Non-Parametric Approach to a Paired Data Hypothesis Test in Trajectory Accuracy Analysis

M. M. Paglione[1]

Summary

Trajectory accuracy plays an important role in evaluating air traffic management decision support tools. To increase the confidence in trajectory accuracy, various statistical approaches have been applied and not all are sufficient in handling data that has multivariate characteristics. Expanding upon previous work, the Federal Aviation Administration's Conflict Probe Assessment Team (CPAT) has developed a practical methodology using a non-parametric paired-data hypothesis test. The technique properly blocks out nuisance factors, focuses the analysis on the factor under study, and is robust in the presence of outliers. This is very useful when air traffic data is very heterogeneous, which is often the case. For example, a given traffic sample will have many flights with various aircraft types, following different routes and altitude profiles, resulting in substantially different accuracy performance. Another practical benefit of the technique is the capability of ranking the individual performance of a given set of flights against a baseline of performance. As a result, the approach supports regression testing as well as overall system measurement.

Introduction

To achieve the goals of Free Flight, broad categories of advances in ground and airborne automation are required. The Federal Aviation Administration (FAA) has sponsored the development of several ground based air traffic management decision support tools (DSTs) to support the en route and terminal air traffic controllers. A fundamental component of a DST's design is the trajectory modeler, upon which its functionality is based. The trajectory modeler provides a prediction of the aircraft's anticipated flight path, determined from sources such as the flight plan and radar track data received from the National Airspace System (NAS) automation. Therefore, the trajectory accuracy, or the deviation between the predicted trajectory and the actual path of the aircraft, has a direct effect on the overall accuracy of these automation tools.

In [2], the Conflict Probe Assessment Team (CPAT) at the FAA's William J. Hughes Technical Center developed a generic method of sampling a set of aircraft trajectories for accuracy measurements, called interval-based sampling. It has been applied to two of the FAA's trajectory prediction tools, the NASA-developed Center-TRACON Automation System (CTAS) and the MITRE/CAASD-developed User Request Evaluation Tool

[1] Federal Aviation Administration, Simulation and Analysis Group; mike.paglione@faa.gov

(URET). Both these systems have been deployed as production systems into the NAS. However as systems like these are upgraded over time for new aircraft types and/or new functionalities, there is a need for testing whether the upgrades have not inadvertently introduced inaccuracies in the trajectory modeling function. This type of testing is often referred to regression testing in the software community. An established inferential statistical technique was presented in [1]. The technique provided a practical approach to state with confidence that the trajectory accuracy has or has not degraded. Expanding on [1], this paper presents the parametric statistical approach, next its assumptions will be discussed, and finally a non-parametric technique will be described that relaxes some of these assumptions.

Hypothesis Testing for Trajectory Analysis

The regression test requires a baseline version of the trajectory modeler software to be run with a given traffic sample. This same traffic sample is then run through the upgraded software, which is referred to as the new release version. Both runs are then processed for trajectory accuracy using the interval-based sampling method as defined in [2]. There are several trajectory accuracy metrics that are examined simultaneously using this process, but for simplicity this paper will focus only on the lateral error, which is the perpendicular distance between the sampled aircraft surveillance position and the time coincident trajectory predicted position. It is expressed in units of nautical miles and is positive signed if the prediction is to the right of the aircraft. To compare the runs, the difference between the baseline and new release sample mean is calculated. Since the sample mean is a statistic and thus a random variable of the true population mean, a statistical hypothesis test is used that considers the variation in both samples. If the true population means were known, the difference between the two means could be calculated exactly. If the difference were zero, it would be concluded that the runs were not equivalent. As described by Devore in [3], the Two-Sample *t* test provides a statistical hypothesis test that provides a criterion to reject the hypothesis that the sample means are not equal. This null hypothesis is expressed in the following equation (1).

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

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

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 the following sections.

Assumption of Independence Sample Runs

Since the same air traffic sample is input into both runs of the trajectory model, the other variables that influence trajectory accuracy are expressed in the variability of flights in the two runs. These flights are the same for each run, so their influence has a proportional effect on both runs. If a specific flight exhibits higher than normal error in the baseline run, it would be expected that the same flight would have similar high error in the new release run. Of course some flights may exhibit better performance in the new release, if indeed the upgrade was to reduce these errors, but on average if the flights perform in this manner, the runs are not independent. In [1], a trajectory accuracy example illustrated this lack of independence between runs, resulting in erroneous conclusions. An alternative technique was recommended and is presented again next.

Application of a Paired t-Test

Instead of taking the difference between the sample means, the sample measurements are paired for the same flight and position. 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 and position from the two runs produces a new statistic, the sample differences. This is expressed in the following equation (2):

$$D_i = x_i - y_i \tag{2}$$

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

Therefore, the hypothesis now is that the sample mean of D_i's is equal to zero. The mean of the difference between two numbers is equal to the difference between the means of the same set of numbers. Therefore, while the hypothesis in equation (1) is the same, the test statistic compared to a Two-Sample *t* test is not (see [1] for details). The following equation expresses the Paired *t* Test's test statistic:

$$t = \frac{\bar{d}}{s_D / \sqrt{n}} \tag{3}$$

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.

