Traffic Complexity in ANSP Evaluation

Applicability of Current Metrics for Benchmarking Purposes

Thomas Standfuß
Institute of Logistics and Aviation
Technische Universität Dresden
Dresden, Germany
thomas.standfuss@tu-dresden.de

Matthias Whittome
FABEC
Langen, Germany
matthias.whittome@fabec.eu

Abstract— **Air traffic control is one of the main pillars in the air transport system, responsible for safe and efficient flight guidance. Due to regulatory purposes regarding the monopolistic Air Navigation Service Providers, economic benchmarking in air traffic management is also a subject of research. One of the key inhibitors of operational performance is traffic complexity, which contributes to a higher workload of air traffic controllers. We used efficiency and regression analyses to show the impact of traffic complexity, calculated by Performance Review Unit, on ANSP performance. We will show that the expected interdependency between complexity and performance cannot be proven. In consequence, the score should neither be used in ex-post performance analyses, nor as a weighting factor in economic modeling, nor for regulatory purposes. Thus, we recommend revising the calculation of complexity fundamentally, which might be beneficial for target setting in the next regulation period.**

Keywords: Performance Benchmarking; Complexity; Airspace; Air Traffic Management; Air Navigation Service Providers

I. Background

European airspace is one of the busiest in the world with high air traffic demand and density. Although the current demand in 2020 is severely limited due to the COVID-19 pandemic, the efficient utilization of airspace capacity still represents one major objective function in today's air traffic management (ATM). Furthermore, the Air Navigation Service Providers (ANSPs), which are responsible for safe and efficient flight guidance, represent companies in a monopolistic market structure and thus are subject to regulations to offer services at minimum costs or resources.

Therefore, economic benchmarking aims to assess operational (e.g. flight hours per sector hour) or financial (e.g. costs per flight-hour) performance of the ANSPs. It reveals absolute or relative deviations from the "best practice" and thus enables to derive targets for costs or resource utilization. Numerous methods have been developed for economic benchmarking, including index number methods, Data Envelope Analysis (DEA), and Stochastic Frontier Analysis (SFA). The resulting performance score may express productivity as well as efficiency. As an example, the Performance Review Unit (PRU) of EUROCONTROL applies a productivity metric which uses an aggregated measure of controlled flight hours and airport movements as an output ('Composite Flight Hours', see also [1]) and the required number of Air Traffic Control Officer (ATCO) hours to control these flights as an input. The corresponding performance score is "ATCO-productivity". EUROCONTROL publishes annual reports addressing the operational and financial performance of ANSPs, e.g. in [2], [3], or compare European and US-American Air Navigation Services (ANS), such as [4], [5]. More recently, authors use alternative methods to calculate ANSP efficiency [6], or addressing effects like economies of scale [7].

The operational performance of an ANSP is not only determined by endogenous variables but is also influenced by exogenous factors. As one of these factors the complexity of traffic and/or airspace was identified, which was initially addressed as a proxy for safety measurement. In recent years, studies approached air traffic "complexity" from various directions and for various purposes, e.g. [8–11]. PRU developed a complexity score reduced to a limited number of features for economic benchmarking, reflecting the distribution and movements of airspace users [12] [13]. The main advantage is, that the PRU complexity score is published on a daily, monthly and annual basis, and on several operational levels. Thus, the complexity values are available for the use in official reports and academic studies. As an example, [14] used the PRU complexity score for economic modeling to set performance targets for Reference Period 3 (RP3). Since this target setting might be used for regulatory purposes of the ANSP as a monopoly, the PRU complexity score gained a high political impact on the regulation of European ANSPs.

Despite the frequent use of this score in efficiency analyses, there is no scientific legitimation for the use of this metric yet. In other words, the plausible derivation of complexity as a descriptive measure for a higher ATCO workload was not specifically questioned in the context of economic efficiency. It has to be considered, that (more complex) model approaches in the sense of workload studies could already prove such a correlation. Therefore, we will investigate whether the PRU complexity score, as a proxy for ATCO workload, is suitable for economic benchmarking and if it correctly offsets all essential factors. The second section deals with the calculation method of the PRU score and its sub-components, assessing it carefully. Section 3 investigates, whether there is an interdependency between performance and complexity. Since the complexity score represents an ex-ante metric while the performance is calculated ex-post, it is expected that the interdependency between complexity and productivity or efficiency be not established properly. Since the score is also being used to assess and/or cluster ACCs, the interdependency between ACC productivity and complexity is analyzed as well. Section 4 discusses potential shortcomings, addressing e.g. database (planned data versus actual data) and potentially missing factors (principal components of complexity). Section 5 summarizes the findings.

II. The PRU Complexity Metric

This section is intended to give a short overview of the calculation method of the complexity score, applied by the Performance Review Unit. A comprehensive description is provided by [12] and [13].

