AAS 10-149

GOODNESS-OF-FIT TESTS FOR SEQUENTIAL ORBIT DETERMINATION

John H. Seago* and David A. Vallado†

Goodness-of-fit tests – sometimes called consistency tests – are useful for investigating the lack of optimality of an estimator. Statistical hypothesis tests involving observation residuals are often recommended; however, the most common diagnostic tests may have limited applicability to general orbit determination. In this paper, some less-familiar test statistics are presented and their usage is modified to apply them to observation residuals that are irregularly spaced with time. The supplemental test statistics are assessed using simulated time-series and sequential-estimation results based on genuine satellite tracking data.

INTRODUCTION

Goodness of fit implies the degree to which an experimental outcome reasonably meets probable expectations. Specifically, sample data should not strongly disagree with the probability law that they are presumed to follow under the status-quo operating condition that is supposed to exist (the so-called *null hypothesis*). Goodness-of-fit techniques therefore involve the testing of statistical hypotheses, which help to make subjective decisions more objective by using a testable "statistic" having a pre-supposed distribution that is usually informative of the null hypothesis.[1]

To conduct a statistical hypothesis test, an analyst chooses *critical values* for a statistic to be tested; appropriate critical values will represent extreme or unlikely outcomes for the statistic's assumed distribution. If the statistical outcome does not fall between the chosen critical values, then the analyst concludes that the assumed operating conditions did not exist (the hypothesis test *fails*); otherwise, he presumes the null hypothesis holds due to a lack of evidence (the hypothesis test *passes*).

Whenever the null hypothesis is true, the probability $\Pr\{\cdot\}$ that a random statistical outcome will end up between the critical values is called the *confidence level* of the test, and the probability of an accidental failure at this critical value is called the *significance level* of the test ($1 - \Pr\{\cdot\}$). For example, if one were testing a sequence of supposedly random outcomes at the $1 - \Pr\{\cdot\} = 1\%$ significance level, a "powerful" statistical test would reject 1% of the outcomes if the null hypothesis were true. Rejection rates much higher than 1% over many outcomes could provide evidence that the operating assumptions are being violated in some way; rejection rates much less than 1% might provide evidence that the test is unable to reasonably reject the null hy-

* Astrodynamics Engineer, Analytical Graphics Inc., 220 Valley Dr., Exton, Pennsylvania, 19341-2380.

† Senior Research Astrodynamicist, Analytical Graphics Inc., Center for Space Standards and Innovation, 7150 Campus Dr., Suite 260, Colorado Springs, Colorado, 80920-6522.

pothesis. It is therefore possible to obtain insight into the validity of a statistical test by repeatedly evaluating simulated outcomes having well understood distributions: in the long run, one expects Pr{·} successes and 1 - Pr{·} failures for the assigned critical value(s).

GOODNESS-OF-FIT (CONSISTENCY) CRITERIA FOR OPTIMAL FILTERS

For sequential orbit determination, the desired estimator is usually the *optimal* one - one that is "best in a certain sense."[2] Regardless of how one might define "best", sources seem to agree that the following criteria are generally necessary for optimal sequential estimation:[3]

(1) State errors should have zero mean and have magnitude commensurate with the state covariance as yielded by the filter.

(2) Predicted residuals should have zero mean and have magnitude commensurate with the residual variances as yielded by the filter.

(3) Predicted residuals should be white (uncorrelated over time).

Bar-Shalom *et al*. state that a filter is *inconsistent* (demonstrates a poor fit to the measurement data) if it does not satisfy the above criteria. Wright suggests more enumerative criteria for his definition of filter optimality specific to the orbit determination problem, specifically adding McReynolds' filter-smoother consistence test.[4] Also, most treatments of the sequential estimation problem presume Gaussian-distributed errors; sufficiently powerful tests of Gaussian distribution (normality) are discussed in an earlier manuscript.[5]

To satisfy Criterion (1), Bar-Shalom *et al*. propose that the normalized mean-estimation error can be tested for normality with mean zero and variance 1/*m* (*m* being number of estimates contributing to the statistic). Unfortunately, the average error can be assessed only through simulation because it requires knowledge of the truth from which it deviates, so this test cannot be used with genuine experimental data where true result is unknown.

Predicted Residuals (Innovations)

Our acknowledgement of tests for state errors and state differences has been mainly for completeness, because the testing of actual tracking residuals will be the primary topic of this manuscript. The following notation (due to Maybeck, 1979) specifies the definition of predicted filter residuals, or *innovations*, which will be tested in the sequel.[6] For a linear estimator such as the Kalman filter, let the measurement update equations be expressed as:

$$\mathbf{K}(t_i) = \mathbf{P}(t_i^-)\mathbf{H}^{\mathrm{T}}(t_i)\left[\mathbf{H}(t_i)\mathbf{P}(t_i^-)\mathbf{H}^{\mathrm{T}}(t_i) + \mathbf{R}(t_i)\right]^{-1} \tag{1}$$

$$\hat{\mathbf{x}}(t_i^+) = \hat{\mathbf{x}}(t_i^-) + \mathbf{K}(t_i)\left[\mathbf{z}_i - \mathbf{H}(t_i)\hat{\mathbf{x}}(t_i^-)\right] \tag{2}$$

$$\mathbf{P}(t_i^+) = \mathbf{P}(t_i^-) - \mathbf{K}(t_i)\mathbf{H}(t_i)\mathbf{P}(t_i^-) \ . \tag{3}$$

where $\hat{\mathbf{x}}(t_i^-)$ is the state (-correction) estimate array, $\mathbf{P}(t_i^-)$ is state-error covariance matrix prior to the measurement update, $\mathbf{K}(t_i)$ is the Kalman filter gain, $\hat{\mathbf{x}}(t_i^+)$ is the state (-correction) estimate array, and $\mathbf{P}(t_i^+)$ is the state-error covariance matrix updated by measurement vector $\mathbf{z}_i$ at time (t_i). The matrix $\mathbf{H}(t_i)$ is realized according to the analytical observation-state relationship at time (t_i), *e.g.*,

$$\mathbf{z}(t_i) = \mathbf{H}(t_i)\mathbf{x}(t_i) + \mathbf{v}(t_i) \;\; ; \;\; E\{\mathbf{v}(t_i)\} = \mathbf{0} \;\; , \; E\{\mathbf{v}(t_i)\mathbf{v}^{\mathrm{T}}(t_j)\} = \begin{cases} \mathbf{R}(t_i); & t_i = t_j \\ \mathbf{0}; & t_i \neq t_j \end{cases} \;\; , \tag{4}$$

where $\mathbf{v}(t_i)$ is an array of white measurement noise having zero expected value and variance $\mathbf{R}(t_i)$. The array of predicted filter residual is defined as the difference between the actual measurement and the best prediction of the measurement just before it is actually taken:

$$\mathbf{r}(t_i^-) = \mathbf{z}_i - \mathbf{H}(t_i)\hat{\mathbf{x}}(t_i^-) \;\; , \tag{5}$$

which has mean and variance:

$$E\{\mathbf{r}(t_i^-)\} = \mathbf{0} \;\; , \;\; E\{\mathbf{r}(t_i^-)\mathbf{r}^{\mathrm{T}}(t_i^-)\} = \mathbf{H}(t_i)\mathbf{P}(t_i^-)\mathbf{H}^{\mathrm{T}}(t_i) + \mathbf{R}(t_i) \tag{6}$$

Because the filter gain is based on the filter-calculated error covariance per Eq. (1), an incorrect covariance yields an incorrect gain. It follows that goodness-of-fit evaluation not only evaluates estimator optimality, but it is closely related to the evaluation of covariance realism.

The Kalman Filter as a Whitening Filter for Residuals

Because the sequence of residuals at times t_i are linear functions of previous measurements $\mathbf{z}_{i-1}$, $\mathbf{z}_{i-2}$, … by definition, each residual at time t_i is independent of all measurements prior to t_i, and each residual $\mathbf{r}(t_i^-)$ is independent of all previous residuals such that the sequence of residuals is *white*. Therefore, in a "truly optimal" filter — one based upon a complete and perfectly calibrated model — the residuals should approximate a white Gaussian sequence, whereas a suboptimal or mis-calibrated filter will exhibit a time-correlated residual sequence.[7] Additionally, the linear filtering problem can be turned around so that it is possible to think of a Kalman filter as a whitening filter for predicted residuals; that is, for the system described by Eqs. (2), (4), and (6), the output $\mathbf{r}(t_i^-)$ of residual sequence will be a *white* process.[8] This justifies Criterion (3) above.

