Applicability of Current Complexity Metrics in ATM Performance Benchmarking and Potential Benefits of Considering Weather Conditions

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

Judith Rosenow
Institute of Logistics and Aviation
Technische Universität Dresden
Dresden, Deutschland
judith.rosenow@tu-dresden.de

Abstract—**Air traffic control is one of the most important stakeholders in the air traffic system and responsible for a safe and efficient air traffic management and traffic flows. Developments in air traffic and changes in the market structure induces the necessity of performance evaluation in air traffic management, done by institutional and academic studies. One of the key influencing factors on air traffic productivity and efficiency is expected to be the air traffic complexity, which contributes to the workload of air traffic controllers and thus limit the airspace capacity. The Performance Review Unit of EUROCONTROL developed a metric to assess the air traffic complexity in order to subject complexity in benchmarking activities. This paper analyses the calculation and the applicability of this complexity score by applying different methods and considering multiple operational levels. It appears that the complexity score contains an unexpectedly high correlation to other metrics (for example, adjusted density), while expected correlations, for example to productivity, weather phenomena or the airspace structure used, do not occur. These significant shortcomings tend to interpret the complexity score as a tendency only.**

Keywords—ATM, Complexity, Benchmarking, Performance, Weather in ATM

I. Introduction

Performance evaluation in air traffic management is addressed by numerous institutional [1] [2] and academic studies, e.g. [3] [4]. Primarily, these investigations focus on performance indicators (PI), such as operational productivity and cost-efficiency in order to benchmark and compare a set of Air Navigation Service Providers (ANSPs). An ANSP is a public or a private legal entity and manages air traffic on behalf of a company, region or country. Therefore, one or more outputs (e.g. total number of flight hours) are set in relation to those resources used to provide the goods or services (e.g. the number of Air Traffic Control Officer (ATCO) hours to control those flights). The main purpose of this kind of benchmarking is to identify best practices of ANSPs, to set cost targets, and to analyse endogenous effects, such as operational differences, characteristics in the air space structure, heterogeneous demand or typical weather phenomena. Furthermore, exogenous and partially exogenous effects and their possible significant influence on the performance are investigated. One of the key factors is represented by traffic complexity, which contributes to the ATCO workload and thereby inhibit capacity.

In 2006, EUROCONTROLS Performance Review Unit (PRU) developed a metric to assess the complexity induced by airspace users, to carry out exactly this benchmarking in the European context [5] [6]. Furthermore, several studies dealt with a possibility to quantify complexity based on traffic characteristics, such as quantity and distribution [7] [8] [9]. Although there are many new approaches, official EUROCONTROL reports still stick to the PRU score, based on the calculation method created in 2006. This PRU score is still used to explain differences in performance scores of ANSPs and to cluster Area Control Centers (ACCs) for productivity ranking [1]. In contrast to ANSPs, ACCs are air traffic control centres which control Enroute air traffic in a specific allocated airspace (the so-called Flight Information Region, FIR). ACCs are divided into those for lower airspace (up to about 8 km, flight level (FL) 245) and those for upper airspace, called Upper Area Control Centre (UAC). Additionally, the score was used to set performance targets for Reference Period 3 [10]. After using the economic model of purely comparing outputs and inputs of an ANSP, PRU weighted the output of each ANSP with their corresponding complexity score. The calculated performance gaps represent operational and economic efficiency scores and are thereby to be distinguished from horizontal or vertical flight efficiencies. In consequence, the metric may have a major impact on operational- and policy-decision-making. This method is only applicable, if the complexity score considers all relevant inputs with significant impact on the complexity.

From an operational perspective. the applicability and comprehensiveness of this metric have not yet been verified. Neither the suitability of the metric for weighting and clustering has yet been investigated, nor the completeness and the dimensioning of all essential inputs has been scrutinised. In this paper, several shortcomings regarding method and model which might affect the meaningfulness of the score has been identified.

II. The PRU Complexity score

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 Enroute airspace based on controller workload. Therein, it is assumed, that the main contributors to complexity are:

- **Air traffic density** as number of aircraft of a traffic flow per route unit at a time

978-1-7281-9825-5/20/$31.00 ©2020 IEEE

- **Air traffic flow** as number of aircraft crossing a given traffic area or line (as a boundary case of area) per unit of time
- **Vertical traffic** as number of aircraft with a positive or negative vertical speed
- **Air traffic mix** as composition of different aircraft types

Other complexity contributing factors like sectorization or the local route structure are not mentioned to be exogenous assumed and thus not reflected in the score. Those contributions are to be considered under the control of the ANSP. Furthermore, military traffic and the interface with adjacent units are considered to have an influence but are not considered due to missing quantification.

