AIAA Guidance, Navigation, and Control Conference
10 - 13 August 2009, Chicago, Illinois

AIAA 2009-6079

Effectiveness of Pairing Flights When Evaluating The Accuracy of a Conflict Probe

Mike M. Paglione*
Federal Aviation Administration, William J. Hughes Technical Center, Atlantic City, NJ 08405

Robert D. Oaks†
General Dynamics Information Technology, William J. Hughes Technical Center, Atlantic City, NJ 08405

I. Introduction

Despite the current economic slow down, most air traffic service providers (ATSPs) across the globe continue to expect significant growth in air traffic demand in the future. If no action is taken, it is generally accepted that this growth will outpace the capacity limits of their aviation systems, resulting in greater congestion and inefficiency. In areas of the northeastern United States as well as Western Europe, these conditions may already have reached their capacity limits under peak demand. In unprecedented proportions, industry and ATSPs have responded by developing comprehensive plans requiring broad advances in ground-based and airborne automation.

In the United States, the interagency Joint Development Planning Office (JPDO) foresees a traffic demand increase by 2025 up to three times the number of flights of today's traffic[1]. The JDPO, as established in their charter under the "Vision-100" legislation (Public Law 108-176) signed by President G. W. Bush in December 2003, has mandated a next generation operational concept of the National Airspace System (NAS) for 2025[1]. This next generation NAS envisions a trajectory-based separation management system that requires precise management of the aircraft's current and future position. The separation function of today, relying heavily on the cognitive skills of the air traffic controller to visualize aircraft trajectories on the radar display and issue resolutions via voice instructions to pilots, will be replaced by a distributed system of separation management components, implementing performance-based separation standards. This future system will rely heavily on enhanced automation with conflict resolutions that are communicated digitally between air and ground and between aircraft.

A key automation component promoted in the JPDO's operational concept is the development of decision support tools (DSTs). These tools are envisioned to help mitigate many of the capacity and workload constraints of the system if effectively integrated with advanced automation solutions in the air and ground systems. These tools have many purposes and typically serve to reduce the cognitive workload of the airspace problems faced by the current human decision makers operating the system. They include tools that serve to predict future conflicts between aircraft, both for ground based controllers or airborne pilots, allowing more strategic separation management of aircraft. Air traffic management DSTs include capabilities that forecast where and when traffic workload would stress the system, allowing air traffic supervisors to make more efficient adjustments to either avoid the condition or alter staff and/or airspace accordingly. Such tools also include air traffic metering tools to efficiently sequence aircraft into en route and arrival flows, maximizing the capacity of the system. A common thread in all these DSTs is the accurate and timely modeling of the aircraft's current state and anticipated future path. This modeling function is referred to as the trajectory predictor (TP) process.

A. Relationship of Trajectory Predictor to Higher Level Applications

The aircraft trajectory is the actual or future 4-dimensional path of the aircraft. TP accuracy can be measured by post flight comparisons of predicted and observed aircraft trajectories. Since the predicted trajectory is the fundamental input that sustains the DST's capabilities and functions, the accuracy of the trajectory prediction has a direct impact on the DST's overall performance and usability. In order to attain the specified accuracy requirements of a DST, it is necessary to validate the DST's TP. Ref. 2 presents a TP validation methodology that can drive the performance of a TP toward a targeted level. Ref. 3 defines system metrics used within this methodology and shows how these metrics can assess a TP's impact on a DST. Figure 1 summarizes the process graphically by illustrating the connection a TP has within a client application and how the TP's output metrics are then input for the application. For example, the time or longitudinal error associated with a TP's aircraft trajectory predictions will have a direct impact on the stability of a time-ordered schedule output from a metering DST. If a flight actually arrives significantly later than predicted, the DST's estimated time of arrival and associated order in the metering list

* Project Lead, Conflict Probe Assessment Team, Simulation and Analysis Group, Senior AIAA Member.
† Senior Computer Scientist, Conflict Probe Assessment Team, Simulation and Analysis Group, AIAA Member.

This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

will be have to change. If the changes are frequent and sufficiently large, the utility of the generated schedule and the entire metering function will come into question.

The first step in improving the TP process within a DST is identifying the factors that influence performance. The late professor and modern management theory pioneer, Peter F. Drucker, is attributed with the famous quote, "If you can't measure it, you can't manage it." Consistent with these words, this paper will integrate standard statistical techniques, the data-driven TP validation methodology presented in Ref. 4, and the system engineering approach in Ref. 3. Through several descriptive examples, it will provide detailed guidance on how to make statistical inferences on the TP's prediction accuracy and resulting impact on their higher level application's performance using actual recorded aircraft information.

B. Experimental Design's Blocking Principal

Experiments are performed by most researchers and scientists in practically all disciplines. An experiment is defined as "a test or series of tests in which purposeful changes are made to input variables of a process or system so that we may observe and identify the reasons for changes in the output response."[5] To illustrate this further, Fig. 2 presents the general model of a process under study as adapted from Ref. 5. An input stimulus is entered into a process with a set of controllable factors. These are the factors or independent variables in the experiment that are manipulated to study the output or response variables. The uncontrollable factors are not easily manipulated, but through experimental design techniques such as blocking and randomization can be removed from the experiment. The output response variables are the dependent variables of the experiment. They are often determined by application of a metric or measured by a sensor device.

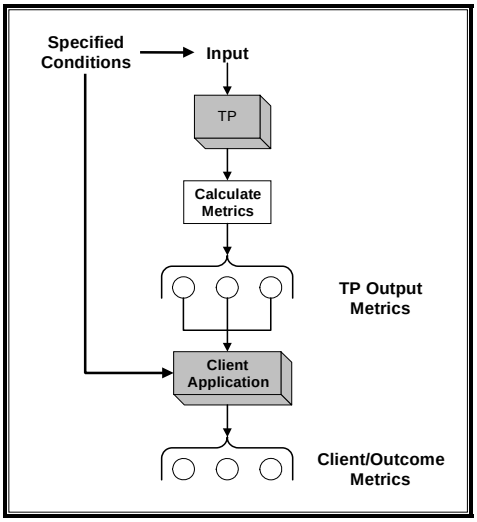

Figure 1. Illustration of Application of Trajectory Prediction Metrics to Higher-Level Applications[‡]

For this study, the primary experiment presented is to determine if the DST's accuracy or TP's accuracy within the DST has degraded or improved following the release of a new version of the system. It is not sufficient to simply measure the overall performance and conclude the system is acceptable. It is necessary to isolate the sources of any differences and at minimum determine the suspected factors involved. However, variability of between measurement units affected by the uncontrollable/extraneous factors can make it very difficult to identify the errors precisely. Randomization requires using many samples picked in an unbiased way from the population under study. This technique will systematically balance the extraneous factor's effects on the experiment.

Blocking is another experimenter technique that serves the same purpose but in a different way. It requires careful analysis of the context of the experiment and sources of these extraneous factors. A block, as defined in Ref. 6, is a portion of experimental material (e.g. two seeds in the same pot, four tires on a single automobile, and two shoes on a single student) that is expected to be more homogenous than the rest of the sample population. For aviation data, the central block under study is typically a single aircraft flight. Thus, if TP accuracy measurements are the sample units being taken (i.e. the response/dependent variable), a factor within the flight represents the focus of the study (i.e. the independent variable). For example, the within flight factor could be the two different runs of the system or change made in the TP algorithm. Therefore, by focusing on the effects of factors that are within the block, in this case within flight, all the variability between flights that are composed of many uncontrollable factors are effectively removed from the experiment (e.g. aircraft model type, origin-destination path, aircraft equipage, etc.). Overall, randomization should be used for time order decisions or any selection of the experimental units of the study, including the blocks themselves. Blocking is needed to remove unwanted uncontrollable factors by aggregating alike experimental units and concentrating their variation between them. As Ref. 6 recommends, "block what you can and randomize what you cannot."

[‡] Adapted from Figure 8 in Ref. 3.