In 2003, The ATM Cost-Effectiveness (ACE) working group initiated efforts to create a complexity metric for benchmarking purposes. The objective was to define a set of high-level complexity indicators for En-route airspace based on controller workload. The score is intended to cover exogenous effects that contribute to the ATCO workload, excluding endogenous effects such as ANSP decision making. PRU implied that traffic density, traffic flow, vertical traffic, and the traffic mix are the main contributors to complexity. They exclude potential factors like sectorization and route structure, military traffic and interface with adjacent units, special events like large-scale military exercises, as well as predictability and variability of traffic

To calculate complexity, PRU divided the European airspace into cells of equal size of 20x20 nautical miles horizontally and 3000 feet vertically (range: FL85-FL385). The size of the cells allows the assignment of each cell to an individual ACC or ANSP. To avoid boundary effects, the grid is shifted four times horizontally (10nm steps) and three times vertically (1000ft steps). Subsequently, the final score reflects an average of twelve values, as described in [12].

The score is based on planned data and thus does not consider actual trajectories. It measures the potential interactions within a one-hour time period. An interaction is defined by the presence of two (or more) aircraft in one cell. Interactions are always considered bilaterally, which means, that the presence of two aircraft results in two interactions, three aircraft analogous lead to six interactions. The used traffic data represents the initial flight plan, provided by the Network Manager (NM). In a next step, PRU calculates the hours of the interactions. The calculation is performed for each pair of aircraft in the specific cell. The sum of the durations provides the hours of potential interactions for that cell.

The final complexity score consists of four components, representing the most influencing traffic characteristics assumed by PRU: Adjusted Density (*AD*), Vertical Interactions (Vertical Different Interacting Flows, *VDIF*), Horizontal Interactions (Horizontal Different Interacting Flows, *HDIF*), and Speed Interactions (Speed Different Interacting Flows, *SDIF*).

The density of traffic is usually defined by the number of vehicles in a defined area or volume. However, the distribution of aircraft in the airspace is nearly always spatially concentrated. The division of airspace into cells enables the identification of traffic “accumulations”, leading to the ‘adjusted’ density. The adjusted density value is calculated by summing up all durations of interaction and dividing it by the number of total flight hours *f* of the ACC or ANSP.

The ‘vertical Interactions’ i_v cover flights in different flight phases (climbing/cruising/descending). No interaction is considered for flights with the same attitude (e.g. two climbing aircraft). An aircraft is “cruising” if the vertical speed is less than +/- 500 ft/minute. The corresponding indicator (VDIF) sums up the hours of vertical interactions and divide it by the number of flight hours. Different headings of the flights are considered for the ‘horizontal interactions’ i_h. The threshold is 20 degrees divergence in the heading. The corresponding indicator (HDIF) is the ratio of the horizontal interaction hours and flight hours. The ‘speed interactions’ i_s consider aircraft with a difference in speed of more than 35 knots. The speed is provided by Base of Aircraft Data (BADA) profiles. The sum of speed interaction hours is divided by the sum of flight hours to calculate the corresponding indicator (SDIF).

While *AD* represents a measure of concentration, *VDIF*, *HDIF*, and *SDIF* reflect flow characteristics. Due to the mathematical background, the flow indicators are a subset of the adjusted density and therefore there is a high correlation between both categories. To avoid this effect, PRU divides the flow-indicators by the adjusted density, leading to horizontal, vertical, and speed scores (HS, VS, SS, also see Table I). Thus, PRU wants to enable not only to evaluate which unit is complex but also if the complexity is basically due to density or flow characteristics.

The sum of the (*AD*-normalized) horizontal, vertical, and speed component is defined as “structural index” *s*. The overall complexity score *c* of an ANSP or ACC is represented by the product of structural index and adjusted density. Mathematical transformations lead to formula (1) for the structural index and formula (2) for the complexity score.

$$s = \frac{VDIF}{AD} + \frac{HDIF}{AD} + \frac{SDIF}{AD} \tag{1}$$

$$c = AD \cdot \left(\frac{VDIF}{AD} + \frac{HDIF}{AD} + \frac{SDIF}{AD}\right) = \frac{(i_v \cdot d) + (i_h \cdot d) + (i_s \cdot d)}{f} \tag{2}$$

