DOT/FAA/TC-14/3

02_0019_CDRL_1414_19_20130823
SE-2020-TASK 0019 CDRL
Validation Report
Aug 23, 2013

Air Traffic Control/Technical Operations Human Factors Program Support

Validation Report

Aug 23, 2013

SE-2020 – Task 0019: AGENTFLY Validation

02_0019_CDRL_1414_19_20130823

Prepared under:

DTFAWA-11-D-00002

Task Order # 0019, CDRL # 19

Prepared for:

Federal Aviation Administration

TABLE OF CONTENTS

1. Introduction **6**

1.1 Background *6*

1.2 Human performance modeling in ATC *7*

1.2.1. Alion's Center Controller Performance Model (CCPM) 8

1.2.2. MicroSaint ATC 8

1.2.3 Man-machine Integration Design and Analysis System (MIDAS) 8

1.3 The AGENTFLY project *8*

1.3.1 AGENTFLY architecture 9

Workload algorithm 9

1.3.2 Model parameters 10

1.4 Aims of the AGENTFLY validation *10*

2. Methodology **12**

2.1 Criterion study *12*

2.2 Data sources *12*

2.2.1 HITL data 12

HITL Common Message Set (CMS) tables 13

Human performance data 13

Eye tracking data 13

SME workload ratings 13

2.2.2 AGENTFLY data 14

AGENTFLY Common Message Set (CMS) tables 14

AGENTFLY Scanner module data 14

Workload tables 15

2.3 Metric selection *15*

2.4 Data analysis: General approach *18*

2.4.1. Assumptions 18

3. Results **21**

3.1 Human performance *21*

3.1.1 Workload 21

3.1.2 Dynamic Density 22

3.1.3 Visual scan 24

3.1.4. Number of aircraft under control 31

3.1.5 Aircraft mix, climbing and descending 33

3.1.6 Handoffs 34

HITL 34

AGENTFLY 35

3.1.7 Flight attitude at sector entry 36

HITL 36

AGENTFLY 36
3.1.8 Flight attitude at sector exit 37
HITL 37
AGENTFLY 37
3.1.9 Flight Level (FL) occupancy rate 38
3.1.10 Radio telephony (RT) rate 39
3.1.11 Radio Telephony (RT) message duration 40
3.1.12 Conflict alert frequency 40
3.1.13 Keystrokes required per command 40
3.2 System performance *41*
3.2.1 Traffic density, average sector flight time, average 41
3.2.2 Traffic density, average sector flight time, per flight 42
3.2.3 Aircraft performance, average deviation from cleared trajectory 43
3.2.4 Minimum separation distance 44
4. Discussion and conclusions 45
5. References 48
Appendix A: Workload in AGENTFLY 50
Appendix B: AGENTFLY parameters (pilot and ATC parameters) 54
Appendix C: Final metric list 63
Appendix D: Simplified Dynamic Density (SDD) 64

LIST OF ACRONYMS

ANOVA	Analysis of Variance
AOI	Area of Interest
AP9	Action Plan 9
ARTCC	Air Route Traffic Control Center
ATC	Air Traffic Control / Air Traffic Controller
ATWIT	Air Traffic Workload Input Technique
BADA	Base of Aircraft Data
CCPM	Center Controller Performance Model
CMS	Common Message Set
DESIREE	Distributed Environment for Simulation, Rapid Engineering and Experimentation
DV	Dependent Variable
DD	Dynamic Density
E-OCVM	European Operational Concept Validation Methodology
EPOG	Eye Point of Gaze
ERAM	En Route Automation Modernization
FL	Flight Level
GLM	General Linear Model
JEDI	Joint En Route Decision Support System Infrastructure
HDI	Highest Density Interval
HITL	Human-in-the-Loop
HPM	Human Performance Modeling
ICAO	International Civil Aviation Organization
KPA	Key Performance Area
KPI	Key Performance Indicator
MAP	Monitor Alert Parameter
MIDAS	Man-machine Integration Design and Analysis System
NAS	National Airspace System
RDHFL	Research Development and Human Factors Lab
RT	Radio Telephony
SA	Situation Awareness
SCI	Sector Complexity Index
SD	Standard Deviation
SDD	Simplified Dynamic Density
SME	Subject Matter Expert
TAAM	Total Airport and Airspace Modeler
WAK	Workload Assessment Keypad
WJHTC	William J. Hughes Technical Center

ACKNOWLEDGMENTS

We wish to thank FAA's ANG-C41 for their support throughout this validation effort, especially Albert Schwartz, Christina Young and Andrew Fabian. We are grateful for their feedback, technical support, and patience.

We would also like to thank the RDHFL for their support with their HITL data, especially Albert Macias, for providing us with the eye tracking data, Daniel R. Johnson and Kenneth A. Schulz for their advice on the SepMan2 experimental design and data.

We also thank CTU for their programming effort and technical support, especially David Šišlák and Přemysl Volf for their continuous help in deciphering the AGENTFLY output.

Last, but certainly not least, we would like to thank Duc Nguyen of Drexel University, for his help in conducting fast time simulations.

1. Introduction

Computer modeling refers to a broad class of techniques used to capture an abstract representation of the mechanisms, algorithms and functions underlying a given system, and to simulate (perhaps in fast time) the performance of that system. When modeling human – machine systems such as Air Traffic Management (ATM), attention is typically focused on describing both the physical dynamics and properties of the system, and elements of human performance. These elements have traditionally included physical (ergonomic) properties, but more attention has increasingly been paid to the cognitive processes—such as perception, information processing, and response—underlying human performance, especially those based on subjective reports.

1.1 Background

Although necessarily simplified, computer models of ATM provide a number of potential benefits, and play an important role in ATM research. Human Performance Modeling (HPM) is of particular importance to development of the Next Generation Air Transportation System (JPDO, 2007). HPM represents ATM human agents as computer entities, and facilitates testable predictions about how human agents (especially controllers) will behave under hypothetical air traffic scenarios. Traditionally, formal representation of the human element in ATM—namely the controller(s), flight crew and other agents—have necessarily aimed to replicate only limited aspects of human behavior, in a simplified environment (Abkin, 2001).

There are at least two roles that HPM can play in NextGen research: first, to limit the need for HITL simulation, and second, to complement HITL research. HITL simulation is resource intensive and, particularly in preliminary or high-level investigations, a moderate fidelity model can suffice, particularly in the earliest stages of concept development. Whereas human-in-the-loop (HITL) studies are often thought to provide the strongest level of empirical evidence, there are obvious drawbacks to such research. Human participants tend to be relatively expensive, difficult to recruit, and prone to fatigue. Computer modeling, on the other hand, can allow us to cost-effectively run large-scale interactive simulations. Because HPM by definition involves simplification of the actual context, its outputs tend to be relatively reliable, as only a few aspects of human behavior are typically simulated at a given time. Such reliability can be particularly useful for uncovering subtle phenomena, which by definition have a statistically small effect size and a tendency to be masked by random variability, and thus do not lend themselves to HITL observation.

In the past, these models tended to rely on strict engineering approaches that treat cognitive functions (e.g. memory and attention) as "black box" phenomena (O'Donnell, 1991). The past few decades, however, have seen among the computer modeling community an increased focus on cognitive processes. For example, approaches such as SOAR (e.g., Cheng, 2004) and MIDAS (Gore, 2002; Laughery & Corker, 1997) have attempted to capture the cognitive mechanisms underlying controllers' behavior.

Most human behavior is not separable from the underlying cognitive processes, and attempts to simulate human behavior will inevitably fail if cognition is not taken into account. Because only behavior is observable, the underlying cognitive processes typically have to be inferred. Thus, human performance models that have attempted to incorporate cognition—i.e., decision-making, attention and memory—into the simulation of human behavior, have to rely on accepted information-processing stages. Nonetheless, HPM programmers need to explicitly incorporate those cognitive processes into the models. The hope is that by incorporating those "invisible" cognitive processes, simulations can produce more reliable outputs, that closely match actual human performance as accessed via either HITL simulation, or operational data. Ultimately, the goal is to improve the models' ability to predict human behavior in a real world setting.

The most ambitious of these modeling approaches has tried to incorporate unified theories of cognition to capture subtle cognitive activities and represent emergent processes. Some years ago, Boeing ATM created an ATM human agent model controller, including a combination of task network and cognitive model. Boeing used subject matter expertise to document en route and TRACON controller actions, and converted these into task network diagrams. Boeing attempted to integrate into their cognitive model the concept of limited attentional resources (visual, auditory, spatial cognitive, response selection, motor, and vocal responses). Such an approach is sometimes referred as to *reductionist* (Leiden, Kopardekar and Green, 2003), in that it decomposes general tasks, in a top down fashion, into component subtasks that can be sequentially mapped.

The following section highlights some previous applications of HPM in the area of ATM, and introduces the AGENTFLY model.

1.2 Human performance modeling in ATC

The U.S. National Airspace System (NAS) is divided into sectors, and these sectors are assigned to Air Traffic Controllers (ATCs). ATCs have to perform several tasks, while managing the assigned sector. Such tasks include, but are not limited to: accepting aircraft into their sector; checking aircraft; issuing verbal

or data instructions and clearances, efficiently communicating with pilots and so on. Given the relevance of the ATC's role in the efficiency of air traffic, it is not surprising that many models and simulations have included an ATC component. Below, we report a few notable examples of models that incorporate or focus on the role of the controller in air traffic simulations.

1.2.1. Alion's Center Controller Performance Model (CCPM)

CCPM is a rule-based HPM able to simulate the behavior of controllers and pilots and has been applied to a dynamic airspace configuration concept. Several controller tasks—including handoff, conflict resolution, communication, metering, clearance issuance and coordination— are implemented, and each task decomposed into sub-task elements. Controllers are represented as multi-channel processors (Wickens, 1984) with four channels: Visual (V), Auditory (A), Cognitive (C) and Psychomotor (P).

1.2.2. MicroSaint ATC

MicroSaint is a model developed for NASA, and it includes both physical ATC tasks and eye tracking. The model consists of 2 components: ATC tasks and airspace. Like CCPM, this model is able to simulate several controller tasks, each requiring a certain amount of workload (Wickens, 1984) in each of the involved modalities (VACP), coming from different input modes (Trackball, Keyboard, Radio, Landline, EyeFocus).

1.2.3 Man-machine Integration Design and Analysis System (MIDAS)

MIDAS is a sophisticated HPM environment that represents human ATM agents as computer entities with underlying computational cognitive structure representing human capabilities and limitations (Gore, Hooey & Foyle, 2008). The model involves human perception and attention, as well as models of working and long-term memory.

1.3 The AGENTFLY project

AGENTFLY is a "complex multi-agent system for modeling and simulation of air-traffic, involving both manned and unmanned aircraft" (Šišlák, Volf, Jakob, & Pěchouček, 2009). AGENTFLY is currently being developed for the FAA, in a collaborative effort between Czech Technical University (CTU) and Drexel University. Part of this effort is dedicated to developing an agent-based simulator for use with Next Generation Air Transportation System (NEXTGEN) projects (Cannon, Nguyen & Regli, 2012). One of the main goals of AGENTFLY is to accurately simulate the behavior of controller agents, especially under

increased traffic scenarios, as anticipated over the NextGen development timeline. AGENTFLY focuses on en route airspace, and has the potential to be a NAS-wide simulation tool, capable of accommodating center and sectorization splits. The formal validation of AGENTFLY (especially the Air Traffic Control simulation agent within AGENTFLY) is the focus of the current report and will be described in the next section.

1.3.1 AGENTFLY architecture

AGENTFLY uses an agent-based approach that incorporates both time-step and event-driven simulation (Sislak et al, 2012). It includes high fidelity models and interactions of air traffic controllers and pilots, and can perform in either real-time or fast-time mode. The system implements controllers, pilots, aircraft and even tools such as ERAM as distinguishable actor agents, and aircraft performance aims to conform with EUROCONTROL's Base of Aircraft Data (BADA) performance models. Most of the initial AGENTFLY data parameter values were based on the Microsaint ATC model

Workload algorithm

AGENTFLY's ATC agent is represented as a computer information-processing entity with four concurrent processing channels (Visual, Cognitive, Auditory, and Psychomotor, together abbreviated VCAP), in line with Wickens' (1984) model of attention and mental workload. Workload is generally seen as a multidimensional construct that is the end product of a number of input factors, and is typically considered a function of both internal and external factors. These include not only characteristics of the traffic and task demands, but also an individual's response to those factors. When we speak of workload, it is therefore important to keep clear the distinction between task load (the input presented to the controller, such as traffic load) and the resulting workload experienced by a given controller. In practice, research often blurs this distinction, and relies on task load measures (e.g. number of aircraft under control) as implicit or proxy measures for workload. The concept of workload has long played a primary role in ATM evaluations, and is often the chief criterion by which new tools and procedures are assessed. Indeed, one of the main goals of human performance models has been to accurately predict workload as a function of MAP values (Leiden, Kopardekar & Green, 2003).

Given that AGENTFLY explicitly incorporates a workload monitoring system, and further given the central role that workload played in this validation, it is important to note a major simplifying assumption made by the AGENTFLY developers. One of the cornerstones of the Wickens' workload model concerns differential interference between tasks. That is: tasks can be performed in parallel, and some tasks (depending on input, output and processing characteristics of the tasks) can be timeshared,

whereas others cause detrimental interference. AGENTFLY has implemented a simplification of this model based on exclusive resource usage, and does not currently permit multitasking or differential task interference. Under AGENTFLY, task interference is a function of time proximity. During performance of a given task, the system cannot perform the next (serial) task until a timeout expires. Appendix A provides an overview of the AGENTFLY workload algorithm.

In addition to monitoring workload, the ATC agent can perform a number of possible actions, including:

- Scan;
- Monitor;
- Analyze;
- Handoff aircraft;
- Issue traffic advisory;
- Perform airspace avoidance;
- Point out;
- Check conflict; and
- Resolve conflict.

When a conflict occurs, AGENTFLY chooses the least complex solution, with complexity being defined on the basis of aircraft coordination, sector coordination, number of actions and positive separation.

Sectors that are outside of the simulated controller's responsibility are simulated by a *ghost agent*, which does not simulate the air traffic controller per se but instead simulates the interaction that would occur with adjacent sectors (Cannon, Nguyen & Regli, 2012). The ghost agent can perform a limited set of tasks, such as initializing, scanning and handoff.

Lastly, AGENTFLY includes a *pilot agent*, which controls the simulated airplane executing its flight path (ibid). The only task assigned to the pilot agent is communication.

1.3.2 Model parameters

AGENTFLY is parameterized around three main agents: Controller, Pilot, and Intelligent Conflict Resolution (ICR). Appendix B presents the parameters associated with the ATC and pilot modules, current as of spring 2013.

1.4 Aims of the AGENTFLY validation

Validation ideally involves both a clear definition of the real world ("criterion") context, as well as a clear understanding of the model's intended use (DoD, 1996). With these thoughts in mind, we set out to [1] compare AGENTFLY's modeled output against results from a representative data source obtained from

human in the loop (HITL) simulation. Additional aims of the effort were to [2] help parameterize and calibrate the baseline AGENTFLY system, based on empirical human performance data; and [3] provide, from a human factors perspective, a preliminary assessment of the theoretical soundness of the AGENTFLY underlying human model.

2. Methodology

2.1 Criterion study

Data from several HITL simulations were available from the RDHFL. Therefore, we adopted a number of criteria in order to select the most appropriate dataset. The main selection criterion was the similarity between the HITL conditions and AGENTFLY capabilities. We therefore excluded studies focused on aspects inconsistent with AGENTFLY data. We also sought a dataset with eye tracking results, to allow a comparison between eye point of gaze (EPOG) measures and those from AGENTFLY's visual scanner module, as discussed in section 2.2.2.

In the end, we selected the baseline condition of the **Separation Management II** (hereafter SepMan2) study (Zingale, Willems, Schulz, & Higgins, 2012), recently conducted at the RDHFL with the aim of evaluating conflict probe location and format and workstation display alternatives in the context of the En Route Automation Modernization (ERAM) system. The dataset included the simulation specifics and human performance data for a total of ten participants, tested individually. Six of these participants were tested in one flight scenario and four of them were tested in an alternative flight scenario.

2.2 Data sources

In order to assess the validity of AGENTFLY, the team gathered data from several sources, including the SepMan2 HITL simulation, Subject Matter Expert (SME) self-report data, and AGENTFLY model output. The following summarizes these data sources.

2.2.1 HITL data

Among the conditions included in SepMan2, we chose the Baseline condition (ibid), because the conflict probe notifications—which were the original focus of the study—were presented outside of the R-side controller workstation. According to the report, "R-side controller workstations were equipped with a high-resolution, 2K (2,048 x 2,048) radarscope, keyboard, CRD, and trackball" (ibid). The Distributed Environment for Simulation, Rapid Engineering, and Experimentation (DESIREE) was used to simulate the Baseline ERAM. Participants in the SepMan2 baseline condition completed a 45-minute, high traffic flight scenario.

HITL Common Message Set (CMS) tables

The CMS is the format used for inter-computer communications (FAA, 2004) and a CMS-formatted scenario file is an ASCII file consisting of CMS messages (Paglione, 2011). Data recorded from various communication sources is collected into CMS XML files and can be extracted as part of the FAA's Conflict Probe Assessment Team (CPAT) scenario generation process into derivative tables. Only derivative tables containing relevant data were provided to us. We further parsed the tables to arrive at the desired format.

Human performance data

As part of its HITL procedure, RDHFL already collects a variety of system and human behaviors. Data are logged to file, and capture such aspects as keyboard activity, interface settings, onscreen activities, and RT interaction between ATC and pseudopilot. The rich dataset consisted of several files relevant to the validation effort, including but not limited to workload, aircraft count, handoffs, radio communication times and the complete set of actions performed by the participants on the aircraft included in the scenario. Of great importance for the evaluation, were the measures obtained using the Air Traffic Workload Input Technique (ATWIT), which required participants to press one of seven buttons to indicate instantaneous workload during the simulation. ATWIT workload data were collected every two minutes.

Eye tracking data

The RDHFL is equipped with an eye tracker system (consisting of infrared pupil capture, and magnetic head tracking device), for capturing time-stamped point-of-gaze. Raw eye tracking data were provided from the SepMan2 HITL trials. The data consisted of all the eye movement recordings (i.e., fixations, saccades[1], dwell time etc). The system also provided for each fixation a timestamp, pixel coordinates, and probable fixation target (onscreen object).