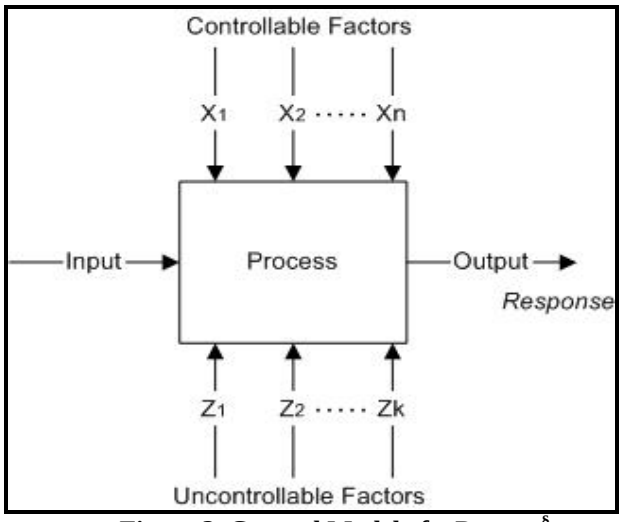

Figure 2. General Model of a Process[§]

II. Experiments with Blocks of Size Two – Pairing

A special case of an experiment with a block that contains only two levels is typically called a Paired Data t-Test or Paired t-Test for the Student t distribution used in its evaluation.[6,7] As described in the previous section, a typical application of such a test is the evaluation of a new release of the system compared to the current operational system. When dealing with systems that perform safety functions any change to the operational system must be sufficiently tested even if the change was intended to fix an existing problem. For software systems, this evaluation is called a regression test. It serves to provide confidence and thus reduce the risk associated in making the upgrade to the system. This section will present details of these tests, the statistical approaches and considerations with a working example from real field data recorded from DSTs operated by the Federal Aviation Administration (FAA).

A. Standard Two Sample Mean t-Test

Before presenting the Paired t-Test with a block, the basic Two Sample t-Test will be presented. It is the standard statistical method used in practice and will serve to highlight the merits of the paired approach. For example, in the en route airspace in the United States the air traffic control system is required to predict the future path of an aircraft based on various inputs including radar surveillance data and submitted flight plans. To evaluate its accuracy, a sampling methodology and set of metrics have been documented in detail in Ref. 4. One such metric, called the cross track error, measures the side-to-side or lateral error associated with the future prediction of the aircraft's path. The Two Sample t-Test compares the sample means of the two independent population means.[7] For this example, the objective is to determine if the cross track error's sample mean taken from the new system is greater than the existing operational system. If this is true, the new system cannot be deployed and further evaluation is warranted.

The statistical hypothesis test is designed to allow the experimenter to balance the two possible testing errors: failing the null hypothesis (Type I error or α) when it is true and failing to reject the null hypothesis when it false (Type II error or β).

$$H_o : \mu_b - \mu_n = 0 \tag{1}$$

where μ_b is the population mean of the baseline run of the operational system and μ_n is the population mean of the new release run.

The alternative hypothesis is the difference in population means is not equal to zero. Thus, the difference is less than zero (degraded) or greater than zero (has improved). The following test statistic is presented in Ref. 7.

[§] Adapted from Fig. 1-1 in Ref. 5.

$$\textit{Test}\text{ statistic: } t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_b^2}{m} + \frac{s_n^2}{n}}} \tag{2}$$

where $\bar{x}$ is the sample mean of the baseline run and $\bar{y}$ is the sample mean of the new release run, s_b^2 is the sample variance of the baseline run and m is the sample size of the baseline run, and s_n^2 and n are the same for the new release run, respectively.

The rejection region of the Two-Sample *t* Test is expressed in the following Eq. (3):

$$\text{Reject null hypothesis if } \quad t \le -t_{\alpha/2,\upsilon} \text{ or } t \ge t_{\alpha/2,\upsilon} \tag{3}$$

where $t_{\alpha/2,\upsilon}$ are parameters taken from the student-t distribution, α is the significance level of the test, and υ is the degrees of freedom for this test. The test assumes the trajectory measurements from each run are normally distributed random variables, and the runs are independent from one another. We will explore both these assumptions in more detail in the subsequent sub-sections of this paper.

The υ, degrees of freedom, can be estimated from the sample data by the following Eq. (4) as defined in Ref. 7.

$$\upsilon = \frac{\left(\frac{s_1^2}{m} + \frac{s_2^2}{n}\right)^2}{\frac{\left(s_1^2/m\right)^2}{m-1} + \frac{\left(s_2^2/n\right)^2}{n-1}} \tag{4}$$

Figure 3 provides a graphic summary of the relationship of the sample size to several key elements of the test. Smaller effect sizes are harder for the test to detect and in turn require larger sample sizes to identify. The smaller the Type I error or α chosen for the test will require larger sample sizes as well. The higher the power or smaller the Type II error (β) also requires larger sample sizes to achieve. As the variability of the sampled population increases for a test, the larger the sample size is required. As the test is implemented, the α, β, and effect size should specified considering the expected variability and sample size.

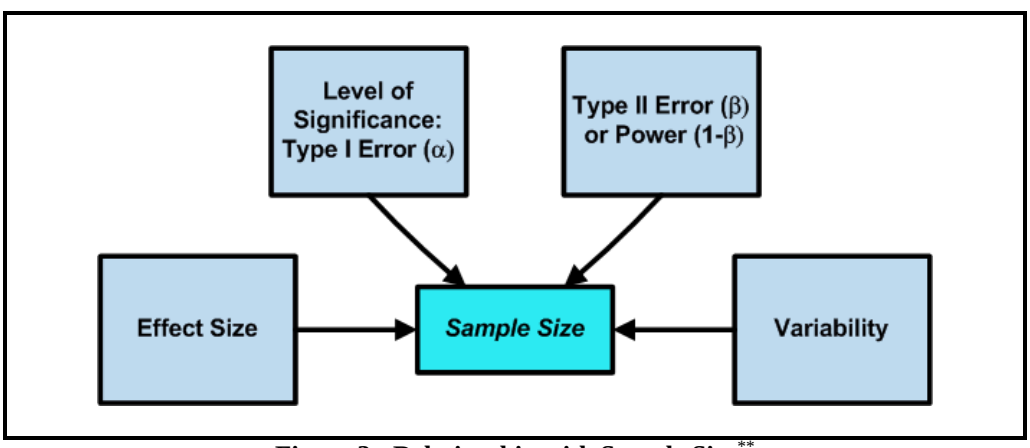

Figure 3: Relationship with Sample Size[**]

Continuing with the example discussed previously, a data collection from two trajectory predictor systems using a common set of input was planned with an objective of collecting over 800 samples per system. A recent data collection from these two trajectory predictor systems produced a sample of 118 flights of mean cross track errors per system. The test implemented requires an α of 0.05 and power of at least 0.90 to detect a shift of 0.1 nautical

[**]Adapted from slide 15 on page 1-13 in Ref. 8.

miles. From previous tests, the expected variability (standard deviation per system) is 0.6 nautical miles. The sample of the baseline or legacy system had an overall mean cross track error of 0.74 nautical miles and standard deviation of 0.64 nautical miles. The new system had an overall mean cross track error of 0.64 nautical miles and standard deviation of 0.57 nautical miles. By applying Eq. (2), the test statistic is 1.209. Application of Eq. (3) with degrees of freedom from Eq. (4) specifies a rejection criterion of ± 1.97. Thus, there is not sufficient statistical evidence to reject the hypothesis that the means are different ($1.209 < 1.97$ and p-value 0.113[††]). However, other considerations should be examined. A priori several estimates were proposed and once the test was executed differences were noted as listed in Table 1. First, it was expected that over 1600 samples (800 per run) were to be collected, yet only 236 actually resulted. The standard deviation of 0.6 nautical miles and mean shift to detect of 0.1 nautical were reasonably close to what was sampled. Thus, the difference in sample size had a considerable impact on the power of the final test. The lower sample size resulted in a test with a retrospective power of only 0.25, when the expectation was 0.90. This means out of 100 tests only 25 can be expected to detect a real difference in sample means of 0.1 nautical miles. Therefore, the immediate conclusion would be to repeat the test and ensure a larger sample size is achieved on the order of the original expectation of 1600 samples. The subsequent sections will examine other aspects of this data sample and may offer a better alternative.

