AAS 09-304

COVARIANCE REALISM

David A. Vallado[*] and John H. Seago[†]

Covariance information from orbit determination is being relied upon for space operations now more than ever. There have been scattered claims and discussions of realistic covariance, but not enough detailed studies to demonstrate the actual performance against independent references using real data. This paper discusses some statistical tests that could be used to help study predicted covariance accuracy. To illustrate the methods, the authors estimate prediction error by comparing predictions to a precision orbit estimated after the fact. The predicted covariance is analyzed relative to the sample error estimates using the methods described.

INTRODUCTION

The covariance matrix from the solution of orbit determination problems has relevance as a measure of parameter uncertainty under rather restrictive assumptions. The use of covariance to assess confidence in astrodynamical operations, such as tracking acquisition for scheduling operations, conjunction probability, relative motion operations, *etc.* will be of limited value when the assumptions are unrealistic. The notion of covariance realism is not without precedent, but a clear definition has been often out of reach. Most often, covariance realism is synonymous with covariance *accuracy*, gauged by comparing the propagated covariance with positional differences found after an accurate reference ("truth") orbit is generated over the time interval of interest.

Estimates of satellite-location uncertainty, by way of the covariance matrix, are particularly useful for computing the probability of collision between two orbiting bodies. For example, for the operational SOCRATES-GEO program producing conjunction probability calculations for many geosynchronous-satellite owner-operators, the use of an accurate covariance becomes a significant differentiator to limiting unnecessary maneuvers and thereby has the potential to extend the operational lifetimes of spacecraft.[1] Orbital safety may then be the most significant driver for realistic covariance.

Linear combinations of independently distributed standard Gaussian variates are also Gaussian distributed.[2] Once the approximate normality of observation errors can be assumed, then appropriately weighted linear combinations of these errors are also approximately normal. However, Junkins *et al.*, Alfriend *et al.*, and others, suggest that whenever satellite positions are forecast for very long, or if the errors are very large, error mappings become more non-linear through time, and thus the error distributions should no longer be supposed Gaussian. For this reason, there is a theoretical expectation that orbit errors will eventually become abnormally distributed in some situations.[3, 4, 5, 6, 7] In particular, if the covariance is propagated in rectangular coordinates, then the disparity between the covariance and the propagated errors grows more rapidly due to the nonlinearity of the dynamics, preventing the covariance from being a good indicator of the orbit-error uncertainty.

---

[*] Senior Research Astrodynamicist, Analytical Graphics Inc., Center for Space Standards and Innovation, 7150 Campus Dr., Suite 260, Colorado Springs, Colorado, 80920-6522.
[†] Astrodynamics Engineer, Analytical Graphics Inc., 220 Valley Dr., Exton, Pennsylvania, 80920-6522.

In contrast, propagation studies using orbits computed from US Space Surveillance Network (SSN) tracking data have concluded that, with few exceptions, propagated error distributions are normally distributed, although the scale (volume) of the covariance may be incorrect.[8, 9] When this is the case, "covariance realism" only needs to address the scale differences of the propagated covariance, rather than the shape.

This paper suggests several statistical tests that could be used to help assess the accuracy of covariance. It also provides a glimpse into the behaviors of covariance matrices using real satellite data to lay the foundation for additional study of covariance realism for predicted satellite states. One goal is to see if predicted covariance tends to match the sample error estimates (based on *post-priori* estimates of "truth") for the SOCRATES conjunction-analysis program. Another goal is to see if specific populations of satellites have covariance-accuracy behaviors that are similar enough to reliably categorize them into special classes.

MATHEMATICS OF COVARIANCE MATRIX

Vallado shows a mathematical description of the covariance matrix.[10] The notation uses $\boldsymbol{P}$ as the covariance, $\boldsymbol{A}$ as the partial-derivative matrix (partial derivatives of the observations with respect to the estimated parameters), and $\boldsymbol{W}$ as the measurement-noise matrix:

$$\boldsymbol{P} = (\boldsymbol{A}^T \boldsymbol{W} \boldsymbol{A})^{-1} \tag{1}$$

State errors are advanced through time using the state-error transition matrix $\boldsymbol{\Phi}$, and the process noise ($\boldsymbol{Q}$) which is mathematically defined for sequential estimators, such as the Kalman Filter. The covariance is then propagated in time with the following equation:

$$\bar{P}_{k+1} = \Phi \bar{P}_k \Phi^T + Q \tag{2}$$

A key to covariance realism may be the process noise $\boldsymbol{Q}$, the implementation of which can often invoke vigorous discussion (and is taken to be zero in many applications). One method of specifying $\boldsymbol{Q}$ is by trial and error to see what "works best" within a given estimation system. This is the traditional approach but it can be prone to failure. The second approach approximates the process-noise matrix using the uncertainty of parameters within the acceleration / force model.[11, 12, 13] This technique has been successfully implemented and operationally used for many years, and is the process-noise method used for this paper.[14]

Covariance propagation for sequential estimators is also a function of the update processing (the Extended Kalman filter application being most common for orbit determination). The update equation is shown below, where $\boldsymbol{K}$ is the Kalman gain, and $\boldsymbol{H}$ is the measurement-state partial matrix (analogous to $\boldsymbol{A}$ above):

$$\hat{P}_{k+1} = \bar{P}_{k+1} - K_{k+1} H_{k+1} \bar{P}_{k+1} \tag{3}$$

The differences between a batch-least-squares estimator and a Kalman filter are well known.[15, 16] Unlike the ability to align numerical propagation methods between programs, aligning entire orbit determination processes is more difficult, if for nothing else, due to mathematical differences in the approaches.[17] Therefore, different programs will arrive at different orbital state with covariance estimates that would propagate to a different outcome in different programs.[*]

STATISTICAL TESTS OF HYPOTHESES

To assess the "realism" (accuracy) of uncertainty measures such as covariance matrices, we sometimes rely on statistical tests of hypotheses that can be used to reject specific assumptions about sample data. A

[*] The covariance matrices will be different when they arise from different software models. Different force models will result in different accelerations and state estimates. However, even if the force models are identically programmed, the software may generate different answers for the final state due to the particular implementation choices of the user.

common example related to orbit determination is the practice of outlier rejection. In this situation, if an individually observed measurement ***m*** is exceedingly far away from its expected value, then the measurement is ignored. However, "exceedingly far away" is a subjective notion that will vary from one analyst to the next, so this approach is made more objective by introducing a statistical test. Specifically, if the magnitude of an individual residual $|\Delta m|$ divided by its uncertainty σ is greater than some threshold, say, $C = 3$, observation ***m*** is ignored.[18]

The rejection threshold C, or ***critical value***, is chosen presuming that an outcome $|\Delta m| / \sigma < C$ would be highly improbable. If the distribution of the residual ratios $\Delta m / \sigma$ is known, then the probability of accidental rejection can be established at this critical value. If the assumption of normality holds, and $C = 3$, the probability of $|\Delta m| / \sigma < C$ is $\Pr\{C\} = 99.932\%$, and the probability of accidental rejection is $1 - \Pr\{C\} = 0.068\%$. Therefore, the analyst may feel quite justified in concluding that the measurement should be rejected because the probability of rejecting a valid datum is quite low for the critical value $C = 3$.

Testing Hypotheses

The basic elements of a statistical hypothesis test were established in the preceding example:

1. The measurement is not tested directly, but rather a proxy ***test statistic*** is computed based on the value of the measurement (in the previous example, $\Delta m / \sigma$).
2. The test statistic is chosen because it has a testable distribution under the operating assumptions (in the previous example, normality).
3. The analyst chooses an appropriate critical value which corresponds with an improbable outcome for the distribution. The probability of a successful test outcome at this critical value C is called the ***confidence level*** of the test ($\Pr\{C\}$), and the probability of accidental failure at this critical value is called the ***significance level*** of the test ($1 - \Pr\{C\}$).
4. A value of the test statistic is computed from the sample and compared to the critical value; if the test statistic exceeds the critical value, then the analyst rejects the hypothesis being tested (in the previous example, that ***m*** was a valid measurement); otherwise, he embraces the assumptions due to the lack of evidence that they are untrue.[*]

The set of ***status-quo*** conditions underlying the test is known as the ***null hypothesis*** (***null*** implying "no-change"). For the outlier example, the null hypothesis is that every value of $\Delta m / \sigma$ will be normally distributed with zero mean and unit variance, which is equal to saying that Δm will be normally distributed with zero mean and variance σ^2. Whenever an outcome does not pass the test, the analyst will usually question the experimental outcome (***m***), but he may also question the ***status-quo*** conditions underlying the test statistic (***e.g.***, the correctness of assuming normality, zero mean, and unit variance).