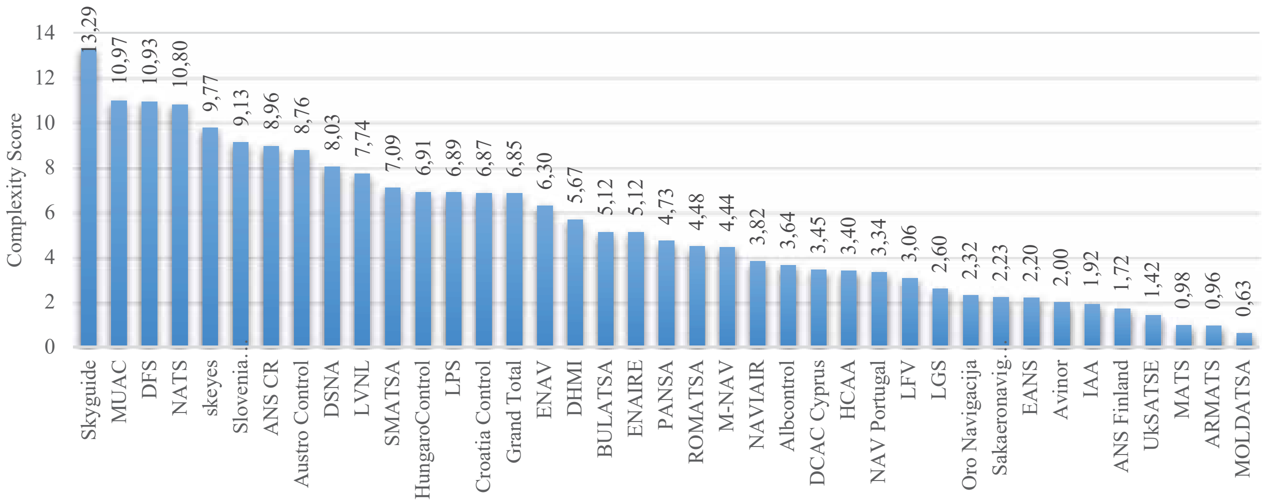

Figure 1. Complexity Scores of European ANSPs, 2019 [16]

As shown Figure 1, the highest score is achieved by Skyguide (Switzerland), which is 21-times higher than the traffic complexity of MOLDATSA (Moldova). Six out of the top ten ANSPs with the most complex airspaces belong to the Functional Airspace Block Europe Central (FABEC) [16].

As already described, the illustrated score consists of four components. Table I shows, that the most contributing factor is adjusted density. For most of the ANSPs, in particular those with high traffic demand, structural index as well as its components, play a minor role only.

TABLE I. Complexity Score and Components for some ANSPs

ANSP	AD	VS	HS	SS	SI	CS
Skyguide	12.80	0.23	0.63	0.17	1.04	13.29
MUAC	11.80	0.23	0.57	0.14	0.93	10.97
DFS	10.84	0.24	0.59	0.18	1.01	10.93
NATS	10.56	0.35	0.46	0.21	1.02	10.80
skeyes	8.12	0.38	0.56	0.27	1.20	9.77
MATS	1.86	0.07	0.37	0.09	0.53	0.98
ARMATS	1.71	0.12	0.36	0.08	0.56	0.96
MOLDATSA	0.95	0.07	0.50	0.10	0.67	0.63

III. Interdependency between Complexity and Performance

A. Approach and Methods

The complexity score was introduced by PRU, intending to measure the exogenous effects of traffic behavior on the workload of ATCOs [12]. Subsequently, there should be an interdependency between the complexity and performance of an ANSP. The key underlying rationale is that a higher complexity results in higher workload, which decreases capacity and productivity.

In the first step, the dependency between the complexity score and ATCO-Productivity (Composite Flight Hours per ATCO-hour) is tested for the year 2017. Complexity and operational data are provided by [15] and [16]. Since the PRU score is primarily aimed to address En-route services, we exclude terminal services in a second step. Therefore, the 'Total Controlled Flight Hours' (Output) were divided by the ACC ATCO hours (Input).

Since the complexity score represents ex-ante calculations and performance is measured ex-post, it is expected that there is no interdependency. In fact, planned data is compared with an already "adjusted" system. However, there might be other reasons for missing interdependency, such as insufficient metrics or an inappropriate analysis method.

The first two steps inhere two potential sources of bias. First, the two-dimensional analysis scheme might not be sufficient since multiple factors could influence productivity. Second, the productivity measure might not be sufficient as a performance indicator. ANSPs use multiple resources (staff, capital, etc.) to produce multiple outputs (flight hours, airport movements, etc). Subsequently, the performance score should consider all elements of economic value creation.

In consequence, we applied regression analyses to identify functional interdependencies between the dependent variable (performance) and multiple independent variables (assumed influencing factors). The results are based on the work of [17], who build up an economic model for ANSPs and identified multiple endogenous, exogenous, and partly-exogenous factors. For the dependent variable, we first use the PRU productivity score. To avoid the discussed disadvantages of the productivity measure (first and second step), we used efficiency scores based on Data Envelopment Analysis which were calculated by [7]. The efficiency scores are available for cross-sectional as well as for panel data. A comprehensive description of the DEA-methodology is provided e.g. by [18].

