METHODOLOGY FOR EVALUATING AND REGRESSION TESTING A CONFLICT PROBE

Mike M. Paglione, Federal Aviation Administration
Robert D. Oaks and Hollis F. Ryan, General Dynamics
William J. Hughes Technical Center, Atlantic City International Airport, NJ 08405

Abstract

A conflict probe is an air traffic management decision support tool that predicts aircraft-to-aircraft and aircraft-to-airspace conflicts. In order to achieve the confidence of the air traffic controllers who are provided this tool, a conflict probe must accurately predict these events. To ensure their continued confidence, the accuracy should not only be assessed in the laboratory before the probe is deployed but continue to be reassessed as the system undergoes upgrades and software changes. Furthermore, it is desirable to use recorded air traffic data to test these tools in order to preserve real-world errors that affect their performance. This paper utilizes a proven approach that modifies surveillance radar track data in time to create traffic scenarios containing conflicts with characteristic properties similar to those encountered in actual air traffic operations. It is these time shifted traffic scenarios that are used to evaluate the conflict probe.

This paper describes the detailed process of evaluating the missed and false conflict predictions, the calculation of the corresponding error probabilities, and a regression testing methodology to examine two runs of the conflict probe to determine if the conflict prediction accuracy has improved or degraded over time. A detailed flight example is presented which illustrates the specific processing involved in conflict accuracy analysis. Next using a scenario of many flights, a methodology utilizing categorical data analysis techniques is applied to determine if a new version of the conflict probe's software significantly improved or degraded in conflict prediction accuracy.

Introduction

In the United States, the overall system of managing and controlling air traffic is known as the National Airspace System (NAS), which is administered by the Federal Aviation Administration (FAA). Detailed procedures involving restrictions on routing, speeds, and altitudes are an integral part of the NAS. These restrictions severely reduce the amount of aircraft traffic that NAS can accommodate, yet are needed to ensure the high level of safety required. A major FAA goal for improving the NAS is to increase the efficiency of aircraft operations while maintaining safety [1]. This is being achieved by introducing technology that both improves safety and allows for reductions in the restrictions imposed by the current NAS. Thus, broad categories of advances in ground and airborne automation are required. One of the most important ground based tools is a conflict detection tool or conflict probe (CP). A conflict probe is a decision support tool that provides the air traffic controller with predictions of conflicts (i.e., loss of minimum separation between aircraft and other aircraft or restricted airspace) for a parameter time (e.g. 5 minutes) into the future. There are two classes of conflict probes: tactical and strategic. These tools both predict conflict events, but the major difference is the time horizon in which the tools make their predictions. The tactical conflict probe is focused primarily on predicting conflict events that are within one to three minutes in the future. In contrast, the strategic conflict probe is focused on predicting conflict events traditionally as much as 20 minutes in the future.

Within each en route center in the United States, air traffic controllers separate and manage aircraft with the aid of the Host Computer System. The Host's Conflict Alert function provides tactical alerts. The upgrade to the Host, still under development, called the En Route Automation Modernization, replaces Conflict Alert with several categories of alerts with the basic function requiring a minimum of 75 seconds warning. The User Request Evaluation Tool (URET), developed by MITRE Corporation's Center for Advanced

Aviation System Development (CAASD), is an example of a strategic conflict probe being deployed by the FAA. It predicts conflicts up to 20 minutes in the future and under normal conditions requiring a minimum of 5 minutes warning. These systems undergo a plethora of testing before implementation, but as these systems are upgraded over time for new aircraft types and/or new functionalities some testing continues. Therefore, an ongoing need exists for testing to determine whether the upgrades have not inadvertently introduced new inaccuracies. This type of testing is often referred to as regression testing in the software community.

At the FAA's William J. Hughes Technical Center, the Conflict Probe Assessment Team (CPAT) within the Simulation and Analysis Group has been evaluating conflict probes for over eight years. This paper presents CPAT's methodology of regression testing a conflict probe. Before the methodology is explained, further description of a conflict probe, accuracy testing methods, test scenarios, and conflict prediction accuracy are presented in detail.

Description of a Conflict Probe

A conflict probe is responsible for predicting both the path an aircraft will fly and potential