SME workload ratings

Though technically not a part of the validation effort, we did rely on SME input to help calibrate the AGENTFLY baseline model. We obtained workload ratings for each of the VCAP modalities. Data were collected through an online questionnaire we compiled. Before giving the ratings, two SMEs were debriefed over the phone and questions were answered as they arose. The questionnaire included 27 questions, one for each of the parameters associated with a specific ATC action within AGENTFLY. SMEs

[1] Saccades are short, coordinated, ballistic movements of the eyes, and tend to be fast (on the order of 200 msec), short (less than 20 degrees angular shift) and independent of head movement.

were asked to indicate, on a scale of 1 to 7, effort by modality for each given task. Scores were averaged and used as needed by CTU to replace missing values in the system.

2.2.2 AGENTFLY data

With a few exceptions, the parsed AGENTFLY dataset closely paralleled that of the SepMan2 HITL simulation.

AGENTFLY Common Message Set (CMS) tables

The same CMS files used by the FAA for inter-computer communications were obtained as output of AGENTFLY. Such CMS files were processed through the CPAT scenario generation process. We further parsed the tables until the desired format was achieved.

AGENTFLY Scanner module data

AGENTFLY is equipped with a visual scanner that aims to simulate human eye movement behavior. Fixations and dwell times for the same simulation scenarios used in SepMan2 were recorded and logged into a file containing the region and time (in msec) of fixation for each time unit. Regions were defined based on their geographic (latitude and longitude) or display (pixel) coordinates within the sector used in the simulation scenario. Fixations were aggregated in distinct time intervals (1 minute, 5, 10 or 15 minutes) and separate files were provided for each time interval.

It is worth noting that AGENTFLY visual scanning is qualitatively different from that of humans. First is the issue of attention: Whereas humans can, and often will, re-fixate on the basis of element salience (e.g. a light appears somewhere in the visual periphery), the AGENTFLY system follows a sequential scan to systematically fixate discrete (grid) elements of the visual field. It then determines, on the basis of activity within the given grid element, whether to dwell at this location. Second, AGENTFLY currently incorporates no object tracking capability. Instead, it successively scans the grid to update its representation of the field of view. Figure 1 shows the spatial grid used by the AGENTFLY visual scanning module.

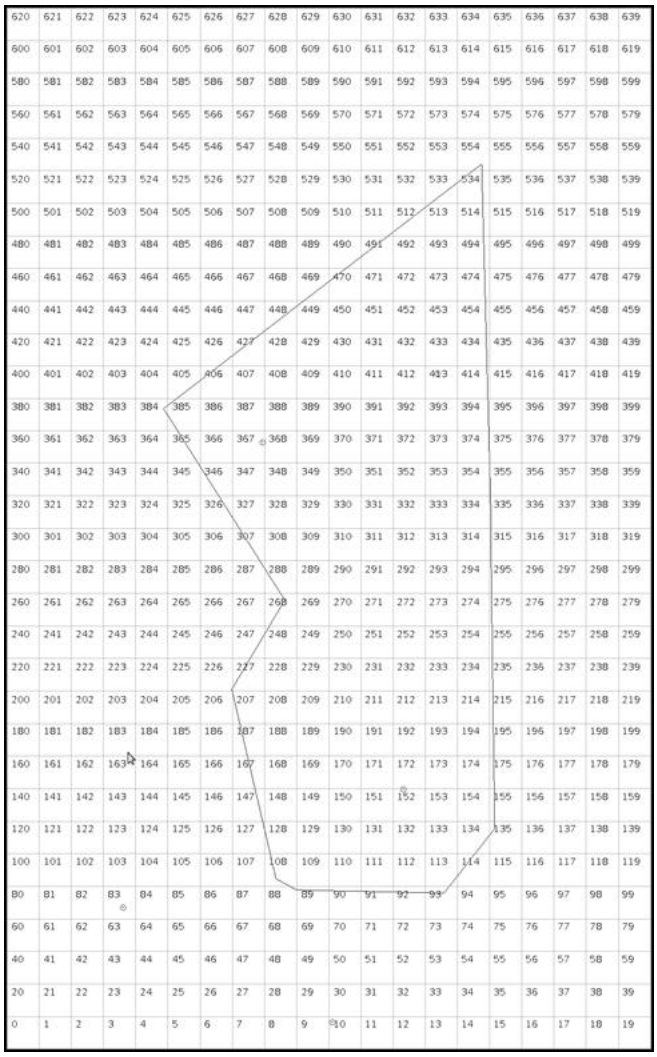


Figure 1. Spatial grid used for visual scan module.

Workload tables

On the basis of the VCAP workload algorithm, as discussed elsewhere, AGENTFLY outputs workload prediction in the form of workload tables. Data are broken out by task element and weighting along the VCAP dimensions. Data are written to file every two minutes.

2.3 Metric selection

There are several measurement issues that must be considered in selecting metrics. These include whether a metric is reliable (repeatable), valid (meaningful), and in fact is even measureable (Holnagel, 1998). When we speak of validity, we are concerned about both internal validity (i.e. is the measure internally consistent?) and external validity (do the results generalize?). We must also consider reliability as it is influenced by both the measurement technique itself (are results repeatable, or does the

technique yield inconsistent results?) and inter-rater differences (in the case of SME derived data, do results vary from one rater to the next?).

An initial list of candidate metrics was generated. The two main sources for this list were FAA and EUROCONTROL, including their collaborative work in the Action Plan 9 (AP9) effort to agree a common set of ATM research metrics. The team did not generate an exhaustive list of all possible ATM metrics. Literature review focused not on the breadth of empirical ATM work, but rather on literature reviews and available metrics lists. The primary source for the initial metrics list was the recent and thorough review by FAA (2010), which itself referenced reviews by Williams et al. (2004); Brueing, Bradford & Liang (2003); Casso & Kopardekar (2001); Allendoerfer & Galushka (1999); Hadley, Guttman & Stringer (1999); and Bradford, Brown & Blucher (1998).

There is a good deal of ambiguity in such metrics lists, in terms of definition, nomenclature, and categorization. Whereas some taxonomies refer to *constructs* and *metrics*, others use the terms *Knowledge Performance Areas* (KPAs) and *Knowledge Performance Indicators* (KPIs). Further, it is not always clear how to categorize a given metric. Measures of complexity, workload and task load can overlap. For instance, workload is a human performance metric, which impacts both capacity and predictability of the system. Air traffic *complexity* is sometimes seen as an input (generally as the composite static and dynamic factors presented to the controller), and sometimes as an output, namely the controller's impact on the traffic pattern. Moreover, the distinction between human and system performance metrics cannot always be drawn. In some cases performance measures refer to the overall human – machine system. Finally, some metrics have units and specific definitions, other "metrics" are methods, such as a particular test.

Human Performance KPAs (primary reference CARE ASAS 2002, others)

1. Workload
2. Situation Awareness
3. System monitoring
4. Controller effectiveness
5. Human error
6. Teamwork
7. Trust
8. User acceptance and usability

System Performance KPAs (from FAA / EUROCONTROL Action Plan 9)

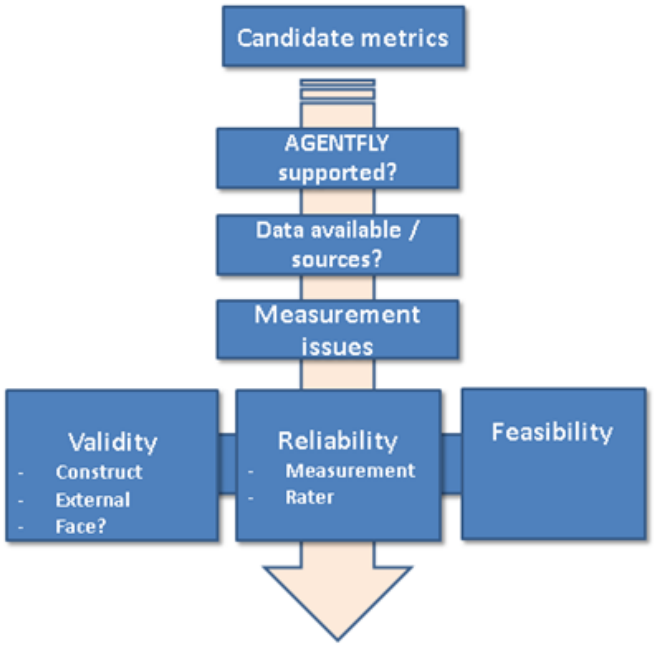
1. Capacity
2. Cost
3. Efficiency
4. Environment
5. Flexibility
6. Predictability
7. Safety

Sources: EURONTROL (2004, 2002a, 2002b); FAA (2010, 2005); FAA & EUROCONTROL (2003)

Given the ambiguity and overlap among metrics lists, we used the following preliminary criteria for identifying candidate metrics:

- Must be focused on en route, as opposed to terminal or ground;
- Must be a metric or method from which a specific measurement procedure and measurement units could be inferred; and
- Must be non-redundant in terms of measurement units (e.g. aircraft per hour, per day, and per 15 minutes were all seen as the same metric: *aircraft per unit time*).

A preliminary list of candidate metrics was reviewed by the validation team. This review resulted in a down-selected final list of metrics, as shown in Appendix C. The strategy used to refine this metrics list is shown in figure 2



.

Figure 2. Metrics selection strategy.

As shown in Appendix C, our final list of metrics can be mapped onto several human- and system-performance constructs, as follows:

Human Performance constructs:

- Workload
- Task load
- Coordination
- Effectiveness and efficiency

System Performance constructs:

- Capacity
- Safety
- Cost

2.4 Data analysis: General approach

The richness of the available data allowed the validation team to perform several analyses, which included subjective evaluation of the output, correlation between the outputs (when possible) and testing of the difference between means (e.g., t-test and ANOVA). Because our goal was to test the similarity between AGENTFLY and human behavior, rather than the difference between the two, we relied on Bayesian statistics, when needed. Bayesian analyses allow one to assign credibility to various hypotheses about the difference between two groups' values. Accordingly, when a hypothesis test did not allow us to reject the null hypothesis—that the difference between the AGENTFLY's output and HITL data equals zero—we used Bayesian estimation to assess the credibility of the null hypothesis (Kruschke, 2012). From the Bayesian analyses, we report the 95% *high density interval* (HDI), which indicates the interval of most credible hypotheses regarding the difference between the two simulations (Kruschke, 2012). An HDI distributed around zero indicates that no difference between the simulations is a highly credible hypothesis.

Several other issues were considered in determining the appropriate analysis approach. For example, some of the candidate metrics (e.g., workload) only made sense in the context of the developing scenario over time. However, it is well known that the interpretation of analyses involving time series can be problematic, due to autocorrelation often observed in this type of data. The next section will describe these issues and how we addressed them.

2.4.1. Assumptions

This section summarizes the most relevant assumptions made in the context of this validation. First, some of the assumptions are necessarily related to the computer model, as any simulation of a complex

phenomenon involves a degree of simplification, especially in the domain of human behavior. Second, some assumptions were made in order to be able to compare different sets of data, some of which were not necessarily collected with the purpose of being compared to human behavior. This list does not aim to be complete, but offers an overview of the major assumptions that were made, to the best of our knowledge.

First, AGENTFLY assumes that the each aircraft is set to 75% of maximum payload and 50% of maximum fuel at takeoff. This is clearly an oversimplification, because in real-world scenarios, aircraft required take-off fuel load is calculated based on the distance to be flown, not on fuel capacity. However, the same parameterization appears to have been used in both the SepMan2 HITL and AGENTFLY.

Other AGENTFLY assumptions and error handling procedures included the following:

- For about half of the aircraft that would run out of fuel in flight, fuel was set between 50% and 100% of maximum fuel;
- For remaining aircraft, fuel was increased over 100% of aircraft capacity, to allow the aircraft to complete the planned flight;
- A few remaining aircraft, for which problems could not be resolved, were removed from the simulation;
- AGENTFLY adjacent ("ghost") sectors automatically handle incoming and outgoing handoffs;
- As discussed in section 2.2.2, AGENTFLY and human visual scanning are fundamentally different;
- As discussed in section 2.2.2, workload computation under AGENTFLY precludes parallel task execution, and differential task interference; and
- Task and RT durations under AGENTFLY do not appear stochastically distributed.

Assumptions made for purposes of the validation include the following:

- Analysis could be robust to the strong correlation between traffic level and session time. Pre-analysis of HITL results showed a strong positive correlation between session time and traffic; that is, traffic load increased steadily over the entire course of the session. At first, this seemed to represent a problem, because in time series analysis a common issue is the autocorrelation of the data. That is, due to the fact that two events closely spaced in time cannot be considered completely independent, some amount of correlation is observed by shifting the time series to be correlated in time. This was confirmed in our dataset. That is, a strong autocorrelation was observed between AGENTFLY and SEPMAN2 workload evaluations. Therefore, we concluded that the correlation would be inflated due to the autocorrelation in the data sets. However, we assumed that this did not impact our analysis because we were not making any inferences regarding the cause of that increase (e.g., aircraft count increase) but rather simply checking on the ability of AGENTFLY to capture task load differences.
- Spatial gaze distribution can be roughly compared to AGENTFLY scanner dwells. The discussed differences between human and AGENTFLY visual scanning meant that certain metrics often

applied to human eye tracking data (e.g. fixation frequency, blink rate) would not applicable to AGENTFLY. Nonetheless, we assumed that spatial gaze distribution (i.e. where gaze was fixated) and dwell times (how long a given region was fixated) could be compared between HITL and AGENTFLY outputs.

3. Results

3.1 Human performance

3.1.1 Workload

AGENTFLY workload output was compared to the workload ratings of SepMan2 HITL participants. Ratings were averaged across runs and subjects, and compared separately for each scenario. Because HITL and AGENTFLY data used different scales, we standardized the outputs as z scores to enable comparison. Results showed a significant correlation between AGENTFLY and participant ratings over time in both scenario B (r = 0.81; p < .001; Figure 3) and scenario D (r = 0.59; p < .01; Figure 4).

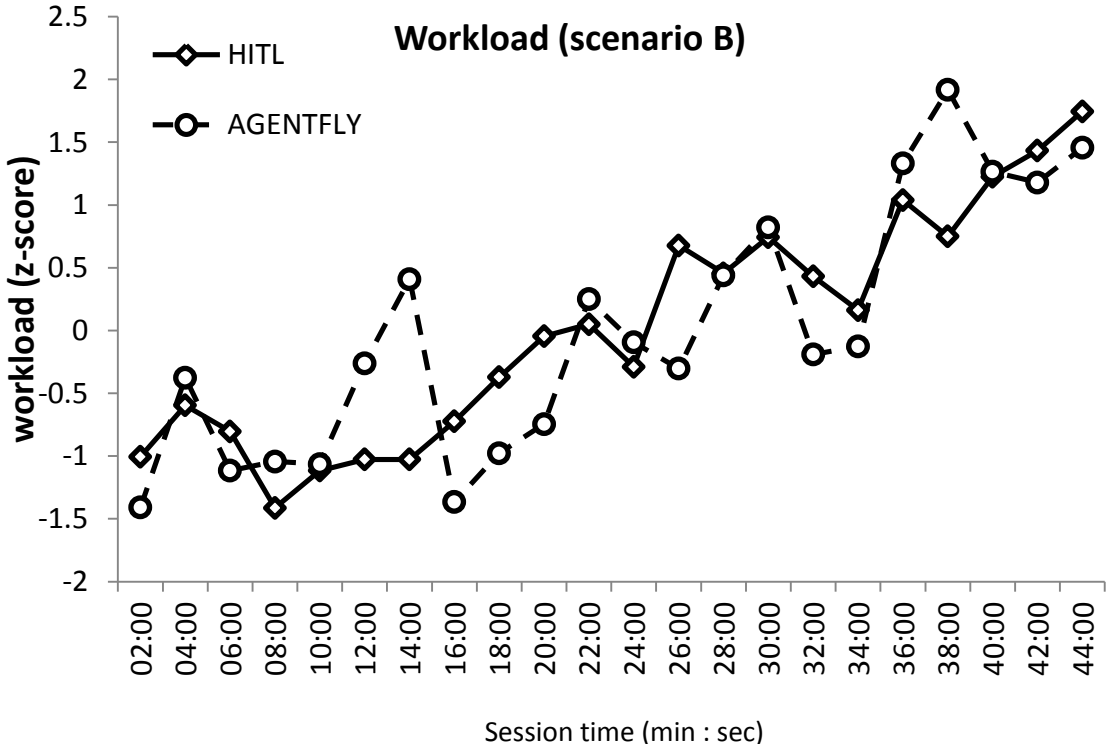


Figure 3. Workload, HITL vs AGENTFLY (Scenario B).

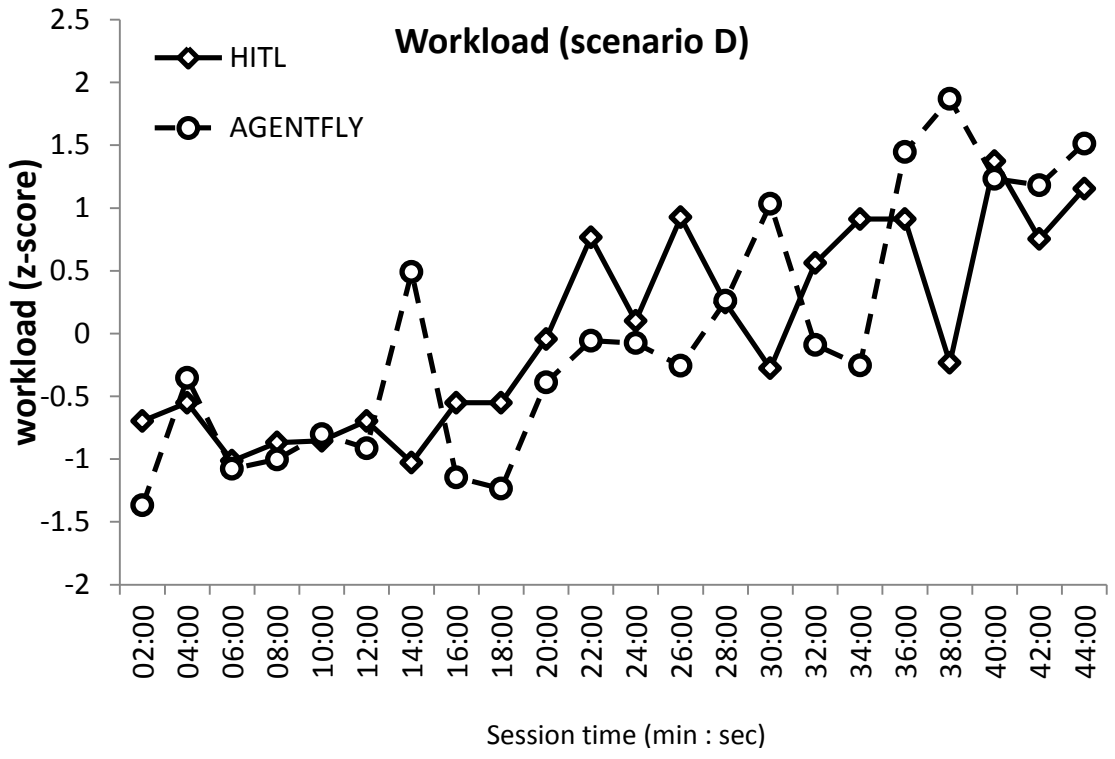


Figure 4. Workload, HITL vs AGENTFLY (Scenario D).

