Data Mining Techniques to Derive Human and System Performance Measures of Air Traffic Control from Operational Data

Esa Rantanen
Ernest Fokoue

Follow this and additional works at: https://corescholar.libraries.wright.edu/isap_2011

Part of the Other Psychiatry and Psychology Commons

Repository Citation
https://corescholar.libraries.wright.edu/isap_2011/26
DATA MINING TECHNIQUES TO DERIVE HUMAN AND SYSTEM PERFORMANCE MEASURES OF AIR TRAFFIC CONTROL FROM OPERATIONAL DATA

Esa Rantanen and Ernest Fokoue
Rochester Institute of Technology
Rochester, NY

The capabilities of the envisioned performance management services of the NextGen may offer means for testing of measures of human performance derived from operational data and allow for longitudinal studies on the effects of operational NextGen on human operators. Time-based metrics offer an attractive solution to measurement challenges of the NextGen, given the dynamic nature of the system, its dependence on severe time constraints, and the role of time in critical aspects of human performance, mainly workload and situation awareness. Due to potentially very high number of variables and complexity of the underlying data structure that render standard statistical techniques inadequate, novel techniques that perform nonlinear regression and pattern recognition along with feature selection and variable selection are potential candidates for such analyses. Promising statistical techniques to handle the many problems associated with these kinds of data include Generalized Linear and Generalized Linear Mixed Models, and Support Vector Machines.

Introduction

The current modernization efforts of the U.S. National Airspace System (NAS)—collectively known as NextGen—represent an unprecedented influx of new technologies and procedures to an immensely complex system that is an integral part of the nation’s infrastructure and economy. As with all changes of this magnitude on systems as complex as the NAS, the final impact of modernization of the air traffic control (ATC) and -management (ATM) remains unknown, in particular on the task environments, working methods, strategies, workload, and performance of the human operators within the system.

Although there are many very good sources for human factors data and specifications for acquisition and implementation of new technologies in aviation systems (e.g., Ahlstrom & Longo, 2003; Cardosi & Murphy, 1995), the fact that the technologies and the procedures required for their use are indeed new will curtail the validity of existing standards. Validation efforts through controlled experiments are not without problems, either. It is difficult and often impossible to mimic operational task environments in simulations, demonstrations seldom and systematic experimentation requires large experimental designs, which in turn are expensive and time-consuming to run. Furthermore, ensuring sufficiently large numbers of participants from the target populations (e.g., airline pilots and air traffic controllers) is very difficult. An alternative to experimental validation of new technologies and procedures is to approach the problem through analyses of operational data in longitudinal studies.

Fortunately the new technologies associated with NextGen also allow for routine gathering of data that could be stored for a myriad of analyses. Radar data (the lat. and long. coordinates plus altitudes of aircraft) can be recorded at very high frequencies (e.g., every 10 s), providing a 4-dimensional picture of all air traffic for accurate reconstruction and analysis using software tools such a SATORI (Rodgers & Duke, 1993, and POWER (Manning, Mills, Fox, & Pfleiderer, 2001). Today, radar data are increasingly used to evaluate NAS operations through joint Federal Aviation Administration (FAA) and National Aeronautics and Space Administration (NASA) program called Performance Data Analysis and Reporting System (PDARS). As laudable and crucial to the NextGen overhaul of the U.S. air transportation infrastructure as these efforts are, however, the current measures predominantly focus on the system performance, with little regard on the human component within it.

About the Measurement of the Air Traffic Controller

Over 30 years after V. D. Hopkin’s paper titled “The measurement of the air traffic controller” (Hopkin, 1980) appeared in the journal Human Factors the intricacies of measurement in ATC/M have, if possible, only been exacerbated by the increase of automation throughout the domain. Consequently, the complexity of the task environments in which controllers work has also increased, making scientific investigation regarding the impact of new technologies more and more difficult due to the escalating number of variables and their interactions in the operational environments as well as the shift from overt performance (i.e. manual control) to predominantly covert behavior (i.e. supervisory control) of the operators. It may also be argued that the traditional measurement
techniques in ATC, that is, subjective, over-the-shoulder (OTS) evaluation of controller performance by other experienced controllers, have become inadequate in the face of the present challenges.

Hopkin (1980) concluded his paper by contemplating “prospects for progress” (p. 555) and laid out two different approaches to help in development of better measures of the air traffic controller. One approach listed several new measures to be developed, including time required to perform handovers and for a new controller to accept the position, time scale of decisions (under high workload controllers supposedly move from strategic planning to tactical decisions and opportunistic control), frequency and extent of use of automated aids, measures of boredom, and measures of trustworthiness of information presented to controllers. Another approach advocated by Hopkin was better application of basic psychological concepts to the measurement problems in ATC.