In order to calculate complexity, PRU divides the European airspace into cells of equal size. The used geographic projection ensures that all cells have the same volume. The dimensions of one cell are fixed to 20x20 nautical miles x 3000 feet within the range of FL85 to FL385. The size of the cells enables the assignment of each cell to an individual ACC or ANSP. Cells that cover multiple ANSPs/ACCs are designated to the one where the center point is located. To avoid boundary effects, the grid is shifted four times horizontally and three times vertically. Subsequently, the final score reflects an average of twelve values.

The score is based on 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. From this follows, that the presence of two aircraft in one cell within one hour results in two interactions (a to b and b to a) and three aircraft, analogous, lead to six interactions. The traffic data is represented by the initial flight plan, based on CFMU data. This assumption covers a worst case scenario, because relating on the initial flight plan, the actual position of each aircraft with a specific cell within one hour cannot be predicted more accurately.

Based on this assumption, PRU calculates the hours of the interactions h [h]. Therefore, the time in which the interacting aircraft (t_a and t_b) spend in the cell is calculated first. Thus, the expected duration d [h] of one interaction i is represented by the product of both times. This term is multiplied with the number of interactions in this cell. Equation 1 represents the duration h for two aircraft spending five minutes (1/12 h) in the cell

$$h = i \cdot (t_a \cdot t_b) = 2 \cdot \frac{1}{12} \cdot \frac{1}{12} = 0{,}014 \qquad (1)$$

The calculation is performed for each aircraft pair in the specific cell. The sum of the durations h provides the hours of potential interactions for that cell.

The complexity score of the PRU consists of four indicators, representing the most influencing traffic characteristics:

- Adjusted Density AD [-]
- Vertical Interactions as Vertical Different Interacting Flows, VDIF [-]
- Horizontal Interactions as Horizontal Different Interacting Flows, HDIF [-]
- Speed Interactions as Speed Different Interacting Flows, SDIF [-]

The density of traffic is usually defined by the number of vehicles in a defined area or volume. In an aviation context, it is expressed by the number of aircraft per airspace unit. However, the distribution of aircraft in the airspace might be spatially concentrated. The division of airspace into cells (see 2.2) enables the consideration of traffic accumulations, leading to the '*adjusted'* density. The key underlying rationale is that complexity tends to increase when aircraft are not evenly distributed in the airspace. The adjusted density value AD is calculated by sum of all durations h of interaction divided by the number of total flight hours f of the ACC or ANSP

$$AD = \frac{h}{f} \qquad (2)$$

Due to the grid structure, only cells with interactions contribute to the overall density score, which thereby is referred to be "adjusted".

The Vertical Interactions i_v cover flights in different flight phases (climbing/ cruising/ descending). No vertical interaction is considered for aircraft with the same attitude (e.g. two climbing aircraft). An aircraft is cruising if the vertical speed is below 500 ft/minute. The corresponding indicator (*VDIF*) sums up the hours of vertical interactions and divide it by the number of flight hours (Equation 3).

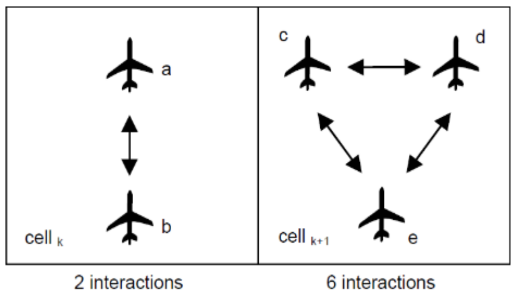

Fig. 1: Horizontal interactions between two or three aircraft [6]

$$VDIF = \frac{i_v \cdot d}{f} \qquad (3)$$

Aircraft with a difference in heading of more than 20 degrees are considered in the Horizontal Interactions i_h. Fig. 1 illustrates those Horizontal Interactions: Aircraft a is interacting with aircraft c, but not with aircraft b. Furthermore, (a) is considered to interact with (d) and (e). That means aircraft (a) has three potential interactions. The corresponding indicator (*HDIF*) is the ratio of the sums of horizontal interaction hours and flight hours (Equation 4).

$$HDIF = \frac{i_h \cdot d}{f} \qquad (4)$$

The speed interactions consider aircraft with a difference in speed of more than 35 knots. Since the score only relies on filed data, the aircraft speed is assumed the speed is estimated using BADA performance tables [11]. The sum of speed interaction hours is divided by the sum of flight hours to calculate the corresponding indicator (*SDIF*), as shown in Equation 5.

$$SDIF = \frac{i_s \cdot d}{f} \tag{5}$$

While adjusted density is addressing a measure of concentration, *VDIF*, *HDIF*, and *SDIF* represent flow characteristics. However, due to the mathematical background (see Equations 2-5), the flow indicators are a subset of the adjusted density and therefore there is a high correlation between both categories. In order to avoid this effect, PRU divides the flow-indicators by the adjusted density. The sum of the normalized horizontal, vertical, and speed component is defined as structural index s (see Equation 6).

