Constructing Numerical Reference Modes from Sparse Time Series

Peter S. Hovmand, Ph.D.
Washington University
George Warren Brown School of Social Work
Campus Box 1196
One Brookings Drive
St. Louis, MO 63130

May 16, 2003

Abstract

Constructing the reference modes is a critical step in system dynamics modeling. Estimating rates from sparse time series presents a unique problem. Specifically, as the counts per unit time approach zero, the time series start to look increasingly like discrete stochastic variables, i.e., not continuous, even though one might in some situations reasonably hypothesize an underlying continuous variable. Smoothing techniques are commonly used to identify patterns in noisy data, but introduce and remove features that could mislead the modeler. There has been considerable research on optimal smoothing techniques for noisy time series. This paper presents initial work toward different approach that side steps the question of optimal smoothing and takes advantage of the emphasis in system dynamics of good models being expected to perform well over a range of conditions.

Having rich numerical data to support reference modes can be a critical element of successful system dynamics model testing and confidence building (Homer, 1997). While numerical data typically represent only a small portion of stakeholders' knowledge of a problem (Forrester, 1980), numerical data that are available should be included. Social services agencies typically maintain one or more case or client level databases, which often include information on dates of key events like opening and closing of cases, and referrals to other agencies, counseling sessions. These event dates can be used to estimate rates such as the number of new cases per day and levels such as the client caseload. When phenomena are "large" relative to the time constants of interest, the number of events per unit time can be reasonably approximated as a continuous variable (Sterman, 2000). Such variables may have noise and require smoothing in order to see the general trends (Forrester, 1961/1999; Randers, 1980). Smoothing, however, is a tricky affair as there is always the risk introducing or removing time series features. Nonetheless, smoothing can help modelers see general patterns, and there is a large body of research into various techniques and the selection of the best approach for various problems. But such approaches, however valid, create special difficulties when considering sparse time series. That is, time series where the frequencies are so

low that the estimated number of events per unit time becomes more discontinuous and discrete.[1] This is likely to be the case with small social service agencies like domestic violence shelters, group homes, small public health clinics, and other situations involving small populations. One might be tempted to abandon the construction of numerical reference modes if faced with sparse time series and rely on hypothetical or qualitative reference modes as an alternative. Doing that, however, would be to ignore important information that already exists. So the question becomes, how might one approach the representation of sparse times series for the construction of numerical reference modes involving rates?

1 Numerical time series

Dates variables can be aggregated over time intervals to produce numerical time series of rates (e.g., number of events per unit time). When preliminary inspection of data sets reveal that the time constants are short relative to the level of aggregation, one is forced to consider shorter time period or aggregation bins. There is a limit, however, to how short the aggregation bins can become before one runs into the problem of sparse time series. Moreover, as the rates (frequency per unit time) gets smaller, the aggregation becomes increasingly sensitive to both the origin of the aggregation bins and the width of the aggregation bins relative to the time constant. (Härdle, 1990).

2 Illustration of sparse time series

The impact of smaller rates on the sensitivity of aggregations to bin size and their origin can be seen through a series of simulations where the expected number of events per unit time decreases. Figure 1 shows three simulated time series of the rate or expected number of cases per day with respective overall means of 100, 10, and 1 (note that, going from left to right, the vertical scaling decreases by a factor of ten for each graph).

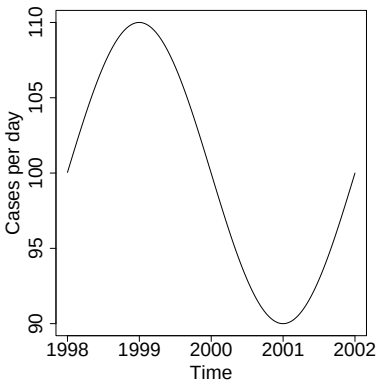
The number of events per unit time is always an integer value equal to or greater than zero. This is typically modeled as a Poisson distribution. One can generate a simulated time series of observations by randomly sampling one case per unit time from a Poisson distribution with the expected rates corresponding to the time series shown in Figure 1. Specifically, $O_m(t) = P(\lambda_m(t))$ where, $O_m(t)$ is the observed value at time t, $P(\lambda_m(t))$ is a random value sampled from the Poisson distribution with a rate of $\lambda_m(t)$. The rate $\lambda_m(t)$ is a function of time and the mean expected value m such that $\lambda_m(t) = m + 0.1 \cdot m \cdot \sin(t \cdot 2\pi/1461)$, where m is the overall mean (i.e., 100, 10, or 1)

