"Thar She Blows!": A New Approach to the Analysis of Yellowstone Geyser Eruptions

Christopher Landauer
Aerospace Integration Science Center
, National Systems Group, The Aerospace Corporation
P.O.Box 92957 Los Angeles, California 90009-2957
Phone: (310) 336-1361
E-mail: cal@aero.org

3 July 1997

Contents

Abstract

This paper is about analyzing a dataset that has been collected off and on for over a hundred years. This dataset records the geyser eruptions in Yellowstone National Park, and while it is not massive in the sense of terabytes or more, it makes up for that in anomalies: it is very heterogeneous, partly missing and somewhat erroneous (almost all data collection is by volunteers who trek out to the geyser basins and watch). It affords a very good (and very stringent) test of any data analysis method, because it has many different kinds of anomalies all concentrated in a small space. In many very large data sets, the anomalies are often simply ignorable, but they are decidedly not ignorable in this case.

The purpose of the data collection and analysis project is to support models of the causal interconnections among the geysers and predictive models of the eruptions (including times and other characteristics). Currently both are based on informal and anecdotal descriptions of their temporal relationships. The models are further complicated by the fact that occasional earthquakes (Yellowstone is in a _very_ active seismic area) may change the subterranean connectivity, and definitely change some of the eruption characteristics.

We will describe some of the data, some preliminary models that have been developed before by others, and explain why we think that Hidden Markov Models will be a good place to start with the interesting problems in this area. This is very much a "work in progress".

1 Introduction

The question at hand, then, is to determine some predictive models of geyser eruptions in Yellowstone National Park. There have been recordings of observations for many years, so there is a rich collection of interesting data to consider. The data is logged by hand by trained observers and arbitrary visitors, then transcribed to computer files (at least, it is now). There are many different kinds of data anomalies introduced by this process. Even without any new problems, it is very difficult to model the geyser systems, because we do not know the underground configurations, the driving heat sources, or the mechanism very well. Some geysers are connected to each other underground, so their eruptions interact, and even when everything else is not considered, some geysers seem to fit a chaotic dynamic system model better than others [9].

Our goal is to find predictive models of eruption characteristics (such as interval between eruptions, duration of eruptions, height of eruptions, amount of water or steams that flows out, etc.), to help us understand the mechanism driving the eruptions.

Other difficulties have been well-described by T. Scott Bryan in [2]. They include the misconception that there are large "chambers" underground that can be treated with the usual "compartment" models of differential equations, the fact that Old Faithful in particular has had a seemingly different distribution of eruption times since the Lake Hebgen earthquake in 1959, and the difficulties expressed by several people who have been attempting to find useful descriptive statistics, such as average interval between eruptions.

The approach taken in this paper, and initially in our project, is to separate the geophysical intuitions from the data analysis intuitions, because our expertise is in data analysis. Many errors and anomalies can be found by considering the data alone. We will use this approach to build models of the eruption intervals.

It is important to note that model "fit" is not the point: almost any model with enough parameters can be made to fit very closely to almost any data. For example, Ptolemy's solar system model was much more accurate than Copernicus' for many years after both existed. In our view, models with more explanatory power are much more likely to be predictively useful than ones that merely "fit".

Our approach to validation of internal consistencies in complex systems leads to many other verification and validation methods [1] [6] as well as other kinds of exploratory data analysis. Of course, there is no guarantee that any exploratory data analysis method reflects any actual physical process; that has to be checked or explained separately. On the other hand, many unexpected properties cannot be found if one makes too many "reasonable" assumptions.

2 "This is _really_ BAD data!": Data and Data Anomalies

In this section, we describe the different kinds of data anomalies that we have observed. The reader should not take this as a set of complaints; data is _very_ hard to get in that area. Rather, these anomalies should be taken as hints that the mathematics we can use needs to be much more interesting.

There is missing data, because there are very few night observations, and bad weather prevents many geysers from being reached (not all of them are near the roads and buildings). The amount of missing data means that we cannot just treat it as a rare occurrence, and continue to use models that assume we have all the data. The data that is missing is not missing randomly, so we can't use any of the usual noisy signal processing methods either.

There are erroneous entries: we found some impossible time values, and a whole day of data with the wrong geyser in the log. We need to perform some serious normalization to prepare the data for analysis.

There are unreliable measurements: visitor reports are in the logs, from untrained visitors. This makes the measurements unreliable, but there is often no other person around to validate (or at least corroborate) the observation.