Table 1: Prospective versus Retrospective Results of the Test

Parameter	Prospective	Retrospective
Type I error (α)	0.05	0.05
Standard deviation (nm)	0.6	0.64, 0.57
Mean of baseline system (nm)	0.7	0.74
Mean of new system (nm)	0.6	0.64
Total sample size from both runs	1600	236
Power (1 - β)	0.91	0.25

B. Evaluating the Test Assumptions

As previously stated, there are two major test assumptions required for the Two Sample t-Test. The test assumes the trajectory measurements from each run are normally distributed random variables, and the runs are independent from one another. We will examine our representative example in more detail to test these assumptions. First a normal distribution is a symmetric distribution around a mean that is not skewed. For the example, the mean cross track error taken for each flight was unsigned to capture the magnitude of the error and compare the two system's performance. As a result, the distributions for both systems are highly skewed as illustrated in Figures 4 and 5.

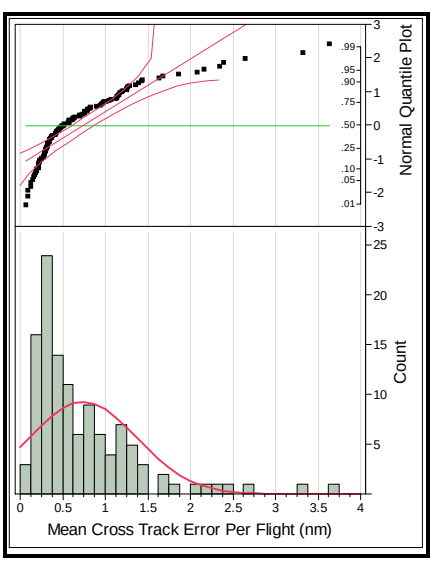

Figure 4: Base Run – Histogram/ Normal Quantile Plot

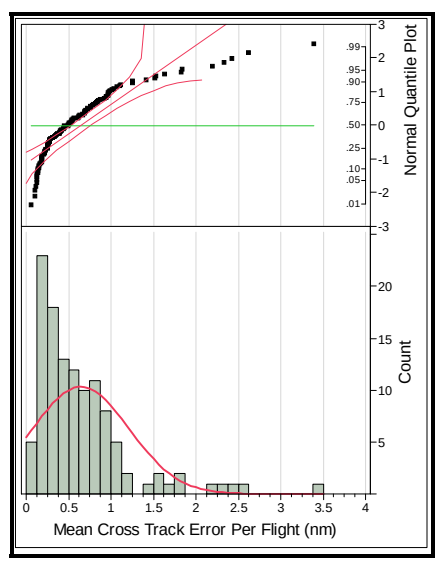

Figure 5: New Run – Histogram/ Normal Quantile Plot

[††]Ref. 7 defines the p-value as the smallest level of significance at which the null hypothesis would be rejected. If the p-value is small and less than the required α value, the null hypothesis should be rejected.

Figure 4 illustrates the baseline run's histogram in the bottom half of the figure and a normal quantile plot in the upper half. From the histogram, the smooth red line is overlaid on the graph, representing an estimate of normal distribution curve for the sample. In contrast the histogram is skewed to the left near zero and produces a long positive tail. The normal quantile plot provides a graphical diagnostic for testing if a data sample matches a normal distribution. In the top portion of Fig. 4, the dotted points are the sample means of the 118 flight samples and plotted on a graph matched to the estimated normal distribution. If the data matches a normal distribution, the data points would fall along the red diagonal line within the dotted red lines. It is clear from the normal quantile plot that the baseline data sample is not normally distributed. The same conclusions can be drawn for the new system from Fig. 5. Therefore, both system's distributions of cross track error are not normally distributed.

To apply the test, a possible solution is to transform the data into a normal distribution, then apply the test to the transformed (presumably normally distributed) data set. Since the data is highly positively skewed, a natural log may provide the solution. Thus, it was applied on all 236 measured data points. The histograms and normal quantile plots are repeated and illustrated in Fig. 6 and 7 for the baseline and new runs, respectively. In both cases, the log function has provided an excellent transformation producing a reasonable match to the normal distribution.

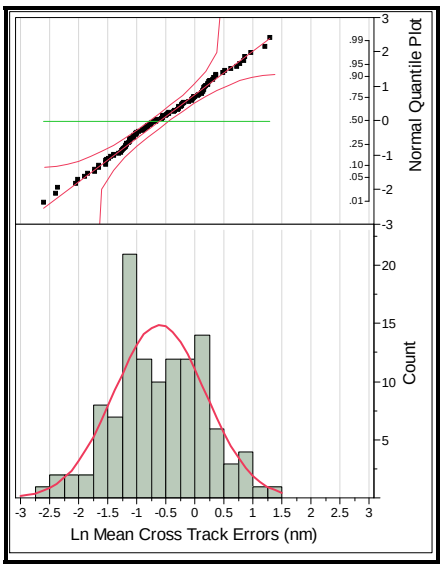

Figure 6: Base Run – Histogram/ Normal Quantile Plot

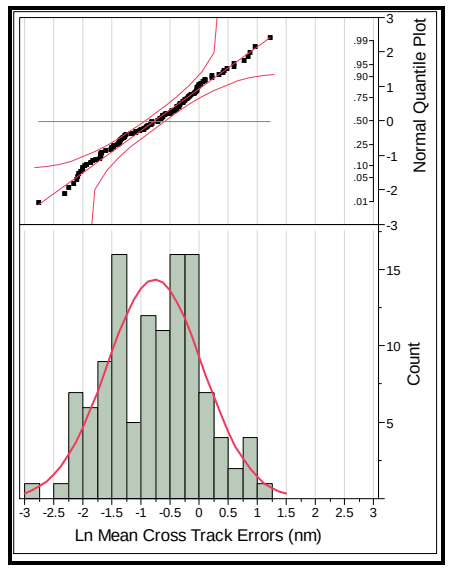

Figure 7: New Run – Histogram/ Normal Quantile Plot

All the calculated statistics also have to be recalculated with the transformed data set. By applying Eq. (2) again, the test statistic is now 1.510. Application of Eq. (3) and (4) remains the same and specifies a rejection criterion of -1.97 and 1.97, respectively. Still as with untransformed data set, there is not sufficient statistical evidence to reject the hypothesis that the means are different (1.510 is less than 1.97, and p-value 0.066). Table 2 also illustrates the prospective and retrospective results of the power of the test and indicates the test still lacks sufficient power to detect the required mean differences. It is only slightly improved with a power of 0.30 over the untransformed 0.25.

Table 2: Prospective versus Retrospective Results of the Test of Transformed Data

Parameter	Prospective	Retrospective
Type I error (α)	0.05	0.05
Standard deviation (nm)	0.8	0.79, 0.82
Mean of baseline system (nm)	-0.35	-0.61
Mean of new system (nm)	-0.50	-0.77
Total sample size from both runs	1600	236
Power (1 - β)	0.96	0.30

The violation of the normality assumption was mitigated by the transformation function, but the power of the test is still low and the second assumption has not been addressed. The second assumption assumed that the two

samples and therefore runs are independent. If you recall, both systems were input with the same set of flights. If a positive correlation between each run exists, the assumption of independence is violated. Discussions with our subject matter experts reveal that many uncontrollable factors could impact the cross track trajectory error metric. Both systems would be affected by these factors and they are different per flight. These factors manifest themselves as large cross track errors and it is likely that both systems will exhibit similar effects. For our example dataset, this is confirmed by the correlation coefficient of 0.87 calculated by comparing like flights from both runs. This indicates that there is an 87 out of 100 percent chance that if the baseline system had a linear increase in cross track error so would the new release run. It also requires a test that does not require the independence assumption. Fortunately, the Paired Data t-Test has this property. Furthermore, it actually performs better (increased power) if the two samples are highly correlated as will be illustrated in the next sub-section utilizing our example.

C. Application of the Paired Data t-Test

Instead of taking the difference between the sample means as applied in the Two Sample t-Test for the Paired Data t-Test, the sample measurements are paired for the same flight. The large variability between flights and linear dependence between runs is effectively blocked out of the experiment. Taking the difference between paired trajectory measurements of same flight from the two runs produces a new statistic, the sample differences. This is expressed in the following Eq. (5).

$$D_i = x_i - y_i \tag{5}$$

where i is the particular flight mean matched for the two runs, x_i is the mean trajectory measurement for the baseline run and y_i is the same for the new release run.