The resulting noisy time series are shown in Figure 2. As m gets smaller, the effects of the distribution being bounded at zero become more pronounced. Figure 2a would typically be seen as a noisy but continuous time series. But as m gets smaller, the result is something that looks much more discrete with quite a few days with zero cases. Figure 2c would be a good example of a sparse time series.

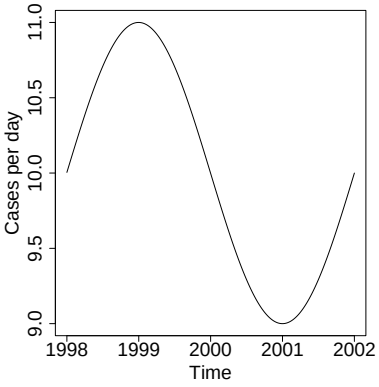
When the number of events per unit time is relatively high, aggregating the noisy time series over quarters results in a reasonable approximation of the original expected values, but not so

[1] This paper will specifically discuss the problem with estimating rates as continuous time series from sparse time series, although the general approach could also be considered for the problem of estimating stocks from noisy time series data.

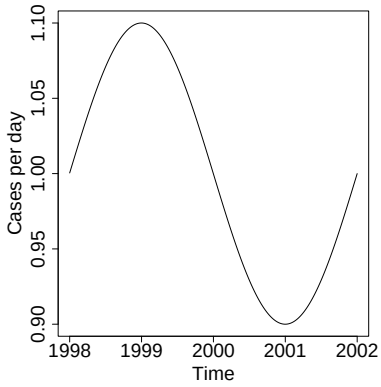
Figure 1: Original Expected Rate of Cases per Day



