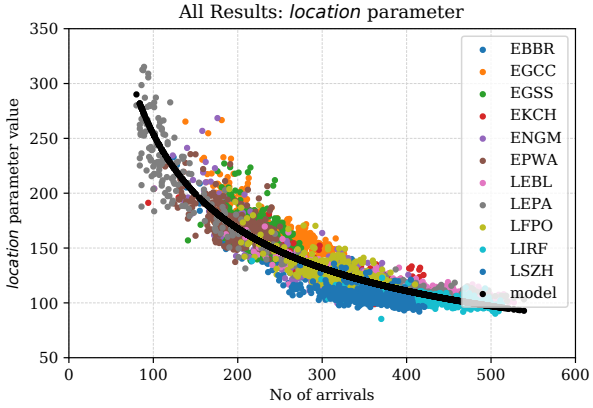

Fig. 8. *Location* parameter values for all airports and regression line representing generalized model

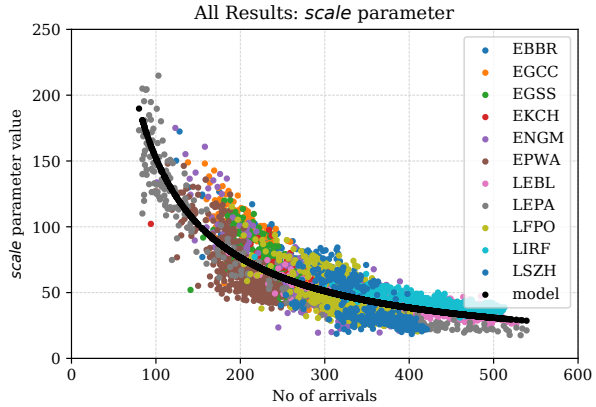

Fig. 9. *Scale* parameter values for all airports and regression line representing generalized model

As can be observed in Fig. 8 and Fig. 9 which present *Cauchy* distribution parameters for each airport together with the regression line, the data for each airport is within the regression line. Thus, the generalized model can be used as an approximation for the initial point when estimating the detailed models as well.

In Table 6 it can be observed that the power function coefficient values have significant differences. This is caused by the fact that for each airport the cluster of points covers a different range of arrivals numbers, what results in different regression of the function. However, TIC and R^2 show that identified models are very good or good. Compared to single airport models, there is an advantage of the generalized model which, thanks to the use of data from various airports, provides coverage of a wide range of arrivals. To efficiently compare the values of regression function coefficient between single airports models or between the single airport model and the generalized model, sampling should be done and TIC or R^2 criterion can be used.

Table 6
Regression fit for power function $\beta_1 x^{\beta_2}$: coefficients values and selected correctness of fit factors for single airports models

		β_1	β_2	R^2	U
EBBR	x_0	3725.49	–0.59173	0.74834	0.03621
	γ	24821.50	–1.05351	0.72257	0.07338
EGCC	x_0	4244.22	–0.58945	0.71782	0.03342
	γ	813838.50	–1.72116	0.74494	0.09224
EGSS	x_0	3581.48	–0.56959	0.37493	0.04081
	γ	129220.10	–1.38688	0.44851	0.09190
EKCH	x_0	2158.12	–0.49447	0.75780	0.03133
	γ	7299.78	–0.86414	0.73713	0.06750
ENGM	x_0	5175.54	–0.64583	0.81943	0.04588
	γ	75439.97	–1.28015	0.80583	0.10272
EPWA	x_0	1585.54	–0.42274	0.47320	0.03465
	γ	101476.31	–1.38098	0.59893	0.09390
LEBL	x_0	1736.23	–0.45293	0.83134	0.02171
	γ	9855.08	–0.92682	0.76967	0.05575
LEPA	x_0	3115.93	–0.55376	0.92135	0.04698
	γ	25898.23	–1.12527	0.93442	0.07586
LFPO	x_0	4378.45	–0.61517	0.59628	0.03649
	γ	627496.60	–1.65392	0.56594	0.11384
LIRF	x_0	499.89	–0.26134	0.43366	0.02124
	γ	5820.07	–0.81029	0.73217	0.03578
LSZH	x_0	860.82	–0.35737	0.31312	0.02886
	γ	4372038.62	–1.99359	0.42515	0.13718

6. CONCLUSIONS

In this study, a method for establishing a generalized aircraft air separation model in function of the number of arriving aircraft, by means of fitting the data to a continuous probability distribution and expressing the distribution parameters using nonlinear regression, was presented. Motivation and introduction to the study were followed by presenting the process of obtaining and processing arrival data, fitting the data to continuous distributions, and choosing the best fit distribution. Afterwards, the nonlinear regression approach to estimate the best fit distribution parameters in function of the number of arriving aircraft was presented, ending with providing the generalized aircraft air separation model and its comparison with real data. Model assumptions and constraints were also provided.

