Analysis of Airspace Complexity Factors’ Capability to Predict Workload and Safety Levels in the TMA

Markus Vogel, Kati Schelbert, and Hartmut Fricke
Chair of Air Transport Technology and Logistics
Dresden Technical University
Dresden, Germany
{vogel, schelbert, fricke}@ifl.tu-dresden.de

Abstract—Research on airspace complexity metrics with the aim of characterizing traffic with respect to air traffic controllers’ cognitive workload has been ongoing for the last 40 years (complexity factors, dynamic density). Nevertheless, simple sector load based on aircraft count remains the key planning figure in ATM & ATFM. The proposed complexity factors lack a common model of causation (pre, peri, post controller intervention), which leads to case-specific workload dependencies. Motivated by EUROCONTROL’s research in Single European Sky and recent THALES advances studying the introduction of dynamic density metrics in an operational ATFM environment, this study analyzes the predictive capabilities of complexity metrics towards (1) controller workload and (2) the level of safety (collision risk). By means of average-filtering and time-shifting, a linear prediction model’s coefficient of determination was improved significantly. For simulated controller workload, a local optimum is situated at the current timeframe ($r^2 \approx 0.7$). For radar-data based collision risk, the best prediction looks 130 s into the future ($r^2 \approx 0.5$). This observation is coherent with the findings for simulation-based collision risk ($150 \text{s}, r^2 \approx 0.5$).

Keywords – terminal airspace complexity factors, dynamic density, human factors, work- and taskload, collision risk model, propensity, agent-based ATM simulation, factor and regression analysis

I. INTRODUCTION

A number of at least 22 distinct air traffic complexity factors (CF), originally defined with en-route sectors in mind, have been proposed during the last 40 years of research [11]. In general, they measure selected traffic characteristics representative of typical air traffic control tasks, targets, or challenges in a statistically straightforward and easy-to-evaluate way. While some target the need for controller intervention (leading, e.g. collision prediction), some target patterns representative of ATC advisories with compliance monitoring (accompanying, e.g. speed and heading changes), and others target patterns that persist after initiation by ATC (trailing, e.g. altitude transitions). Thus, the proposed complexity factors lack a common causation model, which leads to case-specific dependencies to controller workload (WL).

Most widely known and adapted, ‘Dynamic Density’ (DD) complexity formulae evaluate some typical selection of CF in a single weighted sum. There has been an appreciation for some time; however, that individual CF are separately of interest to air traffic controllers and flow managers.

In order to benefit from complexity factors in an operational environment, the problem of CF selection needs to be solved. A structured method of establishing a linear model with the CF of the highest entropy based on traffic samples (surveillance data) has been proposed in [8]. Subsequently, it has been shown that the entropy and hence the selection is dependent on the sector (type, layout), day (traffic load, traffic mix, and atmosphere) and potentially even the controller (individual factors) [9, 11]. Masolonis et al. went to some effort to construct a “generalized set of complexity metrics” [18] applicable across the four FAA centers considered in their study. However, the large variation in sector load complexity formulae present in the literature continues to illustrate the difficulty of finding a single, common metric acceptable in different airspaces, to different organizations, or even to different controllers within any one center.

Regardless of selection, the problem of presentation to the human operator persists. Traffic complexity is always a multi-dimensional and time-variant metric, making interface design an important part of any proposed solution. Masolonis et al. present a scheme to display “multidimensional complexity information” and to toggle this display with that of the standard aircraft count (sector load) [18].

Kistan et al. consider the diversity of proposed selection of CF from the perspective of an Air Traffic Management system vendor targeting customers in different countries [19]. They take a pragmatic, 3-step approach that firstly retains the concept of simple density (standard aircraft count) as a common, baseline indication of sector load. The second step is to annotate (rather than toggle) this measure with an indication of the complexity factors contributing to the load currently being evaluated. Figure 1 illustrates this concept – the length of each bar represents the aircraft count while the color-coded segments annotating the bottom edge of each bar indicates the relative weighted contribution of several specific complexity factors. A drill-down identifies the individual flights involved in each complexity factor. The final, crucial, step is to permit limited user modification of which complexity factors are used to annotate the display of simple density. This starts with the
offline definition of one or several dynamic density formulae, including their terms and weighting factors.

For some organizations this is likely to be the extent of their customization if they mandate standardized formulae for use in particular sectors or across entire centers based on their operational and safety use cases. Standardized formulae are useful for multi-sector planning – such as flow managers making reroute decisions to resolve sector congestion and supervisors planning sector consolidation and de-consolidation impacting several air traffic control positions.

