Selection and evaluation of air traffic complexity metrics

David Gianazza, Kevin Guittet

SELECTION AND EVALUATION OF AIR TRAFFIC COMPLEXITY METRICS

David Gianazza,Laboratoire d'Optimisation Globale, SDER, Toulouse, France
Kévin Guittet, Laboratoire d'Economie et d'Econométrie de l'Aérien, SDER, Toulouse, France

Abstract

This paper presents an original method to evaluate air traffic complexity metrics. In previous works, we applied a Principal Component Analysis (PCA) to find the correlations among a set of 27 complexity indicators found in the literature. Neural networks were then used to find a relationship between the components and the actual airspace sector configurations. Assuming that the decisions to group or split sectors are somewhat related to the controllers workload, this method allowed us to identify which components were significantly related to the actual workload. We now focus on the subset of complexity indicators issued from these components, and use neural networks to find a simple relationship between these indicators and the sector status.

Introduction

How complex is it to control a given air traffic situation? Over the last years, this question has become more and more crucial, as air traffic increased, thus creating bottlenecks in the air traffic control system. There arised a need for performance and complexity metrics, allowing to assess how the current system is operated and how it could be improved.

As a consequence, *air traffic complexity* has become the subject of many studies – often in relation to the controller's workload – and a multitude of complexity metrics have been proposed. The inadequacy of the *aircraft count* to appropriately reflect the traffic complexity has now been acknowledged for a long-time, and complementary indicators such as "traffic mix", "number of potential conflicts" and others, have been (and still are) designed. A linear combination of these variables, often referred to as *dynamic density*, is likely to better fit traffic complexity than individual indicators. It is used throughout most studies, where the correllation of a set of indicators with a quantifiable variable (subjective ratings, number of interactions with the computer,...) assumed to represent the actual traffic complexity, is maximized.

A possible shortcoming of this methodology is that potentially non-linear relations between indicators are missed (see [1] and the concern of Eurocontrol when writing calls for proposals). But, more importantly, the choice of the dependent variable is crucial to determine how well complexity is actually measured. Indeed, physical activity, as used in [2] and [3], miss the important cognitive part of the controller activity. On the other hand, physiological indicators ([4], [5]) seem difficult to exploit and how well they relate to traffic complexity is unclear. Finally, widely used subjective ratings ([6], [7]) provide high quality data (as they obviously relate to the kind of complexity investigated), but are often seen as subject to biases (such as the recency effect denounced in [5], and the possibility of raters errors in the case of "over-the shoulder workload ratings" [8]). In all of these cases, data are very expensive to collect, as they require the active participation of controllers. Databases are often small and might exhibit low variability, which may in turn harm the statistical relevance of the results. This phenomenom is acknowledged in [7], where the overfitting of data is clearly a consequence of a lack of observations rather than a misspecification of the neural network. Finally, as these complexity metrics may be used to design computer-assisted control tools or traffic management tools, and to help organizing airspace, it is surprising to notice that the question of the relevance of the complexity measured to the final goal is scarcely discussed. The question really is to understand which complexity is measured and how well it relates to the foreseen application (benchmarking, improvements in airspace organization, design of new tools...).

Our research is initially motivated by former studies on optimal airspace sector configurations ([9], [10]) and intends to improve the criterion used therein to evaluate sector configurations. The basic idea, introduced in [11], is that the decisions to split a sector, mostly taken when the controller is close to overload, are linked to traffic complexity and may therefore provide an acceptable dependent variable. Interestingly enough, collecting data on sector configurations does not require controllers active participation, as current outcomes from control centers can be used, while related flight informations are available from recorded radar tracks. As such, raw data needed in our study are noticeably cheap to collect and might be produced in large quantities. The price to pay is that these data are noisy, as we may not be sure that a sector splitting (resp. merging) decision is directly related to overload (resp. underload). Other factors might distort data, such as training of unexperienced controllers, meteorological hazards, military airspace use... However, we will assume that the impact of these phenomena on the accuracy of the results is limited, particularly because of the kind of complexity we are looking at here. Indeed, this work is conducted in the perspective of future pre-tactical applications (e.g. sector planning) and thus does not ask for as much details as studies of instantaneous workload would (on the opposite, benchmarking of ATC centers would require an even coarser granularity, as indicators are averaged on wide temporal and geographical horizons [12], [13]).

This paper is a continuation of [14], which started with a principal component analysis on a set of indicators found in the literature, and where we proposed an original method allowing to select which components were actually related to the controller's workload, considering the status of the control sector. Neural networks were used to investigate the link between complexity indicators and sector configurations, as non-linear interactions were suspected.

For practical purposes, we would rather avoid to compute all the indicators and then apply a transformation matrix in order to find the relevant components. So, our aim now is to predict the sector status directly from a small subset of relevant complexity indicators, issued from the previous selection of principal components.

The paper is organized as follows. Section I briefly describes the indicators used throughout the study, while section II presents the raw data from which the final database is built. Section III summarizes the results of the Principal Component Analysis (PCA) which was performed in [14]. Neural networks are introduced in section IV. The results on the complexity metrics selection and evaluation are presented and discussed in sectionV. Section VI concludes.