Most Powerful Statistical Tests

One way to obtain insight into the validity of a statistical test is to repeatedly evaluate samples from a known distribution. For an assigned critical value C, in the long run, one can expect $\Pr\{C\}$ successes and $1 - \Pr\{C\}$ failures because the null hypothesis always holds for the simulation case. For example, if one were testing random outcomes at the $1 - \Pr\{C\} = 5\%$ significance level, he would expect a "most powerful" statistical test to reject 5% of the outcomes should the null hypothesis be true.[†] Rejection rates much less than 5% would provide evidence that the test may be unable to reasonably reject the null hypothesis; that is, the test "lacks power against" the null hypothesis and may be inappropriate to use.

---

[*] A limitation of statistical hypothesis testing is that it doesn't actually *prove* anything; it can only give evidence for rejecting claims based on the improbability of their occurrence under the working assumptions.
[†] A significance level of 5% is quite common is statistical testing.

COVARIANCE REALISM

For the forecasted position covariance to be considered "realistic", the mean error should be close to zero (unbiased) and the error spread in all directions should be consistent with the covariance volume. Because tests of bias and scale are most powerful when distributional assumptions hold, the condition of normality should be satisfied foremost. Thus, it is desirable to test at least three hypotheses to assess covariance realism:

1. whether the distribution of predicted satellite location tends to be normal,
2. whether the mean error of the predicted satellite location tends to be zero, and
3. whether the spread of the error in predicted satellite locations is consistent with the predicted covariance.

TESTS OF NORMALITY BASED ON EMPIRICAL DISTRIBUTION FUNCTIONS

Although there are many tests for normality, the Kolmogorov-Smirnov D statistic (or, *KS test*) seems to be commonly used for testing the normality of predicted orbital deviations.[19] Foster and Frisbee (1998) used the KS test to assess the normality of predicted orbit errors, and Jones and Beckerman use the same test for a similar purpose.[8, 9] Sinclair *et al*. suggests using the KS test of normality to gauge the nonlinearity of estimation systems, citing orbit determination as the case of interest.[20]

Kolmogorov-Smirnov Test-of-Fit

The Kolmogorov-Smirnov D_n statistic belongs to a wider class of test statistics based on the empirical distribution function (EDF). The statistic measures the discrepancy between a continuous distribution function $F(x)$ and a supposed estimate $F_n(x)$ based on a sample of size n.[21] A benefit of the KS test is that it allows the construction of error bounds about a distribution function, and thereby lends itself to graphical methods of testing distributional assumptions which are easy to apply and interpret. For example, the critical value D_n may be added and subtracted from every value of a distribution to form a set of error bounds; should the sample cumulative distribution of size n remain with in these error bounds, then the analyst accepts the hypothesis that the empirical distribution came from the theoretical distribution.

Another potential advantage is that the KS test is "nonparametric" This means that the validity of the test statistic is not limited to a particular distribution being tested. Critical values of the KS test may therefore be used to compare any empirical sample against ***any*** theoretical distribution, not just the normal distribution. However, this flexibility turns out to be a significant shortcoming for the KS test, because nonparametric tests are notorious for lacking statistical power.[22] Again, lack of power implies that the test will unreasonably favor the null hypothesis (***e.g.***, the test is more prone to indicate normality even when it is not true).

Another significant disadvantage is that the KS test assumes that the location, scale, and/or shape parameters of the theoretical distribution have not been estimated from the empirical distribution being tested. Comparing an empirical distribution to a normal distribution whose mean and variance equals the sample mean and variance of the empirical distribution is a misapplication of the KS test. Monte Carlo studies have shown that standard critical values of the KS test should be reduced by approximately 50% if the population mean and variance are estimated from the test sample.[23]

To illustrate the lack of power of the KS test, Figure 1 plots the empirical distribution functions of thirteen simulated normal samples, each of size $n = 30$. Included in the figure is the continuous normal distribution function through the center of the data, about which D_n error bounds are drawn and labeled "Lower – KS" and "Upper – KS". From standard statistical tables, we find that the critical value D_n for sample size $n = 30$ and probability level 80% is $D_{30} = 0.190$ (two-tailed).[24, 25] To implement the test, we therefore add 0.190 above and subtract 0.190 below the normal distribution curve. The KS test fails for a given sample if its empirical distribution function crosses either bound (thus the term "two-tailed" test). In this figure, each sample has been normalized by its empirical mean and variance as estimated from its $n = 30$ members.

Lack of power is illustrated in at least two basic ways in Figure 1. First, the upper bound is undefined toward the right-hand portion of the figure, while the lower bound is undefined toward left-hand portion.

Therefore, it becomes practically impossible for the KS test to reject certain abnormal tail behaviors, particularly "heavy" tails. Because Winsor's principle (an empirical analogue of the CLT) suggests that the centers of "real-data" distributions tend to appear Gaussian more often than the tails, this implies that the KS test lacks statistical power where it is most needed - in the tails.[26]

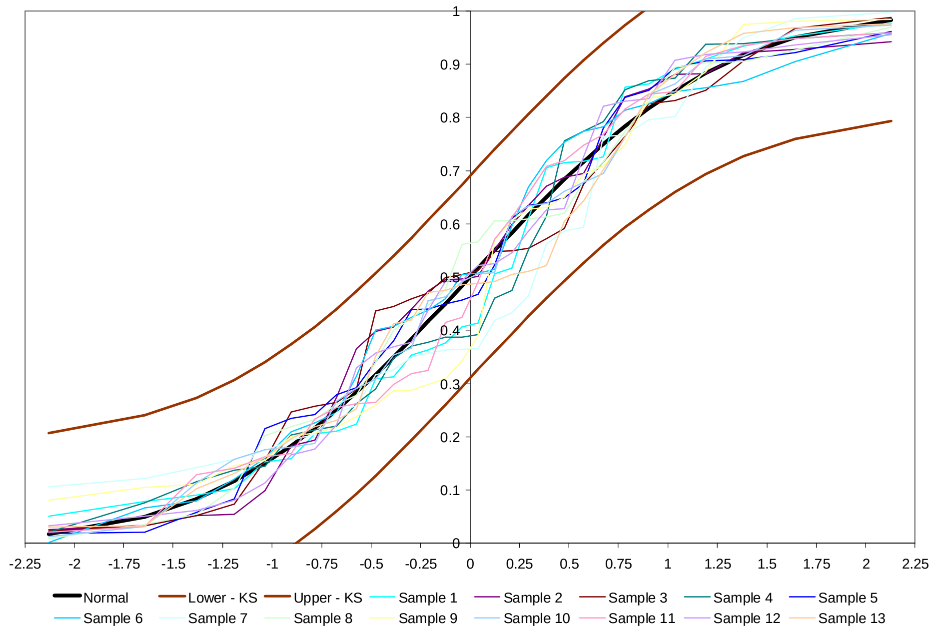

Figure 1. Empirical Distribution Function for Thirteen Normally Distributed Samples. Each sample set contains 30 variates. Also included in this figure are the continuous normal distribution function and the KS-test critical values for probability level 80%**.**

Another illustration of lack of power is that there are no failures in these simulated samples. The probability that thirteen independent samples from the same population would *all* pass a (powerful) statistical test at 80% confidence level is $(0.8)^{13}$, or ~ 5%, a rather improbable outcome. This fact becomes more significant once it is noticed that none of the simulated sample distributions come close to crossing the 80% KS-test error bounds.

Examples of the Kolmogorov-Smirnov Test-of-Fit to Orbital Analyses

Jones and Beckerman (1999) provide a somewhat comprehensive analysis of predicted orbit errors with tracking data from the US space surveillance network.[9] Their sample histograms of predicted-orbit-error estimates (Figure 2) appear somewhat unusual in the sense that longer predictions seem to provide visual evidence of more abnormal behavior, particularly in the radial and transverse (in-track) directions. The transverse error estimates also seem to be rather significant, approaching the one-kilometer level after 36 hours.

Their application of the KS test statistic at the 80% confidence level suggested the normality of every histogram in Figure 2. A 20% significance level was used "to construct conservative tests that reject the null hypothesis as easily as possible" and "confidence limits constructed in this manner provide the greatest opportunity for our a priori notions regarding the distribution of data to be demonstrated incorrect." However, there are some concerns with the conclusions from the ORNL study .

1. The study cites a large-sample approximation for $D_n \approx 1.07/\sqrt{n}$ for testing at 80% confidence level, which is appropriate only if location, scale, and/or shape parameters of the theoretical distribution have *not* been estimated from the empirical distribution being tested. However, the report tests

normalized residual ratios, where the population variances are presumably estimated from the samples. Monte Carlo studies by Lilliefors show that the true 80% confidence level for the KS test is $D_n \approx 0.736/\sqrt{n}$ for normally distributed data.[23] Our extrapolation of Lilliefors' tables suggests that the $1.07/\sqrt{n}$ critical value probably corresponds to only ~0.2% significance level, not 20% as reported. The normality conclusions of the ORNL study are therefore not very significant, because the adopted error bounds used would likely allow for quite a bit of variation from the normal distribution without ever failing the test.