A Taxonomy of ATC/M Measures

Rantanen (2004; Rantanen & Nunes, 2003) took an emphatically systematic and comprehensive approach to the measurement problem in ATC. The approach served a dual purpose, (1) to perform periodic and thorough reviews of past and current research efforts and to organize the findings in a manner that facilitates the use of existing knowledge for a basis of future evolvement of ATC measurement, or, to avoid ‘reinventing the wheel’, and (2) to proceed cautiously on an issue as complex as ATC measurement and consider carefully all the constraints, assumptions, and threats to validity that may emerge. A result of this work was a taxonomy and a database for ATC measures.

There are several possible taxonomies for ATC measures. One is the dichotomy of measurement of system performance and the measurement of an individual controller or a team of controllers (Buckley et al., 1983; Hopkin, 1980). System measures are defined in system terms (i.e., capacity, throughput, delays, and channel occupancy times), and although they are greatly influenced by human performance, they are usually insufficient for the measurement of the performance of an individual controller. Possible measures of individual controller include identified task performance, human activity, errors, omissions, physiological and biochemical indices, and subjective assessment (Hopkin, 1995). Task performance measures compare the controller’s output to that which is required in the task and encompass broad measures of errors and omissions. Human activity measures passively record what occurs in the task, such as radio transmissions, equipment usage, and communication and coordination with other sectors in terms of times, frequencies, and sequences of the activities.

Direct and Indirect Measures

The main division of measures in the ATC measures taxonomy (Rantanen, 2004) was between direct and indirect measures. Direct measures were defined as those that can be explicitly measured. Examples of such measures include a direct observation of a controller’s action, measurement of a response latency, or count of aircraft in a sector at a given time. Indirect measures are those that cannot be measured directly but must be inferred from directly measurable variables. For example, certain actions of a controller may be indicative of his or her performance, response latency can be used to make inferences on some covert cognitive processes, and a number of aircraft in a sector can be used to signify sector complexity.

Identification of direct measures in the literature was a relatively simple task. Their classification was straightforward as well. In the resulting taxonomy, the separately reported direct measures were grouped under 65 distinct classes, down to a sixth level in some cases. Altogether 37 indirect measure classes were identified in the literature. The latter, however, rested on very narrow theoretical foundation for making the associations between direct and indirect measures. This is not to say that the measures reported in the literature are not valid; rather, insufficient information was provided to fully assess their validity. Only a minority of all articles reviewed explicitly justified making inferences about indirect variables based on direct measures by citing past research where such associations have been established. Those articles that did typically cited the same sources, making the body of supporting research literature for ATC measurement remarkably small. Finally, few articles took advantage of the opportunity to add to the foundational body of research in addition to reporting results on their particular topics and thus help validating measures used towards their ends (cf., Vicente & Torenvliet, 2000).

Primary and Secondary Measures

Yet another aspect of classification of measures that warrants discussion is the differentiation between what may be termed primary and secondary measures. Primary measures are those that are measured directly, for example, count of aircraft in a sector, or number of heading changes per aircraft. Secondary measures are those
derived from primary measures, for example an average number of traffic in a sector, its variance, or range. In the case of the average, the criteria are implicit (the time duration or interval during which the aircraft were counted and the number of samples) but nevertheless have an impact on the eventual measure. It is clear that descriptive statistics are crucial for reducing and making sense of the data; yet, it is very important to acknowledge how such techniques might obscure some aspects of the data and skew others. These concerns are seldom effectively dealt with in the literature (Vicente & Torenvliet, 2000). This particular facet of measurement will be of especially great consequence when multiple measures representing various aspects of an indirect variable of interest are combined in some sort of an index, such as workload (Hart & Staveland, 1988) or dynamic density (Laudeman et al., 1998).

**Operational ATC Data**

Data from operational ATC has been available for detailed analysis for decades. The SATORI software was developed in the early 1990s to recreate air traffic situations for incident investigation purposes (Rodgers & Duke, 1993). The FAA also collected System Activity Recordings (SAR) that stored all flight and radar information in Air Route Traffic Control Centers (ARTCCs). These data were processed by two other software applications, the National Track Analysis Program (NTAP; FAA, 1991) and the Data Analysis and Reduction Tool (DART; FAA, 1993), which produced a number of text-based output files for further analysis.

**Performance and Objective Workload Evaluation Research (POWER)**

The NTAP and DART output files from the SAR recordings served as input data for yet another software tool, the Performance and Objective Workload Evaluation Research (POWER; Mills, Manning, & Pfeiderer, 1999; Manning et al., 2000, 2001; Manning, Mills, Fox, Pfeiderer, & Mogilka, 2002). The POWER program derived over 40 separate measures that described a variety of aspects of ATC, including such measures as traffic count, control duration, and variability in aircraft headings, altitudes, and speeds, as well as latencies of handoff initiation and acceptance. A number of different controller activities were also recorded.