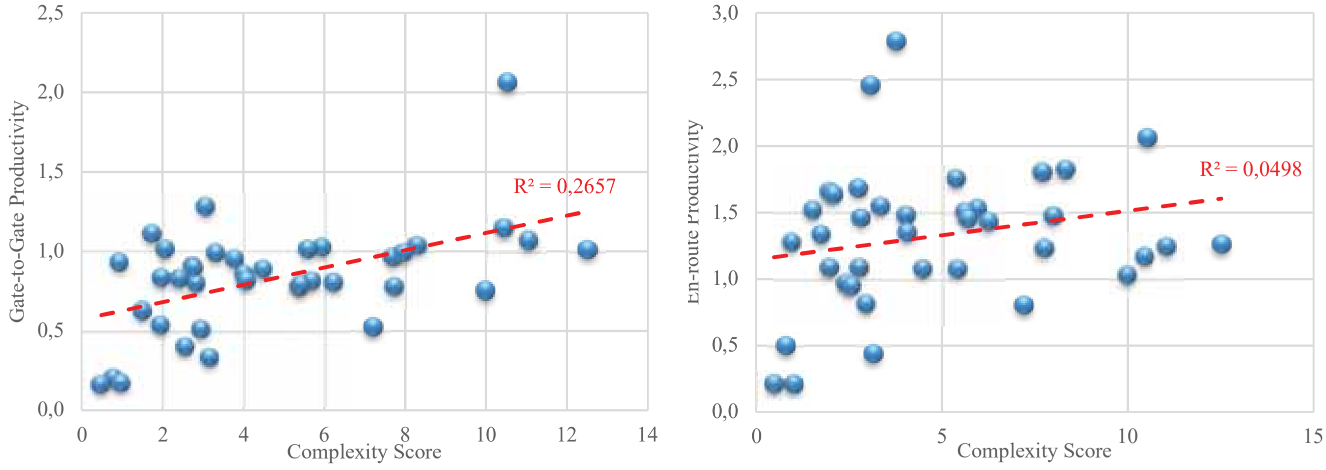

Figure 2. Dependency between complexity score and performance gate-to-gate (left) and for En-route services only (right), 2017

In the last step, we analyzed whether there is an interdependency between complexity and productivity on the ACC level. Intuitively there might be a higher probability for correlation since some ANSPs consists of multiple ACCs and the subsequent ANSP-figures represent cross-sectional average data. PRU provided operational data for 62 ACCs Interdependency Analysis on ANSP-Level

B. *Correlation Analysis*

In the first analysis step, the complexity scores are plotted against the ATCO productivity score in a gate-to-gate perspective. In general (and not considering the ex-ante vs. ex-post data problem) one would assume, that there is a negative interdependency since complexity increases workload. However, Figure 2 (left) shows that productivity increases with complexity, which is rather unexpected. However, the slope of the (linear) trend is near zero, which means that there is no interdependency at all. Furthermore, the statistical significance is weak, since the coefficient of determination (R^2) is below 0.27. These findings might be influenced by Maastricht Upper Area Control Centre (MUAC), which represents an outlier in the dataset. However, eliminating this unit does not change the overall picture.

As discussed in the past section, the unexpected result might be caused by the fact, that the complexity measure was primarily aimed to assess En-route traffic. Using a gate-to-gate perspective, terminal services are included in the productivity score, which may lead to the missing interdependency. However, as shown in Figure 2 (right), there is a slightly positive slope, indicating a non-negative interdependency. Again, the coefficient of determination is very low (0.05). In comparison with the productivity score containing both services, there is even less evidence that performance depends on traffic complexity.

The missing interdependency might be further due to methodological reasons. The productivity score of PRU uses only one input (ATCO hours) and one output (Composite Flight Hours) [2]. The latter represents an aggregated value of flight hours and airport movements, which is weighted by the unit cost ratio of en-route and terminal services [1]. Depending on the quality of the data, there is a possible source of error: If only one of the values is measured imprecisely, it can significantly change the productivity value and thus the ranking of ANSPs as well [19]. In order to eliminate the potential bias caused by an insufficient performance metric, we used alternative scores as well (see next section).

C. *Efficiency and Regression Analysis*

The economic model of an ANSP consists of multiple resources, multiple services, and multiple environmental variables. As an example, the inputs should include all relevant ressources (staff, capital, etc). Statistical tests help to choose the factors, which are included in one benchmarking model simultaneously. Substitutions of variables enable the consecutive consideration of alternative factors (in separate benchmarking models). By comparing the results, it is possible to ensure robustness. Fundamental analyses of inputs and outputs within benchmarking models are provided by [7] .

Academic studies primarily use methods that can take multiple inputs and outputs into account, such as Data Envelopment Analysis or Stochastic Frontier Analysis, e.g. [20], [21]. Applying the non-parametric approach of DEA, the efficiency of units (e.g. ANSPs) is calculated in the first step. The second step analyses the influence of factors on the performance, such as complexity. Although there may be other methods to investigate and quantify influences of multiple factors on one or more dependent variable, regression analysis is a common and widely used method, especially in combination with DEA efficiency analysis.

