JOURNAL OF AIRCRAFT
Vol. 37, No. 2, March–April 2000

Neural Network Forecasting of Service Problems for Aircraft Structural Component Groupings

Lars H. Nordmann* and James T. Luxhoj†
Rutgers University, Piscataway, New Jersey 08854-8018

The Federal Aviation Administration (FAA) in the United States is responsible for regulating aircraft traffic and safety. A significant increase in domestic air traffic coupled with an aging population of aircraft has led the FAA to initiate new aircraft safety research efforts. These efforts are intended to provide the FAA's aviation safety inspectors (ASIs) with the means to evaluate and to control appropriate surveillance levels for aircraft operators. One of the FAA's databases is the service difficulty report (SDR) system. It provides FAA inspectors with reliability and airworthiness statistical data necessary for planning, directing, controlling, and evaluating certain assigned safety and maintenance programs. Neural network forecasting models are developed that predict the number of SDRs for aircraft structural component groupings, referred to as Air Transportation Association chapters. Data are used from two specific operators with homogeneous fleets, that is, same aircraft make. It is an extension of previous SDR forecasting research in that it stratifies forecasts by structural component groupings.

I. Introduction

THE Federal Aviation Administration (FAA) in the United States is responsible for regulating aircraft traffic and safety. An expected increase in usage of domestic flights in the next few years coupled with an aging population of aircraft has led the FAA to initiate new aircraft safety research efforts.[1] An important aviation safety performance measure is the number of service difficulty reports (SDRs), which is an airworthiness or maintenance-related performance measure.

The service difficulty information provides FAA safety inspectors with aircraft component reliability and airworthiness statistical data necessary for planning, directing, controlling, and evaluating certain assigned safety and maintenance programs.[2] This system also provides FAA managers and inspectors with a means for measuring the effectiveness of the self-evaluation techniques being employed by certain segments of the civil aviation industry. The completion of an SDR requires careful review of the reported discrepancy and supporting data. An effective evaluation of the extent of the problem and its causes is essential for determining corrective action. If the opportunity exists, the inspector usually reviews prior reports for possible trends, for example, vendor problems, manufacturer equipment problems, training, and/or procedural problems. However, there are currently no systematic or quantitative techniques used for identifying possible trends. Trending analysis is based on visual inspection of graphical data plots. Thus, aviation safety inspectors (ASIs) in the field repeatedly expressed strong interest in reliable forecast techniques for the number of SDRs. Such forecasts would be a helpful tool for the management of human resources, their effective placement, and scheduling.

Estimation of the total number of SDRs in a given time interval that a particular airline would be expected to have, stratified by structural aircraft component groupings, could help to identify situations in need of heightened level of surveillance by the FAA's safety inspectors, for example, if the airline's number of SDRs is far above or below what should be expected. An excessive number of SDRs in a given time period could suggest mechanical, operating, or design problems with certain aircraft. Whereas too few SDRs reported in a given time may not necessarily be problematic, an expert panel of safety inspectors noted that a very low number of SDRs for an airline in a given time period could possibly suggest organizational or management problems, lack of regulatory compliance, airline maintenance cutbacks, or financial or labor problems. Both high and low numbers of SDRs would merit closer scrutiny by FAA safety inspectors.

Transport Canada, the Canadian counterpart of the FAA, has recently developed and is currently implementing an automated trending and monitoring system (ATMS). This system was developed to identify changes to normal levels in the number of SDRs. The paradigm of the ATMS was borrowed from the quality control discipline. In the first phase, the ATMS utilizes a control charting technique, comparing the number of SDRs for a given month against an upper control limit established from previous data. In a second phase, the ATMS analyzes the aggregate number of SDRs over the last series of months to detect an aggregate trend in the number of SDRs. The upper control limit used in the first step is calculated based on the Chebyshev inequality. The Chebyshev inequality provides distribution-independent bounds for the marginal probabilities of a stochastic process. As such, it is a powerful inequality that does not require any knowledge about the distribution of the underlying process. Unfortunately, estimates obtained from the Chebyshev inequality are relatively weak estimates. Nonetheless, applying the ATMS to historical data has shown that ATMS is capable of identifying potential problems. Currently, the FAA is collaborating with Transport Canada to refine the ATMS and to test and eventually to implement it in the United States. Nonetheless, the ATMS, based on quality control principles, is a reactive tool, not a proactive tool.

Neural networks naturally lend themselves to time series and forecast analysis of complex processes. They, too, do not require any knowledge of an underlying distribution, nor do neural networks require exact mathematical models. Rather they attempt to recognize relationships based on training patterns. They can be trained to recognize relationships in the time domain as well as correlations among a host of other variables. Even though neural network technology is becoming a mature technology, its application to the aviation industry in general and SDR forecasting in particular has not yet been thoroughly investigated. Some earlier research results are discussed next. This paper is aimed at extending these results and the application of neural networks to the forecasting of SDRs stratified by Air Transportation Association (ATA) chapter codes.