Other organizations may permit controllers to make an online selection from amongst several offline-defined dynamic density formulae, referred to as “themes” in [19]. Individual themes are typically subsets of related complexity factors such as those relevant to aircraft density, those relevant to aircraft maneuvers, to their potential for conflict, to airspace structure etc. Hybrid themes may also be created that draw on several, unrelated factors in order to approximate traditional dynamic density formulae.

In this paper, we refine the CF selection process proposed in [8] and adapt it for the proposed human-in-the-loop modeling. With the objective function of modeling (maximized correlation to the reported workload) this method also has a sound internal validation criterion. This way, the process of developing “themes” (CF sets) gets a scientific foundation and a fully automated process. The development of a theme for high traffic demand is demonstrated here. To obtain further themes, the selection of reference data needs to be performed manually – the rest is a mere matter of computing resources and analytical skills.

The following aspects of the previous study [11] influence the objectives of this study: First, the possibility of different controllers’ individual factors (skills and experience, individual strategies) interfering with the model was not considered. In this study, traffic is also simulated (agent-based fast-time simulation) in this study, replacing humans with a standardized controller model which does not exhibit such characteristics. Second, no significant correlation between traffic complexity and collision risk was observed ($r^2 = 0.45$). The traffic throughput was well below the capacity limit, leading to the assumption that no significant workload peaks were present. In this study, the ten busiest operating hours were selected (surveillance data) and a scenario with increased traffic (130%) was simulated (fast-time agent-based simulation). Third, none of the previous studies considered the possibility, of a time offset between complexity (CF), controller workload (WL) and finally the level of safety (Propensity), not taking into account relevant theories of the human cognitive information processor (e.g. perception, attention, decision-making, as expressed in Hollnagel’s ‘Contextual Control Model’ [21]). Safety-relevant temporal effects of generating and implementing ATC advisories were studied in detail in [13, 14].
Considering the lead-lag inhomogeneity of CF, modeling a time offset in the correlation hypothesis is crucial for the CF selection. Fig. 2 delivers a brief outline: Airspace Complexity is evaluated over all known CF at any timeframe $t$. The controller’s reaction does not necessarily follow suit (limited attention, e.g. other tasks on hand, limited perception e.g. visual estimation, other possibly unknown factors, e.g. individual factors), leading to a time lag $\tau_0$ before the subjective measure of workload is even influenced. After the advisory has been issued, another time lag takes effect: $\tau_1$ denotes the time to implement the advisory onboard the aircraft until the reaction becomes visible to the surveillance technology.

II. Metrics Used in this Study

A. Airspace Complexity and Complexity Factors

Complexity factors (CF) describe airspace complexity from the perspective of the controller’s perceived workload. It is generally agreed that an increasing traffic volume must drive complexity [9], [8]; hence, there is an increased demand for traffic management commonly expressed as ATC-WL [8]. A significant amount of research regarding the field of airspace complexity has been undertaken in the past and present – especially in the scope of the Single European Sky – to find a way to quantify ATC-WL drivers and systematically increase safety [8], [9], [16]. As a result of this scientific work, 19 complementary, potentially overlapping complexity factors (CF), defined by various researches to describe airspace complexity in an analytical way [5] - [10], were consolidated as they were considered relevant to the studies performed (tab. I).

<table>
<thead>
<tr>
<th>No.</th>
<th>Description of Complexity Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>CF18</td>
<td>Speed disorder</td>
</tr>
<tr>
<td>CF19</td>
<td>Two dimensional speed disorder</td>
</tr>
</tbody>
</table>

B. Workload

The TMA sector controllers’ workload (WL) is influenced by numerous contributing factors, e.g. task load, individual (physical and mental) factors of the performing controller, etc. As proposed in the introduction, workload is best determined with real-time human-in-the-loop simulation experiments, see [8] for a detailed description. Some of the methodology and expertise gained in these experiments were integrated in Air Traffic Safety research at TU Dresden (TUD) and in parts of TUD’s agent-based fast-time simulation, which is briefly described in Sect. IV.

