A PROGRAM FOR DETECTING AND CORRECTING ERRORS IN LONG SERIES OF TIDAL HEIGHTS

by B. D. Zetler
U.S. Coast and Geodetic Survey

and G. W. Groves
Institute of Geophysics and Planetary Physics
University of California at San Diego, La Jolla, California

The rapid development of electronic computers has opened the door to tide investigations that previously have been physically impossible from a manpower and financial point of view. The authors are dealing with some tidal series of over fifty years, roughly one half million hourly heights in each. Obviously personal plotting and/or scanning is an impossible chore and therefore the checking techniques must use the computers. In constructing a program for this purpose, the aim was to emulate insofar as possible what a thinking human would do in the same assignment.

A human would plot and then scan the data, looking at a relatively small sample at one time and examining, more rigorously, significant anomalies from a smooth curve. In essence, he is using the group as a whole to interpolate one or more values at a time. The program therefore would examine a window (or set) of data, then move on to another window, probably allowing a small overlap to provide continuity. This essentially is the task assigned to the computer in the " ERROR " program.

Among the methods considered for the scanning procedure were polynomial approximation for various sized windows, the subtraction of predicted tides (using harmonic constants), and the consideration of the autocovariance structure of the data to determine Wiener type predictors or interpolators by minimizing the mean square error of prediction.

The last method was found to be the most satisfactory. What is more, very little additional improvement resulted from first subtracting the predicted tide (and thus producing a residue whose energy was a very small part of the original energy) before applying the Wiener method to the residue.

Use of the Wiener method assumes (1) that the time series is stationary, (2) that the significant measure of accuracy is the mean square discrepancy, and (3) that the prediction is to be based on a linear operator. The first assumption, that the statistical properties of the record do not change with time, is ordinarily true only for tide observations devoid of severe storms, tsunamis, etc. The criterion of minimum root mean square error admits small discrepancies and discriminates against large errors; this would appear highly desirable in the present problem. The use of a linear operator reduces computing time; it hardly seems worthwhile to use a more elaborate scheme.

Superimposed on the assumptions implicit in the Wiener predictor is a further assumption that, in a given sample of M values being considered at one time, not more than one of the values may be considered wrong. It might have been assumed that at most N values could be in error, and each possible combination of N values then could have been predicted from the remaining M-N values in each window. In this case the computing time would have been greatly increased and it is not at all certain that the accuracy of predictions would be any better. A somewhat similar result was achieved by the following scheme. Any time a value is replaced, each value within the same window is re-examined. This scheme is repeated until all the values are within tolerance, at which point the program moves on to the next window.

The program takes a significant sample, ordinarily 7 500 values, averages the sample, subtracts the mean from each value and then calculates the autocovariances for lags from 0 to M—1 hours. The autocovariances supply the necessary data for a matrix solution of M equations in M unknowns for each of the M positions in the window, the resulting factors being determined as those that produce a minimum root mean square error of prediction.

A listing of the program as it was used with a 16-year series of sea level hourly heights at San Francisco will be furnished upon request. The series is too long for storage in the memory of the computer and therefore various subroutines for reading a record at a time but retaining needed data for backward slides had to be improvised. Six parameters, that permit considerable flexibility to the user of the program, are read in at the beginning. These (LAUTSR, M, NADV, L, THETA, and THETM) are explained in the text that follows.

A sample of hourly tide data, consisting of the first LAUTSR values, is used to calculate the autocovariances and determine the interpolator weights. Note that LAUTSR must not exceed the memory size of the computer nor the length of the series. The first LAUTSR values are then corrected. The corrected LAUTSR values are then used again to re-evaluate the autocovariances and interpolator weights, which are subsequently used to correct the entire series. In this way, bad data are not allowed to seriously spoil the effectiveness of the interpolator scheme.

To examine and correct the data, a small " window " consisting of M consecutive values is considered at any one time. Each value within this window (with the exceptions noted below) is compared to the value the interpolator computes on the basis of the remaining M-1 values. The entire window is examined before making any substitutions. If any discrepancy exceeds THETA times the root mean square expected discrepancy, then the interpolated value is substituted for the value having the greatest discrepancy. Each time a value is changed, the entire M values are re-examined. As soon as all the M values have passed the correctness test, another set of M values are examined in the same way. NADV is the number of data values by which the window advances along the series. That is, values 1, 2, ..., M are examined. Then values 1 + NADV, 2 + NADV, ..., M + NADV, then values 1 + 2NADV, ..., M + 2NADV, etc., until all the data have been examined.

Interpolation is used rather than prediction because of its greater precision. Thus, the end values of a window are never replaced. However, they are examined in order to keep bad values from influencing the examination of the other values. Actually, as the program now stands, neither the last nor the penultimate value is replaced. If either of these values fails to pass the correctness test, *and* has the greatest discrepancy, then the window is shifted backwards by one data value and re-examined. For comparison purposes, the error of prediction by a human computer was tested, one of the authors looking at a progressive series of M—1 consecutive hourly heights and predicting the Mth value. Many human predictions were good but some were miserably bad, the human RMS error being appreciably worse than that of the computer.

At any time the values in the window turn out to be correct, or can be made correct by replacements, then the advance of NADV is resumed. Otherwise, repeated backward shifts will occur until data previously examined occupy the window, making it useless to proceed further in this way. At this point the window is shifted forward just one value ahead of the point where the backward shifts began, and if the difficulty with the end values persists, then forward advances of one value are made until replacements can be made or good data found.

According to this procedure, it is possible that a series of bad data may be skipped without any replacements being made. However, the user will have knowledge of this from the printouts.