conflicts the aircraft will have with other aircraft or with restricted airspace. As illustrated in Figure 1, the aircraft's trajectory (i.e. four dimensional path of the aircraft) and any conflict predictions are based on the flight information and track data (i.e. smoothed radar surveillance reports) from the Air Route Traffic Control Center's (ARTCC) Host Computer System (HCS), weather forecasts from the National Weather Service, and detailed adaptation databases. The databases include aircraft modeling information and system information relating to the airspace and procedures. In general, the conflict probe uses the flight plan and tracked position information generated from the HCS to build and maintain an aircraft trajectory that predicts the flight path of the aircraft. This process includes monitoring the tracked position compared to the trajectory and rebuilding it when necessary. The key element in maintaining a trajectory is that the original predicted path or trajectory is changed as more information becomes available. For example, an amended flight plan is received and trajectory is updated to match the resulting route change. By using these trajectories for all the active aircraft, the conflict probe predicts future conflicts with other aircraft and restricted airspace [2].
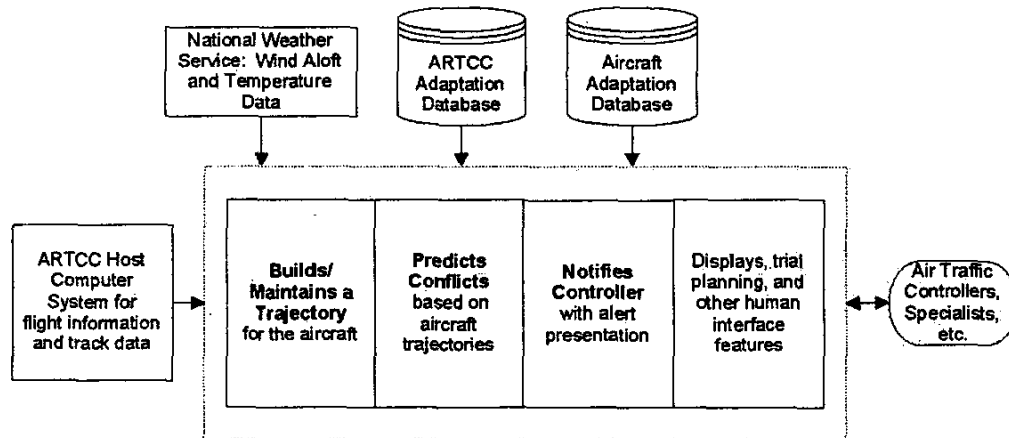

Figure 1. Components of a Conflict Probe's Processing [2]

There are considerable differences between the specific architecture for tactical and strategic conflict probes, as well as the resulting accuracy needs. The major difference is the modeling requirements for the tactical conflict probe predictions are considerably less than the strategic version due to the difference in look-ahead time. However, due to the imminence of the conflict, the tactical conflict probe's accuracy is much more critical to safety. This need for high accuracy is feasible under normal conditions, since the underlying trajectory prediction of an aircraft is significantly higher at lower look-ahead times. There is also an implicit relationship between the modeling requirements and calculation time of a conflict probe. The general rule is a tactical conflict probe will have a relatively simple trajectory modeler but make frequent predictions. A strategic conflict probe will need a more complex trajectory modeler, since its predictions are much longer in look-ahead time resulting in a somewhat longer calculation time.

The following sections will not differentiate between tactical and strategic probes, referring to a conflict probe in general. The assumption is the accuracy methodology presented can be applied to both tactical and strategic conflict probes.

Accuracy Testing Methodology

Accuracy testing of a conflict probe is focused on three main areas of measurement:

- trajectory accuracy,
- conflict prediction accuracy,
- and conflict notification timeliness.

A conflict probe uses its predicted trajectories to determine future separation violations, i.e., to predict conflicts. Thus, the trajectory accuracy, or the deviation between the predicted trajectory and the actual path of the aircraft, has a direct effect on the accuracy of the conflict prediction. Conflict prediction accuracy is measured by several error probabilities that are used to quantify whether a predicted conflict actually occurred, and whether an actual conflict was predicted. The conflict predictions must not only be accurate in terms of the existence of a separation violation, but the conflict needs to be predicted in a timely manner.

Conflict notification timeliness quantifies the amount of lead-time the probe provides in the conflict predictions.

To apply these accuracy metrics, a set of input test scenarios is generated. The test scenarios are assembled to be representative of the air traffic that the conflict probe would confront in the field. The analysis could be performed with controllers in a simulation environment, but the methodology presented in this paper uses a recorded traffic scenario in real time without operators. The conflict probe alerts produced are matched with the actual conflicts in the scenarios. Statistical tests determine whether or not, within a certain confidence, the new release of the conflict probe performed as well as or better than the baseline system.

This paper focuses only on conflict prediction accuracy metrics for aircraft-to-aircraft conflicts but is applicable to airspace conflict events as well. It will describe the metrics, their rationale and method of determination, their application on an individual flight, and a statistical approach for regression testing with a scenario of many flights. A detailed description of trajectory accuracy is presented in References [3] and [4]. Conflict notification timeliness will be left for future publication.

Test Air Traffic Scenarios

As described earlier, test air traffic scenarios are generated as input into the new and baseline conflict probes. For accuracy testing it is important to cover all of the likely types of conflicts, while still providing realistic aircraft flight profiles.

Weather data and a recording of actual messages sent from the Host Computer System (HCS) to the conflict probe are made. The HCS messages include (1) the flight plans and their amendments of all the IFR (Instrument Flight Rule) aircraft, (2) any interim altitude clearances, and (3) the radar position and velocity reports for every aircraft. Since the air traffic controllers ensure the aircraft are separated, there are no aircraft-to-aircraft conflicts in the recorded scenario. Therefore, conflicts are induced by time shifting the individual flights in the recording [5].

