Conf-9205190--1

PNL-SA--19989

DE92 015253

REAL-TIME STATISTICAL QUALITY CONTROL AND ARM

D. K. Blough

May 1992

Presented at the
American Society for Quality Control 46th
Annual Quality Congress
May 18-20, 1992
Nashville, Tennessee

Work supported by
the U.S. Department of Energy
under Contract DE-AC06-76RLO 1830

Pacific Northwest Laboratory
Richland, Washington 99352

REAL-TIME STATISTICAL QUALITY CONTROL AND ARM

David K. Blough
Senior Research Scientist
Pacific Northwest Laboratory
Richland, WA 99352

ABSTRACT

An important component of the Atmospheric Radiation Measurement (ARM) Program is real-time quality control of data obtained from meteorological instruments. It is the goal of the ARM program to enhance the predictive capabilities of global circulation models by incorporating them more detailed information on the radiative characteristics of the earth's atmosphere. To this end, a number of Cloud and Radiation Testbeds (CART's) will be built at various locations worldwide. Each CART will consist of an array of instruments designed to collect radiative data. The large amount of data obtained from these instruments necessitates real-time processing in order to flag outliers and possible instrument malfunction. The Bayesian dynamic linear model (DLM) proves to be an effective way of monitoring the time series data which each instrument generates. It provides a flexible yet powerful approach to detecting in real-time sudden shifts in a non-stationary multivariate time series. An application of these techniques to data arising from a remote sensing instrument to be used in the CART is provided. Using real data from a wind profiler, the ability of the DLM to detect outliers is studied.

TEXT

ARM and the Need for SQC

Real-time statistical quality control of data coming from meteorological instruments is a vital component of the Atmospheric Radiation Measurement (ARM) program. The dynamic linear model (DLM) is a powerful tool suited to processing large volumes of this type of data as it is collected.

The ARM Program is a long-term investigation initiated by the U.S. Department of Energy (DOE) to collect and evaluate data on radiative characteristics of the Earth's atmosphere. DOE is concerned with the possible ramifications of global warming on environmental conditions and therefore, the appropriate response in national energy policy. In order to assess and forecast the effects of increasing greenhouse gases in the Earth's atmosphere, global general circulation models of the atmosphere are used. It is the purpose of ARM to increase the predictive capabilities of these models by including in them the appropriate radiative physics developed from the data collected by the ARM program. The first ARM site is scheduled to be operational in Spring of 1992. It will consist of, among other things, a large array of meteorological instruments (the so-called Cloud and Radiation Testbed, or CART) collecting vast amounts of data virtually continuously.[1]

It is the purpose of this paper to discuss real-time statistical quality control of data coming from the instruments of CART. Due to the large amounts of data that will be generated by CART (one terabyte per day), it is imperative that a real-time, automated quality control scheme be implemented. It is apparent that the quality control techniques implemented in CART will be unique in two ways: the usual statistical quality control techniques used in industry are not applicable here; in fact, the goal of maintaining the process within acceptable limits is not relevant because the process (the atmosphere) is not under our control. Rather, the goal is to monitor the instruments in real time, allowing for non-stationary behavior, and yet providing information on "anomalous" behavior. Additionally, the fact that the data arrive as a time series (univariate, vector-valued, or matrix-valued) must be taken into account; that is, time autocorrelations need to be incorporated into the quality control scheme to enhance its performance.

Quality control will be implemented at multiple points in the CART data environment. One point will be in the conversion of Level 0 (L_0) data to Level 1 (L_1) data. L_0 data is essentially "raw" data from the instruments. The implementation of quality control procedures at this point will detect spurious data (outliers), and instrument failure (including slow or sudden shifts out of calibration). Procedures will then flag suspect data, provide notification to users or operators, and possibly rectify the data; the resulting data is then considered L_1 data. These techniques will also be used in the conversion of L_1 data to L_2 data. This involves the comparison of instrument output with other instruments measuring the same meteorological variables, comparing output with computer models of the instrument, and correlating the output with similar instruments measuring related variables.

