MCDONNELL DOUGLAS TECHNICAL SERVICES CO.
HOUSTON ASTRONAUTICS DIVISION

CR 151363

SPACE SHUTTLE ENGINEERING AND OPERATIONS SUPPORT

DESIGN NOTE NO. 1.4-8-018

SMOOTHING OF ORBITAL TRACKING DATA

MISSION PLANNING, MISSION ANALYSIS AND SOFTWARE FORMULATION

16 MAY 1977

This Design Note is Submitted to NASA Under Task Order No. D0710, Task Assignment J, in Partial Fulfillment of Contract NAS 9-14960.

PREPARED BY: J. D. Vedder
J. D. Vedder
Group Engineer
488-5660, Ext. 293

APPROVED BY: Wm L. Gibson
W. L. Gibson
Task J Manager
488-5660, Ext. 293

APPROVED BY: Peggy M. Dugge
P. M. Dugge
Technical Manager
Navigation Software Formulation
and Analysis

APPROVED BY: W. E. Hayes
W. E. Hayes
Project Manager
Mission Planning, Mission Analysis
and Software Formulation

1.0 SUMMARY

This note discusses the problem created by the presence of "wild" or outlying data points among orbital tracking data. Consideration is given to the pernicious effects of such outliers on the orbit determination process, and new methods for minimizing or even eliminating these effects are proposed after reviewing previous methods. Some preliminary efforts implementing these new methods are described, and the results thus far obtained are summarized. Based on these ideas and results, recommendations are made for future investigation.

2.0 INTRODUCTION

The basic input to an orbit determination (OD) computer program consists of some form of orbital tracking data. These data may be of a single type, or they may be of a variety of types such as range, range-rate, angles, etc. Such data may come from a single earthbound tracking station, a network of such stations, or they may originate from a satellite - to - satellite indirect tracking system. But whatever their origin or type may be, or however sophisticated the OD program may be, the quality of the tracking data will decisively affect the quality of the final results.

The quality of orbital tracking data is a function of several factors. Firstly, it depends on the inherent capability of the measurement process to produce accurate results. Secondly, the tracking data quality depends on the noise superimposed on the data. Lastly, it is affected by the presence of "wild" data points or statistical outliers whose cause is something other than noise. OD results can generally be improved by lessening the deliterious effects of any of these factors affecting input, as well as by improving the OD algorithms and techniques themselves. This paper will focus on the last factor mentioned - the problem of outliers.

3.0 DISCUSSION

3.1 Background, Recent Solutions and Their Limitations

Considering the effect of statistical outliers among tracking data on the OD process, it is immediately apparent that the dividing line between very noisy data points and outliers is rather indistinct. However, this fact need not preclude considering outliers a problem distinct from noise since their borderline can be defined rather arbitrarily without any damaging consequences. Whatever the cause of their deviation, any data differing significantly from the central tendency of the bulk of the data can be considered outlying by definition. If something definite is known about tracking measurement errors or noise, this knowledge may be used to define outlying data points more precisely. But all that actually need be assumed is that the noise does not completely obscure the tendency of the data to cluster or concentrate about a slowly varying function.

Having thus roughly defined tracking data outliers, their effect on the OD process can be considered more specifically. If their presence is essentially ignored, then they will seriously affect the convergence of the differential correction process. At best the rate of convergence will be decreased, while at worst convergence will become erratic and may not even occur. Such effects will almost always be observed when data with outliers is used as input to a least squares OD program. This occurs largely because the least squares (LS) criterion minimizes the sum of squared deviations about the mean, and both this sum as well as the mean itself are very strongly affected by outliers. Nevertheless, for historical and other reasons the LS method continues to dominate OD programs. Possible alternatives will be considered later.