There are many kinds of imprecise measurements, such as "sometime between yesterday and now", which need to be taken into account. We cannot therefore use models that require precise times of events. This problem gets even more interesting; the kinds of imprecision are different, and we need to consider different ways to integrate uncertainties. This is a difficult research problem in general.

There are illegible logs, with some of the entries simply unreadable (all are manually entered and then transcribed). We can treat these as very imprecise measurements, if we can be sure that they record an eruption and not some other event (which they mostly do).

There are explicit non-event intervals, such as "no eruption between then and now". We therefore can't use a simple model that says all of our data is positive instances of the events we are considering, and our models must accommodate data of this kind.

So one of our technical problems is to estimate the completeness of the observations. The Hidden Markov Model approach will do that, and we expect to conduct some studies of the extent to which the models reflect the reality.

Finally, there are some other characteristics of the problem that are not so much data anomalies as difficulties with collecting appropriate or sufficient data.

There is a very wide range of time scales, from minutes to months, depending on the geyser (actually minutes to decades, but there isn't nearly enough data to model those ones).

The models are further complicated by the fact that occasional earthquakes (Yellowstone is in a _very_ active seismic area) may change the subterranean connectivity, and definitely change some of the eruption characteristics.

There is also some possibly useful extra information: ambient temperature (though we expect this to be relatively less important, because the interesting temperature effects occur far from the usual air temperature range), precipitation (similarly, there are underground water sources), and atmospheric pressure (this may be the most interesting, because pressure is relevant at some level). What we don't know is whether or not these are actually relevant at all.

These difficulties with the data require some very sophisticated mathematical techniques, which we will begin to describe in the rest of this paper.

3 "What's Happening Down There?": Geophysical Speculations

We start with a very short description of what might be relevant geophysics, primarily couched as questions that may help in making better models. It is our expectation that many of these questions have immediate answers for those who understand the geophysics, but we expect that some otheres will be harder to resolve at our present state of knowledge.

One of the main questions is how the underground configuration is partitioned, at least in terms of its effects, and to determine how one of these partitions (not "chamber" according to Scott Bryan) influences another.

Another question is whether or not all activity is caused by steam pressure from below, and how the steam pressure is driven by heat flow. Also, what is the "incoming" heat pattern?

Are the subsurface heat / pressure conditions such that large pockets of gas can exist? Pocket models consider a graph of pockets and passages, the dynamics of heating, the likelihood of an initial bubble (if there are bubbles), propagation speed and constrictions, and the likelihood and speed of "infection".

Is explosive decompression the driving mechanism? Consider a reservoir of superheated water, in which high pressure from above keeps the water in liquid form, but with heat from below raising the temperature. If a sudden drop of pressure occurs, then the water could vaporize in a large batch, greatly increasing the volume (really big bubbles?).

Does precipitation affect the system? Maybe the existing ground water makes it irrelevant.

Does ambient temperature affect the system? Maybe the underground heat source makes it irrelevant.

Does air pressure affect the system? Maybe by adjusting the thresholds.

What is the mechanism that reduces the pressure? Maybe heat increase makes water level rise, and then spillover into disconnected reservoirs removes some of the constraining water.

4 Data Analysis Strategies

In this section, we describe our general strategies for data analysis.

Our first main point is that we need a repertoire of methods: no one method can suffice for complex data [10]. We use a style of analysis that we call Model-Based Data Analysis. We are actually more interested in the model analysis, not so much data analysis, but we take the two to be essentially the same, because in our view, there is no data without models and situations. Every algorithm and every dataset presume some properties. Those properties define models, implicitly, perhaps loosely, presumably consistently. Making those models _explicit_ is very important to our approach, because we can study those models effectively in their abstract setting. There are many important formal spaces we can use for these modeling spaces.

The three themes listed for the workshop are relevant: classification contains clustering as one technique, and there are many different spaces in which clustering can be used to look for commonalities and groupings, consensus is a powerful tool for integrating multiple disparate models (though we haven't seen it applied in this way yet), and pattern detection, identification, and recognition lie at the heart of the geyser problem. The lack of generally accepted predictive models means that exploratory pattern analysis is important, to determine unexpected patterns that may be interesting.

We mentioned above that uncertainty model integration is needed to collate the different kinds of uncertainties in the records, and more generally, complex heterogeneous collections of models require a much more interesting kind of integration than simpler models [7] [8].

In general, we believe that meta-information (i.e., information "about" the data) organizes the "terrain", and that data occupies the terrain and sometimes shapes it. This notion allows us to consider many different kinds of formal spaces as terrain, and to map the data into those spaces in many different ways, while still allowing data dependence of the formal space when appropriate. For example, if there are many data items, not very structured, then it is more sensible to map them to points in a space, but if there are fewer much larger items, then the space needs to be more structured.

