# Pump Up the Volume!

Lieke Daley
Dirk Helgemo
John Kenny
Ripon College
300 Seward St.
P.O. Box 248
Ripon, WI 54971

Advisor: Robert J. Fraga

## Summary

We estimate the flow $f$ of water out of the municipal water tower of a given small town at all times during a specific day and determine the amount of water used by the community in that day. We are given the dimensions of the water tank and measurements of the water levels (in feet) at specific times (in seconds); we graph the data in Figure 1.

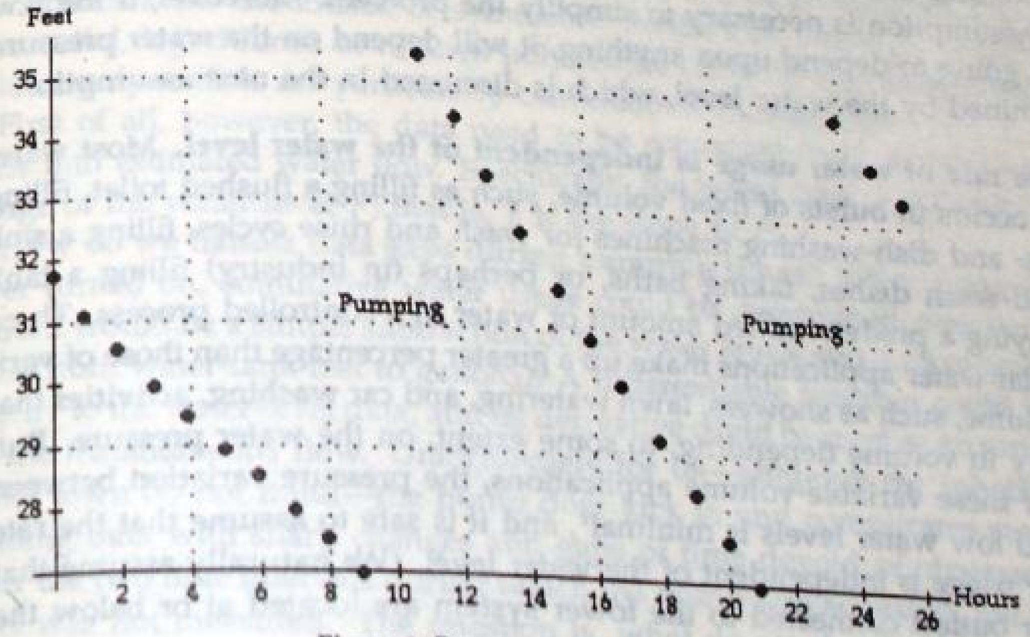

Figure 1. Data: water level vs. time.

From these data we generate a set of slope estimates, which we designate as $P_0$. We pass a parabola through the water-level data, three points at a time, and use the first derivative of the parabola to approximate the instantaneous flow rate for particular times.

We use the least-squares technique to fit a single polynomial to the set of flow estimates $P_0$ as a preliminary approximation to $f$. To improve the fit, we split $P_0$ into two smaller sets and apply lower-degree polynomial fits to each of them. The best fit is the compound function $f_1$, consisting of two separate polynomial fits to two subsets of $P_0$. Specifically, we fit polynomials to the subset $P_1$ (the first 13 data points) and to the subset $P_2$ (the last 14 data points). The two subsets intentionally overlap, to encourage the two polynomial fits to overlap.

To determine the total water usage per day, we use a combination of the original water-level data and polynomial fits to slope estimates during pumping times. We determine cumulative water usage during periods of pump inactivity by subtracting the water-level measurement at the start and end of the period; during periods of pump activity, by integrating polynomial fits to the slope estimates at times close to the pumping times. By taking six different 24-hour blocks out of the data (the data cover 26 hours), we come up with 330,000 gallons as the approximate daily water usage of the observed town.

## General Assumptions

**The rate of water usage is independent of the the state of the pump.** This assumption is necessary to simplify the problem. Moreover, if the flow out is going to depend upon anything, it will depend on the water pressure, determined by the water level, which is discussed in the next assumption.