The hypothesis now is reduced to a single sample mean of D_i's assumed to equal zero on average. In reality, the mean of the difference between two numbers is equal to the difference between the means of the same set of numbers. Referring to the difference between sample means in Eq. (2), the $\bar{x} - \bar{y}$ is equal to the sample mean of D_i's or $\bar{d}$. The Paired t-Test's test statistic is illustrated in the following Eq. (6). The numerator in both Eq. (2) and now in Eq. (6) are equivalent, but the denominators are not the same at all.

$$\textit{Test}\text{ statistic: } t = \frac{\bar{d}}{s_D / \sqrt{n}} \tag{6}$$

where the s_D is the sample standard deviation of the differences (i.e. the D_i's) and the n is the sample size of these differences. Note, the n sample size for the Paired t-Test is half the total sample size applied in the Two Sample t-Test.

The rejection region of the Paired t-Test is expressed in the following Eq. (7).

$$\text{Reject null hypothesis if } \quad t \le -t_{\alpha/2,n-1} \text{ or } t \ge t_{\alpha/2,n-1} \tag{7}$$

Back to the example of two runs and their flight's mean cross track errors, the paired $\bar{d}$ and s_D are 0.095 nautical miles and 0.317 nautical miles, respectively. Application to Eq. (6) produces a test statistic of 3.26 and Eq. (7) produces a rejection region of values of (-2.27, 2.27). Thus, unlike the previous Two Sample t-Tests, the Paired t-Test does reject the hypothesis (with same exact difference measured) and has a very small p-value of 0.0015.

Table 3: Prospective versus Retrospective Results of Paired Data t-Test

Parameter	Prospective	Retrospective
Type I error (α)	0.05	0.05
Standard deviation (nm)	0.3	0.317
Mean of baseline system (nm)	0.7	0.74
Mean of new system (nm)	0.6	0.64
Total sample size from both runs	800	118
Power (1 - β)	1.0	0.92

The prospective versus retrospective power analysis of the Paired t-Test produces much better results. The retrospective power of the Paired Data t-Test is a very acceptable 0.92. Again this was due to blocking on the

extraneous factors in the experiment and pairing the flight means between runs. It did not require more samples only a different approach in the analysis. The property of the Paired Data t-Test to have much higher precision and thus power to detect a difference in the sample means of the test is explained in Eq. (8). It is derived in Ref. 6 and 7, but simply illustrated in this paper to explain precision improvement by pairing the data. Since the correlation coefficient term, ρ, in Eq. (8) reduces the overall variance of differences, a highly positively correlated set of runs will have a significantly reduced variance. The variance is the denominator in the test statistic in Eq. (6) and the smaller its value increases the power of the test.

$$V(\overline{X} - \overline{Y}) = \frac{V(D_i)}{n} = \frac{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}{n} \tag{8}$$

where $V(\overline{X} - \overline{Y})$ is the variance of the difference in means of the two samples, $\frac{V(D_i)}{n}$ is the sample variance of the paired differences, σ_1^2 and σ_2^2 are the variances per sample mean for each run, and ρ is the correlation coefficient between paired differences (it measures the direct linear relationship between two variables).

There is a trade off between reduction in sample size and lowering effect of the variance. By pairing the data, half of the flight samples are pooled and it becomes a test of "n" versus "2n" from the previous test. Thus, if the two runs and associated flights were truly uncorrelated (small ρ), then the Two Sample t-Test would be preferable. However, for analysis of trajectory data where two systems are input with the same flights, the samples are typically highly correlated and thus the Paired Data t-Test is preferred.

D. Application of Non-Parametric Paired Hypothesis Test

Even though the working example had sufficient power by using the Pair Data t-Test correctly, the result in the previous example is interesting, since the difference in sample means was only 0.1 nautical miles. Further inspection of the data showed that about ten measurements of the 118 total were more than 2 to 3 standard deviations larger than the sample mean of the differences. Removal of these ten outliers produced even better results with a test statistic of only 4.61, well below the rejection criterion. The power to detect a 0.1 difference had improved even further to 0.997.

Devore in Ref. 7 offers some insight into why the Pair data t-test was so sensitive to the outliers in the example. The underlying student-t distribution used in the test statistic is approximately normally distributed with large sample sizes, which is often the case with trajectory accuracy measurements. Normally distributed parametric tests can perform poorly when the underlying distribution has heavy tails. These tests depend on sample mean that can be very unstable in the presence of heavy tails caused by outliers. Alternative non-parametric approaches relax the assumption of normality and rely on a more robust metric, the sample median of the observed values.

If the null hypothesis is true, both the baseline and new release will have equally likely positive and negative measurements. Thus, it can be assumed that the sample differences of trajectory accuracy from the baseline and new release measurements are symmetric around a point of symmetry, namely the median. For both sets of measurements to be equally likely, the null hypothesis has a median equal to zero. A procedure is presented in Ref. 7 that provides a non-parametric technique to test the median and requires only that the distribution of differences is continuous and symmetric. This procedure is called the Wilcoxon Signed-Rank Test. To perform this procedure, the signed rank sum is calculated, which includes the following:

1. First, the absolute values of the trajectory accuracy differences are calculated and ranked in ascending order.
2. Next, the ranks of the positive measurements are summed, referred to as S+.

The S+ statistic is a random variable that can be calculated exactly if the sample size is small. It is approximates a normal distribution if the number of samples is greater than twenty. For trajectory accuracy measurements, the samples are often in the hundreds. The test statistic for a S+ calculated from a large sample is expressed in the following Eq. (9).

$$Z = \frac{S_+ - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}} \tag{9}$$

where S_+ is the ranked sum defined above and n is the sample size of trajectory differences.

The rejection region of the Wilcoxon Signed-Rank Test with a large sample size is expressed in the following Eq. (10):

$$\text{Rejection the null hypothesis: } Z \geq z_{\alpha/2} \text{ or } Z \leq -z_{\alpha/2} \tag{10}$$

where $z_{\alpha/2}$ or $-z_{\alpha/2}$ are parameters taken from the normal distribution, and α is the significance level of the test.

For the working example of 118 flights paired between the two system runs, from Eq. (9) the Wilcoxon Signed-Rank test statistic Z is 1396.5 and extremely significant with a p-value of 0.0001.

The key benefit to the pairing the data and using the Wilcoxon Signed-Rank test as the statistical test is both the independence and normality assumptions are relaxed. Ref. 7 indicates that not much is lost in terms of detection power for using this test when the underlying distribution is normally distributed and it is significantly improved when it deviates. Ref. 7 also reports on a metric called the asymptotic relative efficiency (ARE). It is the limiting ratio of sample sizes necessary to obtain identical error probabilities for the two tests. For large sample size, the ARE for the Wilcoxon test is approximately 0.95 if the underlying distribution is normal. This means that only about five percent more samples are required to get the same performance as the Paired t-Test. Ref. 7 also states that if the underlying distribution is any distribution, the Wilcoxon test will be at least 0.86 but is often larger than 1 (more efficient than a test knowing the underlying distribution). Therefore, the additional complexity in calculating the Wilcoxon Signed-Rank test statistic is outweighed by the robustness it provides when the underlying distributions departure significantly from the normal distribution. Most off-the-shelf software packages calculate both tests with little effort for the analyst, so test complexity is not a significant concern.

III. Implementation of Paired Hypothesis Test in Practice

The application of the statistical analysis approach presented in Section II of this paper provides guidance on how to organize your analysis and data. It is useful to put the analysis approach in context of the study objectives. The previous section assumed the main objective was comparative in nature. One system was being compared to the performance of another. For the example, the legacy system was being tested against a new release. The strength of the approach was the common input data (i.e. flight data) being provided to both systems and the pairing process on the output measurements. The method was an example of an experimental design approach called blocking as described in Section I.A. and its strength rests in the variance reduction property in Eq. (8). The method effectively removes extraneous or nuisance factors from the experiment. The method can be used for a larger set of test objectives.

A critical objective of evaluation of aircraft trajectory predictions is not only knowledge on the performance of the TP, but the decision support tool the TP supports. Fig. 1 illustrates this graphically by drawing a linkage between the TP performance and the broader client functions. The primary example is the relationship of aircraft-to-aircraft conflict prediction to the trajectory predictions that drive them. If the forecasted future path of the aircraft provided by the TP is perfectly realized for a pair of aircraft that happen to come close together, then the conflict predicted should be more accurate than another conflict prediction of the same event with some level of TP error. Thus, the paired trajectory differences as measured in Eq. (5) can be correlated to differences in the conflict predictions that are realized by the same aircraft flights involved. This method could be applied to other client services such as metering, capacity forecasting tools, or any higher level application with a functional relationship with its TP trajectory predictions.