This two-step methodology is applicable to our problem, no matter whether the performance measure is based on an index figure (PRU) or another benchmarking methodology (DEA). In order to avoid biased regression results due to a potential inaccuracy of the productivity score, we additionally use the DEA-scores of four different benchmarking models calculated in [7]. The scores are based on both cross-sectional and panel data. For the second step, we ran regression analyses, including multiple endogenous and exogenous factors. The selection of factors is crucial for a valid result: considering insignificant factors or not taking significant factors into account may result in OVB as well, leading to an over- or underestimation of the factors concerning their influence on performance. We used multiple variables, such as airspace size, ownership forms, volatility, socio-economic determinants, and complexity, following the model of [17]. Thus, we avoid omitted variable bias (OVB).

Different benchmarking methods require different regression models: For the PRU productivity score, an Ordinary Least Squares (OLS) regression was applied. Since efficiency scores based on DEA are limited between 0 and 1 (respectively 0% and 100%), OLS is not sufficient. Alternatively, censored and truncated regression methods were applied. For methodological background, see [22]. Since data is available for multiple years, it is advisable to run panel models as well, increasing the data points and thus increase regression model quality. Therefore, the influence of factors (including PRU complexity score) was quantified using Fixed and Random Effects Models (FEM, REM) as well as pooled regression (see also [23]).

The regression analysis was applied in three steps. First, the complexity was represented by the corresponding PRU score. Second, the aggregated score was broken down into the components adjusted density, horizontal score, vertical score, and speed score. Third, in order to check whether the productivity indicator might be insufficient, the dependent variable was substituted by efficiency scores, based on the different DEA models described above.

None of the regressions show statistical evidence that the complexity score influences productivity or efficiency. In panel analysis (FEM and REM), complexity shows statistical significance, but a positive influence on productivity and efficiency. These results are intuitively wrong, but confirm the observed effects shown in Figure 2 In consequence, we analyzed whether there is evidence that the components influence performance.

Using PRU productivity as an independent variable, the components were not statistically significant in most cases. The horizontal score was significant in some cases but had a positive sign mostly. The same applies to adjusted density. Using panel data and PRU productivity, all components were significant. However, the vertical score and adjusted density still had a positive sign.

As an example, Table II shows the results for a regression model using cross-sectional data. The PRU productivity score represents the dependent variable. We maximized the model quality by variable reduction. Therefore, we used a threshold for the p-value of ≤ 0.33. The other components reflect endogenous, exogenous, and partly exogenous factors. A comprehensive description of the method and the variables is provided by [17].

The result shows that speed and vertical interactions have no influence on productivity (p-values > 0.33). The optimized model contains Horizontal Interactions (*HI*) and Adjusted Density (*DENSITY*) only, the latter is statistically significant as well. The sign is positive for both coefficients, leading to the (contra-intuitive) conclusion of an increase in productivity in case of a growing complexity. The overall model quality (adjusted R²) is very high.

Using efficiency scores as the independent variable, the results were equivalent. The panel regression analysis showed a significantly negative impact of the speed score and significantly positive influence of the vertical score on performance.

TABLE 2. REGRESSION RESULT FOR CROSS-SECTIONAL DATA

INT	-2,69 (1,684)
NONA	-0,454 (0,328)
TIME (I)	0,213 (0,178)
DELATM	-0,138 (0,071*
AIRP	0,576 (0,142)***
JSC	0,129 (0,073)*
STATE	-0,065 (0,065)
SIZE (I)	0,125 (0,038)***
OCEAN	0,211 (0,095)**
COORD	0,019 (0,018)
OVER	0,587 (0,15)***
DOM	-1,8 (0,568)***
GINI	-3,069 (0,597)***
DENSITY	0,026 (0,013)*
HI	0,619 (0,361)
COSTS	0,002 (0,001)**
RES	-0,003 (0,002)
Adj. R²	0,86
Akaike	-36,38
N	38

Standard error in brackets. Significance level 90 %*, 95 %**, 99 %***

The reduction of variables to maximize model quality is only one way of building the regression model. In [15], alternative methods were applied as well, considering endogenous, exogenous, and partially exogenous variables in consecutive iterations. However, the results do not differ significantly regarding the influence of complexity on performance.

The regression analysis confirmed that there is no evidence that either the PRU complexity score or its components have a statistically significant influence on performance. For the quantified regression analysis, the results were neither robust nor intuitively correct in most of the cases. The lack of interdependency raises doubts about whether the used metric is sufficient to reflect complexity.