I. Air traffic complexity indicators

Indicator	Definition	Used in
Nb	Number of aircraft	[17] [7] [18] [6]
Nb^2	Squared number of aircraft	[18] [6]
σ^2_{gs}	Variance of ground speed	[7] [6]
N_{ds}	Number of descending aircraft	[2] [17] [7] [6]
N_{cl}	Number of climbing aircraft	[6] [2] [17] [7]
$\frac{\sigma_{gs}}{\overline{gs}}$	Ratio of standard deviation of speed to average speed	[7] [6]
F_5	Incoming flow (hozizon 5mn)	[11]
F_{15}	Incoming flow (hozizon 15mn)	[11]
F_{30}	Incoming flow (hozizon 30mn)	[11]
F_{60}	Incoming flow (hozizon 60mn)	[11]
vprox_1	Vertical proximity C_6 of [7]	[7] [18] [6]
vprox_2	Vertical proximity C_8 of [7]	[7] [18] [6]
hprox_1	Horizontal proximity C_9 of [7]	[7] [18] [6]
Dens		[19]
track_disorder		[19]
speed_disorder		[19]
Div		[19]
Conv	*See appendix A*	[19]
sensi_d		[19]
insen_d		[19]
sensi_c		[19]
insen_c		[19]
inter_vert		[13]
avg_vs	*See appendix A*	[13]
inter_hori		[13]
creed_ok	*See appendix A*	[20], [21]
creed_pb		[20], [21]

TABLE I

CHOSEN SUBSET OF AIR TRAFFIC COMPLEXITY INDICATORS

The accuracy of the results of a study related to air traffic complexity is strongly dependent of the diversity and quality of the chosen individual complexity indicators. Many have been suggested to help describe the controllers workload, and it is hardly possible to implement the entire pool. In order to limit the number of variables to be (re)programmed and present indicators that are representative of the *dynamic density* literature, we focused on the ones selected by Kopardekar [6] in its unified complexity metric[1]. These indicators, and references to studies where they were used and where definitions may be found, are presented in Table I. We also implemented several indicators inspired by studies conducted elsewhere in the SDER (former CENA). Definitions are indicated in appendix A[2]. Finally, we also used incoming flows as explanatory variables, as they may be a significant factor in the decision to split (or merge) a sector.

II. Input data

The indicators are computed every round minute of the day, using recorded radar data, environnement data (sector description), and recorded sector configurations of the five french ATC centers. The sector configurations are recorded every round minute of the day, which explains our choice concerning the frequency at which we compute the indicators.

Radar data is available in several forms: records made by each center, with one position every twelve seconds, in average, and a global record of the five centers, with one position every three minutes. Several months of global records were available, whereas the centers local records were not readily available, at least for a sufficiently long period of time. So we used the global records (made by the IMAGE system), and interpolated the aircraft positions in order to get one position per minute. As many trajectory changes may occur within three minutes of flight, the computed positions are not highly accurate, and this may introduce a bias in the indicators values. However, this bias is most probably of small importance in our problem: we just want to predict when a sector will be merged into another one, or split in several smaller sectors. We are not considering the instant workload, which may require a very high level of accuracy on the aircraft position, speed, and so on. To be sure that this bias is small, we should compare the computed positions, and maybe also the indicators values, using local centers records, and global records, on small data samples. This is left for future work.

Several months of recorded traffic are available. However, considering the volume of data, it would be tedious to run several experimentations on very large data samples. So, we have restricted our choice, at least for the moment, to one day of traffic (1st june, 2003). Once we have found the most significative complexity indicators, it will be possible to re-train the neural network on larger data samples.

On the chosen day, 103 different sectors were armed. The term "sector" means here either an elementary sector, or a set of elementary sectors merged together, and handled on a single controller's working position. The air traffic complexity indicators were computed for each of these sectors, every minute of the day, together with the sector status (*merged*, *armed*, or *split*). This data was split into two sets : about sixty percent was randomly selected in order to *train* the neural network, and the rest was used to *test* the trained network on fresh data.

This single day of traffic already provides a big volume of data, as detailed in table II, with a great diversity of geographic sectors, and with enough data in each class of sector status.

	Total	Merged	Armed	Split
Train	71270	46.6%	27.0%	26.4%
Test	47513	46.4%	27.0%	26.6%

TABLE II

NUMBER OF MEASURES AVAILABLE, ON THE 1ST JUNE OF 2003.

III. Principal component analysis

Before applying the neural network to complexity indicators and sector statuses, let us first give a brief summary of the *principal component analysis* results. The reader may refer to [14] for the details.

The aim of the *principal component analysis* is to find the colinearities between the indicators, and thus reduce the dimensionality of our problem. From the initial 27 complexity indicators, 6 main components were identified (corresponding to eigenvalues greater than 1), that covered more than 76 % of the variance of the data set.

[1] Though we were not always able to find an explicit formula, and thus missed seamingly important indicators like, e.g., "MET_airspace structure". Note that this difficulty to get clear definitions is also reported by Eurocontrol in [15].

[2] Further informations and discussions about indicators are to be found in the internal note [16].

These components are interpreted as follows: C_1 may be seen as a "size factor" and is strongly representative of the *aircraft count*, C_2 is related to the *ground speed variance*, and the *aircraft vertical evolutions*, C_3 is highly correlated with *incoming flows*, C_4 seems mostly related to *converging flows* and *anticipation of conflicts*, C_5 is linked with *divergent flows*, and C_6 is strongly correlated with the vertical proximity measures ([7]), and could stand for the *monitoring of vertical separation* (near the minimas).

Notice that we extracted only 6 components, thus significatively less than the 12 components (briefly) described in [6]. This might be explained by the lack, at the point of the project, of indicators related to the sector geometry.

