Accounting for random errors in linear regression: A practical guide

By E. C. KENT* and P. K. TAYLOR
Southampton Oceanography Centre, UK

(Received 5 May 1998; revised 28 April 1999)

KEYWORDS: Observation error  Regression  Statistical methods

We read with interest Tolman's (1998, hereafter T98) analysis of the effect of observation errors on validation of marine winds. His demonstration of the importance of the correct treatment of the errors in comparative analysis of wind-speed (and other types of) data is welcome, as examples of inaccurate comparisons are common, particularly for satellite data validation which is often published in non-refereed reports. Kent et al. (1998) show that apparent trends of the order reported can be directly attributed to errors in both datasets and that the trend appears smaller and in the opposite sense if the satellite wind speed rather than the ship wind speed is used as the independent variable for plotting.

T98 states that 'in special cases, where the ratio of . . . errors can be estimated, more advanced regression techniques can be used'. Our aim in this note is to advocate a simple, established method of data analysis which can lead to reliable comparisons of pairs of nearly co-located and simultaneous observations from two sources, both containing random errors. This method enables the use of these advanced regression techniques since the ratio of random errors in the datasets to be compared is estimated. The error estimates can then be verified by using the effects of errors on bin-averaged analyses highlighted by T98. Following T98 and Kent et al. (1998), we shall discuss satellite and in situ wind-speed data comparisons, but again expect the results to be more widely applicable.

(a) Estimation of errors in both datasets

For spatially distributed datasets (e.g. from Voluntary Observing Ships (VOS) or satellites) the semi-variogram technique can be used (Lindau 1995; Kent et al. 1999) to separate the random error in the observations from the spatial variability. For all nominally simultaneous pairs of observations the mean squared difference is plotted as a function of spatial separation. The mean squared difference can then be extrapolated back to zero separation to obtain the intrinsic random error variance within the observation pairs excluding any spatial error variance. A division by two is necessary to get the mean squared error in an individual observation, and then a square root will yield the root-mean-square random error. As any spatial variability is removed, this extrapolation gives a better estimate of the error than simply using observations within a certain separation. More data may be used without degrading the error estimate. A similar technique can be used for time series data (e.g. from a buoy or Ocean Weather Ship) by plotting mean squared differences as a function of time difference between the observations and extrapolating to zero time difference. Alternatively the root-mean-square difference from a slowly varying mean can be used as an estimate of the random error; however, in this case, the resulting error estimates should be checked by a binned comparison as detailed below and adjusted if necessary.

(b) Regression

As noted by T98, once the random error has been determined an unbiased regression can be obtained in various ways (Graybill 1961).

(c) Validation of error estimates

Once the estimates of the random error have been determined (for example, as above or from the literature), the estimates can be verified using a bin-averaged analysis (Kent et al. 1998). T98 demonstrates that the behaviour of bin-averaged data are predictable and proposes a method of correction. We suggest that this predictable behaviour can be exploited to verify the error estimates. An independent dataset is required, this may be a set of observations or simulated. The only requirement is that the distribution of wind speed has broadly similar characteristics to the observation dataset. Two simulated sets of paired wind
speeds are then generated by adding random errors to each observation in turn. The two sets differ in that the errors should be sampled from distributions having widths defined by the error estimates obtained for the two data-types being simulated. If the bin-averaged analysis of the differences of pairs of observations in the simulated dataset shows similar behaviour to that of the observation dataset, it can be concluded that the error estimates obtained reasonably represent random errors within both sets of observed winds speeds. Kent et al. (1998) present an example of this technique which confirms T98's conclusion that the errors due to incorrect analysis can be large. They found that for a comparison of VOS and ERS-1* scatterometer winds the VOS wind-speed data had a random error four times greater than the scatterometer wind speeds, and the gradient of the unbiased regression was 20% different from the geometric mean regression.

REFERENCES


* European Remote-sensing Satellite.