TESTING THE MEAN AND VARIANCE OF RESIDUALS*

Criterion (2) above suggests that predicted residuals should have zero mean and their magnitude should be commensurate with the residual variance of Eq. (6). Each residual divided by the square-root of its variance should therefore provide a sequence that has zero mean and variance of unity. If the sequence is also normally distributed, these properties can be tested with the following statistical hypothesis tests.

Chi-Squared Test of Equal Variances. An unbiased "textbook" estimator of variance for discrete samples X_i at times $\{t_1, t_2, \ldots t_n\}$ is:[9]

$$s^2 \cong s^2 \equiv \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \hat{m}_X\right)^2 \;\; , \{X_i, \, i = \; 1, 2, \ldots n\}, \tag{7}$$

when the mean $\hat{m}_X$ is estimated from the sample:

---

* From this point, discussion focuses on predicted residuals (innovations), or, residuals divided by their square-root variance. We will treat each measurement type independently, and thereby drop the vector notation and represent the scalar time series generically for each measurement type, *e.g.*, $\{X_i; \, i = \; 1, 2, \ldots\}$.

$$\hat{m}_X \equiv \frac{1}{n}\sum_{i=1}^{n} X_i \ , \{X_i, i = 1, 2, \ldots n\}. \tag{8}$$

In this case, the distribution of $(n-1)\, s^2/\sigma^2$ is expected to follow a χ^2-distribution of $n-1$ degrees of freedom.[10] Thus, to satisfy Criterion (2), s^2 should not be considered significantly different than σ^2 at, say, the $\alpha = 1\%$ significance level, if $\chi^2_{1-\alpha/2,h} < (n-1)\, s^2/\sigma^2 < \chi^2_{\alpha/2,h}$, where $\chi^2_{1-\alpha/2,h}$ and $\chi^2_{\alpha/2,h}$ are the critical values of the χ^2 cumulative probability distribution, and $h = n-1$.* When dealing with predicted residual ratios, σ^2 is expected to be unity, and for "large enough" n, the χ^2-distribution approaches a normal distribution.

One Sample t-Test of Equal Means. A *t*-test of equal means determines whether the sample mean $\hat{m}_X$ is significantly far away from a presupposed value (*i.e.*, a "one-sample" test).[11] For normally distributed samples, the distribution of $\hat{m}_X / s^2$ follows a *t*-distribution of $n-1$ degrees of freedom. Thus, to satisfy Criterion (2), $\hat{m}_X$ should not be considered significantly different than zero at, say, the $\alpha = 1\%$ significance level, if $t_{\alpha/2,h} < (n-1)\ \hat{m}_X / s < t_{1-\alpha/2,h}$, where $t_{\alpha/2,h}$ and $t_{1-\alpha/2,h}$ are the critical values of the cumulative *t*-distribution, and $h = n-1$.† For "large enough" n, the *t*-distribution approaches a normal distribution.

ESTIMATORS OF CORRELATION

For theoretically continuous distributions, the *variance* of a random variable X is:

$$s^2 = \mathrm{Var}(X) \equiv E[(X - m_X)(X - m_X)] = \int_{-\infty}^{+\infty} (x - m_X)^2 f(x)\, dx \ , \tag{9}$$

where $\mu_X = E[X]$ is the expected value of the random variable X, and $f(x)$ is the probability density function evaluated for $x = X$.[12] For discretely sampled data, Eq. (7) may be used.

(Cross) Covariance and (Cross) Correlation

The covariance between pairs of random variables X and Y is defined as:

$$\mathrm{Cov}(X,Y) \equiv E[(X - m_X)(Y - m_Y)]. \tag{10}$$

The discrete-sample estimator of covariance analogous to the textbook estimator of Eq. (7) is:[13]

$$\mathrm{Cov}(X,Y) \cong \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \hat{m}_X)(Y_i - \hat{m}_Y) \ . \tag{11}$$

For analyses it is useful to standardize the covariance between two random variables by dividing by the product of their respective standard deviations. This yields the *correlation coefficient*:

---

* The Microsoft Excel spreadsheet function CHIINV(α, h) provides appropriate critical values from the χ^2 distribution. It is also convenient to use the inverse function CHIDIST($(n-1)\, s^2/\sigma^2$, h) and reject the null hypothesis should this value be too small or too large (say, less than 0.5% or greater than 99.5%).

† The Microsoft Excel spreadsheet function TINV(α, h) provides appropriate critical values from the t^2 distribution. It is also convenient to use the inverse function TDIST($(n-1)\, \mu_X / s^2$, h, 2) and reject the null hypothesis should this value be too small or too large.

$$r(X,Y) \equiv \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}} \ , \tag{12}$$

which is a measure of the strength of the linear relationship between two random variables.[14] The correlation coefficient may be conveniently estimated using the Pearson product-moment form:

$$\hat{r}(X,Y) = \frac{\sum_{i=1}^{n} (X_i - \hat{m}_X)(Y_i - \hat{m}_Y)}{\left(\sum_{i=1}^{n} (X_i - \hat{m}_X)^2 \right)^{1/2} \left(\sum_{i=1}^{n} (Y_i - \hat{m}_Y)^2 \right)^{1/2}} \ . \tag{13}$$

This is simply the ratio of the covariance expressed by Eq. (11) over the product of the sample standard deviations s_X and s_Y as expressed by Eq. (7), such that $-1 \le r(X,Y) \le 1$. The distribution of $\hat{r}(X,Y)$ is normal for "very large" n.

Autocorrelation (Serial Correlation)

In the previous equations, if Y_i is replaced with a value of X_{i+k} at a time some fixed lag Δt_k away from t_i, it is known as the *autocorrelation* (or serial correlation) at that interval. Thus, autocorrelation may be defined analogously from Eq. (12) as:

$$r(X_i, X_{i+k}) \equiv r(k) = \frac{\mathrm{Cov}(X_i, X_{i+k})}{\mathrm{Var}(X)} \ . \tag{14}$$

TESTING RESIDUAL WHITENESS USING SERIAL CORRELATION ESTIMATES

Bar-Shalom *et al.* and Crassidis & Junkins suggest a test of whiteness of residuals by estimating the "time-average" sample autocorrelation according to this expression:

$$\hat{r}(k) = \frac{\sum_{i=1}^{n-k} (X_i)(X_{i+k})}{\left(\sum_{i=1}^{n-k} (X_i)^2 \right)^{1/2} \left(\sum_{i=1}^{n-k} (X_{i+k})^2 \right)^{1/2}} \ , \tag{15}$$

where X_i is the residual at time t_i.[15] The number of points used in the numerator of Eq. (15) is the same as that in the denominator, providing a Pearson product-moment correlation limited to a sub-sample of size $n - k$. This estimator is often recommended because, for large enough $n - k$, there is a tendency for $\hat{r}(k)$ to be normally distributed with variance of $1/(n - k)$ whenever the expected value of $\hat{r}(k)$ is zero. Therefore, the hypothesis of residual whiteness should not be rejected at the 1% significance level if $-2.58 < \sqrt{n-k}\ \hat{r}(k) < +2.58$. However, there are a few notable cautions:

- Some authors advise against the use of Eq. (15) "on the grounds that […] it is not a satisfactory estimate when a set of estimates is required for the first m autocorrelations."[16] The demoninator, which intends to serve as a normalizing factor, changes with lag k, which may lead to curious behavior in the estimate of a sample spectrum, possibly making comparisons of $\hat{r}(k)$ problematic for differing values of k.

- The sample size $n - k$ being used to estimate the correlation necessarily decreases as lag k increases. As lag k increases, $n - k$ becomes smaller and the normality assumption requiring "large enough" $n - k$ eventually fails.