The overall Complexity score c of an ANSP or an ACC is represented by the product of structural index s and the adjusted density (Equation 7) [12]. Thus, it is possible not only to evaluate which unit is complex but also if the complexity is basically due to density or due to flow characteristics.

$$s = \frac{VDIF}{AD} + \frac{HDIF}{AD} + \frac{SDIF}{AD} \tag{6}$$

$$c = AD \cdot \left(\frac{VDIF}{AD} + \frac{HDIF}{AD} + \frac{SDIF}{AD}\right) = \frac{(i_v \cdot d) + (i_h \cdot d) + (i_s \cdot d)}{f} \tag{7}$$

Fig. 2 shows the complexity scores for European ANSPs in 2019 [13], based on Equation 7 and sorted by complexity. The six FABEC ANSPs belong to the units with the ten highest complexity scores. Note, the highest complexity score is achieved by the Swiss ANSP Skyguide, which is 21 times higher than the complexity of the ANSP of the Republic of Moldova. MOLDATSA. In 2016, Skyguide managed approximatively 1,198,663 flights which is around 3,285 flights per day on average. In contrast to MOLDATSA, who guided 40,944 flights in 2016, on average 112 flights per day. The most contributing factor on the Complexity score is the adjusted density *AD*, as shown in Table 1. For most of the ANSPs, in particular for those with high traffic demand (listed at the top of the table), the structural index s only plays a minor role.

III. Interdependency between Performance and Complexity

A. Gate to Gate Perspective

The complexity score of PRU was introduced to measure the exogenous effects of traffic behavior on the workload of ATCOs [6]. Subsequently, there should be an interdependency between the metric and the PRU productivity score of the ANSP. Productivity is a Key Performance Indicator (KPI) of an ANSP and defined by the number Instrument Flights Rules (IFR) flight hours dived by the number hours of ATCOs in operations. Hence, both metrics complexity and productivity depend on the number of aircraft within an airspace and on flow characteristics, as described by *VDIF*, *HDIF* and *SDIF*.

The key underlying rationale is, assuming two identical spatial units concerning shape and size: the higher the complexity, the higher the workload and the lower the capacity, thus the lower the productivity. In the first step, the study investigates whether this negative interdependency between complexity and productivity could be proved.

The dependency between productivity and complexity score is tested for 2017, based on data provided by PRU Dashboard (Complexity Scores) and the Onesky ACE Working Group Teamsite (Operational Data). PRU productivity is determined by the sum of 'Composite Flight Hours' per ATCO-hour [2] [1].

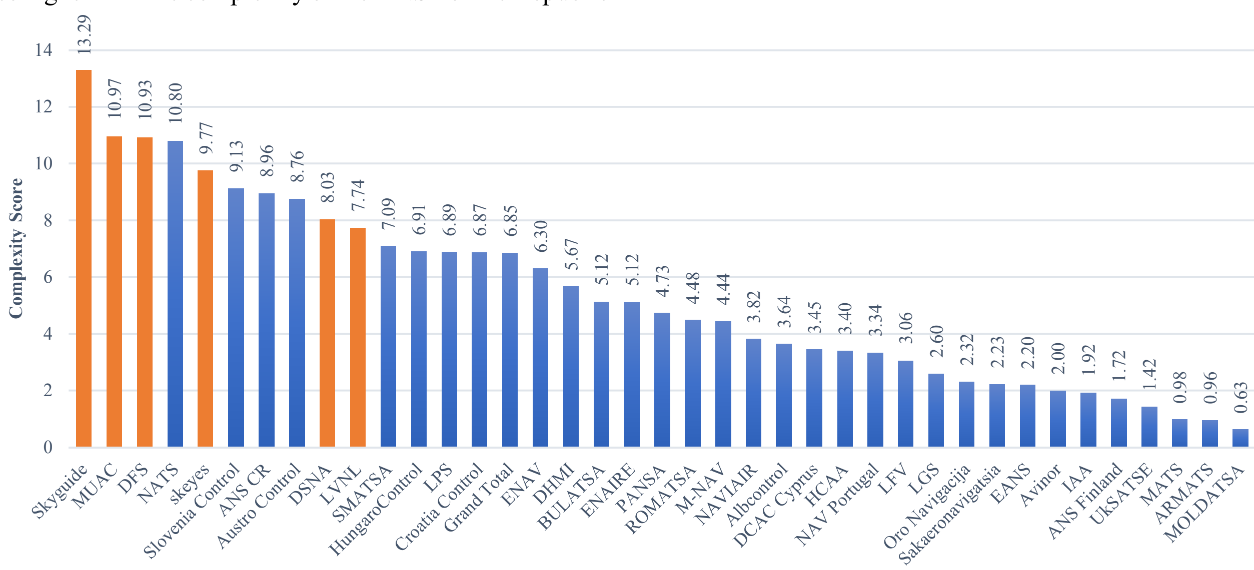

Fig. 2: Complexity Score of European ANSPs in 2019.