For this study, WL is based on 3 trace outputs from the agent-based fast-time simulation: (1) Radio channel utilization time [percentage delivered every 60 s], (2) ATC Agent task model’s idle proportion [percentage delivered every 10 s], and (3) ATC Agent task model’s deferred tasks [absolute count delivered every 10 s]. Metrics (1) and (2) are equivalent, but with another time base, while (3) is the opposite of (2). Tasks are only deferred if there is no time left; in this case (2) is zero. For the following analysis, (1), (2) and (3) are combined into a weighted sum (using the first principal component derived from a reference dataset) to represent WL.

C. Safety, Collision Probability, Propensity

Air Traffic Safety research at TU Dresden has a history of several years [1], [2], [15]. Amongst others, one result of this research is the implementation of collision risk model (CRM) which yields a collision probability estimate as an important ATM safety metric [1]. The collision probability is defined as the probability of a mid-air collision between two aircraft at a specific location in time. In this process, the magnitude of collision probability is driven by the actual navigation performance (ANP) of each aircraft. Implemented as a Java-based tool [1], it focuses on the TMA of major airports and uses ANP modeled after the results of detailed radar data analyses of approaching and departing traffic. The real deviation of aircraft from the nominal track was quantified for all three spatial dimensions and modeled in form of Gaussian tracking error probability distribution functions [3], referred to as along-track tolerance (ATT), cross-track tolerance (XTT), and vertical-track tolerance (VTT). As the best-fitting distribution function is standard-normal, all tracking tolerances are defined as a $2\sigma$ interval in line with the ICAO performance-based navigation concept, which requires aircraft to adhere to a defined $2\sigma$ tracking interval more than 95% of all time. Collision probability is evaluated by pairwise spatial integration of the location probability density functions (essentially following Reich’s work of the 1960s and ICAO Doc. 9274 [17]).

The term Propensity was conceived within the EUROCONTROL’s INTEGRA project [22] to describe a simplified version of the CRM as laid out above, compare [15]. Since the collision probabilities are usually very small, let Propensity denote the natural logarithm of the collision probability further.
III. APPROACH AND METHODOLOGY

A. Overview

This study follows the approach originally published in [8] and demonstrated in [11]. A Factor or Principal Component Analysis is used to select CF based on their contribution to total variance (and hence, their entropy). Fig. 3 gives an outline.

B. Selecting / Extracting Reference Data

From surveillance data (radar-recorded trajectories), reference data is extracted for specific operating conditions (representatives of the class). As laid out above, the CF selection is dependent on sector, day, and controller, so the results are highly dependent on the reference data selection. Depending on sector type, a broad selection of operating conditions might yield acceptable results, but our results show that for the TMA sector, the selection must be made in a fine-grained manner (taking traffic load and mix as well as wind conditions into account). Finally, this will be an economic decision in practice: the tradeoff is between system complexity (number of themes) and prediction quality (correlation with workload). By taxing theme generation, implementation, training and application of themes in daily operations with costs, a cost-benefit analysis becomes feasible. This process is shown in the topmost horizontal flow in fig. 3 (step 1).

C. Generating a Theme from Reference Data

Based on the reference data, the CF are first evaluated following their mathematical definitions. Redundancies in the CF dataset are then reduced by means of Principal Component Analysis (PCA), which yields new factors in a rotated coordinate system (the principal components). As the CF are inhomogeneous in definition and hence magnitude, standardization is mandatory before performing the PCA (z-scoring is used). The principal components (PC) are ranked by their contribution to total variance. A selection retaining as few as possible PC but also retaining as much variance as possible is made. Two scientific criterions exist for this tradeoff: (1) stop when the contribution to total variance is less than one (2) stop when the gradient of variance contribution drops to ‘almost horizontal’. Criterion (2) is usually augmented with so-called scree-plots. The selected PC are termed ‘super factors’, and a varimax rotation on the loadings matrix yields the contributing CF to each super factor for interpretation by a human (e.g. to create training material). Interpretability is crucial for the ‘drill-down’ to contributing CF and further down to contributing flights in the later operational system. This process is shown in the topmost horizontal flow in fig. 3 (steps 2-4).

D. Obtaining Sample Data

The sample data should be obtained in human-in-the-loop experiments, but fast-time computer simulations or surveillance data are valid options as well. While Propensity can be calculated with the trajectories alone, in most cases, obtaining WL metrics can be challenging (voice-com logs, instantaneous self-assessment ISA [8], biophysical measurements, or observation). This process links the topmost with the middle horizontal flow in fig. 3.