Realizing that the inclusion of tracking data outliers will usually lead to serious convergence problems as well as produce results disproportionately affected by these outliers, it would appear that their elimination prior to further OD processing would be wise. However, to eliminate them it is first necessary to detect them and this in effect transforms the problem into what is commonly called data smoothing or pre-filtering or editing. Solutions to this latter type of problem have shown that although it is not difficult to devise simple and effective means to detect and expunge strongly deviant data, accomplishing this task for slightly deviant data is much more difficult. Outlying tracking data points, for example, may differ from their proper values by only one part in 10^5 or so, and thus very sensitive detection methods are required. But many sensitive techniques cannot simultaneously detect a wide range of deviant points. The development of effective and fast data editing algorithms capable of smoothing orbital tracking data is therefore a formidable task, especially if they must smooth data whose spacing and density vary widely and whose outlying points deviate over a very wide range.

Before discussing possible new solutions to the problem posed by tracking data outliers, it might be well to consider how NASA's GSFC has approached this problem. Details of their approach are described in Ref. 1-4. Their approach as summarized in Reference 4 consists essentially of fitting low order (third or fourth) polynomials to groups of sequential tracking data points. A LS method is used for the polynomial fits and a 2.5 sigma data rejection criterion is used to reject "wild" data. The midpoint values of the polynominals replace the original data, thus achieving data compression as well as smoothing. This smoothed and compressed data is then used as input to a weighted LS OD program. Among the chief limitations and disadvantages of this approach are the following. Low order polynominal fitting of data only lends itself well to short areas of fairly dense tracking data. Long arcs of dense data must be divided into numerous short arcs for smoothing by this technique, and long arcs of sparse data can hardly be handled at all. Thus in the first case the global cohesiveness of the data is substantially lessened and hence a new source of noise is effectively introduced, while in the latter case smoothing is hardly possible at all. Furthermore, the use of least squares polynomial fitting assumes that tracking data deviations are normally distributed, a frequently questionable assumption, especially when outliers are numerous. This assumption is in fact what necessitates the effort to eliminate outliers from the tracking data before using the data as input to a LS OD program, since the data with outliers is not even approximately normally distributed. These facts emphasize the point (also made in Reference 6) that LS methods are in general ill-suited to applications involving leptokurtic distributions, that is, distributions having "fatter tails" or appreciably higher fourth central moments (kurtosis) than the normal distribution. Hence even as the Goddard approach illustrates, such applications, if successful, will generally be found to use LS methods iteratively in some way. The contrived nature of such schemes is usually obvious, and even their relative efficiency is decreasing as the computational advantages of LS algorithms relative to alternatives steadily diminishes.

Reference 5 documents a NATO technical group's approach to smoothing or pre-filtering of the orbital tracking data for an geosynchronous satellite. Their approach, essentially similar to Goddard's, additionally included using the smoothed analytical representations of the tracking data directly as input for their OD program. They reported that this latter procedure produced very sizable overall reductions in computer running time for OD.

Several deductions can be made from the discussion so far. These include the following:

1. The problem of outliers in orbital tracking data cannot be safely ignored.
2. Outliers should either be eliminated prior to OD processing (Approach 2), or the OD process should be modified to assimilate outliers without convergence problems resulting (Approach 1).
3. Polynomials are not naturally suited for smoothing

tracking data via Approach 2.

4. LS techniques are poorly suited to handle outlier contaminated data, and in particular cannot be used directly in Approach 1.
5. The problem of outliers can best be solved by analyzing it in its proper statistical context instead of attempting to apply ad hoc remedies.
6. Manual data editing should only be considered as a last resort due to its temporal inefficiency.

Before examining the implications of these deductions, it is necessary to briefly review some basic statistical ideas and results. Reference 6 is an excellent source for a deeper discussion of the following concepts.

