CHALLENGES IN EUROPEAN AEROSPACE
5TH CEAS AIR & SPACE CONFERENCE

CEAS

Context-Dependent ATC Complexity Metric

Gustavo Mercado
PhD Candidate
*G.A.MercadoVelasco**

Clark Borst
Assistant Professor
*C.Borst**

Delft University of Technology
Faculty of Aerospace Engineering
Address
Delft, The Netherlands
**@tudelf.nl*

ABSTRACT

Several studies have investigated Air Traffic Control (ATC) complexity metrics in a search for a metric that could best capture workload. These studies have shown how daunting the search for a universal workload metric (one that could be applied in different contexts: sectors, traffic patterns, controllers, or ATC tasks) can be. We propose that complexity metrics should be tailor made to the task they were meant to be used for, and focus in the elicitation of parameters that could best capture complexity of the task at hand. For the ATC task of rerouting aircraft a selection of relevant context parameters was made, and based on these a new complexity metric called SSD Composite is proposed. The metric is based on the Solution Space Diagram (SSD) concept. As a pilot study, a low fidelity simulation in which ATM experts operated as controllers and provided self assessments of workload was conducted. Data obtained from the simulation runs was used to compare traditional metrics (Traffic Load and NASA/FAA's Dynamic Density) against the SSD Composite metric by means of mixed-effects linear regressions. Results showed that SSD-based metrics were much less sensitive to sector or traffic pattern effects and that the amount of observed workload variation accounted for by the SSD Composite fixed effects was higher than that of the Dynamic Density metric.

1 INTRODUCTION

Air Traffic Controller (ATCo) workload is considered to be one of the main limitations in the capacity of the current Air Traffic Management system. The limited options available for implementing changes in this system and their possible effects on ATCo workload have made it subject of many investigations.[1] One of these topics is the search for traffic complexity metrics that can be representative of ATCo workload, which finds its motivation in the reduced costs of implementation when compared to physiological workload monitoring stations, its non obtrusive nature, and its possible implementation as an evaluation metric for ATC sector design or as a real time metric that can be used by automated aids.

Several sector traffic and complexity metrics have been proposed for estimating ATCo workload, being Traffic Load and Dynamic Density by far the most studied ones[1]. Reliable prediction of ATCo workload based on objective metrics has, however, showed to be a great challenge. Hilburn [1] elaborated on how complexity metrics developed so far depend heavily on the sector in which they've been developed, making its direct use in other sectors unreliable for workload estimation. Hilburn [1]

concludes by saying that no complexity indicator is context-free. This statement can take us to further conclude that a complexity indicator should be context-dependent.

The problem is the fact that context in ATC can be described very broadly. A universal ATC workload metric might prove to be a daunting research goal, but a metric based on the ATC task it is meant to serve (the task context) seems more feasible. In that sense, the focus of this study is the ATC task of rerouting aircraft and sheds a light on the issue of developing workload metrics for context-dependency.

In order to further progress in the refinement of complexity models, perhaps not only should they be applied in various types of airspace (as mentioned by Hilburn [1]), but also different new task load indices that have the potential to capture contextual information regarding the difficulty of the task at hand should be explored. A complexity metric that has shown its potential to capture task difficulty is one based on the Solution Space Diagram (SSD). The SSD is a two-dimensional representation that covers all heading/speed combinations possible for a specific aircraft, indicating which velocity vectors offer *safe solutions* and which velocity vectors lead to an impending conflict with another aircraft [2]. Hermes et al. [3] and van Paassen et al. [2] have shown that the area that represents *unsafe solutions* in the SSD has a significant correlation to subjective workload ratings. Hence, apparently, the SSD effectively "captures" elements of the environmental context that predict task difficulty. This research investigates its possible use as context-dependent ATC workload metric.

The purpose of this study was to identify task context related variables that best capture the difficulty of the ATC task of rerouting traffic. The research question that we pretended to answer was: does the use of SSD-based indices in complexity metrics enhance the correlation with subjective assessments of workload? Evaluating generalization across subjects, sectors or traffic samples is a topic that has not received much attention in the body of literature. Such assessment was also a part of our research contribution.

The paper is structured as follows. Section 2 provides a background on previous related work. Section 3 explains the methodology followed to elicit context-relevant parameters and describes the complexity metric we propose. The experimental procedure for evaluating the proposed metric is detailed in Section 4, and its results are discussed under Section 5, to finalize with the discussion and conclusions in Section 6.

2 PREVIOUS RELATED WORK

Several studies have shown a strong relationship between complexity factors and controller workload (cited by Hilburn [1]: Hurst & Rose, 1978; Stein, 1985; Grossberg, 1989; Laudeman et al., 1998). A wide range of benefits are possible with this relationship between workload and a set of objective indices. For example, non intrusive real time evaluation of workload is a valuable asset under dynamic sectorization concepts. The (re)design of ATC sectors is often based on estimates of the workload the new design would induce in controllers and the evaluation of safety issues. Another area that can certainly benefit from these estimates is the implementation of levels of automation, under which the computer assumes more responsibility and control of air traffic than what is nowadays available or even allowed (due to safety issues).

The list of complexity indices studied by researchers is extensive. The literature review performed by Hilburn [1] can be seen as a quite complete compilation of such indices, from which we take as

reference the two most studied ones: Traffic Load and NASA/FAA's Dynamic Density.