Luxhoj et al.[3] and Shyur et al.[4] report on SDR forecasting models for aircraft. The data used in their research investigation included a subset of the SDR database that had been merged with the aircraft utilization aviation research and support database for the same fleet.

Received 17 February 1999; revision received 11 October 1999; accepted for publication 20 October 1999.

*Ph.D. Candidate, Department of Industrial Engineering, 96 Frelinghuysen Road. nordmann@rci.rutgers.edu.

†Associate Professor, Department of Industrial Engineering, 96 Frelinghuysen Road; jluxhoj@rci.rutgers.edu.

A data grouping method is used to obtain a population model. Multiple regression and neural network models were studied, and the forecasting accuracy for each method was reported. In their studies, the original ungrouped data set appeared to be noisy. A population concept proved to be a very effective modeling technique for both the regression analysis and in the construction of neural networks for determining strategic safety inspection indicators. Whereas the population concept is constructive for developing models to predict national norms for SDR reporting, there is a loss of information in grouping the data.

In another research effort, Luxhoj[5] refined the population model analysis by focusing on specific operators with homogeneous fleets, that is, same aircraft make. It is an extension to the previous research,[2] but now provides more meaningful information because the aircraft data in the earlier study were composed of numerous operators with mixed fleets and differing operating and maintenance policies. Again, multiple regression and neural network models were the principal two forecasting methods examined. Autoregression, exponential smoothing, and moving average forecasting techniques were also evaluated.

II. Problem Domain: Motivation and Goal

The current research effort was sparked by interest of ASIs who expressed that a further specification in the SDR forecast would be extremely useful for direct field application. In this sense, the current research is a further extension to studies by Luxhoj et al.,[3] Shyur et al.,[4] and Luxhoj.[5] It builds on the experience and results from these studies, but introduces yet another data stratification. The current research, again, used data from specific operators with homogeneous fleets, that is, same aircraft make. However, instead of forecasting the overall number of SDRs, the current research focuses on the number of SDRs by structural and functional aircraft component groupings (as grouped by ATA chapters and described subsequently).

Structural and functional components common to most aircraft are coded in the ATA specification 100 code or the Joint Aircraft System Component (JASC) code. Further logical, syntactical, and historical clarification on these coding systems is given next.

The first aircraft structural/functional component coding system was developed by the FAA in the mid-1960s and became known as the FAA Aircraft System/Component Code. It was an eight-digit alpha–numeric code developed around the computer technology of that period. It consisted of a four-digit numerical code plus four alpha characters.

Later the ATA specification 100 code (ATA code) was developed and maintained by an FAA contractor. Advances in computer technology made a reduction from an eight- to a four-digit code possible. The ATA code is a four-digit code, of which the first two digits reference a major structural/functional system or component grouping on an aircraft, whereas the third digit references a subsystem/subcomponent. The fourth digit is not referenced.

During the late 1980s, the FAA initiated research efforts to develop their own coding system. The resulting product is the JASC code. It was introduced in May 1991 and is now suggested for use. Once implemented, it will be mandatory for reporting in all major FAA databases, such as the SDR or accident/incident data system. Coding of historic records in those databases will be updated to reflect the new coding system. In most cases, the first three digits of the JASC code match the first three digits of the ATA code. The JASC, however, because of later development does divert in some areas from the ATA code to reflect technological advances that led to a significant increase in some structural areas and made a subdivision of others more meaningful. For example, the FAA code 5301SXBD (body, section structure) has been expanded to 20 items due to the high rate of reporting in this area for the year 1989 (8021 reports were received). Also, the JASC code divides the engine section into two code groups to separate turbine and reciprocating engines. As with the ATA code, the JASC code is a hierarchical code. Its first two digits specify the major structural/functional aircraft component groupings and are referred to as the JASC chapter, whereas the last two digits specify the subsystem/subcomponent.