3.1.2 Dynamic Density

Dynamic Density, or DD (Laudeman, Shelden, Branstrom, & Brasil, 1998; Masalonis, Callaham & Wanke, 2003) has been under development for over 15 years as an accurate and robust indicator of sector traffic complexity. In its original form, DD was defined as "*a measure of control related workload that is a function of the number of aircraft in the complexity of traffic patterns in a volume of airspace*" (Laudeman et al., 1998). Magyarits & Kopardekar (2001) later defined it as "*a collective effect of all factors that contribute to the sector level air traffic control complexity or difficulty.*" DD, however, has been criticized as unwieldy. Klein, Rodgers and Leiden (2009) studied a smaller representative set of Dynamic Density components, with the aim of creating a Simplified Dynamic Density (SDD) metric, which consists of the weighted sum of such components. These SDD components include: sector occupancy counts (absolute or relative to MAP value); proximities in a sector altitude transitions in a sector (aircraft climbing or descending at a rate > 500 ft/min); transfers across sector boundaries, both horizontal and vertical (piercing); number of aircraft per sector volume; variance of aircraft headings in sector; and variance of cruising aircraft speeds in sector.

FAA provided CMS derivative tables which included the parameters necessary to derive the SDD metric. SDD was computing as the weighted sum of the following seven parameters, in order of weighting (Klein, Rodgers and Leiden, 2009):

- Traffic density
- Occupancy counts
- Sector boundary crossing
- Proximities (4 levels)
- Altitude changes
- Heading variance
- Speed variance

A parameter was sampled every 300 seconds of the simulation for both AGENTFLY and SepMan2, in each of the two scenarios used for this validation (B and D). Subsequently we computed the correlation between the two scenarios, across each simulation (i.e., AGENTFLY and HITL). Results (see Figure 5) show that the SDD values in scenario B perfectly correlated with those of scenario D, in both AGENTFLY (r= 1; p < .001) and SepMan2 (r= 1; p < .001). Most important, the SDD values obtained in AGENTFLY also correlated highly with those obtained in SepMan2, separately for the B and D scenarios (both rs= 0.98; p < .001).

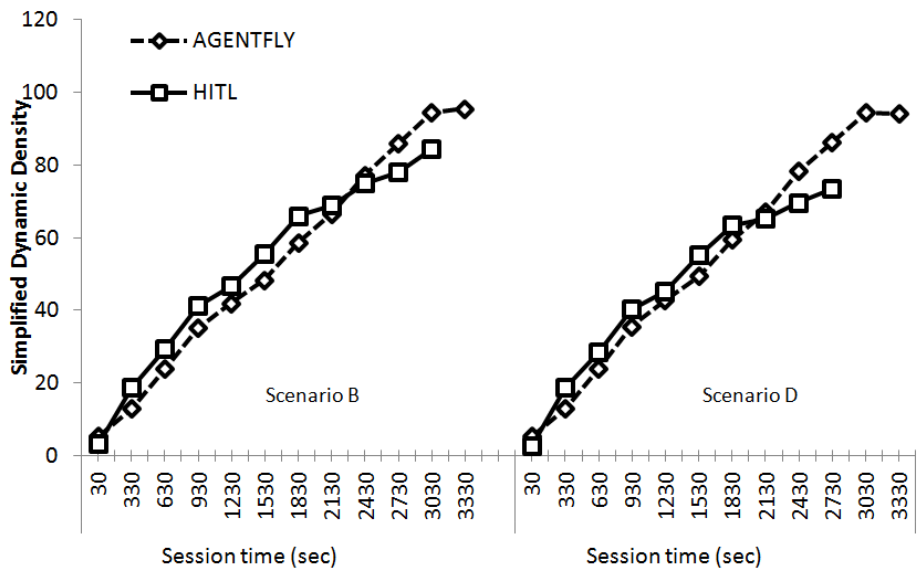


Figure 5. Simplified Dynamic Density (SDD), HITL vs AGENTFLY, separately for B and D scenarios.

3.1.3 Visual scan

To compare HITL and AGENTFLY visual scan, we first grouped the eye tracking data to match the scanner's output. Fixations, which were originally recorded as they occurred, were grouped into 15 minute time intervals (0-14 minutes; 15-29 minutes; 30-45 minutes), from start of simulation. To permit comparison between the different HITL and AGENTFLY time scales, we converted fixation times to percentages. Specifically, we calculated the percentage of time spent on each area of the screen, compared to the total time spent looking at the screen. Given inconsistencies in the screen setting across subjects and resolution of the eye tracker, we determined that a numerical analysis was not appropriate. Instead, we opted for a subjective evaluation of the similarities between the eye tracking data and scanner output. We mapped the spatial distribution of visual scan over time, as a means of comparing spatial attention of AGENTFLY and HITL. Ideally, the system should in the end mimic human allocation of attention, to those parts of the visual field (radar screen) that contained task-relevant information (e.g. aircraft plots, data labels, and sector critical information). This was done by constructing fixation "heat maps" built on the basis of time-weighted fixations (ie, points-of-gaze for HITL data, scan location for AF), defined using an AA x BB grid.

Based on preliminary investigation, each 45 minute session was subdivided into three 15 minute epochs. AGENTFLY scanning cannot be time locked to the human eye scan, but over a sufficient interval it should approximate eye tracking data (too long an interval, however, would produce blotted, indecipherable plots). Separate heat maps were generated for each epoch, for each of the 10 HITL participants. Figures 6-8 show the AGENTFLY scan pattern for a given epoch, and corresponding HITL patterns, for scenario B. Figures 9-11 show the same results, for scenario D.[2]

[2] These data are derived from the initial AGENTFLY run, as the scanning module showed extremely consistent behavior across runs.

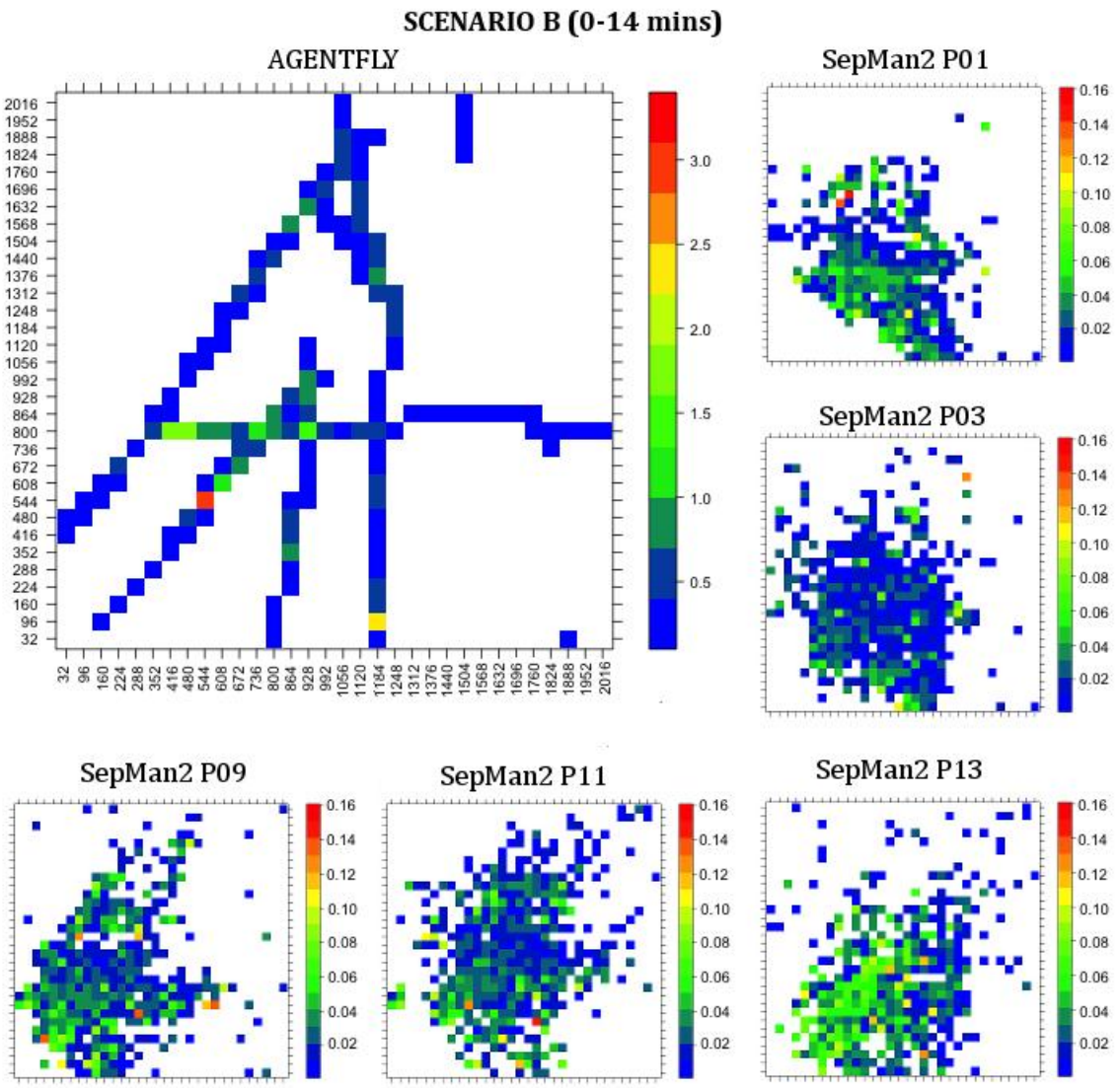


Figure 6. Heat maps of dwell time for AGENTFLY (large panel) and each participant (small panels), in SCENARIO B, first 15 minutes of the simulation.

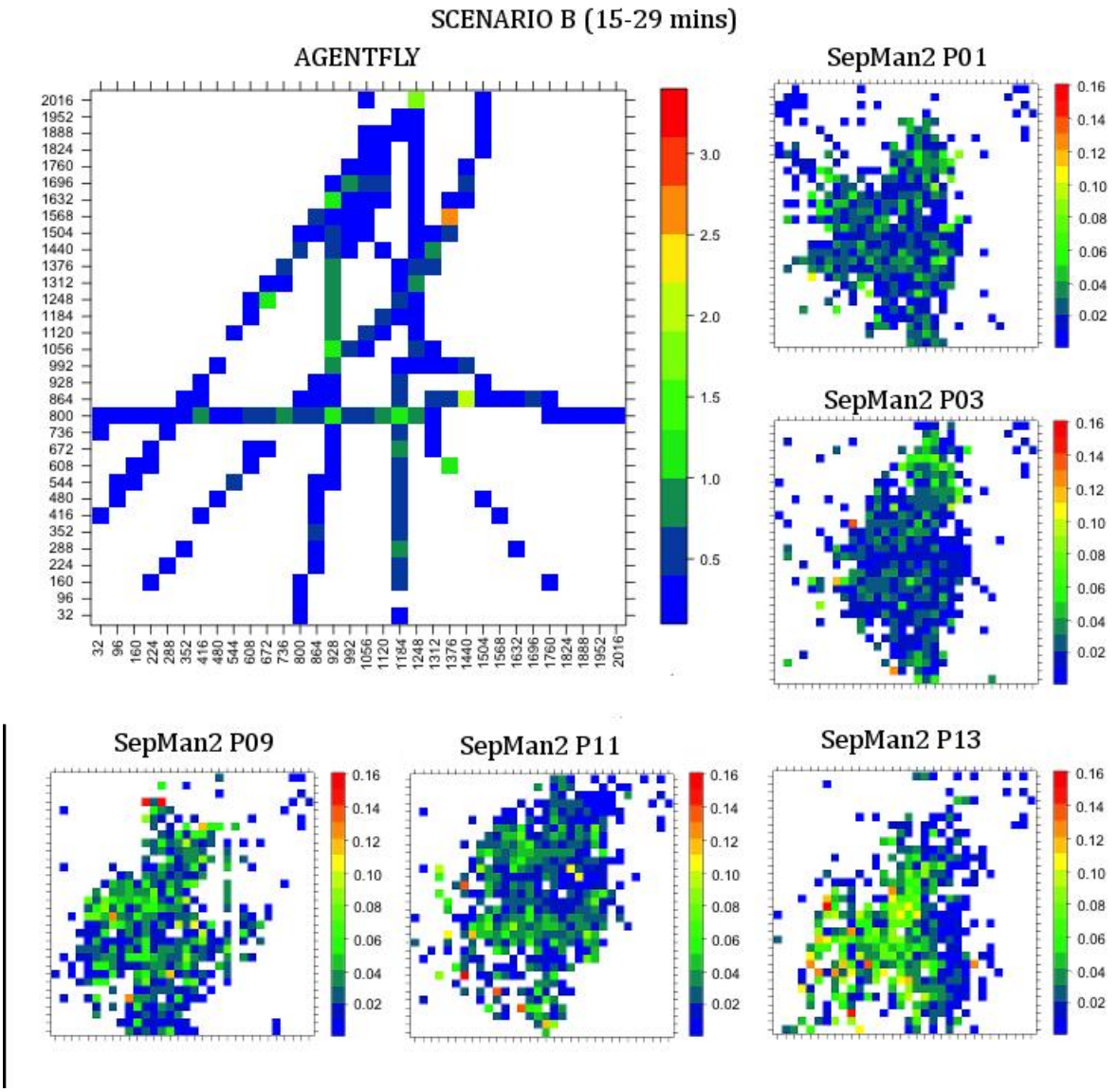


Figure 7. Heat maps of dwell time for AGENTFLY (large panel) and each participant (small panels), in SCENARIO B, minutes 15-29 of the simulation.

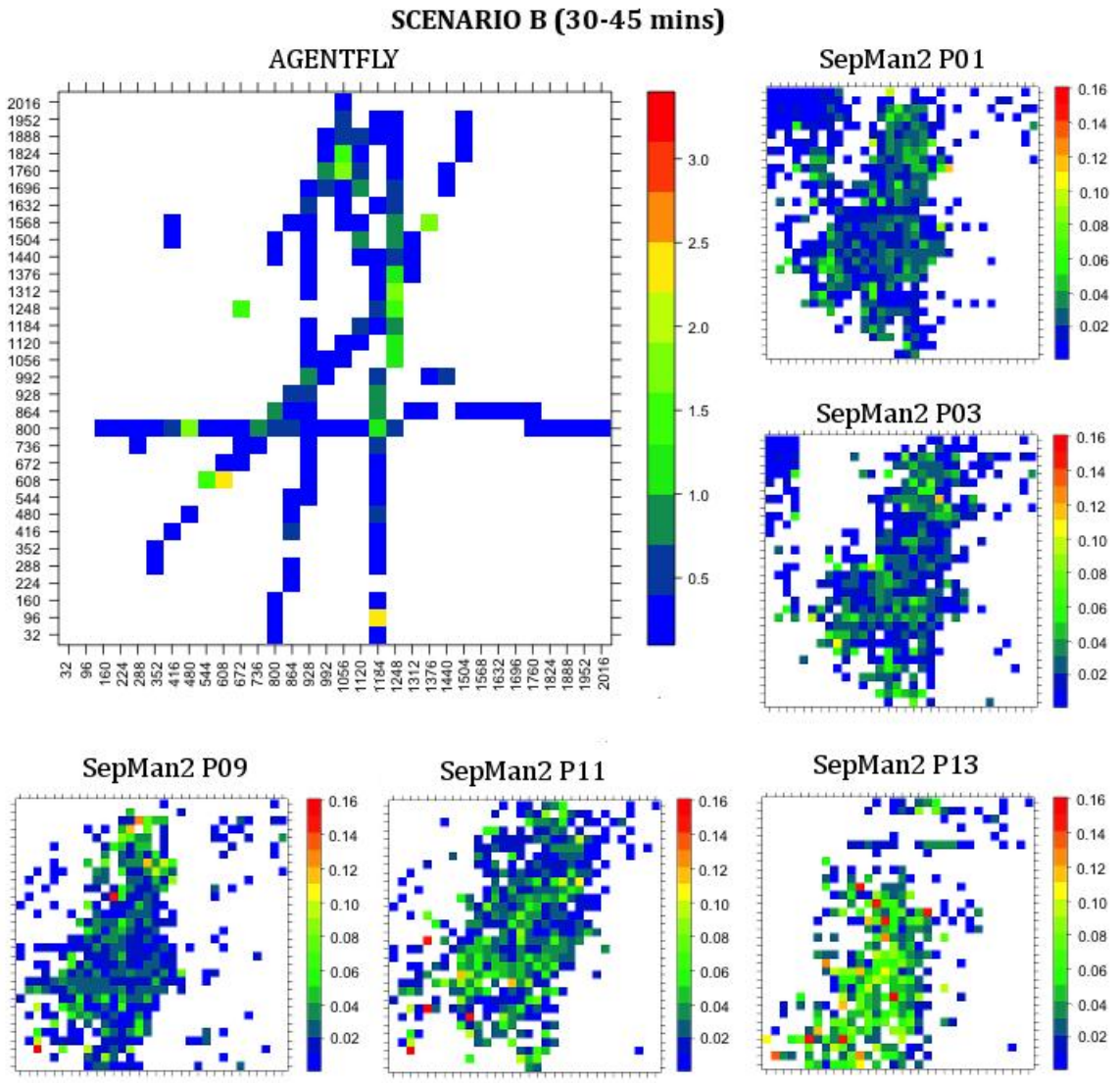


Figure 8. Heat maps of dwell time for AGENTFLY (large panel) and each participant (small panels), in SCENARIO B, minutes 30-45 of the simulation.