Introduction to the Bayesian DLM

In order to assess the quality of incoming data, a model-based approach is necessary, and in particular the sequential generation of one-step ahead forecasts with known distributional properties is needed for real-time processing. The Bayesian dynamic linear model satisfies these requirements and the specific model used can be reasonably generated by simple assumptions about the physical behavior of the atmosphere (locally constant in time). The Bayesian DLM is a state space representation of a time series that can accommodate univariate, vector or matrix observations; it can model non-stationary time series and take time autocorrelations into account. Estimation of model parameters is done recursively (Kalman filtering), and in the Bayesian context can be viewed as a forecast step made prior to an observation being taken, followed by an update step made posterior to an observation being taken. It is thus suited to real-time applications.[2]

DLM methodology includes that of dynamic regression, transfer functions, seasonal time series, the non-linear DLM, noise models, and polynomial trend models. It is particularly useful for automated intervention analysis: when anomalous data is encountered (this is determined via a likelihood ratio-type test), the model can be made to signal the change, automatically adapt to the change, and then continue tracking the data. More complex models are available; these are the so-called multi-process DLM's. These models explicitly unite a collection of models, each of which has a prior probability of being the one generating the data at a given time. The forecast is then a weighted average of the forecasts generated from each individual model, posterior probabilities being the weights. The nature of the collection of models under consideration is determined by the types of anomalous behavior one is interested in modeling: outlier, sudden level change, gradual

drift, etc. Multi-process models not only allow for tracking of the series itself, but also tracking the probabilities of outlier, sudden level change, etc. Hence, for quality control purposes, they prove to be highly informative.[3]

For the purposes of this paper, the likelihood ratio-type monitoring scheme will be discussed. Following a brief description of the model, its application to real data will be presented. The data comes from a remote sensing instrument of central importance to CART--the wind profiler. A description of the instrument will be given along with data obtained from a profiler in Colorado. Finally, the performance of the model on the data will be discussed.

Application of the Model to SQC

The Bayesian DLM is a state space formulation of a time series model. Thus, it consists of two components, the system equation and the observation equation:

$$\Theta_t = G_t \Theta_{t-1} + \Omega_t$$

$$Y_t' = F_t' \Theta_t + v_t'$$

The notation and development given here are consistent with that of West and Harrison[4]. For a more detailed view of the DLM, this text is recommended. The matrix Θ_t is known as the state matrix at time t, and Y_t is the observation vector at time t. The matrices G_t and F_t are assumed known at each time t. The matrix Ω_t is a random matrix assumed to have a matrix normal distribution; the vector v_t is a random vector assumed to have a multivariate normal distribution and be independent of Ω_t for all t. The variance-covariance matrix of v_t is denoted by Σ, assumed to be constant over time. A conjugate prior is assumed for Σ (an inverse Wishart distribution) and its estimation is done recursively as each observation arrives using standard Bayesian analysis. As observations Y_t arrive, estimation of the state matrix Θ_t is done via the Kalman filter. This can be viewed as an application of Bayes theorem. One of the important products of this model formulation is a one-step ahead forecast. When the next observation arrives, it can be compared to the forecast value and if the two are too different, an outlier can be flagged. The actual criterion used to determine the degree of similarity between the forecasted and observed value is called the Bayes factor; it is a likelihood ratio test to ascertain if the variability in the observation has significantly increased. The distribution of the difference between the forecast and the observation is known to have a multivariate-T distribution. Hence, the criterion is

$$h_t = (k^q)\left((1 + \frac{e_t R_t^{-1} e_t'}{k^2 n_t}) / (1 + \frac{e_t R_t^{-1} e_t'}{n_t}) \right)^{((n_t+q)/2)}$$

where e_t is row vector of differences between the one-step ahead forecast and the observation, R_t is the estimated variance-covariance matrix of e_t, q is the number of elements in the vector of observations Y_t, n_t is degrees of freedom (the number of observations at time t minus 1), and k is the factor by which the standard deviation is believed to have increased, causing a potential outlier. The choice of

k is not crucial; usually k in the range of 1.5 to 4 is sufficient for most data sets.

When h_t is small, this is an indication that an outlier has been encountered. A flag is set and then the system noise is increased to allow for possible adaptation of the model to future observations. In this way, the model tracks the series, signals possible outliers, and then adapts to these sudden changes in the series to continue further tracking.