Table 1 ATA chapters and codes

ATA chapter code	Category
	Aircraft
11	Placards and markings
12	Servicing
18	Helicopter vibration
	Airframe systems
21	Air conditioning
22	Autoflight
23	Communication
24	Electric power
25	Equipment/furnishing
26	Fire protection
27	Flight controls
28	Fuel
29	Hydraulic power
30	Ice and rain protection
31	Instruments
32	Landing gear
33	Lights
34	Navigation
35	Oxygen
36	Pneumatic
37	Vacuum
38	Water/waste
45	Central maintenance system
46	Airborne auxiliary power
51	Standard practices/structures
52	Doors
53	Fuselage
54	Nacelles/pylons
55	Stabilizers
56	Windows
57	Wings
	Propeller/rotor systems
61	Propellers/propulsors
62	Main rotor
63	Main rotor drive
64	Tail rotor
65	Tail rotor drive
67	Rotor flight control
	Powerplant system
71	Powerplant
72	Turbine/turboprop engine
73	Engine fuel and control
74	Ignition
75	Air
76	Engine controls
77	Engine indicating
78	Engine exhaust
79	Engine oil
80	Starting
81	Turbocharging
82	Water injection
83	Accessory gearboxes
85	Reciprocating engine

As of today, the ATA code is still in place, and thus the analysis was based on that ATA code. A list of chapter codes is provided in Table 1. ATA chapter codes are generally divided into four categories, aircraft, airframe systems, propeller/rotor systems, and powerplant systems.

III. Neural Networks

Neural network technology mimics the brain's own problem solving process. Just as humans apply knowledge gained from past experience to new problems or situations, a neural network takes previously solved examples to build a system of neurons that make new decisions, classifications, and forecasts.

Neural networks look for patterns in training sets of data, learn these patterns, and develop the ability to correctly classify a new pattern or to make forecasts and predictions. Neural networks excel at problem diagnosis, decision making, prediction, and other

classifying problems where pattern recognition is important and precise computational answers are not required.

McCulloch and Pitts[6] introduced the fundamental idea of neural networks. Neural networks gained in popularity at the beginning of the 1980s.[7,8] Neural networks (NNs) consist of relatively simple processing elements (nodes or units) connected by links. A unit receives the signal from the input links and computes an activation level that it sends to the next layer along the output links. The computation can be divided into two parts. The first part is a linear function that computes the weighted sum of all of the input variables. The second part is a nonlinear function (activation function) that decides whether the output is be greater than a threshold or not.[9]

The NN structure, in which a particular NN connects its perceptrons and generates outputs, is also referred to as its architecture. There are many distinguishing aspects of NN architectures, such as the number of hidden layers, recurrent vs feedforward networks, discrete vs continuous output networks, jump vs nonjump connection, etc. However, a general aspect of NNs is that of the neural connections that mimic the brain's dendrites. A connection is characterized by its starting and ending neuron and a particular weight that determines how much activation is passed through that connection.

Training an NN, or learning, is the process of adjusting neuron connections such that the prediction error is minimized. Whereas human brains can also build new and break old dendrites, artificial NNs can only adjust the weights of the connection.

In NN terminology, the set of all variables is divided into inputs and outputs. Inputs (or independent variables) are those variables whose values are commonly available. Outputs (or dependent variables) are those that, eventually, one wishes to classify or predict based on known input values.

A set of associated variable values is referred to as a pattern. The set of all available patterns defines the pattern set. This pattern set is divided into three distinct subsets, the training set, the test set, and the production set.

In training an NN, it will repeatedly look at the patterns in the training set. Starting with some initial settings for all of the connection weights, the NN predicts values of the output variables based on the values of the input variables only. An internal algorithm, also referred to as the learning algorithm, will then compare the predicted output values to the actual output values and attempts to adjust the connection weights of the NN such that prediction errors are minimized. Eventually, if the problem can be learned, a stable set of weights adaptively evolves. However, if the artificial NN were only shown the same training set often enough, like a human brain, it would finally lose its capability to reason and instead start to memorize. Such a network would then yield excellent prediction results on the memorized patterns, but would fail on new, unseen patterns. To avoid memorization, or overfitting, in training the NN, from time to time it will be evaluated on the test set. This process is referred to as calibration. Finally, to evaluate the trained network's power, it can be tested on the production set. This is a reserved set of patterns that the NN has never seen before. If the underlying problem can be learned and the NN was able to learn primary variable associations, then it should also yield good results on this production set.

As already indicated, NNs can be distinguished between continuous-output and classification networks. Which one to use is determined by the underlying problem. However, there are many practical continuous-output situations where a classification output is more meaningful than a continuous output, for example, in fuzzy logic applications. In this research, we have used a classification NN because the knowledge of the exact number of SDRs expected is not as important as knowing whether the expected SDR count is high vs low vs average. Further justification on this decision is given later in Sec. IV. A popular classification network is the probabilistic NN (PNN).[10,11] PNNs are known for their ability to train quickly on sparse data sets. They are supervised networks that separate data into a specified number of output categories. PNNs are three-layer networks wherein the training patterns are presented to the input layer and the output layer has one neuron for each possible category. There must be as many neurons in the hidden layer as there are training patterns. In training the network, the membership of a pattern to any one of the output categories is defined by 0–1 indicator variables. The trained network, on the other hand, produces activations in the output layer corresponding to the probability density function estimate for that category. The highest output represents the most probable category and is commonly selected as the network's classification prediction.