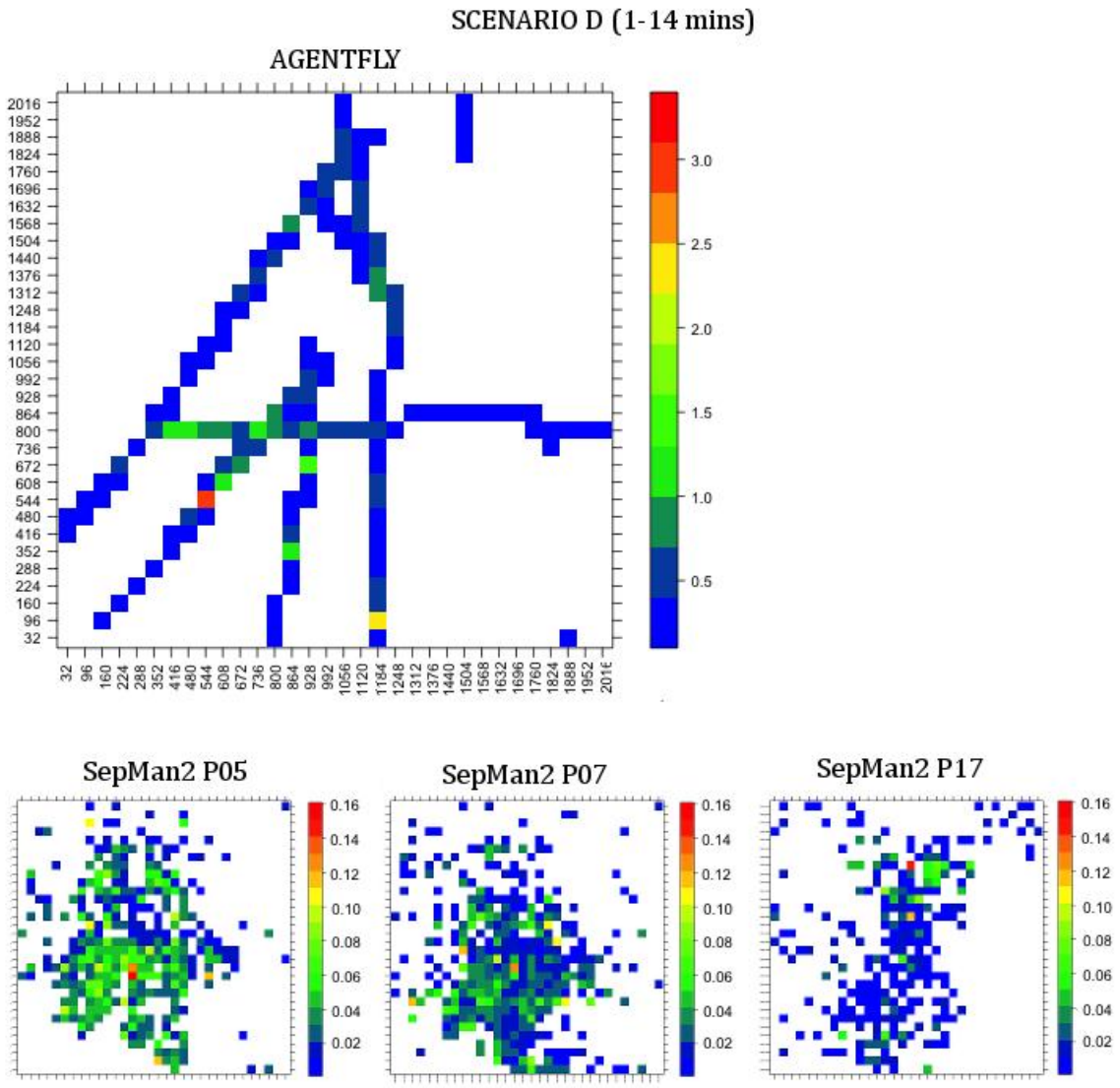


Figure 9. Heat maps of dwell time for AGENTFLY (large panel) and each participant (small panels), in SCENARIO D, first 15 minutes of the simulation.

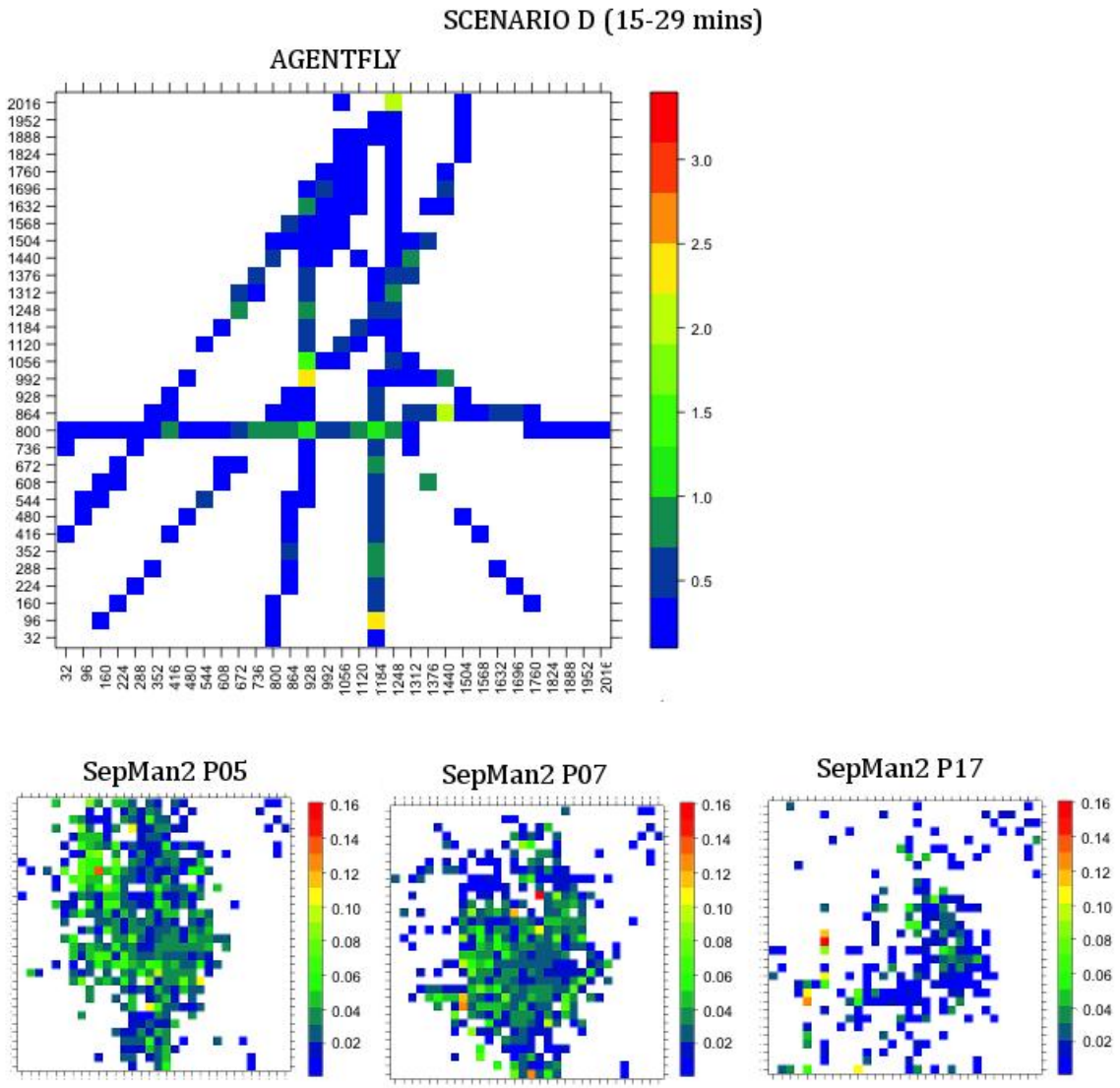


Figure 10. Heat maps of dwell time for AGENTFLY (large panel) and each participant (small panels), in SCENARIO D, minutes 15-29 of the simulation.

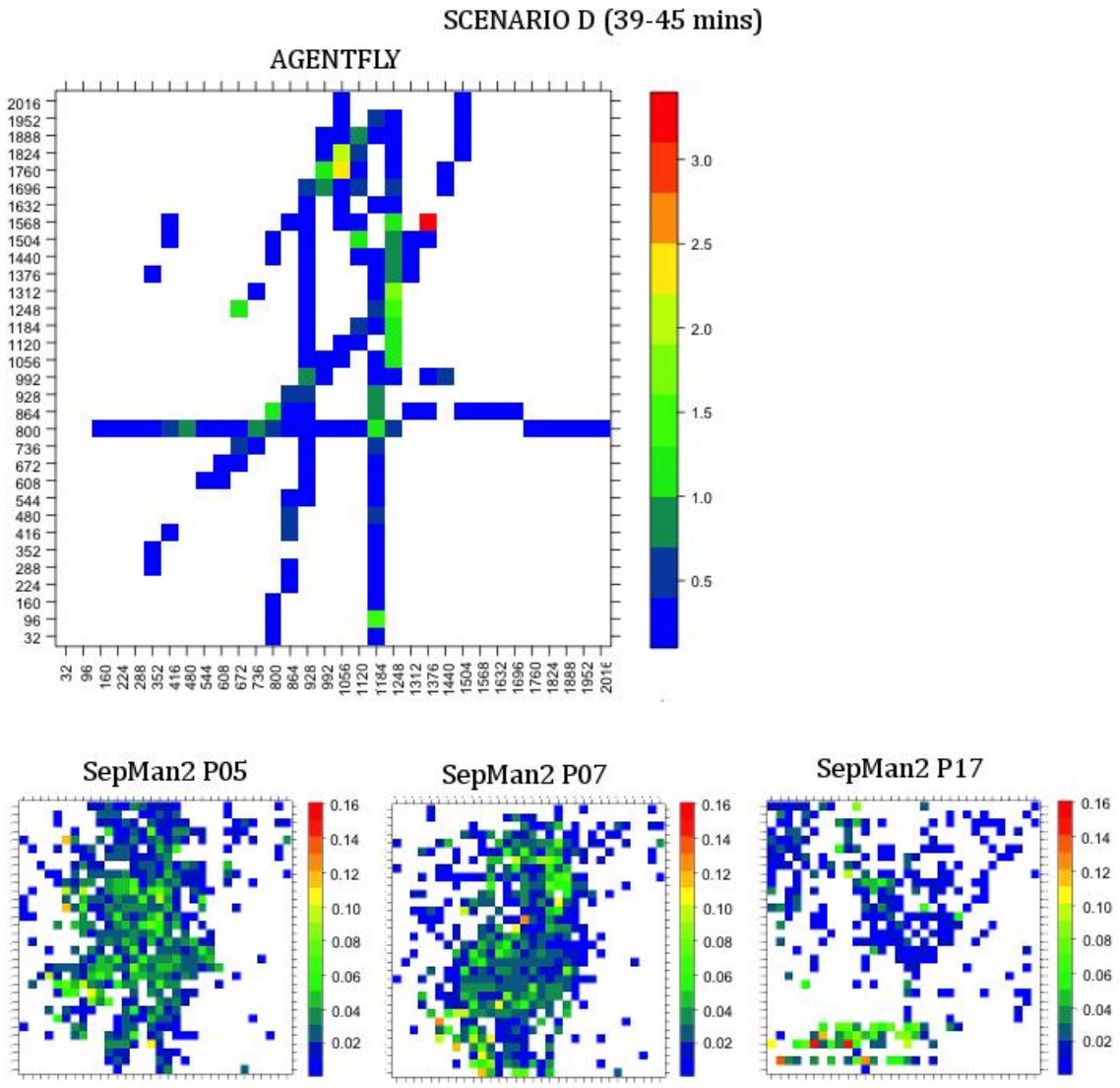


Figure 11. Heat maps of dwell time for AGENTFLY (large panel) and each participant (small panels), in SCENARIO D, minutes 30-45 of the simulation.

A quick visual inspection of the AGENTFLY heatmaps suggests that the system's scan corresponds to a predominant set of major flows. This pattern (notice the main flows at roughly headings 360, 040 and 090, and reciprocals) is less obvious in the SepMan2 scans. Moreover, AGENTFLY's scan pattern was far more defined (i.e., less dispersed) than that of humans. It is not clear to what extent this difference is due to normal human scanning behavior or data collection methods (all but the most invasive eye tracking methods—which immobilize the participant's head—can introduce some measurement variability).

Overall, some of the scanning differences between HITL and AGENTFLY seem to reflect known discrepancies between the scanner's behavior and the typical pattern of human eye movement. Whereas AGENTFLY realistically simulates some aspects of human attention (for example that of dwelling when an interesting perceptual event takes place), it also fails to capture some more complex phenomena (for example that of disengaging attention from areas where the same events occur repeatedly over time, such as in the case of aircraft flying through the intended path). These differences will be further addressed in the Discussion, found in Chapter 4.

3.1.4. Number of aircraft under control

Maximum instantaneous traffic count

Traffic load is known to be one of the prime drivers of task load (and workload) in ATC. Everything else equal, the more aircraft under control, the higher the task load. Given the fundamental role of traffic level, we were therefore interested in first confirming that AGENTFLY could reasonably reproduce the traffic levels experienced under SepMan2.

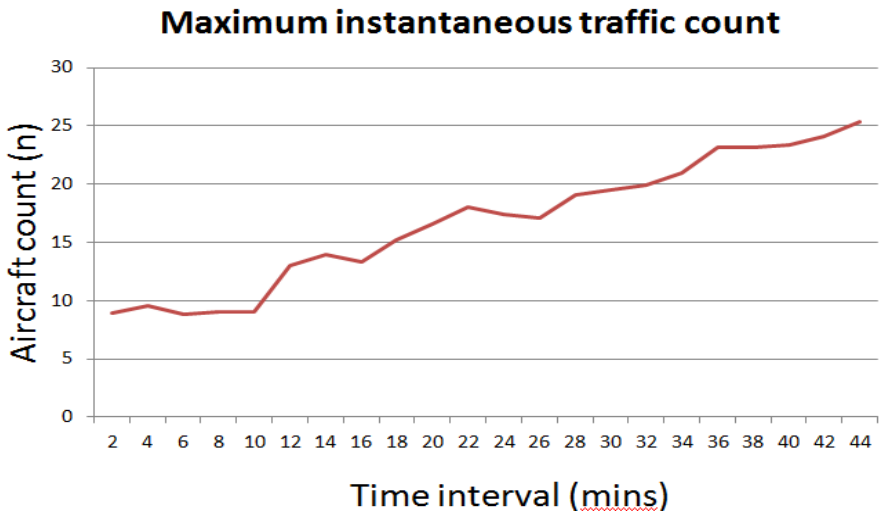


Figure 12. Maximum instantaneous traffic count, averaged across participants for each two-minute interval (from HITL).

For each of the 10 HITL participants, maximum instantaneous traffic load (i.e., the maximum number of aircraft simultaneously under sector control) was computed over successive two-minute intervals. These interval values were then averaged across all 10 participants. As shown in figure 12, traffic load

increased fairly steadily over the 45 minute HITL sessions. Average maximum traffic load ranged from 8.9 aircraft (at the very beginning of the scenarios) to 25.3 aircraft (at the end).

Traffic count, HITL vs AGENTFLY

As shown in table 1, the overall average traffic count (i.e., number of aircraft simultaneously both within sector and under control) was 16.75 and 17.55 for HITL and AGENTFLY, respectively. Over session time, AGENTFLY was also able to closely match the HITL traffic count (figure 13). As shown in figure 13, AGENTFLY-derived traffic count generally fell within the range of HITL values (notice the dashed minimum and maximum lines around the HITL average.

	Avg	Min	Max	Std Dev
HITL	16.75	8.80	25.30	5.47
AGENTFLY	17.55	7.00	27.00	5.85

Table 1. Traffic count, HITL vs AGENTFLY (n)

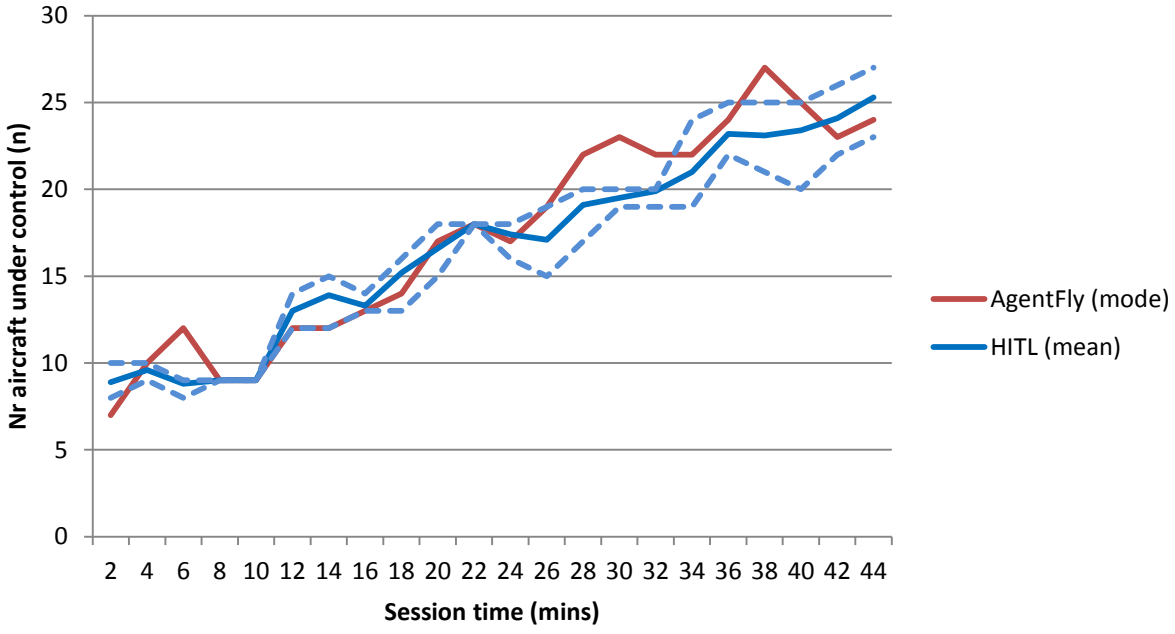


Figure 13. Traffic count over time, AGENTFLY (mode) vs HITL (mean and range). Note HITL range (minimum and maximum values in dashed blue line).