2. Use of a 20% significance level is a highly unusual convention for statistical hypothesis testing, because such a high level of significance tends to cause unreasonable false alarms if the hypothesis were indeed true. The adoption of a high significance level, coupled with a lack of failures, provides compelling evidence that the results are unrealistically good. This, plus the additional fact that the large-sample histograms "look" abnormal, lends credence to the belief that the test used may not have enough power to reject the null hypothesis of normality.

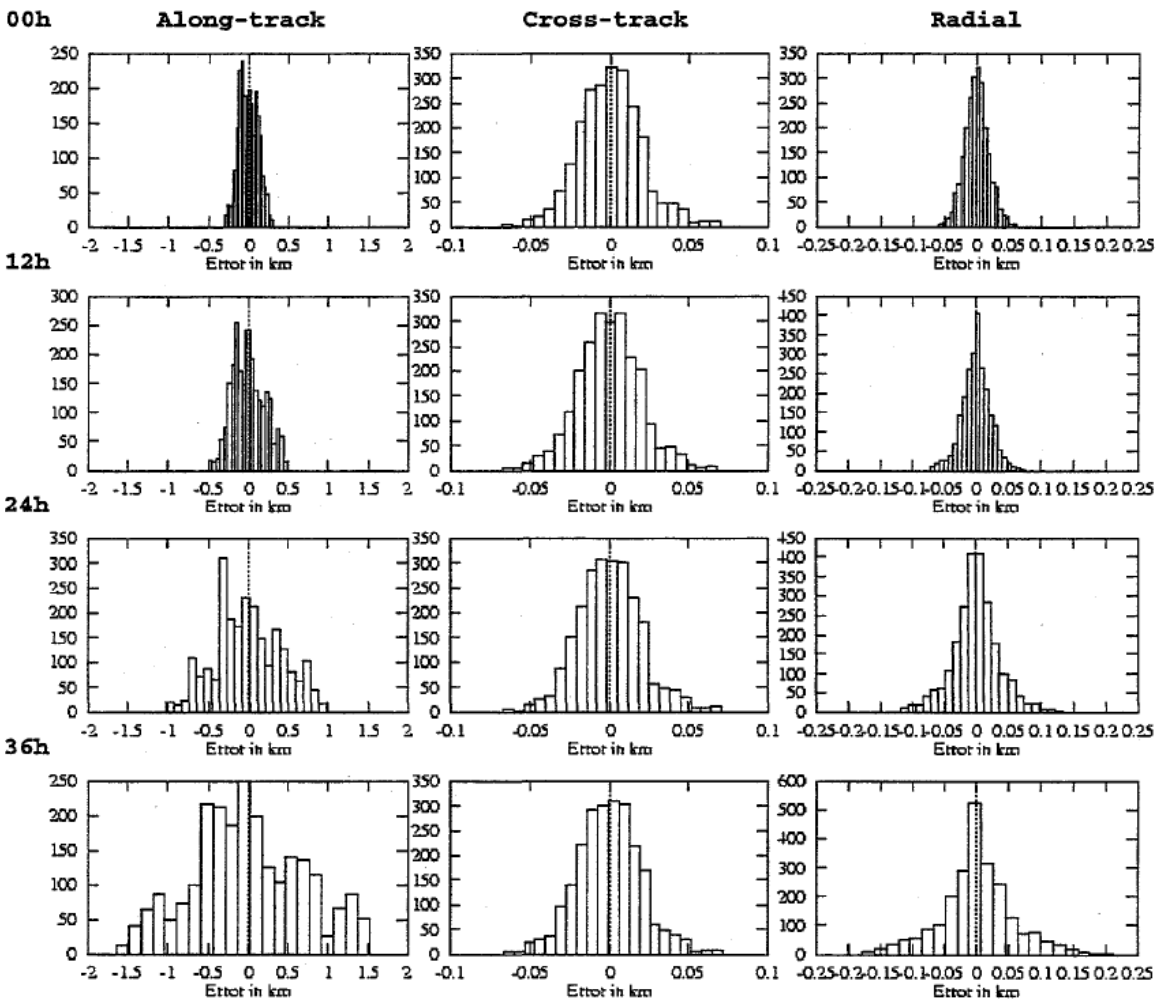

Figure 2. Histograms of Predicted Satellite Position Errors (from Jones and Beckerman, 1999)

KS testing at an 80% confidence level was also used by Foster and Frisbee (1998) for testing the normality of predicted-orbit-error estimates. The means and variances of the theoretical normal distributions were also estimated from the samples being tested, such that all the prior caveats regarding the ORNL testing procedure apply. Of additional note is that a supposed outlier was deleted when estimating the radial, in-track, and cross-track population means and variances; however, the outlier was *not* deleted from the samples tested for normality and the three contaminated empirical distributions still passed the KS test of normality. This described behavior again points to an apparent weakness of the KS test for testing normality and raises the possibility that the null hypothesis of normality might have been rejected had a more powerful test been used.

More Powerful EDF Test-of-Fit for Normality

The KS test likely is employed for orbital analyses because of its ease of use. However, ease of use is not a characteristic of the KS test alone; it generally applies to EDF test-of-fit statistics. Therefore, this paper proposes the use of Michael's D_{SP} test statistic as an alternative to the KS test.[27] Michael's D_{SP} is similar to D_n in that it enables significance limits to be drawn directly onto a normal probability plot instead of the empirical density function plot. An attractive feature of the normal probability plot is that D_{SP} acceptance regions become straight lines on the figure by means of a so-called variance-stabilizing transformation to the EDF test-of-fit statistic. A similar transformation is also applied to the order statistics (sorted data) of the empirical sample to be tested. The transformed data and D_{SP} critical values make up the so-called *stabilized probability plot*.[28]

Functionally, the D_{SP} test is assessed the same way as the KS test: should the plotted data contact or exceed the plotted confidence limits implied by Michael's D_{SP} statistic, the hypothesis of normality is rejected. D_{SP} is more powerful than D_n, particularly in the tails where outliers reside, and is reportedly surpassed in power only by the Shapiro-Wilk W test and the Anderson-Darling A^2 tests of normality.[29] Royston's method for computing D_{SP} critical values was adopted for this study.[30]

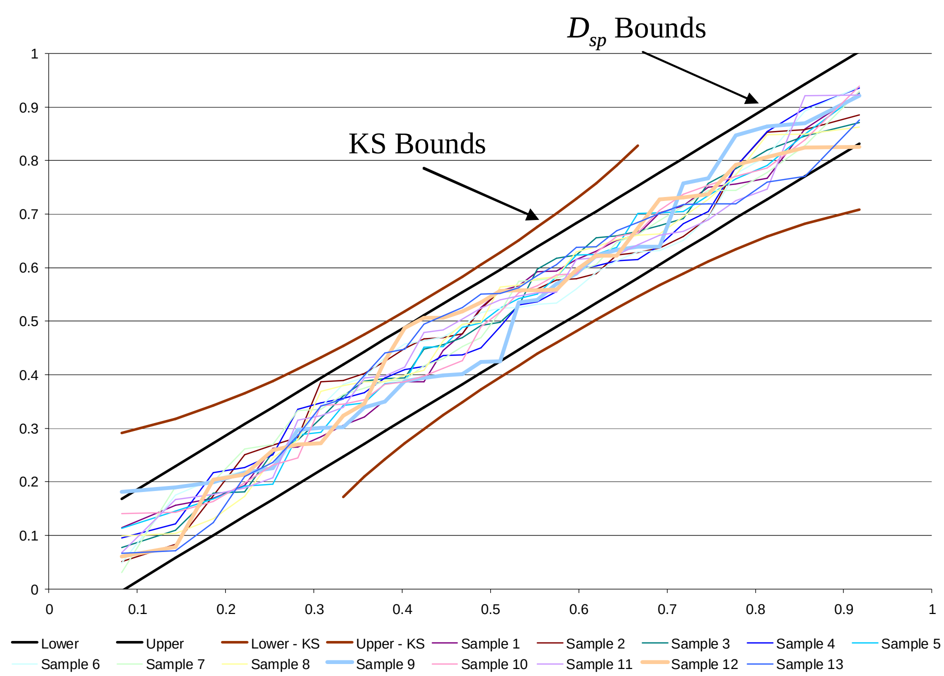

Figure 3. Stabilized Probability Plot. Thirteen samples of size 30 are tested at the 80% confidence level for both Michael's D_{SP} and the KS D_n. Only D_{SP} demonstrates power against the null hypothesis.

An example of the stabilized probability plot is given in Figure 3. Thirteen samples of size 30 were drawn from a normal population and tested at the 80% confidence level, or, 20% significance level, the same level adopted by the ORNL study. The D_{SP} critical values are black lines in the figure labeled "Upper" and "Lower". At this significance level, one would expect a 20% failure rate if the null hypothesis were true, or 2.6 failures out of 13 samples. In this figure, two of the thirteen samples exceeded the critical values, which is not unexpected if D_{SP} were powerful.