(a) Mean = 100

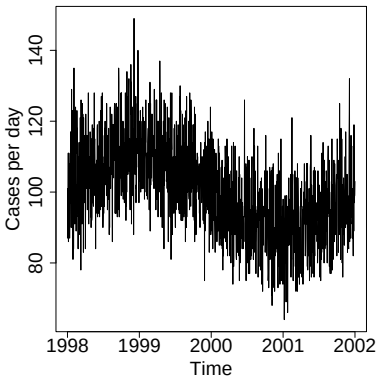


(b) Mean = 10

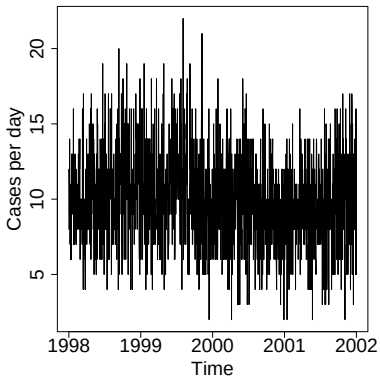


(c) Mean = 1

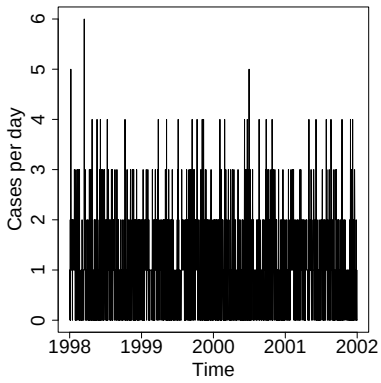
Figure 2: Noisy Time Series



(a) Mean = 100



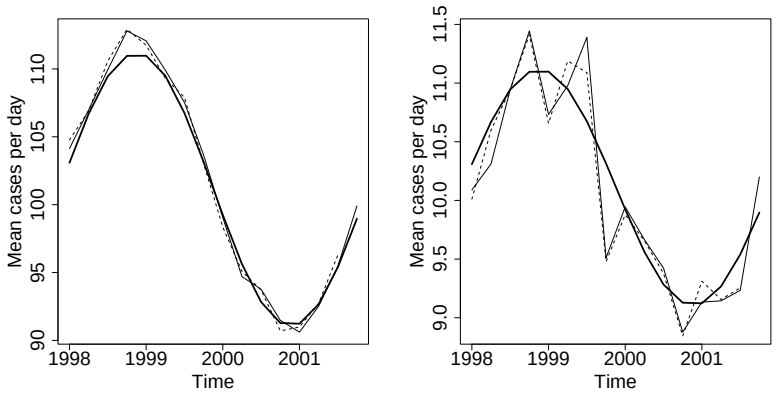
(b) Mean = 10



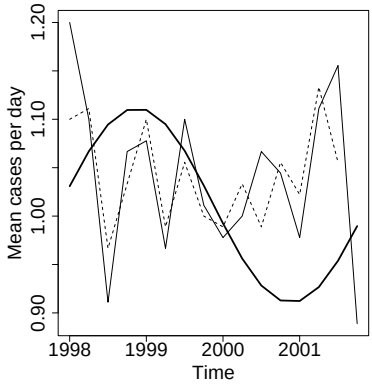
(c) Mean = 1

for sparse time series. Figure 3 shows the result of aggregating the counts over quarters. The thick solid line shows the original expected values, the thin solid line shows the aggregated counts over quarters (divided by the number of days per quarter to standardize the values), and the thin dashed line shows the effect of offsetting the origin of the bins by as little as nine days.[2] Figure 3a illustrates how aggregating the number of cases per day on quarters results in a reasonable approximation of the original expected values and is insensitive to minor variations in the origin of the aggregating bins. However, as *m* gets smaller, the aggregated time series does a worse job of approximating the original time series and becomes more sensitive to variations in the origin of the bins. Figure3b might still be a reasonable approximation, but one would be hard pressed to identify the original time series in Figure 3c. Figure 3c also shows how even a small change in the origin of the bins of only nine days can affect quarterly totals, especially the first, second, and third quarters of 2000.

Figure 3: Quarterly Aggregated Time Series



(a) Mean = 100 (b) Mean = 10

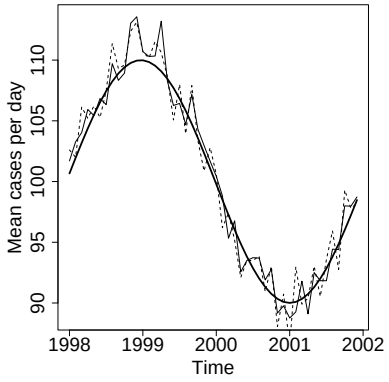


(c) Mean = 1

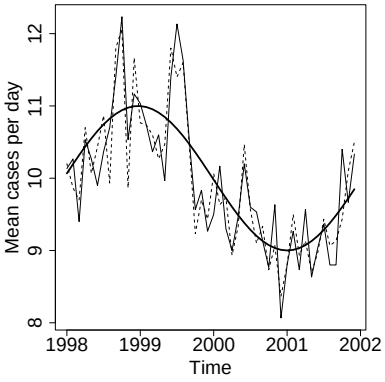
[2]The choice of offsetting the origin by nine days is arbitrary. The main point is that the offset is small relative to the width of the bins.

The effects of small time series on aggregated values get worse as the size of the aggregating bins get smaller. Figure 4 shows the results of aggregating the number of cases over months which are, again, divided by the width of the aggregating bin to standardize the values. While Figure 4a still appears to be a reasonable approximation of the original time series, Figure 4b is starting to become questionable and showing more effects from varying the origin of the bin size. Figure 4c is unrecognizable.

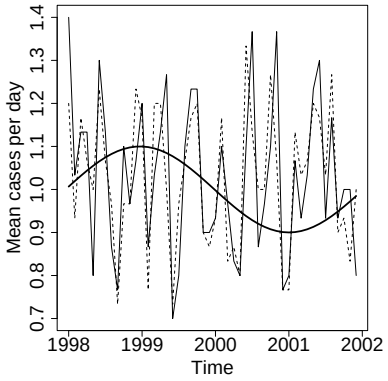
Figure 4: Monthly Aggregated Time Series