Moreover, there was a strong positive correlation between HITL and AGENTFLY, in terms of traffic count (r=.95, p<.001). This is shown below in the scatter plot of figure 14.

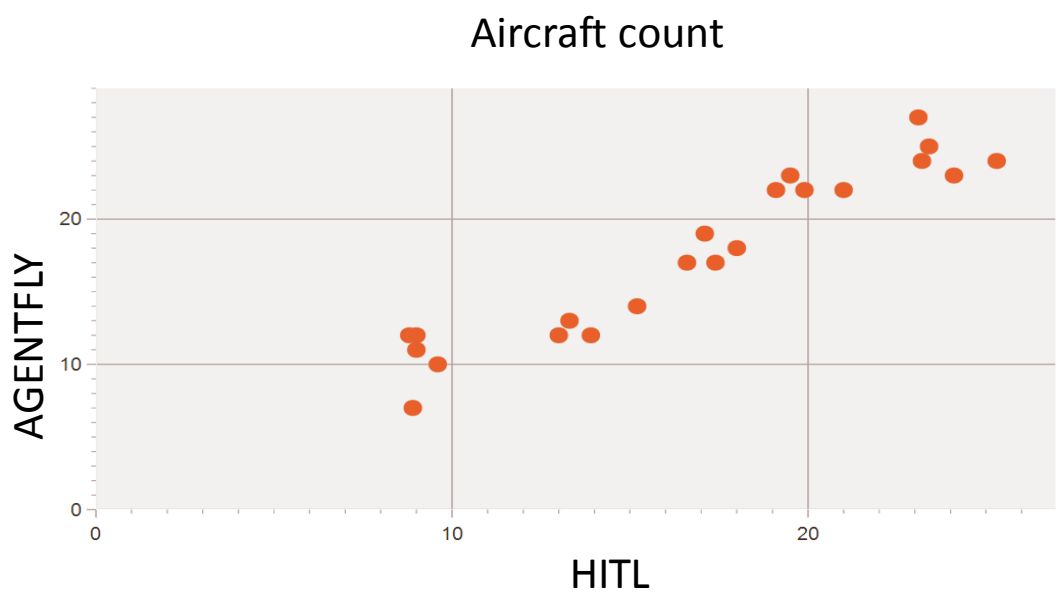


Figure 14. Traffic count, HITL vs AGENTFLY (r=.95, p<.001).

3.1.5 Aircraft mix, climbing and descending

In addition to traffic load, flight attitude (whether aircraft are climbing, descending or level) is known to influence the controller's task load. Everything else equal, a traffic scenario consisting of many climbing and descending aircraft is more difficult to handle than one in which all aircraft are level in cruise. We therefore identified the mix of flight attitudes, as an important traffic flow characteristic that the AGENTFLY model should ideally capture. For HITL data, this mixture was defined (for each participant) as the respective percentage of all under-sector-control flight time (summed across aircraft) that represented ascending, descending, or level flight.

	Ascending	Descending	Level
seconds	7240	61770	467980
percentage	1.3%	11.5%	87.1%

Table 2. Relative time under the three flight attitudes (HITL).

These values were remarkably consistent across participants, as shown in Table 3.

Participant	Ascending	Descending	Level
1	1.20%	14.30%	84.50%
3	1.60%	14.50%	83.80%
5	1.30%	12.20%	86.50%
7	0.80%	11.20%	88.00%
9	1.30%	10.10%	88.60%
11	1.20%	9.50%	89.30%
13	1.30%	10.90%	87.70%
15	1.70%	11.40%	86.90%
17	2.00%	11.40%	86.60%
19	0.80%	10.70%	88.50%

Table 3. Relative flight attitude time, by individual participant (HITL).

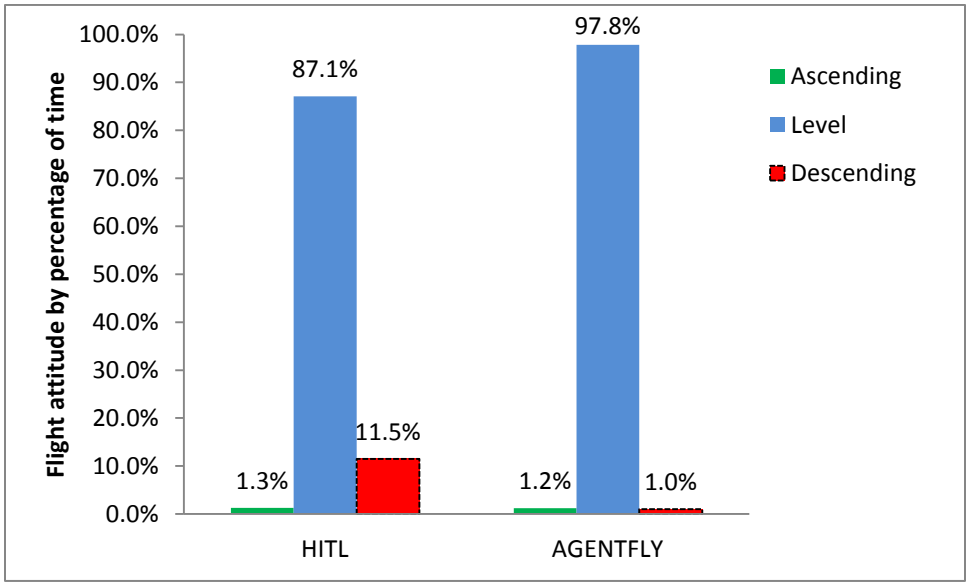


Figure 15. Relative flight attitude time, HITL vs AGENTFLY.

3.1.6 Handoffs

HITL

Coordination requirements were captured by the number of handoffs accepted into and out of the sector. In all cases the main simulation sector interacted with the same six nearby sectors. On average, HITL participants accepted 58.3 handoffs into the sector, and handed off 38 aircraft out of the sector. This works out to a rate of 1.3 inbound handoffs per minute, and 0.84 outbound handoffs per minute. As shown below in table 4, handoff rate was quite similar across participants.

Participant	Inbound nr (45 mins)	Inbound rate (per minute)	Outbound nr (45 mins)	Outbound rate (per minute)
1	60	1.33	38	0.84
3	57	1.27	38	0.84
5	56	1.24	35	0.78
7	58	1.29	36	0.80
9	56	1.24	38	0.84
11	65	1.44	47	1.04
13	56	1.24	35	0.78
15	58	1.29	37	0.82
17	57	1.27	36	0.80
19	60	1.33	40	0.89
1	60	1.33	38	0.84
3	57	1.27	38	0.84
5	56	1.24	35	0.78
7	58	1.29	36	0.80

Table 4. In- and outbound handoff count and rate (per minute), across participants (HITL).

AGENTFLY

AGENTFLY handoffs were identified from "Accept incoming handoff" and "Request outgoing handoff" syntax in the keyboard input data, eight each for the B and D scenarios (thus 16 runs in all). Again, AGENTFLY results showed very slight variability over runs, and averaged 1.07 (sd = .01) handoffs inbound, and .89 (sd = .001) handoffs outbound, per minute.

Run	Inbound nr (60 mins)	Inbound rate (per minute)	Outbound nr (60 mins)	Outbound rate (per minute)
B1	65	1.083333	53	0.883333
B2	64	1.066667	54	0.9
B3	66	1.1	52	0.866667
B4	64	1.066667	53	0.883333
B5	62	1.033333	53	0.883333
B6	65	1.083333	54	0.9
B7	65	1.083333	53	0.883333
B8	65	1.083333	53	0.883333
D1	64	1.066667	54	0.9
D2	65	1.083333	53	0.883333
D3	65	1.083333	53	0.883333
D4	64	1.066667	53	0.883333
D5	65	1.083333	53	0.883333
D6	64	1.066667	54	0.9
D7	64	1.066667	53	0.883333
D8	64	1.066667	53	0.883333

Table 5. In- and outbound handoff count and rate (per minute), across sim runs (AGENTFLY).

	In	Out
HITL	1.30 (.06)	.84 (.08)
AGENTFLY	1.07 (.01)	.89 (.001)

Table 6. Handoffs per minute, average (and SD), for HITL and AGENTFLY.

3.1.7 Flight attitude at sector entry

HITL

Sector entry was defined as any time an aircraft transitioned into the simulation sector (notice that it was possible for a given aircraft to enter the sector more than once). Flight attitude (either ascending, descending or level) was then determined for each of these sector entries. As shown in table 7, 94.2% of aircraft entered the sector in level cruise flight, and roughly 4% and 2% entered in descent and ascent, respectively.

		raw counts			as percentage of total		
Participant	**Total entries**	**Ascending**	**Descending**	**Level**	**Ascending**	**Descending**	**Level**
1	65	1	4	60	1.5%	6.2%	92.3%
3	62	1	5	56	1.6%	8.1%	90.3%
5	61	1	2	58	1.6%	3.3%	95.1%
7	62	1	1	60	1.6%	1.6%	96.8%
9	61	1	0	60	1.6%	0.0%	98.4%
11	69	1	0	68	1.4%	0.0%	98.6%
13	61	1	3	57	1.6%	4.9%	93.4%
15	63	1	3	59	1.6%	4.8%	93.7%
17	61	1	6	54	1.6%	9.8%	88.5%
19	63	1	2	60	1.6%	3.2%	95.2%
MEAN					**1.6%**	**4.2%**	**94.2%**

Table 7. Flight attitude at sector entry (HITL).

AGENTFLY

AGENTFLY entries were almost exclusively level at time of sector entry. Only one entry in each session (and this aircraft was different for B and D scenarios) entered in ascent. None entered in descent. The rest (roughly 98.6%) entered in level flight.

	Ascending	Descending	Level
HITL	1.6%	4.2%	94.2%
AGENTFLY	1.4%	0.0%	98.6%

Table 8. Flight attitude at entry, HITL vs AGENTFLY.

3.1.8 Flight attitude at sector exit

HITL

Sector exit was defined as any time an aircraft transitioned out of the simulation sector (which a given flight could do more than once). Flight attitude (either ascending, descending or level) was then determined for each of these sector exits. As shown in table 9, 82.6% of aircraft exited the sector in level cruise flight, and 17.1% and 0.3% exited in exited in descent and ascent attitudes, respectively.

		raw counts			as percentage of total		
Participant	**Total exits**	**Ascending**	**Descending**	**Level**	**Ascending**	**Descending**	**Level**
1	66	0	8	58	0.0%	12.1%	87.9%
3	43	0	7	36	0.0%	16.3%	83.7%
5	40	0	12	28	0.0%	30.0%	70.0%
7	59	0	6	53	0.0%	10.2%	89.8%
9	71	0	8	63	0.0%	11.3%	88.7%
11	102	3	10	89	2.9%	9.8%	87.3%
13	38	0	8	30	0.0%	21.1%	78.9%
15	57	0	14	43	0.0%	24.6%	75.4%
17	52	0	9	43	0.0%	17.3%	82.7%
19	71	0	13	58	0.0%	18.3%	81.7%
MEAN					**0.3%**	**17.1%**	**82.6%**

Table 9. Flight attitude at sector exit (HITL).

AGENTFLY

For AGENTFLY, only 1.4% of outgoing traffic (i.e. a single aircraft) exited the sector under transition (i.e. non-level cruise): this aircraft exited under descent.

	Ascending	**Descending**	**Level**
HITL	0.3%	17.1%	82.6%
AGENTFLY	0.0%	1.4%	98.6%

Table 10. Flight attitude at exit, HITL vs AGENTFLY.

Notice that under AGENTFLY, aircraft tended to enter and exit the sector at level cruise. In the HITL, on the other hand, aircraft were more likely to enter / exit the sector in vertical transition. As discussed later, AGENTFLY did tend to use altitude clearances on aircraft while in sector, and roughly 8.4% of aircraft were cleared to an interim clearance while in sector.

3.1.9 Flight Level (FL) occupancy rate

Flight Level occupancy was assessed from the absolute number of cleared altitudes, sampled for each aircraft every second of the simulation. The percentage of times each altitude was sampled across participants was calculated and compared to the percentage of times each altitude was sampled across the AGENTFLY runs. The percentage was calculated together for both scenarios (Table 11), after verifying that there were no differences between the two.

	AGENTFLY	HITL
21000	0.0%	0.0%
22000	0.0%	0.0%
23000	0.0%	0.0%
24000	0.0%	17.6%
25000	1.7%	0.4%
26000	0.0%	0.0%
27000	2.5%	1.5%
28000	5.2%	2.5%
29000	2.8%	2.2%
31000	0.0%	1.0%
32000	0.0%	1.6%
33000	5.1%	5.6%
34000	8.3%	9.3%
35000	14.6%	8.7%
36000	12.3%	13.5%
37000	12.1%	8.7%
38000	7.7%	8.4%
39000	15.3%	9.2%
40000	3.8%	6.3%
41000	3.1%	1.0%
43000	5.6%	2.3%

Table 11. Percentage of time the altitude was in use during the simulation, across all participants and all runs of AGENTFLY.

Results revealed a significant positive correlation (r = 0.58; $p < .001$) between the use of altitudes in AGENTFLY and HITL. After noticing that most of the difference occurred at the lowest altitudes (i.e., up to 24000 ft), we repeated the analysis using only altitudes higher than 25000. This also showed a significant positive correlation (r = 0.87; $p < .001$), suggesting an improvement in AGENTFLY's ability to predict which altitudes are going to be used the most. Overall, this result suggests that AGENTFLY is fairly able to predict which altitudes controllers will use in comparable traffic conditions.

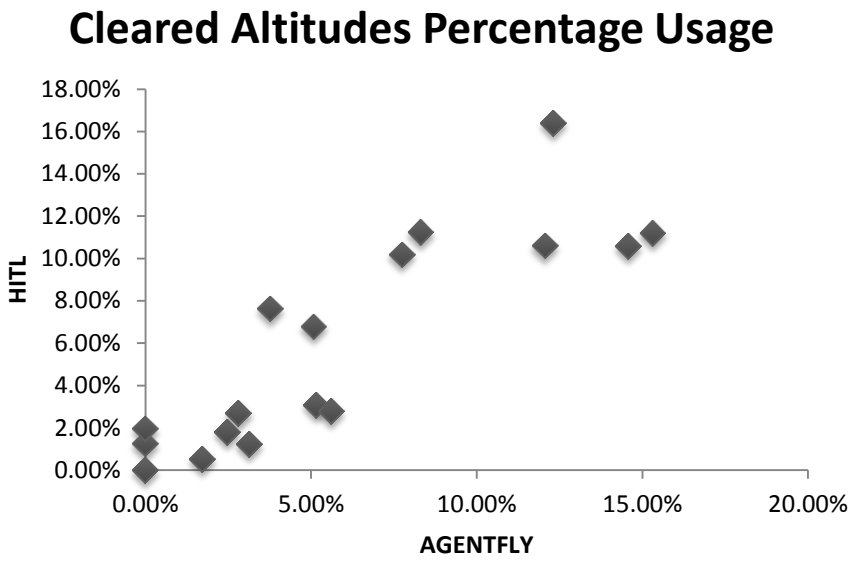


Figure 16. Percentage of time each altitude was in use during the AGENTFLY (x-axis) and SepMan2 HITL (y-axis) simulations respectively.

3.1.10 Radio telephony (RT) rate

One of the chief sources of controller physical activity is communicating via RT, primarily with aircraft. We wanted to confirm that AGENTFLY could reasonably reproduce the RT activity as seen in the HITL simulation.

	Outgoing		Incoming	
Participant	**Rate (calls / min)**	**avg dur (msec)**	**Rate (calls / min)**	**avg dur (msec)**
1	3.04	3520	3.89	3423
3	3.56	2806	3.40	2962
5	3.02	3580	3.42	3465
7	3.02	4315	3.71	3452
9	3.18	3780	3.98	3178
11	3.04	3589	4.11	2563
13	2.96	3045	3.16	3295
15	3.18	2755	3.91	3244
17	3.02	5075	4.09	3227
19	3.16	3711	4.13	2878
MIN	**2.96**	**2755**	**3.16**	**2563**
MAX	**3.56**	**5075**	**4.13**	**3465**
AVG	**3.12**	**3618**	**3.78**	**3169**

Table 12. Rate and duration of RT communications: both outgoing (i.e. ground-to-air) and incoming (air-to-ground), from HITL.

AGENTFLY produced roughly 3.5 incoming (air-to-ground) calls per minute, and 3.63 outgoing (ground-to-air) calls per minute. These RT rates seem very close between HITL and AGENTFLY results. Notice, however, that the pattern is reversed: the majority of AGENTFLY RT calls were outgoing, and the majority of HITL RT calls were incoming.

	RT rate outgoing calls (calls / min)	RT rate incoming calls (calls / min)
HITL	**3.12**	**3.78**
AGENTFLY	**3.63**	**3.12**

Table 13. RT rate, HITL vs AGENTFLY.

3.1.11 Radio Telephony (RT) message duration

RT effort is of course not only determined by how many messages must be conveyed, but also by how long each lasts (as a proxy measure for message complexity, or communication difficulty). As shown in table 14, AGENTFLY tended to underestimate message duration by about 33% and 20% for outgoing (i.e. ground-to-air) and inbound (air-to-ground) messages, respectively.

	RT duration outgoing calls (msec)	Std dev	RT duration incoming calls (msec)	Std dev
HITL	3618	698	3169	289
AGENTFLY	2438	1723	2545	1688

Table 14. RT average duration, HITL vs AGENTFLY.

3.1.12 Conflict alert frequency

We suspected at the outset, and preliminary investigation confirmed, that losses of separation were very rare during HITL sessions. Across all participants, a total of only three conflicts occurred in the simulation sector.

We therefore decided that the data were not rich enough to allow analysis of conflict frequency, or other related metrics.