For PNNs, calibration can either be turned off or set to either iterative or genetic adaptive. With calibration turned off, the user has to specify a smoothing factor that will be the same for all links in the network. In some instances this feature might be helpful, for example, when dealing with multiple outputs that exhibit different sensitivity to the set of inputs. In general, however, it is recommended to let the NN choose a smoothing factor via calibration. With iterative calibration, the network training is divided into two parts. In the first part, the network is trained with the data in the training set. In the second part, calibration is used to test a whole range of smoothing factors, trying to converge on one that works best on the test set. The smoothing factor will be the same for all inputs. Thus, iterative calibration should only be used when it is justified to assume that all inputs have the same impact on predicting the output. An advantage of iterative calibration is that training proceeds quickly. With genetic adaptive calibration, the network determines an individual smoothing factor for each input. Again, network training proceeds in two parts as for the iterative option. The only difference is that for the genetic adaptive option, the second part uses calibration to find a whole range of smoothing factor combinations that works best on the test set. In further research, such individual smoothing factors can be used as a sensitivity analysis tool. The genetic adaptive method will produce networks that work much better on the test set but will take much longer to train. Both calibration methods were evaluated in this research, and the genetic adaptive method produced the best results for the data used in our study.

Genetic algorithms use a fitness measure to determine which of the individuals in the population survive and reproduce. Thus, survival of the fittest causes good solutions to evolve. For PNNs, the breeding pool size describes the number of smoothing factor combinations that are currently the fittest and which will be further refined through adjustments defined by an internal algorithm. The parameter breeding pool size can be set by the user. Larger breeding pool sizes result into potentially better networks. However, a large breeding pool size slows down training process. Generally speaking, larger breeding pool sizes only marginally improve the network's quality, whereas genetic adaptive calibration significantly improves the network's quality vs iterative or no calibration.

In this research, we utilized the NeuroShell 2 (Ref. 12) software to construct and evaluate NNs.

IV. Modeling

Because NNs are essentially black boxes, the modeling does not center around the development of a structural (or physical) model. Rather, NN modeling can be divided into the following subtasks: selection of model inputs/outputs, choice of an appropriate neural network architecture, data pre- and postmanipulation, and training of the NN.

This section will describe in detail the steps involved in the actual network modeling. We begin with a data description, continue with a preanalysis, discuss logical consequences, and describe the NN design.

A. Data Description

The variables that are thought to be logically related to the SDR analysis and that were found to be statistically relevant in previous studies[3–5] can be divided into five classes.

1. Fleet information consists of statistics regarding the fleet of a particular carrier. It includes the fleet's size, the fleet's composition, the fleet's average age, the fleet's operational region, and the fleet's usage, such as average number of cycles per month and average flight time per cycle. Only data on fleet size were available for this current study.

2. Operations (OPS) surveillances indicate the number of unfavorable OPS surveillances for a given month as well as the total number of OPS surveillances in that month. OPS surveillances focus on an operator's procedural knowledge to follow safety regulations (e.g., preflight checks, inflight operations, emergency measures, etc.) and the performance abilities of the crew (e.g., the flight attendants' knowledge of their jobs, the pilot's need to maintain certification, etc.). We utilized both the total number of OPS surveillances and the number of unfavorable OPS surveillances.

3. Airworthiness (AW) surveillances indicate the percentage of unfavorable AW surveillances for a given 1-month period. AW surveillances focus on the maintenance-related aspects of safety performance, that is, procedures, log books, equipment, preventive maintenance schedules, etc. We utilized both the total number of AW surveillances and the number of unfavorable AW surveillances.

4. Event information consists of numerical and narrative information about any event as defined by FAA regulations. These regulations distinguish between accident, incidents, and occurrences. The distinguishing criterion is the severity of the event. Whereas accidents are defined through either personnel injury or major damage to the aircraft, incidents are events with less severe damages, and occurrences are characterized by either minor damages or action that did not result into a damage but were intended to avert damage such as an aborted takeoff. In our study, we utilized the event counts comprising number of accidents, number of incidents, and number of occurrences.

5. SDR information consists of numerical and narrative information about submitted SDRs as described earlier. In our study we utilized the number of SDRs by ATA chapter code. All counts as utilized in this study are monthly aggregates.

Because of the new concept of data stratification by ATA chapter codes in the current study, a natural data scarcity was imposed. Instead of having total numbers of SDRs, the records were now disaggregated according to their ATA chapter code, leaving each of these categories with a significant lower number of SDRs.