Table 1. Impact of the contributing components on the Complexity Score for a selection of ANSPs.

ANSP	Adjusted Density	Vertical Score	Horizontal Score	Speed Score	Structural Index	Complexity Score
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

In Fig. 3 both metrics of the identical time period and investigation area were subjected to a regression analysis. Although the trend is not significant, it seems that productivity increases with complexity, which is rather unexpected. However, the slope of the linear trend near zero, may suggests, that there might be no interdependency at all. Furthermore, the statistical significance is weak, since the coefficient of determination is $R^2 = 0.27$. These findings might be influenced by the ACC MUAC, which is an outlier (top right) in the dataset. However, eliminating the unit does not change the overall picture, Still, there is a positive dependency visible. An exponential trendline leads to a higher coefficient of determination but is still positively correlated. The analyses were executed for several years and led to equivalent results.

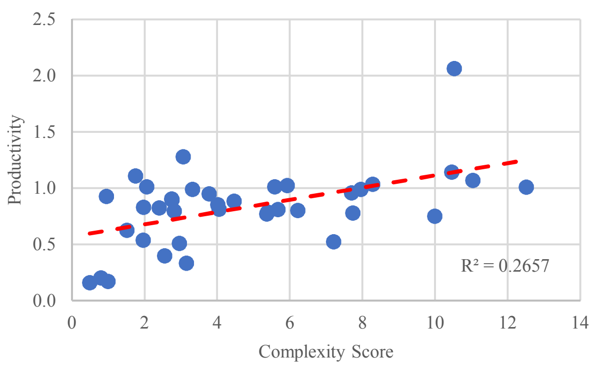

Fig. 3: Interdependency between Complexity and Gate-to-Gate productivity.

B. Enroute Services

The reason for the unexpected results might be that the complexity measure is primarily aimed to assess Enroute traffic regarding complexity [6]. Using a gate-to-gate perspective (as in the previous section), terminal services are included in the productivity score, which may lead to the missing interdependency. In consequence, the complexity score is compared with a metric for Enroute productivity only. Therefore, the total controlled flight hours were divided by the ACC ATCO hours.

The highest productivity scores for Enroute services are achieved by PANSA, NAV Portugal, and MUAC. As shown in Fig. 4, the trend still has a slightly positive slope, indicating a non-negative interdependency. The slope is (again) near zero and the coefficient of determination is about $R^2 = 0.05$. In comparison with overall productivity, there is even less evidence regarding an interdependency between complexity and Enroute productivity.

The strong dependency between adjusted density and complexity may lead to the assumption, that the PRU complexity score is primarily triggered by the overall demand. This conjecture is supported by Figure 8, which illustrates the dependency between the controlled flights and the PRU complexity score. A higher coefficient of determination indicates a statistically significant interdependency using a potency formula.

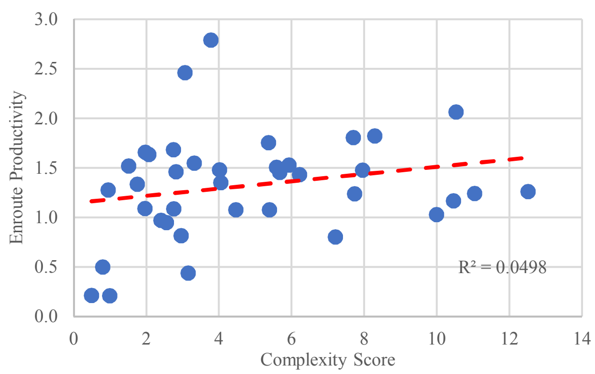

Fig. 4: Interdependency between Complexity and Enroute productivity.

C. Regresssion Analysis

In the previous sections, no evidence was found regarding an interdependency of PRU complexity and productivity. This might be due to three reasons.

- First, the two-dimensional analysis scheme might not be sufficient since multiple factors might influence productivity and capacity.
- Second, the productivity measure might not be sufficient as a performance indicator. ANSPs uses multiple resources to produce multiple outputs. Subsequently, the performance score should consider all elements of economic value creation.

- Third, the complexity score itself might be not sufficient. This section addresses the first and second opportunity.

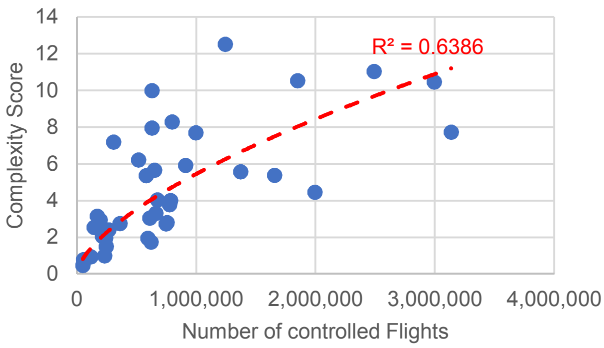

Fig. 5: Dependency between demand and complexity score.