To further illustrate the lack of power of the KS test for normality, a variance-stabilizing transformation was also applied to the D_n critical value at the 80% confidence level and labeled "Upper - KS" and "Lower - KS" in Figure 3. Two characteristics seem to confirm what has already been discussed.

1. The distance between the transformed-D_n bounds is much greater than that between the D_{SP} bounds, even though both are supposedly testing at the same significance level. The conclusion is that D_n critical values are optimistic compared to D_{SP}. Comparisons suggest that D_n critical values of 80% confidence correspond to D_{SP} critical values of ~ 98% over the central portion of the distribution.

2. The transformed-D_n critical values curve away from the transformed data near the tails and terminate prematurely, such that there is no correspondence between the transformed-D_n and D_{SP} statistics in the tails, reinforcing the notion that the KS test especially lacks power in the tails of the distribution.

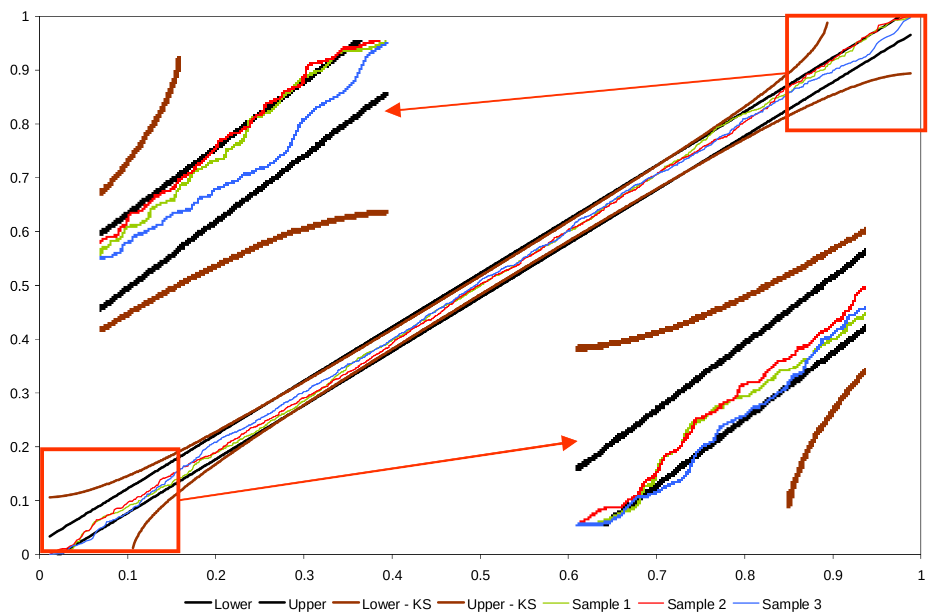

Figure 4. Test of Normality on Contaminated Normal Data. Three sets of 1500 N(0,1) random deviates were contaminated with 75 N(0,2) random deviates and tested for normality. D_{SP} at 5% significance correctly rejected all three samples based on tail behavior (insets), but the KS D_n at 20% significance) did not**.**

Figure 4 carries the comparison of tail power further. Three random samples of size 1500 were drawn from a normal population, and then 5% (75) of these values were replaced with random draws from a normal population having twice the standard deviation (four times the variance). This "5% contaminated normal mixture" is no longer a normal distribution, but has just slightly heavier tails than a regular normal distribution. Two sets of critical values were added to Figure 4: 95% confidence level for Michael's D_{SP} test, and 80% confidence level for the KS test. The insets of Figure 4 magnify the tail behavior of the stabilized probability plot, showing that all three contaminated distributions were rejected by the D_{SP} test at 95%. The probability of three accidental failures in a row at 5% significance is $(0.05)^3$ ~ 0.013%, an extremely rare outcome if normality were indeed true. Therefore, D_{SP} seemingly has power to reject slight distributional abnormalities where D_n cannot.

One concern not addressed by Michael's original paper is whether D_{SP} maintains its power when the mean and variance are estimated from the empirical distribution. The authors therefore conducted a small study that repeated the 80% confidence test of Figure 3 where the population means and variances were estimated from the sample distributions. After testing 78 samples of size 30 at 20% significance, we found the rejection rate to be 19.2%, leading us to conclude that Michael's D_{SP} seems to have excellent power even when the sample size is relatively small and the mean and variance are initially unknown.

NORMALITY-BASED STATISTICAL TESTS FOR COVARIANCE ASSESSMENT

Multivariate normality tends to be difficult to test; however, a sample that is truly multivariate normal will also be marginally normal in all dimensions. Therefore, it is common to assess the univariate normality of all components in order justify the assumption of multivariate normality.[31]

As it relates to testing covariance realism, one expects propagation errors to have a mean of zero, and the ratio of sample variance to population variance to be unity (where population variance comes from the diagonal elements of the covariance matrix). If the assumption of normality is not rejected for a sample, it is reasonable to perform additional statistical hypothesis tests on the sample that assume normality.

One Sample T-Test of the Equal Means

A test of equal means determines whether the value of the mean of a sample distribution is significantly far away from an independently assumed value (*i.e.*, a "one-sample" test).[25] The ***t***-test of equal means is reasonably powerful under the assumption of normality and is often used for this purpose when the sample data are normal and the population variance is estimated from the sample.

Chi-Square Test of Equal Variances

A test of equal variances determines whether the value of the mean of a sample distribution is significantly far away from an independently assumed value (*i.e*., a "one-sample" test).[25] This is equal to testing whether the ratio of the sample variance over the assumed variance is significantly far from unity. The chi-square test of equal variances (two-tailed test) is reasonably powerful under the assumption of normality and is often used for this purpose when the data are normal.

Filter-Smoother Consistency Test

The *filter-smoother consistency test* is useful for model validation in estimation problems, and basically serves as a type of goodness-of-fit test for orbit determination. McReynolds (1984) proved that the difference between a filtered state and a smoothed state is normally distributed in k dimensions, where k is the size of the state-difference vector.[32] He also showed that the variances and correlations of the state-difference vector are equal to the filter error-covariance minus the smoother error-covariance. This leads to the following theorem and test statistic.[33].

Filter-Smoother Consistency Theorem. Let the array $\boldsymbol{x}_f(t)$ be an $n \times 1$ filtered estimate at time t having $k \times k$ error-covariance $\boldsymbol{P}_f(t)$, and let the array $\boldsymbol{x}_s(t)$ be its smoothed estimate at time t having error-covariance $\boldsymbol{P}_s(t)$. Then, assuming the state-estimate errors of $\boldsymbol{x}_f(t)$ and $\boldsymbol{x}_s(t)$ are multivariate normal:

- The $k \times 1$ statistic $\Delta\boldsymbol{x}_{(f\text{-}s)}(t) = \boldsymbol{x}_s(t) - \boldsymbol{x}_f(t)$ is multivariate normal at time t, and has $k \times k$ covariance $\Delta\boldsymbol{P}_{(f\text{-}s)}(t) = \boldsymbol{P}_f(t) - \boldsymbol{P}_s(t)$.
- The time sequence of $\boldsymbol{z}_{(f\text{-}s)}(t) = [\Delta\boldsymbol{x}_{(f\text{-}s)}(t)]^{\mathrm{T}}[\boldsymbol{P}_{(f\text{-}s)}(t)]^{-1}[\Delta\boldsymbol{x}_{(f\text{-}s)}(t)]$, $t = \{t_0, t_1, t_2 \ldots\}$ provides an (auto-correlated) sample population over the estimation interval upon which the null hypothesis of multivariate normality can be tested.

Filter-Smoother Consistency Test. If the sequence of $\boldsymbol{z}_{(f\text{-}s)}(t)$ supports the null hypothesis of multivariate normality, then the hypothesis of consistency between the filter and smoother models is accepted. If the sequence $\boldsymbol{z}_{(f\text{-}s)}(t)$ does not support the null hypothesis of multivariate normality, then the hypothesis of consistency between the filter and smoother models is rejected.

There are at least two difficulties in accessing the test statistic $\boldsymbol{z}_{(f\text{-}s)}(t)$. First, $\boldsymbol{z}_{(f\text{-}s)}(t)$ is multivariate normal, which is harder to test than a univariate normal statistic. More critically, $\boldsymbol{z}_{(f\text{-}s)}(t)$ is not independently distributed, but is strongly correlated; therefore, any statistical test assuming the independence of $\boldsymbol{z}_{(f\text{-}s)}(t)$ will tend to fail.