3.1.13 Keystrokes required per command

The number of keystrokes required for each command was indirectly derived from two different files in SepMan2 and AGENTFLY datasets. The possible use of macros in HITL meant that these data had to

sometimes be decomposed into associated keystrokes. In the end, it was determined that the command types recorded by AGENTFLY differed greatly from those of SepMan2, making the two incomparable. Some of the differences included, but were not limited to: the absence of macros in AGENTFLY and the different classification of specific commands.

3.2 System performance

3.2.1 Traffic density, average sector flight time, average

Average sector flight time was based on total time each aircraft spent in sector 8 (the sector of interest). Exit time was subtracted from entrance time, to obtain the total flight duration (in seconds) within the sector. The average was then calculated for both AGENTFLY and SepMan2 tables, and then compared by means of a t-test. Average time-in-sector for SepMan2 (*M* = 73.28 seconds; *SE* = 2.25) was significantly lower than the AGENTFLY average (*M* = 92.95 seconds; *SE* = 2.93; see Figure 17).

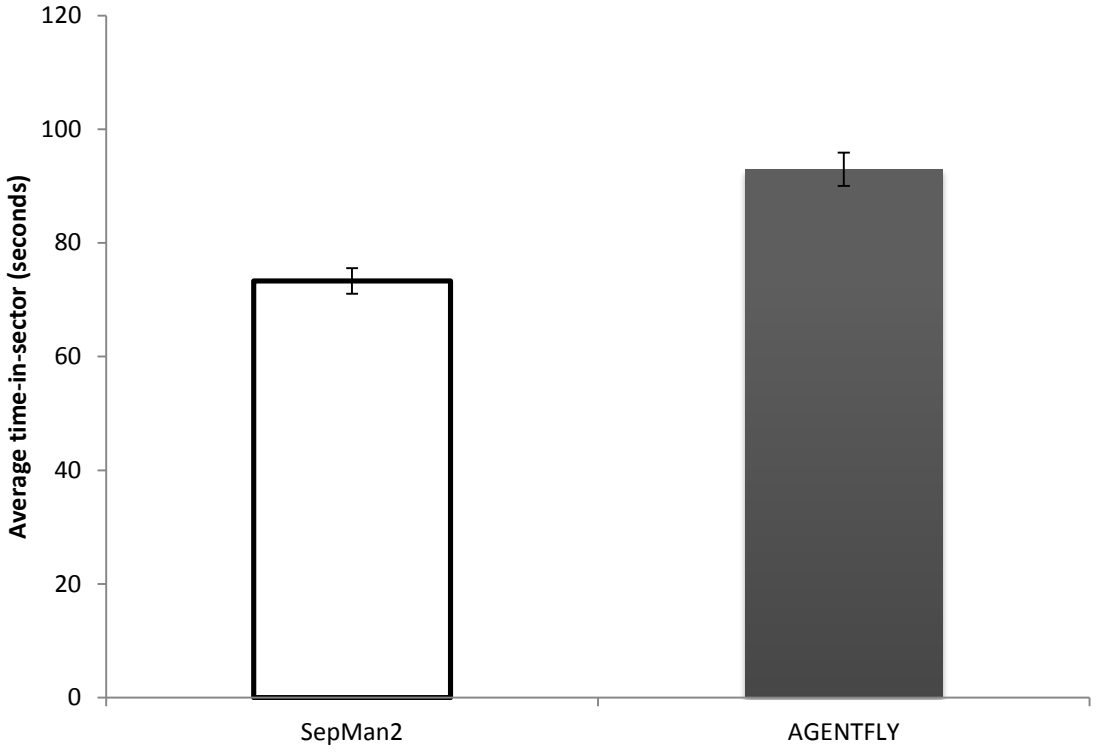


Figure 17. Average time that aircraft spent in sector, HITL vs AGENTFLY.

3.2.2 Traffic density, average sector flight time, per flight

Transit time was calculated for each aircraft in each of the two scenarios, using a table, derived from the CMS through the CPAT tool and filtered by sector 8—that is, considering only aircraft that at some point in the simulation entered the sector under the participant's control.

A very high positive correlation was found between HITL and AGENTFLY flight time for both scenarios, B ($r=.92$; $p < .001$) and D ($r=.84$; $p<.001$).

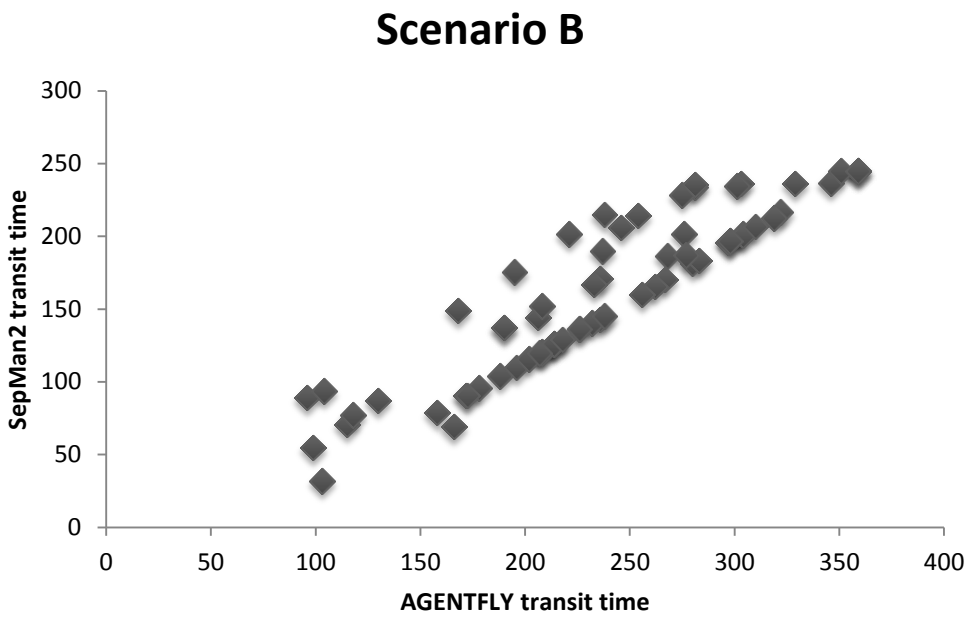


Figure 18. Transit time for each aircraft in scenario B, calculated by AGENTFLY (x axis) and as the result of the participants' performance (y axis).

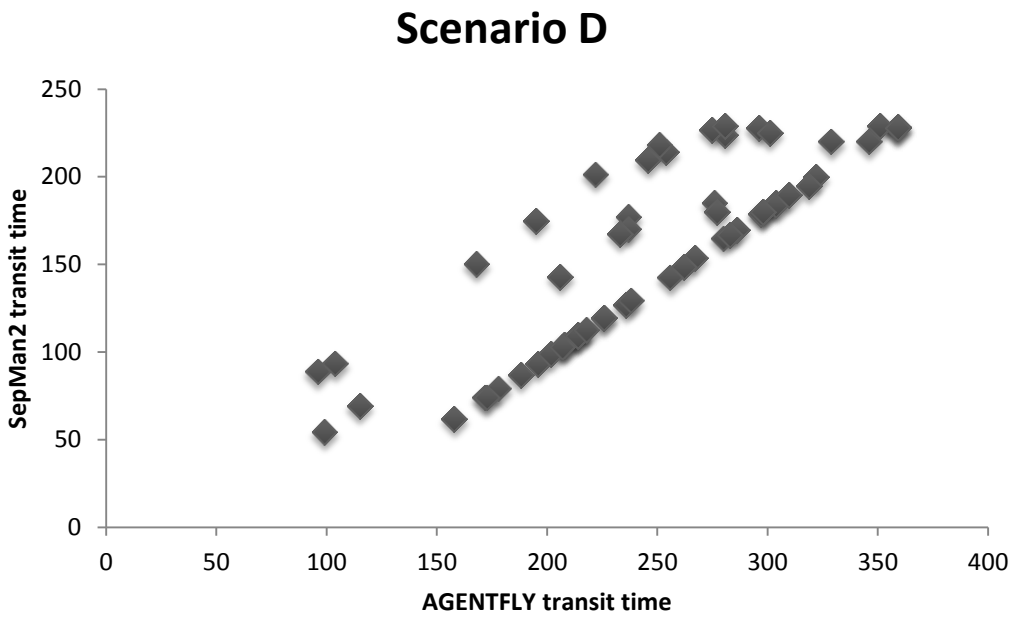


Figure 19. Transit time for each aircraft in scenario D, calculated by AGENTFLY (x axis) and as the result of the participants' performance (y axis).

3.2.3 Aircraft performance, average deviation from cleared trajectory

Both lateral and vertical deviation from cleared trajectory was calculated starting from the CMS derivative table. Average deviation was calculated for both AGENTFLY and SepMan2, in each scenario. Arithmetic means did not seem to differ across scenarios, and data were therefore collapsed across the B and D scenarios.

Levene's test indicated inequality of variance in both metrics (all *p*s<.01); therefore, the degrees of freedom in the t-test were adjusted accordingly. In terms of lateral deviation, HITL and AGENTFLY did not statistically differ, $t(9.12)=0.69$, $p > .1$. In terms of vertical separation, AGENTFLY showed a significantly lower deviation from cleared trajectory than did SepMan2, $t(9) = -18.9$, $p < .001$.

HITL:	mean = 707.40	*SD* = 113.54
AGENTFLY:	mean = 29.18	*SD* = 0.31

.

Table 15: Vertical deviation from cleared trajectory

We confirmed this result via a Bayesian t-test (Kruschke, 2012). Results indicate that credible values for the difference in means were between −24.5 and 30 (95% HDI), suggesting that no difference between AGENTFLY's output and the human participants' data is the most likely hypothesis.

3.2.4 Minimum separation distance

Minimum separation distance was calculated on the basis of the CMS derivative table, which listed all encounters (MinMax ratio < 1) between aircraft that occurred within sector 8—that is, one of the aircraft was in the sector under ATC responsibility.

Minimum separation was calculated for both vertical and horizontal distance. In addition, an analysis on the Minmax ratio was also performed. Neither AGENTFLY nor SepMan2 differed across scenarios ($p > .05$), which were therefore collapsed.

Levene's test indicated inequality of variance in all three metrics (all *p*s<.001); therefore, the degrees of freedom in the t-test were adjusted accordingly. The t-test did not reveal differences between SepMan2 and AGENTFLY on either the average minimum horizontal separation, $t(9.18) = 0.43$, $p < .1$, or the MinMax ratio, $t(9.29) = -1.71$, $p < .1$. However, it revealed that the average minimum vertical separation in AGENTFLY (M = 1941.25; SE = 2.17) was significantly higher than the one in SepMan2 (M = 1606.64; SE = 2.17), $t(9.07) = -9.83$ $p < .001$.

To confirm the similarity between the two simulations, we also performed a Bayesian t-test for the horizontal separation and the MinMax ratio. Results indicate that credible values for the difference in means fell between −0.64 and 0.9 (95% HDI) for the horizontal separation and between −0.03 and 0.005 (95% HDI) for the MinMax ratio, suggesting in both cases that no difference between AGENTFLY's output and the human participants' data is the most likely hypothesis.

4. Discussion and conclusions

The main goal of this validation was to assess how well AGENTFLY could predict human behavior, based on previous HITL data. It is important to keep in mind that AGENTFLY is still under development. Improvements to its predictive and modeling capabilities are ongoing. With that in mind, we tried to capture both strengths and weaknesses of the model objectively, and tried where possible to offer advice on how to improve it.

One of the most important metrics we evaluated in this validation effort was workload. AGENTFLY demonstrated an excellent ability to predict controller workload (and this given the previous caveat about the workload model's simplification of the original Wickens model). As discussed in the introduction, workload is a useful construct in light of the current prediction regarding the forthcoming increase of traffic in the airspace (JPDO, 2007). It could be used by the FAA for budgeting the hire of new controllers, by determining what traffic levels keep the workload rating acceptable or for assigning the appropriate task load to currently employed controllers. Determining those standards by using HITL simulations might be expensive and time consuming, as well as unsustainable in the long run (especially given the likelihood of dramatically new air traffic scenarios over the coming years). Therefore, having a fast time simulation capability, able to calculate workload in these scenarios using low-cost, low-impact solutions might be important. AGENTFLY seems an excellent candidate for a fast-time workload prediction method.

Past research has shown a high positive correlation between SDD and controller workload (r=.84 (Leiden, 2009)), and on the basis of our results SDD seems a reasonable proxy for subjective methods, especially for developmental purposes. SDD values of AGENTFLY and HITL showed a nearly perfect correlation in our analysis. Notice that workload differed (between subjects) across the two scenarios, and this could not be attributed to traffic pattern differences (SDD was comparable across scenarios). Therefore, we are left with at least two possibilities: 1) that subjective workload captures an aspect of the air traffic scenario that neither the number of aircraft under control, nor the SDD metric, is able to capture; or 2) that differences between participants are responsible for this result.

Similarly, we observed a high correlation between the transit time simulated by AGENTFLY and those observed in SepMan2, suggesting that AGENTFLY adequately replicates the amount of time controllers spend on each aircraft.

AGENTFLY proved to accurately predict the rate of RT communication between air and ground. In addition, results showed a quite comparable rate of handoffs (both accepted and outgoing), and comparable flight attitude, with over 90% of aircraft entering the sector in level cruise flight, and over 80% of aircraft exited the sector in level cruise flight, in both simulations.

Results also show that overall AGENTFLY's decision-making patterns were compatible with those of (HITL) controllers. One of the variables that showed great similarity with controllers was the choice of altitudes for the aircraft involved in the simulation. The Flight Level (FL) occupancy rate used by AGENTFLY was very similar to that of altitudes used by ATC, with the exception of the relatively lower altitudes (in the range used within this simulation). This difference might depend in a number of factors, including parameter specification in AGENTFLY or specific instruction given to the participants of SepMan2. Understanding the specific causes for this difference would, however, require further investigation. In addition, results indicate that there was no difference between AGENTFLY and human participants in the minimum distance kept between each pair of aircraft.

Differences did emerge between HITL and AGENTFLY data. Average time in sector differed. In addition, the command types recorded by AGENTFLY differed greatly from the ones of the HITL, and after careful analysis it was determined that such differences made the comparison task impossible. Some of the differences included, but were not limited to: the absence of macros in AGENTFLY and the different classification of the specific commands.

Lastly, several differences between AGENTFLY and human behavior were quite apparent. First, the scanner initiates a fixation every time some event of interest occurs in the display. When no event is recorded, AGENTFLY continuously scans the display as discrete boxes, searching for the next interesting event. The process of visual search, as performed by humans, is obviously quite different. We do not simply scan sequential portions of the visual field. Instead, our eyes might also land on areas of salience (defined for instance by boundaries, contrast regions, brightness). Early studies (e.g., Buswell, 1935) have pointed out a marked central bias in humans, suggesting that when nothing particularly salient in ongoing in the visual scene, humans tend to stare at the center of the display rather than incessantly scanning the visual scene. Second, when humans are performing any visual task that involves locating objects in the visual scene, they tend to ignore locations that were previously scanned and considered not relevant for the current goals. This phenomenon is known as Inhibition of Return (Posner and Cohen, 1984) and it has often been considered evolutionarily advantageous, because it may allow

humans to disregard irrelevant objects or locations. In contrast, AGENTFLY is programmed to scan each location with equal probability, irrespective of whether something relevant was found in that location.

In conclusion, it is clear that AGENTFLY is an excellent candidate for modeling aspects of the human controller (especially workload), and its use would likely be most apparent under early developmental research. In the end, it is important to keep in mind the rationale for using computer modeling (JPDO, 2007). Any computer model necessarily involves some degree of simplification in terms of the tasks and underlying processes. In a sense, the question to be asked is not "does the model perfectly predict HITL results?" but rather "does it predict well enough to add value?"

.

5. References

Abkin, M.H. et al. (2001). Development Of Fast-Time Simulation Techniques To Model Safety Issues In The National Airspace System. CY01 Final Report (Phase III) . Prepared for: National Aeronautics and Space Administration, Ames Research Center. Contract No. NAS2-99072, NEXTOR RR-01-4

Allendoerfer, K.R. & Galushka, J. (1999). Air Traffic Control System Baseline Methodology, DOT/FAA/CT-TN99/15.

Bradford, S., Brown, L.M. & Blucher, M.J. (1998). Use of Performance Metrics in Airspace Systems: US Perspective, 2nd USA/Europe Air Traffic Management R&D Seminar, Dec. 1998.

Breuing, T., Bradford, S. & Liang, D. (2003). Standardizing Performance Metrics, 5th USA/Europe Air Traffic Management R&D Seminar, June 2003.

Buswell, G. T. (1935). How people look at pictures: A study of the psychology of perception in art. Chicago: University of Chicago Press.

Cannon, C.T., Nguyen, D.N. & Regli, W.C. (2012). AGENTFLY NEXTGEN Validation Report, OTA DTFATC-09-OTA-0001.

Casso, N. & Kopardekar, P. (2001). Air Traffic Management System Development and Integration (ATMSDI).

Cheng, V.H.L. (2004). Phase III capacity-increasing concept SOAR (Surface Operation Automation Research) (Technical report OSS-0202-13). Palo Alto, CA: Optimal Synthesis Inc.

Department of Defense (1996). *VV&A Recommended Practices Guide*. Office of the Director of Defense Research and Engineering – Defense Modeling and Simulation Office.

EUROCONTROL (2004). Air Traffic Complexity: Literature Review, Controller Resolution Assistant (CORA) project deliverable D1.

EUROCONTROL (2002a). CARE/ASAS Activity 2: Validation Framework Project – WP2: System Performance Metrics Report.

EUROCONTROL (2002b). CARE/ASAS Activity 2: Validation Framework Project – WP3: Human Performance Metrics Report.

FAA (2010). Air Traffic Modeling of Operational Concepts: Performance Measures and Metrics. Final report, prepared for FAA WJHTC by CSSI.

FAA (2005). Airspace Management Handbook, Metrics, v2.2. Prepared by the MITRE Corporation.

FAA (2004). WJHTC/AOS-300 ARTCC HOST Computer System Interface Requirements Document (IRD) for Air Traffic Applications, NAS-IR-8217-0001.

FAA / EUROCONTROL (2003). Al Concept Validation Strategy Document (OCVSD), December 2003.