The POWER program was never—as far as we know—used beyond a couple validation studies. One such study used 12 traffic samples from four sectors in Kansas City ARTCC (Manning et al., 2002). In another study, Rantanen, Naseri and Neogi (2007) derived several additional measures of controller performance from operational data obtained from Indianapolis (ZID) ARTCC. For this research, the POWER program was augmented by the Medium Term Conflict Detection (MTCD) algorithm (MTCD Library, n. d.), which allowed for derivation of several critical metrics, including counts of aircraft pairs at the same altitude and time to loss of separation and duration of conflict. Another algorithm (Laudeman et al. 1998; Sridhar, Sheth, & Grabbe, 1998) implemented in the POWER program calculated a dynamic density value for every minute. For demonstration purposes, operational data from three different sectors of ZID ARTCC at two different (busy and slow) times were analyzed and different metrics derived from the data. Both the different sectors and the time periods were clearly distinguished by the measures, attesting to their sensitivity and validity.

**Performance Data Analysis and Reporting System (PDARS)**

PDARS is based on several FAA initiatives to measure the performance of the NAS and NASA Distributed Air/Ground (DAG) Traffic Management (TM), or DAG-TM, project (Prevot & al., 2003). The system automatically and continually collects radar track and flight plan data from both terminal and enroute ATC computer systems. From these data, variables such as traffic counts, travel times, travel distances, traffic flows, and in-trail separations can be derived. In addition, several events can also be identified and recorded, for example, takeoffs and landings, sector and facility boundary crossings, tops of climb and descent, fix crossings, reroutes, handoffs, and holding patterns. Several reports are also generated automatically (Den Braven & Schade, 2003; Nehl & Schade, 2007). Given the sophisticated data collection and analysis capabilities of PDARS, the system could plausibly be augmented by a number of additional metrics that focus on human variables.

**Human Performance Measurement**

An example of derivation of objective human performance metrics from ATC data is given in Rantanen (2009). The data were collected from the FAA’s evaluation simulations of Future En route WorkStation (FEWS) under different air traffic load configurations at the FAA Technical Center at Atlantic City International Airport, NJ. The simulations were extremely realistic including a complete set of tasks for the participating controllers to
perform. The simulations were conducted using the FAA Target Generation Facility (TGF), an emulator for the
HOST computer system and the Display System Replacement (DSR), and the Center-Tracon Automation System.
The simulation airspace was the Genera ARTCC with IFR in effect. The data were first processed by the FAA into
event and trajectory files. The former contained a time line of all controller and pilot actions and events during the
run to millisecond accuracy. A separate data processing program was developed to derive temporal task
performance variables, the opening and closing of a window of opportunity to perform a given task, and when the
task was initiated and completed. From these variables it was possible to determine air traffic controller’s task
prioritization schemes (first come, first served) as well as the effect traffic density had on their performance
(Rantanen, 2009).

These examples demonstrate the feasibility of extracting useful metrics from complex, objective data. They
fall short of systematic examination of the effects of different traffic characteristics (e.g., in different sectors at
different times) or the effects of new tools of controllers had on their working methods or performance (e.g.,
FEWS), however, for they merely provide a snapshot of a particular hour or so. Furthermore, the data processing
and algorithms employed were very time-consuming and thus not suitable for longitudinal research. New tools must
therefore be developed to realize the goals advocated in this paper.

Challenges

Theoretical Obstacles

As has been discussed above, the primary challenge in deriving meaningful, reliable, and valid human
measures from operational data lies in the matching of psychological research done in laboratories (controlled
experiments) with the operational task environments where air traffic controllers work. This challenge is not
insurmountable, however, as has been demonstrated before (Rantanen et al., 2007; Rantanen, 2009). The richness of
data available from operational ATC through programs like PDARS plausibly allows for identification of common
variables between the two domains, that is, independent and dependent variables in psychological research literature
that are comparable to the task demands and performance indices identifiable in the operational world.

A prime example of such common variable is time. Time is also common to the human, the task, and the
environment and thus offers a common unit of measurement of human performance in the context of the task. Time
is central to several critical aspects of human performance. Time pressure is a key element of task load and one of
the primary drivers of subsequent mental workload (Hendy, 1995; Hendy, Liao, & Milgram, 1997; Hancock &
Chignell, 1988; Laudeman & Palmer, 1995; Loft, Sanderson, Neal, & Mooij, 2007). Time is also central to what is
arguably the most critical construct of human performance in ATC, situation awareness (cf., Rantanen, 2009). Thus,
accurate time stamps associated with myriad of variables collected from operational ATC provide a promising point
of contact with vast amounts of existing psychological research.