Traffic Load (some studies call it Aircraft Count, others call it Traffic Density) is perhaps the most studied task load index. It is simply defined as the number of aircraft currently under control. The literature has shown that it can be a good indicator of complexity that, however, is not able to capture the richness of what controllers find complex (cited by Hilburn [1]: Kirwan et al., 2001; Mogford et al., 1995; Athenes et al., 2002).

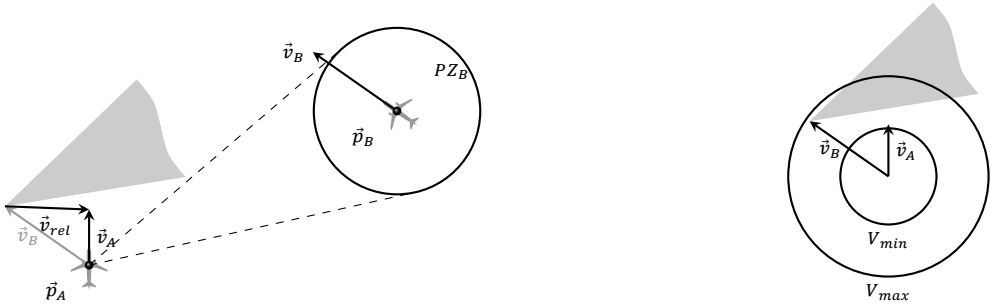
The NASA/FAA's Dynamic Density metric [4] is a composite of traffic load and several traffic complexity indices that has been developed and validated with operational data, showing to be highly correlated with workload. Nonetheless, a big shortcoming of the metric is the fact that the factor weightings, obtained by means of multiple regressions, may not show the same correlation levels for the sectors in which they were not collected.

The high dependency of complexity metrics to the sector in which they were experimentally tested is a general problem. In fact, no task load index (or composite of indices) can be equally applied for different ATC situations [5]. Based on these and other studies, Hilburn [1] concludes his literature review by stating that no complexity indicator is context-free. This statement can take us to further conclude that a complexity indicator should be context-dependent. In order to further progress in the refinement of complexity models, perhaps not only should they be applied in various types of airspace (as mentioned by Hilburn), but also different new task load indices that have the potential to capture contextual information regarding the difficulty of the task at hand should be explored.

If we take the task contextual information into consideration, the solution space diagram (SSD) is perhaps a complexity index capable of capturing the difficulty of the conflict detection and/or resolution of a two-dimensional ATC traffic situation. Having its foundations in the Velocity Obstacle theory [6], the SSD is a two-dimensional representation that covers all heading/speed combinations possible for a specific aircraft, indicating which velocity vectors offer *safe solutions* and which velocity vectors lead to an impending conflict with another aircraft [2]. Hermes et al. [3] and van Paassen et al. [2] have shown that the area that represents *unsafe solutions* in the SSD has a significant correlation to subjective workload ratings. Hence, apparently, the solution space effectively "captures" elements of the environmental context that predict task difficulty.

Mercado Velasco et al. [6] elaborate in detail on the construction of the SSD. The simple diagram shown in Figure 1 can be used to have an idea of the SSD concept. Consider a controlled vehicle A and an observed (moving) obstacle B with circular protected zone PZ_B, as shown in the Figure. A *relative velocity cone* can be constructed from the traffic geometry, with its edges originating in A, and tangent to PZ_B. By adding the velocity of B to the relative velocity cone, and then drawing it in the velocity plane, a simple *solution space diagram* is created. Note that the large and small circles represent the aircraft performance limits, i.e., the maximum and minimum speed, respectively.

Several studies have shown a strong relationship between workload and a SSD-based metric [2, 3, 7], even though an important caveat on the metric definition can be observed. The metric can be described as the ratio of the area between the minimum and maximum speed circles in the SSD covered by velocity cone(s) to the entire area between the minimum and maximum speed circles, averaged across all aircraft under control. This implicitly assumes that each aircraft under control has an equal contribution to workload, which cannot be true since aircraft that just entered the sector cannot pose the same level of difficulty as aircraft involved in a conflict that must be solved in a very short period of time.



(a) Velocity Obstacle for A, imposed by aircraft B ($VO_{A|B}$). (b) Solution space diagram for aircraft A.

Figure 1: Basic solution space construction.

The literature explains that a single task load index (or composite of indices) cannot be equally applied in every single situation. We believe that research should be instead done aiming at task load metrics tailored to the task it is meant to evaluate, and based on the identification of relevant task context parameters. The following section elaborates on the such identification methodology, tailored to the ATC task of rerouting aircraft.

3 CONTEXT-BASED COMPLEXITY METRIC

As discussed previously, we propose that complexity metrics need to be based on indices that strongly depend on the task context in which they were meant to serve, so that as much context information is captured as possible. Important questions that arise are related to the identification of such context parameters, and how to measure their ability to capture context information.

This section focuses on providing a list of objective parameters that can be deemed as relevant aspects of the ATC task of rerouting aircraft for capturing controller workload. Based on this analysis, a simple composite metric is proposed, and is, in following sections, compared against widely known complexity metrics. This comparison will test evaluate the correlation of all metrics to subjective assessments of workload. A description of relevant task context parameters is provided first, to end with the proposed complexity metric equation.

