ANALYSIS
OF
LARGE SCIENTIFIC DATA SETS

D. J. Downing V. V. Fedorov J. Finger R. E. Flanery
W. F. Lawkins R. Lee M. D. Morris E. M. Oblow
G. Ostrouchov

Account: 3321-0720
Computer Science and Mathematics Division

EXECUTIVE SUMMARY

The goal of the project was to begin development of a new approach to the statistical analysis of scientific data, appropriate for and applicable to the very large data sets now being produced by advanced measurement systems and computational models. Because most analytical and even graphical paradigms of data analysis in current use have been developed in contexts motivated by smaller data structures, they are often impractical or ineffective when the data set of interest is of gigabyte or terabyte size. The point of this work was to develop methods which would identify those segments of a very large data set which are most unusual in one or more significant characteristics. Our specific focus was on very long time series, both measured and generated by computer models, where interest is in finding time intervals corresponding to changes and/or perturbations in the monitored or simulated process. An additional goal was to develop preliminary software which would implement the techniques developed.

The primary products of this work include the development of statistical procedures, primarily for detecting changepoints and classifying process states, which are computationally feasible for very large time series data sets. In the early stages of this project, we had hoped to enlarge this effort to include techniques for highly

structured multivariate time series problems, such as long sequences of digital images. However, it became clear to us that the first and perhaps greatest challenge in establishing a new approach for analysis in this class of problems involved efficient treatment of the time dimension, and that became our ultimate focus. We did, however, generalize one method (Bayesian changepoint detection) for application to multivariate problems of relatively modest size, implemented another multivariate method based on Mahalanobis distance, and our most recent work has led us to understand how other methods (nonlinear dynamics and regression approaches) can be formulated for multivariate problems as well. Software development did not proceed as far as we had initially hoped, but implementations of all methods were produced using AVS5, an enhanced computer visualization environment.

The work supported by this project has resulted in three ORNL technical reports. In addition, a demonstration of our software was presented at a technical meeting, a talk describing our work was given at another, and a second talk is planned for an upcoming meeting. A proposal entitled "Enhancement of Statistical Methods and Algorithms for Remote Sensing" was submitted to the DOE NN-20 program, but was not funded due to budgetary constraints.

1. SCIENTIFIC AND TECHNICAL BACKGROUND

Improved measurement and computing technologies have led to rapid growth in the size of scientific (and other) data sets. Most conventional methods of data analysis are inadequate either because they are not flexible enough to treat the complex structure of many such data sets, or are too computationally demanding to be practical. This latter point, the sheer size problem, has been the focus of this project. Many conventional data analysis techniques which require modest computational power for data sets of 100 values are for practical purposes impossible to use with data sets of 10,000,000 values because the computational effort or storage required often increases much faster than linearly with the sample size. In many cases, the data sets which require analysis are too large to allow complete storage on all but specialized collections of computer mass storage devices.

This dramatic increase in the size of many scientific data sets has resulted from correspondingly dramatic improvements in measurement and computing technologies. Remote sensing systems, such as satellite imaging networks, now routinely produce so much measurement data that transmission and storage have become limiting operational capabilities. At the same time, the "data" used in many scientific studies are now generated by computer models of physical systems rather than the physical systems themselves, A model of ocean general circulation, for example, may produce output containing several physical values at each of several hundred or thousand physical locations for one or more times in each simulated day, where simulations covering 1000 years may be of interest.

Hence both measured and simulated data sets now exist which far exceed sizes which, at least in practice, mark the edge of practical utility for many commonly used analysis tools. Even when clever and effective graphical procedures are employed to consolidate numerical information into a more "visual" format an overwhelming array of changes, many of them redundant or unimportant, is produced. The amount of data which must be scanned quickly dilutes the ability of the investigator to confidently identify all the meaningful patterns and trends which may be present.

The aim of our research has been the initial development of statistically motivated techniques which can be used to "select" a relatively small subset of such a large data set for examination by the investigator. The intent is that an investigator who uses conventional techniques to examine only the data in this subset can gain a sufficient understanding of the important features of entire data structure.