(a) Mean = 100



(b) Mean = 10



(c) Mean = 1

2.1 Conventional approaches to handling noisy data

Approaches to handling noisy time series fall into three general classifications: filtering, smoothing, and prediction (Anderson & Moore, 1979). Filtering works in real time, trying to recover the original value from the noisy signal $O(t)$ at time t. The tuner on a radio or television is a classic example of a filter. Smoothing uses a range of values, $(t-\Delta t, t+\Delta t)$, to estimate $O(t)$. There are many approaches to smoothing, but perhaps the most familiar is a moving average. Prediction uses

a set of past values, $(t-\Delta t, t)$, to estimate $O(t)$ for a set of future values, $(t, t+\Delta t)$: for example, by fitting an equation to past values in order to predict future values.

When generating reference modes from real data, one is generally concerned with identifying trends and working with time series spanning the entire time horizon. That is, for each point in time t, one can use information coming both before and after t to estimate some value at t. Thus, one is generally looking to select a method for smoothing noisy time series. If one knows a priori the underlying probability distribution of a given time series, then one can apply regression techniques to estimate the parameters of the underlying distribution. Some variables in this study might fit one of the many available parametric probability distributions. But this cannot, in general, be assumed because of the nature of feedback.

One possibility is the use of non-parametric techniques to estimate the density distribution (Härdle, 1990; Scott, 1992). Such techniques do not assume a specific underlying distribution, but rather, are based on the empirical distribution of a given time series. However the application of smoothing techniques introduces some problems. Smoothing always distorts oscillations in time series in one or more ways. Frequencies may be entirely eliminated, peaks flattened out, and delays introduced. Smoothing also introduces the problem of truncating the first and last parts of a time series since most procedures use an interval of values to estimate a given point. Thus, one is often not sure whether or not some important features have been removed by smoothing the time series, or whether the features that one is trying to model are in fact, not just artifacts of the smoothing algorithm.

2.2 Solution to the sparse time series problem

In system dynamics modeling, one is generally concerned with identifying and studying the underlying structure generating a particular pattern of behavior over time. One uses numerical time series, not to identify the system of equations as one might in traditional time series modeling, but to test the model's behavior against real data. The purpose of such a test is generally not to see whether or not the model yielded the precise values of the observed data, but whether the overall behavior pattern is realistic. That is, one can make a distinction in system dynamics between the *real numeric values* and *realistic numeric values* of a variable. The real numeric value of a variable is the actual value that the variable takes on at a given point in time. If the variable is the number of warrant requests on a given day, then there is a real theoretical numeric value for the expected number of warrant requests on that day. The problem of estimating the density distribution of the number of warrant requests per day is concerned with estimating that numeric value, and it is this distribution that is sensitive to the particular technique used to smooth the time series. A realistic numeric value, however, is something more general. For any point in time, there can be many values that would be realistic, only one of which is the real numeric value of that variable at that time.

A good robust model should be able to describe the dynamics over a range of situations. Given a variety of inputs, the model should be able to reproduce corresponding outputs that are realistic. In system dynamics, we frequently test our models using a variety of inputs to explore the model's behavior over extreme conditions, oscillations, random perturbations, and so forth (Forrester, 1961/1999, 1971; Forrester & Senge, 1980; Richardson & Pugh, 1986; Sterman, 2000). Careful inspection of the outputs and feedback loop behavior often reveals structural flaws or important insights into the model's behavior. These inputs are typically idealized in some form as step

functions, pulses, sine waves, or random samples from a parametric distribution. But there is no reason why a robust model should not also be able to handle a much wider range of inputs, including the results of not just a specific smoothing algorithm, but the results from an entire collection of smoothing algorithms.

More specifically, a robust system dynamics model should be able to produce reasonable approximations of smoothed numeric time series outputs given corresponding smoothed numeric time series for an input, provided that both the input and output numeric time series have been smoothed using the same algorithm. That is, a robust system dynamics model should be able to reproduce the qualitative behavior of two different time series, each smoothed using a different algorithm or with different smoothing parameters, provided that the model's input (if there is any) was also smoothed using the same algorithm and smoothing parameters.

