Practitioner's Commentary: The Outstanding Water Tank Papers

David Reboussin
Statistics Dept.
University of Wisconsin
1210 W. Dayton St.
Madison, WI 53706

Being a lover of data analysis, I could not resist making some exploratory plots and fitting some simple models to the problem data. As I did this and read the winning papers, my opinion of the best way to do the required estimation changed many times. This is a dataset that resists exact analysis yet succumbs to a sensible, pointed attack. The investigation raises a number of subtle points concerning the analysis of observational data.

The winning papers show a remarkable similarity in their solutions, considering the complexity of the problem. The three point-estimates for total water consumed (338,000, 326,600, and 330,000 gallons) differ by about 3%, and the two teams that provided some sense of the variability in their estimate easily cover this range. All the teams began by estimating the derivative of the level measurements and then fitted a model to estimate the flow rate. The three resulting curves are not identical; but considering the differences in the methodologies, each captures the same essential features. My own somewhat less intensive analysis agrees well with these results.

In addition, each of the winning solutions had one or more of what I would consider particularly inspired features. The difficulty in finding an adequate global model for the flow rates led the team from Alaska—Fairbanks to consider a spline model, which is essentially a series of good local approximations. (A spline is a piecewise polynomial whose pieces have matching derivatives of a certain order.) The team from Hiram made the most thorough analysis of the flow rate during pumping times, by making an explicit assumption about the pumping rate: that it was constant. This allowed some detailed computation of the flow rate during the times when no levels could be measured, and also provided an additional check of their model. Finally, in the only acknowledgement that the original measurements contain random error, the team from Ripon perturbed the original data to produce two new "worst-case" datasets, to check the sensitivity of their method to known variation in the data.

The fundamental dilemma is whether to treat the problem as a "real" problem or as an academic exercise. On the one hand, the data are actual measurements from a small town; and if we imagine ourselves writing a

report to a small-town water utility, the requirements are clear: keep to the point, do something sensible that answers the question, and avoid unnecessary technical detail. On the other hand, this is a contest; and if we know that the judges will be looking for some mathematical prowess as well as "correct" solutions, there are other considerations: demonstrate a deeper understanding of the issues, use a little finesse in developing solutions, and apply techniques that are likely to be at least as sophisticated as our peers'.

This conflict is not as artificial as it may seem. Mathematicians and statisticians working on applied problems usually collaborate with specialists in the subject matter, and these specialists may have a lot of mathematical background or none at all. There is always tension between using a more complex method that is more accurate but harder to present and explain, and using a less complex method that is easily grasped and gives sufficient (if less) precision. Mathematicians tend toward the former.

On a less philosophical level, the problem still presents a number of difficulties. Certainly, there is no single estimate of the flow rate to be made. We are given "one day" (actually about 26 hours) of data and told nothing about the day; that is, whether it is typical in some sense, or was chosen for some special reason having to do with the water use. Are we interested in this particular day, or in some general daily pattern? It is natural to expect the latter, but can we even define "daily pattern"? Does the day represent typical water use on weekdays or weekends, in the summer or the winter, during drought or flood?

These are not idle questions: indeed, each of the winning solutions made specific assumptions about how the flow rate, denoted $f(t)$, is to be interpreted. This has a direct bearing on the estimation procedure, since a general pattern should reappear on other days. It need not reproduce the data on this particular day, or even approximate it very well, if a closer fit is at the expense of smoothness. A smooth $f(t)$ ought to be more representative of general trends, and smoothness translates directly into continuity and the existence and magnitudes of derivatives. If we fit every blip and wiggle of the data (even ignoring any random errors in the measurements), our estimate of $f(t)$ will be correspondingly contorted.

The Hiram and Alaska—Fairbanks teams took the view that some underlying pattern should be estimated, while the Ripon team confined its reference to the day on which the measurements were taken. This difference might have led to very different estimates of $f(t)$; but a tendency to overfit, in an effort to find a reproducible pattern, resulted in great similarity in the final estimates despite the wide range of methods.

The most difficult question is how to account for errors in the measurements. We are told that the level in the tank is recorded with 0.5% accuracy, and this information ought to have some effect on the analysis; yet none of the winning solutions made really good use of it (though the Ripon team

did use it in an ad hoc manner). As a statistician, I take strong exception to the failure to account for random error in the analyses: in fact, it is sorely apparent that no one on these three teams is familiar with the practical use of linear regression. Let me soapbox only long enough to offer a single piece of advice: failure to account properly for stochastic variation in observational data will often lead to serious misinterpretation of the results.