3.1 Identification of relevant context parameters

Huang and Gartner [8] have elaborated on the selection of relevant context parameters and its use in context-aware applications. They mention that in order to identify relevant parameters, every subtask should be studied in terms of its specific goals and requirements. In that sense, the Control Task Analysis (CTA) is perhaps the best starting point for eliciting the information required for this analysis. The CTA is part of a methodological framework developed within the discipline of Cognitive Systems Engineering. The purpose of the CTA is helping to better understand the actions required for completing those control activities needed for realizing the domain's functional purposes (in our case, these functional purposes are *Route planes efficiently*, and *Route planes safely*), by making use of an analytical tool called the Decision Ladder [9].

We start by discussing the CTA of the task at hand, mention all parameters that we deem as relevant for the estimation of workload, and we then propose a selection of the most important.

In a previous study, Kilgore et al. [10] have performed a CTA on the same controlling goal that falls within the scope of this project. The study provides a list of information processing steps that need to be completed in order to achieve the main controlling goal:

1. Scan for aircraft presence in area of responsibility.
   The goal of this initial subtask is identify the aircraft that are (or soon will be) under control.
2. Determine future flight vector for each aircraft.
   The goal of this task is determining which aircraft within the area of responsibility have intersecting flight paths.
3. Predict future, time-based location states for aircraft on convergent paths.
   In this task, the operator intends to determine whether converging aircraft will arrive at point of convergence within a similar time frame. It can be thought of a filtering stage for saving cognitive resources.
4. Determine criticality of pending convergence.
   Here the controller establishes whether or not the future distances between converging aircraft will constitute a loss of separation. This becomes a distance calculation exercise, based on extrapolated trajectory paths.
5. Choose to modify aircraft flight path(s) to address future problem.
   The goal of this subtask is identifying which aircraft flight path(s) must be modified to eliminate a potential loss of separation.
6. Select specific strategy for accomplishing rerouting of aircraft.
   The goal of this subtask is achieving a desired aircraft rerouting strategy.
7. Convey flight modifications to aircraft for execution.
   The final goal here is to communicate the desired changes to the relevant aircraft pilots.

Further analysis on the information needed to successfully complete these subtasks reveals relevant context parameters. For example, different type of converging tracks (same, reciprocal or crossing tracks, as defined in ICAO Doc. 4444) can be associated with different resolution strategies, acording to the difficulty of handling them. The separation at – and time to – Closest Point of Approach (CPA) should also be considered as relevant parameters.

A monitoring tool that makes use of CPA calculations and type of converging tracks has been developed by NATS [11] and is called *Separation Monitor*. In it, aircraft pairs are being displayed in a Cartesian plane having on the horizontal axis the time to CPA ranging from 0 to 10 min, and in the vertical axis the separation at CPA ranging from 0 to 15 NM. Such selection of the time span and distance range should also be a point of attention and deemed as relevant parameters. Another study [12] made use of a controlled simulation environment to evaluate the probability of a controller to identify an aircraft pair as being conflictive as a function of separation at CPA, time to CPA, and angle of convergence. They've found that even with a separation of 10 NM at CPA, a 30% of chance of identifying the pair as conflictive was present when time to CPA was around 8 min. Establishing the right values for the time span and distance range for the CPA calculations is important for capturing workload at the most relevant moments in time: when and how far from each aircraft conflict detection and resolution is generally performed by controllers.

Based on all this information that we consider relevant for capturing the complexity of the rerouting task, we propose the following complexity metric.

3.2 Traffic factor selection

A total of four air traffic complexity factors were identified from the contextual information provided in the previous section. These factors were selected to capture the complexity of four major subtasks: identification of the aircraft pairs on which conflict detection should be executed (factor 1); conflict detection (factor 2); conflict resolution (factor 3); conveying flight modifications (factor 4).

The following traffic complexity factors were identified:

1. Number of intersecting pairs (IP) - The number of aircraft pairs that have a time to CPA of more than zero and less than 10 minutes, with a separation at CPA of less than 15 NM; and having both of the aircraft in every pair inside the sector under control, or one aircraft and the location at CPA of any aircraft inside the sector under control.
2. Average SSD covered area for conflict detection ($SS10$) - The average covered area of the SSD diagram of all intersecting aircraft pairs (time to CPA > 0 min), having a time to CPA of less than 10 min, and a separation at CPA of less than 10 NM.
3. Average SSD covered area for conflict resolution ($SS5_S$, $SS5_R$, $SS5_C$) - The average covered area of the SSD diagram of all intersecting aircraft pairs (time to CPA > 0 min) having a time to CPA of less than 10 min, a separation at CPA of less than 5 NM, averaged across different types of converging tracks (same - $SS5_S$, reciprocal - $SS5_R$, or crossing - $SS5_C$).
4. Heading change (HC) - The number of aircraft that made a heading change of greater than 15 degrees during a sample interval of two minutes. This factor was extracted from the Dynamic Density metric developed by Laudeman et al. [4]. Similar to that study, the current experiment required a considerable amount of vectoring to solve conflicting pairs; which, according to Laudeman et al., explains why this parameter received the highest weight in the multiple regression they performed.

We combine these factors into a metric we call SSD Composite:

$$SC = \beta_1(IP) + \beta_2(SS10) + \beta_3(SS5_S) + \beta_4(SS5_R) + \beta_5(SS5_C) + \beta_6(HC) \quad (1)$$

The following section elaborates on the experimental setup used for evaluating this metric against workload measures.