In order to test influencing factors on performance, [14] ran several regressions. This type of quantitative analysis is able to identify functional interdependencies between one or more dependent variable(s) and multiple independent variables. The dependent variable is represented by the performance score of an ANSP or ACC. The assumed influencing factors represent the independent variables [15].

For the dependent variable, there are different ways to calculate productivity and efficiency. The PRU metric represents an index-number method, dividing one output (Composite Flight hours) by one input (ATCO hours). More appropriate methods like Data Envelopment Analysis (DEA) or Stochastic Frontier Analysis (SFA) consider multiple inputs and outputs [16] [17]. The cited study tested both measures: PRU productivity and the efficiency score based on DEA.

The different approaches and measures require different regression methods. For PRU productivity, 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 in the case of DEA. Alternatively, censored and truncated regression methods were applied. For methodological background, see [18].

Furthermore, a panel analysis, considering the data of multiple periods have been investigated because of two advantages. First, there are more data points available. Second, it is possible to determine temporal effects (e.g. how overall efficiency has changed over the years). For panel analysis, influencing factors were quantified using fixed effect (FEM) and random effect models (REM) as well as pooled regression [19].

The big advantage of applying multiple regression is the consideration of cross effects between influencing factors. The regression model estimates the performance score based on the used independent variables. The selection of factors is crucial for a valid result: Taking insignificant factors into account or not taking significant factors into account may result in an omitted variable bias (OVB) leading to an over- or underestimation of the factors concerning their influence on performance. This is exactly what would happen if only complexity were considered as an influencing factor.

D. Disaggregation

As shown in Fig. 3 and Fig. 4, there is no interdependency between PRU complexity and ANSP productivity. Since the complexity metric is calculated for the ACC level as well, this section intends to investigate whether there is an interdependency on this operational 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 provides operational data for 62 ACCs. The highest productivity scores are achieved by Warszawa, Lisbon, and MUAC, the lowest by Chisinau, Kyiv, and Dnipropetrovs'k.

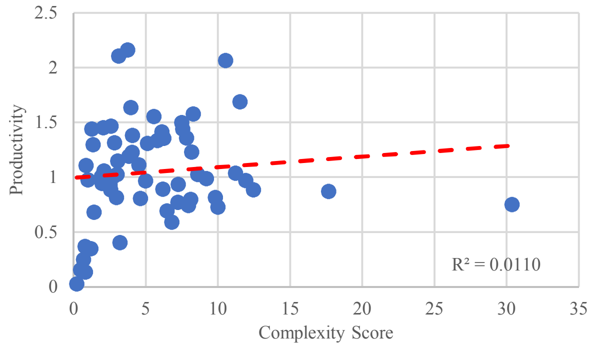

Fig. 6. Interdependency between Complexity and Enroute productivity on ACC level.

Complexity data was provided by PRU on request. FABEC ACCs 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.

The dependency between complexity and productivity on ACC level is shown in Fig. 6. Analogous to the aggregated level, no interdependencies are discernible. The slope of the trend line is positive but close to zero. Both the rise and the low coefficient of determination suggest that no correlation can be proven for ACCs. The weak regression proofs that the trend line is influenced by two extreme values with very high complexity scores, belonging to Istanbul and London TC. However, excluding London TC and Istanbul from the dataset leads to a similar result.

Since there is a high influence of *AD* on the complexity score, it might be assumed that the complexity score is mainly determined by the demand itself. Therefore, the interdependency between the complexity and controlled flight hours is shown in Figure 11.

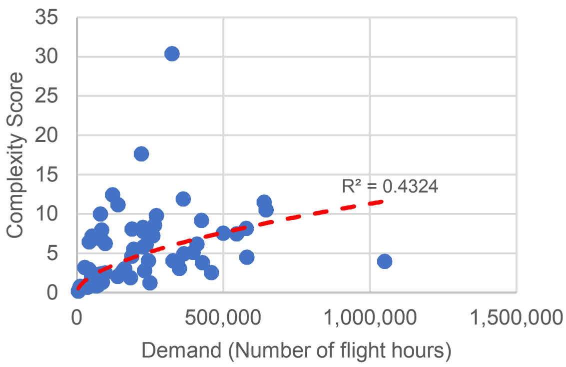

Fig. 7. Interdependency between demand and complexity on ACC level in 2017

The coefficient of determination of the exponential estimator yields $R^2=0.43$. The correlation might be clearer when eliminating London TC and Istanbul (with the highest complexity scores), who distort the regression due to terminal services.