B. Data Preanalysis, Consequences, and Data Manipulation

From previous studies, it is known that best results are achieved for carriers with homogeneous fleets. This is true because focusing on homogeneous fleets eliminates effects related to fleet composition. Naturally it is expected that different aircraft types exhibit different mechanical problems. Therefore, the current study analyzes the same two carriers as in the previous studies, both with homogenous fleets.

To avoid the mentioned problems with data scarcity due to the disaggregation of the SDRs by ATA chapter codes and to keep attention on the concept, we have focused only on SDRs with the four most frequent ATA chapter codes: 53 fuselage, 52 doors, 33 lights, and 57 wings. The study period was chosen according to data accessibility, that is, five years from March 1993 through February 1998.

To construct an effective and more accurate NN, a forecast of actual numbers of SDRs was not intended, nor is it possible or meaningful. Such a forecast would obviously be unreliable and without practical use. Instead, categorizing numbers of SDRs (by ATA chapter codes) into high (H), medium (M), and low (L) will yield more useful information. This, of course, raises the question of how to partition between H, M, and L. Such divisions are obviously based on historical records inasmuch as H, M, and L imply a relative meaning. However, there is nothing like an optimal statistical division, and the categorization is strictly defined by the functional requirements. To uncomplicate the current analysis and avoid basing it on wrong assumptions, the current study assumes a straightforward categorization as described in the following. For every ATA chapter code (and each of the two considered aircraft operators), the number of SDRs was plotted over the 5-year study period. A trend line was computed, and the deviation of the actual numbers of SDRs from the trend was used to compute a sample standard deviation. The residuals of the actual numbers relative to the trend line (the fit) roughly resembled a normal distribution. Thus, with the trend line as a function of the expected number of SDRs over time, the computed standard deviation, and the assumption of a normal distribution, a data categorization was possible. Two more trend lines (and upper trend line T_u and a lower trend line T_l) were computed such that statistically 33.3% of all observations fall above T_u, 33.3% fall between T_u and T_l, and 33.3% fall below T_l (see Fig. 1).

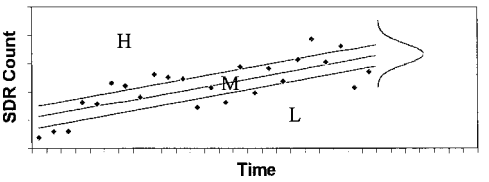

Fig. 1 Categorization of SDR counts into H, M, and L.

Because the number of unfavorable surveillance inspections is only meaningful with respect to the total number of surveillance inspections, these two variables were combined to define one new variable, the percentage of unfavorable inspections (for OPS and AW).

Furthermore, the numbers of incidents and occurrences were collapsed because the distinction between them is not always clear and sometimes confused during reporting. To account for first- and second-order time effects as well as for autocorrelation, the data have also been lagged one and two months. To level the effects of data spikes, data were once analyzed using actual monthly data and then a second time by using 3-month moving averages instead.

C. NN Model

As suggested by the data and the underlying phenomenon, SDR counts were categorized, and supervised PNNs were selected, trained, and tested. Experiments with different data and network architectures did confirm that best results are achieved with PNNs.

For comparison, the data were not categorized in H, M, and L, and three continuous-output network architectures were tested: a three-layer standard backpropagation NN (BPNN), a Jordan–Elman recurrent BPNN, and a general regression network (GRNN).

Standard BPNNs are NNs where each layer is connected to only the immediately previous layer. These networks are the most commonly used network architectures and are known to perform especially well for pattern recognition.

Recurrent BPNNs are similar to standard BPNNs, except the only difference in structure is that there is one extra slab in the input layer that is connected to a hidden layer just like the other input slab. This extra slab holds the contents of one of the layers as it existed when the previous pattern was trained. In this way, the network sees previous knowledge it had about previous inputs. This extra slab is sometimes called the network's long-term memory. Note that feeding the input layer from one pattern into the input layer of the next pattern is similar to but more powerful than giving the network previous values of each of the inputs. Recurrent networks that feed the hidden layer back into the input layer are commonly called Jordan–Elman networks. Recurrent networks are known for their ability to learn on sequences (time series), which makes them an invaluable tool for the analysis of data with a temporal structure. If there is no temporal structure, using a recurrent BPNN will not work as well as a standard BPNN because the long-term memory slab will introduce random noise into the network.

A GRNN is a three-layer network with one hidden neuron for each training pattern. Each output is evaluated independent of the other outputs, and the network trains in one pass of the training data. GRNNs are known for their ability to train quickly on sparse data.

All three architectures were trained on and applied to the raw SDR counts, but none of the architectures produced reasonable results.

