Determining Air Traffic Controller Proficiency
Identifying Objective Measures Using Clustering

de Jong, T. P.; Borst, Clark

DOI
10.1016/j.ifacol.2022.10.120

Publication date
2022

Document Version
Final published version

Published in
IFAC-PapersOnline

Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnline CONFERENCE PAPER ARCHIVE

IFAC PapersOnLine 55-29 (2022) 7–12

Determining Air Traffic Controller Proficiency: Identifying Objective Measures Using Clustering

T. P. de Jong, Clark Borst [1]

Control and Simulation, Faculty of Aerospace Engineering, TU Delft, 2629 HS, Delft, The Netherlands

Abstract: Air traffic control (ATC) is a complex and demanding job reserved for highly-trained professionals. Training ATC candidates is challenging as trainees are subjectively assessed by instructors who are biased by their own ways of working. As an effort to determine control expertise objectively, this study employed clustering techniques on an existing data set in which course and professional controllers participated in a medium-fidelity simulation experiment. Results identified a set of eight measures that formed two distinct and stable expertise clusters. A subsequent sensitivity analysis was able to reveal how far (or close) each course participant was positioned from the expert cluster and on which measures those participants deviated from the experts. At this stage, however, it is difficult to translate these results into specific advice on how to improve underdeveloped skills. Despite the small sample size and limited generalizability of the results in this exploratory study, the method appears to be a promising demonstration in determining objective factors that describe ATC expertise, warranting further research.

Keywords: Decision making and cognitive processes, Human centred automation, Shared control, cooperation and degree of automation

1. INTRODUCTION

Air traffic control (ATC) is a complex and demanding job reserved for highly-trained professionals who need to process large amounts of dynamically changing information while maintaining a good balance between safety and efficiency (Schuver-van Blanken et al. (2010); Oprins et al. (2006)). Not surprisingly, training ATC candidates is challenging and this community is currently facing relatively high drop-out rates in the three to four years of training due to insufficient expertise (Schuver-van Blanken et al. (2010)). This is undesirable, because it drives the cost of training and it could eventually lead a shortage of controllers (Federal Aviation Administration (2013)).

A contributing factor in reduced training efficiency is that a trainee's expertise level is mainly established by *subjective* assessments done by ATC instructors who are biased by their own ways of working (Federal Aviation Administration (2013)). Additionally, these assessments are performed during high-fidelity simulator sessions at a relatively late stage in training. Ideally, *objective* assessments should be done earlier and more frequently during training. This would enable the assessor (and the trainee) to get more insight into the developed skills over time (e.g., learning curves) and subsequently provide trainees with specific guidance toward areas of improvement (see Figure 1).

In this paper a first step is made in this direction by devising a method that seeks a set of objective measures that would help to establish the expertise level of a trainee. To find such a set, clustering analysis (i.e., a branch of machine learning) is applied on an existing data set recorded in medium-fidelity ATC simulator sessions in which university staff, who completed a five-day ATC course, and retired ATC professionals participated. Note that the application of machine learning techniques to determine expertise levels have already been successfully performed in other fields, such as surgery (Watson (2014)). However, no prior research has been conducted in objectively establishing the expertise level of air traffic controllers (ATCos).

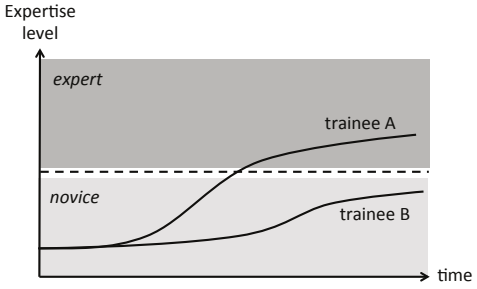


Fig. 1. Possible hypothetical learning curves, where trainee A is most promising in acquiring sufficient control expertise within the training time.