For this problem, consideration of random errors does not make the analysis simple, but it does clarify some points. Suppose we denote the observed levels at time t_i as L_i, and the process that is generating the observations as $L(t)$. Although the magnitude of the errors is given as a percentage, which implies that they are multiplicative, for simplicity let us model them as additive. (This simplification could be avoided by taking the logarithms of the levels and modeling those instead.) Let us denote for each time t_i the *true* level of the water in the tank by Λ_i and the error by ϵ_i. Then

$$L(t_i) = L_i = \Lambda_i + \epsilon_i.$$

If the errors are the accumulation of a large number of inconsiderable influences, then it may not be unreasonable to treat them as independent of one another, as having the same variation at all times, and as following a normal distribution. These three assumptions, in conjunction with the high degree of smoothness in the original measurements, make a regression model for the levels an attractive line of attack.

Two problems would have been solved by this analysis. One is that the random errors in the level measurements could be removed by replacing the observed values with values predicted by the regression model. Another is that estimates of the flow rates based on derivatives from the regression model will be smooth, if the model is appropriately chosen. A disadvantage of this approach is that it can be applied only between pumping times.

Two of the teams used least-squares in part of their analysis. Least-squares is a popular and easily implemented method of fitting models to data; and in these solutions, the models of choice were polynomials. A variation of the spline model used by the Alaska—Fairbanks team could also have been fit by least-squares. Instead of regressing onto a polynomial basis, a basis of splines could be formed by chosing a set of join points (*knots*), which should be denser in regions where the function is changing more rapidly. This approach would have retained the local nature of the approximation without insisting that the data be interpolated. A further step in this direction would be to use a *smoothing spline*. Smoothing splines are solutions to penalized least-squares problems, in which there is some penalty, typically on the integrated derivative of some order, which forces the estimate to be smoother than a polynomial fit by least-squares. A smoothing spline would produce an estimate of the levels which is local and which accounts for the randomness in the observations.

In fact, any reasonable function fit to the level before the first pumping and between the two pumping times will correlate extremely well with the observations, since the magnitude of the errors is so small. Taking the regression model as our best estimate of the true level in the tank, derivatives of the level either at the observed times or at hourly intervals results in flow rates. But this solves only the simplest aspect of the problem. The major stumbling block for this problem is how much water was used during the pumping times and at what rate.

Before turning to the details of the methods in the winning solutions, we should consider their key assumptions; that is, the ones allowing them to make particularly good use of the data. The Hiram team assumed that the pump rate is constant. This excludes certain plausible but unlikely possibilities; for example, that the pump has some sort of emergency high gear for times of extraordinarily heavy water use. The team from Alaska—Fairbanks assumed that the flow rate has continuous second derivatives, which is essential to their spline approximation. The Ripon paper is perhaps more distinguished by what the team members did not assume. Unlike the other two solutions, which depend to some extent on the accuracy of the level measurements, they made no assumption about the accuracy but checked their results against what they would have concluded if the data had been perturbed in some unfortunate way.

Rather than model the level in the tank, as I might have done, the teams chose to convert the levels to rates, using numerical estimates of the derivative of $L(t)$. Fitting a global model to these estimated rates, instead of to the level (or volume) measurements, has the strong advantage that flow rates during the pumping times can be estimated by interpolation.

Each team took a different approach to this estimation. The Hiram team used the crude rates, that is, $(L_i - L_{i-1})/(t_i - t_{i-1})$, which is a numerical approximation to the derivative of $L(t)$ at $(t_i + t_{i-1})/2$. The team from Alaska—Fairbanks began by taking more sophisticated numerical approximations to the derivative of $L(t)$, which are smoother than the crude rates, since they incorporate information from more points. The Ripon team developed their own method for estimating the derivative. To consecutive sets of three points, they fit quadratic polynomials, assigning the slope of the parabola at the middle point as their estimate of the derivative of the cubic there.

But there are some disadvantages to estimating rates. The random errors are given as 0.5% of the level, or about 0.16 feet. This amount may be negligible with respect to level measurements, but the flow rates lie mostly in the range from 0.5 to 1.0 feet per hour. The crude rates involve two level measurements, so their error is on the order of 0.3 feet, certainly not negligible with respect to the size of the flow rates. In addition, any two consecutive crude rates share a level measurement. This introduces some serial correlation, and least squares fits to non-independent measurements

are more difficult to interpret. Other estimates of the rates have the same disadvantages as the crude rates: the random error has been ignored, and it is appreciable compared to the rates. However, since the more-complicated numerical approximations are effectively averaging over three or more points, the propagation of the random error is less exaggerated.