Component	Variables
C_2	*avg_vs*, $\frac{\sigma_{gs}}{\overline{gs}}$, σ^2_{gs}, N_{cl}, N_{ds}
C_3	F_5, F_{15}, F_{30}, F_{60}
C_4	*inter_hori, insen_c, Conv, creed_ok, creed_pb*
C_5	*insen_d, Div*
C_6	*vprox_1, vprox_2*

TABLE III

VARIABLES BELONGING TO THE PRINCIPAL COMPONENTS

Table III lists the indicators providing the best explanation for each component between C_2 and C_6. For C_1, we will simply use the number of aircraft Nb, as it is the most representative of the overall size of the problem. For C_2 and C_3, the selected variables have an absolute value of correlation above 0.4 (that is the correlation with the component axis). For these two components, the chosen variables are highly representative of the corresponding components. The choice is less easy for component C_4, where several indicators with the highest correlation values are, in fact, not very meaningful, as they are already involved in previous components. This means that we probably should refine again the definition of some of the indicators, in order to have a clearer interpretation of this component. For the time being, we have selected the variables with the highest absolute values of correlation, and which did not appear in previous components. For C_5 and C_6, the selected variables are well correlated with the component (0.4 to 0.7).

IV. Neural networks

A. Presentation

Artificial neural networks are algorithms inspired from the biological neurons and synaptic links. An artificial neural network is a graph, with vertices (neurons, or units) and edges (connections) between vertices. There are many types of such networks, associated to a wide range of applications: pattern recognition (see [22] and [23]), control theory,...

Beyond the similarities with the biological model, an artificial neural network may be viewed as a statistical processor, making probabilistic assumptions about data ([24]). Some *train* data is used to determine a statistical model of the process which produced this data. Once correctly trained, the neural network uses this model to make predictions on new data.

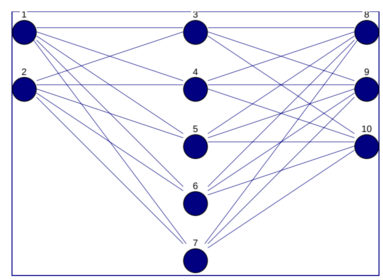

Figure 1. Example of a feed-forward network with one hidden layer

Neural networks are closely related to the Bayesian probabilities. They may be used for unsupervised learning (*density estimation* problems), and, mainly, supervised learning problems (*regression*, *classification*). Density estimation is not in the scope of our paper, so we will not detail it. The aim of *regression* is to find a statistical model producing an *output* y from *input variables* (let us denote them by x), so that the output y is as close as possible to a *target variable*, which we shall denote by t. In the case of *classification* problems, the target variables represent class labels, and the aim is to assign each input vector x to a class.

We will use a specific class of neural networks, referred to as feed-forward networks, or multilayer perceptrons (when the activation function is logistic). In such networks, the units (neurons) are arranged in fully-connected layers: an *input layer*, one or several *hidden layers*, and an *output layer*. Figure 1 shows an example of such a network.

For our problem, we have chosen three-layers feed-forward networks, denoted $I_\alpha H_\beta O_\gamma$ in the rest of the paper, with α units in the input layer, β units in the hidden layer, and γ units in the output layer. The input variables are normalized, by substracting the mean value and dividing by the standard deviation.

The output vector $y = (y_1, ..., y_k, ..., y_q)^T$ is expressed as a function of the input vector $x = (x_1, ..., x_i, ..., x_p)^T$ as follows:

$$y_k = \Psi(\sum_{j=1}^{q} w_{jk}\Phi(\sum_{i=1}^{p} w_{ij}x_i + w_{0j}) + w_{0k}) \quad (1)$$

where the w_{ij} and w_{jk} are weights assigned to the connections between the input layer and the hidden layer, and between the hidden layer and the output layer, respectively, and where w_0j and w_0k are biases (or threshold values in the activation of a unit). Φ is the sigmoid logistic function :

$$\Phi(z) = \frac{1}{1 + e^{-z}} \quad (2)$$

For the transfer function Ψ applied to the output layer, we have chosen the *softmax* function:

$$\Psi(z_k) = \frac{e^{z_k}}{\sum_{m=1}^{C} e^{z_m}} \quad (3)$$

With this choice for Ψ, the error function minimized during training is the following log-likelihood, known as *cross entropy*:

$$E(w) = -\sum_{n=1}^{N}\sum_{k=1}^{C} t_k^{(n)} ln(y_k^{(n)}) \quad (4)$$

where $t^{(n)}$ and $y^{(n)}$ are the n^{th} target and output vectors, respectively.

The error is a function of the network's parameters w (weights and biases). The *training* aims at choosing these parameters, so as to minimize the chosen output error function. As we adress a classification problem – assign each input vector (a list of complexity indicators values) to a class representing the sector status (*merged*, *armed*, or *split*) – it is best to minimize the cross-entropy instead of the quadratic error often used in regression.

The neural network is designed with one output unit per class (*merged*, *armed*, or *split*). A target vector $t^{(n)}$ with value $(1, 0, 0)$ means that the considered sector was merged with other sectors when the n^{th} measure of the vector of complexity indicators was made. Armed sectors will be represented by $(0, 1, 0)$. A value of $(0, 0, 1)$ means that the sector was previously split in two or more sectors at the time x was measured. Of course, the actual output $y^{(n)}$ of the neural network will not be made of exact values 0 or 1. It will be a triple (a, b, c) of floating-point values between 0 and 1, each value being the probability to belong to the corresponding class. The input vector $x^{(n)}$ will be assigned to the class of highest probability.

The number of input units of the network is the cardinal of the evaluated set of metrics. For example, if we consider $\{V, Nb\}$ we will use a network with two input units. The hidden layer comprises 15 units.

The `nnet` package of the `R` language was used (see `http://www.r-project.org/` for details on the `R` language and environnment). In this package, developped by Pr B. D. Ripley, a quasi-newton minimization method (BFGS) is used for the network's learning. We have taken the same `nnet` parameters as in [14].