3 Procedure

The basic procedure for such an approach is first to generate a reasonably diverse family of time series for each rate variable of interest. This can be done by varying parameters of a smoothing algorithm (e.g., the order of an exponential smooth and/or the delay). But, the family of time series could also include the results of different types of smoothing algorithms. The main point is that one is generating a family of numeric time series to drive and test the model such that the model's response should be realistic. This can be stated more formally.

Let S_α be a function that smoothes a given numeric time series such that $P_{\alpha,i}(t) = S_\alpha(O_i(t))$, where $O_i(t)$ is the i-th observed time series and $P_{\alpha,i}(t)$ is the resulting smoothed time series from applying the function $S_\alpha(\cdot)$ to $O_i(t)$. Then one selects and applies a family of smoothing algorithms, F, with various parameters such that $\alpha \in F$. If $O_1(t)$ represents an input used to drive the system and $O_2(t)$ an observed dependent or endogenous variable, then the claim is that a good model should be able to reproduce $P_{\alpha,2}(t)$ when driven by $P_{\alpha,1}(t)$ over the family of smoothing algorithms where F. If the models behavior to $P_{\alpha,1}(t)$ to $M_{\alpha,2}(t)$, then $M_{\alpha,2}(t)$ should be a realistic representation of $P_{\alpha,2}(t)$ over the entire family of smoothing algorithms F.

4 Limitations

There are clearly a number to this approach. First and foremost, speaking of sets of time series for a single variable could easily make stakeholders more weary and suspicious of the modeling effort. This is a serious problem if the purpose of including numeric time series is to help ground the model in stakeholders' perception of their problem. Second, there is the problem of trying systematically evaluate the results. That is, one is now faced with the problem of having to reduce all of these comparisons in some way that meaningfully relates the comparison to variations in smoothing parameters or algorithms. This could be partially resolved by using reducing Theil (1966) inequality statistics to a single index. For example, by taking the product of (a) the root mean square error, and (b) the angular difference between the observed and desired inequality statistics. Third, there is a heavy computation requirement in generating these families of smoothed time series, simulating the model, and analyzing the results. Fourth, the approach really only works in cases where one is driving one part of the model with real data and studying the model's

response.

5 Conclusion

Considering the resulting smoothed numerical time series as a family of curves for the reference mode takes advantage of the system dynamics emphasis on sensitivity analysis and robust modeling. The procedure described in this paper is being used with some success to identify key features, calibrate and test the models, and make subsequent refinements to both the structure of the data and the model. There are clearly a number of limitations, so such an approach should not be considered without careful consideration of the underlying phenomena being studied.

References

Anderson, B. D. O., & Moore, J. B. (1979). *Optimal filtering.* Englewood Cliffs, NJ: Prentice-Hall, Inc.

Forrester, J. W. (1971). *Principles of systems.* Waltham: Pegasus Communications, Inc.

Forrester, J. W. (1980). Information sources for modeling the national economy. *Journal of the American Statistical Association, 75*(371), 555–566.

Forrester, J. W. (1999). *Industrial dynamics.* Waltham: Pegasus Communications, Inc. (Original work published 1961)

Forrester, J. W., & Senge, P. M. (1980). Tests for building model confidence in system dynamics models. In G. P. Richardson (Ed.), *Modelling for management: Volume ii: Simulation in support of systems thinking.* (pp. 413–432). Brookfield, VT: Dartmouth Publishing Company, Ltd.

Härdle, W. (1990). *Smoothing techniques: with implementation in s.* New York: Springer-Verlag.

Homer, J. B. (1997). Structure, data, and compelling conclusions: notes from the field. *System dynamics review, 3*(2), 293-309.

Randers, J. (1980). Guidelines for model conceptualization. In J. e. Randers (Ed.), *Elements of the system dynamics method* (pp. 117–139). Cambridge, MA: Productivity Press.

Richardson, G. P., & Pugh, A. L. (1986). *Introduction to system dynamics modeling with dynamo.* Cambridge, MA: MIT Press.

Scott, D. W. (1992). *Multivariate density estimation: theory, practice, and visualization.* New York: John Wiley and Sons, Inc.

Sterman, J. D. (2000). *Business dynamics: Systems thinking and modeling for a complex world.* Irwin McGraw-Hill.

Theil, H. (1966). *Applied economic forcasting.* Chicago: Rand McNally and Company.

Back to the Top