The result is consistent with the observation in Fig. 4 and proofs that the PRU complexity score mainly depends on the demand, regardless of whether expressed in flight hours or flights. This is mainly because traffic density has a relatively large share in the score. Since the size of each cell is constant the density increases due to higher demand and thus does the complexity score.

IV. Assessment of the PRU complexity score

The PRU methodology represents one approach to measure airspace complexity for benchmarking purposes. The method itself has not been further developed since the year of implementation (2006). This section is intended to address the methodological strengths and weaknesses of the PRU approach.

PRU differentiates complexity dimensions into two essential components: density and flows. The flows are further differentiated into a horizontal, vertical, and speed component. The scores are applied on a grid of cells, which are shifted horizontally and vertically. The procedure and mathematical calculations are valid and enable the aggregation of scores to a specific area (ANSP or ACC). Furthermore, the shift of cells avoids boundary effects.

A. Analysis of the input data in the PRU complexity score

The speed interactions are based on aircraft type-specific BADA performance tables. Therewith, the score implies that all aircraft that all aircraft always follow the manufacturer's recommendations. Flow-related ATC requirements are neglected. Subsequently, changes in speed of a specific aircraft are also not considered. Therefore, the speed score might inhere some inaccuracies. Furthermore, a threshold of 35 knots indicating a speed interaction might be seen as too high, especially for ANSPs which mainly serve lower airspaces. However, as shown in Table 1, the speed score has only a minor influence on the total score.

It is not clear how the single flights are identified to avoid double counting. Aircraft that are in the same cell for two consecutive periods of time, might be double-counted if they are not clearly identifiable. Especially in cells with high traffic density, this contributes to inaccuracy. However, this is mainly influenced by the type of aircraft and the current flight phase. Assuming a cruising flight, an aircraft would need approximately three minutes or 20 NM (cell edge) to four minutes or 28 NM (cell hypothenuse) to cross the cell, not considering changes in height which might extend this time. In lower airspaces, the time extends to five (20 NM) or seven (28 NM) minutes. For the last case, all flights which are located in one cell e.g. between 00:53 and 01:06 would be double-counted (other hours analogous).

Table 1 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 normalized by this value (see Equation 6). In consequence, the influence of a heterogenous distribution might be underestimated for small entities (such as MOLDATSA) and overestimated von large ANSPs (such as DSNA). From this follows, a clear differentiation between the contribution of density and flow to complexity is hard to evaluate. For example, differences in the structural index for MUAC (higher *s*) than for LVNL (lower *s*) are not explainable. The ACC LVNL is responsible for the lower airspace only. But the vertical score of LVNL is lower than the vertical score of MUAC, although the share of cruising traffic of LVNL is 7%, but 54% for MUAC (NEST Data). Subsequently, the share of vertical traffic is much higher for the Dutch ANSP.

The past sections showed the missing interdependency between performance indicators and PRU complexity scores on different operational levels. In consequence, this section is intended to investigate the shortcomings of the score. It discusses significant factors contributing to ATCO task load or flexibility in the capacity provision, which directly affects productivity performance as well. Furthermore, the overall approach and meaningfulness are addressed. In a first point, the assumptions of PRU are checked, which were described in [6].

B. Missing Input Varaibles in the PRU Complexity Score

The lack of correlation between PRU complexity and productivity and results of the structural index *s* that are difficult to follow for some entities indicate missing input variables in the PRU complexity score. The traffic variability and predictability are not considered in the complexity score, although both measures hamper the ATCO's productivity.

Military traffic and adjacent units are not considered in the PRU complexity score, because there is no possibility of quantification. This might be the case for the year 2006. Nevertheless, the currently available databases enable an approximation. For example, neighbouring units can 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). It is assumed that military traffic contributes to complexity significantly. This limitation is probably due to the use of planning data as well.

C. Analysis of the Assumptions in the PRU Complexity score

The PRU complexity score clearly distinguishes between exogenous and endogenous factors. This is a good assumption, since the score should not contain any internal heterogeneities, mainly because productivity additionally depends on human factors (e.g. individual capabilities of the controller). Furthermore, route structure and sectorization are assumed to be endogenous. However, especially the route structure is influenced by the demand (and its preferences) itself and might be considered as partly exogenous.

Maybe for reasons of data processing, the PRU complexity score relies on the planed flight data. Well known deviations of the actual flights from the filed route are not considered. This fact creates a further inaccuracy with regard to speed. The use of real flight tracks and the True Air Speed (TAS) contained therein would allow the consideration of wind speed and wind direction in the complexity score. Especially deviations between planned and actual traffic (e.g. due to a change in flight intentions by the airline or the provision of directs in an upstream sector) pose challenges for capacity planning. It might be assumed that these factors are not considered due to the data (initial flight plan), which leads to several other shortcomings (see next subsection).