3.2 Some Pertinent Statistical Concepts

During the past two centuries numerous methods have been proposed for estimating the central tendency and spread of a group of independent measurements of a given quantity. Among the best known point estimators of the central tendency are the arithmetic average, the median, the mode, and the midrange. Among the best known point estimators of spread are the variance or standard deviation, the semi-range, and the average absolute deviation. Statistical theory shows that each of these estimators is associated with a particular distribution of errors in the measurements. In particular, the arithmetic average and standard deviation are the best estimators for normally distributed measurement errors, the midrange and semi-range are the best estimators for uniformly distributed errors, while the median and average absolute deviation are the best estimators when errors follow a double-exponential distribution. A natural generalization of these facts is the following. If the probability density function f of the measurement errors $(u-\hat{u})$ has an exponential form given by

$$(1)\quad f(u-\hat{u}) = c_1\exp(-c_2|u-\hat{u}|^p)$$

where c_1 and c_2 are constants and p is a positive real number, then the best estimator, $\hat{u}$, of the quantity being measured is the value which minimizes Q as defined by

$$(2)\quad Q = \sum_{i=1}^{n} |u_i-\hat{u}|^p$$

where the sum extends over all n measurements. The best estimator, $\hat{s}$, of the spread of the measurements around $\hat{u}$ is given by

$$(3)\quad \hat{s} = \left(\frac{Q}{n-1}\right)^{1/p}$$

Thus the median, mean and midrange are equal to $\hat{u}$ in (2) for p = 1, 2 and infinity respectively. Equations (1) - (3) also imply that deviant measurements receive greater weight in determining $\hat{u}$ as p increases. More specifically, the case of p = 1 constitutes a natural dividing point: for $p > 1$ deviant points contribute more to Q than the measure of their linear deviation from $\hat{u}$, while for $p < 1$ deviant points contribute less to Q than the measure of their linear deviation from $\hat{u}$. When $p = 1$, $\hat{u}$ is equivalent to the median of the measurements, and obviously the median is determined solely by the number of measurements on both sides of it, not by the magnitude of their devitations. Hence even intuitively it is clear that the median is unaffected by the magnitude of outliers, though it does depend somewhat on their number. This property implies that the median is an excellent point estimator of the central tendency of a group of measurement data whenever the majority of the data is reasonably clustered about a central value while a minority of the data consists of outliers or "wild" points. Such a data distribution corresponds quite closely to the kind of distribution usually encountered with real orbital tracking data. The insensitivity of the median to the magnitudes of outliers is in marked contrast to the arithmetic mean, which shows a direct dependence on these magnitudes. Furthermore, the median is far less sensitive to bias than is the mean. For example, if a majority of data is well clustered about some value while a minority of biased data is outlying in one direction, the sample mean would be strongly shifted towards the biased data while the sample median would be only weakly shifted. And even in the case of normally distributed data with no outliers, it has been shown that the sample mean is only slightly superior to the sample median as an estimator of the mean.

Although the foregoing ideas were developed only for one dimensional distributions, they can readily be extended to multi-dimensional distributions and multiple regression as discussed in Reference 6. This suggests that there may be significant advantages to replacing LS algorithms ($p=2$ in (1) - (3)) with least sum of absolute deviations (LSAD) algorithms ($p=1$ in (1) - (3)) in some of the data smoothing and OD programs mentioned earlier. The feasibility of this approach has been greatly enhanced in recent years by the development of fast algorithms for solving overdetermined sets of linear equations in the best $p=1$ sense, and so the theoretical advantages of this choice are no longer overwhelmed by computational handicaps when compared to LS. Details of these new algorithms can be found in Reference 7 - 10.

3.3 New Solutions

Reverting now to Approach 1 for dealing with the problem of outliers in orbital tracking data, it would appear worthwhile investigating whether replacing the weighted LS algorithm in the differential correction process with a

LSAD algorithm might largely solve the problem. Total replacement of second moment concepts in an OD program with their analogous first absolute moment concepts would probably entail some fairly extensive theoretical development to produce viable analogs of such entities as propagable covariance matrices. But such total replacement need not be the sine qua non of using LSAD algorithms for at least determining solutions of the differential correction problem without concern for the effects of outliers. Such LSAD solutions could then be meshed in various ways with subsequent p=2 procedures in the OD process. Approach 1 therefore warrants development and testing.