In practice, the normality of $\boldsymbol{z}_{(f\text{-}s)}(t)$ is accessed in a very heuristic way that nevertheless seems rather effective. First $\boldsymbol{z}_{(f\text{-}s)}(t)$ is replaced by a subset of its k univariate components:

$$\Delta x_{(f\text{-}s)} \ / \ \sigma_{(f\text{-}s)} = (x_{\text{filter}} - x_{\text{smoother}}) \ / \ (\sigma_{\text{filter}} - \sigma_{\text{smoother}}) \tag{4}$$

where x is the parameter estimate and σ is the element of the covariance corresponding to that x. Usually the parameter x is the radial, transverse (in-track), and cross-track components of the Cartesian position difference. Next, a time series of the univariate filter-smoother consistency test statistic $\Delta x_{(f\text{-}s)} \ / \ \sigma_{(f\text{-}s)}$ is plotted and examined by an analyst. Filter-smoother consistency is claimed when the scatter of this metric stays within ± 3 over the fit interval (Figure 5). If the spread seems too large or too small, then the multivariate normality of $\boldsymbol{z}_{(f\text{-}s)}(t)$ must be questioned. Note that normality of the filter estimate is one of the presumptions; if $\boldsymbol{z}_{(f\text{-}s)}(t)$ is considered normal, then there is no evidence to question the normality of the parameter estimates or the general correctness of the scale of the filter covariance.

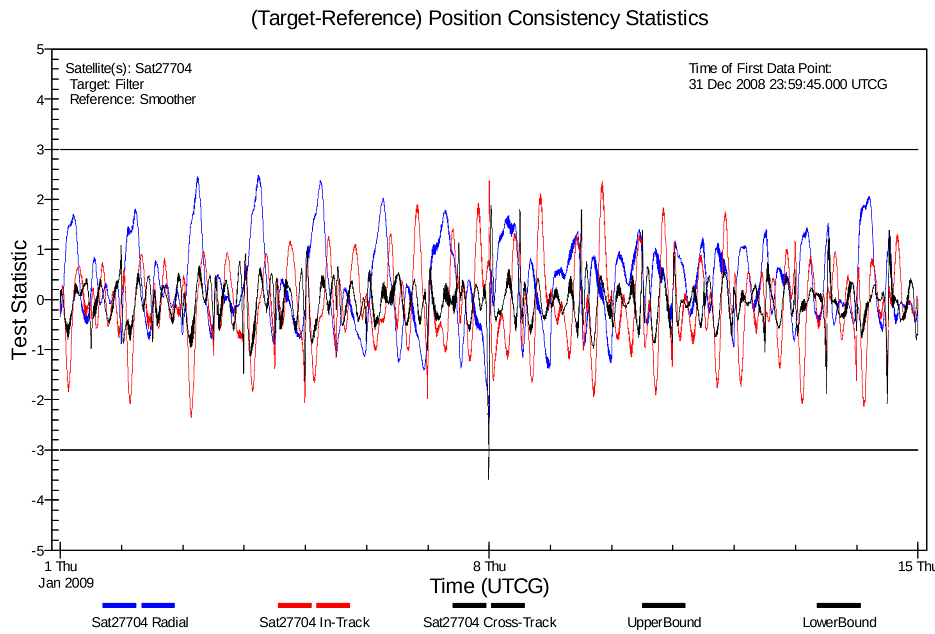

Figure 5. GPS Satellite 27704 Filter-Smoother Consistency Test Statistic.

AN INVESTIGATION OF COVARIANCE REALISM USING NON-SIMULATED DATA

The overall approach for examining results based on actual tracking data follows essential elements of a previous paper which serves as a starting point for some of the analyses in this paper.[34] Satellites were first grouped in major orbital populations; only the results of the GPS (MEO) satellites are presented in this paper. The GPS satellites have long, precise ephemerides which made testing quite simple. Independently generated reference orbits, in the form of Precision Orbit Ephemerides (POE's), provide an excellent means by which to test the accuracy of propagated results. The reference orbits are considered to be accurate to within ~10 cm or better; for the analysis of prediction error they are considered "truth".

The analyses used Analytical Graphics, Inc. Orbit Determination Toolkit (ODTK).[14] Each GPS satellite required the estimation of additional parameters solar radiation pressure coefficients. Measurement-residual ratios were examined to determine the variability of the input data, and plots of estimated position uncertainty were examined to understand of how close the filter matched either the observations or the input ephemeris. Filter-smoother consistency tests were used to determine whether the solution was adequately parameterized. Some studies have developed extensive algorithms to test and ensure the fit span is adequate, but because the underlying technique for ODTK is a real-time filter, there is no fit-span. Thus, no investigation for "optimal" fit span was required.

The propagation span for each ephemeris was kept to fourteen days. This excessive prediction span was done out of thoroughness (owner-operators MAY make decisions about four to seven days in advance). Differences between the prediction and the "truth" ephemerides are computed at several times along the prediction span to provide the analyst with time-varying trends.

Experimental Parameters

The population of GPS satellites permitted an analysis at about thirty satellites. Some of the following software settings were used for this study:

- Satellite mass = 1100 kg
- 70×70 EGM-96 gravity
  - Variational Equations of Degree 8
  - Solid and time dependant tides
- Sun and Moon third-body perturbations
- Solar radiation pressure(SRP) ROCK model with dual cone shadow-boundary mitigation
  - Solve-for Solar radiation pressure scale and *y*-bias coefficients
  - Parameter half-life of 240 min
- RK 7/8 Integrator, relative error of 1×10^{-15}, 1-360 sec step sizes
- Additional radial velocity sigma process noise = 0.0001 cm/s
- Initial R/I/C uncertainties: 5 / 10 / 2 m position and 0.006 / 0.004 / 0.002 m/s velocity

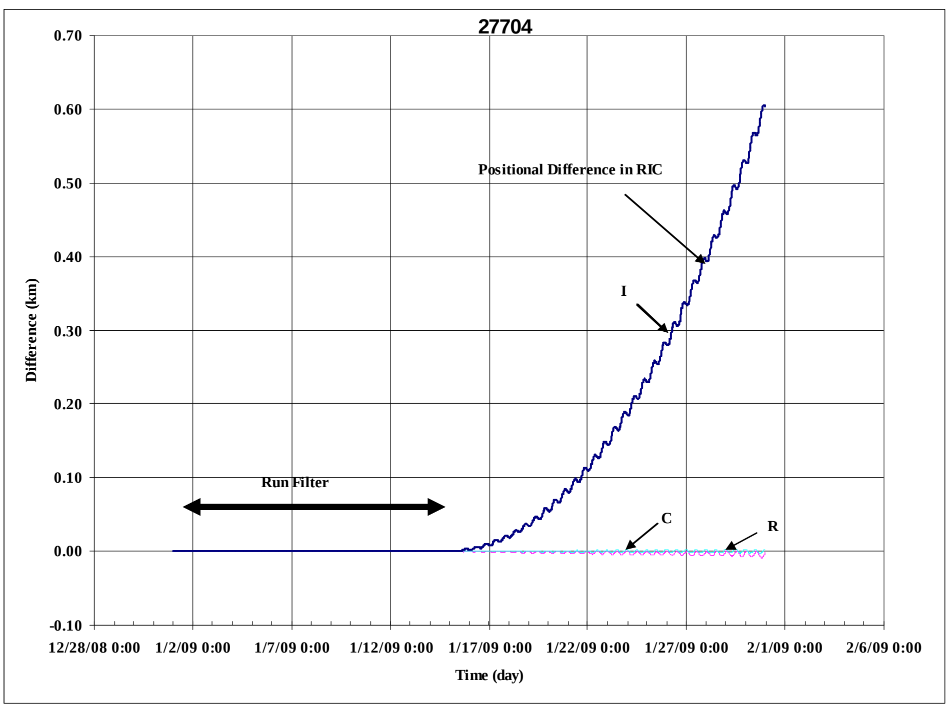

Figure 6. GPS Satellite 27704 Positional Error Estimate. A positional error estimate is found by differencing the precision orbit ephemeris and the predicted ephemeris after the filter has run.

The filter fit fourteen days of POE ephemeris positions as observations. Using this general setup, and iterating on the SRP scale and *y*-bias values, all the satellites showed very similar results. Satellite 27704 is shown as an example. The filter-smoother consistency test statistic in three directions was almost completely within ±3; this is considered as being a sufficient outcome for this test (Figure 5).

To create a population of position error estimates, the final filter state prediction is forecast for fourteen days and differenced from the GPS POE throughout the prediction interval (Figure 6). As a point of comparison, the one-sigma covariance elements generally demonstrate the same growth pattern as the filter propagation, growing to about 400 m at the end of the analysis interval considered in Figure 6. These error estimates were computed across the available satellite population to create a sample of thirty orbital error estimates at various forecasts spans.

Test of Normality

The first goal was to test the normality of prediction error samples, because other tests are based on this assumption. Thirty GPS satellites were included in the samples of the initial analysis. The results for the test of normality are shown in Figure 7, Figure 8, and Figure 9. Departures from a straight line through the center of the bounds indicate a lack of fit of the data relative to a normal distribution. Lack of fit was considered excessive if a given sample distribution exceeded 95% confidence / 5% significance critical values of the D_{SP} test. The various populations represent estimated prediction error sampled at various prediction times, *e.g.*, 3 hours, 6 hours, *etc.*

The lack of normality in these tests prompted us to reexamine our general filtering diagnostics. We determined that the filter-smoother consistency test was excessively irregular across our population of fitted ephemerides, which caused us to refine the parameterization of our orbit-determination runs. Ultimately, a small amount of process noise was added in the radial direction, and the individual scale and *y*-bias coefficients for solar radiation pressure were iterated additional times for each satellite. After limiting ourselves to the twenty satellites that best demonstrated filter-smoother consistency (for the sake of time), the normality of resulting sample distributions improved greatly, as shown in Figure 10, Figure 11, and Figure 12.