2. AIR TRAFFIC CONTROL PROFICIENCY

The main job of controllers is to ensure a safe, orderly and expeditious flow of air traffic in their sector (Oprins et al. (2006)). The expertise level required for this job is determined by a set of related competences. After

[1] E-mail: c.borst@tudelft.nl

10.1016/j.ifacol.2022.10.120

conducting a literature survey, it can be concluded that expert controllers:

- have a set of 'best practices' in solving conflicts (Fothergill and Neal (2008); Kallus et al. (1999))
- are consistent in the type of instructions (Kallus et al. (1999))
- adopt safety buffers to account for uncertainty (D'Arcy and Della Rocco (2001))
- minimize sector disruptions (Kirwan and Flynn (2001); Oprins et al. (2006))
- create expeditious traffic flows (D'Arcy and Della Rocco (2001))
- create solution spaces (Schuver-van Blanken et al. (2010))
- show less variability in procedures (Schuver-van Blanken et al. (2010))
- have efficient visual scanning patterns (van Meeuwen et al. (2014))

Additionally, anecdotal evidence from interviews with ATC professionals indicated that experts are proactive in handling traffic. That is, ATCos tend to organize traffic upon sector entry in patterns that prevent conflicts from emerging.

Based on these metrics, objective measures can be devised to determine control expertise. For example, experts tend to vector slow aircraft behind the faster aircraft in solving conflicts. The adoption of safety buffers can be measured by the separation distances between aircraft from radar data. Sector disruptions and expeditious traffic flows can be measured by route deviations (causing additional flown track miles) and sector outflow, respectively. A way to measure solution spaces could be operationalized by measuring the available opportunities in speed, heading and altitude to solve conflicts. Further, deviation and/or compliance with sector rules (e.g., the exit speed and/or altitude of aircraft leaving the sector) as well gaze patterns on sector areas of interest, measured by eye-tracking equipment, could indicate expertise.

It is clear that many possibilities in measures exist that could signal expertise. However, it is not clear which set of measures would articulate control expertise best. To find answers, clustering analysis on radar and instruction data can shed light on this matter.

3. DATA DESCRIPTION

3.1 Participants

The radar and instruction data used in this research consists of data from four different air traffic scenarios solved by ten different controllers, recorded in previous research (Somers et al. (2019)). Four participants were retired controllers (representing the Professional group) and six participants, all Delft university staff, completed a multiple-day ATC course (representing the Course group). The experience of the professional group ranged from 33 to 35 years of experience. Two of them were Area Controllers (ACC) and two were Approach Controllers (APP).

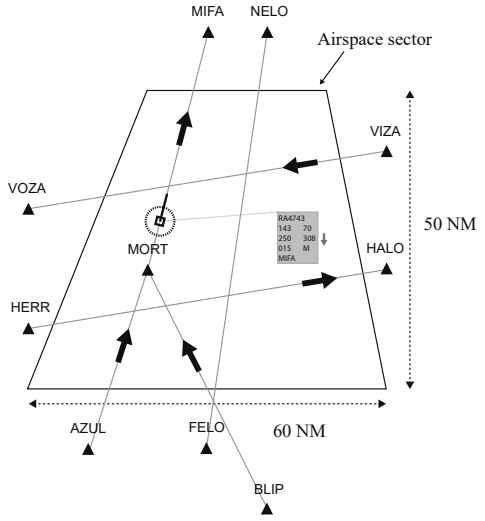


Fig. 2. Overview of the simulator screen showing the sector layout, waypoints, standard routes and one aircraft visible in the middle (Adapted from Somers et al. (2019))

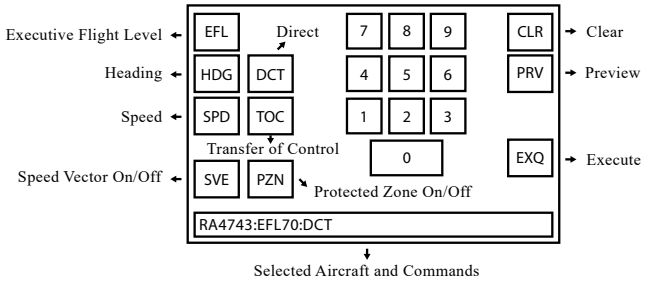


Fig. 3. The command window participants had to use to control the aircraft (Adapted from Somers et al. (2019))

3.2 Simulation environment

A simplified, medium-fidelity desktop ATC simulator, developed by Delft University of Technology, was used which showed a sector comparable to the Amsterdam South Sector in the Netherlands (Figure 2). The participants could control the traffic using a separate control window, which could be operated by using a mouse or a touchscreen (Figure 3). The traffic was controlled by clicking on the aircraft blips (or clicking on the flight label) and then providing a clearance using the command window. Aircraft could be controlled by altitude (EFL), heading (HDG), and/or speed (SPD) commands. Another way to control the flight direction was by giving aircraft "direct-to" (DCT) clearances toward their designated exit waypoints. Aircraft could also be transferred to the adjacent sector by a transfer-of-control (TOC) instruction.