**The rate of water usage is independent of the water level.** Most water usage occurs in bursts of fixed volume, such as filling a flushed toilet, filling clothes- and dish-washing machines for wash and rinse cycles, filling a sink to hand-wash dishes, taking baths, or perhaps (in industry) filling a tank or spraying a predetermined amount of water in a controlled process. These particular water applications make up a greater percentage than those of variable volume, such as showers, lawn watering, and car washing, activities that will vary in volume depending, to some extent, on the water pressure. But, even for these variable-volume applications, the pressure variation between high and low water levels is minimal[1], and it is safe to assume that the rate of water usage is independent of the water level. (We naturally assume that all water outlets connected to the tower system are located at or below the tower base.)

[1] If the water pressure for the town is provided only by the force of gravity, the pressure at a water outlet at the foot of the tower would vary from 8,060 to 10,600 pascals (11.7 to 15.4 psi) from low to high water levels—a change of only 31%. The pressure supplied to water outlets below the foot of the tower would have greater pressure but a lower percentage change. Many water towers, however, have pressure-regulation devices that augment the pressure provided by gravity (to around 50 psi) and regulate that pressure to a constant value.

The pump does not turn on between observations, except as recorded; that is, the pump cycles on and off only twice, and only during the times given. Because this town's population is between 2,000 and 4,000 people[2], we can safely assume that pumping is not required between successive measurements (as would be required if the tower supported a significantly larger population or water-hungry commercial applications).

$F$, the water-level curve, and $f$, its derivative, are piecewise-continuous functions. Discontinuity may appear when the pump is turned off and on. (Note: When discussing functions in general, we shall use $f$ with no subscripts. Subscripts will denote specific functions that we have found in answer to the problem.)

The times at which water levels are taken are accurate to within one second, as recorded.

## The Approach

We determine a function $f$ that estimates water flow at all times during the observed day, and we also estimate the total amount of water used by the community in one day. A convenient and standard way of doing this is to fit a smoothing curve to the data. Specifically, we use polynomial fits; the data do not show monotonic or symmetric trends that would motivate an exponential, logarithmic, or rational polynomial curve fit or transformation. Besides, polynomials are particularly convenient for integration.

First of all, however, the data need to be converted from sample water levels into estimated water flow. Specifically, the water flow $f$ is the rate of change of the water level $F$; that is, $f$ is the slope of $F$.

How do we handle data gaps during pumping times? If the pump were never turned on, cumulative water usage could be determined directly: the amount would be a simple subtraction of the water levels and a conversion of units (from water-tank feet to gallons). A differentiable function could be fit directly to the water-level data, and its derivative could be used a an estimate for the instantaneous flow. Unfortunately for us—although the inhabitants of the town do not mind—the pump does kick in and complicates matters, yielding data with sharp changes and gaps of time devoid of observations. Over the two time intervals during which the pump was activated, the water level was not measured. The question is, what do we do about these data gaps?

**Answer:** We ignore them. Knowing the water levels during pump activity might help us better approximate water usage. But without these data

[2] Assuming that the residential and commercial water usage is typical of U.S. towns and cities, the approximate population is 3,100 to 3,200—see the section on Estimating Water Usage.

points, and without knowing beforehand the output rate of the pump, we cannot estimate flow rates over those periods without relying on a curve-fit based on water flow before and after pump activity. However, doing this does not improve the quality of our final curve-fit; because such estimates would be determined by the curve-fit, they contribute nothing to a re-evaluation of that curve.

## Converting Water Level to Water Flow

We determine the slopes of the line segments connecting consecutive water-level data points, locating each slope at the average time of the pair of points in question (see Figure 2).