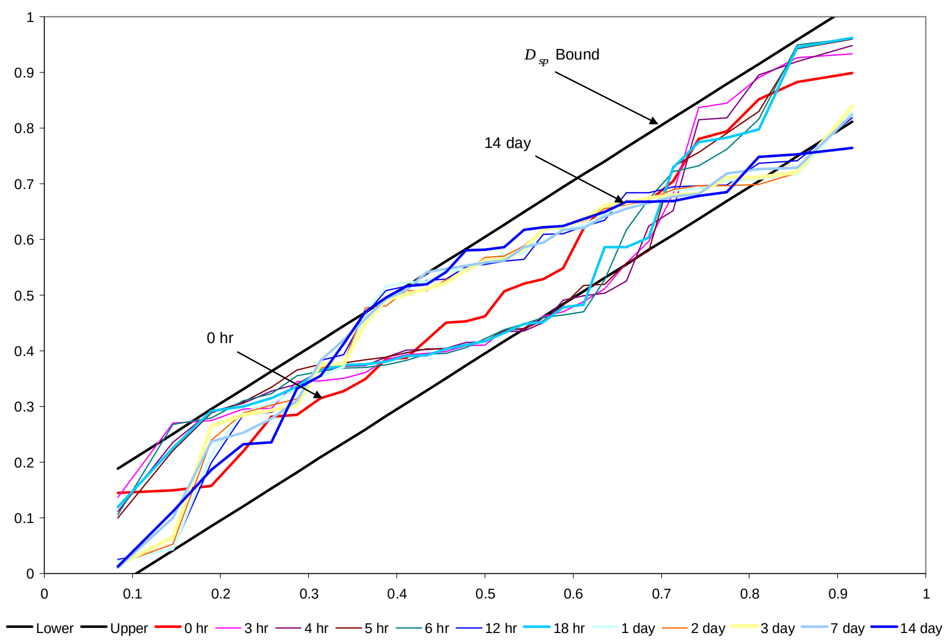

Figure 7. Stabilized-Probability-Plot Test of Normality for Radial Orbit Errors. Based on a sample of thirty satellites with poor filter-smoother consistency.

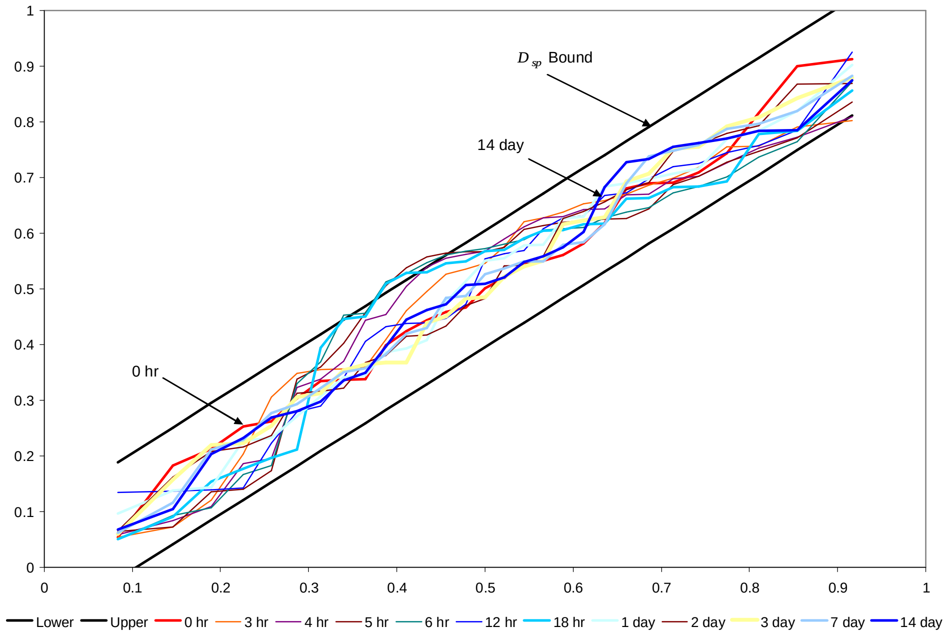

Figure 8. Stabilized-Probability-Plot Test of Normality for Transverse (In-track) Orbit Errors. Based on a sample of thirty satellites with poor filter-smoother consistency.

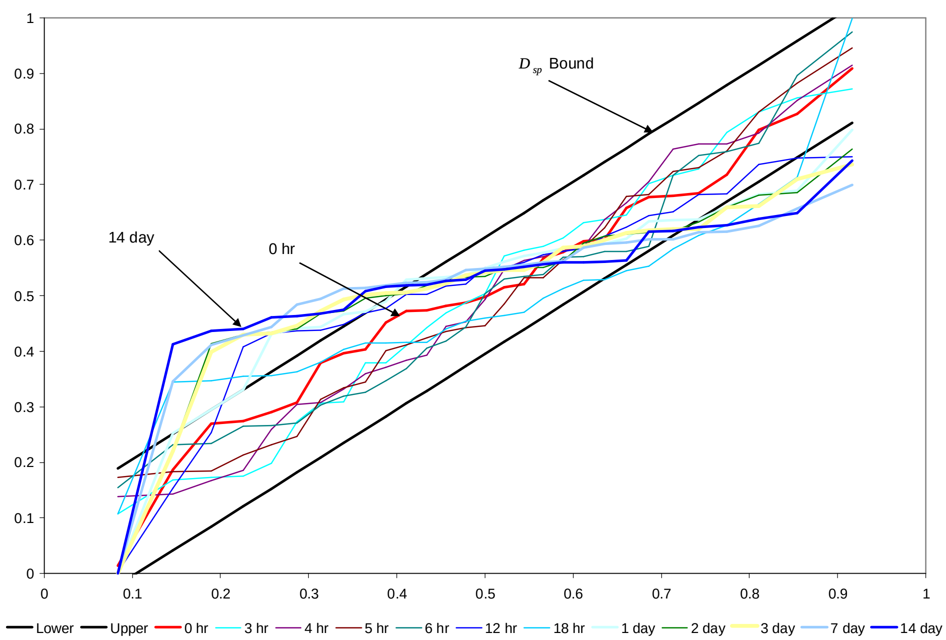

Figure 9. Stabilized-Probability-Plot Test of Normality for Cross-Track Orbit Errors. Based on a sample of thirty satellites with poor filter-smoother consistency.

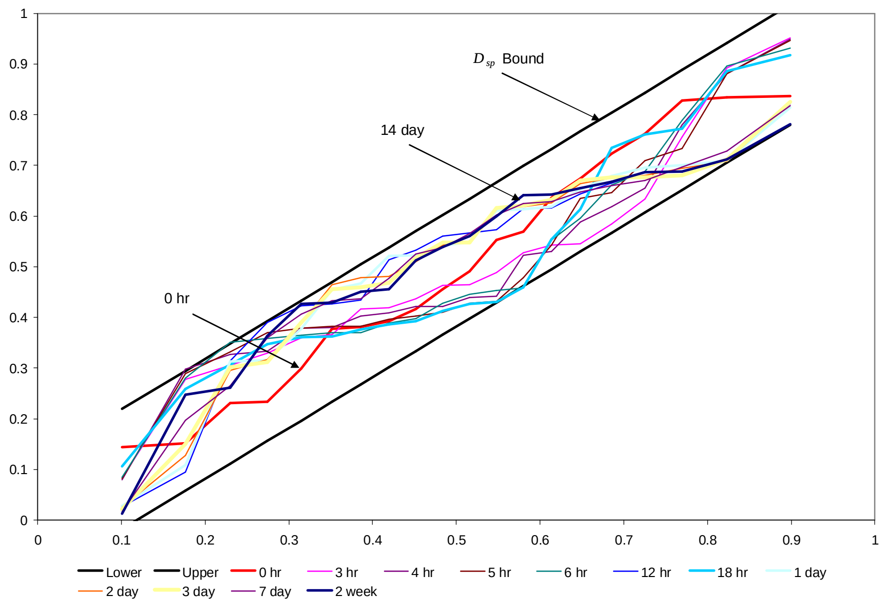

Figure 10. Stabilized-Probability-Plot Test of Normality for Radial Orbit Errors. Based on a sub-sample of twenty satellites with excellent filter-smoother consistency.