B. *Evaluation of the neural network's outputs*

A well-known problem, when using neural networks (or other regression methods), is *overfitting*: with enough parameters and enough training cycles, it is always possible to find a good fit for a given data set. So one may find a perfect fit for a chosen data sample, and then feel disappointed when the trained network makes wrong predictions on fresh data. So, we will systematically proceed as follows: train the network on a randomly chosen data sample (called *train*), then check the results, first on the same data sample, and second on a fresh data sample (called *test*) that was not used for the training.

In order to evaluate the outputs of several different models, we have to compare the neural networks predictions to the actual target values. We may use the fit criterion (*cross entropy*) but it does not reflect the influence of the number of weights (and biases) in the neural network. It is known (see [24]) that a network with too few weights may not be able to capture all the variations of the response to the input x, whereas a network with too many weights will more likely be subject to *overfitting*. In the next sections, we will compare several sets of input variables, of various sizes. Consequently, *the number of weights in the network will not remain constant*, and this variation will bias the results.

In [14], we used the *Akaike information criterion* ([25]) to select the best model: $AIC = 2\lambda - 2ln(L)$, where λ is the number of unadjusted parameters of the model (i.e. the number of weights and biases of the network), and $ln(L)$ is the log-likelihood. However, fully connected neural networks are very greedy, as concerns the number of parameters: the number of weights and biases highly increases when adding new input variables. So we will rather use the Schwartz Bayesian information criterion ($BIC = 2\lambda.ln(N) - 2ln(L)$), which assigns a higher penalty to the number of parameters in the model.

In our case, the BIC is written as follows, where N is the size of the data sample:

$$BIC = 2\lambda.ln(N) - 2\sum_{n=1}^{N}\sum_{k=1}^{C} t_k^{(n)}.ln(y_k^{(n)}) \quad (5)$$

In our previous work, we were able to compare predictions made on the *train* test and on the *test* set, which are of different size, by dividing the AIC by N, the number of data items: $AIC_{avg} = \frac{AIC}{N}$. In this paper, we will use $BIC_{avg} = \frac{BIC}{N}$. However, we must be aware that the correcting factor $ln(N)$ is different for both data sets, so it is meaningless to compare *train* and *test* data with this criterion. This is not of much importance, as we have already observed that the results on both data sets were highly consistent. In fact, we will make our model selection with the BIC criterion computed with the network's outputs predicted on *test* data only.

In addition to the numerical results provided by the *Bayesian information criterion*, we shall also consider the global proportion of correctly classified input vectors, and also the percentage of correct classifications for each class. One must be aware, however, that the rate of correct classifications *is not* the criterion being maximized by the neural network, so we should remain cautious when comparing the different classification rates. However, these percentages are easily understandable and may allow us to make some interesting statements on the results.

V. Results

A. Components selection

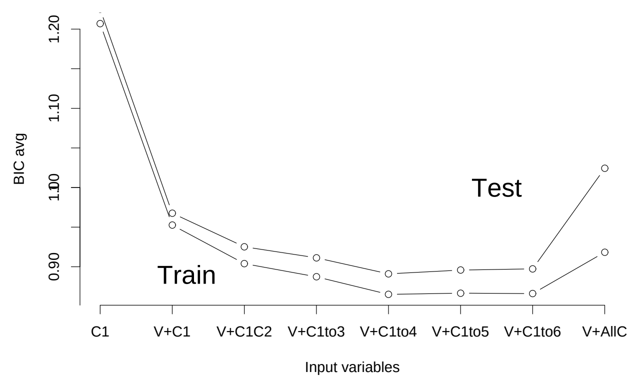

Figure 2. Values of the BIC_{avg} criterion for the different sets of components

Figure 2 shows, for *train* and *test* data, the evolution of the averaged Schartz Bayesian Information Criterion for several combinations of components. The mean values, averaged on five runs, are presented. This BIC curve confirms the similar results found in [14] with the AIC.

The best trade-off between the model complexity (number of parameters in the neural network) and the benefit provided by the additional components is obtained with the $\{V; C_1; ...; C_4\}$ model, that is: the sector volume, the "overal size" factor (aircraft count), the ground speed variance and vertical evolutions, the incoming flows, and the flow convergence and conflicts anticipation. Component C_5 (*flow divergence*) is either irrelevant, or redundant with C_4 (*flow convergence* and *conflicts anticipation*). Component C_6 (*monitoring of aircraft vertical separation*) is most probably irrelevant: in normal traffic situations, we may expect that all aircraft in close vertical proximity have previously been laterally separated by the controller. The same remark applies for horizontal proximity. This was already stated in [7]. So the aircraft in close proximity are not a significant factor in the decision to split or merge a sector. The *anticipation* of future conflicts may be more relevant.

This selection of the best model is very useful for a qualitative evaluation of the metrics. However, for practical purposes, it is not very convenient to compute the 27 complexity indicators, and then apply a transformation matrix to get the components. It would be useful to predict the sector status directly from a smaller subset of relevant complexity indicators.

B. Metrics selection

Let us now select the relevant metrics from the components of the best model $\{V; C_1; ...; C_4\}$. The initial variables subsets – the variables most correlated to each component axis – are detailed in table III of the PCA. We must be aware that the relative weight of a variable within a component (the linear correlation with the component axis) is not necessarily related to its relative influence on the quality of the neural network's prediction, as the relationship between the sector status and the explanatory variables may be non-linear. So we will again select the variables according to the Bayesian information criterion.

An iterative process is used to select the relevant variables. Each component is considered in turn. We have chosen Nb, the number of aircraft, to represent component $C1$, the overal size of the problem.

Starting with a basic set of variables $\{V, Nb\}$, we build the best sequence of C_2 variables, by choosing at each step the variable which minimizes the BIC criterion. To do so, we consider each C_2 variable in turn (together with V and Nb), select the best one, add it to the sequence, and then consider the remaining variables, and so on until all variables have been tested.