The three-layer back-propagation network produced R^2 values ranging from 0.00 to 0.82. In continuous-output NN modeling, R^2 compares the accuracy of the model with the accuracy of a trivial benchmark model, where the prediction is simply the mean of all sample patterns. A perfect fit would result in an R^2 of 1, a very good fit near 1, and a poor fit near 0. If the NN model predictions are worse than one could predict by just using the mean of the sample case outputs, R^2 will be 0. Although not precisely interpreted in the

same manner as the R^2, the coefficient of multiple determination in regression modeling, the R^2 in NN modeling can be used as a measure of model fit. NeuroShell2 uses the following formula for R^2: $R^2 = SSE/SS_{YY}$, where

$$SSE = \sum (y - \hat{y})^2, \qquad SS_{YY} = \sum (y - \bar{y})^2$$

and where y is the actual value, $\hat{y}$ is the predicted value of y, and $\bar{y}$ is the mean of they values.

The R^2 values for the recurrent network ranged from 0.01 to 0.57, and the R^2 values for the GRNN ranged from 0.09 to 0.62. Note that though the upper ends of the R^2 ranges seem high, these R^2 values were obtained only for the forecast of one particular group of SDRs (SDRs stratified by one ATA code). Forecasts of the other seven groups resulted in R^2 values as low as 0.00. This indicates that the network architectures might fit one group of SDRs particularly well due to overfitting or random data effects. In general, however, these network architectures appear not to be suited for SDR forecasting. Therefore, no further experiments were made with these types of network architectures.

The analysis has been divided into two separate parts, one for and with actual monthly data and one for and with smoothed (3-month moving averages) data. In each of these two parts, the two considered aircraft operators have been analyzed separately. For each of the two operators, the four most frequent ATA chapter codes have been considered separately. For each of these four ATA chapter codes, three separate neural networks have been trained and tested, one with a forecast horizon of 1 month, one with a forecast horizon of 2 months, and one with a forecast horizon of 3 months. Thus, altogether 48 different neural networks have been trained and tested.

All NNs have a set of 21 inputs and 3 outputs each. The inputs are as follows: fleet size for months t, $t-1$, and $t-2$ (3); percent unfavorable OPS surveillance for months t, $t-1$, and $t-2$ (3); percent unfavorable AW surveillance for months t, $t-1$, and $t-2$ (3); number of accidents for months t, $t-1$, and $t-2$ (3); number of incidents plus occurrences for months t, $t-1$, and $t-2$ (3); and three indicator variables for the category of SDR count (H, M, L) for months $t-1$ and $t-2$ (6).

The outputs are three indicator variables for the category of SDR count (H, M, L) for the respective forecast.

The 24 neural networks developed for the actual monthly data and the 24 neural networks developed for the smoothed (3-month moving average) data differ only in that the first use actual data for the inputs and outputs, whereas the later ones use smoothed data for all inputs and outputs.

Each of the 48 data sets (for actual and smoothed data, for each of the two operators, for each of four ATA chapter codes, and for all three forecast horizons) had 52 patterns total, corresponding to the 52 in the 5-year study period. Note that the first four and the last four months had to be cut off due to 3-month averaging and data lagging.

Each of the 48 data sets has been divided into training, test, and production sets. A division of 60% (training set), 20% (test set), and 20% (production set) was made and resulted into 32, 10, and 10 patterns, respectively, for the training, test, and production sets.

The actual selection of patterns was random because experiments with the selection indicated no crucial sensitivity and randomness is deemed the best choice in absence of other reasoning. However, to allow for comparison among the 48 models, the same randomly chosen pattern division was applied to all 48 models.

Furthermore, the NNs were robust to changes in the default parameters. Thus default parameters were used for the training of all 48 models. The parameter settings for the PNN are summarized as follows: slab 1 (layer 1) 21 neurons and linear [0, 1] scale function, slab 2 (layer 2) 32 neurons, slab 3 (layer 3) 3 neurons, vanilla Euclidean distance metric, genetic adaption calibration, and a breeding pool size of 20.

V. NN Modeling Results

After training, each one of the NNs was applied once to a set of all patterns and once to the production set only. Tables 2–5 summarize the results of the 24 predictive models for the smoothed (3-month averages) data applied twice, once to the pattern set and once to the production set. There are another 24 predictive models for the actual (not smoothed) data, but the results are not shown here.