Testing Whiteness Using Gaussian Critical Values

To help deal with small samples sizes with increasing lag k, Fisher's variance-stabilizing transformation may be applied to Eq. (15), which is reputed to approach normality much faster than $\hat{r}(k)$. This transformation amounts to taking the hyperbolic tangent of the estimated coefficient:

$$z = \tanh^{-1}\left(\hat{r}(k)\right) = \frac{1}{2}\ln\left(\frac{1+\hat{r}(k)}{1-\hat{r}(k)}\right). \tag{16}$$

Fisher's z statistic has an expected value of $\tanh^{-1}(\rho(k))$ with variance of approximately $(n - k - 3)$. Thus, under the null hypothesis of zero correlation, the hypothesis of whiteness cannot be rejected at the 1% significance level if $-2.58 < \sqrt{n-k-3}\,\tanh^{-1}\left(\hat{r}(k)\right) < +2.58$. Note however, that as $\rho(k)$ approaches zero, $\tanh^{-1}(\rho(k))$ approaches $\rho(k) = 0$, such that the transformation may not be very beneficial under the null hypothesis of zero correlation. It is therefore useful to pursue alternative statistics.

THE SAMPLE SEMI-VARIOGRAM

Let the variance of the k^{th} lag between X_i and X_{i+k} be denoted as $\gamma(k)$.*

$$g(k) \equiv \tfrac{1}{2} E\left[\left(X_{i+k} - X_i\right)^2\right]. \tag{17}$$

Known as the *semi-variogram*,† this function characterizes the second-order dependence properties of a time series alternative estimator of the variance of a time series.[17] The sample estimator of the semi-variogram is:

$$2\hat{g}(k) = \frac{1}{n-k}\sum_{i=1}^{n-k}\left(X_{i+k} - X_i\right)^2. \tag{18}$$

Unlike Eq. (7), the semi-variogram estimator of Eq. (18) beneficially annihilates the series mean because μ_X differences out when X_i is subtracted from X_{i+k}.

Expanding Eq. (17) using algebra of the expectation operator, and considering that $\mathrm{Var}(X_i) = \mathrm{Var}(X_{i+k})$, it can be shown that (*c.f.*, Eqs. (9) and (10)):

$$\gamma(k) = \mathrm{Var}(X_i) - \mathrm{Cov}\,(X_i, X_{i+k}). \tag{19}$$

---

* The k^{th}-lagged semi-variogram is related to the so-called "Allan variance" from precision-timing community, and may also be known by slightly different terminology depending upon the discipline.

† Usage varies surrounding application of the term *variogram*; technically it refers to the quantity $2\gamma(k)$, but it is often convenient to drop the "semi-" prefix when discussing $\gamma(k)$.

Therefore, if a series is white, Cov (X_i, X_{i+k}) = 0 and the sample semi-variogram becomes an alternative estimator for the sample variance. Also, from Equations (14) and (19), it is possible to show that, for stationary processes, the semi-variogram is simply related to the sample autocorrelation function (*i.e.*, correlogram) through the relationship:

$$\hat{r}(k) = 1 - \frac{\hat{g}(k)}{s^2}, \tag{20}$$

where s^2 is the sample variance of X_i from Eq. (7). Thus, the sample semi-variogram and the sample autocorrelation provide the same information for stationary time series, albeit in slightly different forms. Also note that Eq. (20) does not suffer from a normalizing factor that changes with lag k like Eq. (15) does.

WHITENESS TESTING USING THE SAMPLE SEMI-VARIOGRAM

Because the semi-variogram is an alternative estimator of variance under the null hypothesis of whiteness, it seems reasonable to employ a statistical test for the equality of variances using $\hat{g}(k) / s^2$ as a test statistic. There are at least three straightforward, if approximate, approaches that one might take toward this end.

F-Test of Equal Variances Applied to the Semi-Variogram/Variance Ratio

The *F*-distribution may apply to the ratio of two independent χ^2 variables. To pass the *F*-test of equal variances at, say, the 1% significance level, $F_{0.005,(n-k,\, n-1)} < \hat{g}(k) / s^2 < F_{0.995,(n-k,\, n-1)}$.[18] Technically, the *F*-test applies when $\hat{g}(k)$ and s^2 are based on independent samples; however, in our situation $\hat{g}(k)$ is based on a subset of the same values used to estimate s^2; thus, the *F*-test can only be expected to provide approximate results.

Chi-Squared Test of the Semi-Variogram

For normally distributed data, the distribution of $(n-k)\ \hat{g}(k) / \sigma^2$ may be expected to have a χ^2 distribution of $n - k$ degrees of freedom under the null hypothesis of zero correlation. To pass the χ^2-test of equal variances at, say, the 1% significance level, $\chi^2_{0.005,(n-k)} < (n-k)\ \hat{g}(k) < \chi^2_{0.995,(n-k)}$, where $\sigma^2 \equiv 1$ for residual ratios. This approach is exact; however, there is a subtle problem with assuming that $\sigma^2 \equiv 1$ in that it is common practice to automatically reject outlying measurements whenever the magnitude of the residual ratio exceeds some threshold, say ±3. Unfortunately, complete outlier rejection followed by an ordinary analysis of the variance of the remaining data tends to bias the sample variance downward (even when the error-generating distribution is outlier-prone), resulting in a value for s^2 that is smaller than unity on average.[19] When $\sigma^2 \equiv 1$ is not representative of a (censored) sample, it is difficult to distinguish if the χ^2-test is failing because $\sigma^2 \equiv 1$ is an improper assumption for the censored data, or because there are significant serial correlations. (Note that sample correlation estimators like Eq. (13) lack this problem because they are scaled by actual ***sample*** variances rather than ***expected*** variances.)

Chi-Squared Test Applied to the Semi-Variogram/Variance Ratio

A hybrid, or compromise, statistical approach might substitute σ^2 with s^2 while still employing a χ^2-test. Thus, to pass a χ^2-test of equal variances at, say, the 1% significance level, $\chi^2_{0.005,(n-k)} < (n-k)\ \hat{g}(k) / s^2 < \chi^2_{0.995,(n-k)}$. This substitution seems reasonable based on prior arguments, especially when n is large. A disadvantage is that the χ^2-test (in contrast with the *F*-test) does not account for the sample uncertainty in the estimate of s^2 (*i.e.*, the test is valid for small $n - k$, but not

necessarily small n). However, the supposed advantage of the F-test may not be as virtuous as it first appears, because s^2 and $\hat{g}(k)$ are not based on independent samples, as noted before. Also, improper scaling of s^2 can be tested separately with an equal-variance test.

TREATMENT OF IRREGULARLY SPACED OBSERVATIONS

The Correlogram and Semi-Variogram

A plot of the sequence of $\{ \hat{r}(k), k = 1, 2, \ldots n\text{-}1\}$ versus lag k or Δt_k is known as the sample autocorrelation function, or, *correlogram*. The (semi-)variogram may be similarly plotted as a function of lag k. Interpretation of such graphical devices (or any other serial correlation) estimator usually hinges on the assumption that X_i and X_{i+k} are always matched pairs; that is, the interval of time Δt_k between all t_i and t_{i+k} will be constant for the k^{th} lag.

The tests of correlation presented so far have not presented any explicit requirements on the time-spacing of the residual differences being processed. This is partly because, under the assumption of whiteness, the correlation between values in a series is expected to be zero ***regardless*** of the time interval between them. It therefore seems permissible to test for whiteness without strictly regarding the sampling rate as even; however, the alternative hypothesis (that non-zero serial correlations exist) should be more easily detected if sample data are evenly spaced in time. Unfortunately, evenly spaced observations rarely exist in actual orbit determination problems; even tracking data that are expected to be regularly spaced over time (such as those coming from a space-borne GPS receiver) may face tracking outages or outlier rejections, thereby requiring a process that accommodates irregular time gaps.

The Pseudo-Correlogram and Pseudo-Variogram

Because one cannot presume that Δt_k will be constant for the orbit-determination problem, we approach the problem by dividing the estimation timeline into regularly spaced time grids that are small enough to occupy either ***one*** measurement time t_i, or, ***no*** measurement time tag. Each measurement can thereby be uniquely assigned to the time of the grid that contains it, which is equal to replacing the unevenly-spaced measurement times with evenly-spaced grid times Δt_{grid}. In so doing, the k^{th} lag now corresponds to measurements separated by k evenly-spaced grids.