It is necessary to specify the precise form of the model; that is, the structure of the system and observation equations needs to be determined. Many of the instruments on CART collect data over time with relatively short intervals between each observation. For example, the wind profiler data to be presented below represents wind velocities collected at six-minute intervals. For this reason, a locally constant model will be assumed. Thus,

$$\Theta_t=(\mu_{t1}\ \mu_{t2}.....\mu_{tq})$$

where

$$Y_t'=(Y_{t1}\ Y_{t2}.....Y_{tq})$$

Also, $F_t=1$ and $G_t=1$. Finally,

$$\Omega_t=(\omega_{t1}\ \omega_{t2}.....\omega_{tq})$$

$$\nu_t'=(\nu_{t1}\ \nu_{t2}.....\nu_{tq})$$

In terms of the i^{th} component Y_{ti} of the observation vector Y_t, we thus have the state space formulation

$$\mu_{ti}=\mu_{(t-1)i}+\omega_{ti}$$

$$Y_{ti}=\mu_{ti}+\nu_{ti}$$

This simple model will be shown to be quite effective at detecting outliers in multivariate time series.

Example: The Wind Profiler

CART consists of a suite of meteorological instruments. The wind profiler is considered here because it is one of the most difficult instruments in terms of developing a data quality control scheme. A profiler is a remote sensing instrument which uses radar to construct a vertical profile of wind speed and wind direction in a vertical column of the atmosphere directly over the instrument. This column

typically is from 5 to 15 kilometers in height. Radar pulses are sent and received by each of three beams; two beams are tipped at an angle of 16.3 degrees from the vertical, one to the east and one to the north. The third beam is oriented vertically. Observations consist of 32 radial velocities equally spaced along each beam. Each beam produces such a profile every six minutes. Using trigonometry, it is then possible to combine the information in the three beams to produce a vertical profile of horizontal wind speed and direction.[4]

The data from the three beams is collected sequentially; that is, first the east beam is activated (it takes two minutes to then construct the profile), then the north beam (two more minutes), then the vertical beam (two more minutes). This cycle is then repeated. Hence, observations are obtained from a given beam every six minutes. For this reason, the quality control algorithm above is applied to each beam separately. This real-time approach differs from more traditional after-the-fact, non-stochastic profiler editing schemes.[5] As an example of such data, Figures 1 through 3 portray the output of the wind profiler located in Platteville, Colorado from 10:06 a.m. to 2:00 p.m. on December 16, 1990. Thus for each beam we have 40 observations, each observation being a vector of 32 radial velocities.