The amount of traffic data used depends on the goals of the particular accuracy analysis. Typically

a traffic scenario is several hours in length with 1500 to 3000 flights and over 100 aircraft pair conflicts. Building the test scenarios and generating the proper mix of aircraft pair conflicts requires significant effort. The performance of a conflict probe (as measured by missed/false alert rates) is strongly influenced by the characteristic properties of the conflicts themselves. For example, it is relatively easy for a conflict probe to correctly detect an opposing (encounter angle near 180 degrees) collision conflict (zero distance at closest approach) between two cruising aircraft. Conversely, it is relatively difficult to correctly

detect a trailing (encounter angle near 0 degrees) grazing conflict (separation just below the minimum standard) between a climbing aircraft and a descending aircraft. Hence a conflict probe will perform poorly if evaluated with a traffic scenario that contains a large percentage of "difficult" conflicts. The process of generating these test scenarios with the proper mix of conflict properties is described in detail in [5] and [6]. The following accuracy analysis assumes that the input test scenarios are produced in this manner.
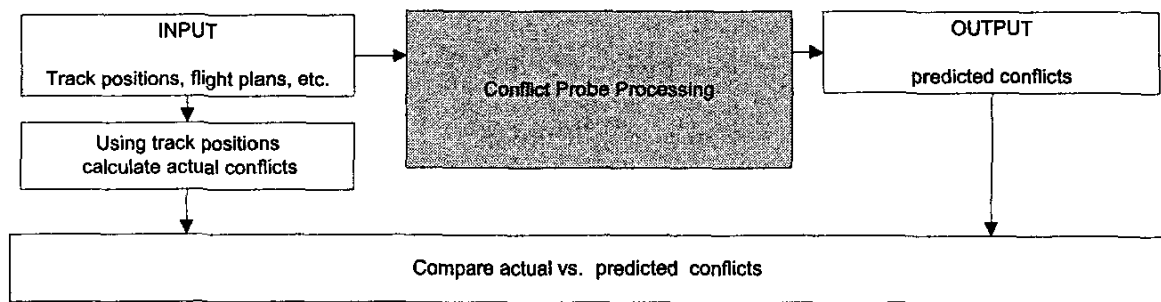
Figure 2. General Conflict Prediction Accuracy Processing [7]

Conflict Prediction Accuracy

The measurement of the accuracy of a conflict probe's predictions of aircraft-to-aircraft is referred to as conflict prediction accuracy. This is probably the most operationally significant metric category, since the major purpose of a conflict probe is to support the separation management of aircraft. Conflict prediction accuracy quantifies the fundamental error probabilities that are directly related to the probe's central goal: detecting conflicts.

In Figure 2, the conflict prediction accuracy metric isolates the conflict probe processing as a black box. Such an approach is only concerned with the input (i.e. the positions of the aircraft) and the output (i.e. predicted conflicts). A post-processing tool must first determine the actual conflicts using the aircraft position data, and then these conflicts are compared to the predicted conflicts.

Aircraft-to-Aircraft Conflicts and Encounters

Once the traffic scenarios are generated the HCS surveillance track positions are checked for reasonableness and processed for determination of aircraft pair conflicts and encounters. A conflict or encounter between two aircraft occurs when their separation drops below established minima [8]. In en route airspace, while operating under IFR, aircraft are required to be at least five nautical miles horizontally separated or vertically separated by at least 1000 feet up to and including Flight Level (FL) 290, and by 2000 feet above [9]. In this paper, the test scenario's time-shifted aircraft that violate these standard separations are considered in conflict.

It is also necessary to consider aircraft that approach each other but do not violate separation standards. In general, these include all the non-conflicting aircraft pairs that have some time overlap in the test scenario. However, for analysis

purposes encounters are often restricted by horizontal and vertical separations thresholds, such as less than 25 nautical miles horizontally and vertically less than 4000 feet up to and including FL 290 and 5000 feet above.

Software tools process the scenario and generate a relational database of the aircraft-to-aircraft conflicts and encounters. The fields consist of the aircraft pair's identification codes, start and end times, and other attributes of the conflict. Conflict attributes include horizontal and vertical minimum separations, vertical phase of flight (e.g. climb-climb, cruise-cruise) and pop-up category. These pop-up categories are used to excuse conflict predictions that are notified late.

Fundamentals in Evaluating Alerts

When the conflict probe predicts that a future conflict will occur between two aircraft, it posts an alert to the air traffic controller's display. The alert remains posted until the conflict is past or is no longer predicted. Usually the controller will redirect one of the aircraft so that the conflict will not occur. The probe automatically reads this change in flight path and deletes the alert.

The alert may be updated (in time and/or space), while it is posted to the controller's display. The initial posting of the alert and its final deletion form a notification set which can be matched to an actual conflict.

As documented in References [2, 10, 11, 12], the conflict probe is not perfect – it does make mistakes. For example, it can miss a conflict (Missed Alert) or it can predict a conflict that never occurs (False or Nuisance Alert). The four possible situations are shown in Table 1.

For a real time system, it is important that an alert be given sufficiently earlier in time of the actual conflict so corrective action can be taken. In other words, an alert must be timely as well as accurate. To ensure timeliness in conflict predictions, a conflict probe is often required to have some lead-time or actual warning time. This *Minimum Warning Time (MWT)* ranges from 1 to 5 minutes depending on the particular type of conflict probe being evaluated.