D. Disaggregation on ACC Level

After we found no evidence that complexity influences productivity on the ANSP level, we check whether the observations can be confirmed on the ACC level. The highest productivity scores (Flight hours per ATCO hour) are achieved by Warszawa, Lisbon, and MUAC, the lowest by Chisinau, Kyiv, and Dnipropetrovs'k. ACCs of FABEC represent eight out of ten ACCs with the highest score. The lowest complexities were calculated for ACCs which are located in Europeans eastern and northern periphery.

We found out that the slope of the trend line is positive but close to zero (Figure 3). Both the rise and the low coefficient of determination suggest that no correlation can be proven for ACCs. However, the dataset is determined by two extreme values: London TC and Istanbul. Despite the R² becomes larger when excluding both ACCs from the dataset, the functional interdependency leads to a similar result. Assuming a parabolic trend (green line), the R² is higher, implying that productivity initially increases with increasing complexity (e.g. due to higher demand and thus adjusted density). After a turning point, however, productivity decreases as complexity increases.

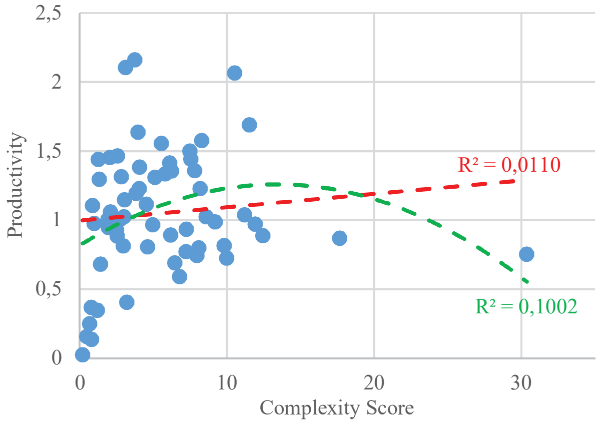

Figure 3. Dependency between complexity score and performance (ACC)

Since EUROCONTROL uses the score to cluster ACCs into four groups [2], we also checked the corresponding interdependencies within these clusters. The results (R²) were insignificantly better for two of the clusters, but significantly worse for the other two groups.

IV. Shortcomings and Missing Contributors

A. Modeling

The past sections showed the missing interdependency between performance indicators and the PRU complexity scores. By applying different methodologies, we could show that this (missing) effect is neither due to the performance indicator, nor the analysis method, nor the included services (En-route / Terminal). In consequence, the reason is the complexity metric itself. Therefore, we provide an overview of potential shortcomings and missing contributors.

The differentiation between exogenous and endogenous factors of complexity is necessary. A traffic complexity score should not contain any internal heterogeneities. However, route structure and sectorization are considered to be endogenous, which is only partly correct. Especially the route structure is influenced by the demand itself and might be considered as exogenous. It might be assumed that these factors are not considered due to the data (initial flight plan).

It is also stated that military traffic and adjacent units are not considered, because there is no possibility of quantification. Nevertheless, the currently available databases might enable an approximation. For example, neighboring units could be evaluated by the number or length of the boundaries. Military activities can be depicted e.g. by the number and opening hours of Temporarily Reserved Areas (TRAs).

The speed interactions are simulated based on BADA data. Thus, the score implies that all aircraft fly BADA compliant. Subsequently, changes in the speed of a specific aircraft are not considered. Therefore, the speed score might inhere some inaccuracies. Furthermore, a threshold of 35 knots might be seen as too high, especially for ANSPs which mainly serve lower airspaces. However, as shown in Table I, the speed score has only a minor influence on the total score

The size of the airspace cells might be debated as well. PRU [12] states the “20nm cell size was chosen because it mapped the ACC boundaries more closely than a larger cell size while maintaining a macroscopic view. It is not discussed why the cell hight was set to 3000ft and whether Reduced Vertical Separation Minimum (RVSM) was / should be considered. In further studies, it might be beneficial to run sensitivity analyses by varying cell sizes.

The pairwise separation of three aircraft result in six interactions in the PRU metric. However, this is the maximum number. In reality, the number of interactions might be lower (e.g. due to vertical separation). The multiplication of interaction hours would imply independency of the traffic (flows), which is not the case in reality.

Table I showed, that adjusted density has a high impact on the overall complexity score. This is mainly due to the fact, that the components of the structural index are divided by a. In consequence, the influence of spatial distribution might be underestimated for small entities (such as MOLDATSA) and overestimated for large ANSPs (such as DSNA). A clear separation into the contribution of density and flow to complexity, as intended by PRU, is therefore questionable. As an example, it is not clear why the structural index for MUAC is higher than for the Dutch LVNL. Although LVNL is responsible for lower airspace only, the vertical score is lower than the one of MUAC (upper airspace). This is intuitively wrong since the share of cruising traffic of LVNL is 7%, but 54% for MUAC [15]. Further, the ex-ante calculation of AD might be biased, since ATCOs will avoid traffic accumulations.