A predicted loss of separation occurring within 120 seconds was made visible by changing the aircraft color to orange. A predicted loss of separation within 60 seconds changed the aircraft color to red. All aircraft had the same 5 NM protected zone. Furthermore, an option was present to turn the protected zone circle (PZN) and the speed trend vector (SVE) on/off to aid the participants separate the traffic. However, this was not logged in the data.

A few simplifications were made compared to a high-fidelity ATC simulator. First, interaction with aircraft was done by a computer mouse and thus no voice R/T was simulated. Second, aircraft entering the sector did not follow the sector routes, but needed to be vectored toward their exit waypoints. Third, clearances given through the command interface were always executed by the aircraft immediately (i.e., no pilot delay). Fourth, there were only three aircraft categories, which were shown in the aircraft label: light, medium or heavy. Fifth, aircraft motion was simulated by kinematic equations and flight performances (e.g., climb and descent rates, min/max speeds, acceleration, turn rates, etc.) were modeled by look-up tables.

3.3 Task and scenarios

The task of the controllers was to separate traffic and hand them over to the adjacent sectors at certain predefined flight levels (see Somers et al. (2019)). Before the aircraft left the sector a 'transfer of control' (TOC) had to be instructed.

Each participant solved four different scenarios. The differences between the scenarios was mainly characterized by the number of aircraft and the traffic mix entering the sector along the routes. Each scenario had a duration of 20 minutes.

3.4 Data set and limitations

Due to simulator limitations, not all indicators of expertise were directly measured and/or available. For example, voice R/T was replaced by the command window inputs. No eye-tracking equipment was used in the experiment, resulting in no data about visual scanning patterns. The obtained data consist of two files per controller per scenario. One file contains the given clearances to the aircraft using the command window, including a timestamp in seconds. The other file contains the radar data of the aircraft in the scenario. This file includes, per logpoint, among others, the aircraft position, (cleared) flight level, (cleared) heading and (cleared) speed. Based on this information, expertise measures such as safety buffers, added track miles, solution space areas needed to be reconstructed. A logpoint was recorded every 3 seconds during the experiment. With 10 participants, this resulted in 40 radar and 40 clearance files.

4. METHODOLOGY

A graphical overview of the data analysis methodology to elicit the best subset of expertise measures is provided in Figure 4. In a nutshell, the Feature Selection Wrapper does the bulk of the work. There, a genetic algorithm selects the set of measures that maximizes a fitness function based on clustering results. Given that two expertise groups (i.e., professional and course participants) exist in the data, the set of measures that result in two distinct and robust clusters will indicate the candidate measures that would articulate expertise best.

4.1 Measures

A total of 59 measures are used in which a set of measures can be extracted by the genetic algorithm (Table 1). These

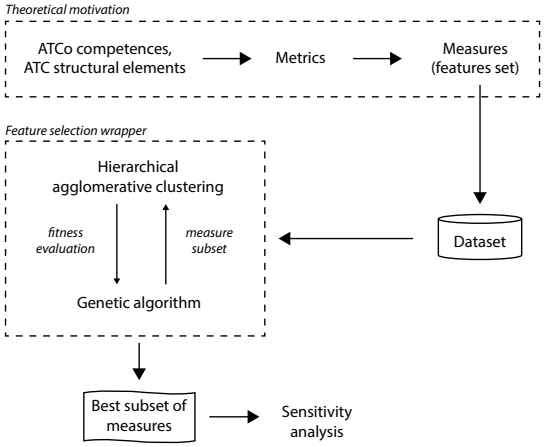


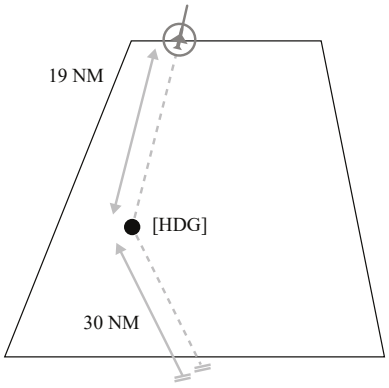
Fig. 4. Flow diagram of the measure selection process