4 EXPERIMENTAL DESIGN

A body of literature shows that several complexity metrics are analyzed by means of linear regressions to evaluate their correlation to either subjective ratings of workload or activity counts elicited from human-in-the-loop experiments [1, 2, 3, 4]. Random elements (like subject-related or scenario-specific characteristics) can however influence large part of the variation explained by a regression equation, influencing the correlation levels and possibly providing wrong estimates of the

metric's ability to capture complexity. Laudeman et al. [4] reported, for example, that the dynamic density metric was a multiple regression exercise across several TRACON facilities, and showed to have accounted for 50% of the total variance in air traffic controller activity. They do not provide, however, an estimate of how much of this variance can be explained due to differences across facilities, controllers, or period of sampling.

Based on these findings, the current study attempts to identify the amount of variation explained by such random effects by making use of linear mixed-effects models; and compares the amount of variance explained by the fixed effects of a number of complexity metrics. The results were obtained from a human-in-the-loop low-fidelity simulation in which controllers had to issue speed change or vector commands in order to solve conflicts.

4.1 Sectors' layout

In order to evaluate the effect of the differences between sector layouts and traffic samples, two different sectors (shown in Figure 2) each with 4 different traffic samples were simulated. The main goal of the simulation was to safely clear aircraft to their respective exit point without performing any changes of altitude.

(a) Sector 1. (b) Sector 2.

Figure 2: Experimental sector design.

Both sector had three streams of incoming traffic, but laid out in different ways:

- *Crossing points*: Sector 1 had three crossing points clustered near the sector border, while Sector 2 had two crossing points with ample spacing between them.
- *Converging tracks*: Sector 1 had routes with intercept angles of 45°, 90°, and 120°, while Sector 2 had intercept angles of about 90°.
- *Sector shapes*: Sector 1 had an area approximately 30% smaller than Sector 2.

4.2 Subjects and instructions

A total of nine male subjects with ages between 29 and 51 ($\mu = 37$, $\sigma = 8.6$) participated in the experiment. None of them were air traffic controllers, but all of them had received an ATC introductory course and had therefore similar basic level experience.

Subjects were instructed to safely clear all aircraft (with a separation of 5 NM) to their designated sector exit points (provided on their labels) without making use of altitude changes, but making sole use of speed and/or heading adjustments. In order to aid them with the estimation of the minimum separation, every time an aircraft was selected a 5 NM radius protected zone was displayed around it.

Some color coding was used in order to provide some more queues to support the controllers in their task:

- selected aircraft were colored white;
- aircraft that had not yet received any clearance at all or a clearance that would not lead them to their exit point were colored gray; aircraft that had received a clearance towards their exit point were colored green;
- aircraft pairs with a time to loss of separation of 3 minutes were colored red;
- a magenta circle around a selected aircraft indicated the intended speed; and
- a magenta line shown on selected aircraft indicated the intended heading.

Aircraft would only accept new clearances when inside the controlled sector. Clearances were given by selecting the relevant aircraft and dragging the heading line with the mouse to a new heading and/or scrolling the mouse scroll wheel up or down for speed changes. To confirm the new clearance, the Enter key had to be pressed.

ISA ratings of workload had to be provided every 60 seconds by clicking on a scaled bar that changed gradually from 0 (no workload) to 100 (full workload). This scale appeared on the top side of the display and was presumed to be less intrusive than typing a number on a keyboard.

4.3 Independent variables

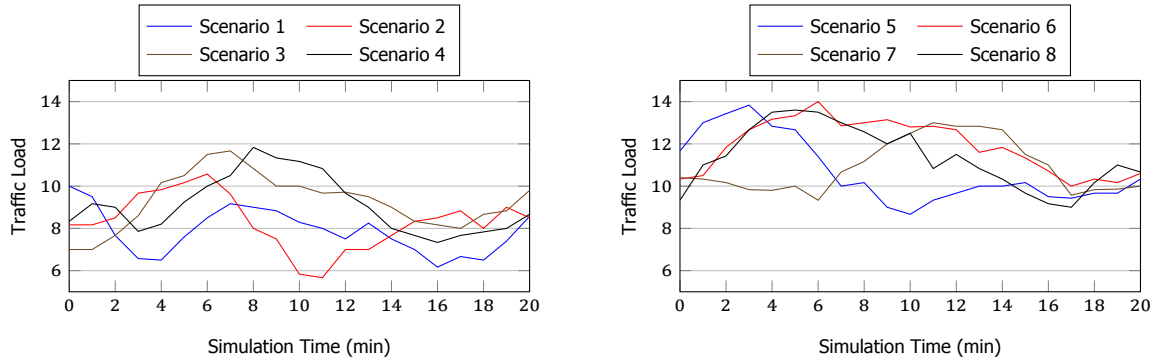
Both sectors were simulated with four different traffic sequences. These were treated as eight scenarios. The median Traffic Load of these eight scenarios is shown in Figure 3. In order to counter balance any learning effects, all scenarios were randomly presented to all subjects.

4.4 Dependent variables

Six different measures of workload were extracted from the simulated scenarios. Three of them were based on the Dynamic Density metric as proposed by Laudeman et al. [4]. The reader is referred to that article for more information on the complexity indices included in the metric. Those indices are mentioned here just for explanation purposes.

The following metrics of workload were extracted:

1. Traffic Load (TL) - The number of aircraft under control;
2. Average SSD covered area (SSD) - The ratio of the area between the minimum and maximum speed circles in the SSD covered by velocity cone(s) to the entire area between the minimum and maximum speed circles, averaged across all aircraft under control.