If NADV is less than M, then values near the beginning of the window will have been examined in the previous window. Whether or not some overlap is desirable is left to the user. L specifies the number of values to be re-examined in each window. For example, if L = 0, then each value would normally be examined just once.

Because prediction is so much less precise than interpolation, the RMS expected error for the end points of the window are much greater than for the interior points. For this reason, the user is allowed to select a different criterion for correctness of the end points. For interior points, values will be considered erroneous if they differ by more than THETA times the RMS expected error from their corresponding interpolated values, while the end points will be considered erroneous if they differ by more than THETM times the RMS expected error.

It was found that the predictor weights sometimes attained values close to or exceeding unity. This seems more likely to occur for the predictor of the end points than for the interpolator of an interior point. Whenever a weight exceeds unity, an erroneous value (which happens to be paired with the weight in the prediction scheme) could make a good value (the one being predicted) appear worse than the erroneous value. For this reason it is desirable to determine the predictor in such a way that the absolute value of any weight is not too large, or at least does not exceed unity. To state the criterion properly would entail quite a complicated procedure, possibly an iteration method. A more direct method was used. The problem is not posed properly and it is more difficult and uncertain for the user, but the procedure is greatly simplified.

Instead of minimizing $< (x_i - \hat{x}_i)^2 >$ to determine the predictor, $< (x_i - \hat{x}_i)^2 > + \lambda \overline{w_j^2}$ is minimized. Here x_i are the data values, $\hat{x}_i$ are the predicted or interpolated values, $< >$ indicates the expected value (for each position within the window), w_j are the weights, and $\overline{w_j^2}$ is the mean square value of all the weights. If $\lambda = 0$, then the straight Wiener criterion is used. If λ is very large, then the weights are forced to lie close to zero even though this gives a poor prediction. If the tidal heights are in feet and the tidal range is ordinary, the value, $\lambda = 1$, used in the program illustrated for the end positions is acceptable. However, if the units used are not in feet or the tidal range is unusual, lambda should equal about 20 % of the variance. The variance can be approximated readily from the tidal harmonic constants as one half the sum of the squares of the amplitudes of the various constituents. Lambda is set equal to zero for the interior positions. The value of lambda in the program can be changed by modifying the control card specifying " AL " in statement 101 of subroutine DETCNV.

Suggested values of the parameters

(length of sample)	LAUTSR = size of available memory or the length of the series, whichever is smaller.
(length of window)	M = 13 for the predominantly semidiurnal tidal data. However, M = 25 gives substantially better results when there is a strong diurnal inequality.
(advance)	NADV = somewhere between 1/2 and 3/4 of M.
(overlap)	L = 1 to 3, 3 the more conservative. Under no circumstances should L exceed M — NADV — 2.
(allowed standard deviations)	THETA = 5
(allowed standard deviations for end point)	THETM = 3

Other parameters to be specified are indicated in comment cards in the FORTRAN program and the way they are to be read in is indicated.

While testing the program, the changes made in the data were scanned by the authors to insure that the program was in fact achieving the desired result. Furthermore, the Coast and Geodetic Survey re-examined the original records for three years of San Francisco tide data and one year of Willets

Point data to check the validity of the changes. The program evolved over a period of time as its output was evaluated and more and more data were reviewed.

When the computing program is completed the human must take over again, primarily to seek out windows where multiple errors in original data may have caused complications. Inasmuch as the output identifies the number of changes in a given window, these are easy to locate. The heights in the window and, when necessary, adjacent heights are plotted and the changes evaluated. Some examples of multiple adjacent bad values spoiling good values in the window were found. Inasmuch as the output of the ERROR routine is on tape, programs for inserting necessary additional changes have to be prepared. This involves making a new tape, retrieving most of the old data and including proper changes in correct places.

The tolerated difference between observed and predicted values for San Francisco (using THETA = 4) varied between 0.3 and 0.4 foot except in the end positions. The Coast and Geodetic Survey review of the changes for three years of record found that changes of about this amount were sometimes due to the hour marks being off the mean curve due to seiche, or to an intermittent change in the curve characteristic (a falling tide dropping more sharply than usual), and therefore were not errors at all. Going back to our original standards, our studies can tolerate errors of this size as they are not significant unless they are very numerous. Nevertheless, future programs will use a slightly larger tolerated difference, THETA = 5 rather than 4 and THETM = 3 rather than 2.5. The review also showed some instances of multiple adjacent or nearby errors where the human inspection failed to identify properly the erring values. In one case, 22 consecutive hours had been tabulated 0.5 foot too high; the program smoothed the end of the section but did not change the beginning or intermediate values. Admittedly then, multiple adjacent errors may not always be corrected by the program; it has been demonstrated that a human may also fail in this area. The authors believe the program is a workable and useful method of always correcting an isolated large error and, in most cases, of detecting and reporting multiple errors that are close together in time. We hope that it is also applicable to other types of geophysical time series but this has not yet been tested.

The computing time of the program, taking about 16 years of hourly heights, a window (M) of 25 hours, an advance (NADV) of 18 hours, and an overlap (L) of 3, was about 30 minutes on the CDC 1604 at the University of California at San Diego. This used a data input on magnetic tape (BCD) and furnished a corrected tape and a listing and identification of the changes made.

The authors are grateful to Prof. Walter MUNK, University of California at San Diego, for interesting us in the problem, to Prof. MUNK and Mr. George HEXT, Stanford University, for many helpful discussions, and to Capt. Kenneth ULM, Chief, Marine Data Division, U.S. Coast and Geodetic Survey, for evaluating the program output by arranging checks of original marigrams.

This work was supported by U.S. Navy contract Nonr 2216 (12).