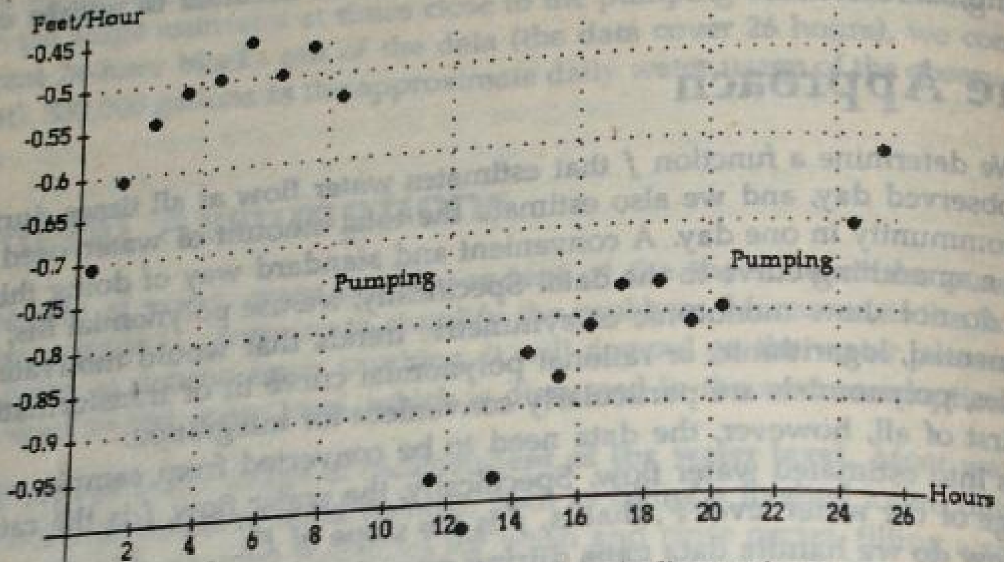

Figure 2. Water flow vs. time—piecewise-linear estimate.

This piecewise-linear fit to the height data does not make for a very good water-flow estimate. We note that the data are slightly scattered; and the polynomial fit to the set is somewhat inaccurate, with error sum of squares (SSE) of 0.0238. Moreover, merely finding the slope of the segment connecting two points from the original data set is not necessarily the best approximation of rate. In fact, in determining the error involved in these water-flow estimates, we find that there is a great deal of variability about the given points.

So, instead of using averages, we find the instantaneous slope at each point $t_i$, by looking at the points $t_{i-1}$ and $t_{i+1}$ and finding the parabola that passes through all three points. The derivative of this parabola provides the instantaneous slope for the central point at $t_i$. In this case, the error per point is much less than in the piecewise-linear approach (see the section on Error Analysis for more detail).

*Pump Up the Volume!* 213

Endpoints require special treatment because they cannot depend on either the preceding or the following point. (Because of the gaps in the data, we consider any point directly preceding or following pump activity to be an endpoint.) For endpoints, we use the same parabola as was used for either the previous point or the following point (depending on whether the endpoint is at the right or left edge of the graph) but evaluate the parabola at the point to be estimated.

The resulting estimates appear significantly smoother and conform better to a polynomial fit (see **Figure 3**), having a significantly lower SSE of 0.0117. The remainder of our analysis uses this set $P_0$ of instantaneous slopes.