Model of this kind is a novelty and similar models have not been observed in the literature. As described within the article, aircraft separation models in the form of probability distribution can improve results, ATM modelling and optimization activities, as the use of artificial data can lead to insufficient coverage of algorithms testing and biased results. The methodology used in this research has been successfully used and produced accurate results in the modelling of aircraft motion for full flight simulators [18, 20] or atmospheric phenomena, e.g., icing conditions [26].

Modelling approximates reality and the models will always contain some inaccuracies. A number of error sources were presented in this study and future studies will focus on the minimization of errors impact, which will result in a more accurate model. As mentioned in Section 3, DDR2 data has two disadvantages: insufficient resolution and occurrence of flights which contain flight plan data, instead of real data. Other data sources, e.g. OpenSky Network can be considered which is of higher resolution, however several other issues are present, like validity and missing flights, as described in [12]. DDR2 data still should be used as the reference data, so ideally a merge of DDR2 and OpenSky Network data could give high-confidence high-accuracy data. Given the significant number of flights analysed within this study, all must be done with care to avoid mistakes and requires significant computing capabilities.

Future studies may also focus on building models which depend on other factors than the number of arrivals, like weather or airport configuration. Moreover, more sophisticated models can be developed which consider the traffic characteristics of each airport.

Finally, it is worth noting that the model estimation and assessment techniques presented within this article can be applied to various scientific disciplines, whereas the generalized air separation modelling approach can be applied to other transport fields, like railway or road traffic, where the need for modelling becomes higher due to autonomous technologies [27]. Also, more specific traffic aspects can be addressed, e.g.: vehicle separations in intermodal terminals [28].

REFERENCES

[1] EUROCONTROL, "Performance Review Report, An Assessment of Air Traffic Management in Europe during the Calendar Year 2019," Performance Review Commission, EUROCONTROL, 96, rue de la Fusée, B-1130 Brussels, Belgium, Tech. Rep., Jun 2020.

[2] R.E. Bryant, R.H. Katz, and E.D. Lazowska, "Big-Data Computing: Creating revolutionary breakthroughs in commerce, science, and society," 2008, a white paper prepared for the Computing Community Consortium committee of the Computing Research Association. [Online]. Available: http://cra.org/ccc/resources/ccc-led-whitepapers/, Accessed: 2021-25-08.

[3] V. Dhar, "Data Science and Prediction," *Comm. ACM*, vol. 56, no. 12, pp. 64–73, Dec 2013, doi: 10.1145/2500499.

[4] M. Motylewicz and W. Gardziejczyk, "Statistical model for traffic noise prediction in signalised roundabouts," *Bull. Pol. Acad. Sci. Tech. Sci.*, vol. 68, no. 4, pp. 937–948, Aug 2020.

[5] G. Carrozza, M. Faella, F. Fucci, R. Pietrantuono, and S. Russo, "Engineering air traffic control systems with a model-driven approach," *IEEE Software*, vol. 30, no. 3, pp. 42–48, 2013, doi: 10.1109/MS.2013.20.

[6] Y. Xu, H. Zhang, Z. Liao, and L. Yang, "A dynamic air traffic model for analyzing relationship patterns of traffic flow parameters in terminal airspace," *Aerosp. Sci. Technol.*, vol. 55, pp. 10–23, May 2016, doi: 10.1016/j.ast.2016.05.010.

[7] L. Yang, S. Yin, and M. Hu, "Network flow dynamics modeling and analysis of arrival traffic in terminal airspace," *IEEE Access*, vol. 7, pp. 73 993–74 016, Jun 2019, doi: 10.1109/ACCESS.2019.2921335.

[8] D. Delahaye and S. Puechmorel, *Modelling and Optimization of Air Traffic*. London, UK: Wiley-ISTE, 2013.

[9] M. Prandini, L. Piroddi, S. Puechmorel, and S.L. Brázdilová, "Toward air traffic complexity assessment in new generation air traffic management systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 3, pp. 809–818, Sep 2011.

[10] P. Gołda, T. Zawisza, and M. Izdebski, "Evaluation of efficiency and reliability of airport processes using simulation tools," *Eksploat. Niezawodn. – Maint. Reliab.*, vol. 23, no. 4, p. 659–669, 2021, doi: 10.17531/ein.2021.4.8.