Table 1. Measures corresponding to ATCo expertise

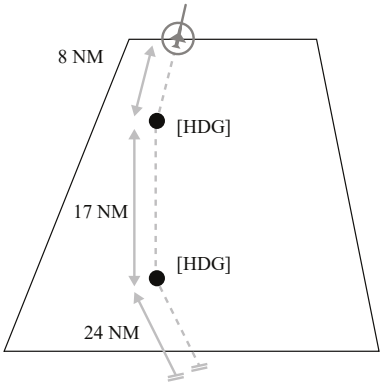
	Clearances	**Measures**
1	Consistency	Number of DCT, EFL, HDG and SPD commands [-]
2	Workload	Total number of commands; Number of level changes per aircraft [-]
	Safety	**Measures**
3	Safety buffers	Relative distance between aircraft [nm]
4	Availability of solution space	Mean occupied solution space area at each given clearance [-]; Total occupied solution space area of every aircraft in the sector [-]
	Efficiency	**Measures**
5	Maximize efficiency	Time spent in sector [s]; Additional track miles [nm]; Number of aircraft reaching their waypoint [-].
6	Moment of traffic handling	Track penalty when using EFL, HDG, DCT or SPD commands [nm^2].
7	Expeditious traffic flows	Outflow of traffic in the sector [s^{-1}]
	Procedural compliance	**Measures**
8	Variability in procedures	Altitude of aircraft leaving the sector [ft]

59 measures are gathered per ATCo per scenario. The majority of the 59 measures are of central tendency (like the mean) or of variability (like the standard deviation, maximum and minimum values). Also, ratios, summations and mean squared errors (MSE) are used. These measures aim to summarize the generated data of each individual ATCo. The majority of measures speak for themselves, but the track penalty and solution space areas requires more explanation.

Considering that ATCos are proactive, it is expected that more experienced ATCos will give all level, heading or speed changes far before the aircraft leaves the sector. A possible way to represent this metric is by summing the squared track miles when a level, heading or speed command is given is used for each aircraft (see Equation 1). Figure 5 shows an example of this track penalty when using heading clearances for a single aircraft. For each aircraft, and each clearance type, the sum of the squared track miles



(a) ATCo A issues one heading change and introduces a track penalty of $30^2 = 900$ NM^2



(b) ATCo B issues two heading changes and introduces a track penalty of $24^2 + (24 + 17)^2 = 2257$ NM^2

Fig. 5. ATCo A handles the aircraft far before the aircraft leaves the sector compared to ATCo B. This is represented by a lower track penalty for ATCo A compared to ATCo B

can be obtained. These sums are taken together to get a single sum of squared track miles for each command type. This results in a track penalty when using level, heading or speed commands. The ratio between these track penalties and the sum of these track penalties are used as measures, resulting in 4 measures.

$$\text{Trackpenalty: } \sum_{\substack{\text{squared track miles} \\ \text{at command}}} = a^2 + b^2 + ... \quad (1)$$

a Track miles of aircraft at first command
b Track miles of aircraft at second command

According to the findings of Schuver-van Blanken et al. (2010), ATCos create solution spaces or use the solution space that is already available. To assess this metric, the occupied Solution Space area was used (Somers et al. (2019)), where it is expected that experienced controllers create solution spaces and therefore minimize the occupied solution area (or, maximize the available solution area). The occupied area is represented by a ratio of the total available area with a value between 0 and 1. The total area is the total annular area bounded by V_{min} and V_{max} as shown in Figure 6. The mean occupied SSD area of every aircraft in the sector at each given command could be used to see to what extend the solution space changes between commands. The mean, standard deviation and the ratio

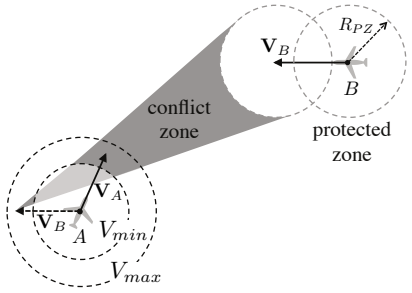


Fig. 6. The Solution Space Diagram (SSD), where the triangular conflict zone (i.e., velocity obstacle) caused by aircraft B reduces the available solution area (i.e., conflict-free heading and speed options between V_{min} and V_{max}) of aircraft A.

between an increase or a decrease in solution space are used as measures.