Table 1. CP Alert and Conflict Event Combinations [2, 12]

	CONFLICT OCCURS	CONFLICT DOES NOT OCCUR
ALERT	CP predicts conflict and it occurs (VA – valid alerts)	CP predicts conflict and it does not occur (FA – false alert)
NO ALERT	CP does not predict conflict and it occurs (MA – missed alert)	CP does not predict conflict and it does not occur (NC – correct no-calls)
Total Number of Alerts	Total Number of Conflicts	Total Number of Non-Conflicts (Encounters that did not have conflicts)

As summarized in Table 1, a notification set is evaluated as a Valid Alert when the conflict probe correctly predicts the conflict and when it is posted in a timely manner. If the notification set is not presented at all or correctly predicts the conflict but is not posted soon enough, it is called a Missed Alert. The lateness of the alert may be excused only if the conflict is considered a pop-up, which is defined in detail in the later Section *Definition of Pop-Up Conflicts*. A notification set determined to be a Missed Alert due to lateness is also referred to as a Late Missed Alert. A notification set presented late but excused is referred to as a Late Valid Alert.

A notification set that predicts a conflict when no conflict occurs is a False Alert. However, a False Alert withdrawn before the predicted conflict start time is also called a Retracted False Alert. A False Alert matched to an encounter not a conflict may be excused under certain circumstances.

Simply counting the number of times each of the events occur for a suitable mix of aircraft conflicts is not possible. It is necessary to match the alerts to the actual conflicts. There may be multiple conflicts between two aircraft. This occurs when the two aircraft are flying on close, nearly parallel paths and move in and out of conflict. Similarly there may be multiple alerts generated by the conflict probe for the same aircraft pair.

The test scenario and limitations of the conflict probe introduce additional complications. The scenario recording has a specific start time and end time. Alerts that span the start time or end time have to be treated as special cases. Also, the radar track data may be missing at the predicted conflict location. Adjustments are made for the inability of any conflict probe to predict future actions of

controllers. Therefore, what appears initially to be a simple and straightforward analysis, due to the many special cases and limitations of the test scenarios and the conflict probe, ends up being quite complicated.

Taking all these factors into account, the best way to present the methodology of measuring the conflict prediction accuracy is to describe the specific process used to quantify these error events. First it is necessary to provide some definitions of key concepts. In the next section, pop-up conflicts will be defined. In the subsequent two sections, the conflict prediction processing and the error probabilities will be described.

Definition of Pop-Up Conflicts

For a conflict prediction to be considered correct and labeled a Valid Alert, it must be presented to the controller at least a threshold number of minutes prior to the actual conflict start time. This threshold is the MWT defined earlier, which again ranges from 1 to 5 minutes. This conflict timeliness requirement for a Valid Alert is relaxed if the conflict is considered a pop-up. A pop-up conflict occurs if the probe is not provided with MWT threshold of continuous surveillance

data or prediction for either of the associated flights. There can be several reasons for a conflict being labeled a pop-up. Some examples include:

- The conflict starts within MWT of the start of either aircraft's HCS track. For example, this occurs when the conflict starts as one of the associated aircraft enters the scenario.

- The conflict starts within MWT of a recorded clearance message.

- The conflict starts within MWT from the time either aircraft exit an inhibited airspace not modeled for aircraft-to-aircraft conflicts. For strategic conflict probes, these airspace boundaries usually include terminal areas where separation rules differ from en route airspace.

- The conflict occurs when either aircraft is less than an adapted altitude (e.g. 300 feet) from a cleared interim or hold altitude at conflict start.

These situations allow relaxation of the Valid Alert conflict timeliness requirement, since under these conditions a conflict probe would not be expected to predict the conflict beyond the MWT threshold. However, regardless whether the conflict is a pop-up, a Valid Alert still needs to be posted prior to the actual conflict start time.
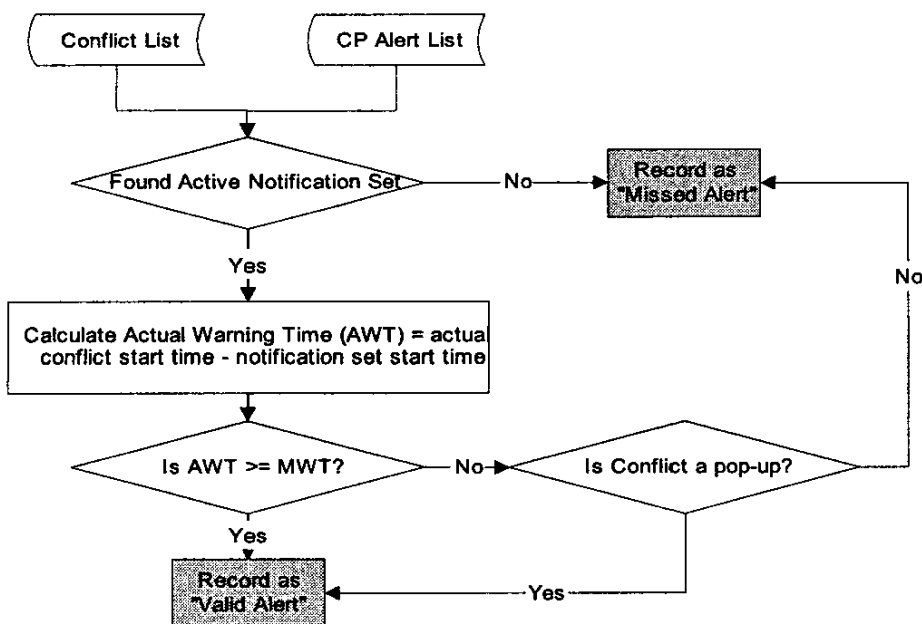