We refer to a correlogram or semi-variogram where pairing is done by time grids instead of measurement times as a ***pseudo-correlogram*** or ***pseudo-variogram***,[*] respectively. We differentiate them in name because, unlike an ordinary correlogram or semi-variogram, we do not actually make any strong claims about the interpretation of these "pseudo-estimates" when correlations exist in the residual time series. Rather, we only claim that the pseudo-estimates are a generalization that ought to provide a more powerful test of serial independence against the alternative hypothesis of correlated data relative to an ordinary correlogram or semi-variogram if the measurements are irregularly spaced.[†]

---

[*] It would be more accurately called the ***pseudo-semi-variogram-variance ratio***, but the shorthand descriptor seems less cumbersome.

[†] However, if the data are evenly spaced, and if the fixed grid size Δt_{grid} is chosen to be equal to the sampling rate of the time series data, then the pseudo-estimators provide the same result as the ordinary estimators.

Estimation of the Pseudo-Variogram and Pseudo-Correlogram

The pseudo-variogram has a rather simple form. For a time-sorted series, the following nested loops (illustrated with FORTRAN code) suggest how the data might be associated and stored:

```
DO I = 1,N
      DO J = I+1, N
            K = NINT(ABS(T(I) - T(J))/DTGRID)
            SOS(K) = SOS(K) + (X(I) - X(J))**2
            NPAIR(K) = NPAIR(K) + 1
      ENDDO
ENDDO
```

The semi-variogram at lag k would then be calculated as half the sum of squares `SOS` divided by the number of pairs `NPAIR`, once a grid spacing Δt_{grid} (`DTGRID`) is chosen. Our experience so far suggests that the precise value of Δt_{grid} is not terribly critical under the null hypothesis, so long as it remains smaller than minimum successive difference encountered. Our preference has therefore been to divide the median of all successive time differences $\Delta t_1 = |t_{i+1} - t_i|$, $\{i = 1, 2, \ldots n-1\}$ by an integer multiple of say, two or four, to get a sufficiently fine grid spacing to guarantee no more than one measurement time per grid. The number of differences computed is $n\,(n-1)/2 \approx \frac{1}{2}n^2$, and the total of variable `NPAIR(K)` replaces $(n-k)$ in estimators of $\gamma(k)$ and $\rho(k)$ in Equations (20) and (15).

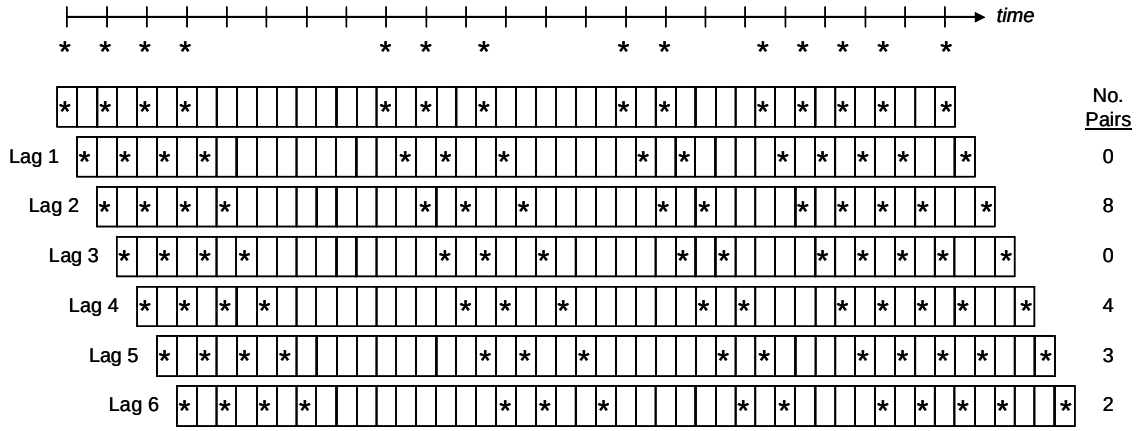


Figure 1. Example of Pseudo-Correlogram Time Gridding.

Figure 1 illustrates how time gridding works for either the pseudo-correlogram or pseudo-variogram for data that are *almost* regularly spaced, yet some gaps exist due to extended lapses in tracking or the rejection of outliers. In the example, fourteen measurements (represented by asterisks on the timeline) are separated on the timeline by intervals of {10.1, 9.8, 9.9, 50.1, 9.8, 15.0, 34.8, 10.0, 24.9, 9.9, 9.9, 9.9, 15.0}. The median time difference of this set is found to be $\Delta t_{\text{median}} = 10.0$, and the minimum time difference is observed to be $\Delta t_{\text{min}} = 9.8$. To create a sufficiently small grid, a divisor Δt_{grid} is needed that exceeds $\Delta t_{\text{median}} / \Delta t_{\text{min}} \approx 1.02$. For this example, an integer divisor of two (2) is chosen such that $\Delta t_{\text{grid}} = \Delta t_{\text{median}}/2 = 5$. Each measurement is then assigned to a grid that is five time units wide, and lagged pairs are identified according to grid time rather than actual time. At lag $k = 1$ ($\Delta t = 5.0$) then, no pairings exist; at lag $k = 2$ ($\Delta t = 10.0$), eight pairings exist; and so on, as indicated in Figure 1. Generally, as the grid size Δt_{grid} decreases, the measurements will appear more regularly spaced to the estimator. However, this

comes at a price, for the number of matched pairs at each lag may also decrease. It would not be uncommon for some gridded lags to have *no* available estimate of correlation because no measurements could be associated. This suggests another need for statistical tests that must be accurate for small samples.

TESTS FOR SHORT-TERM SERIAL CORRELATION

Whiteness testing of sample correlogram or semi-variogram coefficients is applied to the estimate at each lag. Therefore, a test of short-term correlation easily follows by testing whether $\hat{r}(k)$ or $\hat{g}(k)$ is significantly far from its expected value for small k, especially $k = 1$ (the first lag).

Mean Squared Successive Difference Test. A specialized test often used for testing the null hypothesis of randomness or trends in a data set is the *mean squared successive difference test* (MSSD). This test supposes that the $k = 1$ semi-variogram ratio for an uncorrelated, normally-distributed series will approach the normal distribution with mean and variance:[20]

$$E\left[\frac{\hat{g}(1)}{s^2}\right] = 1 \;;\; Var\left[\frac{\hat{g}(1)}{s^2}\right] = \frac{n-2}{n^2-1}. \tag{21}$$

Therefore, if $-2.58 < ((n^2 - 1) / (n - 2))\ \hat{g}(1)\ / s^2 < +2.58$, then the hypothesis of residual whiteness at $k = 1$ cannot be rejected at the 1% significance level. The advantage of this test is its simplicity: it uses confidence intervals based on a normal distribution and it does not presume regularly spaced measurements under the hypothesis of whiteness.

OMNIBUS TESTS OF OVERALL SERIAL CORRELATION

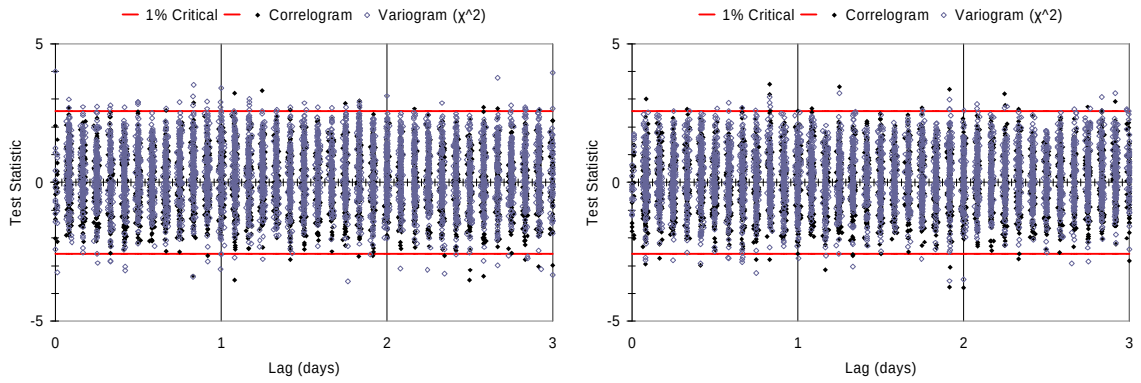
Testing individual correlation coefficients does not provide a conveniently single criterion for gauging whiteness "overall". This inconvenience is addressed by the *Ljung–Box test*, which tests whether any of a group of autocorrelations of a time series is significantly different from zero.[21] The test is specifically intended as a goodness-of-fit test of time-ordered residuals when the Pearson product-moment form is used to estimate the sample correlogram, but for small n (< 100), the test is reputed to have limited power.[22] A similar test, which is possibly more sensitive, is noted by Brockwell and Davis (1991), based on the correlogram of the sequence of *squared* residuals.[23]