The BIC curve is then drawn for this sequence, and the subset of variables corresponding to the minimum of this curve is added to the set of relevant variables. The process is then repeated with the next component, starting with the new set of relevant variables, and so on until all components have been considered.

The neural networks are trained on the *train* data, and the BIC is computed on *test* data. Five runs are made for each subset of variables. We will take the minimum value among these five runs, and use it in the selection process.

In the previous experiments ([14]) on components selection, the mean values and the minimum values lead to the same results. In the metrics selection, we have stated small differences between the two, probably due to the fact that some metrics belonging to a same component give fairly close results, like *avg_vs*, $\frac{\sigma_{gs}}{gs}$, and σ^2_{gs} in component C_2, for example. As the neural network with the best fit is closer to the "true" model than when considering the averaged value, we will now use the minimum value.

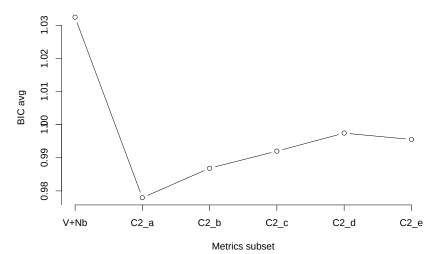

Figure 3. Evolution of the BIC_{avg} criterion for component C_2, for the sequence *avg_vs*, $\frac{\sigma_{gs}}{gs}, \sigma^2_{gs}, N_{cl}, N_{des}$

Figure 3 shows the evolution of the BIC_{avg} criterion when adding successively the variables of component C_2 to the basic set $\{V, Nb\}$, in the order of the best sequence found by the BIC minimization process: *avg_vs*, $\frac{\sigma_{gs}}{gs}, \sigma^2_{gs}, N_{cl}, N_{des}$. The x axis labels on this figure are detailed in table IV.

Subset	Indicators
C2_a	V, Nb, *avg_vs*
C2_b	V, Nb, *avg_vs*, $\frac{\sigma_{gs}}{gs}$
C2_c	V, Nb, *avg_vs*, $\frac{\sigma_{gs}}{gs}$, σ^2_{gs}
C2_d	V, Nb, *avg_vs*, $\frac{\sigma_{gs}}{gs}$, σ^2_{gs} , N_{cl}
C2_e	V, Nb, *avg_vs*, $\frac{\sigma_{gs}}{gs}$, σ^2_{gs} , N_{cl}, N_{ds}

TABLE IV

SUBSETS OF THE SEQUENCE SHOWN ON FIGURE 3 (COMP. C_2)

The curve's minimum is reached with the combination $\{V, Nb,$ *avg_vs*$\}$. This does not mean that the other C_2 variables are irrelevant: the results with $\{V, Nb, \sigma^2_{gs}\}$, and $\{V, Nb, \frac{\sigma_{gs}}{gs}\}$ were fairly close to the best result. So the ground speed variance and the ratio of standard deviation of speed to average speed are simply redundant with the average vertical speed *avg_vs*. This is also true for the numbers of climbing and descending aircraft, although they proved slightly less performant.

These results, showing the relevance of ground speed variance and altitude changes, are in contradiction with previous results presented in [26]. In this study, Masalonis, Callaham, and Wanke applied a logistic regression to subjective ratings and *dynamic density* metrics. The contribution of each metric was then assessed, for each ARTCC, by dropping the metric from the model. In particular, they found that the speed and altitude changes could be dropped without significantly affecting the model's predictions.

Our results contradict these statements, as we find that the component representing the speed variance and the altitude changes actually improves the BIC criterion. We have no ready explanation for this. It may be that, in [26], the sector volume alone captures the fact that some sectors are dedicated to a specific type of traffic: pre-approach sectors with evolutive aircraft are generally smaller than pure en-route sectors. As we use sectors or groups of sectors of various sizes and statuses (merged, armed, or split), there is no correlation between the sector volume and the nature of the traffic within the sector. So the speed variance and the altitude changes become relevant in that context. But, as we don't know the particulars of [26], it is difficult to be sure.

Let us now consider the variables of component C_3, applying the same procedure, and starting from the new set of relevant variables $\{V, Nb, \textit{avg\_vs}\}$.

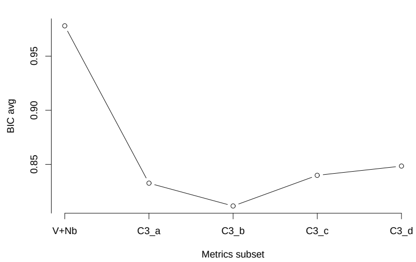

Figure 4. Evolution of the BIC_{avg} criterion for component C_3

Figure 4 shows the evolution of the BIC_{avg} for the best sequence found for component C_3 (see table V). The minimum is reached for the subset $\{V, Nb, \textit{avg\_vs}, F_{60}, F_{15}\}$.

Subset	Indicators
C3_a	V, Nb, *avg_vs*, F_{60}
C3_b	V, Nb, *avg_vs*, F_{60}, F_{15}
C3_c	V, Nb, *avg_vs*, F_{60}, F_{15}, F_5
C3_d	V, Nb, *avg_vs*, F_{60}, F_{15}, F_5, F_{30}

TABLE V

SUBSETS OF THE SEQUENCE SHOWN ON FIGURE 4 (COMP. C_3)

The incoming flow with a time horizon of 15 minutes belongs to the subset of relevant variables. This is consistent with the fact that strips are sent to the controller's working position about 15 minutes before the aircraft actually enters the sector.

We were rather surprised to see that the incoming flow of traffic entering the sector in the next hour also belongs to the set of relevant variables. This incoming flow is probably less linked to the variations of the number of aircraft, and it may somehow smooth these variations.