E. Analyzing the Predictive Power of the Theme

For the sample data, the CF and the safety metric Propensity are evaluated according to their respective definitions. WL is assumed to be present in the sample. Following the initial research hypotheses, the resulting data vectors are subjected to filtering: (1) averaging (sliding/moving average, sophisticated windowing functions) and (2) time shifting (on the allegedly dependent variables only). Using the filtered CF, the ‘super factors’ are evaluated using the reference PCA rotation matrix. Finally, a linear regression model CF → WL / Propensity is fitted using the filtered data vectors. The predictive power is expressed by $r^2$, the coefficient of determination (which is the square coefficient of correlation $r$ in case of the linear model). To maximize $r^2$, local or global optima are searched in the space of averaging and time shifting (3D, 2x averaging, 1x time shifting). In this study, the search space was defined first, and all data points were calculated ‘brute-force’. Optima are defined by their filtering parameters (to be applied with the theme in the later operational system) and the coefficient of determination (the internal validation criterion). This process is shown in the middle and bottom horizontal flows in fig. 3 (forming a loop).

IV. DATA SOURCE AND IMPLEMENTATION ISSUES

A. Radar Dataset

ADS-B transponder and flight plan augmented radar trajectories of aircraft within the TMA of Munich Airport, a typical representative of a Central European hub airport was used for this study. The recording is based on multiple primary and secondary radar systems integrated with the Flight Track and Aircraft Noise Monitoring System (FANAMOS) of the German ANSP Deutsche Flugsicherung GmbH (DFS) and was provided in a standardized ASCII format. The dataset spans a period of six months (May - October 2008) and contains 81,084 flight movements. The spatial resolution is < 1 m, the temporal resolution equals 4 s. There has only ever been an insignificant adaptation in airspace layout since the recorded year; and during busy hours, the runway configuration already served at maximum capacity. With this respect, the empirical foundation for this study is to be considered up to date and applicable for this study. Due to work-regulation issues under German jurisdiction, it is very difficult to obtain written or
audio logs of ATC voice communication with radar data (which would be extremely useful to evaluate ATC strategies or workload).

B. Agent-Based Fast-Time Simulation

The agent model comprises function allocation, standard operating procedures and operational performance (aircraft dynamics, human performance) relevant to today’s operations in the terminal airspace. It is characterized by asynchronous, time-discrete interaction events between intelligent agents and physical entities. A time-synchronous simulation of the physical entities (aircraft flying) drives the interaction process. Details on modeling and simulation as well as case studies can be found in [1], [12], [13], [14].

ATC-to-pilot communication is modeled as ‘verbal’ (with *scripted* text messages, e.g. ‘CS123, turn left immediately, heading X-Y-Z’). A dedicated *RadioChannel* Agent implements a blocking resource for all participants (to model today’s standard of R/T communication in ATC). This agent also determines the time needed to verbalize a given message with an estimator derived from the *jACT-R* implementation (50 milliseconds per syllable, typical English syllable length of 3 characters, 100 milliseconds between words) plus additions (300 milliseconds listening for a free channel, 150 milliseconds to formulate a sentence, spelling of callsigns and numbers, etc.). The radio channel utilization forms one component of the WL metric used for this study.

The *ApproachController* agent’s strategies and the resulting task model are described in detail in [1]. The task model is pre-emitive with a fixed cycle time (set to 10 s for this study). Cognitive tasks are *not yet* taxed with processing time; the verbal communication is the only time-consuming process. For this reason, the WL-components (1) and (2) are indeed equivalent (compare chapter II.B).

For modeling and simulation, we use CTU Prague’s agent middleware platform *AglobeX* [23] and the simulation environment *AgentFly* [24].

C. Extraction of Reference and Sample Data

The workflow of trajectory-based safety assessment has been ‘put into software’ (*SafetyDatamanager*), easing extraction of sample data from the radar dataset, the calculation of CF, Propensity and further metrics. The Java-based tool also visualizes trajectories (Module *SafetyVisualization*, using OpenGL) and interfaces with GNUPlot to visualize the metrics graphs. Images are included in [11]. The plugin architecture for further analyses allowed for the filtering step to be added as a module (*SafetyMetricsExtraction*).

D. Statistical Analyses

All generic statistical analyses were implemented in R-Stats [20]. The data analysis process shown in fig. 3 was fully automated by scripting calls to *SafetyDatamanager* and *R*. Runtime performance has been quite an issue; 1000 data points take 2.5 days to evaluate (1000·3=10 h = 30’000 h of TMA flight operations). Since the same data is evaluated under parameter variation, any caching of intermediate results would result in a significant speedup.