However, these omnibus tests do not necessarily apply to a *pseudo*-correlogram. Therefore, we consider that the significance level of any statistical test equals the overall false alarm rate whenever the null hypothesis is true. For example, if $\{ \hat{r}(k),\ k = 1, 2, \ldots\}$ were all being tested for whiteness at the 5% significance level, we would expect the cumulative number of failures to be ~5% if whiteness existed. However, simulation studies suggest that the actual cumulative failures for white Gaussian deviates might vary in the neighborhood of 4% to 6% depending on such factors as the measurement spacing, the number of measurements, and the gridding value Δt_{grid} adopted.

Example

To illustrate the omnibus test, we consider 2220 irregularly-spaced satellite tracking residuals originating from an orbit-determination filter that are seemingly "non-optimal" due to the detection of short-term correlation (that is, the short-term correlation test statistics MSSD and first-lag pseudo-correlation $\hat{r}(1)$ exceeded their 1% significance critical values). The measurements, spaced at $\Delta t_{median} = 16.7$ seconds, are evaluated using a pseudo-correlogram and pseudo-variogram with grid spacing of $\Delta t_{grid} = 4.2$ seconds to create approximately 7000 pseudo-

coefficients with at least h = 5 pairings. From the filter residuals, we computed and plotted $\sqrt{h}\ \hat{r}(k)$ ("Correlogram" of Figure 2a), where h is the number of available pairs at lag k and $\hat{r}(k)$ is pseudo-correlogram coefficient estimated according to Eq. (15). Critical values of ±2.58 (indicating 1% significance assuming normality) are superposed. Also superposed in Figure 2a is a plot of a transformation of $1 - \hat{g}(k) / s^2$ (labeled "Variogram"), which recalling Eq. (20), transforms the semi-variogram coefficient into a correlogram-like value.* We now compare these outcomes to Figure 2b, where the tracking residuals are replaced by simulated Gaussian white noise at the same measurement times, such that no significant correlations should exist. Here we might notice that the scatter in the real-data statistics seem to be slightly greater than that for the simulated white noise, resulting in more "failures" (values exceeding the 1% significance levels) relative to white noise. But since it is hard to tell from inspection, how should this be quantified?



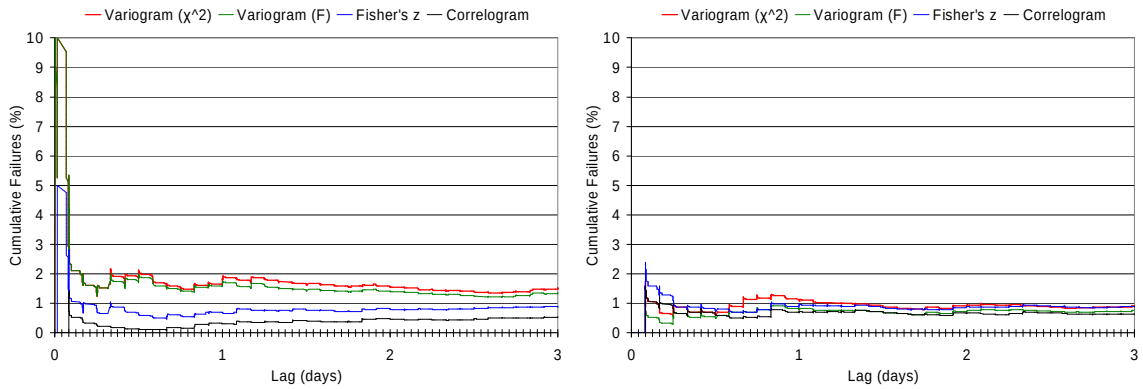
(a) Actual Tracking (b) Simulated White Noise

Figure 2. Correlation Test Statistics for Sample Tracking Tested at 1% Significance.

Figure 3 illustrates a plot of the accumulated number of failures versus the number of lags tested. Accumulated failures using normal critical values for Fisher's *z*, as well as critical values for the pseudo-variogram based on the *F*-distribution, are also added to Figure 3. It may be observed that, for the simulated data where zero correlation is known to exist, all four test statistics experience cumulative failures close to the 1% significance level used, as expected. Also, both types of test statistics (correlogram and semi-variogram coefficients) seem to detect that short term correlations exist in these tracking residuals, typified by the higher-than-expected failure rates for short lags. However, the number of overall failures appears higher than expected when semi-variogram coefficients are used as the test statistic, where the pseudo-correlogram is much lower than expected. This example suggests that the critical values of the pseudo-variogram estimator (based on the χ^2- and *F*-distributions) may detecting additional correlations that the pseudo-correlogram estimators cannot.

---

* Critical values of the semi-variogram statistic (based on either the χ^2- and *F*-distributions) are non-linear functions of the degrees of freedom *h*. For Figure 2, we rescale the semi-variogram-ratio statistic by multiplying by $\pm 2.58\ h / \chi^2_{\beta,h}$ (where β = 0.5% if $\hat{g}(k) / s^2 < 1$, and β = 99.5% if $\hat{g}(k) / s^2 > 1$) to display a statistic whose 1% significance critical values are also at ±2.58. This transformation is not operationally necessary, but is done solely for the purpose of creating a side-by-side illustration with the correlogram estimate of Eq. (13).



(a) Actual Tracking

(b) Simulated White Noise

Figure 3. Cumulative False Alarms for Sample Tracking Tested at 1% Significance.

Simulated Time Series

In cases where simulated time series are truly white, there tends to be little difference in the outcomes of statistical tests using the pseudo-correlogram of Eq. (15) or the pseudo-variogram of Eq. (20). In simulation cases where significant autoregressive correlations are introduced (*e.g.*, short-term recursions such as $X_{i+k} = a\, X_i + \nu_i$, where a is some small constant and ν is white noise, *etc.*), the two types of estimators tend to behave similarly, although the critical values of the semi-variogram seems more sensitive to detecting the presence of correlations. We are therefore of the opinion that semi-variogram-based estimators show more power at rejecting the null hypothesis of whiteness when it is false, and may be preferred for that reason.

Spectral analyses of real-tracking residuals also suggest that un-modeled accelerations in non-optimal estimates may present slight periodic (per revolution) signatures. However, it is often the case that binning / gridding process of the pseudo-estimators can cause periodic signal to alias into other frequencies, with statistically significant failures appearing at lag values that are not at the actual period of the signal. This aliasing is quite acceptable for our purposes, since our use of pseudo-estimators is to simply identify the presence of significant correlation, rather than investigate the structure of the correlation. However, differences in behavior between the conventional and pseudo-estimators in the presence of periodic signals in one reason why we cannot generally equate pseudo-estimates with those of conventional correlograms and semi-variograms.

Simulated Orbit Determination

The Orbit Determination Took Kit (ODTK) simulator-filter-smoother was used to briefly investigate the correctness of the mean, variance, and whiteness tests. ODTK maps the uncertainties of physical force models into covariance process noise, which is more physically realistic than ad-hoc tuning methods.[24] Ordinarily, ODTK uses the filter-smoother consistency test as its primary goodness-of-fit test (discussed in the Appendix), but statistical hypothesis testing of the residuals are not ordinarily employed.

Our limited goodness-of-fit testing of residual mean, variance, and whiteness suggested that the ODTK filter-smoother provides optimal (good-fitting) results in simulation cases where the filter and smoother parameterizations reasonably match the simulator parameterization. We observed that the test of the variance and the tests of short-term correlation are most sensitive to changes in tracking system parameterization, and thereby may be most useful for calibration.