Gore, B.F. (2002). An emergent behavior model of complex human-system performance: An aviation surface related application. VDI Berichte 1675, 313-328. Düsseldorf, Germany: VDI.

Gore, B.F., Hooey, B. L., & Foyle, D. C. (2008). Meeting the Challenge of Cognitive Human Performance Model Interpretability Through Transparency: MIDAS v5.x. The 2nd International Conference on Applied Human Factors and Ergonomics, July 14-17 2008, Las Vegas, NV.

Hadley, G.A., Guttman, J.A. & Stringer, P.G. (1999). Air Traffic Control Specialist Performance Measurement Database.

Kruschke, J. K. (2012). Bayesian estimation supersedes the t test. Journal of Experimental Psychology: General, 142(2), 573-603. (doi: 10.1037/a0029146).

Leiden, K., Kopardekar, P., & Green, S.M. (2003). Controller workload analysis methodology to predict increases in airspace capacity. ATIO, Denver, Colorado.

Laudeman, I. V., Shelden, S. G., Branstrom, R., & Brasil, C. L. (1998). *Dynamic density: An air traffic management metric* (NASA/TM-1998-112226). Moffett Field, CA: NASA Ames.

Laughery, K.R. & Corker, K. (1997). Computer Modeling and Simulation of Human/System Performance, in Gavriel Salvendy [ed], Handbook of Human Factors and Ergonomics. New York: Wiley & Sons, 1375-1408.

JPDO (2007). Concept of operations for the next generation air transportation system, Version 2. NextGen. Joint Planning and Development Office, Washington, DC: JPDO.

Magyarits, S. & Kopardekar, P. (2001). Dynamic Density: The Measure Of Air Traffic Control Sector Complexity For The En Route Environment. Phase II Experiment Plan. WJHTC: FAA.

Masalonis, A.J., Callaham, M.B. & Wanke, C.R. (2003). Dynamic density and complexity metrics for realtime traffic flow management. In *Proceedings of the 5th USA/Europe Air Traffic Management R & D Seminar*.

O'Donnell, R.D. (1991). Directions In Performance Assessment: Results Of The Workshop On Future Metrics And Models For Assessment Of Human/System Performance In Advanced Military Systems. CSERIAC report 91-2, University of Dayton Research Institute.

Paglione, M. (2011). CMS_ASCII/CMS_XML Format Definition, Revision F [CMSMemo20110427F.doc]

Posner, M.I. & Cohen, Y. (1984). Components of visual orienting. Chapter in *Attention & Performance X*, [Bouma, H. & Bouwhuis, D., eds]. New York: Erlbaum, 531-556.

Šišlák, D., Volf, P., Pechoucek, M., Cannon, C.T., Nguyen, D.N. & Regliet, W.C. (2012). Multi-agent Simulation of Enroute Human Air Traffic Controller. In Proceedings of the 24th IAAI Conference, 22-26 July, Toronto: Association for the Advancement of Artificial Intelligence.

Šišlák, D., Volf, P., Jakob, M., & Pěchouček, M. (2009). Distributed platform for large-scale agent-based simulations. In Agents for Games and Simulations (pp. 16-32). Springer Berlin Heidelberg.

Stein, E.S. (1985). *Air traffic controller workload: An examination of workload probe* (Report FAA/CT-TN90/60). Atlantic City, New Jersey: FAA Technical Center.

Wickens, C.D. (1984). Processing resources in attention. In R. Parasuraman & D.R. Davies [eds.] Varieties of Attention. New York: Academic Press, 63-102.

Williams,A., Bradford, S., Liang, D., Post, W. & Pomeret, J.M. (2004). Common Metrics Framework for ATM Performance Assessment and Monitoring", Digital Avionics Systems Conference, DASC 04, 2004.

Zingale, C. M., Willems, B., Schulz, K., & Higgins, J. S. (2012). Separation Management (SepMan2): Human Factors Evaluations of Conflict Probe Location and Format and Workstation Display Alternatives. Available through the *National Technical Information Service (NTIS), Springfield, VA.*

Appendix A: Workload in AGENTFLY

The following diagram describes the calculation of workload in AGENTFLY.

An **atomic task** (*t*) executed by the ATC component of AGENTFLY is defined by three variables:

- VCAP resources necessary to complete the task
- Weight $w \in$ <0,7> for each resource r, and
- Task duration *d*

An ATC **activity** (A) is a sequence of ATC tasks:

- Its duration is defined by the sum of durations of each task
- Every activity in the simulation is defined by
  - The **trigger time** A_{trg}
  - The **execution time** from A_{run} to $A_{run}+A_{dur}$
- An activity can be triggered (initiated) by
  - **Scan trigger** (incoming HO, Sector Exit, ...), which is associated with 3 points in time
    - Creation time – when it appears on the screen
    - Trigger time – when ATC notices it
    - Execution time – when ATC starts reacting
  - **Cognitive trigger**
    - Activities memorized by ATC (postponed CR, …)
    - Is associated with **trigger** and **execution** times

Different activity triggers use different VCAP resources.

Scan trigger

- Uses VC (parameter SCAN_TRIGGER_VCAP)

Cognitive trigger

- Uses C (parameter COGNITIVE_TRIGGER_VCAP)

Let tasks $(t_1,t_2,...,t_n)$ be executed in simulation minute m. The **base workload** (without conflicts) of minute m for resource r is defined by

(1) $\sum w_{i,r} * (d_i / 60000)$ for $i \in \{1,...,n\}$

where $w_{i,r}$ is weight of task t_i for resource r and d_i is duration of task t_i in miliseconds. There are 2 types of conflicts in AGENTFLY:

- Incoming Radio Conflict

    - when the currently executed task is interrupted by an incoming radio message
    - Current task is interrupted and finished after receiving the radio message.
- Trigger Conflict
    - when an activity is triggered during the execution of another task
    - The triggered activity is scheduled for future execution but current task is not interrupted.

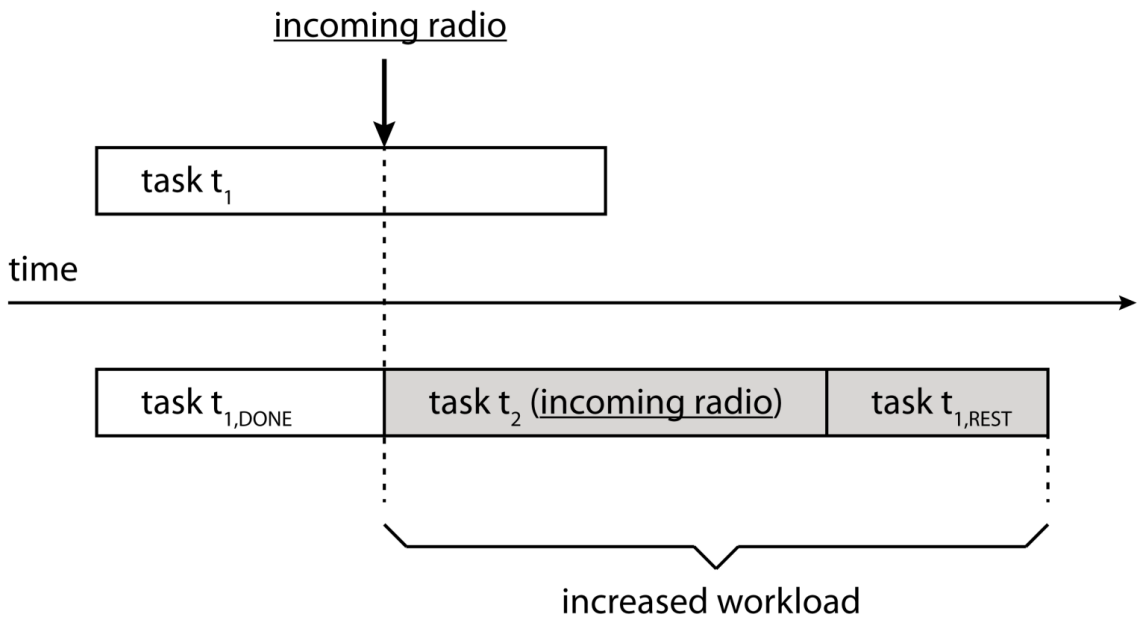


Figure 1. Schematic description of the sequence of events leading to an incoming radio conflict. Task t_1 is interrupted by incoming radio (task t_2) which has to be executed immediately. Task t_2 and the rest of t_1 are executed with increased workload. Only weights for resources **shared** by both the incoming radio task and the interrupted task are increased by the ratio from the Conflict Ratio Table.

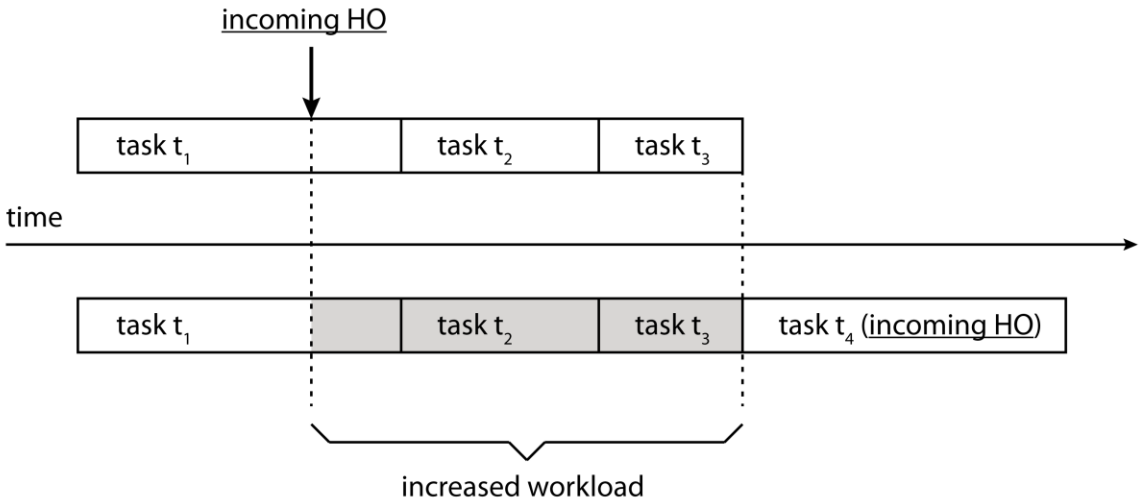


Figure 2. When task t_4 is triggered (noted by ATC) within execution time of another task t_1 then all the tasks (t_1,t_2,t_3) from that time up to the start of actual execution of t_4 are executed with increased workload. Only weights for resources used **both** by the trigger (scan or cognitive) and increased workload tasks are increased.

Conflict Ratios Table

- Weights of tasks executed with increased workload are increased by the ratios from the table below (adjustable in configuration)
- In the case of more conflicts, ratios are summed (e.g. for 2 triggers, conflict in V is 0.2)

(parameters INCOMING_RADIO_CONFLICT_RATIOS, TRIGGER_CONFLICT_RATIOS)

	V	C	A	P
Inc. Radio conflict	0.2	0.4	0.2	0.2
Trigger conflict	0.1	0.2	0.1	0.1

Maximum Trigger Conflict Ratios

- In the case of more triggers, conflict ratios are summed and their sum can get *very high*
- Thus, it is possible to set the maximum conflict ratio by which triggers can increase workload
- Using parameter MAX_TRIGGER_CONFLICT_RATIOS

Minute Workload

- Let tasks ($t_1, t_2, ..., t_n$) be executed in simulation minute m.
- Workload (with conflicts) of minute m for resource r is

(2) $\sum((1 + c_{i,r}) * w_{i,r}) * (d_i / 60000)$ for $i \in \{1,...,n\}$

where $w_{i,r}$ is weight of task t_i for resource r, $c_{i,r}$ is the task conflict ratio for resource r and d_i is duration of task t_i in miliseconds.

Appendix B: AGENTFLY parameters (pilot and ATC parameters)

ATC PARAMETERS

Parameter name	Unit	Probabilistic	VCAP	Description
CD_SECTOR_EXIT_EXTENSION_NM	nm	-	-	defines a distance for which ATC model check for a conflict behind the sector boundary for airplanes which are still inside and are under its control
CHECK_DSA_PROCESS	sec	P	VC	how long does it take to check DSA constraints (DSA = Departure (climb to filed altitude) and Standards Arrival (SOP)) for an airplane entering the sector?
CLIDES_PROCESS	sec	P	-	ATC process duration: the time to notice that an airplane is starting to climb/descend
CLR_ACK_PROCESS	sec	P	C	ATC process duration: the time to process a clearance read back to dermine that it is correct
CLR_CROSS_FIX_OFS_RMD	sec	P	-	radio message duration: the time additional to CLR_MANDATORY_RMD used when the clearance includes a cross fix restriction including offset (like “2 nm south of ORF”)
CLR_CROSS_FIX_RMD	sec	P	-	radio message duration: the time additional to CLR_MANDATORY_RMD used when the clearance includes a cross fix restriction, e.g. {ACID}, climb and maintain {XXX} FL, cross {ABC} at or above {YYY} FL.
CLR_MANDATORY_RMD	sec	P	-	radio message duration: the duration of a radio clearance application message; its manadatory part e.g. {ACID}, climb and maintain {XXX} FL
CLR_REMOVED_KBD	sec	P	VCP	ATC process duration: the time to type the remove interim input to a keyboard
CLR_RL	count	P	-	clearance retry limit (how many times is a clearance re-issued); if it is not successful after CLR_RL attempts, it is logged and AC is removed from the simulation
CLR_RSD	sec	P	-	radio silence duration: the time an ATC will wait before transmitting a radio message (while listening to the radio that the frequency is free of other transmittion)
CLR_SET_KBD	sec	P	VCP	ATC process duration: the time to type the clearance input (set interim or new filed altitude) to a keyboard
CLR_TO	sec	-	-	clearance acknowledgement timeout (any clearance which is not confirmed (read backed) within this time is re-issued)
COLLISION_MINIMUM_DURATION_SEC	sec	-	-	a minimal duration of a collision to avoid numerical instability; the value has to be set close to zero
COLLISION_SAFETY_RANGE_NM	nm	-	-	a horizontal separation limit used by ATC model while searching for conflict resolution
COLLISION_TOO_CLOSE_TO_SECTOR_ENTRY_THRESHOLD_SEC	sec	-	-	a threshold to identify collisions that happen too close (after/before) sector boundary; airplane causing such conflict is logged and removed from the systém

CONTACT_ACK_RMD	sec	P	-	radio message duration: the duration of a contact confiration radio message
CONTACT_ACK_RSD	sec	P	-	radio silence duration: the duration an ATC will listen free frequency before transmitting a radio contact confirmation (while listening to the radio that the frequency is free of other transmittion)
CONTACT_PROCESS	sec	P	C	ATC process duration: the time to process a radio contact message (think about it)
CR_EVAL_PROCESS	sec	P	VC	an additional time to the conflict resolution added for every solution candidate (that is usable for resolution) considered in the conflict resulution process
CR_OPTION_PROCESS	sec	P	VC	an additional time to the conflict resolution per each conflict resolution option generated
CR_PROCESS	sec	P	VC	mandatory time to resolve a conflict of two airplanes
DECONFLICTION_CHECK_RANGE_NM	nm	-	-	a horizontal separation limit used by ATC model while detecting potential future loose of separation
DECONFLICTION_EXFL_SHIFT_STEP_SEC	sec	-	-	a duration by which vertical clearance can be shifted iff there could not be solved a conflict using immediate changed; used to postpone vertical clearance comming from SOP for the sector or to postopone climb to the filed altitude
DIRECT_TO_KBD	sec	P	VCP	the time to type the “DIRECT TO” input to a keyboard when an airplane is skipping a fix during a horizontal return
DISPLAY_VELOCITY_VECTORS_KBD	sec	P	VCP	time required to type the keyboard input to display velocity vectors; time to move up to the next level (1 min, 2 mins, 4 mins and 8 mins after each hit)
DISTANT_COLLISION_THRESHOLD_SEC	sec	-	-	defines a lookahead horizon for which ATC model check for a conflict; conflict occuring after this horizon are checked later
EXPANDED_BOUNDING_BOX_MARGIN_EAST_NM	nm	-	-	extension towards east direction (right)
EXPANDED_BOUNDING_BOX_MARGIN_NORTH_NM	nm	-	-	defines what area is visible (is scanned) by ATC agent around the sector boundary; the visible area is extended bounding rectangle covering sector which by defined margin to north direction (top)
EXPANDED_BOUNDING_BOX_MARGIN_SOUTH_NM	nm	-	-	extension towards south direction (bottom)
EXPANDED_BOUNDING_BOX_MARGIN_WEST_NM	nm	-	-	extension towards west direction (left)
FREQ_CHNG_ACK_PROCESS	sec	P	C	similar to TURN_ACK_PROCESS but for a frequency change radio request
FREQ_CHNG_RL	count	P	-	similar to TURN_RL but for a frequency change radio request
FREQ_CHNG_RMD	sec	P	-	similar to TURN_RMD but for a frequency change radio request
FREQ_CHNG_RSD	sec	P	-	similar to TURN_RSD but for a frequency change radio request; e.g. {ACID}, contact {AB} on {XXX.XX}
FREQ_CHNG_TO	sec	-	-	similar to TURN_TO but for a frequency change radio request
HANDOFF_TIME_MAX_THRESHOLD_SEC	sec	-	-	the earliest time of an airplane to a sector boundary when a handoff procedure is initiated (handoff is initiated depending of controller workload but has to be initiated before HANDOFF_TIME_MIN_THRESHOLD_SEC)
HANDOFF_TIME_MIN_THRESHOLD_SEC	sec	-	-	the latest time of an airplane to a sector boundary when a handoff procedure