V. Results

A. Selection of Reference Data

As reference data, the ten busiest operating hours in the ‘08’ operating direction were extracted from the radar dataset and combined into one dataset.

B. Generating a Theme of ‘Super Factors’

As a result of the PCA performed, the 19 CF listed in tab. 1 were reduced to six independent ‘super factors’, listed in tab. II. Fig. 4 shows the scree-plot. The first stopping criterion is fulfilled after PC 5 (slope ‘almost horizontal’), the second (variance < 1) after PC 6. Six ‘super factors’ were chosen.

Due to slight changes in analytical methods (e.g. interpolated intermediate steps removed) and modified input data (radar data samples refer to comparable condition, e.g. operating direction ‘08’ only), the results according diverge slightly from [11]. Nevertheless, the most influencing CF (according to proportion of variance) in the Density and Horizontal ‘super factors’, are equal in both PCA (tab. II & [11]). In addition to the results in tab. II, CF could also have a minor influence to another than the listed ‘super factor’ (not included for clarity, a small cutoff value was set in the varimax rotation).
C. Sample Data

Three scenarios were generated as sample data: (1) the same ten busiest operating hours in the '08' operating direction from the radar dataset, (2) the same ten hours simulated (simulated aircraft have the same parameters and enter the sector at the same time and location as in the radar dataset; then, the trajectories evolve according to the dynamics of the agent-based model, forming an alternative, but comparable reality), and (3) the same ten hours with 130% traffic (every third aircraft gets a 'follower' trailing the 'original' at 120 s separation; soon enough, the aircraft get different advisories from the ApproachController agent, resulting in different trajectories).

Only the scenarios (2) and (3) deliver a WL figure (the one described in chapter II.B). In the results described below, the scenarios (2) and (3) did not exhibit any significant difference for any of the analyses. For this reason, we will only discriminate between scenario (1), denoted as ‘radar data’ (10 one-hour instances), and scenarios (2) and (3) combined, denoted as ‘simulated data’ (10+10 one-hour instances).

D. Analysis of the Model’s Predictive Power

The analysis was performed for both scenarios – radar data and simulated data following the methodology outlined in sect. III and using the fully automated process that allows for searching the best filtering parameters for an optimized coefficient of determination. The figures for the results are presented systematically on the next page: the left column shows the Propensity component of the analysis; the right column shows the results for WL.

First, the search space was defined. The search space is comprised of the relevant parameters for metrics filtering, which are: (x) the width of the moving average applied to the CF, (y) the width of the moving average applied to the dependent variables (Propensity, WL), and (z) the time shift between CF and dependent variable. x, y and z are values of time with seconds as unit. Trial runs were performed to grasp the distribution of $r^2$ values in the search space. By careful consideration, the boundaries were defined as follows (tab. III).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Range for Propensity</th>
<th>Range for WL</th>
</tr>
</thead>
<tbody>
<tr>
<td>CF moving average [s]</td>
<td>0..100 s</td>
<td>0..100 s</td>
</tr>
<tr>
<td>Dependent variable moving average [s]</td>
<td>0..100 s</td>
<td>0..100 s</td>
</tr>
<tr>
<td>Time-shift between X and Y [s]</td>
<td>0..200 s</td>
<td>-100..100 s</td>
</tr>
</tbody>
</table>

Additionally, it was later decided to run z from 100 s to 20 min with $x = y = z = 0$ s and $x = y = 100$ s in order to verify if correlations with predictions further into the future do exist. The reasoning for this step follows the analytical consideration of correlation: depending on the type of ATC advisory considered, the manifestation in workload or reduced safety could happen significantly later than the tactical horizon of 100 / 200 s proposed as search space. Consider a scenario where multiple flight level adaptations result in a conflict: depending on climb performance & FMS level transition sink rates the timing of a potential conflict can be calculated, yielding the possibility of long-term correlations between ATC advisories, CF and dependent metrics (i.e. several minutes). The maximum (of 20 min) was chosen because this is the time needed for approach/ departure flights to vacate the Munich TMA.

The search space was then evaluated for all 30 scenario instances. Fig. 5 shows the results for the radar data scenario instance no. 5. The 3D coordinate system is defined by the filtering parameters (compare tab. III). The relative improvement of $r^2$ (compared with the baseline value obtained without any filtering) is mapped to the cubes’ size and color (heatmap). Two connected local maxima are visible in red at $x = 100$ s, $y = 100$ s and $z = 60$ s and 130 s for this instance. The visual analysis of all scenario instances revealed that these results are highly case-dependent: local optima are always present but visual appearance, shape, and location varies considerably between instances. Nevertheless, some recurrent patterns were noticed, which motivate further research (performing more analyses and applying clustering methods to the results).