Approach 2, or the idea of efficiently eliminating outliers before further OD processing, exhibits definite promise as well, and has in fact already been developed and tested in a preliminary manner using LSAD techniques. This approach and its development thus far can be described as follows. Considered as functions of time, tracking data variables such as range and range-rate can usually be represented to a high degree of fidelity or accuracy by a linear combination of simple analytical functions such as sinusoids etc. This suggests that some simple ideas from the theory of finite dimensional vector spaces might be usefully applied to the problem. The actual application made was to consider the range and range-rate functions from a given tracking station over selected time intervals as vectors which could be represented as linear combinations of a small number of basis vectors chosen because of their appropriateness from a physical viewpoint. The coefficients of these basis vectors were then determined for a best fit in the LSAD sense, rather than the more commonly chosen LS sense. Residuals were then calculated as the absolute differences between the best fit and the actual data, and points with anomalously large residuals were identified as outliers. Although this technique definitely resembles the Goddard approach described previously, it also differs with it in several important respects which overcome the limitations of the Goddard approach. Since the basis vectors are chosen for their natural suitability for tracking data representation, much longer data arcs can be successfully represented and smoothed than with polynomials. Another consequence of this basis choice is that both dense and sparse data can be smoothed. Furthermore, global cohesiveness of the data is retained because the data arcs need not be subdivided. Last but not least, determination of best LSAD rather than LS fits enables simultaneous detection and comparison of outliers of widely varying magnitudes.

The primary difficulty encountered with the previous technique was the determination of suitable sets of basis vectors. Initial guidance was supplied by such sources as Reference 11 - 12, but the limitations of these sources necessitated further investigation. This effort resulted in the delineation of two distinct methods for the determination of effective basis vector sets. The first is a theoretical approach such as Reference 11 - 12 which attempts to determine analytical representation of the tracking data functions from basic kinematic theory of satellites. The second approach is purely statistical and consists of a principal components analysis (=Karhunen-Loeve analysis = proper orthogonal decomposition = intrinsic analysis) of an ensemble of tracking function vectors generated numerically by a realistic orbit simulation program; the ensemble must correspond to an orbit similar to that of the data to be smoothed. This second approach leads (by eigenvector determination of a second moment matrix) to an optimum orthogonal basis for the ensemble analyzed, and since the relative importance of each basis vector is simultaneously determined along with a measure of the fidelity of representation, the least important basis vectors can be discarded to reduce the size of the set to a reasonable extent. The basis vectors so determined can either be used directly in the data smoothing technique or simple analytical approximations of them can be determined by inspection and used instead. This second approach lends itself more easily to semi-automation since it is largely numerically, rather than analytically, implemented. The first approach can become analytically unwieldy, though it does perhaps offer the possibility of greater insight. Results thus far have been based on the first approach, though both warrant further development and testing.

4.0 RESULTS

Application of some of the foregoing ideas has thus far been limited to attempts to smooth range and range-rate tracking data from a single tracking station. More specifically, these data were generated by the Madrid station for the ATS-6 geosynchronous satellite over an eight consecutive day period from 16 July 1975 through 23 July 1975. The method used for detecting outliers was to fit the data with various sets of simple basis vectors, with the process of fitting being optimum in the LSAD sense. Residuals or differences between the actual data and the calculated fitting function were then determined, and significant residuals were then identifed as outliers. Results obtained by this method were then compared with a semi-manual analysis of this same data accomplished by W. L. Gibson and previously reported in Reference 13.