(a) Sector 1.

(b) Sector 2.

Figure 3: Traffic load per scenario.

3. Dynamic Density unit weighted:

$$DD_{UW} = (HC) + (CP40) + (CP70) + (MD5) + (MD10) + (AC) + (TD)$$

4. Dynamic Density regression weighted, based on the weights reported by Laudeman et al.:

$$DD_{RW} = 2.17(HC) + 1.85(CP40) + 1.85(CP70) + 1.02(MD5) + 1.18(MD10) + 0.88(AC) + 0.79(TD)$$

5. Dynamic Density with new regression weights:

$$DD'_{RW} = \beta_1(HC) + \beta_2(CP40) + \beta_3(CP70) + \beta_4(MD5) + \beta_5(MD10) + \beta_6(AC) + \beta_7(TD)$$

6. SSD Composite metric, copied from Equation 1:

$$SC = \beta_1(IP) + \beta_2(SS10) + \beta_3(SS5_S) + \beta_4(SS5_R) + \beta_5(SS5_C) + \beta_6(HC)$$

The regression weights for metrics 5 and 6 were calculated ex-post following the methodology explained in Section 5.

4.5 Procedure

All subjects were initially briefed on the experiment objectives and the simulator they were going to use. Two training scenarios (each 10 minutes long) allowed them to learn how to make use of the low-fidelity simulator. Subjects then performed in eight randomly selected scenarios (mentioned earlier in this section), each with a total duration of 25 minutes and run 4 times faster than real-time. During all simulation runs, participants had to provide and estimate of their workload using a scale that appeared on top of the simulator screen.

5 RESULTS

In the current experiment, different sources of variability had to be taken into consideration. Furthermore, the multiple observations per subject narrowed down the number of analysis methods that could be applied due to the failure of satisfying the assumption of independence of samples for parametric methods. Linear mixed-effects models (extensions of linear regression models) are powerful and useful approaches to account for varying sizes of experimental units, multiple sources of variation that can be modeled as random variables that follow a predefined distribution, or correlations between multiple observations in repeated measures studies [13]. It was therefore the analysis method of choice for our experimental study.

All task load metrics extracted from the simulation were fitted into a linear-mixed effects model of the form:

$$y_{ijk} = \beta_0 + \sum_{h=1}^{p} \beta_h x_{hijk} + \gamma_k + \alpha_j + \epsilon_{ijk}, \quad (2)$$

where y_{ijk} is the ith ISA workload rating provided by the jth subject when performing on the kth scenario, β_0 is the regression intercept, β_h is the regression weight of the hth predictor[1], x_{hijk} is the ith value of the jth subject when performing in the kth scenario for the hth predictor, γ_k is the scenario-specific effect from a normal distribution of scenario-specific effects with mean of zero and variance of σ_γ^2, α_j is the subject-specific effect from a normal distribution of subject-specific effects with mean of zero and variance of σ_α^2, and ϵ_{ijk} is the residual from a normal distribution of scenario-specific effects with mean of zero and variance of σ_ϵ^2.

Note from Equation 2 that an intercept β_0 was used in the model as part of the fixed effects. Additionally, a random intercept γ_k to account for scenario-specific effects, and a random intercept α_j to account for subject-specific effects were taken into consideration as sources of uncorrelated variability that follow a normal distribution.

After fitting all metrics into the model, the standardized residuals were used to identify outliers. Specific observations that would fall outside of a normal distribution where tested with the following criteria: 5% should fall outside 1.96 standard deviations, 1 % outside 2.58 standard deviations, and 0.1% outside 3.29 standard deviations. No observation was filtered out from the analysis since, as shown by Figure 4, only a few number of observations fell outside of these limits.

All considered metrics showed to have passed the tests of model assumptions. The normality of the standardized model residuals was confirmed by mean of histograms and quantile plots.

As goodness-of-fit indicators, the amount of "variance explained" (R^2) and the Akaike Information Criterion (AIC) were calculated. Results are depicted in Figure 5. The SSD complexity metric showed to explain the least amount of observed variance, followed by the TL metric. On the other side, the DD'_{RW} showed to explain the highest amount of variance, closely followed by the SC metric we propose.

[1]Note from Section 4.4 that the TL, SSD, DD_{UW}, and DD_{RW} metrics have a single predictor; while DD'_{RW} and SC have 7 and 6 predictors, respectively.