The incoming flow within the next 5 minutes does not provide any improvement. The first step of the BIC minimization process showed that the BIC_{avg} criterion increased when adding $F5$ to the subset $\{V, Nb, \textit{avg\_vs}\}$, whereas all other flows improved the criterion. So it is either highly redundant with the previous variables, or irrelevant.

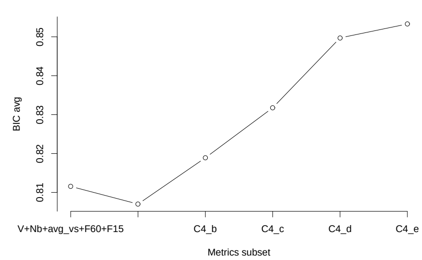

Figure 5. Evolution of the BIC_{avg} criterion for component C_4

Figure 5 and table VI show the results for component C_4. The minimum for this component is reached when adding *inter_hori*. The additionnal variables of the rest of the best sequence increase the BIC criterion. However, the other variables also performed quite well, when put in first position in the sequence. This means that they are not irrelevant, but redundant with *inter_hori*.

Subset	Indicators
$C4_a$	V, Nb, *avg_vs*, F_{60}, F_{15}, *inter_hori*
$C4_b$	V, Nb, *avg_vs*, F_{60}, F_{15}, *inter_hori*, *creed_pb*
$C4_c$	V, Nb, *avg_vs*, F_{60}, F_{15}, *inter_hori*, *creed_pb*, *creed_ok*
$C4_d$	V, Nb, *avg_vs*, F_{60}, F_{15}, *inter_hori*, *creed_pb*, *creed_ok*, *Conv*
$C4_e$	V, Nb, *avg_vs*, F_{60}, F_{15}, *inter_hori*, *creed_pb*, *creed_ok*, *Conv*, *insen_c*

TABLE VI

SUBSETS OF THE SEQUENCE SHOWN ON FIGURE 5 (COMP. C_4)

An interesting result is that the metrics related to the *anticipation of conflicts*, for converging aircraft in the horizontal plane, do not bring any supplementary benefit, when added to *inter_hori*, the number of potential crossings with angle greater than 20 degrees. Notice that all these metrics – *anticipation of conflicts* and *flow convergence* – were already put together in a same component by the PCA. This leads to think that the main feature captured by this set of metrics is the fact that the pairs of aircraft within the sector are in a converging configuration, or not.

We have a feeling that other *conflict detection* indicators (not only for converging flights, and not only in the horizontal plane) may be useful, and may be more orthogonal to the indicators we have implemented so far. This may be the subject of future work.

To conclude on the metrics selection, we end up with a subset of only 6 most relevant indicators $\{V, Nb, \textit{avg\_vs}, F_{60}, F_{15}, \textit{inter\_hori}\}$ among the 27 metrics, plus the sector volume, that were considered in this study. This allows to establish a direct relationship (equation IV-A, with the weights and biases of the trained network) between these six indicators and the probability to belong to a sector status class.

C. Classification rates

Table VII shows the correct classification rates obtained, on *train* and *test* data, with the subset of most relevant indicators. The results with *test* data are rather consistent with the ones obtained with *train* data, as was already the case when studying the components ([14]).

Set	Global	Merged	Armed	Split
Train	83.47%	89.07%	65.58%	91.91%
Test	83.24%	89.14%	64.56%	91.75%

TABLE VII

CORRECT CLASSIFICATION RATES FOR THE $\{V, Nb, \textit{avg\_vs}, F_{60}, F_{15}, \textit{inter\_hori}\}$ MODEL, WITH A $I_6H_{15}O_3$ NETWORK

VI. Conclusion

In conclusion, the proposed method, using neural networks and sector status records, allows selecting the most relevant air traffic complexity metrics, among several indicators proposed in the literature.

In a previous work ([14]), a principal component analysis reduced the scope of the study to 6 main components , starting from 27 metrics to which we added the sector volume. It was then found, by applying neural networks to the principal components and to the sector statuses, that only the first four were significantly related to the decisions to split or merge sectors.

In the current paper, we focused on the subset of metrics belonging to the four relevant components. Still using neural networks and sector statuses, we proposed a selection process, minimizing the Bayesian information criterion, which allowed to select a small subset of 6 relevant metrics: the sector volume V, the number of aircraft within the sector Nb, the average vertical speed *avg_vs*, the incoming flows with time horizons of 15 minutes and 60 minutes (F_{15}, F_{60}), and the number of potential crossings with an angle greater than 20 degrees (*inter_hori*). Let us note that the other metrics may be either irrelevant, or simply redundant with the most relevant indicators.

We established a fairly simple equation (equation IV-A with the weights of the trained network), allowing to predict the sector status (*merged*, *armed*, or *split*) from these relevant metrics. The rates of correct predictions were above 83% (global rate), and highly consistent when considering *train* and *test* data.

The neural network approach used in this study seems appropriate for the granularity we are interested in and the foreseen applications, either strategical (sector design) or pre-tactical (sector planning). We are also fairly confident that decisions to split or merge sectors may allow to assess the instantaneous workload as well, and could therefore be used to improve tactical tools (PRESAGE). To this end, other statistical methods should be investigated to take into account the serial correlation of sector status, looking closely at the sector splitting times. We plan to tackle this issue in a close future, using dynamic discrete choice models.

Finally, another issue that we intend to address, in relation to the complexity indicators, is the prediction of optimal sector configurations. Previous works ([9], [10]) proposed several algorithms to compute optimal sector configurations, using sector capacities and *incoming flows*. The output of the neural network is a triple of probabilities, allowing to decide when a sector should be split, or merged. We may derive a realistic workload indicator – and also threshold values – from these probabilities, which could be used to compute optimal sector configurations.