For the same scenario instance, scatter plots showing the quality of Propensity prediction by the linear model of ‘super factors’ are depicted in fig. 6. On the x-axis, the calculated Propensity value is plotted (reminder: in ln(p) scale, see chapter II.C). On the y-axis the ‘super factor’ linear model’s prediction of Propensity is shown. The $y = x$ diagonal has been included to indicate the target of ‘perfect prediction’. Considering the targeted operative environment, the diagonal discriminates ‘missed alerts’ ($x > y$, the lower triangle), valid predictions ($x = y$, on the diagonal), and false alerts ($y > x$, the upper triangle). In real-life applications, a tolerance band would have to be defined for alerting. The left panel shows the predictive performance without any filtering applied (the baseline case). The right panel shows the improved prediction with parameters matching the local optimum ($x = y = 100$ s, $z = 130$ s). Although the improvement is clearly visible, the results are not too promising for application in an operational system: a lower rate of false alerting is bought with the lowered variance introduced by average filtering. The vertical sections of the ‘ant trails’ cause additional concerns by indicating changes predicted Propensity ($y$) while the calculated Propensity is actually stable ($x$). From the definition of Propensity and previous experience, we conclude, that the ‘verticals’ evolve from aircraft being separated at the boundary of regulated separation minima. As has been noted in previous publications, e.g. [11], the low predictive power of CF towards Propensity is not problematic per se: it is the controllers’ well-paid skill to ensure that the proposed transfer function is as negligible as possible.

Fig. 7, showing the same results for radar data scenario instance 7, has been included to show a ‘false fit’. The algorithms have matched timeframes with low complexity at the beginning of the scenario instance to timeframes with low Propensity at the end of the scenario instance (the time shift is 1110 s). This fitting neuters all other Propensity predictions while obtaining a good prediction for the single transition into the safe region. It has been noted previously that performing statistical model fitting in the transformed ln(Propensity)-space has to be applied with caution [12] as the residuals in the well-safe region ($< 1E-20$) are over-emphasized in comparison to the residuals in the barely-safe region ($< 1E-9$).
Figure 5: Visualization of $r^2$ improvement for the Propensity prediction of radar data (scenario instance 5). Size and heat of the cubes indicate improvements over the baseline $r^2$ at (0, 0, 0).

Figure 6: Improvement of prediction for radar data scenario instance 5.

Figure 7: False improvement of prediction for radar data scenario instance 7 (note the large time shift value of 1110 s, which ‘fits’ one ‘ant trail’).

Figure 8: Parameters for best prediction and improvement of $r^2$ for radar (2 panels on the left) and simulated (2 panels on the right) data.

Figure 9: Visualization of $r^2$ improvement for the workload prediction of simulated data (scenario instance 5). Size and heat of the cubes indicate improvements over the baseline $r^2$ at (0, 0, 0).

Figure 10: Improvement of prediction for simulation data scenario instance 7.

Figure 11: Improvement of prediction for simulation data scenario (all instances were combined, and a common prediction model was built).

Figure 12: Parameters for best prediction and improvement of $r^2$ for the simulated data scenario (2 panels on the right; WL is not present in the radar data).
Fig. 8 summarizes the results for the Propensity part of the analysis. The box-plot in the first panel shows the filtering parameters for the local maxima of $r^2$ improvement over the real data scenario (10 instances). The mentioned instability of results is clearly visible, esp. for the width of the Propensity moving average and time shift. We interpret the large box width with the fact in mind that a single traffic situation may govern the entire model fitting (as demonstrated by fig. 7). Nevertheless, the most plausible explanation is that traffic, environmental and controllers’ individual factors are too diverse in the sample data selected. The second box-plot shows $r^2$ improvement for the radar data scenario. The improvement achieved with optimized filtering parameters at the local optimum is small but significant (paired T-test with ‘greater’ as alternate hypothesis). The large diversity of sample datasets hampers the definition of one single filtering parameter set applicable to all scenario instances (denoted as ‘tradeoff’ in the plot). Methodologically, the paired T-test’s T-value was maximized in this step, which yielded $x = 100$ s, $y = 100$ s, $z = 130$ s.