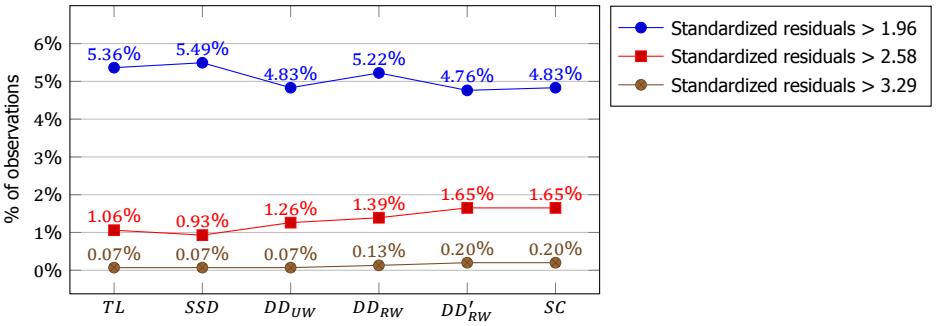
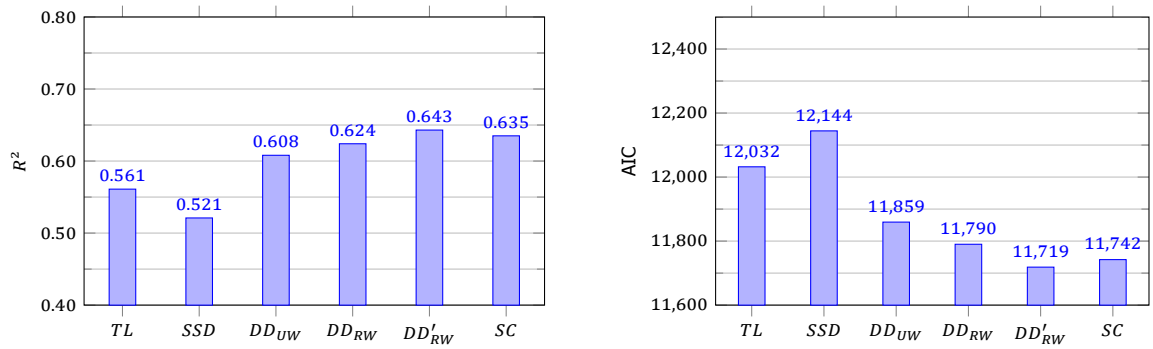


Figure 4: Normal distribution analysis on standardized residuals.

Interestingly, all Dynamic Density variations showed to explain more than 60% of the observed variance, while the results published by Laudeman et al. [4] indicated an explained variance of 50%. This might be attributed to the fact that several TRACON facilities, and therefore controllers from very different population groups, were considered by Laudeman et al. in their experiment.

The overall increase in explained variance obtained when fitting new regression coefficients (the DD'_{RW} metric) instead of using the regression weights published by Laudeman et al. (the DD_{RW} metric) was of 1.9%. On the other side, the amount of variance explained by the new metric we propose (SC) was 0.8% less than the one obtained with the newly fitted coefficients (the DD'_{RW} metric). These results show only minor differences in the total amount of variance explained by these top three metrics.

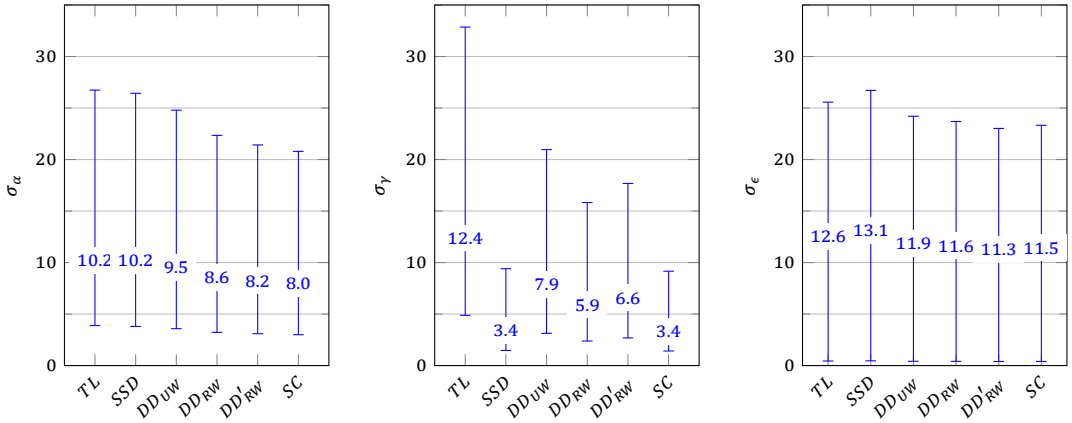


(a) Total variance explained. The calculation is based on the most common definition based on the residual sum of squares and the total sum of squares [14].

(b) Akaike Information Criterion (AIC) comparison. Lower values represent models with better fit.

Figure 5: Goodness-of-fit indicators for the different complexity metrics.

All considered workload metrics showed significant variance in intercepts across subjects and scenarios. The estimates of the standard distribution of these random effects are shown in Figure 6. The modeled random variations induced by subject-specific effects showed to be the least under the proposed metric (SC) and the Dynamic Density metric with new regression coefficients (DD'_{RW}), in that order. The random variations induced by scenario-specific effects showed to be the least under both Solution Space Diagram-based metrics (SSD and SC), and the highest under the TL metric. The random residual variations showed to be the least under the SC and DD'_{RW} metrics.



(a) Standard distribution estimates of intercepts across **subjects**.

(b) Standard distribution estimates of intercepts across **scenarios**.

(c) Standard distribution estimates of **residuals**.

Figure 6: Standard distribution estimates of random effects.

Based on the estimates of the standard distribution of the random effects, the amount of variance explained by the fixed effects was calculated as proposed by Nakagawa and Schielzeth [14]:

$$R^2_{(m)} = \frac{\sigma_f^2}{\sigma_f^2 + \sigma_\gamma^2 + \sigma_\alpha^2 + \sigma_\epsilon^2},$$

$$\sigma_f^2 = \mathrm{var}\left(\sum_{h=1}^{p} \beta_h x_{hijk}\right),$$

where σ_f^2 is the variance calculated from the fixed effect components of the linear mixed model, m in the parentheses indicates marginal R^2 (i.e. variance explained by fixed factors). Note that σ_f^2 should be estimated without degrees-of-freedom correction. The calculation results are depicted in Figure 7.