Figure 3. Process A – Valid and Missed Alert Processing

Methodology of Conflict Prediction Accuracy

The Missed, Valid, and False Alerts, as defined in Table 1, are determined in two sub-processes. In Process A (see Figure 3), conflicts are evaluated in order of actual conflict start time and matched against eligible notification sets. To even be eligible for matching to a specific conflict, a notification set must have a posting time prior to the start of the conflict and must have an end or delete time after the start of the actual conflict. Thus, the notification must precede the conflict and must be active at the start of the conflict. The result is listing of Valid Alerts and Missed Alerts associated to all the input conflicts provided by the test scenario.

The remaining notification sets not matched, as Valid Alerts, are potentially False Alerts. In Process B (see Figure 4), the remaining notification sets are evaluated to determine which of them are truly False Alerts and which can be discarded. Unlike the Missed Alerts, there are several reasons for discarding False Alerts. The potential False Alert is discarded if either aircraft does not have HCS track data present at the predicted conflict start time (PCST). With a lack of HCS track data, the False Alert error is unverifiable and thus excused. In many of these cases, the discarded notification sets represent alerts predicted beyond the end of the traffic scenario.

If the potential False Alert is retracted due to an air traffic control clearance, the notification set is discarded. The potential False Alert can also be discarded if the notification set was posted after the last actual conflict start time (ACST) between the associated aircraft. This can only happen if a conflict actually occurs between these aircraft and another alert is presented after it starts. When the conflict probe is operating in the NAS, once the actual conflict started, alert predictions would have little value and other more tactical procedures would be utilized. This is event is mainly an artifact of the test scenario resulting from the time-shifting process.
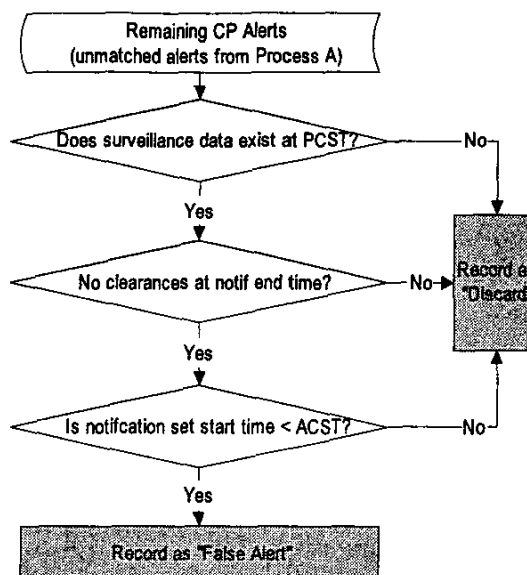

Figure 4. Process B – False Alert Processing[1]

Example Flight Analysis

A flight example, referred to in this paper as ABC100, was selected from a Memphis ARTCC (ZME) test scenario. This same flight was first presented in Reference [4] in December 2001 to illustrate how the trajectory prediction accuracy methodology is applied. The focus of this paper is on conflict prediction accuracy, so the analysis of this flight's actual and predicted conflicts are presented. The conflict probe used for this example is an FAA laboratory prototype available to the authors. Flight ABC100 is an over flight, entering the ZME airspace at Flight Level 350 (FL350), descending to FL310, and then exiting the ZME airspace at this altitude. The aircraft is cleared to descend to FL310 at 14:25:05 and the resulting Top Of Descent (TOD) time is at 14:25:10 (51910 seconds). For ZME and the analysis, the flight concludes at 14:48:00 (53280 seconds), when air traffic control of ABC100 is passed to the Fort Worth ARTCC (ZFW).

---

[1] PCST is the predicted conflict start time of the notification set and ACST is the actual conflict start time of the true conflict.

In this time shifted test scenario, as shown horizontally in Figure 5, the flight XYZ200 is cruising at FL310 crosses ABC100's route at an encounter angle of 38 degrees. As shown vertically in Figure 6, this crossing encounter occurs while ABC100 is descending to FL310 causing a test conflict with a minimum horizontal and vertical separation of 4.8 nautical miles and 1050 feet,