A. Quality Performance Measures

Because all of the models are categorical networks, standard measures of prediction accuracy, such as the mean square error or R^2, were not computed. Instead two other measures of prediction quality were defined:

The first, success rate, is measured as

$$\text{S-rate} := \frac{(\text{number of correctly classified patterns})}{(\text{total number of patterns})}$$

The second performance measure is the α value of rejecting the NN as a random generator. This performance measure builds on the theory of hypothesis testing. In hypothesis testing, a type I error is committed if a null hypothesis is rejected when it is, in fact, true. The

Table 2 Correct NN classifications, by ATA chapter code and by forecast horizon (operator A, for pattern set, smoothed data)[a]

	Number of correct classifications											
Parameter	41	42	43	44	45	46	47	48	49	50	51	52
ATA number												
53	——	——	——	——	——	$t+2$	$t+3$	$t+1$	——	——	——	——
33	——	——	——	$t+3$	——	$t+2$	——	——	——	——	$t+1$	——
57	——	——	——	——	$t+2, t+1$	——	$t+3$	——	——	——	——	——
52	$t+3$	——	——	——	$t+1$	$t+2$	——	——	——	——	——	——
S-rate, %	79	81	83	85	87	88	90	92	94	96	98	100

[a]Here 0% α error.

Table 3 Correct NN classifications, by ATA chapter code and by forecast horizon (operator B, for pattern set, smoothed data)[a]

	Number of correct classifications											
Parameter	41	42	43	44	45	46	47	48	49	50	51	52
ATA number												
53	$t+1$	——	——	$t+3$	——	——	——	$t+2$	——	——	——	——
33	——	——	——	——	$t+2$	——	——	——	$t+3$	——	$t+1$	——
57	$t+2$	——	——	——	——	$t+1$	——	——	$t+3$	——	——	——
52	——	——	——	——	——	$t+3$	——	——	$t+2$	$t+1$	——	——
S-rate, %	79	81	83	85	87	88	90	92	94	96	98	100

[a]Here 0% α error.

Table 4 Correct NN classifications, by ATA chapter code and by forecast horizon (operator A, for production set, smoothed data)

	Number of correct classifications										
Parameter	0	1	2	3	4	5	6	7	8	9	10
ATA number											
53	——	——	——	——	——	$t+3, t+2$	——	——	$t+1$	——	——
33	——	——	——	——	$t+3$	——	——	$t+2$	——	$t+1$	——
57	——	——	——	——	——	——	$t+3, t+1$	——	$t+2$	——	——
52	——	——	——	——	——	$t+3, t+2$	$t+1$	——	——	——	——
S-rate, %	0	10	20	30	40	50	60	70	80	90	100
α-error, %	100	98	90	70	44	21	8	2	0	0	0

Table 5 Correct NN classifications, by ATA chapter code and by forecast horizon (operator B, for production set, smoothed data)

	Number of correct classifications										
Parameter	0	1	2	3	4	5	6	7	8	9	10
ATA number											
53	——	——	——	$t+1$	$t+3$	——	——	$t+2$	——	——	——
33	——	——	——	——	——	——	——	$t+3, t+2$	——	$t+1$	——
57	——	——	——	$t+2$	——	——	——	$t+3$	$t+1$	——	——
52	——	——	——	——	——	$t+3$	——	$t+2$	$t+1$	——	——
S-rate, %	0	10	20	30	40	50	60	70	80	90	100
α-error, %	100	98	90	70	44	21	8	2	0	0	0

probability of committing this type of error is commonly denoted by α. In other words, α is the probability that the observed results could have occurred under the null hypothesis. In our case, the null hypothesis is the hypothesis that the NN is a random generator that picks any predicted value at random out of the possible output value domain (H, M, L). Then, under the null hypothesis, the probability of picking k out of n output values correctly is

$$\alpha = \binom{n}{k} \cdot \left(\frac{1}{3}\right)^{k} \cdot \left(\frac{1}{3}\right)^{(n-k)}$$

B. Quantitative Findings

Tables 2–5 summarize the prediction quality of the NNs for the smoothed (3-month moving average) data. Each table summarizes the quality of the three forecast horizons ($t+1, t+2, t+3$) for each of the four ATA chapter codes. The top row indicates how many predictions are accurate for the respective NN model. The bottom two rows (Tables 4 and 5) indicate the prediction quality of the respective model with regard to the earlier defined two performance measures of the prediction quality. Tables 2 and 3 summarize the prediction quality on the pattern set (for operators A and B). Tables 4 and 5 summarize the prediction quality on the production set (for operators A and B).

For example, Table 3 summarizes the prediction results of the NNs on the pattern set (52 patterns) for operator B with smoothed (3-month moving average) data. The NN for ATA code 33 and a 2-month forecast horizon, for instance, classified 45 out of the 52 patterns correctly, corresponding to a 87% success rate and a zero probability that such accuracy could have been randomly achieved.

C. Qualitative Findings

As already indicated, the NNs were robust to changes in the default parameters and not sensitive to the selection of training, test, and pattern sets. This suggests that the underlying problem is learnable and that the NNs were able to learn variable associations.