2. PROJECT OBJECTIVES

The problem of "reinventing data analysis" for arbitrarily large or complex data structures is, of course, far too general for any research project. Our first step in defining a more specific problem was a decision about the sense of "large" we would consider. In many large data sets, records are indexed by an extensive system of variables which can themselves be quite complex in their definition. Examples include library catalogues which are based on primarily textual information and categorical keys, and financial transaction databases which contain numerical, categorical, and textual information, indexed by location and time of transaction. Data sets of interest in scientific studies often follow a more "physical" structure, and may contain one or a number of quantitative variables, indexed by time and/or location. For example, computational models of environmental groundwater migration produce simulations defined by physical variables at each of a (typically) large number of spatial grid point, over a (typically) large number of time steps. However, for the purposes of this study, an even narrower scope was required to allow us to define a generic problem of broad utility. We specifically selected problems involving relatively few variables (or physical locations of a single variable) recorded over very long time series as the general form of data set we would address. While this does not include every scientific problem of interest, it does represent many of importance. Additionally, because the first step in analyzing large multivariate (e.g. image) data is often dimension reduction to a smaller number of values, methods which concentrate on a few variables in long time series may also be useful in analyzing highly multivariate problems after some preprocessing.

The second step in defining the focus for this project was in selecting the particular variety of "analysis" to be considered. Taken broadly, since the range of potential questions or hypotheses to be entertained is practically without limit, the class of potentially applicable analyses is practically beyond a single definition. Our focus on large data sets, and in particular data sets which owe their size primarily to repeated values over time, suggests that in many cases the first analytical activity of practical interest will be the identification of times or time intervals in which the observed or simulated process "changes" or is temporarily "perturbed" relative to its "normal" behavior. Appropriate definitions of these terms are not particularly obvious in this context; even "normal" behavior may be expected to change over the course of time. The methods we have developed, described in the next section, are based on different modeling schemes which tacitly but precisely define "change", "perturbation", and "normal" in different ways, each of which will be more appropriate than the others in certain situations.

Following these ideas, the objectives of this research, stated in terms which are only slightly modified from our proposal, are to develop methods and computer tools which will allow computational scientists to identify sequences of "critical" time-frames of output or measured data for careful examination or dynamic visualization. These time-frames will be selected based on the corresponding data being

1. an "outlier", or fundamentally different from that at other times,
2. the beginning or ending of a significant trend or other feature in the output, and/or
3. sufficient to allow appropriate understanding of what is represented in the time-frames not selected.

We have selected this as a logical starting point for the development of new analytical paradigms for very large data sets, for the reasons stated above. This clearly leaves much to be done, beginning with different senses of "large" and "analysis" than those selected here. Another important feature of very large data sets, which is not addressed at all in our work but which is at least as important as sheer size, is the issue of data set complexity. This is especially true in "meta-analysis" settings where data sources of different origins must be simultaneously considered. Nevertheless, we continue to view our stated objectives as a logical and appropriate beginning.

3. SIGNIFICANT RESULTS AND CONCLUSIONS

3.1. METHODS DEVELOPED

In the course of this investigation, a number of statistical methods have been developed to address the objectives of identifying important features of very long time series data structures. Some of these are modifications of methods which have been in use for some time, while others are the result of our own basic research activities during this period. Our intent has been to assemble a collection of methods which are complimentary, and which taken together provide a powerful set of tools useful in a variety of applications.

The following is a brief description of each method developed in the course of this work.

Bayesian changepoint detection

Our research in this area is based on the working assumption, common in much of the statistical literature on time series analysis, that the observed data are generated by a stationary stochastic process. Our approach to detecting a change or unusual segment of data is by comparison of two "predictions" of the series, one based on data from earlier time points (e.g. a forecast) and the other based on data from later time points (e.g. a "backcast"). A statistical test has been developed to assess the degree of difference in quality of these two predictions; significant differences are interpreted as evidence of a change in the underlying process. We have generalized this method for multivariate settings, i.e. data structures of parallel correlated time series.