4.2 Feature Selection

The feature selection wrapper selects from the measures a subset of measures that lead to the most distinct clusters describing the different ATCo expertise groups. First, an initial subset from the measures is constructed. The wrapper loop then evaluates the performance of the formed clusters. This performance is used as a fitness criterion for the genetic algorithm. Based on the fitness evaluation, the genetic algorithm creates a new subset of measures which is, again, used as an input for the clustering algorithm to determine the performance of this subset. The genetic algorithm creates new subsets based on the current best performing subset and when the new subset is better than the current best-performing subset, the new subset becomes the best performing subset. After a stop criterion has been reached, the output of the wrapper will be the best performing subset of measures.

Since the number of samples (ATCos and scenarios) is small, it is desired to directly get information about the relationship between the samples. Therefore, hierarchical clustering was used which organizes the data in a hierarchical structure according to a normalized distance matrix. It is also desired to create clusters containing ATCos with similar experience levels and that the ATCos within each cluster are close to each other. Therefore, the total within-cluster sum of squares should be minimized, which is something Ward's method aims to achieve. Finally, the Euclidean distance measure wasused in Ward's method.

5. RESULTS

5.1 Identified measures and clusters

A set of eight measures were found that best describe the expertise level of the ATCos in the data, see Table 2. Clustering results of the complete set (training and test) using the eight measures from Table 2 can be summarized graphically in a dendrogram with heatmap (Figure 7). From the dendrogram, it can be seen that two clusters are formed: one cluster with only pro ATCos (the pro-cluster) and one cluster with mainly course ATCos (the course-cluster). While the accuracy of the pro-cluster is

Table 2. Measure set identified by the feature selection wrapper

M1	Ratio between the number of given DCT commands and the total number of given DCT, HDG, EFL and SPD commands
M2	Ratio between the number of given EFL commands and the total number of given DCT, HDG, EFL and SPD commands
M3	Ratio between the number of given HDG commands and the total number of given DCT, HDG, EFL and SPD commands
M4	Ratio between the total sum of squared track miles when a level command is given and the total sum of squared track miles when a level, heading or speed command is given
M5	Ratio between the total sum of squared track miles when a heading command is given and the total sum of squared track miles when a level, heading or speed command is given
M6	The aircraft with the highest flight level of all aircraft flying to waypoint MIFA
M7	The mean over all logpoints of the average time to closest point of separation per logpoint
M8	The maximum over all logpoints of the average time to loss of separation per logpoint

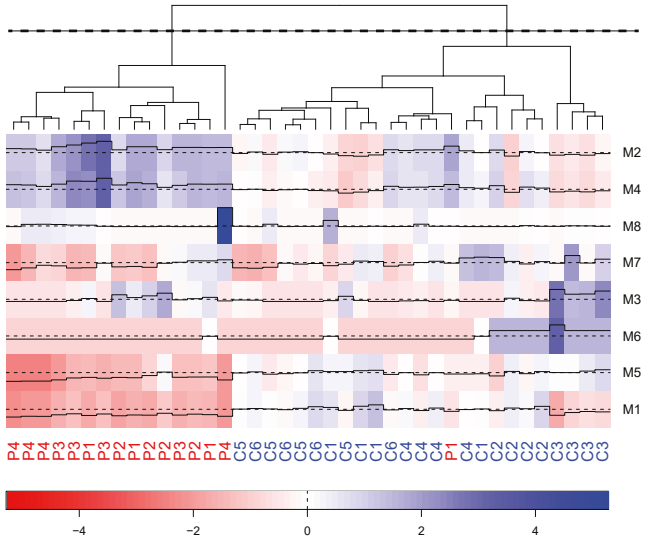


Fig. 7. Heatmap and dendrogram containing the clustering results and individual measure results

perfect, the course-cluster is not 100% accurate due to the presence of a single pro-ATCo. Further, each row in the heatmap corresponds to the scaled value of the corresponding measure for each ATCo and each scenario. The dashed line in each row corresponds to the average value of the course-cluster. The color is red when the measure is below course-cluster average and blue when the measure is above it. The black continuous line in each measure row shows the size of the standardized value relative to the other ATCos and scenarios in the same measure row.