CASE STUDY USING REAL TRACKING DATA

Our next goal was to apply goodness-of-fit tests to genuine satellite-tracking residuals processed by the ODTK filter-smoother. For our case study, we used two-way ranging observations from two tracking stations relayed through the transponders of a geosynchronous satellite.

System Initialization. We first established an initial filter parameterization whose solution converged on the measurements but was not necessarily an optimal fit. A realistic satellite parameterization (mass, area, attitude, initial conditions, *etc*.) was required to begin the estimation process. Parameterizations related to the force modeling, such solar radiation pressure, ballistic coefficients, transponder characteristics, *etc*., were also carefully researched and set, as well as tracking-station characterizations (location, coordinate frame, bias and noise characteristics, refraction and attenuation modeling, *etc*.) Discussions with the owner operators and initial filter-smoother runs were conducted to set expectations for the parameters being explored, and a special program was written to vary these parameters in a systematic way.

Transponder Calibration. The transponder was thought to operate according to its manufacturer's configuration as it is not changed by the satellite's owner / operator unless problems are observed. The estimated transponder noise was initially set to be a few meters based on reports of manufacturer's pre-flight testing, but other parameters characterizing drifting ranging bias are generally unknown ahead of time. We therefore processed several weeks of tracking data with ODTK to obtain other preliminary estimates of transponder performance (transponder bias uncertainty and transponder bias half-life).

Tracking Station Calibration. The tracking stations had relatively well-known locations, but little was known about their bias and noise characteristics. Even while the random noise may be very small, the bias value may drift significantly over time. Three parameters were therefore initialized based on early ODTK runs using available tracking data: the ranging-system noise level (white-noise sigma), the ranging bias (drift) uncertainty, and ranging bias half-life.

Approach

After preliminary parameter values were set, we varied each of the transponder and tracking-station parameters one at a time, and plotted the outcomes of the statistical tests versus the parameter values tried. Possible trends in the plots were used to identify what parameter values might be tried next.[*] Matching residual-test outcomes with successful filter-smoother consistency was more difficult to automate because plots of the filter-smoother consistency statistics tend to be assessed subjectively (visual examination by an analyst). Some sample trend curves are shown in Figure 4 illustrating how the test statistics can change as the parameterization changes. In this case, it is apparent that as the bias (drift) sigma is increased, the variance statistic moves downward and the MSSD and 1st lag statistics move upward. Nearly "successful" test outcomes appear close to unity in this plot.

[*] Early attempts tried "mesh grids" over *all* calibrated parameters (transponder bias sigma and half-life, and tracker white noise sigma, bias sigma, and half-life); however, plotted results did not provide definitive guidance for subsequent trials.

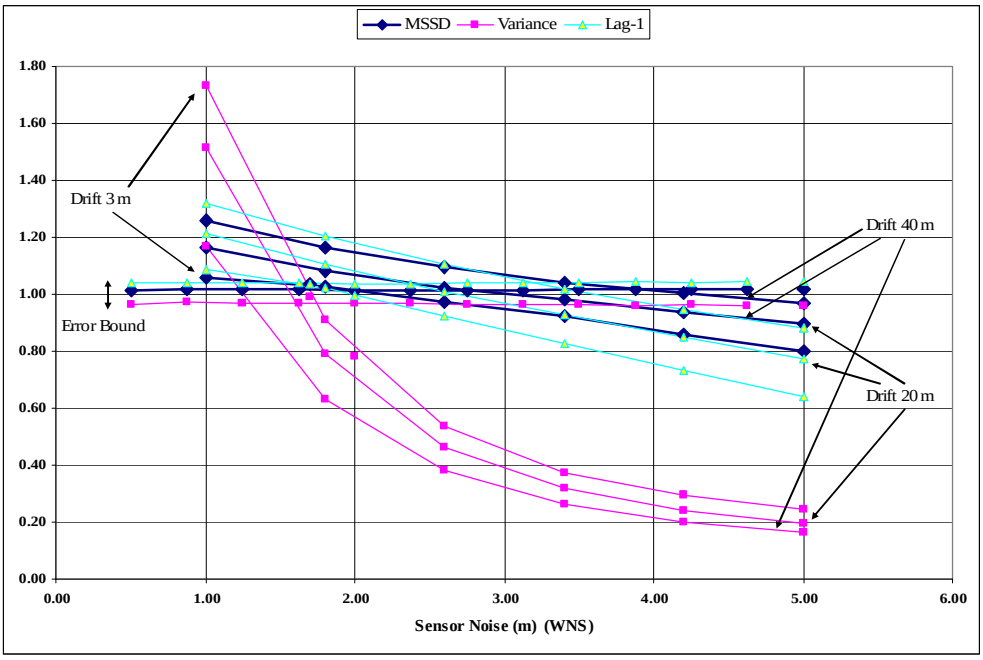
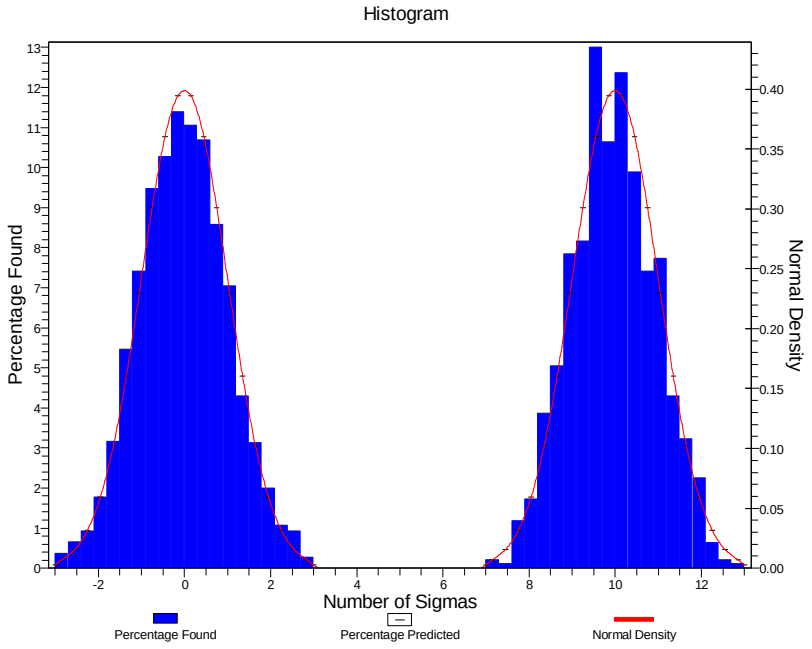


Figure 4. Tracker Parameter Trends.



(a) Tracker #1 (b) Tracker #2

Figure 5. Histograms Tracking-Station Residual Ratios.

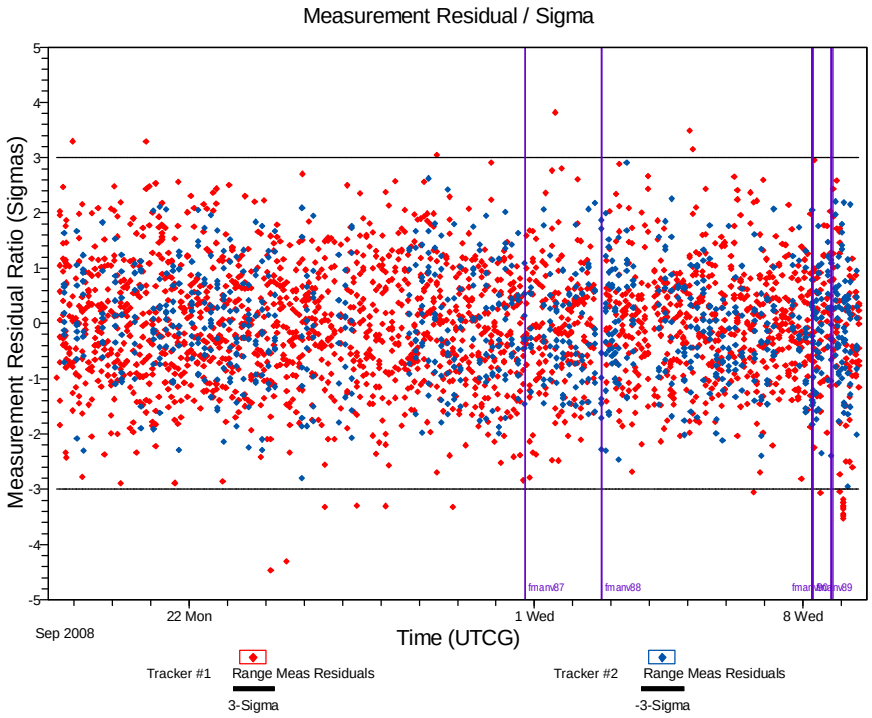


Figure 6. Tracking-Station Residual Ratios.