Based largely on Reference 11, subsets of basis vectors or functions for representing the tracking data over various intervals were chosen from the following: $\{t^m$, cos(nat), sin(nat), t cos(Mat), t sin(Mat), t^2cos(Mat), t^2sin(Mat), cos(Nct), sin(Nct), t cos(Mct), t sin(Mct), t^2cos(Mct)$\}$, where m = 1, 2, ..., 7, n = 1, 2, 3, 4, M = 1, 2, N = 1, 2, 3, $a = 2\pi/T_s$, $c = \pi/T_1$, T_s = earth's sidereal period, T_1 = lunar sideral period, and t = time. The cardinal numbers of the subsets chosen ranged from 3 to 22, and the intervals of representation ranged from one to eight days. The data analyzed consisted of 43 distinct clusters of points irregularly spaced over the eight day period, and hence there was appreciable variation of data spacing and density. More details of the data distribution are given in Reference 13.

Since the results obtained are far too voluminous to be given in detail, only a summary will be given here. Qualitatively, the performance of a given basis vector set was observed to be a function of the span of the data arc and more weakly of the spacing and density of its component points. Thus larger basis vector sets were required for longer data spans, an expected result in view of the fact that long-period perturbations only become manifest in a longer span of data, e.g., lunar perturbations can be represented as a linear time function during a one day interval but are better represented as sinusoids over an eight day interval.

The weaker data spacing and density dependence accorded well with that expected purely on the basis of statistical expectations, i.e., uniform spacing of noisy data generally conveys more information about its functional form than does spacing which is highly irregular. Another noteworthy result was that basis vector sets which were too large for the quantity of data to be represented yielded results which were mathematically valid but physically meaningless. This suggested that it might well be worthwhile to try LSAD solutions with appropriate side conditions or constraints in some circumstances rather than the unconstrained solutions which clearly had too many degrees of freedom. This possibility is especially attractive in light of the fact that algorithms for such constrained solutions have recently been developed (Reference 14) by the same author, R.D. Armstrong, who kindly furnished the FORTRAN listing for the unconstrained LSAD algorithm actually used.

Some specific results are worth reporting. For range and range-rate data spanning 19-29 hours, the following two basis vector sets yielded excellent results: $\{1, t, \cos(at), \sin(at), \cos(2at), \sin(2at)\}$, and $\{1, \cos(at), \sin(at), t\cos(at), t\sin(at), \cos(2at), \sin(2at)\}$. Solutions obtained with these two sets yielded results which agreed very closely, both qualitatively and quantitatively, with those of Reference 13; this held true even for the size of the outlier residuals as determined by these completely different methods. Such comparison for longer data spans was rendered difficult because the results in Reference 13 were based on one day solutions only, and the variation in these solutions from day to day indicated that these solutions were not sufficiently cohesive to afford a basis of comparison with the results obtained by the LSAD smoothing technique for long data spans. However, fair agreement was obtained between the results of smoothing four to eight day data spans using sets of 14 to 18 basis vectors with the results obtained in Reference 13. Inclusion of basis vectors representing lunar perturbations definitely proved necessary for such longer spans. Finally, it should be mentioned that the same basis vector sets generally proved effective for both range and range-rate tracking data, though results for range data were slightly better than for range-rate. Both types of data were used in their raw form. Attempts to analyze simple functions of range and range-rate as suggested in Reference 3, such as range squared and the product of range and range-rate, did not lead to superior results, but there is every reason to believe that this suggestion would improve results for data from low altitude satellites.

5.0 CONCLUSION

The problem caused by the presence of outliers in orbital tracking data is too severe to be ignored. Outliers must either be detected and expunged before the usual OD processing, or the OD process must be modified to make it insensitive to outliers while retaining all desirable properties. Manual data editing assisted by repeated OD solutions is woefully inefficient in terms of both man-hours and computer time required, and should be considered only as a last resort. Existing methods of outlier detection and tracking data smoothing are of limited applicability and appear to suffer from a mismatch between the effect desired and the statistical techniques used to bring it about. The analysis of the problem described in this paper leads to the conclusion that strong consideration should be given to replacing some of the LS solutions with LSAD solutions. The results of some preliminary efforts in this direction as given previously appears to confirm this conclusion. Representing the tracking data by linear combinations of simple analytic functions chosen for their natural suitability from a physical standpoint and fitted by the LSAD criterion appears to lead to a fast and efficient means of outlier elimination. Because this approach is also capable of handling data arcs of widely varying lengths and densities from a single tracking station, it appears to be even better suited to the coming era of tracking by TDRSS which will supercede the existing STDN network of earthbound stations. However, the other approach of introducing LSAD solutions directly into the OD process should prove effective as well.