B. Meaningfulness

Since the airspace size of a single unit (e.g. Skyguide), as well as the spatial allocation of (large) airports, is nearly constant, it could be assumed that the density factor, and thus the complexity score as well, might be mainly determined by the traffic volume itself. Comparing the monthly figures of demand and complexity of the Swiss ANSP skyguide, both curves have a similar shape. This result was confirmed for other ANSPs as well: 24 of 38 ANSPs show a high correlation (> 80%) between demand (flight hours) and complexity, with a median of 88.9%. These findings may lead to the conclusion that the PRU complexity score can be substituted by the demand itself, raising doubts whether there is an added value of calculating the score.

Some of the points of criticism discussed are due to the calculation of scores, which are based on ex-ante simulations (flight plan data). However, the assessment of performance is ex-post. It seems much more useful to calculate complexity scores ex-post as well, based on actual trajectories. The utilization of planning data (instead of actual data) leads to several problems.

Ex-post analyses inhere all adjustments of traffic- and airspaces structures, as well as the used resources, especially ATCO hours (since the score is provided on ANSP and ACC level only, the airspace adjustments are not relevant). This may hamper the comparability of data significantly. In consequence, the parallel use of both metrics (performance score and PRU complexity score), as well as the underlying datasets, could be insufficient. This also speaks against the usage of complexity as a weighting factor, which may be done for cost target setting in [14].

C. Comprehensiveness

A major limitation by using ex-ante data is the missing inclusion of important influencing factors. One main contributor to the ATCO workload is military traffic. This contains primarily the number and opening times of Temporary Reserved Areas (TRAs), which reduce the available airspace for civil air traffic (especially commercial airlines) and the possibility to provide directs. Thus, TRAs might inhere traffic detours and delays. In addition to the TRAs, military traffic contributes to complexity by crossing civil airspaces (to and from TRA). The predictability of military demand is low, which amplifies the negative impact on capacity provision.

Complexity is further dependent on environmental conditions, as discussed by [24]. Traffic flows are dependent on winds and, if existent, convective clouds/areas, which will be avoided due to the presence of high turbulences. With regards to severe weather areas, flight crews may adjust their routing to the new situation. These actions are dependent e.g. on the changes in the intensity of cells of severe weather (vertically and horizontally), the situational awareness of the flight crew and routing decisions they take based on the display on their weather radar, the altitude of aircraft, the onward routing of the aircraft ([25]).

In Terminal Maneuvering Areas, the complexity is amplified by strong winds, shifting wind directions, low visibility, rain- or snowfall, and convective weather. These conditions directly affect airport capacity, which may lead to additional unanticipated traffic in surrounding sectors (e.g. for holdings). In consequence, heavy weather may lead to lower predictability and/or higher short-term volatility.

Complexity might further be influenced by general aviation (GA) traffic. This type of traffic uses lower airspaces, which increase potential interactions with descending and climbing commercial traffic. Thereby, GA traffic causes a higher task load for ATCOs and thus induce more complexity. GA using visual flight rules (VFR) primarily crosses uncontrolled airspaces. Nevertheless, they contribute to the ATCO workload when crossing e.g. airspaces, near airports. In future, there might be a further contributor, represented by Unmanned Air Vehicles (UAV), such as drones.

V. Conclusions and Outlook

We investigated the methodology and components of the PRU complexity score, which was introduced for economic benchmarking of air navigation service providers [12]. It is expected that different traffic situations (e.g. climb, descent) require more controller input and thus have to be considered in the productivity analysis. The PRU complexity score was developed as a proxy to provide a weighted metric to cover these different traffic situations to be handled by local operators.

Dealing with traffic complexity based on the PRU score raised four questions:

- Is the metric meaningful?
- Is there an added value?
- Is the score comprehensive?
- Is the score valid for benchmarking?

Our investigation points out that the current PRU metric seems not to be an appropriate candidate for the evaluation of the performance of Air Navigation Service Providers. In particular, we find no statistical evidence for interdependency between complexity and productivity. This result was constant across methods and models at both ANSP and ACC level. In consequence, the ex-ante complexity metric should not be used for ex-post performance benchmarking, especially in economic modeling.

Thus, concerning the study on performance targets in RP3, it must be noted that weighting according to complexity can lead to a distorted efficiency measure and thus performance targets. As shown before, complexity increases at almost the same rate as total demand, weighting the output with complexity means multiplying demand with demand.

It is recommended to revise the complexity measure due to the numerous potentials of improvement showed in this paper. Current academic studies on components and calculation methods may be taken into account. Updating the assessment of complexity could further improve the performance benchmarking itself, e.g. with regard to the target setting for RP4.

REFERENCES

[1] EUROCONTROL, *Composite flight-hour*. Performance Review Unit: https://ansperformance.eu/definition/composite-flight-hour/ (29.04.2020), 2020.