Nonlinear dynamics changepoint detection

Our work here is motivated by ideas related to recent research in chaotic time series analysis. The underlying assumption is that the observed process (time series) generally reflects what could be called a “normal” or quiescent mode, but from time to time departs from this pattern during less stable or “excited” periods. The process is modeled in state space, generally an augmented version of the minimal space needed to represent the data, and the “attractor” which defines normal behavior of the process is identified from the data. Following the process as a “trajectory” in state space, excited periods are identified as those during which the process behavior is not reflected by the attractor. The idea can be extended to processes which are nonstationary, e.g. for which the definition of “normal” changes with time, by allowing the attractor to be time-dependent.

Dynamic Linear Model

Kalman filtering [11] is an established methodology based on recursive estimation which allows the parameters of the system state to vary with time. By appropriately selecting different possible systems (i.e. models of the state of the process) one can effectively model various scenarios including trends or jumps. The Kalman filter can be used to detect changes in level or trend through model parameters and, in addition, the probability of each scenario being the true (as opposed to the apparent) state of the system can also be obtained. Our use of this analysis paradigm has been through the more more recent “dynamic linear model” of [8], which is specifically tailored to identify situations in which a time series is not changing structurally, is changing in level, is changing in slope, or is displaying a transient change of state.

Regression-based data summarization

Linear regression is one of the most widely used of all statistical techniques. Some of our recent work has been in experimenting with new applications of this method for identifying important segments of very large data sets. Our approach utilizes local regression, i.e. regression models fitted to subsets of data, as a method of dimensionality reduction so as to allow more focused identification of unusual segments. We have examined how various forms of process perturbations affect local regression analysis, effects of window selection, and some of the relationships between regression techniques and the other approaches we've considered. Potential advantages of methods based on regression analysis include the very well documented properties of the technique in a wide variety of settings, the relatively low computational burden required, and ease of developing multivariate extensions to useful univariate techniques.

Multivariate Outlier Detection Using Mahalanobis Distance

Observations which are atypical in a single component may be identified by their "sticking out" at one end of the ordered data. Multivariate observations which are atypical are those whose pattern of relationships with other variables are different from the majority of the data. The detection of these atypical observations have been hampered using classical techniques due to the fact that they exert large influences on the estimates of the mean and covariances which are used in calculating the decision criteria. Thus, robust estimation techniques (see [2]), using maximum likelihood type estimators, referred to as M-estimation, are used to estimate the mean and covariances. This is done by weighting each observation, with extreme or atypical observations getting small weights. The weight is calculated based on some measure of central tendency, in this particular case we used the Mahalanobis distance, which is a generalization of the Euclidean distance that allows the distance between points whose components may have different metrics (e.g. temperature, tensile strength, thickness, and so on)[1]. The weighted observations are used to estimate the mean and covariance parameters, which are then used to recalculate the distances used to reestimate the weights and this process iterates until the mean and covariance estimates converge. The final weights can be used to identify atypical observations. This procedure is computationally intensive and not feasible for large data sets or data that must be dealt with in real-time.

State characterization via cluster analysis

The data features identified by a sequence of "critical" time frames (that are identified by one of the above methods) may still represent a very large amount of data. Typically, however, these features will naturally cluster into a relatively small number of similarity classes (or states). By selecting or computing a few key variables that characterize the data in each time frame, multivariate cluster analysis can be used to group the time frames for a more organized examination or visualization. The regression-based data summarization technique is particularly suitable for cluster analysis post-processing because the regression parameters and goodness of fit measures already provide a few key variables that characterize the data in each time frame. We have obtained the best clustering results with some of the more recently developed model-based clustering methods of Banfield and

[1]The Mahalanobis distance is the distance between the observation and the mean of the data weighted by the inverse of the covariance matrix. This distance is then used to determine a weight corresponding to the observation, with large distances receiving small weights

Raftery [1]. Although these methods require a lot of computation, this is feasible in post-processing analysis.