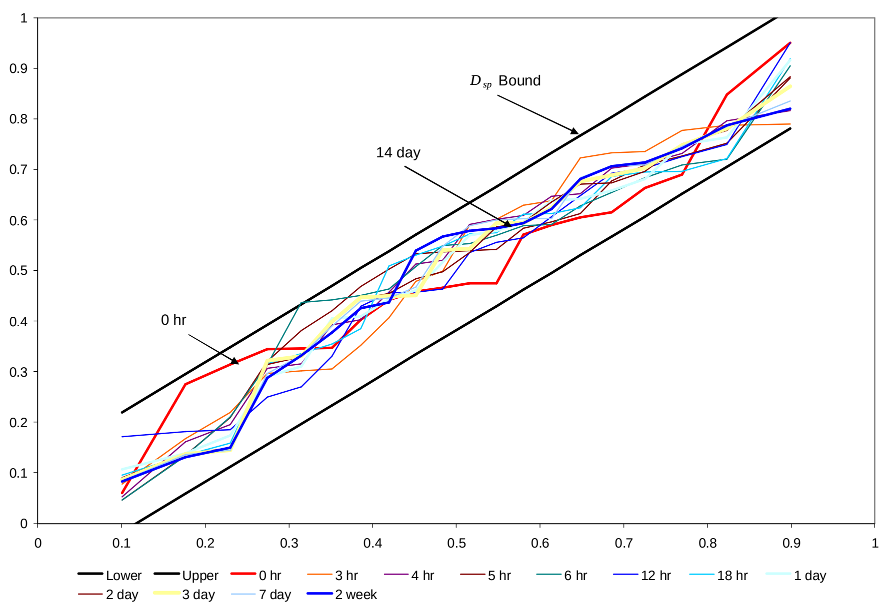

Figure 11. Stabilized-Probability-Plot Test of Normality for Transverse (In-track) Orbit Errors. Based on a sub-sample of twenty satellites with excellent filter-smoother consistency.

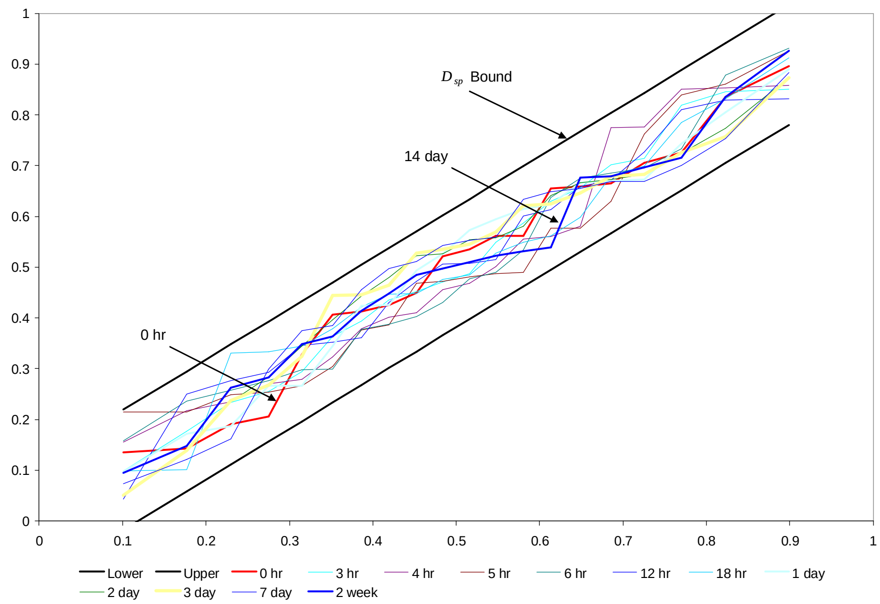

Figure 12. Stabilized-Probability-Plot Test of Normality for Cross-Track Orbit Errors. Based on a sub-sample of twenty satellites with excellent filter-smoother consistency.

Tests of Bias and Scale

With tests of normality generally satisfied for a sub-sample of the original population, we used the *t*-test for the equality of mean and chi-squared test for the equality of variance to determine if the location and scale of the covariance elements were correctly scaled and unbiased. We tested at both the 95% and 99% confidence levels, seeing that a larger-than-expected number of radial samples were marginally rejected at the 95% level.

Radial Component. Results of the estimated errors for the radial component are listed in Table 1. Generally speaking, the test results indicate that the mean is significantly biased and the variance scale generally becomes too large for the GPS samples under consideration after just a few hours.

Transverse (In-Track) Component. Results of the estimated errors for the in-track component are listed in Table 2. Generally speaking, the test results indicate that the mean is not significantly biased and the variance ratio changes significantly from unity for the GPS samples under consideration, although seemingly the change is not as great as what was experienced in the radial and cross-track directions.

Cross-Track Component. Results of the estimated errors for the cross-track component are listed in Table 3. Generally speaking, we conclude that the mean is not significantly biased yet the variance scale changes significantly from unity for the GPS samples under consideration.

Table 1. Radial Deviations Compared to Truth Ephemeris

Prediction interval	Mean Error	σ Ratio	D_{SP} of Normality		t-test of Mean		ψ^2 test of σ	
R	(m)	(m)	95%	99%	95%	99%	95%	99%
0 hr	1.28	0.97	+	+	×	×	+	+
1 hr	0.43	0.96	+	+	+	+	+	+
2 hr	0.40	1.40	+	+	+	+	×	+
3 hr	0.55	1.84	+	+	×	+	×	×
4 hr	0.76	2.17	×	+	×	×	×	×
5 hr	0.93	2.28	+	+	×	×	×	×
6 hr	0.99	2.15	×	+	×	×	×	×
12 hr	-0.42	0.91	+	+	+	+	+	+
18 hr	0.81	1.75	×	+	×	×	×	×
1 day	-0.60	1.21	+	+	×	+	+	+
2 day	-0.87	1.62	+	+	×	×	×	×
3 day	-1.10	1.91	+	+	×	×	×	×
4 day	-1.29	2.17	+	+	×	×	×	×
5 day	-1.48	2.39	+	+	×	×	×	×
7 day	-1.80	2.77	+	+	×	×	×	×
14 day	-2.60	3.43	+	+	×	×	×	×

Table 2. Transverse (In-track) Deviations Compared to Truth Ephemeris

Prediction interval	Mean Error	σ Ratio	D_{SP} of Normality		t-test of Mean		ψ^2 test of σ	
I	(m)	(m)	95%	99%	95%	99%	95%	99%
0 hr	0.91	1.65	+	+	×	×	×	×
1 hr	0.91	1.35	+	+	×	×	×	+
2 hr	0.62	1.00	+	+	×	+	+	+
3 hr	0.37	1.10	+	+	+	+	+	+
4 hr	0.15	1.44	+	+	+	+	×	×
5 hr	-0.09	1.76	+	+	+	+	×	×
6 hr	-0.34	2.01	×	+	+	+	×	×
12 hr	-0.33	1.09	+	+	+	+	+	+
18 hr	-0.22	1.13	+	+	+	+	+	+
1 day	-0.32	1.03	+	+	+	+	+	+
2 day	-0.24	1.08	+	+	+	+	+	+
3 day	-0.21	1.18	+	+	+	+	+	+
4 day	-0.19	1.28	+	+	+	+	+	+
5 day	-0.17	1.37	+	+	+	+	×	+
7 day	-0.14	1.55	+	+	+	+	×	×
14 day	-0.03	2.04	+	+	+	+	×	×

Table 3. Cross-track Deviations Compared to Truth Ephemeris

Prediction interval	Mean Error	σ Ratio	D_{SP} of Normality		t-test of Mean		ψ^2 test of σ	
C	(m)	(m)	95%	99%	95%	99%	95%	99%
0 hr	0.06	2.29	+	+	+	+	×	×
1 hr	0.10	2.39	+	+	+	+	×	×
2 hr	0.12	2.42	+	+	+	+	×	×
3 hr	0.16	2.39	+	+	+	+	×	×
4 hr	0.20	2.42	+	+	+	+	×	×
5 hr	0.04	2.38	+	+	+	+	×	×
6 hr	-0.20	1.88	+	+	+	+	×	×
12 hr	-0.24	0.85	+	+	+	+	+	+
18 hr	-0.04	1.08	+	+	+	+	+	+
1 day	-0.29	1.09	+	+	+	+	+	+
2 day	-0.24	1.31	+	+	+	+	+	+
3 day	-0.30	1.59	+	+	+	+	×	×
4 day	-0.39	1.68	+	+	+	+	×	×
5 day	-0.43	1.79	+	+	+	+	×	×
7 day	-0.45	1.93	+	+	+	+	×	×
14 day	0.28	2.81	+	+	+	+	×	×

CONCLUDING OBSERVATIONS

The authors have proposed that, for the forecast position covariance to be considered "realistic" (accurate), the mean error should be close to zero (unbiased), the error spread in all directions should be consistent with the covariance volume, and the errors should be normally distributed. These three characteristics can be evaluated using statistical hypothesis tests. However, the ability of sample data to pass a statistical test of normality does not necessarily mean the data are normally distributed *if the statistical test lacks power*.