The rejection region of the Paired *t* Test is expressed in the following:

$$\text{Reject null hypothesis if } t \geq t_{\alpha/2,n-1} \text{ or } t \leq -t_{\alpha/2,n-1} \tag{4}$$

where $t_{\alpha/2,n-1}$ or $-t_{\alpha/2,n-1}$ are parameters taken from the student-t distribution, α is the significance level of the test, and *n-1* is the degrees of freedom for this test (number of samples minus one).

Example Application of the Paired t Test

To test the hypothesis defined in equation (1) for the measurements of trajectory lateral error, two runs were performed on a NAS trajectory modeler and the lateral error was measured at a look-ahead time of 15 minutes. The sample scenario was based on two-hours of recorded traffic data from Indianapolis en route center in May 1999. The trajectory modeler produced over 5000 trajectories for each of the runs. The baseline run produced a sample mean of 0.60 nautical miles of lateral error and a sample standard deviation of 5.58 nautical miles (square root of the sample variance). The new release run produced a sample mean of 0.56 nautical miles and sample standard deviation of 5.62 nautical miles. The sample mean of the differences is 0.038 nautical miles and sample standard deviation of the differences is 0.559 nautical miles. Since the same traffic sample was run through the trajectory modeler, both runs are balanced with the same quantity of 832 measurements of lateral error.

As shown in [1] and discussed previously, the Paired *t* test offers significance precision due the heterogeneity in the runs. By applying equation (2) on the above values, the test statistic *t* equals 1.99. The rejection region from equation (3) equals ± 1.96, using a significance of 0.05 and 831 degrees of freedom. This value is found in Table A.5 from [2] as the critical value taken from a student t distribution. Therefore, the hypothesis that the mean horizontal error of the two runs is equivalent can be rejected (i.e. t is $\geq t_{0.025,831}$ or $\leq -t_{0.025,831}$). Therefore, the upgrade or new release trajectory model is considered statistically different to the previous baseline version. In this case, it has slightly less error.

As discussed in [3] and shown explicitly in [1], the Paired *t* Test has a property of improving the precision of the test statistic when there is a correlation between runs and significant heterogeneity between samples (in this example the difference between flights).

Assumption of Normality of Samples

Even though the data was paired correctly, the result in the previous example is surprising, since the difference in sample means was only 0.038 nautical miles. Further inspection of the data showed that six measurements of the 832 total were more than six standard deviations larger than the sample mean of the differences. Removal of these six outliers produced very different results with a test statistic of only 0.116, well below the 1.96 rejection criterion.

Devore in [3] offers some insight into why the Pair 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.

Application of the Wilcoxon Signed-Rank Test

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 [3] 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].
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 equation (5).

$$Z = \frac{S_+ - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}} \tag{5}$$

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 equation (6):

$$\text{Reject null hypothesis if } Z \geq z_{\alpha/2} \text{ or } Z \leq -z_{\alpha/2} \tag{6}$$

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

[2] Refer to Section 15.1 in [3] for a technique to handle the ranking of tied values.

Example Application of Wilcoxon Signed-Rank Test

Now repeating the same example as before but applying the Wilcoxon Signed-Rank Test, the sample median is 0.00 nautical miles. Using the Statistical Analysis System (SAS) JMP Software package, the Signed-Rank statistic, $S_{+,}$ is 14390.5. By applying equation (5), the Z statistic equals 0.856. The rejection region from equation (6) again equals ± 1.96, using a significance of 0.05. Therefore, the hypothesis that the mean lateral error of the two runs is equivalent cannot be rejected, since 0.856 is not greater that +1.96 or less than –1.96 (i.e. is not $Z \geq z_{0.25}$ or $Z \leq -z_{0.25}$). This is a very different result from the application of the Paired t-test presented in the previous example using the same data set. Like the Pair t-test, the samples are paired, providing the same benefits in precision, but this procedure is much less sensitive to outliers. The result is now consistent with data inspection with 80 percent of the data within ± 0.002 nautical miles.

Conclusion

In conclusion, the development and later maintenance of the trajectory modeling function of FAA decision support tools requires frequent regression testing between baseline and new releases of the software. To perform this testing effectively, it is recommended that the trajectory accuracy measurements between runs be paired for the same flight and position. It was shown in [1] and repeated in this paper that the Paired t Test can be used, which has the property of improved precision by reducing the variance in the samples. This allows the runs to be correlated, but still requires the samples to be normally distributed. The Wilcoxon Signed-Rank Test is the recommended procedure, since it also pairs the sample differences but relaxes the normality assumption. As a result, it is much less sensitive to outliers.

References

1 Paglione, M, Charles, L. (2003): "Applying Pairwise Hypothesis Testing In Trajectory Accuracy Analysis," 48th Air Traffic Control Association Annual Conference Proceedings.

2 Cale, M, Liu, S, Oaks, R, Paglione, M, Ryan, H, Summerill, S. (2001): "A Generic Sampling Technique for Measuring Aircraft Trajectory Prediction Accuracy," Presented at the 4th USA/Europe Air Traffic Management R&D Seminar, Santa Fe, NM.

3 Devore, Jay L. (2000): *Probability and Statistics for Engineering and the Sciences, 5th Edition*.