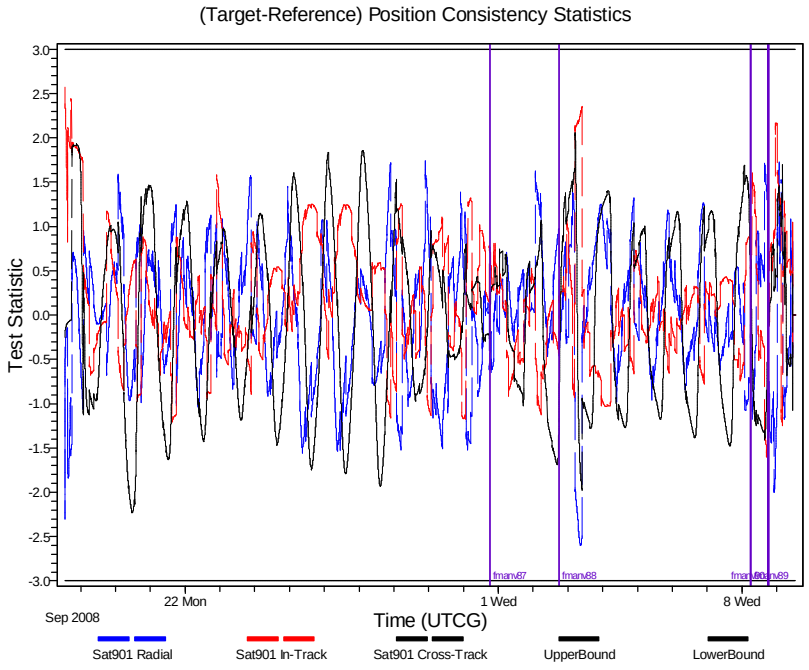


Figure 7. Filter Smoother Position Consistency Test.

Best Results Encountered

A "best-case" outcome was selected after testing extensive adjustments to the tracking-station and transponder parameters. *Best-case* is in quotes because the statistical hypothesis uncertainty of the tests can allow for a *range* of acceptable parameterizations. For example, with one of the tracking stations, a 1-meter bias (drift) sigma and 81-minute half-life seemed just as acceptable as a 5-meter bias sigma and a 1440-minute half-life (the values eventually adopted as "best").

For our adopted best case, the histograms suggest that the measurement residual ratios are approximately normally distributed (Figure 5) and the residual ratios appear sufficiently random at the proper scale (Figure 6). The observation interval included several maneuvers that we did not estimate using ODTK, because there was no noticeably adverse effect on the filter-smoother consistency test using the maneuver estimates provided by the satellite's owner-operator (Figure 7).

Table 1. Case-Study Results: Overall Whiteness of Residual Ratios

Test Statistic	Distribution for Critical Values	Cumulative Failures Tested at #1 Significance	
		Tracker #1	*Tracker #2*
Pseudo-Variogram	χ^2	1.59% (120 of 7524)	1.16% (128 of 11049)
Pseudo-Variogram	*F*	1.42% (107 of 7524)	1.11% (123 of 11049)
Pseudo-Correlogram (Fisher's *z*)	Normal	0.89% (67 of 7524)	0.81% (90 of 11049)
Pseudo-Correlogram (Pearson Product-Moment)	Normal	0.54% (41 of 7524)	0.28% (32 of 11049)

Table 2. Case-Study Results: Mean, Variance, and Short-term Whiteness of Residual Ratios[*]

Test Statistic	*2143 Residual Ratios from Tracker #1*				*955 Residual Ratio from Tracker #2*			
	1% Lower Critical	Outcome	1% Upper Critical	Significance (%)	1% Lower Critical	Outcome	1% Upper Critical	Significance (%)
Mean	-0.056	**0.014**	0.056	*52.5*	-0.084	**-0.044**	-0.084	*17.4*
Variance	0.923	**1.058**	1.080	*6.19*	0.885	**0.960**	1.124	*39.7*
MSSD	0.944	**1.043**	1.056	*4.58*	0.916	**1.070**	1.084	*3.23*
$\hat{g}(1)$	0.918	**1.081**	1.086	*1.57*	0.878	**1.062**	1.131	*20.6*

Table 1 and Table 2 summarize the results of statistical hypothesis tests on the residual ratios for the best case. The results of an omnibus test of overall whiteness (Table 1) seems to affirm that critical values from a normal distribution used with the sample autocorrelation coefficients from Eq. (15) can under-report the presence of correlations relative to the other statistics. This

[*] The critical values and significance values assume that the mean is normally distributed, and the variance, mean-square successive difference, and first-lag pseudo-variogram are χ^2 distributed.

affect was noticed during our simulation studies, leading us to believe that this statistic is less likely to detect correlations should they exist. If so, then the residuals of Tracker #1 might not be as white as they could in this case, since the pseudo-variogram statistics are slightly above a 1% false alarm rate anticipated under white noise conditions.

Table 2 summarizes test results for the mean, variance, and short-term correlation of the residual ratios. Basically, we consider the filter's parameterization to be in the neighborhood of optimality whenever all/most of the test statistics fall between their critical values. The column "Significance (%)" is the significance of the critical value at which the test statistic would have failed if it had been tested as that level; thus, while all of short-lag whiteness statistics pass a 1% significance test, some would not always pass a 5% significance test; this again suggests that there might still be room for very small improvements in our calibration of Tracker #1.

POSSIBLE AREAS OF FUTURE STUDY

One concern regarding Fisher's z is that it may have less sensitivity to the presence of significant serial correlations. Another transformation has been developed by Hotelling that employs higher-order corrections than Fisher's z, thereby providing a more nearly normally distributed statistic than Eq. (16) for small samples.[25] Future work may explore the use Hotelling's z^* as an alternative to Fisher's z.

Other future work may consider the use of variance estimators that more accurately compensate for the fact that outliers have been rejected from the sample.[26] Automating the search process for optimal parameters also appears feasible, and is being explored.

SUMMARY RECOMMENDATIONS

An optimal filter should generate predicted measurement-residual ratios that are zero mean, unit variance, and uncorrelated. While the need for white residuals is often noted, this condition seems to be rarely tested in practice; this is probably because the most commonly recommended test - namely significance testing of the sample autocorrelation function coefficients - may not be useful to orbit determination problems where tracking data are irregularly sampled with time. In this paper, we propose methods of estimating autocorrelation coefficients by pairing measurements according to regularly spaced time grids so they can be time-paired in an approximately even way. Called the *pseudo-correlogram* or *pseudo-variogram* in this paper, these methods are thought to have better power at detecting serial correlations in irregularly sampled time series.

We conclude that pseudo-variogram estimates $\hat{g}(k)$ demonstrate reasonable power at detecting correlations in real and simulated data, and we therefore suggest the following goodness-of-fit tests for residual ratios:

1. Test for zero mean using the critical values based on a t- or normal distribution.
2. Test for unit variance using the critical values based on a χ^2-distribution.
3. Test for short-term autocorrelation using χ^2 critical values for the first-lag pseudo-variogram ratio $\hat{g}(1) \; / \; s^2$ (supplemented with mean-square-successive-difference test).
4. Test of significant autocorrelations overall by monitoring the cumulative failure rate over many lagged pseudo-variogram ratios $\{ \hat{g}(k) \; / \; s^2, \; k = 2, 3, \ldots \}$.