respectively. The conflict is rather short starting approximately at 14:27:10 and ending at 14:27:40. Once again, this aircraft-to-aircraft conflict is not real, but induced in the test scenario simply by time shifting the flights. However, the conflict probe tested with these flights is expected to predict the conflict as if it were real.
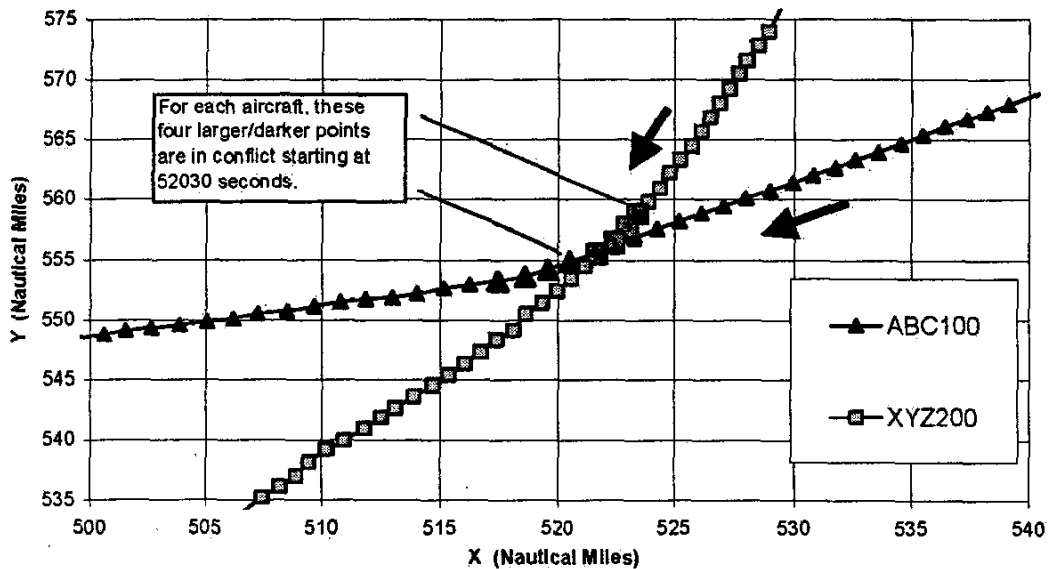

Figure 5. Horizontal Profile of ABC100 and XYZ200 Conflict Example
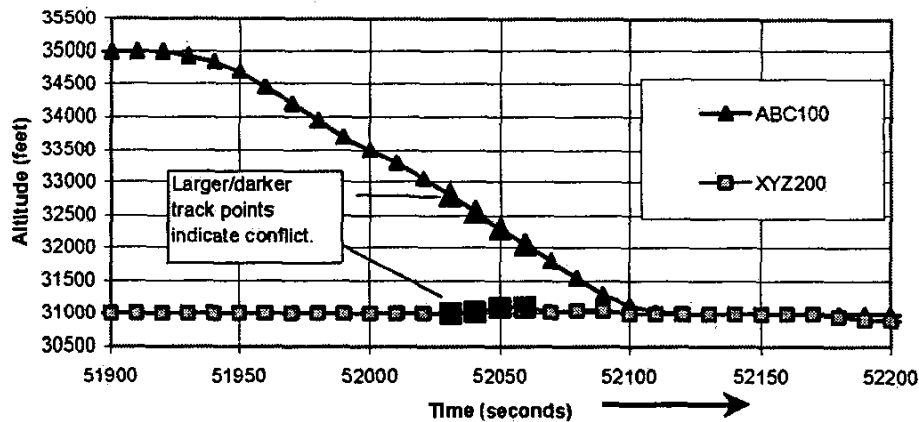

Figure 6. Vertical Profile of ABC100 and XYZ200 Conflict Example

As presented in the Table 2, the probe presents four notification sets, where the first three are all retracted before the conflict started. Notification Set A was presented at 14:10:25 but

was retracted only four seconds later at 14:10:29. This notification set was evaluated as a Retracted False Alert. Approximately a minute and a half later at 14:11:58 Notification Set B was presented

and again retracted, producing a second Retracted False Alert. At 14:20:57 yet another notification set, Notification Set C, was presented, but it was

almost immediately retracted at 14:21:00. This retraction was caused by a hold altitude clearance and consequently was discarded.

Table 2. Notification Sets for ABC100 and XYZ200 Conflict Example

Notification Set	Notification Start Time	Notification Set End Time	Predicted Conflict Start Time	Predicted Conflict End Time	Description
A	14:10:25	14:10:29	14:24:35	14:29:47	Retracted False Alert.
B	14:11:58	14:14:29	14:26:21	14:30:00	Retracted False Alert.
C	14:20:57	14:21:00	14:25:24	14:29:50	Retracted False Alert discarded due to clearance.
D	14:25:05	14:29:56	14:26:23	14:29:56	Valid Alert 2:05 before pop-up

Finally at 14:25:05 a fourth notification, Notification Set D, was presented and remained active until the conflict started two minutes in five seconds later at 14:27:10. Thus, this last notification set is a Valid Alert matched to the ABC100 and XYZ200 conflict. As discussed in the previous *Fundamentals in Evaluating Alerts* Section, the conflict probe is normally required to present an alert a threshold of warning time (MWT) before the conflict actually starts, however this is relaxed if the conflict is labeled a pop-up. For this example, the MWT was defined at 5 minutes. The ABC100 flight was cleared to descend at 14:25:05, started its descent five seconds later, and the conflict started roughly two minutes later at 32800 feet. The conflict started within the defined five minutes of a clearance labeling it a pop-up conflict.