In light of the conclusions reached thus far, more extensive development and testing of the foregoing ideas and techniques for overcoming the problem of tracking data outliers appears well warranted and is therefore recommended. Both approaches described should be pursued, and their usefulness and efficiency should be tested and compared with a variety of both direct tracking and indirect (relay, or TDRSS) satellite-to-satellite type tracking data.

6.0 REFERENCES

1. B. Kruger, "Filter Properties of Least Squares Fitted Polynominals," NASA-GSFC, X-551-68-147, TM X-63321, N68-34567, January 1968.

2. T. J. Grenchik and B. H. Pitney, "A Review of Goddard Range and Range Rate System Measurement and Data Processing Techniques," NASA-GSFC, X-551-69-137, TM X-63572, N69-29748, April 1969.

3. T. J. Grenchik and C. W. Murray, Jr., "Smoothing of Functions of Range and Range Rate Measurements from Earth Orbiting Satellites," NASA-GSFC, X-551-72-385, TM X-66126, N73-13881, October 1972.

4. J. O. Cappellari, C. E. Velez, and A. J. Fuchs, editors, "Mathematical Theory of the Goddard Trajectory Determination System," NASA-GSFC, X-582-76-77, TM X-71106, N76-24291, April 1976.

5. G. M. Coupe', "Satellite Orbit Prediction and Determination (Error Analysis and Accuracy and Convergence Tests)," NATO SHAPE Technical Center, The Hague, The Netherlands, STC TM-370, AD-909793, February 1973.

6. H. L. Harter, "The Method of Least Squares and Some Alternatives," USAF Aerospace Research Laboratories, Wright-Patterson AFB, Ohio, ARL 72-0129, AD-752211, Sept. 1972.

7. I. Barrodale and F. D. K. Roberts, "An Improved Algorithm for Discrete L_1 Linear Approximation," SIAM J. Numer. Anal., 10 (1973), 839-848.

8. K. Spyropoulos, E. Kiountouzis, and A. Young, "Discrete Approximations in the L_1-norm," The Computer Journal, 16 (1973), 180-186.

9. R. D. Armstrong and E. L. Frome, "A Comparison of Two Algorithms for Absolute Deviation Curve Fitting," J. Amer. Statist. Ass., 71 (1976), 328-330.

10. R. D. Armstrong and E. L. Frome, "A Revised Simplex Algorithm for the Absolute Deviation Curve Fitting Problem," to be published, private communication.

11. V. J. Slabinski, "Variation in Range, Range-Rate, Propagation Time Delay, and Doppler Shift for a Nearly Geostationary Satellite," in Progress in Astronautics and Aeronautics, Volume 33, The MIT Press, Cambridge, Mass.

12. V. J. Slabinski, "Expression for the Time-varying Topocentric Direction of a Geostationary Satellite," Comsat Tech. Rev., vol. 5, no. 1, Spring 1975, 1-14.

13. W. L. Gibson, "Final Report: ATS-6 Satellite Data Editing," McDonnell Douglas Technical Services Co. - Houston Astronautics Div., Houston, Texas, Working Paper E914-8J-007, TM. No. 1.4-MPB-528, January 24, 1977.

14. R. D. Armstrong and J. W. Hultz, "A Restricted Discrete Approximation Problem in the L_1 Norm," Center for Cybernetic Studies, The University of Texas, Austin, Texas, Research Report CCS 242, July 1975.