Stochastic approximation

This approach was developed in the 1940's to solve optimization problems using noisy samples of an unknown function whose minimum needed to be determined. The method has been updated considerably over the years and has received recent attention because of its relevance to understanding nonlinear neural network in machine learning and optimal control. It has an even deeper relationship to Bayesian inference and information-theoretic approaches to representing nonlinear functions. It is this latter connection which is exploited to detect changes in very large data sets.

The approach is to treat a large data set derived from solving a complex set of differential equations (e.g. the global climate community model) as a noisy data set derived from an unknown but simpler nonlinear function. This unknown function is found by an optimization process over a class of potential nonlinear functions in the presence of noise. The resulting optimal function represents a lousy compression of the original data set into a known function plus noise. Given a suitably large class of functions over which to optimize, the result is also equivalent to a minimum size description of the data set or a probabilistic learned approximation to the data.

The resulting minimum function can then be used by humans to analyze the entire data graphically using the compressed function. This nonlinear form now represents the real data in a stochastic manner. That is, it can predict an as yet unseen data point to within a prescribed error tolerance (there is a small but known chance of being incorrect). A complementary data set, comprised of points not predicted by the minimum function, can be further studied as an anomalous data set for additional insights into the data being studied.

Hurst coefficient

This method, sometimes called rescaled range analysis, was developed in the 1940's by Harold E. Hurst [10], for estimating the optimal reservoir capacity. Rescaled range analysis measures the bias or trend in a time series. Hurst observed experimentally that for many natural phenomena the rescaled range analysis followed certain relationships. A Hurst coefficient of 0.5 is indicative of randomness, while a Hurst coefficient greater than 0.5 is indicative of a memory effect that is biased toward reinforcement of the trend. A Hurst coefficient less than 0.5 is indicative of a negative bias or a tendency to reverse the trend. We note that the reciprocal of the Hurst coefficient is an estimate of the so called fractal dimension of the series. One can calculate H through time using equal sized blocks of data which overlap. When H changes this is indicative of a change in the series.

3.2. SOFTWARE DEVELOPED

Some of the above methods were implemented in Splus [12] while others with AVS5 and POSTGRESS. Splus was used for implementing the regression-based method as well as for producing many of the graphics in our reports. It contains many standard statistical building blocks in a programming language interface as well as a good graphical capability. AVS5 is the Application Visualization System by Advanced Visual Systems, Inc. which provides an interactive 3D visualization system and a sophisticated graphical user interface (see `http://www.avs.com/`). POSTGRES is a public domain, portable Data Base Management System (see `http://s2k-ftp.CS.Berkeley.EDU:8000/postgres/`). It allows back-end functions to better utilize the CPUs of the host machine, and provides object oriented extensions that allow useful Large Objects to be stored with the data.

AVS5 and POSTGRES was our testbed for designing a system for the analysis of large datasets. The underlying concept is to pre-process the large dataset with a feature extraction tool, store extracted features in a database and then examine these with a post-processor. The prototype system consists of the following parts:

- a feature extractor toolkit (the pre-processor),
- a query language database (POSTGRES), and
- a visualization and analysis module (the post-processor).

We use AVS5 to create the Graphical User Interface (GUI) and as an interface tool to POSTGRES. Statistical analysis and browsing tools for the data sets are provided via the AVS5 and POSTGRES systems. Graphical interfaces are provided for both the feature extraction stage and the post processing stage of the entire system.

The feature extractor toolkit contains several methods described above and capable of analyzing the data for "critical" time-frames to be more carefully examined by the scientist. These tools operate on the data and identify points where changes are detected. These points along with their context are sent to

POSTGRES. Due to the size of the data sets, they are typically stored on large mainframes or supercomputers. It is crucial to minimize the amount of data down-loaded and processed on the client workstation. Hence, another aspect of this project was utilizing POSTGRES' back-end function capability to do as much processing on the back-end server machine as possible.