The two right panels of fig. 8 show similar results for the simulated data scenario. No notable differences between the originally defined scenarios (2) and (3) are observable. The differences to the radar data scenario are little: the visual results and the data analyses lead to similar results, casting a good light on the simulation system’s validity. The initial hypothesis of obtaining higher $r^2$ by replacing human controllers’ individual behavior with a standardized model (agent-based fast-time simulation) and thus reducing hard-to-explain variance in the data sample holds only partially true. First, only the relative improvement of $r^2$ is better than in the radar data scenario; the absolute values, however, are not as high. Second, the box width for the Propensity moving average is smaller, indicating less variant results across scenario instances. The time shift parameter, however, is not as clearly defined. And third, the loss in $r^2$ by applying one ‘tradeoff’ parameter set to all scenario instances is less grave. The best tradeoff yields $x = 100$ s, $y = 100$ s, $z = 150$ s. The time shift value is 20 s larger than in the real data scenario. The faster response time of the computer yielded $x = 100$ s, $y = 100$ s, $z = 150$ s. The time shift was evaluated for negative values as well because the values delivered by the simulation system are already accumulated (10 s or 60 s respectively). Nevertheless, the highest $r^2$ values are obtained close to the ‘current’ timeframe (the optimal $z$ is 10 s). Most notably, the visual appearance of all results is highly similar across the 20 scenario instances.

For detailed insight, a scatter plot for the WL prediction of scenario instance 7 is included in fig. 10. The explanations of fig. 6 above are applicable to this plot as well. A more significant improvement of the prediction (baseline vs. optimum) is clearly visible. Additionally, no false matches (e.g. fig. 7) were produced for WL.

Fig. 11 shows the prediction quality of all scenario instances combined. The combination is possible because the optimal filtering parameters are quite stable across scenario instances. It is noteworthy that the ‘ant trails’ of fig. 10 are not part of fig. 11, although they are included. The linear prediction is fitted to all scenario instances as sample data, resulting in a different but well-performing model. The improvement of $r^2$ is clearly visible, while some missed alerts remain for higher WL amounts.

The box-plots in fig. 12 summarize the findings with respect to WL. The stability of filtering parameters across scenario instances is recognizable from the comparatively small box widths in the left panel. The $r^2$ improvement is much larger both relatively and absolutely (right panel), and statistically significant (paired T-test). Furthermore, $r^2$ improvement does not suffer much from a ‘tradeoff’ parameter set applicable to all scenario instances (the good quality of the prediction is demonstrated in fig. 11). For an operational ATM/ATFM system incorporating CF metrics to annotate sector load figures with workload predictions, this is quite a promising result.

VI. CONCLUSION AND OUTLOOK

The challenge of defining ‘themes’ of air traffic complexity factors for operational use in ATM/ATFM systems lies in the selection of factors and in conveying this multi-dimensional and time-variant information to the end user. As the complexity factors defined in literature lack a common causation model and target various aspects of air traffic control, an optimal selection cannot be static and must be made with the target audience in mind. With the structured process for theme generation proposed in this paper, themes are built on a selection of relevant sample data and evaluated with regard to the fitness-for-use by means of regression analyses. The aggregation to ‘super factors’ reduces multi-dimensionality while retaining the option for ‘drill-downs’ to contributing factors all the way to the contribution of individual flights.

While [19] discusses the concept of dynamic density themes, it provides little practical guidance in how they should be constructed other than by grouping similar or related CFs into a particular theme. The research presented in this paper advances on this work by proposing a structured process for generating specific dynamic density themes, which we refer to as ‘super factors’.

The CFs contributing to a ‘super factor’ are more tightly defined than the CFs contributing to the more general dynamic density themes of [19] and would not usually be subject to user modification. There are practical benefits to this organization. For example, it is apparent from Figure 1 that presenting too many CFs to the operator at once can be confusing. Presenting a single theme comprised of several carefully selected CFs relieves this visual overload, but still leaves the operator wondering about the picture presented by the other themes. The “Super Factor” concept provides a way to present this information without having to display each theme in turn. We would simply annotate the bar chart of aircraft count in Fig. 1 with the
6 ‘super factors’ listed in Tab. II, providing a simplified, reduced multi-dimensionality overview or meta-representation of the dynamic density load situation. This is essentially an overview of all 6 dynamic density themes / ‘Super Factors’ at once. From this one could drill down into any of the 6 ‘super factors’ to view the relative contribution of each CF comprising the selected ‘super factors’ to the dynamic density load - this view being equivalent to what is currently illustrated in Figure 1. The final level of drill-down to the contribution of individual flights to a "Super Factor" or even to an individual CF remains possible.