In this study, the authors have noted that the Kolmogorov-Smirnov *D* test statistic (*KS test*), which has been previously used to support the hypothesis of the normality of orbit errors, lacks statistical power against abnormality in the tails and is not suited for testing normality when the true population means and variances are initially unknown. We are therefore unsure that prior studies relying on the KS test reach the proper conclusions about the normality of orbit errors. However, the stabilized probability plot is a more powerful test than the KS test and offers the same advantages in terms of ease of use and graphical interpretation; therefore, it can be recommended in place of the KS test in analysis situations that require a graphical presentation of test outcomes.

The ability of our small-sample satellite population to demonstrate a tendency toward normal-error propagations appeared to be somewhat correlated with the quality of the orbit determination as assessed using the filter-smoother consistency test statistic. Our experience suggested that normality of predictions may be sensitive to incorrect scaling of this statistic. In situations where the filter-smoother consistency test statistic greatly exceeded a ±3 limit, solutions especially seemed to demonstrate a lack-of-fit that ultimately affected the normality of the error predictions We tentatively conclude that rather high quality OD methods and tracking data may be necessary to be confident that predicted orbital error estimates will tend to be normally distributed.

Even with normality testing satisfied, the covariance scale may be too large or the mean error may be biased. This is not a new conclusion, and the covariance scale problem may be improved by more attention to satisfying the filter-smoother consistency test statistic. Due to constraints of time, we were also unable to process additional satellite classes, nor assess the impact of including the process noise in the covariance propagation, but more work is planned in these areas.

ACKNOWLEDGMENTS

The authors are grateful to Dr. Paul Schumacher for some helpful discussions regarding this topic.

REFERENCES

[1] Kelso, T. S., S. Alfano (2005), "Satellite Orbital Conjunction Reports Assessing Threatening Encounters in Space (SOCRATES)." Paper AAS 05-124 Proceedings of the AAS/AIAA Space Flight Mechanics Conference, Copper Mountain, CO.

[2] Johnson, N.J., S. Kotz, N. Balakrishnan (1994), *Continuous Univariate Distributions, Vol. 1, 2nd ed.* John Wiley & Sons, New York. p. 91.

[3] Junkins, J.L., M.R. Akella, K.T. Alfriend (1996), "Non-Gaussian Error Propagation in Orbital Mechanics." Journal of the Astronautical Sciences. Vol. 44, No. 4. pp. 541-563

[4] Alfriend, K.T., M.R. Akella, J. Frisbee, J.L. Foster, D.-J. Lee, M. Wilkins (1998). "Probability of Collision Error Analysis." Paper AIAA 98-4279, Proceedings of the AIAA/AAS Astrodynamics Specialist Conference, Boston, Mass., August 10-12.

[5] Majji, M.T. (2008), "Updated J^{th} Moment Extended Kalman Filtering for Estimation of Nonlinear Dynamic Systems." Paper AIAA 2008-7386, Proceedings of at the AAS/AIAA Astrodynamics Specialist Conference, Honolulu, HI.

[6] Hill, Keric (2008), "Covariance-Based Uncorrelated Track Association." Paper AIAA-2008-7211 Proceedings of the AAS/AIAA Astrodynamics Specialist Conference. Honolulu, HI.

[7] Majji, Manoranjan, James D. Turner, and John L. Junkins (2008), "Higher Order Methods for Estimation of Dynamic Systems, Part I : Theory." Paper AAS 08-162, Proceedings of the AAS/AIAA Space Flight Mechanics Conference. Galveston, TX.

[8] Foster, J.L. Jr., J.C. Frisbee (1998), "Position Error Covariance Matrix Scaling Factors for Early Operational ISS Debris Avoidance." NASA, Johnson Space Center/DM33.

[9] Jones, J.P., M. Beckerman (1999), "Analysis of Errors in a Special Perturbations Satellite Orbit Propagator." Oak Ridge National Laboratory Technical Report ORNL/TM-13726, Oak Ridge, Tennessee. 1 Feb 1999.

[10] Vallado, D.A. (2007), *Fundamentals of Astrodynamics and Applications*. 3rd Edition. Springer/Microcosm, Hawthorne, CA.

[11] Wright, J.R. (1981). Sequential Orbit Determination with Auto-Correlated Gravity Modeling Errors. AIAA Journal of Guidance and Control. 4(2): 304.

[12] Wright, J.R. (1994). "Orbit Determination Solution to Non-Markov Gravity Error Problem." Paper AAS-94-176 presented at the AAS/AIAA Spaceflight Mechanics Meeting. Cocoa Beach, FL.

[13] Wright, J.R. (1994), "Analytical Expressions for Orbit Covariance Due to Gravity Errors." Paper AAS-94-3722 Proceedings of the AIAA/AAS Astrodynamics Specialist Conference. Scottsdale, AZ.

[14] Hujsak, R.S., Woodburn, J.W., Seago, J. H. (2007), "The Orbit Determination Tool Kit (ODTK) – Version 5," AAS 07-125, Proceedings of the AAS/AIAA Space Flight Mechanics Meeting, Sedona, AZ.

[15] Tapley, B.D., B.E. Schutz, G.H. Born (2004), *Statistical Orbit Determination*, Elsevier Academic Press, Burlington, MA.

[16] Crassidis, J.L., and J.L. Junkins (2004), *Optimal Estimation of Dynamic Systems*, CRC Press, Boca Raton, Florida.

[17] Vallado, D.A. (2005), "An Analysis of State Vector Propagation for Differing Flight Dynamics Programs." Paper AAS 05-199 Proceedings of the AAS/AIAA Spaceflight Mechanics Conference, January 23-27. Copper Mountain, CO.

[18] Wright, T.W. (1884), *A Treatise on the Adjustment of Observations by the Method of Least Squares, With Applications To Geodetic Work and Other Measures Of Precision.* Van Nostrand, New York, NY.

[19] Thode, H.C. (2002). *Testing for Normality*. Marcel Dekker, Inc. New York, NY.

[20] A.J. Sinclair, J.E. Hurtado, and J.L. Junkins (2006), "A Nonlinearity Measure for Estimation Systems." Paper AAS 06-135, Proceedings of the AAS/AIAA Spaceflight Mechanics Meeting, Tampa, Florida, January 22-26.

[21] Stevens, M.A. (1983), Kolmogorov–Smirnov–Type Tests of Fit, from Kotz, S. and N.L. Johnson (eds.), Encyclopedia of the Statistical Sciences, Vol. 4, pp. 398-401.

[22] Conover, W.J. (1999), *Practical Nonparametric Statistics*, John Wiley & Sons.

[23] Lilliefors, H. W. (1967), "On the Kolmogorov–Smirnov Test for Normality with Mean and Variance Unknown." *Journal of the American Statistical Association*, Vol. 62, p. 399-401.

[24] Hoel, P.G., (1971), *Introduction to Mathematical Statistics, 4th edition.* John Wiley & Sons, Inc. p. 401.

[25] Sheskin, D.J. (2000), *Handbook of Parametric and Nonparametric Statistical Procedures*, 2nd Edition. Chapman & Hall/CRC.

[26] Seago, J.H., M.A. Davis, W.R. Smith (2005), "Estimating the Error Variance of Space Surveillance Sensors." Paper AAS 05-127, from Vallado *et al.*, Spaceflight Mechanics 2005 - Advances in the Astronautical Sciences. Vol. 120, Part I, pp. 367-386.

[27] Nelson, L.S. (1989), "A Stabilized Normal Probability Plotting Technique." Journal of Quality Technology, Vol. 21, No. 3, July. pp. 213-15.

[28] Michael, J.R. (1983), "The stabilized probability plot.", Biometrika , Vol. 70, No. 1, pp. 11–17.

[29] Stephens, M. A. (1974). "EDF Statistics for Goodness of Fit and Some Comparisons." Journal of the American Statistical Association, Vol. 69, No. 347, pp. 730-737.

[30] Royston, P. (1993). "Graphical detection of non-normality by using Michael's statistic." Applied Statistics, Vol. 42, No. 1, pp. 153-58.

[31] Thode, H.C. (2002). *Testing for Normality*. Marcel Dekker, Inc. New York, NY. p. 181.

[32] McReynolds, S.R. (1984), "Editing Data Using Sequential Smoothing Techniques for Discrete Systems." Paper AIAA-1984-2053, Proceedings of the AIAA/AAS Astrodynamics Conference, Seattle, WA, Aug 20-22, 1984.

[33] Seago, J.H., J.W. Woodburn (2007), "Sensor Calibration as an Application of Optimal Sequential Estimation Toward Maintaining the Space Object Catalog." Paper USR 07-S7.1 Proceedings of the 7th US/Russian Space Surveillance Workshop, Naval Postgraduate School, Monterey, California, October 29-November 2, 2007, p. 309.

[34] Vallado, D.A. (2007), "An Analysis of State Vector Prediction Accuracy." Paper USR 07-S6.1 Proceedings of the 7th US/Russian Space Surveillance Workshop, Naval Postgraduate School, Monterey, California, October 29-November 2, 2007, p. 231.