The fixed effects of the TL and SSD metrics showed to explain the least amount of variance, while the SC and and DD'_{RW} fixed effects explained the most variance. The increase in variance explained by the fixed effects obtained when fitting new regression coefficients (the DD'_{RW} metric) instead of using the regression weights published by Laudeman et al. (the DD_{RW} metric) was of 3.7%. This was, however, further increased in 5.9% by making use of the new SSD-based metric we propose (SC).

Tests on collinearity between the SC metric predictors showed slight collinearity between variables IP and HC ($r = -0.311$), between IP and $SS10$ ($r = -.365$), and between IP and $SS5_C$ ($r = -0.362$). For all other fixed effect correlations $r < 0.2$. However, leave-one-out model comparison confirmed that collinearity did not affect any of the significant effects reported in Table 1.

Note from Table 1 that the predictor $SS5_R$ (average SSD covered area for conflict resolution of reciprocal converging tracks) has been removed from the model, since it showed not to be significant.

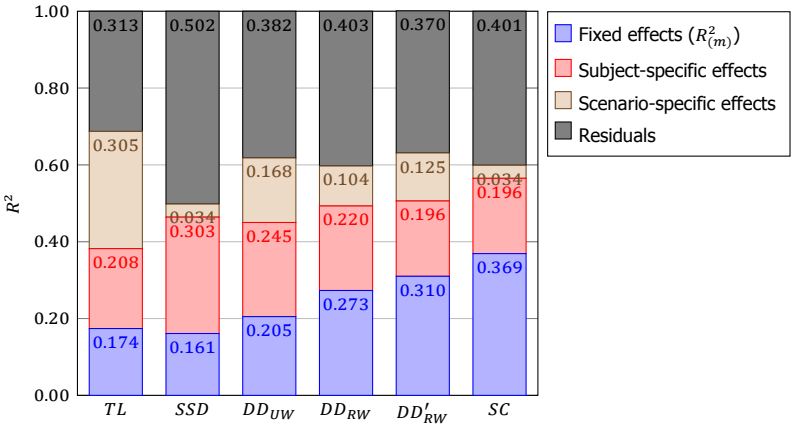


Figure 7: Amount of variance explained by the mixed-effects models.

Table 1: Parameter estimates of the Solution Space Composite (SC) metric.

Parameter	Estimate	SE	t Statistic	DF	p	95% CI	
Intercept	9.14	3.43	2.66	1506	**	2.41	15.88
IP	0.81	0.09	9.14	1506	***	0.64	0.99
$SS10$	11.56	4.80	2.41	1506	*	2.14	20.97
$SS5_S$	10.71	4.19	2.55	1506	*	2.48	18.93
$SS5_C$	6.67	1.89	3.53	1506	***	2.96	10.38
HC	5.01	0.30	16.48	1506	***	4.41	5.60

Note: * $= p < 0.05$, ** $= p < 0.01$, *** $= p < 0.001$.

Reciprocal tracks are opposite (or nearly opposite), with an angular difference of more than 135° and less than 225°. We propose that the lack of significancy of this predictor can be explained by the fact that it is not common practice to have aircraft flying towards each other with these angular differences due to safety reasons. In the current study, those angular differences only took place as part of conflict resolution maneuvers that were initiated between aircraft pairs that did not have opposite (or nearly opposite) tracks in the first place, but became so for short periods of time.

6 DISCUSSION AND CONCLUSIONS

This study proposed a complexity metric called SSD Composite, consisting of four simple task context related predictors that, to some extent, captured the complexity of the ATC task of rerouting traffic. Two of those predictors were based on the Solution Space Diagram (SSD) concept, showing that the use of SSD-based indices in ATC complexity metrics has the potential of enhancing the correlation with subjective assessments of workload.

A comparison was made between the SSD Composite metric and the popular Traffic Load and NASA/FAA's Dynamic Density metrics. These metrics were extracted from a human-in-the-loop simulation in which 9 test subject that had have an introductory ATC course participated. The use of linear mixed-effects regression models allowed to estimate how much of the variation observed in instantaneous self assessments of workload could be attributed to random effects (either subject-specific or scenario-specific), and how much of could be explained by the metric itself (the

fixed effects).

Results showed that Traffic Load was largely influenced by random sector and traffic-specific effects (fixed effects accounted only for 17% of total variance), strongly suggesting that it's use is conditioned to the recalculation of regression weights for any small changes in traffic sample or sector changes. The average SSD covered area, on the other hand, was barely influenced by these scenario-specific effects effects (only 3.4% of total variance), but the large variation of the residuals showed that it failed to capture the complexity of the performed task (fixed effects accounted only for 16% of total variance). We conclude that, as metric of workload, the averaged SSD covered area is as good as the Traffic Load metric, although recalculation of regression weights might not be need for small changes in sector and traffic sample.

The Dynamic Density metrics (unit weighted, with the regression weights reported by Laudeman et al. [4], and with new regression weights) proved to better capture the complexity of the ATC task of rerouting aircraft (fixed effects accounted for 21 to 31% of total variance); but could be somewhat less robust to sector or traffic changes with respect to the SSD-based metrics, since 11 to 17% of the total variance was accounted for by scenario-specific effects.