The reduction in the multi-dimensionality presented to the operator at the HMI level is particularly significant as it should be noted that Figure 1 was originally proposed as a planning tool for supervisors/flow managers before being extended to a decision aid for controllers. Indeed [19] proposed replacing surveillance data with flight plan data in the themes in order to achieve the longer look-ahead times required for this purpose. Since the “Super Factors” in this study consider surveillance data and work with a look-ahead time of several minutes the implication is that controllers in a busy TMA environment will be:

(i) Hard pressed to absorb and interpret the data presented in Figure 1.
(ii) Prone to “tunnel vision” if they focus on any one theme.

Simplification of the HMI is essential in this environment and the eventual tool may end up proving more useful to supervisors and planners than to executive controllers.

Fixed ‘super factors’ would be the default ‘theme’ representation in a deployed system, with user-defined themes reserved for end-user customization.

The theme generation process was demonstrated with radar data from the TMA of Munich Airport, Germany. Sample data was selected for peak traffic demand. The statistical analyses reveal a considerable but low predictive power towards the level of safety (expressed as collision risk, or Propensity) with a look-ahead time of about two minutes when time averaging is applied to the metrics. The predictive power towards simulated workload metrics is comparatively strong; however, the word prediction is misleading because the look-ahead time is zero (explanatory power would be more accurate). Again, average filtering increases the capability.

The methodology presented in this paper exhibits a number of potential weaknesses that need to be assessed:

(1) From the perspective of signal processing, moving average filtering is equivalent to the boxcar windowing function which acts as a low-pass filter. Thus, the variance reduction performed on independent and dependent variables in this paper removes high-frequency information. The model fitting needs to be evaluated in detail, to see if complexity factors are filtered out completely to gain better coefficients of determination. In the worst case, all complexity factors but aircraft count could have been filtered out, rendering the results trivial. However, by comparing the scatter plots with the input data it is evident that the data is too variable to be attributed to sector load alone. Alternative windowing functions (e.g. raised cosine types) are worth investigating.

(2) The fitting of linear models can produce false positives, i.e. spurious correlation, which may not be detected by statistical tests such as the paired T-test applied in this paper. More research, including verifying interpretation of the results with respect to the underlying traffic situations is necessary.

(3) The workload metrics used for this study are by far not a representation of cognitive, but rather observable workload (business of the controller). In this sense, the results may hold true for stress but not for cognitive strain. ‘More cognitive’ analyses will become possible with further development of the agent-based ATC simulation model developed at TUD. Human-in-the-loop studies should be undertaken (or data obtained) to validate the results. The ‘internal validation criterion’ $R^2$ is a statistical figure entirely dependent on the quality of sample data.

(4) Further research is needed to establish validation criteria on the selection of reference data. It remains unclear by what segmentation criteria needs to be applied to surveillance-based trajectory data and which level of detail will be sufficient, i.e. how many themes are required for operational use. This question is connected to the question of algorithmic stability. Detailed analyses of the metrics’ time-series combined with what-if scenario techniques may answer this important question.

Finally, it is to be remarked that the research on complexity factors has sparked a lot of research trying to link traffic complexity to safety. As appealing as these efficient and easy-to-grasp figures are, they were defined to explain cognitive workload. This paper unsuccessfully tried to link workload with the collision risk levels by means of time shifting (alleging a transfer function with constant delay). The much-observed non-correlation emphasizes that the ATM system including all human stakeholders, performs well within the design range. The alienation of air traffic complexity metrics should be relinquished in order to concentrate on the creating of valid predictors for the violation of the safety design range.

REFERENCES


[8] ATC Complexity as Workload and Safety Driver; J. Djokic, B. Lorenz, H. Fricke; 3rd International Conference on research in air transportation; Fairfax, VA, USA; 2008


[10] Dynamic density: An air traffic management metric; I. V. Laudeman, S. G. Shelden, R. Branstorm, C. L. Brasil; NASA/TM 1998-112226; San Jose State University Foundation, San Jose, CA, USA 1998


[16] Thesis: Investigation into ATC Safety Indicators (Subjective Assessment), J. Djokic, University of Belgrade, Faculty of Transport and Traffic Engineering, Belgrade 2005