Finally, we view datasets as being defined with models, and we want to find methods for mapping these superficial or surface structure models (containing, e.g., data types, typical values, exceptional values) to partial models that are deeper and explain more (containing more of the semantics, i.e., more of the meaning of the phenomena).

5 Data Analysis Techniques

In this section, we begin to describe the specific techniques that we are using or will use. First, we describe the initial data analysis process, that is, the processing steps from log data to event data.

The first part of data analysis is "normalization": cleaning up the data. There is a painstaking process of informal local model building, which leads to a syntactic error analysis that can find many internal inconsistencies. We found some errors in the data this way. This careful attention to local syntactic models of data also allows us to determine local models of time intervals and duration precision (e.g., times reported as 4:03, 2 min, more than 3 min, 2-3min, short, long, etc.), and leads to several different uncertainty models for time reporting.

Finally, we "fix" the disparate data formats, by converting them into a form that we find convenient for analysis. We really only have two data analysis problems, when they are considered in the abstract: sequence completion and sequence comparison. Prediction is a kind of completion into the future, and causality is a kind of comparison.

In our context, descriptive statistics are not much help in prediction, because these are nonlinear dynamical systems, for which behavior is not very amenable to description by averages (or any other statistical summaries). On the other hand, we can learn something useful about the local data models, such as the time duration distribution (for example, the use of "short" and "long" as time durations is relatively well-defined by a strongly bimodal time distribution for that particular geyser).

The technique we concentrate on in this paper comes from the following considerations. We assume that we have a state-determined system, so that each state at eruption onset determines the duration of eruption, the interval until next eruption onset (these are the conditions we can see), and also the next state. We assume that there are a fixed finite number of states, and use a Hidden Markov Model [3] [5], which has stochastic state transitions with stochastic observations. These models have become very popular for modeling certain kinds of unobservable processes because they have a very effective parameter estimation algorithm.

6 Hidden Markov Models

Hidden Markov Models are a generalization of simple descriptive statistics: we assume that there is an unknown state sequence underlying our observations. There are stochastic state transitions with stochastic observations. We assume a fixed finite number of states, but we also assume that we don't know the number of states in advance. If there is only one state, then the usual probability model. The stochastic output function is a function from the state to a probability density function of the interval between eruptions, which we can characterize by mean and variance, or some other small set of parameters. Our examples will all use the simplest case of characterizing the output function using mean and variance.

The process of determining model parameters uses a maximum likelihood estimate of the Markov chain parameters called the "Forward-backward" algorithm, which is a special case of the very general EM (Estimation-Maximization) algorithm [4]: for a parametric stochastic process, first estimate missing data from parameters (E-step), then choose parameters that maximize likelihood of generating the data (M-step), and iterate until convergence.

6.1 Abstract Data Analysis Problems

We are considering only two problems: sequence completion and sequence comparison. Sequence completion is the problem of filling in a sequence if intervals while retaining the important characteristics (and presuming that the characteristics are stable). We use this approach to repair datasets according to our models. Sequence comparison is the problem of finding a temporal relationship among sequences (and presuming that causality may be inferred).

An *observation interval* is an interval of time in which the events that occur are actually observed or at least noted, AND the accuracy of the observation is correctly reported. That means that it is a complete record of the eruptions.

We have event time data with large and small gaps. We measure and want to predict intervals between events. We don't always know where data is missing, but we do know that it is not missing randomly.

We want to optimize the number and times of unobserved eruptions, but we need to formulate appropriate criteria for the optimizations. One problem is to determine when we should divide up an interval. For example, a simple constraint is that if m is known to be a lower bound for the interval, then no interval less than $2*m$ is subdivided.

We will start with a model for the case in which all intervals are known, then move to a process that estimates the missing observations. It will find the most likely locations for new points, conditioned on the observed data.

6.2 E-M Algorithm

For our Hidden Markov Models, the EM algorithm is fairly simple [4] (we describe the process in detail in the rest of this section). We start by assuming a given number of states (we will repeat the rest of the analysis for each of several different numbers of states, and choose the best performing number). The first case assumes that there is no missing data, that we have all the intervals in a sequence of steps.

We initialize the transitions to uniformly random (there is an eigenvector calculation that we think can speed up the convergence, but it is not necessary for this problem).