A. Recommended Steps for TP and Associated Higher Level Application Evaluation

From quality engineering domain, the process improvement model of Plan-Do-Study-Act (PDSA) provides a high level four step process of iteratively improving a system or process.[9] As illustrated in Fig. 8, it has been adapted as the high level framework for studying the TP and higher level client application functions discussed in this paper. The first is the planning step. It begins by defining one or more focused objectives of the study and expands to listing the questions that the study needs to answer. Still within the planning stage, the appropriate experiment is designed to achieve the objectives and answer the research questions proposed. This includes defining the factors and levels of those factors to test and associated response variables to measure. The next stage is the study execution step, labeled “Do”. It includes development of a set of air traffic scenarios and other associated input data and then the actual operation of running the TP and client application. It also includes the data collection effort that parses and organizes the output from the TP and client application. The third step is the study, composed of the statistical analysis of the TP output. The statistical analysis was subject of the previous Section II. It includes two powerful analysis approaches: observational and experimental data analysis.

Observational data analysis is the examination of differences between data observed in the past in an attempt to make inferences based on these observations. Since the data may be confounded by many uncontrolled or nuisance factors that were captured as well, the analyst must take special care to either block out these influences. Section II’s presentation of the Paired Data t-Test is an example of such a technique.

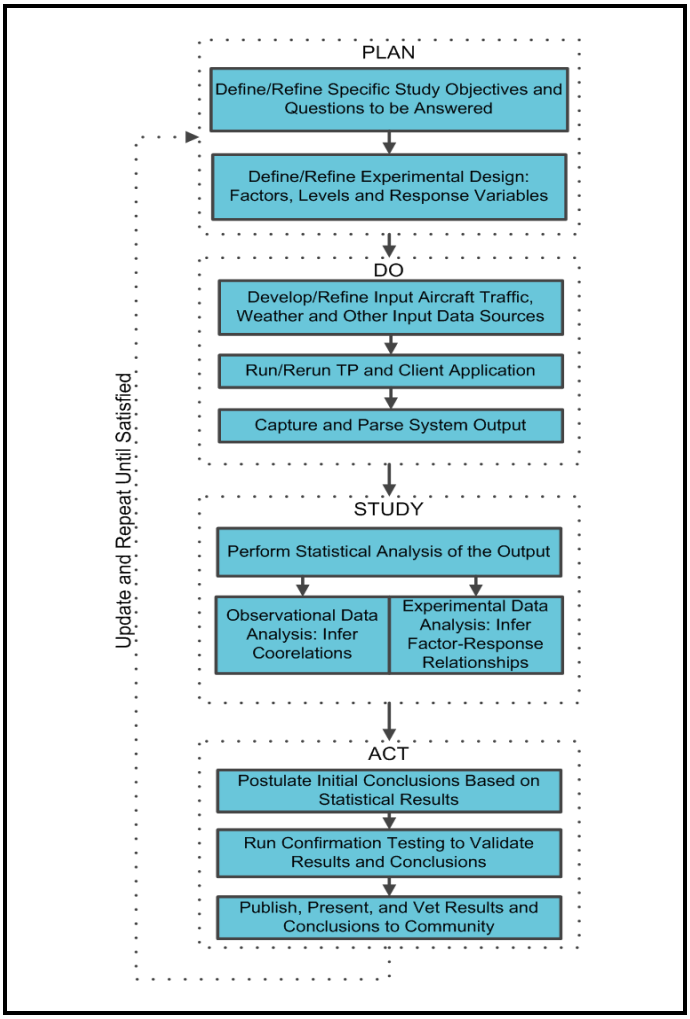

Figure 8: Recommended TP's PDSA Evaluation Process

Experimental data analysis is the approach where the analyst directly manipulates the dependent variables (factors) to determine the influence on the response variables. The design of the experiment can not only eliminate the influence of nuisance factors but maximize the efficiency of the data collected by simultaneously altering the levels of the factors under study. As a result, experimental data analysis can estimate both the main effects and interactions of the factors. For the TP, screening experiments can help determine which factors have statistically significant effects. This technique is beyond the scope of this paper and left for future work.

In Fig. 8, the final stage, labeled "Act", requires the analyst to synthesize the statistical results into a set of conclusions and recommendations based on the original objectives and questions from the planning stage of the study. The steps are notional, so the process can iterate within the stage internally and then expand to publications, presentations and input from the larger community. This stage could also seek out subject matter experts to aid in the interpretation of results. The process does not necessarily end if the feedback indicates otherwise and the analysis can return to the first stage with the lessons learned from the previous iteration. The process is iterative and only ends when the analyst determines it is warranted or like many real world projects the budgeted resources has been exhausted or gains associated with continuing are diminished below the cost of continuing.

B. Important Considerations for the Evaluation Process

When applying this PDSA evaluation in practice, there are several important considerations to follow besides the main steps illustrated in Fig. 8. Listed below are several key lessons documented in the literature and found to be critical in the author's experience to accomplish a successful study.

- As discussed by both Ref. 5 and 6, the analysis should utilize non-statistical knowledge when defining and analyzing the problem as much as possible. For TP and higher level applications, there are subject matter experts that can help explain the relationship of the various factors and response variables. The statistical approaches alone are not likely to explain the relationships between variables. The physical understanding of the factors and their interactions is invaluable in determining the level ranges, number of replications, and in general interpretation of the results under study. For example in Section II.D, several outliers were discussed that represented flights that had significantly larger differences between the two systems under study. Further analysis indicated that these flights exhibited differences in the airspace adaptation between the two systems.
- Ref. 5 and 7 both highlight the difference between practical and statistical significance. Statistical significance is determined by the acceptable Type I error of the test and associated to the sample size and design characteristics of the experiment. Practical significance requires knowledge of the context of the test. It is the actual effect size that matters for the system under study. For example, if Section II's example had reported on vertical trajectory error with a statistically significant effect below 100 feet, the result could have been discarded based on practical significance. The source of altitude data used in the study is reported in 100 foot increments, so an effect below has little meaning.
- Removing uncontrollable or nuisance factor effects in the test can provide increased power to detect actual factor effects for the same sample size and Type I error. This was illustrated in the example in Section II where the Paired Data t-Test provided a significant improvement in sensitivity to detect a shift in the process mean.
- Analysis for TP and their higher level applications, like all good statistical investigations, using the techniques presented is an iterative process.[9] Graphically this was re-enforced in Fig. 8 by the return arrow following the feedback step and returning back to the planning stage. Ref. 5 argues in favor of an iterative and sequential approach to performing the study. This approach maximizes the analyst's chances of meeting the original objectives and even adjusting them if necessary to meet the broader goals that initiated the study from the start. The PDSA Evaluation Process presented assumes a 'learn-as-you-go' principle of discovery with the discipline of a focused and organized process.

IV. Trajectory Performance Correlated with their Higher Level Application

As illustrated in Fig. 1 and stated in Section I, the accuracy of the TP has a direct impact on the DST's overall performance and usability. Using the statistical methods from Section II and applying the PDSA Evaluation Process from Section III, TP performance can be validated and thus improved and these gains in turn can improve the higher level application's function. One very important higher level application is a strategic conflict probe (CP). At its core, the CP has a dedicated TP, producing aircraft trajectory predictions. These forecasts of the aircraft path are used to predict when aircraft pairs could potentially violate safe separation distances – the higher level CP function. A fundamental function of the entire air traffic control system is the safe and efficient operation of aircraft within the airspace of the National Airspace System, so the role of the conflict probe to support the human air traffic controller is an important one. Thus, others have estimated the relationship of CP performance based on the uncertainty associated with TP performance.[10,11] These studies focused on the mathematical modeling of these relationships which are important and consistent with the use of non-statistical subject matter knowledge of the process under study. Others have used simulation modeling to estimate the impact on conflict prediction uncertainty based on trajectory predictability.[12] The focus in this paper is on measuring their relationship empirically through sampling and experimentation.