The SSD Composite metric we propose proved to have the best performance as workload metric. It accounted for 37% of the total observed variance, mainly because only 3.4% of the total variance was accounted for by scenario-specific effects and because the variation of the residuals was much lower than that observed with the averaged SSD covered area metric. This metric showed in this study to be more robust to random effects than the NASA/FAA's Dynamic Density metric. Based on these findings, we conclude that SSD-based metrics have a large potential that could be exploited by context-dependent metrics of workload.

Being this a pilot study, generalizing these results must be done with care. Even though several measurements were obtained (1512 samples), these came from a population of 9 test subjects performing in 2 different sector designs with 4 different traffic samples each. Furthermore, the subjects that participated in the experiment were not actual controllers, but were ATM experts from Delft University of Technology that had previously received an introductory ATC course provided by the Dutch air navigation services provider (LVNL). This research shows, however, that a study at a larger scale with actual controllers and other sources of random variation (like more sectors, traffic samples, or controllers from different training facilities/programs) could show more of the SSD-based metrics potential of capturing complexity in varying contexts.

Our research contribution lies not only in the practical exemplification of the design for context dependency of complexity metrics, but also in the application of statistical regression methods that have not yet been extensively practiced by ATM experts.

References

[1] Brian Hilburn. Cognitive Complexity in Air Traffic in Air Traffic Control - A Literature Review, 2004. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.195.5288.

[2] M. M. van Paassen, Jurriaan G. D'Engelbronner, and Max Mulder. Towards an air traffic control complexity metric based on workspace constraints. In *2010 IEEE International Conference on Systems, Man and Cybernetics*, pages 654–660. IEEE, October 2010. ISBN 978-1-4244-6586-6.

doi: 10.1109/ICSMC.2010.5641823. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5641823'escapeXml='false'/>.

[3] P. Hermes, Max Mulder, M. M. van Paassen, and J. H. L Boering. Solution-Space-Based Analysis of the Difficulty of Aircraft Merging Tasks. *Journal of Aircraft*, 46(6):1995–2015, 2009. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.174.1088.

[4] IV Laudeman, SG Shelden, R Branstrom, and CL Brasil. Dynamic density an air traffic management metric. Technical report, Ames Research Center, Moffett Field, California, 1998. URL http://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19980210764.pdf.

[5] B Kirwan, R Scaife, and R Kennedy. Investigating complexity factors in UK Air Traffic Management. *Human Factors and Aerospace Safety*, 1(2), 2001. ISSN 1468-9456. URL http://trid.trb.org/view.aspx?id=717648.

[6] Gustavo A.; Mercado Velasco, Clark Borst, Joost Ellerbroek, Max Mulder, and M. M. van Paassen. The Use of Intent Information in Conflict Detection and Resolution Models Based on Dynamic Velocity Obstacles. *IEEE Transactions on Intelligent Transportation Systems*, 2014. doi: 10.1109/TITS.2014.2376031. URL http://dx.doi.org/10.1109/TITS.2014.2376031.

[7] S.M.B. Abdul Rahman. *Solution Space-based Approach to Assess Sector Complexity in Air Traffic Control*. PhD thesis, Delft University of Technology, February 2014. URL http://repository.tudelft.nl/view/ir/uuid:abf9f743-ea5c-4995-9371-c0018150b0cd/.

[8] H Huang and G Gartner. Using activity theory to identify relevant context parameters. In *Location Based Services and TeleCartography II*, chapter 3, pages 35–45. Springer Berlin Heidelberg, 2009. ISBN 978-3-540-87393-8. doi: 10.1007/978-3-540-87393-8_3. URL http://link.springer.com/chapter/10.1007/978-3-540-87393-8_3.

[9] Jens Rasmussen, Annelise Mark Pejtersen, and L. P. Goodstein. *Cognitive Systems Engineering*. John Wiley & Sons, Inc., New York, NY, USA, September 1994. ISBN 0-471-01198-3. URL http://dl.acm.org/citation.cfm?id=179248.

[10] R. M. Kilgore, O. St-Cyr, and G.A. Jamieson. From Work Domains to Worker Competencies: A Five-Phase CWA for Air Traffic Control. In *Applications of Cognitive Work Analysis*, chapter 2. Taylor & Francis, 2009. URL http://scholar.google.com/scholar?cluster=3644880407501387269&hl=en&as_sdt=0,5#1.

[11] Melodi Irfan, Michael John Bull, Andrew Trevor Clinch, and Stephen James Pember. Air Traffic Control, 2012. URL http://www.google.com/patents/US20120303253.

[12] A. Vuckovic, P. Kwantes, and A. Neal. A dynamic model of decision making in ATC: Adaptation of criterion across angle and time. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 55(1):330–334, September 2011. ISSN 1071-1813. doi: 10.1177/1071181311551068. URL http://pro.sagepub.com/content/55/1/330.abstract.

[13] AL Oberg and DW Mahoney. Linear Mixed Effects Models. In Walter T. Ambrosius, editor, *Topics in Biostatistics*, chapter 11, pages 213 – 234. Humana Press Inc., Totowa, New Jersey, 2007. ISBN 978-1-58829-531-6. URL https://intranet.pasteur.edu.uy/publico/bonilla/Protocolos/mmb/404-TopicsinBiostatistics.pdf#page=220.

[14] S Nakagawa and H Schielzeth. A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, 2013. URL http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210x.2012.00261.x/full.