As described in [6], the measure is primarily intended to depict Enroute complexity. Subsequently, the method is not applicable to terminal complexity. In consequence, the score should not be used to run a gate-to-gate performance benchmarking, as was the case in [10].

Furthermore, the calculation of score is based on flight plan data and the performance calculation is carried out in advance. It seems much more useful to calculate complexity scores retrospectively, based on actual trajectories. The benefits are:

- The inclusion of more contributors such as weather, military traffic, and volatility would be possible.
- The score should be based on actual conflicts, not potential conflicts.
- The score might be applied to terminal service as well.
- The methodology might be applied on sector level, which increases accuracy.

Note, actual conflicts mean interactions of aircraft which cause a close observation or intervention. Potential conflicts include all interactions, even those that do not require flow measures.

Due to the method-internal weighting, the essential part of the complexity is attributed to density (see Table 1). The scores and components show: The higher the demand, the higher the effects of the adjusted density on the score. In consequence and as shown in Fig. 5 and Fig. 7, there is a strong interdependency between demand and complexity, which was also proven by the latest Performance Review Report [2] [20]. This may lead to the conclusion that the PRU complexity score can be substituted by the demand in relation to the airspace size or even only by the demand itself. However, the weighting is mainly influenced by the calculation (and thereby correlation) of the factors. Subsequently, there should be a revision of the methodology, in particular the calculation of the individual components. In fact, the workload of ATCOs is significantly influenced by conflicting flows [20].

As discussed above, the utilization of planning data (instead of actual flights) leads to several problems. A major limitation is the inclusion of important influencing factors, such as military traffic and weather conditions.

One main contributor to some ATCO's workload is military traffic. This workload is mainly driven by the number and opening times of TRAs because of a reduction of the available airspace for civil air traffic, for especially commercial airlines, and the possibility to provide directs. From this follows, TRAs 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 the capacity provision.

Complexity is further influenced by general aviation (GA) traffic. In general, this type of traffic uses lower airspaces (except business jets like Gulfstream V or Falcon 900X), 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 traffic uses visual flight rules primarily across uncontrolled airspaces but contribute to the ATCO workload when crossing controlled airspaces, e.g. near airports. GA pilots often have less routine in communication with ATC, leading to higher frequency occupancy time. The trajectories are less predictable and less accurate since pilots navigating visually and control the aircraft manually. Demand by medical aviation has to be considered as well, which has to be prioritized. Finally, transport operations to or from oil platforms sometimes require 1:1 service. In the future, there might be a further contributor, represented by Unmanned Air Vehicles (UAV), such as drones.

Whether these demand characteristics influence the complexity and thus on performance is the object of future investigations.

D. Impact of Weather on Complexity

Complexity strongly depends on environmental conditions [21]. In Enroute traffic, flows are designed following optimal conditions of windspeed and wind direction. Convective clouds and turbulent areas, if predicted, must be avoided often without a reliable prediction of size and shape of these areas [22] [23]. From this follows, flight crews usually adjust their routing on a short term. Therewith, the flight crew requests ATC to react vertically and horizontally on the changes in the intensity of convective or turbulent areas. These requests increase the ANSP complexity, but are only considerable, if complexity is measures retrospectively.

This procedure strongly depends on the situational awareness of the flight crew and on onboard technical issues. First, the weather radar information and/or the weather prediction must be correctly. Second, the routing decisions the flight crew does depending on the updated weather information must be correctly. And third, the request to change the route must be communicated accurately. The possibilities of rerouting further depend on the aircraft altitude and flight phase, the aircraft performance limitations and on the onward routing of the aircraft.

The deviating flight routes are considered as unanticipated traffic especially when a crossing of an unplanned sector is required and may lead to several consequences for ATC, such as:

- Reduction in available airspace
- New conflict points
- Increased frequency occupancy time
- Increased manual coordination
- Degradation of the Reduced Vertical Separation Minima (RVSM) capability
- Limited applicability of radar vectoring

Hence, the difference between initially filed and rerouted trajectory strongly contribute to the airspace complexity and could be easily considered in PRU complexity score by including the number of rerouted flight hours per filed flight hour or the rerouted distance per filed distance in the structural index s (Equation 6), if complexity is analysed retrospectively. Thereby, aircraft flying directs should be given more emphasis. Furthermore, the speed interactions SDIF should be weighted by the mean difference between TAS and groundspeed.

In Terminal Manoeuvring Area (TMAs), the complexity is amplified by strong windspeeds, 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 for holdings. In consequence, significant weather alerts always lead to a lower predictability and to a higher short-term volatility [24] [14] [25].