In most cases, the NNs provide good to excellent forecasts. The two prediction quality measures do further indicate some degree of regularity or correlation in the underlying process as well as the capability of the NN to recognize those relationships. Note that the success rate expected by a random generator is only 33.3% (not 50%).

Generally, the forecast quality is excellent on the pattern sets and acceptable on the production sets. Though results from NNs applied to the production set give generally a more realistic picture of the network's quality, note that this study was based on a relatively small sample size of only 52 patterns. This warrants caution in overevaluating the results from the production set, which consisted of only 10 patterns.

As another rule of thumb it can be seen that the forecast quality generally decreases with increasing forecast horizon ($t+1$ fields are predominately on the left, $t+2$ fields are in the middle, and $t+3$ fields are on the left-hand side of Tables 2–5). This is in agreement with what is to be expected by forecast models. That is, the longer the forecast horizon, the lower the prediction accuracy. It further justifies validity of the NNs as well as the hypothesis of existing relationships (in the time domain and in the performance measure domain) in the underlying data.

VI. Conclusions

An NN represents a powerful tool in modeling complex data relationships and generating high-accuracy forecasts without the necessity of understanding the underlying physical relationship. For data with random spikes and outliers, as in the underlying study, PNNs are most appropriate and will yield best results for classifying data patterns. Data smoothing through consideration of moving averages will further remove spikes and increase the NN quality. However, oversmoothing has to be balanced against functional requirements. For the given data in this study, a 3-month moving average is appropriate.

In the current study, many key variables that have been found important in previous studies were no longer available. Nonetheless, the developed NNs provided good to excellent prediction results based on easy-to-interpret, tangible measures of prediction quality. Refinement of the developed models can be achieved through recapture of the lost variables, further selection, and/or manipulation of data. In general, the more the better does not apply to the number of inputs for NNs. Thus, further statistical studies and communication/feedback from field inspectors are important steps for an optimal network design.

The predictive quality of the developed networks indicates the existence of relationships in the underlying data and the network's capability to recognize them. Further research could focus on the analysis of the weights of NNs in an attempt to derive an actual structural, physical, or mathematical model of the actual SDR submission process.

Acknowledgment

The authors would like to acknowledge the support of the Federal Aviation Administration through Grant 97-G-005.

References

[1]"The Federal Aviation Administration Plan for Research, Engineering, and Development, Vol. I: Program Plan (1989)," U.S. Dept. of Transportation, Rept. 100-591, 1989.

[2]Federal Aviation Administration HANDBOOKS/ORDERS 8300.10 03-128, "Process Service Difficulty Report," Vol. 3, U.S. Dept. of Transportation, Washington, DC, Sept. 1989, Chap. 128.

[3]Luxhoj, J. T., Williams, T. P., and Shyur, H., "Comparison of Regression and Neural Network Models for Prediction of Inspection Profiles for Aging Aircraft," *IIE Transactions on Scheduling and Logistics*, Vol. 29, 1997, pp. 91–101.

[4]Shuyr, H., Luxhoj, J. T., and Williams, T. P., "Using Neural Networks to Predict Component Inspection Requirements for Aging Aircraft," *Computers and Industrial Engineering*, Vol. 30, No. 2, 1996, pp. 257–267.

[5]Luxhoj, J. T., and Cheng, J., "Neural Network Modeling of Aviation Safety Field Studies," *Proceedings of the 7th Annual Industrial Engineering Research Conference*, Banff, Alberta, Canada, 1998.

[6]McCulloch, W. S., and Pitts, W., "A Logical Calculus of Ideas Immanent in Nervous Activity," *Bulletin of Mathematical Biophysics*, Vol. 5, 1943, pp. 115–133.

[7]Hopfield, J. J., "Neural Networks and Physical Streams with Emergent Collective Abilities," *Proceedings of the National Academy of Science*, Vol. 79, No. 182, pp. 2554–2558.

[8]Russel, S. J., and Norvig, P., *Artificial Intelligence—A Modern Approach*, Prentice–Hall, Upper Saddle River, NJ, 1995.

[9]Simpson, P., *Artificial Neural Networks*, Pergamon, New York, 1990.

[10]Specht, D., "Probabilistic Neural Networks for Classification, Mapping, or Associative Memory," *Proceedings of the IEEE International Conference on Neural Networks*, Vol. 1, 1988, pp. 525–532.

[11]Specht, D., "Probabilistic Neural Networks," *Neural Networks*, Vol. 3, 1990, pp. 109–118.

[12]Frederick, M., *NeuroShell 2*, Ward Systems Group, MD, 1993.