				must be initiated
HCHECK_AC_PROCESS	sec	P	VC	time to decide whether two airplanes on the screen has potential future collision considering both horizontal routes and verticla profile
HHO_DET_NTC_PROCESS	sec	P	VC	ATC process duration: the time to notice an request for outgoing horizontal handoff (which is thrown HANDOFF_TIME_MAX_THRESHOLD_SEC before airplane will cross sector boundary, and has to be notices by scanning process)
HHO_RQ_DET_NTC_PROCESS	sec	P	VC	ATC process duration: the time to notice an incoming horizontal handoff (flashing in FDB on the radar screen, ATC model need to scan that region)
HIDE_VELOCITY_VECTORS_KBD	sec	P	VCP	time required to type the keyboard input to hide or shorten velocity vector; again time to move one level down
HO_RQ_ACC_KBD	sec	P	VCP	the time to type a handoff accept input to a keyboard (during an incoming handoff)
HO_RQ_KBD	sec	P	VCP	the time to type a handoff request input to a keyboard (during an outgoing handoff)
HORIZONTAL_DIVERSION_DELAY_SEC	sec	-	-	a maximum pilot's delay to apply a horizontal maneuver considered by ATC model (the max time between the moment when the ATC want to do horizontal turn to the moment when airplane start turning); the time covers radio communication (including read back) and manoeuvre application delay
LBD_KBD	sec	-	VCP	the time to type the “drop full datablock” input to a keyboard (that is, to set a limited datablock)
LEVOFF_PROCESS	sec	P	-	ATC process duration: the time to notice that an airplane has finished a climb/descend phase, thus level off at desired altitude
MAX_SPEED_CHANGE_DELTA_KNOTS	kn	-	-	maximal speed change difference for a speed chande for conflict resolution ever issued to a pilot in CAS
MINIMUM_APPLICATION_TIME_SEC	sec	-	-	the minimal time in which a pilot can possibly apply a maneuver (ATC processes & radio communication included); in combination with DISTANT_COLLISION_THRESHOLD_SEC it defines when ATC model checks for collisions again for the same flight; the flight is rechecked after DISTANT_COLLISION_THRESHOLD_SEC-MINIMUM_APPLICATION_TIME_SEC period
NEAR_LEVOFF_PROCESS	sec	P	-	ATC process duration: the time to notice that an airplane is almost done with a climb/descend phase
NUM_SCAN_REGIONS_COLS	count	-	-	the number of columns by which divide the screen into regular grid for the purpose of screen scanning
NUM_SCAN_REGIONS_ROWS	count	P	-	the number of rows by which divide the screen into regular grid for the purpose of screen scanning
PO_APP_ACCEPT_LLMD	sec	P	-	land line message duration, pointout approving ATC: accept a pointout; {ACID}, point out approved {yyy}.
PO_APP_DENY_LLMD	sec	-	-	land line message duration, pointout approving ATC: deny a pointout; E.g. I

				have traffic. Handoff {ACID} to me {yyy}.
PO_APP_INTRO_ANSWER_LLMD	sec	P	-	land line message duration, pointout approving ATC: response to an introduction message; This is {yyy}, go ahead {xxx}.
PO_RQ_ACCEPT_ACK_LLMD	sec	P	-	land line message duration, pointout requesting ATC: acknowledge after approval; {xxx}
PO_RQ_DENY_ACK_LLMD	sec	P	-	land line message duration, pointout requesting ATC: process a pointout denial message; Wilco, {xxx}
PO_RQ_DET_PROCESS	sec	P	-	ATC process duration: the duration to start a pointout procedure; thinks about pointout request
PO_RQ_DISPLAY_FDB_KBD	sec	P	-	the time to type the input to display full data block on a destination controller screen during point out negotiation
PO_RQ_INTRO_LLMD	sec	P	-	land line message duration, pointout requesting ATC: operator introduction message; {xxx}, {yyy} point-out
PO_RQ_REQUEST_LLMD	sec	P	-	land line message duration, pointout requesting ATC: state a pointout request; Point out, {ACID}, clipping your airspace
POINTOUT_MAX_TIME_IN_SECTOR_SEC	sec	P	-	a pointout procedure is initiated when the duration spend in the next sector is smaller than this paremeter
PREP_DSA_CLR_SEQ_PROCESS	sec	P	VC	how long does it take to prepare DSA changes for a later application?
RADIO_INCOMING_VCAP	vcap	-	CA	VCAP resources an operator uses when listening to a radio
RADIO_OUTGOING_VCAP	vcap	-	CP	VCAP resources an operator uses when speaking to a radio
RETURN_ACK_PROCESS	sec	P	C	similar to TURN_ACK_PROCESS but for a resume own navigation or direct to
RETURN_RL	count	P	-	similar to TURN_RL but for a resume own navigation or direct to
RETURN_RMD	sec	P	-	similar to TURN_RMD but for a resume own navigation or direct to
RETURN_RSD	sec	P	-	similar to TURN_RSD but for a resume own navigation or direct to; e.g. {ACID}, resume own navigation OR {ACID}, direct to {ABC}
RETURN_TO	sec	-	-	similar to TURN_TO but for a resume own navigation or direct to
SCAN_DURATION_AWT_CLI_DES	sec	P	-	screen scanning duration: additional time added for every airplane which ATC is waiting to start climbing/descending within a particular grid scannig region
SCAN_DURATION_AWT_CLR_TRG	sec	P	-	screen scanning duration: additional time added for every airplane for which a vertical clearance is planned to be issued within a particular grid scannig region
SCAN_DURATION_AWT_HO_ACC	sec	P	-	screen scanning duration: additional time added for every airplane awaiting acceptance of a handoff request by other secotr (during outgoing handoff) within a particular grid scannig region
SCAN_DURATION_AWT_LEVOFF	sec	P	-	screen scanning duration: additional time added for every airplane which ATC is waiting to level off at particular flight level within a particular grid scannig region
SCAN_DURATION_AWT_NEAR_LEVOFF	sec	P	-	screen scanning duration: additional time added for every airplane which ATC is monitoring climb/descent within a particular grid scannig region
SCAN_DURATION_AWT_RETURN_TRG	sec	P	-	screen scanning duration: additional time added for every airplane for which a

				horizontal return command is planned to be issued within a particular grid scannig region
SCAN_DURATION_AWT_SEC_V_EXT	sec	-	-	screen scanning duration: additional time added for every airplane which ATC is waiting to receive update interim or filed altitude during vertical handoff
SCAN_DURATION_AWT_TURN_TRG	sec	P	-	screen scanning duration: additional time added for every airplane for which a horizontal turn command is planned to be issued within a particular grid scannig region
SCAN_DURATION_FDB	sec	P	-	screen scanning duration: additional time added for every airplane with a full datablock within a particular grid scannig region
SCAN_DURATION_LDB	sec	P	-	screen scanning duration: additional time added for every airplane with a limited datablock within a particular grid scannig region
SCAN_DURATION_SCREEN_EMPTY	sec	P	-	screen scanning duration: the duration to go through an empty whole screen with no airplanes (this time is included in every screen scanning process); the duration for each scanning grid region is determinded from this total duration by dividing it by the number of grid positions
SCAN_PRIORITY_AWT_CLI_DES	priority	-	-	scan region priority: an extra priority added for a region with an airplane has to start climbing/descending
SCAN_PRIORITY_AWT_CLR_TRG	priority	-	-	scan region priority: an extra priority added for a region with an airplane for ATC is planned to issue a vertical command
SCAN_PRIORITY_AWT_HO_ACC	priority	-	-	scan region priority: an extra priority added for a region with an airplane with initated outgoing handoff but waiting for its acceptance
SCAN_PRIORITY_AWT_LEVOFF	priority	-	-	scan region priority: an extra priority added for a region with an airplane is close to its level off FL
SCAN_PRIORITY_AWT_NEAR_LEVOFF	priority	-	-	scan region priority: an extra priority added for a region with an airplane is climbing/descing
SCAN_PRIORITY_AWT_RETURN_TRG	priority	-	-	scan region priority: an extra priority added for a region with an airplane for ATC is planned to issue a horizontal return command
SCAN_PRIORITY_AWT_SEC_V_EXT	priority	-	-	scan region priority: an extra priority added for a region with an airplane for ATC is waiting for receive update interim of riled altitude during vertical handoff
SCAN_PRIORITY_AWT_TURN_TRG	priority	-	-	scan region priority: an extra priority added for a region with an airplane for ATC is planned to issue a horizontal turn command
SCAN_PRIORITY_FLASHING	priority	-	-	scan region priority: an extra priority added for a region with anything flashing information, e.g. incoming handoff request
SCAN_VCAP	vcap	-	V	VCAP resourcess an operator uses when scanning a screen
SEPARATION_CHECK_MIN_HORIZONTAL_NM	nm	-	-	a minimal horizontal separation limit used by internal component checking that there is no separation violation during the simulation
SPEED_CHANGE_ACK_PROCESS	sec	P	C	similar to TURN_ACK_PROCESS but for a speed change command
SPEED_CHANGE_DELAY_SEC	sec	-	-	a maximum pilot's delay to apply a speed change considered by ATC model (the max time between the moment when the ATC want to do speed change

				to the moment when airplane start changing speed); the time covers radio communication (including read back) and manoeuvre application delay
SPEED_CHANGE_KBD	sec	P	VCP	the time to type a speed change input to a keyboard
SPEED_CHANGE_RL	count	P	-	similar to TURN_RL but for a speed change command
SPEED_CHANGE_RMD	sec	P	-	similar to TURN_RMD but for a speed change command
SPEED_CHANGE_RSD	sec	P	-	similar to TURN_RSD but for a speed change command
SPEED_CHANGE_TO	sec	-	-	similar to TURN_TO but for a speed change command
SPEED_MAX_DEVIATION_COEFFICIENT	coef	P	-	uncertainty in modeling true airplane speed; it means that controller model thinks that true airplane speed can differs from the assigned one +- the provided SPEED_MAX_DEVIATION_COEFfICIENT * indicated radar speed
TCPROBE_CLASS_NAME	class	-	-	class name implementing a tactical conflict probe (atc.faa.atm.enroute.ata.tacticalprobe.simple.SimpleTacticalConflictProbe is the simple linear prediction, atc.faa.atm.enroute.ata.tacticalprobe.safe.SafeTacticalConflictProbe is the TCP implemented by Andrew F.); each TCProbe implementation requires own set of configuration parameters, e.g. SafeTacticalConflictProbe requires about 40 parameters
TURN_ACK_PROCESS	sec	P	C	ATC process duration: the time to process a turn read back to dermine that it is correct
TURN_RL	count	P	-	horizontal turn command retry limit (how many times is a command re-issued); if it is not successful after TURN_RL attempts, it is logged and AC is removed from the simulation
TURN_RMD	sec	P	-	radio message duration: the duration of a horizontal turn application message, e.g. {ACID}, turn right/left {XXX} degrees
TURN_RSD	sec	P	-	radio silence duration: the time an ATC will wait before transmitting a radio message (while listening to the radio that the frequency is free of other transmittion)
TURN_TO	sec	-	-	turn acknowledgement timeout (any turn which is not confirmed (read backed) within this time is re-issued)
USE_HORIZONTAL_CR	yes/no	-	-	should horizontal conflict resolution be used?
USE_SPEED_CR	yes/no	-	-	should speed change conflict resolution be used?
USE_VERTICAL_CR	yes/no	-	-	should vertical conflict resolution be used?
VCHECK_AC_PROCESS	sec	P	VC	time to decide whether two airplanes on the screen could have potential future collision considering their vertical profile only
VELOCITY_VECTORS_RANGE_NM	nm	-	-	if the distance between velocity vectors is below this limit ATC model initiates detailed horizontal check considering flights' airplane routes
VELOCITY_VECTORS_TIME_LOOKAHEADE_SEC	sec	-	-	defines lookahead duration within which the

				VELOCITY_VECTORS_RANGE_NM is tested; it has to be the same value as is used for DISTANT_COLLISION_THREASHOLD_SEC
VERTICAL_SEPARATION_FL	FL	-	-	a minimal vertical separation of two airplanes which are not separated horizontally
VHO_CLR_ACK_PROCESS	sec	P	C	similar to CLR_ACK_PROCESS but for a clearance issued during a verticall handoff
VHO_CLR_RL	count	P	-	similar to CLR_RL but for a clearance issued during a vertical handoff
VHO_CLR_TO	sec	-	-	similar to CLR_TO but for a clearance issued during a vertical handoff
VHO_DET_NTC_PROCESS	sec	P	VC	ATC process duration: the time to notice an request for outgoing vertical handoff, similar to HHO_DET_NTC_PROCESS
VHO_RQ_DET_NTC_PROCESS	sec	P	VC	ATC process duration: the time to notice an incoming vertical handoff; similar to HHO_RQ_DT_NTC_PROCESS
WAITING_FOR_RADIO_ACK_VCAP	vcap	-	CA	VCAP resources an operator uses when waiting for an acknowledgment radio message

PILOT PARAMETERS

Parameter name	**Unit**	**Probabilistic**	**Description**
CLR_ACK_CROSS_FIX_OFS_RMD	sec	P	radio message duration: an additional time for a clearance that contain a cross fix restriction with an offset
CLR_ACK_CROSS_FIX_RMD	sec	P	radio message duration: an additional time for a clearance that contain a cross fix restriction
CLR_ACK_MANDATORY_RMD	sec	P	radio message duration: the time to transmit a clearance acknowledgment; mandatory component similar concept like CLR_MANDATORY_RMD, CLR_CROSS_FIX_RMD and CLR_CROSS_FIX_OFS_RMD
CLR_ACK_RSD	sec	P	radio silence duration: the time a pilot will wait for free frequency before transmitting read back of received clearance
CLR_APPLICATION	sec	P	the time a pilot needs to apply a clearance (vertical maneuver); apply to its autopilot/FMS
CLR_RECEIVED	sec	P	the time to process received clearance through radio

CONTACT_ACK_RECEIVED	sec	P	the time a pilot needs to process acknowledgement (response of ATC agent) to a radio contact
CONTACT_RL	count	-	retry limit: how many times should we try to re-transmit a radio contact message; if it is not successful after CONTACT_RL attempts, it is logged and AC is removed from the simulation
CONTACT_RMD	sec	P	radio message duration: the duration to transmit a contact radio message
CONTACT_RSD	sec	P	radio silence duration: the time a pilot will wait for free frquency before transmitting of a contact message
CONTACT_TO	sec	P	timeout for a radio contact (a contact message is re-transmitted when not acknowledge before this time)
FREQ_CHNG_ACK_RMD	sec	P	radio message duration: the duration to transmit an acknowledgment to a frequency change radio command
FREQ_CHNG_ACK_RSD	sec	P	radio silence duration: the time a pilot will wait for free frequency before transmitting a radio acknowledgment
FREQ_CHNG_RECEIVED	sec	P	the duration of processing of a frequency change radio message
HANDOFF_TIME_MAX_THRESHOLD_SEC	sec	-	the time of an airplane to a sector boundary when a handoff procedure is initiated (outside controlled sectors - in Ghost airspace, i.e. when an airplane is going to enter a controlled sector)
NEW_RADIO_FREQ	sec	P	the time a pilot will wait after radio tuning and before transmitting the contact message on the new frequency
RADIO_TUNING	sec	P	the time a pilot needs to tune the radio to a new frequency
RETURN_ACK_RMD	sec	P	like CLR_ACK_RMD but for a horizontal return command
RETURN_ACK_RSD	sec	P	like CLR_ACK_RSD but for a horizontal return command
RETURN_APPLICATION	sec	P	like CLR_APPLICATION but for a horizontal return command

RETURN_RECEIVED	sec	P	like CLR_RECEIVED but for a horizontal return command
SPEED_CHANGE_ACK_RMD	sec	P	like CLR_ACK_RMD but for a speed change command
SPEED_CHANGE_ACK_RSD	sec	P	like CLR_ACK_RSD but for a speed change command
SPEED_CHANGE_APPLICATION	sec	P	like CLR_APPLICATION but for a speed change command
SPEED_CHANGE_RECEIVED	sec	P	like CLR_RECEIVED but for a speed change command
TURN_ACK_RMD	sec	P	like CLR_ACK_RMD but for a horizontal turn command
TURN_ACK_RSD	sec	P	like CLR_ACK_RSD but for a horizontal turn command
TURN_APPLICATION	sec	P	like CLR_APPLICATION but for a horizontal turn command
TURN_RECEIVED	sec	P	like CLR_RECEIVED but for a horizontal turn command

Appendix C: Final metric list

Construct	Metric	AGENTFLY predictor	HITL criterion
Workload	Workload	Computed workload	Subjective workload (WAK)
	Dynamic Density	Computed SDD	Computed SDD
	Visual scan	Scan module: spatial distribution and dwell time	Eye point-of-gaze: spatial distribution and dwell time
Task load	Nr aircraft under sector control	Max concurrent traffic, by two-minute interval	Max concurrent traffic, by two-minute interval
	Aircraft mix, climbing and descending	Percentage climb / descent / level	Percentage climb / descent / level
Coordination	Handoffs, inbound and outbound	Handoff rate, inbound and outbound	Handoff rate, inbound and outbound
	Attitude, flight entry	Percentage climb / descent / level	Percentage climb / descent / level
	Attitude, flight exit	Percentage climb / descent / level	Percentage climb / descent / level
Flow organization	Nr altitudes used	Cumulative percentage occupancy time, by FL	Cumulative percentage occupancy time, by FL
Radio communications	RT rate	RT call rate (calls per min), inbound and outbound	RT call rate (calls per min), inbound and outbound
	RT duration	Mean and std dev of RT call duration, inbound and outbound	Mean and std dev of RT call duration, inbound and outbound
Effectiveness / Efficiency	Distance flown in sector	Total and average distance flown in sector	Total and average distance flown in sector
	Clearances, all	Relative proportion of all clearances	Relative proportion of all clearances
	Trajectory clearances	Relative percentage of altitude, heading and speed clearances	Relative percentage of altitude, heading and speed clearances
	Keystrokes required per command	Total and average keystrokes required	Total and average keystrokes required
System performance, capacity	Sector flight time, across flights	flight time under sector control, averaged across all flights	flight time under sector control, averaged across all flights
	Sector flight time, by flight	flight time under sector control, for each individual flight	flight time under sector control, for each individual flight
	Aircraft spacing	Average deviation from cleared flightpath, lateral and vertical	Average deviation from cleared flightpath, lateral and vertical
Safety and cost	Min separation distance	MinMax ratio	MinMax ratio

Appendix D: Simplified Dynamic Density (SDD)

DD component	Weight
Occupancy counts	2.2
Altitude changes	0.3
Proximities*	0.4
Sector boundary crossing	0.5
Heading variance	0.0005
Speed variance	0.0005
Traffic density	30000

* $P = (4*P1 + 2*P2 + P3 + P4) / 4$

where P1, P2, P3 and P4 indicate the number of Level 1, 2, 3 and 4 proximities, respectively.

From Klein, Rodger and Leiden (2009).