V. CONCLUSIONS AND WAY FORWARD

This paper investigated methodology, sub-models and components of the complexity score according to EUROCONTROL's Performance Review Unit PRU. The metric was examined for validity and applicability. The methodology and the composition have been derived in 2006 and have not been adjusted until today, although the data availability increased in recent years. It was shown that there is no statistical evidence for a correlation between complexity and productivity. This result was constant across methods at both ANSP and ACC level. Even the use of the individual components (horizontal score, vertical score, etc.) represent statistically significant variables in a few cases only. This can be interpreted as a lack of quality of the complexity measure for performance benchmarking purposes.

We further discussed potential improvements regarding the components to be considered in a complexity metric which are easily to be implemented and do not need further data sources. We analysed a change in the data source from plan data to actual data which leading to a significant improvement on several levels of accuracy. This allows the consideration of potential influencing factors, such as weather or military traffic. Furthermore, fluctuations and the volatility in demand could be mapped in this way, GA and pop up flights could be considered both contributing significantly to the ATCO's workload.

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. Since it has been shown that 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. The findings of state-of-the-art studies on components and calculation methods should be taken into account. Updating the assessment of complexity could further improve the performance benchmarking itself, e.g. with regards to the target setting for RP4.

The authors are going to implement both the productivity score and the complexity score into the air traffic simulation environment TOMATO [26] at Technische Universität Dresden in order to improve the complexity score step by step. Thereby, actual flight data will be considered additionally. Furthermore, the suggested improvements for considering weather conditions will be considered.

REFERENCES

[1] EUROCONTROL, "Air traffic management cost-effectiveness (ACE) benchmarking report for 2017," PRU, Brussels, 2019.

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

[3] 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,* p. 50–60, vol. 77 2015.

[4] T. Blondiau, E. Delhaye, N. Adler and S. Proost, "Productivity Measurement of Air Traffic Management in Europe," in *Competition for Air Traafic Management COMPAIR*, 2016.

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

[6] EUROCONTROL, "Complexity Metrics for ANSP Benchmarking Analysis," ACE Working Group on Complexity, Brussels, 2006.

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

[8] J. Djokic, "Investigation into Air Traffic Complexity as a Driver of a Controller's Workload," TU Dresden, PhD Thesis, 2014.

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

[10] 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," in *Performance Review Body*, Brussels:, 2018.

[11] A. Nuic, D. Poles and V. Mouillet, "BADA: An advanced aircraft performance model for present and future ATM systems," *International Journal of Adaptive Control and Signal Processing,* p. 850–866, vol. 24 2010.

[12] EUROCONTROL, "Traffic Complexity Score Dataset - Data description," PRU, Brussels, 2020.

[13] EUROCONTROL, “OneSky Online,” ACE Working Group, [Online]. [Accessed 2020].

[14] T. Standfuß, “Performance Benchmarking in Air Traffic Management - Methodology, Analysis and Evaluation,” PhD Thesis, Technische Universität Dresden, 2020.

[15] J. Welc and P. Esquerdo , Applied Regression Analysis for Business: Tools, Traps and Applications, Ney York: Springer, 2018.

[16] W. Cooper, L. Seiford and J. Zhu, Handbook on Data Envelopment Analysis, Ney York: Springer, 2011.

[17] S. Kumbhakar and C. A. K. Lovell, Stochastic Frontier Analysis, Cambridge: Cambridge University Press, 2003.

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

[19] J. Wooldridge, Econometric Analysis of Cross Section and Panel Data, MIT Press, 2002.

[20] M. Meinecke, “Development and Evaluation of ATCO Task Load Models in a Fast Time Simulation Environment,” DLR German Aerospace Center, Braunschweig, 2014.

[21] J. Rosenow, H. Fricke and M. Schultz, “Air Traffic Simulation with 4D Multi-criteria optimized trajectories,” in *Winter Simulation Conference* , Las Vegas, USA, 2017 .

[22] J. Rosenow, D. Strunck and H. Fricke, “Free Route Airspaces in Functional Air Space Blocks,” in *8th Sesar Innovation Days*, Salzburg, 2018.

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

[24] FABEC, “Volatility Task Force Final Report,” Deutsche Flugsicherung, Langen, 2018.

[25] T. Standfuss, M. Whittome, I. Ruiz-Gauna and F. Knabe, “Volatility in Air Traffic Management – How changes in traffic patterns affect efficiency in service provision,” in *EIWAC Conference*, Tokyo, 2019.

[26] J. Rosenow, S. Förster, M. Lindner and H. Fricke, “Multicriteria-Optimized Trajectories Impacting Today’s Air Traffic Density, Efficiency, and Environmental Compatibility,” *Journal of Air Transportation,* vol. 27, no. 1, 2018.

[27] J. Rosenow and M. Schultz, “Coupling of Turnaround and Trajectory Optimization based on Delay Cost,” in *Winter Simulation Conference* , Gotenburg, 2018.