[11] M. Kowalski, M. Izdebski, J. Żak, P. Gołda, and J. Manerowski, "Planning and management of aircraft maintenance using a genetic algorithm," *Eksploat. Niezawodn. – Maint. Reliab.*, vol. 23, no. 1, p. 143–153, 2021, doi: 10.17531/ein.2021.1.15.

[12] T. Polishchuk, A. Lemetti, and R. Sáez, "Evaluation of Flight Efficiency for Stockholm Arlanda Airport using OpenSky Network Data," in *Proceedings of the 7th OpenSky Workshop 2019*, vol. 67, 2019, pp. 13–24.

[13] A. Pawełek and P. Lichota, "Arrival air traffic separations assessment using Maximum Likelihood Estimation and Fisher Information Matrix," in *Proceedings of the 20th International Carpathian Control Conference*, Kraków-Wieliczka, May 2019, pp. 624–629.

[14] EUROCONTROL, *DDR2 Reference Manual For Generic Users 2.1.2*, EUROCONTROL, Brussles, Belgium, Jun 2015.

[15] A. Pawelek, P. Lichota, R. Dalmau, and X. Prats, "Fuel-Efficient Trajectories Traffic Synchronization," *J. Aircr.*, vol. 56, pp. 481–492, Mar 2019, doi: 10.2514/1.C034730.

[16] R. Dalmau, X. Prats, R. Verhoeven, F. Bussink, and B. Heesbeen, "Comparison of various guidance strategies to achieve time constraints in optimal descents," *J. Guidance Control Dyn.*, vol. 42, no. 7, pp. 1612–1621, Jan 2019, doi: 10.2514/1.G004019.

[17] R. Saez, X. Prats, T. Polishchuk, and V. Polishchuk, "Traffic synchronization in terminal airspace to enable continuous descent operations in trombone sequencing and merging procedures: An implementation study for Frankfurt airport," *Transp. Res. Part C Emerging Technol.*, vol. 121, no. C121, pp. 1–23, Dec 2020, doi: 10.1016/j.trc.2020.102875.

[18] R.V. Jategaonkar, *Flight Vehicle System Identification: A Time Domain Methodology*, 2nd ed., ser. Progess in Astronautics and Aeronautics. Reston, VA: AIAA, 2015.

[19] M. Hazewinkel, *"Maximum-likelihood method", Encyclopedia of Mathematics*, 2001st ed. Springer Science+Business Media B.V. / Kluwer Academic Publishers, 1994.

[20] P. Lichota, "Aircraft system identification using simultaneous quantized harmonic input signals," *Bull. Pol. Acad. Sci. Tech. Sci.*, vol. 68, no. 6, pp. 1351–1362, Dec 2020, doi: 10.24425/bpasts.2020.135397.

[21] Y. Pawitan, *In All Likelihood: Statistical Modelling And Inference Using Likelihood*, 1st ed. The address: Oxford University Press, Mar 2013.

[22] N.L. Johnson, S. Kotz, and N. Balakrishnan, *Continuous Univariate Distributions*, 2nd ed. New York: Wiley, 1994, vol. 1.

[23] J.J. Moré, "The Levenberg-Marquardt algorithm: Implementation and theory," in *Numerical Analysis*, G.A. Watson, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 1978, pp. 105–116, doi: 10.1007/BFb0067700.

[24] N.R. Draper and H. Smith, *Applied Regression Analysis*, 3rd ed. New York: Wiley, 1998.

[25] P. Lichota, F. Dul, and A. Karbowski, "System Identification and LQR Controller Design with Incomplete State Observation for Aircraft Trajectory Tracking," *Energies*, vol. 13, no. 20, p. 5354, 2020, doi: 10.3390/en13205354.

[26] C. Deiler, "Aerodynamic modeling, system identification, and analysis of iced aircraft configurations," *J. Aircr.*, vol. 55, no. 1, pp. 145–161, 2017, doi: 10.2514/1.C034390.

[27] S.A. Bagloee, M. Tavana, M. Asadi, and T. Oliver, "Autonomous vehicles: challenges, opportunities, and future implications for transportation policies," *J. Mod. Transp.*, vol. 24, p. 284–303, 2016, doi: 10.1007/s40534-016-0117-3.

[28] M. Jacyna, R. Jachimowski, E. Szczepański, and M. Izdebski, "Road vehicle sequencing problem in a railroad intermodal terminal – simulation research," *Bull. Pol. Acad. Sci. Tech. Sci.*, vol. 68, no. 5, pp. 1135–1148, Oct 2020, doi: 10.24425/bpasts.2020.134643.