In summary, for this aircraft pair, ABC100 and XYZ200, the laboratory conflict probe produced two False Alerts and one Valid Alert. It illustrates several of the conflict prediction accuracy rules. For the complete regression testing of a conflict probe, thousands of notifications sets would be evaluated. Even with a complete regression test, it is useful to examine several notification sets in this manner to understand the nature of the probe's errors as well as validate the analysis software.

Probability Definitions

The Missed and False Alert counts are normalized by dividing them by the number of conflicts and encounters they are matched to. The resulting ratios are the probability of Missed and False Alerts. Equation 1 defines the probability of Missed Alert. It quantifies the conditional probability that the conflict probe does not predict the conflict when it occurs.

$$P(MA) = \frac{MA}{C}$$

Equation 1

where MA is the number of Missed Alerts and C is the number of input conflicts from the scenario.

The False Alert probability is defined as the likelihood in predicting a conflict when it does not occur. This is defined in Equation 2. False Alert probabilities are partitioned by the minimum horizontal separation of their corresponding encounters. For example, the False Alert probability for one such bin is the probability of falsely predicting a conflict when the aircraft are actually in an encounter separated between 10 to 15 nautical miles.

$$P(FA_i) = \frac{FA_i}{E_i}$$

Equation 2

where is the index of the bin, FA^i_i is the number of False Alerts matched to encounters in a given bin of minimum horizontal separation and E^i_i is the total number of encounters for the same bin present in the input test scenario.

Regression Testing

The objective of regression testing is to verify that the new release of the conflict probe software is as accurate as the baseline system. In the context of conflict prediction accuracy, the comparison is to determine if more Missed and False Alerts are present in the new conflict probe relative to the baseline version. The common approach presented in Reference [13] to determine if this difference is statistically significance is to utilize a binomial distribution and perform a hypothesis test concerning the difference between population proportions. However, this technique assumes that the respective runs are independent. In this context, each conflict probes are not independent, since they are run with the same air traffic scenario and weather files.

An alternative technique is presented in References [14] and [15], utilizing categorical data analysis techniques. For categorical data analysis, we examine the difference in frequencies not proportions. For this study, the frequencies are the counts of Missed and False Alert events. Paired counts that are mutually exclusive and exhaustive, which is required for this test, occur when the error event occurs in one run and the correct event occurs in the other.

For the Missed Alert analysis, the count of interest is the quantity of Missed Alerts in the baseline conflict probe when simultaneously getting a Valid Alert in the new release conflict probe or vice versa for the opposite case. These counts should be equally likely if the two system's accuracy is statistically equivalent. Calculating the ratio of the squared difference between the expected value of each run and the observed value can test this hypothesis. If the hypothesis is true, this ratio will follow a chi-squared distribution with one degree of freedom.

The test statistic is defined as follows: χ^2

$$\chi^2 = \sum_{i=1}^{2} \frac{(O_i - E)^2}{E}$$

Equation 3

where

O_i is the observed frequency in category i

E is the expected frequency

Since the null hypothesis assumes both events are equally likely, both expected frequencies are equal and calculated from the following equation:

$$E = \frac{\sum_{j=1}^{2} O_j}{2}$$

Equation 4

An example will be presented in the next section that will more thoroughly describe the application of this technique.

Example Regression Test

Similar to the previous flight example a laboratory version of a prototype FAA conflict probe was input with a test scenario from Indianapolis ARTCC collected for two-hours on May 26, 1999. The scenario produced over 211 aircraft-to-aircraft conflicts and 1,715 encounters. This test scenario was input in a baseline system producing 37 Missed Alerts, 174 Valid Alerts, and 449 False Alerts. The same scenario was input into the conflict probe with inferior weather forecasts, altered with 60 knots of wind error. This second run is analogous to running a new release of the conflict probe but in this case to study the effect of weather forecast error. It resulted in 43 Missed Alerts, 168 Valid Alerts, and 522 False Alerts. Just as a regression test, the analysis needs to determine if the new release has statistically equivalent conflict prediction accuracy as the baseline run.

Table 3. Comparison of Missed and Valid Alerts

Baseline Run	New Release Run		
	Valid Alert	Missed Alert	Total
Valid Alert	162	12	174
Missed Alert	6	31	37
Total	168	43	211

From the Table 3, the baseline run produced 12 Valid Alerts that were missed in the new release run. Conversely, the new release run produced 6 Valid Alerts that were missed by the baseline system. There were 162 common Valid Alerts and

31 common Missed Alerts. Using Equation 4, the expected value of Valid-to-Missed and Missed-to-Valid Alerts is 9. The resulting test statistic from Equation 3 is 2 and can be expressed as a probability or P-value by assuming a chi-squared distribution with one degree of freedom. In [13], the P-value is defined as the smallest level of significance at which the null hypothesis would be rejected. The P-value is the probability that the null hypothesis has occurred, so a small P-value (less than 0.10) would indicate the null hypothesis unlikely and should be rejected. In this example the P-value is 0.16, so the test provides evidence that null hypothesis is approximately 16 percent likely and cannot be rejected.