In general, it is expected that the resulting measures of the ATCos in the course cluster will be close to the dashed line. This can indeed be observed in the heatmap for all measures in the course-cluster. Furthermore, it is expected that the ATCos in the pro-cluster will differ uniformly from the dashed line. This can also clearly be observed in measures M2, M4, M5 and M1. For the other measures, however, this is not the case.

Table 3. Sum of squared distances between the ATCos and the course- and the pro-cluster centroids.

	C1	C2	C3	C4	C5	C6	P1	P2	P3	P4
Pro	84,0	77,0	117,2	34,6	56,3	45,7	9,4	15,0	10,7	36,4
Course	15,3	19,2	49,3	11,6	12,7	11,5	50,5	40,0	67,6	92,1

By observing the dendogram and corresponding heatmap, many interesting observations can be made. First, the professional group gives far less DCT clearances (M1) and have a low track penalty from HDG clearances (M4) compared to the course group. Second, the professional group also has a higher EFL clearance ratio (M2) and have a higher EFL track penalty (M4) relative to the course group. Especially the higher EFL track penalty was contrary to what was expected, because experts are usually proactive, which should thus result in early altitude clearances, resulting in a low track penalty. One contributing factor to this result is the usage of less DCT and HDG clearances (relative to EFL commands) for the professional group. The expert ATCos were mainly driven by minimizing their workload and therefore accepted slight deviations from the exit waypoints. The course participants were driven by optimizing for performance and therefore aimed to steer aircraft directly toward the exit waypoints as best as possible, something that can be achieved by DCT commands.

A third observation from the dendogram is that participant C3 formed its own separate subcluster. On the majority of measures, C3 deviated from both expertise groups and was also quite consistent in that, resulting in a separate subcluster. Finally, it also seems that measure M8 by itself does not contribute much in forming the two distinct clusters.

5.2 ATCo distances to clusters

For each scenario of an ATCo, the distances of all measures to the centroid of the pro-cluster are squared and summed together. After that, the sums of squared distances are taken together to get a single sum of squared distance for each ATCo. The same is done for the distances of all measures to the centroid of the course-cluster. Table 3 shows the resulting sum of squared distances.

Looking at C3 in Table 3, it can be seen that he or she is relatively distant (117,2 units) from the pro-cluster compared to the rest of the course group. Furthermore, C3 is also relatively distant (49,3 units) to his or her own course-cluster compared to the rest of the course group. Therefore, it can be stated that C3 neither behaved like a course-ATCo nor a pro-ATCo, which echoes the result observed in the dendogram. Omitting C3 will therefore prevent the clustering algorithm to categorize C3 and will therefore lead to higher accuracies.

Considering C4 in Table 3, it can be observed that he or she is relatively closer (34,6 units) to the pro-cluster than to the rest of the course group. Furthermore, C4 is also still close (11,6 units) to the course cluster compared to the rest of his or her course group. Therefore, it can be stated that C4 behaved like a course ATCo, but also

shows signs of moving toward a professional ATCo. Again, this is not clear behavior for the clustering algorithm, just like C3. Omitting C4 will therefore prevent the clustering algorithm to categorize C4 and will therefore lead to higher accuracies.

6. DISCUSSION

The goal of this study was to identify, using clustering techniques, a set of *objective* measures that could establish the level of expertise of trainee ATCos. The underlying motivation is to help improve training efficiency by monitoring trainee learning curves throughout the training and provide trainees specific guidance toward areas of skill and knowledge improvement.

The clustering results of a small dataset, containing 6 ATC course participants and 4 professionals, revealed a set of eight measures that resulted in two stable clusters of high accuracy. Together with a sensitivity analysis inspecting the contribution and distances of each participant to the clusters, it can be traced how close (or far apart) each course participant is from the professional cluster and on which specific measure(s) a participant deviates from the average expert. Performing this analysis more frequently during training would enable an ATC instructor to monitor the trainees' progress and to select trainees who need more attention in terms of guidance and support. At this point, however, it is difficult to provide participants specific advice on how to improve. For example, in the experiment experts gave relatively less DCT clearances. Advising a trainee to simply "give less DCT clearances" is rather meaningless without providing any context within which those clearances should and should not be given.