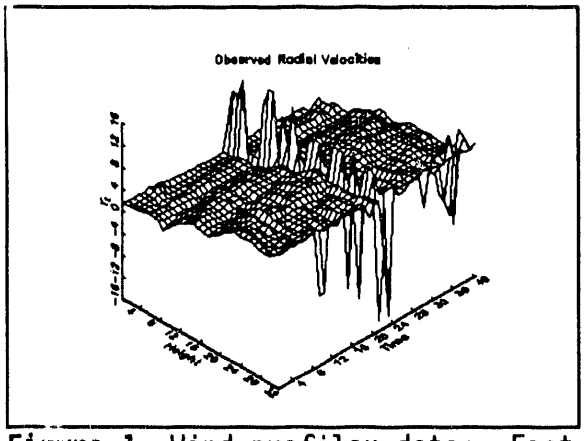

Figure 1 Wind profiler data: East beam

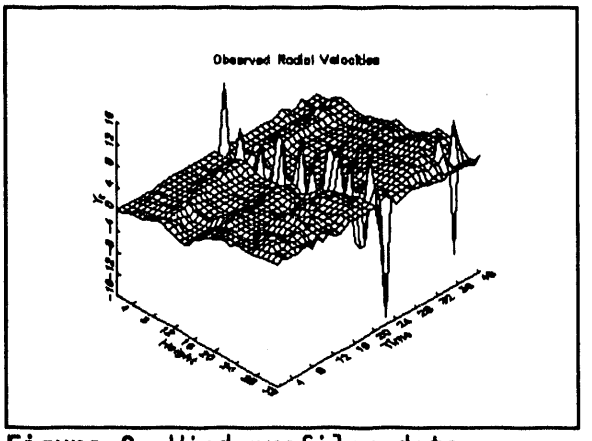

Figure 2 Wind profiler data: north beam

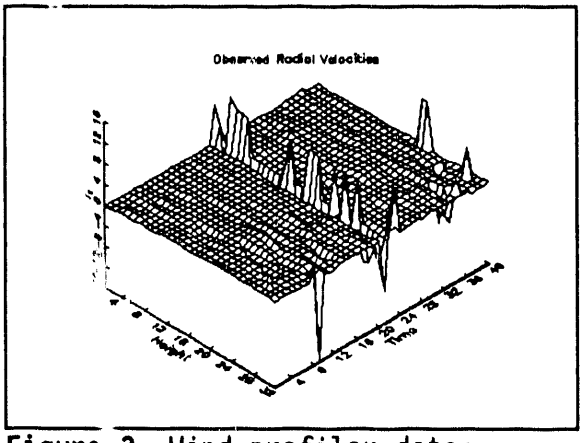

Figure 3 Wind profiler data: Vertical beam

The above model was applied to this data, reading in one observation at a time to simulate real-time processing of profiles. The likelihood ratio-type test (that is, the Bayes factor h_t) was computed as each profile arrived. Thus, the observed profile was compared to the forecasted profile using the multivariate-T Bayes factor. If an outlier was flagged, univariate tests were then performed in order to ascertain at which height(s) the outliers occurred. For each height, the absolute value of the radial velocity forecast error divided by its standard deviation (the square root of the corresponding diagonal element of the matrix R_t) was compared to 2. If the standardized error was larger than two, the radial velocity at the corresponding height was flagged. The resulting output was obtained:

EAST BEAM

Outliers detected at times (multivariate test)	Corresponding heights (univariate tests)
9	32
10	32
22	4, 8, 9, 10, 25, 27, 31, 32
23	32
31	individual tests failed to show significance
35	individual tests failed to show significance
36	31
38	individual tests failed to show significance
39	29

NORTH BEAM

Outliers detected at times (multivariate test)	Corresponding heights (univariate tests)
21	2, 28
22	32
35	individual tests failed to show significance
38	individual tests failed to show significance
39	28

VERTICAL BEAM

Outliers detected at times (multivariate test)	Corresponding heights (univariate tests)
9	32
21	32
34	individual tests failed to show significance
35	individual tests failed to show significance
38	individual tests failed to show significance
40	29

It can be seen by comparing these results with the graphic display of the data that for each beam the model detects and correctly flags nearly every outlier. The superiority of the multivariate approach to outlier detection used here is evident here since individual tests sometimes fail to flag outliers. However, as is often the case in multivariate analyses, ascertaining which component of the observation vector is outlying can be difficult.

FOOTNOTES

1. Atmospheric Radiation Measurement Program Plan, 1990, p. 30-37.
2. West, Harrison, Bayesian Forecasting and Dynamic Models, 1989, chapters 11 and 15.
3. Hubele, "A Multivariate and Stochastic Framework for Statistical Process Control," 1989, p. 129-152.
4. Brewster, "Quality Control of Wind Profiler Data," 1989.
5. Wuertz, Weber, "Editing Wind Profiler Measurements," 1989.

CONCLUSION

The Bayesian dynamic linear model is powerful tool for real-time processing of data obtained from meteorological instruments. It is clear however that these techniques can be applied to any process or production scheme in which data is arriving in the form of a multivariate time series. If observations are infrequent, a higher order model than the one used in this paper can be used; one which might include linear or quadratic trends rather than assuming locally constant trend. There are a number of tuning parameters involved in the above algorithm (for example, the scale factor k), and further work needs to be done to ascertain the effects of varying these parameters on the performance of the model. It is also important to check residuals during periods of in-control behavior in order to assess the assumptions of normality and constant variance-covariance matrix.

BIBLIOGRAPHY

Brewster, Keith A. Quality Control of Wind Profiler Data. Profiler Training Manual #2. Boulder, Colorado: NOAA, 1989.

Hubele, N.F. "A Multivariate and Stochastic Framework for Statistical Process Control." In Keats, J.B., and N.F. Hubele (eds.). Statistical Process Control in Automated Manufacturing. New York: Marcel Dekker, Inc., 1989.

U.S. Department of Energy. Atmospheric Radiation Measurement Program Plan. National Technical Information Service. Springfield, VA: U.S. Department of Commerce, Februray, 1990.

West, Mike and Jeff Harrison. Bayesian Forecasting and Dynamic Models. New York: Springer-Verlag, 1989.

Wuertz, David B., and Bob L. Weber. Editing Wind Profiler Measurements. Technical Report ERL 438-WPL 62. Boulder, Colorado: NOAA, 1989.

LCS: 545:70:991

END

DATE FILMED

7/28/92