We initialize the output parameters for observations, using the interval statistics for complete observation intervals (here, just sample mean and variance). We also need to decide in general on the shape of the output distribution, with choices ranging from nearly uniform to weakly or strongly unimodal, and then choose a one or two parameter family of density functions.

We compute the state sequence that maximizes the probability of producing the observed sequence (this is a dynamic programming problem), use the estimated state sequence to recompute the transition probabilities, and use the observed data to recompute the output distribution parameters. This process cycles until some convergence criteria are met (this process is known to converge).

The EM algorithm begins by assuming a given number of states (the rest of the analysis is repeated for each of several different numbers of states, and the best performing number is chosen).

We can produce new data for missing observations during each cycle also, even though we don't know where or when data is missing, and wait for convergence of that as well. We get a measure of fit for each number of states, and can choose our best estimate of the size of the state space accordingly.

7 Current Model-Building Experiments

We have seen some of the log data for the geysers, and we have chosen several geysers and sets of geysers for experimental models. There are four individual geysers of interest: Old Faithful has the most regular and most complete observations by far, but there is also substantial data for Plume, Little Cub, and Daisy. We will build Hidden Markov Models for these and see if we can find geophysical realities in the number of states and the transitions.

There are two families of geysers of interest: Beehive and Beehive Indicator, because the Beehive Indicator is known to be a predictor of Beehive, and the phenomenon known as the Geyser Hill Wave, which is a collection of geysers whose eruptions are correlated in sequence, and which includes Plume, Little Cub, Beehive and Beehive's Indicator, and several others. We rely here on the close observations over many years of experts: we have no mathematical model of these correlations, but observation of coincidence and delayed coincidence is one of the things humans notice very well. We are not trying to show whether or not these phenomena exist; we know they do. We are trying to find mathematical models that predict the phenomena adequately.

8 Conclusions

This paper seems to be more about model analysis, and not so much data analysis, BUT in our view, there is no data without models of what the data means and situations in which the data was collected. Every algorithm and every dataset presume some properties. Those properties define models, implicitly, perhaps loosely, but presumably consistently. Making those models _explicit_ is very important, because we can study them only if they are explicit.

We described many important formal spaces that we can use for modeling the geyser data, and described one in depth. The most promising of the early models are hidden markov models, and we are still formulating them for this application.

References

[1] Kirstie L. Bellman, "The Modelling Issues Inherent in Testing and Evaluating Knowledge-based Systems", pp. 199-215 in Chris Culbert (ed.), *Special Issue: Verification and Validation of Knowledge Based Systems, Expert Systems With Applications Journal*, Volume 1, No. 3 (1990)

[2] T. Scott Bryan, Discussion of subtleties in Old Faithful eruption patterns Geyser mailing list (11 January 1997)

[3] Eugene Charniak, *Statistical Language Learning*, Bradford Book, MIT Press (1993)

[4] A. P. Dempster, N. M. Laird, D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", *J. Roy. Stat. Soc. London, Series B*, Vol. 39, pp. 1-38 (1977)

[5] Robert J. Elliott, Lakhdar Aggoun, John B. Moore, *Hidden Markov Models: Estimation and Control*, Springer (1995)

[6] Christopher Landauer, "Correctness Principles for Rule-Based Expert Systems", pp. 291-316 in Chris Culbert (ed.), *Special Issue: Verification and Validation of Knowledge Based Systems, Expert Systems With Applications Journal*, Volume 1, No. 3 (1990)

[7] Christopher Landauer, Kirstie L. Bellman, "Integration Systems and Interaction Spaces", pp. 161-178 in *Proceedings of FroCoS'96: The First International Workshop on Frontiers of Combining Systems*, 26-29 March 1995, Munich, Germany (March 1996)

[8] Christopher Landauer, Kirstie L. Bellman, "Model-Based Simulation Design with Wrappings", pp. 169-174 in *Proceedings of OOS'97: Object Oriented Simulation Conference, WMC'97: 1997 SCS Western Multi-Conference*, 12-15 January, Phoenix, SCS International (1997)

[9] Julie Knowles Raye, Kevin M. Short, "Nonlinear Analysis of Yellowstone's Plume Geyser", *Proceedings 1997 Joint Mathematics Meetings*, 8-11 January 1997, San Diego, CA; paper on chaotic dynamical system model of Plume, discussed in geyser mailing list (20 November 1996)

[10] Donald O. Walter, Kirstie L. Bellman, "Some Issues in Model Integration", pp. 249-254 in *Proceedings of the SCS Eastern MultiConference*, 23-26 April 1990, Nashville, Tennessee, Simulation Series, Volume 22, No. 3, SCS (1990)