3.3. APPLICATIONS ENCOUNTERED

From the beginning of this project, we have been convinced that it would be absolutely necessary to establish associations with real problems involving very large data sets so as to insure that the methods and tools developed under this proposal are useful in applied settings. We have succeeded to some degree in this, and note below some particular data sets we have used as "testbeds" in this work. Having said this, we need also to report that one persistent difficulty we encountered throughout this project was in identifying ORNL scientists actually doing research based on the kinds of very large data sets discussed in our proposal. While such data sets clearly exist, some at ORNL (e.g. ARM, described below), our experience is that often the associated scientific goals have not gotten beyond the point of collecting, storing, and occasionally rudimentary quality screening of data. Perhaps because these problems can be difficult enough for truly large problems, questions of actual analysis for the purpose of scientific inference may be viewed as premature by investigators in some areas. We do not see this as evidence that our original aims and goals were premature, but rather as indicating the wide variety of challenges associated with basing scientific inquiry on massive data sets. Analysis is only one area in need of attention; others include management of complex automated sensing networks, data management in large and heterogeneous computing environments, data communication, and data quality assurance, all of which are also fundamentally affected by the scale of the data structures to be handled.

Data sets used as test cases, for which results are reported in the publications mentioned in Section IV, are as follows:

Sunspot data

The National Center for Atmospheric Research maintains a continuing record of the daily sunspot index, the current version of which is a single time series of 52,320 values. The recorded values certainly do not reflect high-technology measurements by today's standards, but the series is unique in that it has been collected continuously since being introduced by Johann Rudolph Wolf in 1848 with no change in the definition of a measurement. Hence it has become a standard demonstration data set for time series methods intended for large-scale applications. We have applied the Bayesian changepoint detection procedure to these data as a demonstration exercise, reported in [5].

EEG data

The example time series is the recording from one channel of a sixteen channel electroencephalogram (EEG) record for an epileptic patient [9]. The complete EEG record is 23 minutes, recorded at the sampling rate 512 samples/second, and it includes a seizure. EEG records typically include a great deal of "artifact," representing head movement, eye movement, muscle tension, grinding teeth, etc., in addition to unmasked neurophysiological activity. If we associate normal neurophysiological activity with the background process, then artifact is a perturbation relative to that background process. In addition to the artifact types already mentioned, because this is an EEG for a patient with epilepsy, there are also epileptiform transients, which are short duration (approximately .5 seconds) electrical discharges, randomly distributed in the record along with other types of unusual neurophysiological activity.

ARM data

The D.O.E. Atmospheric Radiation Measurement (ARM) project manages a large database at ORNL [7]. The time series supplied to us was a record of atmospheric water vapor values recorded near Oak Ridge, Tennessee, over a period of 257 days beginning in March, 1994. The data in this set reflect both periodic and irregular events, and include the effects of "real world" difficulties such as instrument failure in some segments. The observations are taken at 20 second intervals, but the data contain numerous gaps, some of over a day in duration. We have used a segment of 122,786 measurements, covering about 28 days, in demonstration exercises reported in [3].

CHAMMP data

The D.O.E. Computer Hardware, Advanced Mathematics, and Model Physics (CHAMMP) program has sponsored development of an atmospheric global climate model, PCCM2.0 [6]. The model simulates atmospheric physics over time, based on inputs including sea surface temperatures. We were given access to an output data set produced by this model, including four variables (global root mean square values of temperature, potential vorticity, and divergence, and an index related to the maximum time step allowed in the calculation) at intervals of 20 minutes for a simulation of about 1.4 years. We used this as a test problem for the multivariate implementation of the Bayesian changepoint detection method, and found a number of points in the time series which appear to represent unusual behavior. Interactions with an ORNL mathematician involved in the project have led to the hypothesis that at least some of these may correspond to "restarts" of the code which may not have been executed perfectly in the version of the code examined.