Our preference has been to use χ^2 critical values for the test statistic $\hat{g}(k) \; / \; s^2$, instead of the F critical values, the F distribution being less conservative.

APPENDIX - FILTER-SMOOTHER CONSISTENCY TEST

While the whiteness of measurement residuals provides a testable criterion of optimality for real tracking data, Wright acknowledged that McReynolds filter-smoother consistency test should also be satisfied globally as a type of goodness-of-fit test for orbit determination.[4] McReynolds (1984) proved that the difference between a filtered state $\boldsymbol{x}_f(t)$ and a smoothed state $\boldsymbol{x}_s(t)$ is normally distributed in k dimensions, if k is the size of the state-difference vector $\Delta\boldsymbol{x}_{(f\text{-}s)}(t)$.[27] He also showed that the covariance matrix $\Delta\boldsymbol{P}_{(f\text{-}s)}(t)$ of the state-difference vector is equal to the filter error-covariance $\boldsymbol{P}_f(t)$ minus the smoother error-covariance $\boldsymbol{P}_s(t)$. Therefore, the time sequence of $\boldsymbol{z}_{(f\text{-}s)}(t) = [\Delta\boldsymbol{x}_{(f\text{-}s)}(t_i)]^T[\boldsymbol{P}_{(f\text{-}s)}(t_i)]^{-1}[\Delta\boldsymbol{x}_{(f\text{-}s)}(t_i)]$, $\{t_i\ ,\ i = 1, 2, \ldots\ n\}$ provides an sample population over the estimation interval upon which the assumption of multivariate normality can be tested.[28] If the sequence of $\boldsymbol{z}_{(f\text{-}s)}(t)$ supports the hypothesis of multivariate normality, then the filter may be considered optimal. If the sequence $\boldsymbol{z}_{(f\text{-}s)}(t)$ does not support the null hypothesis of multivariate normality, then the filter may be considered non-optimal.

Due to difficulties in accessing the multivariate normality of the test statistic $\boldsymbol{z}_{(f\text{-}s)}(t)$, $\boldsymbol{z}_{(f\text{-}s)}(t)$ is replaced by a subset of its k univariate components:

$$\Delta x_{(f\text{-}s)}\ /\ \sigma_{(f\text{-}s)} = (x_{\text{filter}} - x_{\text{smoother}})\ /\ (\sigma_{\text{filter}} - \sigma_{\text{smoother}}), \tag{22}$$

where x is the parameter estimate and σ^2 is the diagonal element of the covariance corresponding to that x.[29] 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 and filter-smoother consistency is claimed when the scatter of this metric stays within ±3 over the fit interval.

REFERENCES

[1] D'Agostino, R.B., M.A Stephens (eds., 1986), *Goodness-of-fit Techniques*. Dekker. p. 1.

[2] Anderson, B.D.O., J.B. Moore (1979), *Optimal Filtering*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey. p. 5.

[3] Bar-Shalom, Y., X.R. Li, T. Kirubarajan (2001), *Estimation with Applications to Tracking and Navigation*. John Wiley & Sons, New York, p. 232-3

[4] Wright, J.R. (2002), "Optimal Orbit Determination." Paper AAS 02-192, from Alfriend, K.T., *et al.* (eds.), *Spaceflight Mechanics 2002 - Advances in the Astronautical Sciences, Vol. 112, Part II,* Proceedings of the AAS/AIAA Space Flight Mechanics Meeting, San Antonio, Texas, January 27-30, 2002, pp. 1123-34.

[5] Vallado, D.A., J.H. Seago (2009), Covariance Realism." Paper AAS 09-304, from Rao, *et al*. (eds.), *Astrodynamics 2009 - Advances in the Astronautical Sciences, Vol. 112, Part I,* Proceedings of the AAS/AIAA Astrodynamics Specialist Conference, Pittsburgh, Pennsylvania, August 9-13, 2009, pp. 49-67.

[6] Maybeck, P.S. (1979) *Stochastic Models, Estimation, and Control - Vol. 1*. Academic Press, New York, p. 231-2.

[7] Maybeck, P.S. (1982) *Stochastic Models, Estimation, and Control - Vol. 2*. Academic Press, New York, p. 136.

[8] Anderson, B.D.O., J.B. Moore (1979), *Optimal Filtering*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey. p. 231.

[9] Hart, A. (1988), "Standard Deviation." in Kotz, S., N.L. Johnson (eds.), *Encyclopedia of Statistical Sciences, Vol. 8,* John Wiley & Sons, New York. p. 626.

[10] Sheskin, D.J. (2000), *Handbook of Parametric and Nonparametric Statistical Procedures*, 2nd Edition. Chapman & Hall/CRC, p. 91.

[11] Ibid., p. 67.

[12] Bendat, J.S., A.G. Piersol (2000), *Random Data: Analysis & Measurement Procedures*. Wiley-Interscience. pp. 87-9.

[13] Chatfield, C. (1996), *The Analysis of Time Series*, 5th Edition. Chapman & Hall / CRC. Boca Raton, FL, p. 243.

[14] Rodriguez, R.N. (1982), "Correlation", in Kotz, S., N.L. Johnson (eds.), *Encyclopedia of Statistical Sciences, Vol. 2,* John Wiley & Sons, New York. p. 204.

[15] Crassidis, J.L., and J.L. Junkins (2004), *Optimal Estimation of Dynamic Systems*, CRC Press, Boca Raton, Florida, pp. 301-03.

[16] Jenkins, G.M., D.G. Watts (1968), Spectral Analysis and its Applications, Holden-Day, San Francisco. p. 182.

[17] Cressie, N. (1988), "A Graphical Procedure for Determining Non-stationarity in Time Series." *Journal of the American Statistical Association*, Vol. 83, pp. 1108-1116.

[18] Brownlee, K.A. (1965), *Statistical Theory and Methodology in Science and Engineering – 2nd ed.* John Wiley & Sons, Inc. New York, pp. 285-88.

[19] 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*. (eds.), *Spaceflight Mechanics 2005 - Advances in the Astronautical Sciences. Vol. 120, Part I*, Proceedings of the AAS/AIAA Space Flight Mechanics Meeting, Copper Mountain, Colorado, January 23-27, 2005, pp. 367-386.

[20] von Newmann, J., R.H. Kent, H.R. Bellinson, B.I. Hart (1941), "The Mean Square Successive Difference Test." Annals of Mathematical Statistics, Vol. 12, pp. 153-62.

[21] Ljung, G. M., and G.E.P. Box (1978). "On a Measure of a Lack of Fit in Time Series Models". *Biometrika*, Vol. 65, No. 2, : pp. 297–303.

[22] Davies, N. & P. Newbold (1979), "Some power studies of a portmanteau test of time series model specification." *Biometrika*, Vol. 66, No. 1, pp. 153–56.

[23] Brockwell, P.J. & R.A. Davis (1991), *Time Series: Theory and Methods*, 2nd Editoin. Springer-Verlag, New York. pp. 311-12.

[24] Hujsak, R.S., J.W. Woodburn, J. H.Seago, "The Orbit Determination Tool Kit (ODTK) – Version 5," Paper AAS 07-125, from Akella, M. *et al*. (eds, 2007), *Spaceflight Mechanics 2007 - Advances in the Astronautical Sciences. Vol. 127, Part I.* Proceedings of the AAS/AIAA Space Flight Mechanics Meeting, Sedona, Arizona, January 28-February 1, 2007.

[25] Rodriguez, R.N. (1982), "Correlation", in Kotz, S., N.L. Johnson (eds.), *Encyclopedia of Statistical Sciences, Vol. 2,* John Wiley & Sons, New York. p. 204.

[26] Prescott, P. (1979) "A mean difference estimator of standard deviation in asymmetrically censored normal samples." *Biometrika*, Vol. 66, No. 3. pp. 684-86.

[27] 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.

[28] 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.

[29] Vallado, D.A., J.H. Seago (2009), "Covariance Realism." Paper AAS 09-304, from Rao, *et al*. (eds.), *Astrodynamics 2009 - Advances in the Astronautical Sciences, Vol. 112, Part I*, Proceedings of the AAS/AIAA Astrodynamics Specialist Conference, Pittsburgh, Pennsylvania, August 9-13, 2009, pp. 49-67.