Validation of human performance measures such as discussed here presents further challenges. Undoubtedly
many metrics will measure approximately the same thing, or closely related things. Therefore, such
metrics should be closely correlated when derived from the same data set. Metrics that do not agree with others
supposedly measuring the same thing are suspect, warranting closer inspection of their validity. Similarly,
comparison of the same metrics from data sets with known differences will allow for assessment of the sensitivity of
the metrics (sensitive metrics are obviously preferred). Examination of corroborating evidence will thus be the
primary means to validate the measures.

Statistical Challenges

Contexts such as ATC provide a very potent field of study for the application of data mining and machine
learning techniques, primarily because of the potentially very high number of variables. A justification for the use of
these techniques comes from the fact that it is not common to end up with situations where the complexity of the
underlying structure of the data renders standard off the shelf statistical techniques obsolete and inadequate. Due to
the potentially complex nature of the patterns underlying the observed data and considering the fact that in some
cases the sample size is not large enough to adequately cover the input space of interest, nonstandard statistical
analyses are most likely to be needed to better handle this kind of research. For example, it is very unlikely in most
cases that a traditional multiple regression model would provide an adequate representation of the patterns of
interest, even if one were to include all kinds of interactions between the variables.

Even under the assumption of a linear model, regularization techniques that enforce the shrinkage of some
of the coefficients to zero might be needed because the resulting formulation of the statistical problem yields a
severely under-determined system, and therefore an ill-posed statistical problem. It goes without saying that as $p$
increases, it gets harder and harder, both computationally and statistically, to extract the information underlying the
process at hand let alone fully understand it. The curse of dimensionality has been and still is one of the most
important aspects to deal with in statistical analyses. Many techniques are known to work in low dimension, but fail
miserably when $p$ gets large. Therefore we need to more powerful tools to handle such situations.

The internal structure inherent in the input space quickly becomes a pitfall to any technique that does not
take it into account. For example, a naive regression analysis that does address the correlation structure of the input
space is bound to produce suboptimal (at best) results. Techniques such as principal component regression can be
thought of here, but if one really desires to have a handle on the original variables—as one should for model
identification—it is better to think of nonstandard techniques that concurrently achieve regression and variable
selection.

Given the nontrivial, or at least nonstandard, nature of the structure potentially under consideration, it is
very likely that expert’s knowledge would be of great importance in zeroing in on the patterns underlying the data.
The Bayesian provide tools and methods for incorporating such knowledge into modeling in order to achieve a
better extraction of the true structure.

It would be very surprising if the patterns of workload management and situation awareness, for example,
were the same (homogeneous) across the whole population under consideration. We expect large individual
differences in controllers’ techniques and tactics, differences due to different sector and traffic characteristics, and
differences due to regulatory factors (e.g., SOPs and LOAs between sectors and facilities). It is reasonable to
hypothesize that there might be clusters of patterns rather than a single pattern that captures the structure of the
whole population. This is likely to impact even a regression analysis performed on the data. For instance, if one had
to perform a regression analysis, one might have to consider modeling a mixture of regression models in order to
account for the nonhomogeneity of the population being studied. Nonhomogeneity can also influence the correlation
structure of the data, so that one might have to resort to such modern techniques of latent variable analysis as
mixtures of factor analyzers and mixtures of principal component analyzers to extract meaningful new concepts and
constructs from the inherently high dimensional data generated in operational ATC settings.

Modern data mining and machine learning that perform nonlinear regression and pattern recognition along
with feature selection and variable selection are likely to be candidates for this kind analysis. Generalized Linear
Models, Generalized Linear Mixed Models and modern techniques such Support Vector Machines should be
attempted in this kind of studies as these techniques handle complex structures better than standard techniques.

Summary and Conclusion

In this paper we have argued for integration of results from an exhaustive literature review on the
theoretical connections between constructs of interest (e.g., workload and situation awareness) and variables
available from operational data. In other words, to examine human performance in operational contexts will require
matching independent (or predictor) variables from research literature that have been shown to have valid
relationship with dependent variables of interest in laboratory studies with those available from operational data.
Such an effort could be built on the successes of existing systems of routine and largely automated data collection
and analysis systems such as PDARS. Validation of the measures suggested here will require large amounts of data,
also available from programs such as PDARS. All metrics derived from the data must be constantly evaluated for
their coherence, consistency, and reliability, which will require research and development of new statistical
methods.

References

Atlantic City, NJ: FAA Technical Center.


Cardosi, K. M., & Murphy, E. D. (Eds.) (1995). Human factors in the design and evaluation of air traffic control

Den Braden, W., & Schade, J. (2003). Concept and operation of the performance data analysis and reporting system

Federal Aviation Administration (1991). Multiple Virtual Storage (MVS); subprogram design document; National
Track Analysis Program (NTAP) (NASP-9247-PO2). Washington, DC: Author.