Engine Data

This is a data series of pressure measurements from a automobile engine, where the experiment was one in a series designed to test the dynamic effects of running at constant speed using lean fuel to air mixtures. The engine is a standard two stroke per engine cycle design and there is one observation per crank angle, so that there are 720 observations per engine cycle. The data series has approximately 2300 cycles, or 1,656,000 observations. Due to the lean fuel to air mixture, there are anomalous cycles in the record that are related to incomplete combustion. The objective of the analysis is to identify those anomalous cycles, which comprise approximately one percent of all cycles. This data was obtained from Ford Motor Company under the auspices of a Crada between ORNL and Ford (Contact: C.S. Daw, Engr. Tech. Division).

4. PAPERS, PUBLICATIONS, AND PROPOSALS

At this date, three ORNL technical reports describing our work have been written or are in preparation:

- Report [5] defines the univariate version of our Bayesian changepoint detection method, along with demonstration applications based on the sunspot data set and a computer-generated series based on a simple growth model from population biology.
- Report [4] defines our nonlinear dynamics changepoint detection method.

- Report [3] describes our work in applying regression techniques to the problem of identifying "unusual" segments, compares it to other methods developed in this study, and presents a computational exercise in which several of these methods are used to find changes and transient behavior in the EEG and ARM data sets.

This research was featured in an article "Filtering on Interesting Data Features," *Statistical Computing and Graphics*, Vol.6, No.1, pp.11-13 (April 1995). This is a publication of the American Statistical Association.

A proposal entitled "Enhancement of Statistical Methods and Algorithms for Remote Sensing" was prepared and submitted to NN-20/On-Site Systems.

References

[1] Jeffrey D. Banfield and Adrian E. Raftery. Model-based gaussian and non-gaussian clustering. *Biometrics*, 49:803–821, 1993.

[2] N. A. Campbell. Robust procedures in multivariate analysis I: Robust covariance estimation. *Applied Statistics*, 29:231–237, 1980.

[3] D. J. Downing, V. Fedorov, W. F. Lawkins, M. D. Morris, and G. Ostrouchov. Large datasets: Segmentation, feature extraction, and compression. Technical Report 0000, ORNL, 1995.

[4] D. J. Downing, V. V. Fedorov, W. F. Lawkins, M. D. Morris, and G. Ostrouchov. Analysing perturbations and nonstationarity in time series using techniques based on the theory of chaotic nonlinear dynamical systems. Technical Report ORNL/TM-13115, Oak Ridge National Laboratory, Oak Ridge, TN 37831, 1995.

[5] D. J. Downing, W. F. Lawkins, M. D. Morris, and G. Ostrouchov. A method for detecting changes in long time series. Technical Report ORNL/TM-12879, Oak Ridge National Laboratory, Oak Ridge, TN 37831, 1995.

[6] John Drake, B. D. Semeraro, P. Worley, I. Foster, J. Michalakes, B. Toonen, J. J. Hack, and D. L. Williamson. PCCM2: A GCM adapted for scalable parallel computers. In *Fifth Symposium on Global Change Studies*, pages 91–98. American Meteorological Society, 1994.

[7] D. O. E. Atmospheric radiation measurement program plan. Technical Report DOE/ER-0441, U. S. Department of Energy, Office of Health and Environmental Research, Atmospheric and Climate Research Division, National Technical Information Service, 5285 Port Royal Road, Springfield, Virginia 22161, 1990.

[8] K. Gordon and A. F. M. Smith. Modeling and monitoring biomedical time series. *J. Amer. Statist. Assoc.*, 85:328–337, 1990.

[9] L. M. Hively, N. E. Clapp, C. S. Daw, W. F. Lawkins, and M. L. Eisenstadt. Nonlinear analysis of eeg for epileptic seizures. Technical Report ORNL/TM-12961, Oak Ridge National Laboratory, Oak Ridge, TN 37831, 1995.

[10] H. E. Hurst. Long-term storage capacity of reservoirs. *Trans. Amer. Soc. Civil Engrs*, 116:770, 1951.

[11] R. E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82:34–45, 1960.

[12] Statistical Sciences. *S-PLUS Guide to statistical and mathematical analysis, Version 3.2*. StatSci, a division of MathSoft Inc., Seattle, 1993.