References

[1] B. Hilburn and G. Flynn. Toward a non-linear approach to modeling air traffic complexity. In *2nd Human Performance Situation Awareness and Automation Conference*, 2004.

[2] I. V. Laudeman, S. G. Shelden, R. Branstrom, and C. L. Brasil. Dynamic density: An air traffic management metric. Technical report, 1999.

[3] A. Majumdar, W. Y. Ochieng, G. McAuley, J.M. Lenzi, and C. Lepadetu. The factors affecting airspace capacity in europe: A framework methodology based on cross sectional time-series analysis using simulated controller workload data. In *Proceedings of the 6th USA/Europe Air Traffic Management R & D Seminar*, 2005.

[4] J.H. Crump. Review of stress in air traffic control: Its measurement and effects. *Aviation, Space and Environmental Medecice*, 1979.

[5] P. Averty, S. Athènes, C. Collet, and A. Dittmar. Evaluating a new index of mental workload in real atc situation using psychological measures. Note cena nr02-763, CENA, 2002.

[6] P. Kopardekar and S. Magyarits. Measurement and prediction of dynamic density. In *Proceedings of the 5th USA/Europe Air Traffic Management R & D Seminar*, 2003.

[7] G.B. Chatterji and B. Sridhar. Measures for air traffic controller workload prediction. In *Proceedings of the First AIAA Aircraft Technology, Integration, and Operations Forum*, 2001.

[8] C. Mannings, S. Mill, C. Fox, E. Pfleiderer, and H. Mogilka. The relationship between air traffic control events and measures of controller taskload and workload. In *Proceedings of the 4th Air Traffic Management Research & Developpment Seminar*, 2001.

[9] D. Gianazza and J. M. Alliot. Optimization of air traffic control sector configurations using tree search methods and genetic algorithms. In *Proceedings of the 21st Digital Avionics Systems Conference*, 2002.

[10] D. Gianazza, J. M. Alliot, and G. Granger. Optimal combinations of air traffic control sectors using classical and stochastic methods. In *Proceedings of the 2002 International Conference on Artificial Intelligence*, 2002.

[11] D. Gianazza. *Optimisation des flux de trafic aérien*. PhD thesis, Institut National Polytechnique de Toulouse, 2004.

[12] G. M. Flynn, C. Leleu, and L. Zerrouki. Traffic complexity indicators and sector typology analysis of u.s. and european centres. Technical report, Eurocontrol, 2003.

[13] Note de synthèse sur l'indicateur de complexité pru. Technical report, DTI/SDER (ex CENA), 2005.

[14] D. Gianazza and K. Guittet. Evaluation of air traffic complexity metrics using neural networks ans sector status. In *Proceedings of the 2nd International Conference on Research in Air Transportation*. ICRAT, 2006.

[15] Cognitive complexity in air traffic control, a litterature review. Technical report, Eurocontrol experimental centre, 2004.

[16] K. Guittet and D. Gianazza. Analyse descriptive des indicateurs de complexité du trafic aérien à partir des données image et courage. Note nr05-905, DSNA/DTI/SDER, Décembre 2005.

[17] B. Sridhar, K. S. Sheth, and S. Grabbe. Airspace complexity and its application in air traffic management. In *Proceedings of the 2nd USA/Europe Air trafic Management R&D Seminar*.

[18] P. Kopardekar. Dynamic density: A review of proposed variables. Faa wjhtc internal document. overall conclusions and recommendations, Federal Aviation Administration, 2000.

[19] D. Delahaye and S. Puechmorel. Air traffic complexity: towards intrinsic metrics. In *Proceedings of the third USA/Europe Air Traffic Management R & D Seminar*, 2000.

[20] P. Averty. Conflit perception by atcs admits doubt but not inconsistency. In *Proceedings of the 6th Air Traffic Management Research & Developpment Seminar*, 2005.

[21] P. Averty, K. Guittet, and P. Lezaud. *Work in progress, presented at an internal SDER seminar*. Technical report, DTI/SDER (former CENA), 2005.

[22] C. M. Bishop. *Neural networks for pattern recognition*. Oxford University Press, 1996. ISBN: 0-198-53864-2.

[23] B. D. Ripley. *Pattern recognition and neural networks*. Cambridge University Press, 1996. ISBN: 0-521-46086-7.

[24] M. I. Jordan and C. Bishop. *Neural Networks*. CRC Press, 1997.

[25] W. H. Greene. *Econometrics analysis*. Prentice Hall International. ISBN: 0-13-015679-5.

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

[27] F. Chatton. Etudes de nouvelles métriques de complexité

de la circulation aérienne. Master's thesis, Ecole Nationale de l'Aviation Civile (ENAC), 2001.

Email

{lastname}@recherche.enac.fr

APPENDIX A: Complexity metrics

Delahaye and Puechmorel metrics

To present the geometrical indicators introduced in [19], we need to define several quantities:

- The vector representing the distance between two aircraft is denoted by $\overrightarrow{X_iX_j}$ where X_i (resp. X_j) stands for the location of aircraft i (resp. j).
- The "oblical" distance between two aircraft (i and j) is denoted by

$$d_{ij}^{ob} = \sqrt{< \overrightarrow{X_iX_j}, \overrightarrow{X_iX_j} >} \,, \qquad (6)$$

  where $< .,. >$ stands for the appropriate scalar product.
- We denote by $\overrightarrow{v}_{ij} = \overrightarrow{v}_j - \overrightarrow{v}_i$ the speed difference between two aircraft.
- The derivative of the "oblical" distance between two aircraft is denoted by v_{ij} and writes

$$v_{ij} = \frac{< \overrightarrow{X_iX_j}, \overrightarrow{v}_{ij} >}{d_{ij}^{ob}} \,. \qquad (7)$$

- We introduce a weighting function f. As suggested in [27], we used

$$f(d_{ij}^{ob}) = \frac{e^{-\alpha (d_{ij}^{ob})^2} + e^{-\beta d_{ij}^{ob}}}{2} \,, \qquad (8)$$

  with $\alpha = 0.002$, $\beta = 0.01$, the d^{ob} being expressed in nautical miles.