[2] EUROCONTROL, *Air traffic management cost-effectiveness (ACE) benchmarking report for 2017*. Brussels: Performance Review Unit, 2019.

[3] EUROCONTROL, *Performance Review Report - An Assessment of Air Traffic Management in Europe during the Calendar Year 2018*. Brussels: Performance Review Commission, 2019.

[4] EUROCONTROL and FAA, *U.S. - Europe Comparison of ATM related operational performance 2017*. Brussels: , 2019.

[5] EUROCONTROL and FAA, *U.S. - Europe continental comparison of ANS cost-efficiency trends 2006-2016*. Brussels: Peformance Review Unit, 2019.

[6] V. Bilotkach, S. Gitto, R. Jovanović, J. Mueller, and E. Pels, “Cost-efficiency benchmarking of European air navigation service providers,” *Transportation Research Part A: Policy and Practice*, vol. 77, pp. 50–60, 2015.

[7] T. Standfuss, F. Fichert, M. Schultz, and P. Stratis, “Efficiency losses through Fragmentation? Scale effects in European ANS Provision,” *Competition and Regulation in Network Industries*, vol. 20, no. 4, pp. 275–289, 2019.

[8] A. G. Diaconu, V. Stanciu, and O. T. Pleter, “Air traffic complexity metric for en-route and terminal areas,” *Scientific Bulletin University Politehnica of Bucharest*, vol. 76, pp. 13–24, 2014.

[9] J. Djokic, “Investigation into Air Traffic Complexity as a Driver of a Controller‘s Workload,” Technische Universität Dresden, 2014.

[10] E. Salaün, M. Gariel, A. Vela, E. Feron, and J.-P. Clarke, *Airspace Complexity Estimations Based on Data-Driven Flow Modeling*. Anaheim: AIAA Guidance, Navigation, and Control Conference, 2010.

[11] M. Vogel, K. Schelbert, H. Fricke, and T. Kistan, “Analysis of Airspace Complexity Factors’ Capabilitiy to Predict Workload and Safety Levels in the TMA - Explorative Factor and Regression Analyses of Radar and Simulated Trajectory Data,” in *ATM Seminar*, 2013.

[12] EUROCONTROL, *Complexity Metrics for ANSP Benchmarking Analysis*. Brussels: ACE Working Group on Complexity, 2006.

[13] EUROCONTROL, *Airspace Complexity For Regulation Purposes*. EEC Note No. 13/2008, 2008.

[14] European Commission, *EU-wide target ranges for RP 3 - Annex 2. Air Navigation Service Providers: Advice on benchmarking of ANSPs and EU-wide cost targets*. Brussels: Performance Review Body, 2018.

[15] EUROCONTROL, *OneSky Online*. Brussels: ACE Working Group, 2020.

[16] EUROCONTROL, *Traffic Complexity Score Dataset*. Performance Review Unit: http://ansperformance.eu/references/dataset/Traffic_Complexity_Score.html (10.02.2020), 2020.

[17] T. Standfuss, “Performance Benchmarking in Air Traffic Management - Methodology, Analysis and Evaluation of current ATM concepts,” PhD-Thesis, Technische Universität Dresden, 2020 (submitted, to be published).

[18] T. J. Coelli, P. D. S. Rao, C. J. O’Donnell, and G. E. Battese, *An Introduction to Efficiency and Productivity Analysis*. New York: Springer, 2005.

[19] T. Standfuss, F. Fichert, and M. Schultz, *Input and Output measurement in Air Navigation Service Provider Performance Benchmarking - Implementing composite indicators for efficiency analysis using European data*. Seoul: Air Transport Research Society Conference (ATRS), 2018.

[20] R. M. Arnaldo, V. F. G. Comendador, R. Barrangan, and L. Pérez, *European Air Navigation Service Providers’ Efficiency Evaluation Through Data Envelopment Analysis (DEA)*. St. Petersburg: International Council of Aeronautical Sciences Conference (ICAS), 2014.

[21] NERA, *Cost Benchmarking of Air Navigation Service Providers: A Stochastic Frontier Analysis*. London: , 2006.

[22] L. Simar and P. W. Wilson, “Estimation and inference in two-stage, semi-parametric models of production processes,” *Journal of Econometrics*, vol. 136, pp. 31–64, 2007.

[23] J. M. Wooldridge, *Econometric Analysis of Cross Section and Panel Data*. Cambridge: MIT Press, 2002.

[24] J. Rosenow, H. Fricke, and M. Schultz, “Air Traffic Simulation with 4D multi-criteria optimized trajectories,” in *Proceedings of the Winter Simulation Conference*, 2017, pp. 2589–2600.

[25] J. Rosenow, D. Strunck, and H. Fricke, “Trajectory Optimization in Daily Operations,” *CEAS Aeronautical Journal*, 2019.