The False Alert results are quite different. From Table 4, of the 1,715 encounters 167 were correctly not called by the baseline system but were falsely presented by the new release conflict probe. Conversely, the baseline generated 94 False Alerts when the new release correctly did not present alerts for these same encounters. Using Equations 3 and 4 analogous to the Missed Alert analysis, the resulting test statistic was 20.4 and P-value 0.00. Therefore, there is strong evidence to reject the null hypothesis that the runs are equivalent.

Table 4. Comparison of No-Call and False Alerts

Baseline Run	New Release Run		
	Correct No-Call	False Alert	Total
Correct No-Call	1099	167	1266
False Alert	94	355	449
Total	1193	522	1715

In this example, the weather forecast error caused a statistically significant effect to the conflict prediction accuracy resulting in 16 percent more Retracted False Alerts, yet had no significant impact on the Missed Alert error.

Summary

This paper presents conflict prediction accuracy metrics that can be applied generically to any conflict probe. The application of generating a time shifted air traffic scenario with induced aircraft-to-aircraft conflict and encounters, the two-stage process of evaluating Missed and False Alerts, and the calculation of the corresponding error probabilities are presented. A specific flight is presented which illustrates the application of the conflict accuracy measurement rules. A statistical test utilizing categorical statistical analysis is presented for regression testing two versions of the conflict probe that is input with the same scenario. The regression testing is a standard approach required to ensure a system does not degrade after deployment as upgrades are implemented with new functions and features. The approach is demonstrated on an entire scenario of many flights where a conflict probe is shown to have significantly more False Alerts.

Additional considerations for identifying Missed and False Alerts generated from poor flight intent, a major source of error for a conflict probe, metrics for conflict prediction timeliness, and partitioning the errors, as a function look-ahead time, will be explored in future publications.

References

[1] Federal Aviation Administration, May 2003, "FAA Flight Plan 2004-2008," www.faa.gov.

[2] Paglione, M., M. Cale, H. Ryan, Fall 1999, "Generic Metrics for the Estimation of the Conflict Prediction Accuracy of Aircraft to Aircraft Conflicts by a Strategic Conflict Probe Tool," *Air Traffic Control Quarterly*, Vol. 7 (3).

[3] Paglione, M., H. F. Ryan, R. D. Oaks, J. S. Summerill, M. L. Cale, May 1999, "Trajectory Prediction Accuracy Report User Request Evaluation Tool (URET)/Center-TRACON Automation System (CTAS)," DOT/FAA/CT-TN99/10, FAA WJHTC/ACT-250.

[4] Paglione, M., R. Oaks, M. L. Cale, S. Liu, H. Ryan, J. S. Summerill, December 3, 2001, "A Generic Sampling Technique for Measuring Aircraft Trajectory Prediction Accuracy," 4th USA/EUROPE Air Traffic Management R&D Seminar.

[5] Paglione, Mike M., Robert D. Oaks, J. Scott Summerill, August 2003, "Time Shifting Air Traffic Data for Quantitative Evaluation of a

Conflict probe," American Institute of Aeronautics and Astronautics (AIAA) Guidance, Navigation, and Control Conference, Austin TX.

[6] Paglione, Mike M., Robert D. Oaks, Karl D. Bilimoria, November 2003, "Methodology for Generating Conflict Scenarios by Time Shifting Recorded Traffic Data," American Institute of Aeronautics and Astronautics (AIAA) Aviation Technology, Integration, and Operations (ATIO) Technical Forum, Denver, CO.

[7] Estrella, John A., 2004, "Course Notes - Course 316: Software Testing and Inspection Methods," Learning Tree International, pp. 316-1-25.

[8] Bilimoria, K. D., H. Q. Lee, August, 2001, "Properties of Air Traffic Conflicts for Free and Structured Routing," paper presented at the AIAA Guidance, Navigation, and Control Conference, Montreal, Canada.

[9] United States Department of Transportation, Federal Aviation Administration, February 24, 2000, *Air Traffic Control 7110.65M*, FAA.

[10] Bilimoria, Karl, May-June 2001, "A Methodology for the Performance Evaluation of a Conflict Probe," *Journal of Guidance, Control, and Dynamics*, Vol. 24 (3).

[11] Brudnicki, D., W. Arthur, K. Lindsay, April 1998, "URET Scenario-based Functional Performance Requirements Document," MTR98W0000044, MITRE/CAASD.

[12] Cale, M. L., M. Paglione, H. Ryan, D. Timoteo, R. Oaks, April 1998, "URET Conflict Prediction Accuracy Report," DOT/FAA/CT-TN98/8, FAA WJHTC/ACT-250.

[13] Devore, Jay L., 2000, *Probability and Statistics for Engineering and the Sciences, Fifth Edition*, Pacific Grove, CA, Duxbury.

[14] Agresti, Alan, 1996, *An Introduction to Categorical Data Analysis*, New York, New York, John Wiley and Sons, Inc.

[15] Kachigan, Sam Kash, 1986, *Statistical Analysis: An Interdisciplinary Introduction to Univariate & Multivariate Methods*, New York, New York, Radius Press.