These indicators are defined pointwise. To get a value on the controlled airspace, they have to be averaged on the different aircraft. In [19], a density indicator is defined as follows

$$\mathit{Dens}(i) = \sum_{j=1}^{N} b f(d_{ij}^{ob}) \,. \qquad (9)$$

Two indicators are introduced to reflect the variability in headings (*track_disorder*) and speed (*speed_disorder*). There are defined as

$$\mathit{track\_disorder}(i) = \sum_{j \neq i} |\theta_i - \theta_j| f(d_{ij}^{ob}) \,. \qquad (10)$$

$$\mathit{speed\_disorder}(i) = \sum_{j \neq i} \| \overrightarrow{v}_{ij} \| f(d_{ij}^{ob}) \,. \qquad (11)$$

Indicators *Div* et *Conv* respectively describe convergency and divergency of the aircraft in the controlled sector.

$$\mathit{Div}(i) = \sum_{\substack{j=1 \\ j \neq i}}^{Nb} 1_{\mathbb{R}^-}(v_{ij}) \,.\, |v_{ij}| \, f(d_{ij}^{ob}) \,, \qquad (12)$$

$$\mathit{Conv}(i) = \sum_{\substack{j=1 \\ j \neq i}}^{Nb} 1_{\mathbb{R}^+}(v_{ij}) \,.\, |v_{ij}| \, f(d_{ij}^{ob}) \,. \qquad (13)$$

Indicators Sd_+ and Sd_- are designed to set a weight on potential conflicts that are difficult to solve. These "sensitivity" indicators are defined by

$$Sd_-(i) = \sum_{\substack{j=1 \\ j \neq i}}^{Nb} 1_{\mathbb{R}^-}(v_{ij}) \, \| \overrightarrow{\nabla} v_{ij} \| f(d_{ij}) \,, \qquad (14)$$

$$Sd_+(i) = \sum_{\substack{j=1 \\ j \neq i}}^{Nb} 1_{\mathbb{R}^+}(v_{ij}) \, \| \overrightarrow{\nabla} v_{ij} \| f(d_{ij}) \,. \qquad (15)$$

Note that components of the gradient are weighted so as to reflect the difficulty of the respective manoeuvers[3]. As observed in [19], a situation with a high "sensitivity" is easier to resolve for the air controller than one with a low "sensitivity". As these indicators "increase" with the number of aircraft, it is unclear whether they actually are "complexity" or "simplicity" indicators. We thus define a last pair of indicators, *insen_c* and *insen_d* as

$$\mathit{insen\_c} = \frac{\mathit{Conv}^2}{Sd_+} \quad \text{and} \quad \mathit{insen\_d} = \frac{\mathit{Div}^2}{Sd_-} \,. \qquad (16)$$

Modified PRU metrics

The work conducted by SDER-RFM for the Performance Research Unit (citeRFM), though initially designed to compare ATC centers on a daily basis, inspired the following indicators :

- *inter_hori*: number of potential crossings (irrespective of the aircraft direction on their trajectories) with angle greater than 20 degrees.
- *inter_vert*: denote by n_1, n_2 et n_3 the numbers of stable/climbing/descending aircraft. The indicator is then defined as

$$\texttt{inter\_vert} = \frac{(n_1 n_2 + n_2 n_3 + n_1 n_3)}{(n_1 + n_2 + n_3)} \,. \qquad (17)$$

- *avg_vs*: this is simply the average vertical speed of controlled aircraft.

[3]Reasonable weights were given by P. Averty and M. Tognoni.

Metrics inspired from the CREED project

The work of P. Averty on conflict detection [20] inspired a set of indicators. One of the ideas in [20] is that conflict perception is "plannar". The author thus defines for converging pairs of aircraft the following quantities

- *Ed* : minimum horizontal distance between aircraft.
- *Efl* : horizontal distance when the aircraft are vertically separated (after the crossing).
- *Da* : the "anticipation degree", i.e. the distance between the faster aircraft and the intersection of the aircraft trajectories (in the horizontal plan). We replace this variable to a modified *Da*, *DaC*, which stands for the greater distance between one of the aircraft and the point where, horizontally, the distance between aircraft is the smallest. For explanations about this substitution, we refer to [16].

Originally, these quantities are defined to describe conflict perception. To translate the idea of [20] in terms of traffic complexity, we assume that a conflict is all the more critic that the expected separation (*Ed* and *Efl*) and the anticipation (*DaC*) are small. We thus set

$$\texttt{creed} = \frac{1}{\alpha\ \texttt{Da} + (1-\alpha)(\beta\ \texttt{Ed} + (1-\beta)\ \texttt{Efl})}\ , \quad (18)$$

where α and β are parameters in $[0;1]^4$. Finally, aircraft pairs considered in [20] are such that vertical separation occurs prior to separation, as the converse situation is avoided as much as possible by controllers. Accordingly, the complexity associated with these latter pairs is likely to be greater and we distinguished the two kind of conflicts by summing the quantity fintroduced in (18) on both sets of aircraft, thus creating two distinct indicators, *creed_ok* ("good pairs") and *creed_pb* ("bad pairs").

[4] As for now, these parameters are set equal to 0.5, but are meant to be adjusted and possibly vary with *DaC* to reflect the results of ongoing research [21].