A. Sample of Accuracy Metrics for TP and CP Functions

The term accuracy is the condition of being correct, exact, or free from error.[13] Accuracy metrics serve to measure the degree of correctness or exactness of a process. For trajectory and conflict prediction of aircraft within the FAA's air traffic control system, accuracy metrics estimate the error in these predictions by comparing the prediction of the event to the actual event. TP accuracy metrics were defined in detail in Ref. 4. To summarize,

time coincident and spatially coincident metrics were defined as well as an un-biased sampling technique to capture these metrics. Table 4 provides a summary of four selected metrics that cover both the horizontal and vertical dimensions of the trajectory prediction. For example, the cross track error is the same metric used in the example presented in Section II.

Table 4: Summary of Sample TP Accuracy Metrics[‡‡]

Metric	Definition	Units
Horizontal Error	Time coincident straight-line planar distance between the predicted trajectory position and actual aircraft position	Nautical Miles
Cross Track Error	Spatially coincident side-to-side planar distance between the predicted trajectory position and actual aircraft position	Nautical Miles
Along Track Error	Spatially coincident longitudinal on path planar distance between the predicted trajectory position and actual aircraft position	Nautical Miles
Vertical Error	Time coincident altitude distance between the predicted trajectory position and actual aircraft position	Feet

A conflict probe predicts when two aircraft will violate separation standards some time in the future. A violation of separation standards is typically called a conflict. An event where two aircraft pass near each other but not close enough to violate separation standards is labeled an encounter in this study. As documented in Ref.'s 14-17, the conflict probe is not perfect and 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 5.

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. For this study, 5 minutes was required unless the conflict was determined to be a pop-up event. 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. Detailed descriptions of the different situations that cause this to occur are described in Ref. 17.

Table 5: CP Alert and Conflict Event Combinations (adapted from Ref. 14 and 17)

	CONFLICT OCCURS	CONFLICT DOES NOT OCCUR
ALERT	CP predicts conflict and it occurs *(VA – valid alert)*	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)

For this paper, the conflict prediction accuracy metrics consist of counts of the error events, including the false alerts (FA) and missed alerts (MA) in context of the correctly predicted events of valid alerts (VA) and correct no-calls (NC). The focus in this paper and subsequent sub-sections will illustrate the correlation between these events and the TP metrics in Table 4.

B. Observational Data Analysis of TP and CP Performance

Following the PDSA Evaluation Process described in Section III, assume for illustrative purposes that the "Plan" and "Do" stages of the study have been completed and the study stage has begun. The objective of the analysis was the development of performance requirements for the CP. To accomplish this, one of the resulting research questions includes determining how much the TP performance effects CP performance. The observational data analysis can utilize some or all the metrics in Table 4 and their associated statistics like sample mean, standard deviation, maximum, or median to name a few. However, a convenient statistic that combines all the three dimensions (i.e. lateral, longitudinal, and vertical) is the ratio of large sampled trajectory errors of the associated aircraft with TP error metrics listed in Table 4 up to 20 minutes prior to the predicted conflict event's (alert) start

[‡‡] See Ref. 4 for the detailed mathematical definitions of the listed metrics.

time. The ratio is the quantity of trajectory accuracy sampled positions that exceeded a threshold to the total number of events sampled within the same time interval. The various types of alert events defined in Table 5 then can be compared as a group difference. For example, the difference between missed alert events and valid alert events or difference between false and valid events can be statistically tested. This difference can also be compared between two systems (i.e. different CPs with their own TP).

For example, two CP tools were provided the same input traffic scenario using the techniques defined in Ref. 18 and 19. This process includes generation of a recorded aircraft traffic scenario that is time shifted to induce test conflict events and then used to test the CP. For this example, the trajectory predictions are sampled using the interval based sampling technique defined in Ref. 4, and the unsigned cross track, along track, and vertical errors as listed in Table 4 were counted that exceeded the thresholds 1.5 nautical miles, 2.5 nautical miles, and 500 feet, respectively. This was applied to the entire set of false alert (FA), valid alert (VA), and missed alert (MA) events detected for the two CP tools evaluated. The null hypothesis claims that the average ratio for these three groups of alert types is equivalent. The alternative hypothesis claims that they are not equivalent. Subject matter knowledge of the TP and CP algorithms suggest that events that exceed the thresholds of trajectory deviations in any of the three dimensions will increase the likelihood of producing false and missed alerts and lower the chance of making valid alerts.

As the mean value of the ratio of trajectory deviations increases, the slope of the cumulative distribution function (CDF, integral of the probability density function) should increase. This is illustrated in Fig. 9 and Fig. 10 by the two example CP tools' cumulative probabilities as a function of the ratio of trajectory deviation per alert type. For both CPs, the blue farthest left curve is composed of the correct prediction events or valid alerts (VAs). The VA events are farthest to the left because for a given average ratio of trajectory deviation they are more likely to occur at a lower ratio than a higher value relative to the two other alert types. The green curve for the missed alert (MA) events is farthest to the right for both CPs. It is constructed from a very small sample size of events (i.e. five for the CP A and six for CP B) and as a result produced a discrete step function instead of a smooth curve like the other two alert types. Regardless of the discreteness of the curves, both CDF lines in Fig. 9 and 10 for the MA events illustrate the relationship of missed alert events and their associated trajectory deviations. The curves are steep and more to the right than the other alert types indicating the more trajectory deviations are present in the missed alert events. Finally, the third alert type formed from FA events is illustrated as a red colored line in both Fig. 9 and 10. The curves for both example CPs is to the right of their VA CDF curves. Both VA and FA events include conflict predictions by the CP. VA alert types are situations when the prediction can be matched to a test conflict. The prediction is therefore both accurate and timely. However, FA alert types cannot be matched to a test conflict or represents situations where the CP prematurely removed an alert for the test conflict. It represents situations where the CP is falsely presenting an alert or prematurely removing them. The difference between the VA and FA curves is proportional to the difference in quantity of trajectory deviations present in the two sub-groups of conflict predictions. Thus, the FA's CDF that is shifted to the right of the VA CDF indicates that FA events have more trajectory deviations.

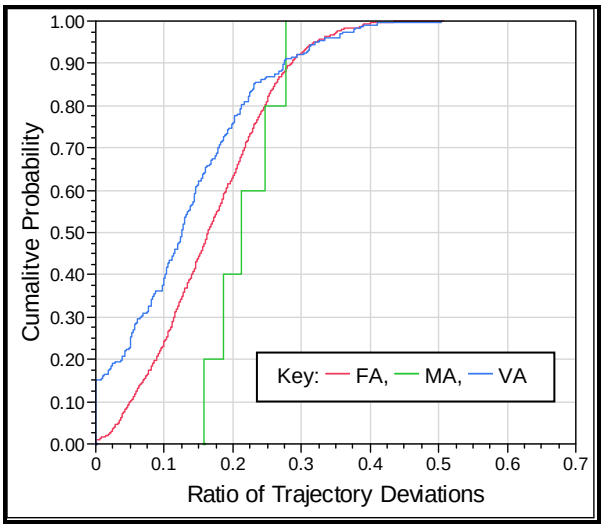

Figure 9: CP A - CDF of Ratio of Trajectory Deviations Per Alert Type

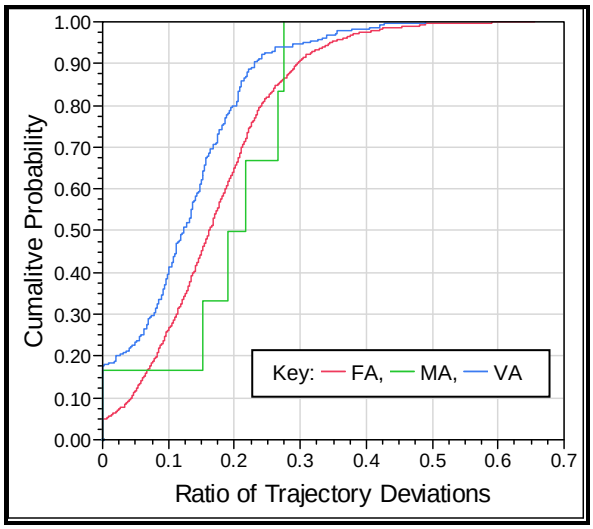

Figure 10: CP B - CDF of Ratio of Trajectory Deviations Per Alert Type

The CDFs give us some indication of the distribution of trajectory deviations as a function of alert type but does not provide a means of testing whether these differences are statistically significance. For the three groups in the current example, there are three pairwise comparisons: FA to VA, MA to VA, and FA to MA. Fig. 11 and Fig. 12 illustrates the comparisons graphically using the SAS JMP® statistical package. The figures present the individual event ratios for each alert type group and summary statistics per group. Specifically, this includes black dots as the ratio of trajectory prediction for each individual event grouped by alert type, a red box plot with 75th, 50th (median), and 25th percentiles for each alert type, blue lines indicating the sample mean and standard deviations for each alert type group, and green diamonds are centered on the sample mean with height proportional to the upper and lower bounds of a 95th level confidence interval. As expected, both Fig. 11 and 12 illustrate the differences in alert type populations. In both CPs, the FA and VA confidence intervals are not overlapping, yet the MA overlaps with the other two. MA has the highest mean and median ratio of trajectory deviations, while FA has the second largest and VA the smallest of the three.