Despite the promising results, this study also has its limitations. First of all, clustering the small dataset might have shown signs of overfitting, indicated by the algorithm including measures such as M6 and M8 that do not seem to contribute significantly in differentiating between the two clusters. These measures are probably selected because they resulted in a higher fitness value for this data. Second, the results only apply to this specific data set. When using another simulator and/or different traffic scenarios, new clustering will be required. To mitigate this, effort should be undertaken to make the measures more context independent. Third, when adding more ATCos to the data set, there is also a possibility that the current measures are not sufficient anymore and the clustering accuracy decreases too much. When this happens, a different set of measures needs to be found that can describe the accuracy of this larger group of ATCos. The advantage is that this new set of measures might describe the expertise of a larger group of ATCos. A greater confidence in the measures can therefore be developed as the data set of ATCos increases.

7. CONCLUSION

In an effort to more objectively establish the expertise level of an ATC trainee, this study employed clustering techniques on an existing data set in which course and professional controllers participated in a medium-fidelity simulation experiment. Results identified a set of eight measures that formed two distinct and stable expertise clusters. A subsequent sensitivity analysis was able to reveal how far (or close) each course participant was positioned from the expert cluster and on which measures those participants deviated from the experts. At this stage, however, it is difficult to translate these results into specific advice on how to improve underdeveloped skills. Despite the small sample size and limited generalizability of the results in this exploratory study, the method appears to be a promising demonstration in determining objective factors that describe ATC expertise, warranting further research.

REFERENCES

D'Arcy, J.F. and Della Rocco, P.S. (2001). Air Traffic Control Specialist Decision Making and Strategic Planning - A Field Survey. Technical report, Federal Aviation Administration, Atlantic City, NJ.

Federal Aviation Administration (2013). Review and Evaluation of Air Traffic Controller Training at the FAA Academy. Technical report, U.S. Department of Transportation.

Fothergill, S. and Neal, A. (2008). The Effect of Workload on Conflict Decision Making Strategies in Air Traffic Control. *Proceedings of the Human Factors and Ergonomics Society 52nd Annual Meeting*, 52(1), 39–43. doi:10.1177/154193120805200110.

Kallus, K., van Damme, D., and Dittmann, A. (1999). Integrated Task and Job Analysis of Air Traffic Controllers – Phase 2: Task Analysis of En-route Controllers. Technical report, EUROCONTROL, Brussels, Belgium. doi:Technical Report HUM.ET1.ST01.1000-REP-04, European.

Kirwan, B. and Flynn, M. (2001). Identification of Air Traffic Controller Conflict Resolution Strategies for the CORA (Conflict Resolution Assistant) Project. Technical report, EUROCONTROL Experimental Centre, Brétigny, France.

Oprins, E., Burggraaff, E., and van Weerdenburg, H. (2006). Design of a Competence-Based Assessment System for Air Traffic Control Training. *The International Journal of Aviation Psychology*, 16(3), 297–320.

Schuver-van Blanken, M.J., Huisman, H., and Roerdink, M.I. (2010). The ATC Cognitive Process and Operational Situation Model - A model for analysing cognitive complexity in ATC. In *29th EAAP Conference*. Budapest, Hungary.

Somers, V.L.J., Borst, C., Mulder, M., and Van Paassen, M.M. (2019). Evaluation of a 3D Solution Space-based ATC Workload Metric. In *14th IFAC Symposium on Analysis, Design, and Evaluation of Human-Machine Systems*, 151–156. Tallinn.

van Meeuwen, L.W., Jarodzka, H., Brand-Gruwel, S., Kirschner, P.a., de Bock, J.J., and van Merriënboer, J.J. (2014). Identification of effective visual problem solving strategies in a complex visual domain. *Learning and Instruction*, 32, 10–21. doi: 10.1016/j.learninstruc.2014.01.004.

Watson, R.A. (2014). Use of a Machine Learning Algorithm to Classify Expertise: Analysis of Hand Motion Patterns During a Simulated Surgical Task. *Academic Medicine*, 89(8), 1163–1167. doi: 10.1097/ACM.0000000000000316.