No matter what kind of smoothing or approximation is done to the crude rates, or to any estimate of the individual rates, the unknown flow rate during the first pumping time causes problems. The change in flow rates appears modest at all times *except* the two-hour period beginning about nine hours after the first measurement and ending about eleven hours after the first measurement. During this period, the flow rate nearly doubles, covering practically the entire range of rates. Of course, we do not know what time of day this may represent, but a good guess would be sometime in the morning, during which the majority of our small town denizens perform their ablutions. Indeed, the Hiram team took the resourceful step of looking at the demand for electricity: there, a dramatic rise is seen between 6 and 9 A.M.

The magnitude of this difficulty is such that one is driven away from a global model for the rates toward a local model. A global model will tend either to smooth this jump so that the overall shallower trend is preserved, or fit the jump together with other smaller (probably spurious) wiggles elsewhere in the data. The global model that the Hiram team chose, an eighth-degree polynomial, tends more toward the latter and results in predicting a second peak in the flow rates around 22 hours after the first measurement—during the second pumping time, a peak that is only weakly supported by the local data.

For a model of the flow rates over time, the Alaska—Fairbanks team chose cubic splines over polynomials. Cubic splines are defined on a set of join points (knots), so that they are cubic polynomials between knots and have matching first and second derivatives at each (interior) knot. A spline fit need not interpolate the data, but this team used an interpolating spline. An important feature for the problem at hand is that, unlike an interpolating polynomial—for which perturbing a single point can have global effects—a spline is a local approximation: its value at any given point is not affected by data sufficiently far away. If the team had not tried to interpolate the data, theirs would have been close to the ideal fit, in my opinion. Since there is still appreciable random noise in the data, interpolation produces a few probably spurious features.

The Ripon team took an intermediate position on modeling the flow rates. Instead of fitting a polynomial to all the estimated derivatives, they divided the data into two (overlapping) parts, roughly before and after the first pumping time. To the first of these, they fit an eighth-degree polynomial, and to the second, a ninth-degree polynomial, by least-squares. These two fits were joined together ad hoc; but the result is a piecewise polynomial, each piece a

a local estimate for its (roughly) half the data. In addition, since the team was not concerned with generalizability and their estimated derivatives were rather smooth, they were able to nearly interpolate the data.

Using a polynomial approximation to flow rate and the assumption t the pumping rate is constant, the Hiram team estimated the flow rate d ing pumping times and the pumping rate itself. This is an attractive featur of their solution: estimating the pumping rate independently during both pumping times gives nearly equal results and makes us confident that the estimate for the flow rates is reliable. Both the Alaska—Fairbanks and the Hiram teams simply used the flow rates predicted by their overall models during the pumping times, and estimated the total water consumed by integrating their models.

The Ripon team, on the other hand, took what in my opinion is the wiser course. While some sort of complicated and error-prone method is necessary to estimate the flow rate, almost all the water consumed can be estimated, quite accurately and with no additional assumptions, by just subtracting levels. For periods when level measurements were made without interruption, use the difference in the first and last; these differences will have errors of about 1% each. Where there are gaps, estimate the flow rate using the flow-rate model, multiply by the length of the gap, and add this to the total. In fact, to estimate the flow rate during the pumping times, the Ripon team did not use their flow rate model but fit two new models to the points surrounding the missing measurements.

The Alaska—Fairbanks team finished their solution by attempting an analysis of the error in their estimate for the total water consumed, and by validating their model on three days of data from their local water utility. Doing such validation is an outstanding idea, and the results are intriguing, since the flow rates in these three days show a lot of noise and not much structure. The Hiram team compared their model of water use to some published data on electricity use. Finally, in the only serious attempt to grapple with the random errors in the level measurements, the Ripon team added to the original data perturbations consistent with the 0.5% errors, creating two "worst-case scenarios." Repeating the modeling process on these fabricated datasets gives some idea in the variability of the solution.

None of these solutions is perfect. More than anything, they cry out for a proper accounting for variability due to random errors in the original measurements. There is also a tendency to include material that has no bearing on the problem, and to extend the analysis beyond what is required. More forgivable is the occasional misuse of a technique or formula. But despite any criticism, all the teams are to be commended for producing estimates of high quality in the face of a truly challenging problem.

About the Author

David Reboussin received a B.A. in philosophy and mathematics from Pomona College in 1982 and an M.S. in statistics from the University of Chicago in 1984. He worked for three years as a quantitative analyst for the RAND Corporation in Santa Monica, California, after which he enrolled in the Ph.D. program in statistics at the University of Wisconsin—Madison. In 1985 he was awarded an NIH traineeship in biostatistics, and he is currently writing and researching his dissertation.