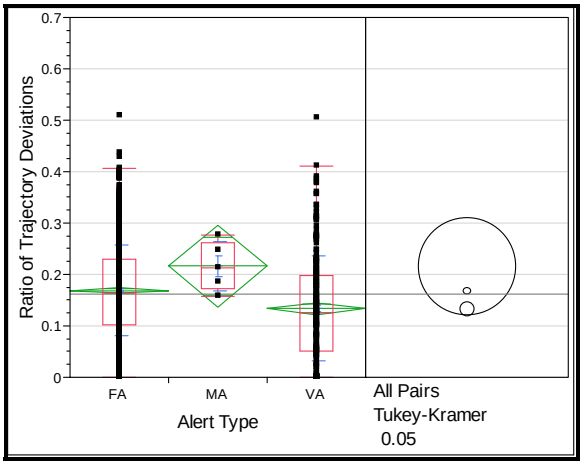

Figure 11: CP A – Comparison of Multiple Group Means and the Tukey-Kramer Test

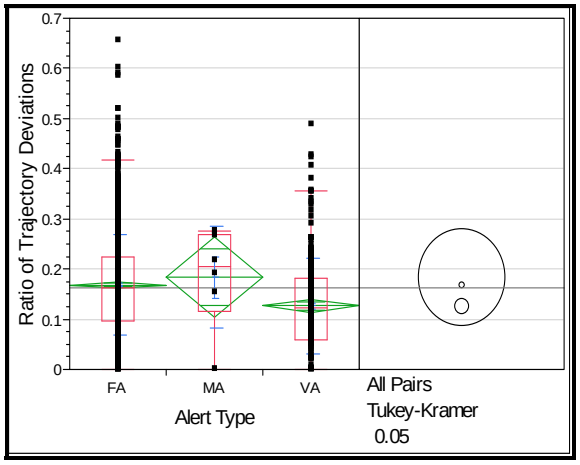

Figure 12: CP B – Comparison of Multiple Group Means and the Tukey-Kramer Test

A specific multiple comparison method is employed to compare which pairs of means are statistically different and which are not. The Tukey-Kramer HSD (Honestly Significant Difference) Test is a conservative test in that it considers all comparisons to control the experimentwise Type I error.[20] It ensures that the entire test as a whole has a specified Type I error. Graphically, the circles that are not overlapping are significantly different, and others that do overlap are not statistically different. For both CPs, the small sample size and large variability of MAs contributes to producing a large circle encompassing the other two alert types. In contrast, the circles for the FA and VA alert types do not overlap indicating there is a statistically significant difference between them. Therefore, still only observational in nature, all four figures from Fig. 9 to 12 consistently show evidence that both CPs have higher incidents of trajectory deviations (trajectory errors exceeding specified thresholds) for MA events, then FA events and then VAs. They also give statistical evidence on the magnitude of these differences. However, the results do not provide evidence that trajectory error induces the conflict prediction errors only that they are correlated. Also, it does not directly compare the two CPs only provides relative information on both systems. The next sub-section will present techniques to compare both CPs directly and utilize the paired data techniques presented in Section II.

C. Observational Comparison of Two System's TP and CP Performance

The Paired Data t-Test presented in Section II compared two releases of the TP with the same input traffic scenario. This allowed the same aircraft to be paired and differences calculated reducing the overall test variability and increasing the power of the test. In a similar manner, it was shown in Ref. 18 and 19 that conflict predictions could be paired between runs as well. The process involves matching the FA, MA, and VA events between the two runs. This results in the list of combined results given in Table 6. If the TP accuracy influences the CP accuracy as expected, the matched results where both CPs have equivalent performance (both VA, MA, or FA) should have negligible TP accuracy differences, compared to when the CPs do not have equivalent performance (e.g. MA_VA, FA_NC). Furthermore, the techniques presented from Section II can be used.

Table 6: Matched Conflict Prediction Results

Conflict Probe A Result	Conflict Probe B Result	Matched Result
VA – valid alert	VA – valid alert	SAME_VA
FA – false alert	FA – false alert	SAME_FA
MA – missed alert	MA – missed alert	SAME_MA
VA – valid alert	MA – missed alert	VA_MA
MA – missed alert	VA – valid alert	MA_VA
FA – false alert	NC – correct no-call	FA_NC
NC – correct no-call	FA – false alert	NC_FA

For example, using the same two CP's presented in the previous sub-section, let's contrast the paired differences of the ratio of trajectory deviations for events when both systems had valid alerts (SAME_VA) and contrast this with the situation when a false alert occurs in CP A and CP B correctly does not make a prediction (correct no-call) to produce the FA_NC matched result. Fig. 13 and 14 presents the SAME_VA and FA_NC paired data results for the ratio of trajectory deviations as defined in the previous Section IV.B. Unlike the previous example, the distributions in Fig. 13 and 14 are the paired differences in ratio of trajectory deviations for the same aircraft pairs matched by event type.

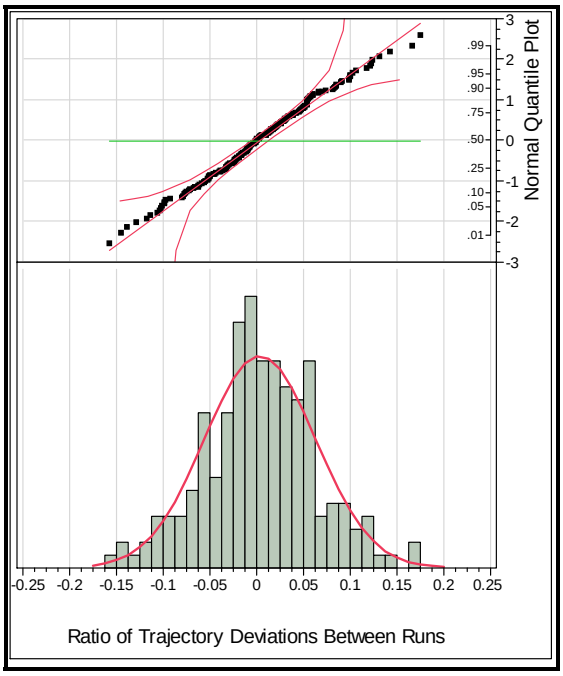

Figure 13: SAME_VA – Paired Data Histogram/Normal Quantile Plot

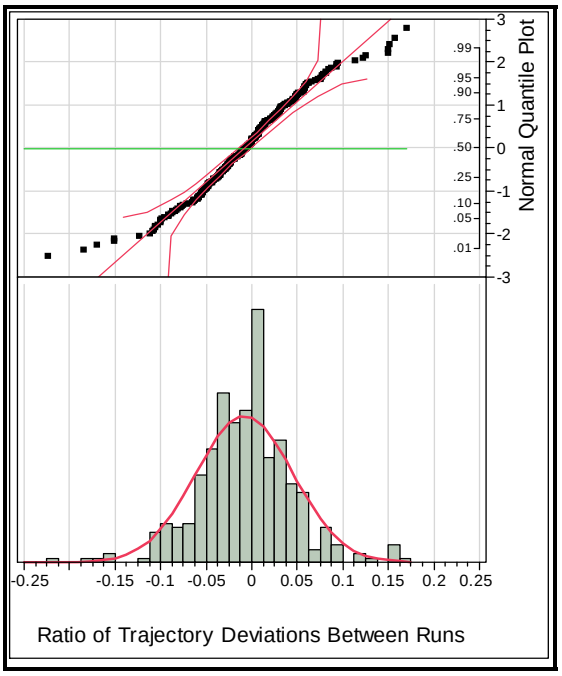

Figure 14: FA_NC – Paired Data Histogram/Normal Quantile Plot

For the Fig. 13, the distribution is approximately normally distributed as illustrated by the normal quantile plot and symmetric histogram. The mean difference in ratio of trajectory deviations is 0.003 and p-value from the Paired Data t-Test is 0.43, indicating the difference is not statistically significant. Thus, for matching valid alert events there is no evidence to conclude the quantity of trajectory deviations is different between runs.

For Fig. 14, the distribution is reasonably symmetric as illustrated by the histogram. However, there is reason to believe the data is not normally distributed as shown by the quantity of data points in the tail regions illustrated as deviations in the normal quantile plot. Thus, for this data set the non-parametric Wilcoxon Signed-Rank Test, presented in Section II.D, should be employed to confirm any results from the Paired Data t-Test. The mean difference in ratio of trajectory deviations is 0.008 and p-value for the Paired Data t-Test is 0.006 and Wilcoxon Signed-Rank Test is 0.004. Thus, both are consistent indicating for the FA_NC data set, the mean difference of trajectory deviations of approximately one percent was statistically significant. Overall, there is may be a question

to whether a one percent different in the quantity of trajectory deviations has practical significance, but the approach of using paired data tests, matching by alert types for the CP, and in general the PDSA Evaluation Process was repeatedly illustrated using observational data analysis techniques.

V. Conclusions

In unprecedented proportions, industry and ATSPs have come together to mitigate the forecasts of increased demand resulting in a capacity constrained aviation system that could potentially degrade in efficiency and ultimately safety if nothing is done. Thus, comprehensive plans requiring broad advances in ground-based and airborne automation have been developed in the United States and as well as in Europe. The 4-dimensional prediction and management of the aircraft path is at the heart of the concepts envisioned by the next generation aviation system (NextGen). Many of the decision support tools will rely on these trajectory predictions and their performance are tightly coupled. This paper showed the linkage between the accuracy of the TP's predictions to higher level applications like a CP which predicts when aircraft will potentially violate separation standards. Notionally, the relationship between the inner functions of the TP and its higher level application is illustrated in Fig. 1.

Powerful statistical analysis concepts should be employed to help quantify the relationship between TP and its higher level functions. This is critical at several levels. First, to derive valid technical requirements for the TP and its higher level applications, it is important to specify both in context of the other. Second, as NextGen functions are developed the current performance of our air traffic control automation will be required to improve to achieve the benefits being proposed. By employing the various TP metrics in Ref. 4, CP metrics from Ref. 19, and statistical methods presented in this paper, the process provides a means to discriminate where improvement is needed and thus can be targeted for re-engineering and re-design.

Utilizing the experimental design technique of blocking out uncontrollable factors out of the study by pairing the data by flight, the Paired Data t-Test and its non-parametric counter part, the Wilcoxon Signed-Rank Test, provide a strong foundation of analysis techniques to validate TP performance from one upgrade to the next. The broader concept of continuous improvement is then presented in Fig. 8 to guide the analyst in all phases of the study. The process is broken up by simple stages adapted from the quality engineering literature, called PDSA or Plan-Do-Study-Act. A key feature of the approach is the iterative nature of the technique. The process promotes an exploratory and incremental approach to study a new or existing system. The study phase encompasses both observational and experimental statistical techniques. The paper focuses on the observational data analysis where inferences are made based on what can be observed from runs of the TP and CP tools.

Future work will leverage on the metrics and analysis techniques presented and illustrate the PDSA Evaluation Process using experimental data analysis. This type of analysis allows direct manipulation one or more independent variables, called factors, and inferences can be drawn on the influence on the dependent or response variables being studied. While the observational analysis presented in this paper does provide evidence of these relationships for the main factors, experimental design techniques take it a bit further in tightly controlling any additional bias or confounding nuisance factors and allow estimation of interaction effects among these variables. Both techniques form a balanced study approach to achieve the research objectives defined in the planning phase of the activity.

Acknowledgments

The authors thank General Dynamics Technical Professional, Shurong Liu, for his tireless support throughout this study in developing several specialized software programs to capture the data for the various analyses. The authors would like to also acknowledge the successful utilization of SAS JMP® professional statistical tool for production of most of the statistical figures and analysis tests within the paper. Finally, the authors would like to acknowledge our FAA sponsors, Pam DellaRocco and Craig Marina, in the FAA's En Route and Oceanic Program Operations Office in support of the Separation Management Project.

References

[1]Joint Planning and Development Office, *Concept of Operations for the Next Generation Air Transportation System, Version 2.0*, [online library], URL: http://www.jpdo.gov/library/NextGen_v2.0.pdf, June 2007.

[2]Paglione, Mike, Carlos Garcia-Avello, Sip Swierstra, Robert Vivona, Steven Green, October 2005, *A Collaborative Approach to Trajectory Modeling Validation*, 24th Digital Avionics Systems Conference, Washington, DC.

[3]Mondoloni, Stéphane, Sipka Swierstra, and Mike Paglione, October 2005, *Assessing Trajectory Prediction Performance – Metrics Definition*, 24th Digital Avionics Systems Conference, Washington, DC.

[4]Paglione, M., R. D. Oaks, "Implementation and Metrics for a Trajectory Prediction Validation Methodology," *Proceedings of the American Institute of Aeronautics and Astronautics (AIAA) Guidance, Navigation, and Control Conference*, August 20-23, 2007.

[5]Montgomery, Douglas C., *Design and Analysis of Experiments, Fourth Edition*, New York, NY, John Wiley & Sons, 1997.

[6]Box, George E. P., Hunter, J. Stuart, Hunter, William G., *Statistics for Experimenters, Design, Innovation, and Discovery, Second Edition*, Hoboken, NJ, John Wiley & Sons, 2005.

[7]Devore, Jay L. 2000, *Probability and Statistics for Engineering and the Sciences*, 5th Edition, ISBN 0-534-37281-3, Duxbury Press, Belmont, CA.

[8]SAS Institute, 2007, *JMP® Software: Classic Design of Experiments Course Notes*, SAS Institute Inc. Cary, NC.

[9]Langley, Gerald J., Ronald Moen, Kevin M. Nolan, Thomas W. Nolan, Clifford L. Norman, Lloyd P. Provost, 2009, *The Improvement Guide: A Practical Approach to Enhancing Organizational Performance*, 2nd Edition, ISBN 978-0-470-19241-2, Jossey-Bass, San Francisco, CA.

[10]Paielli, R. A., and H., Erzberger, "Conflict Probability Estimation Generalized to Non-Level Flight," *Air Traffic Control Quarterly*, Vol. 7, No. 3, 1999, pp. 195-222.

[11]Irvine, Richard, "A Geometrical Approach to Estimating Conflict Probability Estimation," *4th USA and Europe Air Traffic Management R&D Seminar*, Sante Fe, New Mexico, U.S., December 3-7, 2001.

[12]Magill S., 1997, "Trajectory Predictability and Frequency of Conflict-Avoiding Action," *CEAS 10th European Aerospace Conference on Free Flight*, paper 34-1, Amsterdam, 1997.

[13]Random House, 1984, *The Random House College Dictionary*, Revised Edition, Random House, Inc., New York, NY.

[14]Paglione, M., M. Cale, H. Ryan, "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), Fall 1999.

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

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

[17]Paglione, M., R. Oaks, H. Ryan, "Methodology for Evaluation and Regression Testing a Conflict Probe," *Digital Avionics System Conference*, October 2004.

[18]Paglione, M., R. D. Oaks, K. D. Bilimoria, "Methodology for Generating Conflict Scenarios by Time Shifting Recorded Traffic Data," *Proceedings of the American Institute of Aeronautics and Astronautics (AIAA) Technology, Integration, and Operations (ATIO) Technical Forum*, November 2003.

[19]Ryan, H., G. Chandler, M. Paglione, C. Santiago, S. Liu, "Evaluation of En Route Automation's Trajectory Generation and Strategic Alert Processing, Analysis of ERAM Performance" DOT/FAA/TC-TN08/10, Federal Aviation Administration, William J. Hughes Technical Center, April 2008.

[20]SAS Institute, 2007, *JMP® Software: ANOVA and Regression Course Notes*, SAS Institute Inc. Cary, NC.