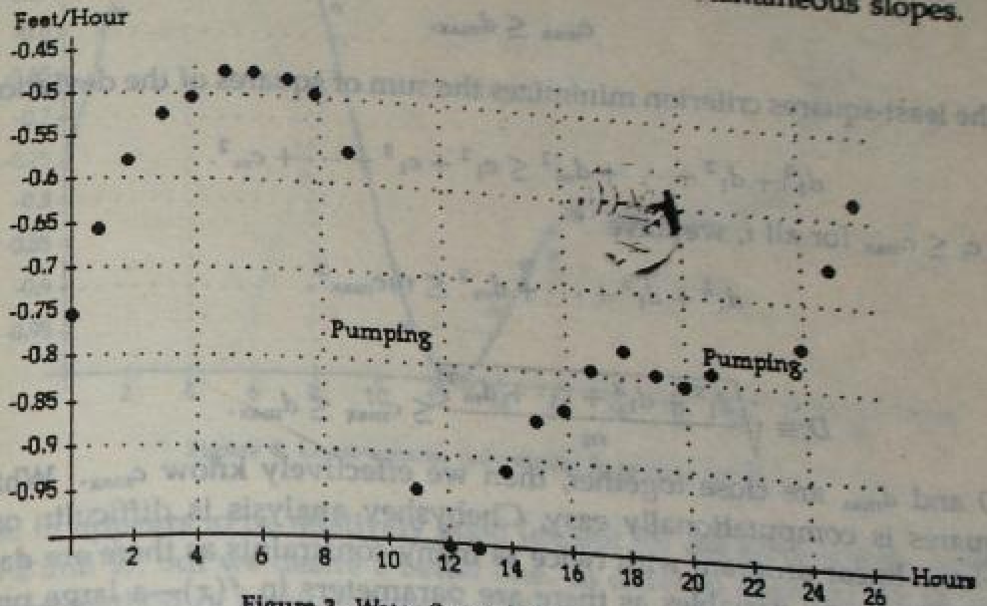

**Figure 3.** Water flow vs. time—parabolic estimate.

## Chebyshev vs. Least-Squares

In trying to find the best way to fit a curve to the data, we consider the Chebyshev approach vs. to the least-squares fit. What follows is a generalization of the Chebyshev approach.

> Given some function type $y = f(x)$ and a collection of $m$ data points $(x_i, y_i)$, minimize the largest absolute deviation $|y_i - f(x_i)|$ over the entire collection. That is, determine the parameters of the function type $y = f(x)$ that minimizes
>
> $$\text{Maximum } |y_i - f(x_i)| \qquad i = 1, 2, \ldots, m.$$
>
> This important criterion is often called the *Chebyshev approximation criterion.*
>
> [Giordano and Weir 1985, 96–97]

We pick as our function type polynomials of a fixed degree $n$. We call the resulting Chebyshev approximation polynomial $f_1(x)$, and we let $c_{\max}$ be the largest of the absolute deviations

The least-squares approximation criterion yields a second polynomial $f_2(x)$. It has absolute deviations

$$d_i = |y_i - f_2(x_i)|, \qquad i = 1, 2, \ldots, m,$$

with largest absolute deviation $d_{\max}$. Since $c_{\max}$ is by construction the *minimal* largest absolute deviation for polynomials of degree $n$, we have

$$c_{\max} \le d_{\max}.$$

The least-squares criterion minimizes the sum of squares of the deviations, so

$$d_1^2 + d_1^2 + \cdots + d_m^2 \le c_1^2 + c_1^2 + \cdots + c_m^2.$$

Since $c_i \le c_{\max}$ for all $i$, we have

$$d_1^2 + d_1^2 + \cdots + d_m^2 \le m c_{\max}^2,$$

or

$$D \equiv \sqrt{\frac{d_1^2 + d_1^2 + \cdots + d_m^2}{m}} \le c_{\max} \le d_{\max}.$$

If $D$ and $d_{\max}$ are close together, then we effectively know $c_{\max}$. While least-squares is computationally easy, Chebyshev analysis is difficult: one must solve a linear program with twice as many constraints as there are data points, and as many variables as there are parameters in $f(x)$—a large program! Thus, it is not worth the effort to calculate $f_1(x)$ using the Chebyshev approximation criterion for the very minimal benefit of reducing one measure of error, the largest absolute deviation. In our case, we find that $D$ and $d_{\max}$ are indeed very close for our final polynomial fit, so we use the least-squares approach to determine $n^{\text{th}}$-degree polynomials to fit the data.

Earlier, we discarded the approach that used a piecewise-linear data set, because of its apparent inaccuracy. The $D$ and $d_{\max}$ values for this set are 0.022 and 0.063. Because there is a large difference, a Chebyshev approach would be necessary; and thus we have another reason for not considering further the piecewise-linear data.

We use the software package Mathematica, with its command Fit, to find least-squares polynomial fits to our slope estimates. To determine the specific degree, we loosely follow the guideline that the power of the polynomial should not exceed $2\sqrt{n}$, where $n$ is the number of data points [Germund and Bjork 1974, 112].

## Results

Our first application of least-squares analysis is to find a polynomial fit for the set of parabolic slope estimates $P_0$. We try many different degrees in the attempt to find the *best* polynomial fit to the slope estimates; an 11th-degree polynomial yields a sum of squares for error of 0.0117. We call this fit $f_0$ (see Figure 4).

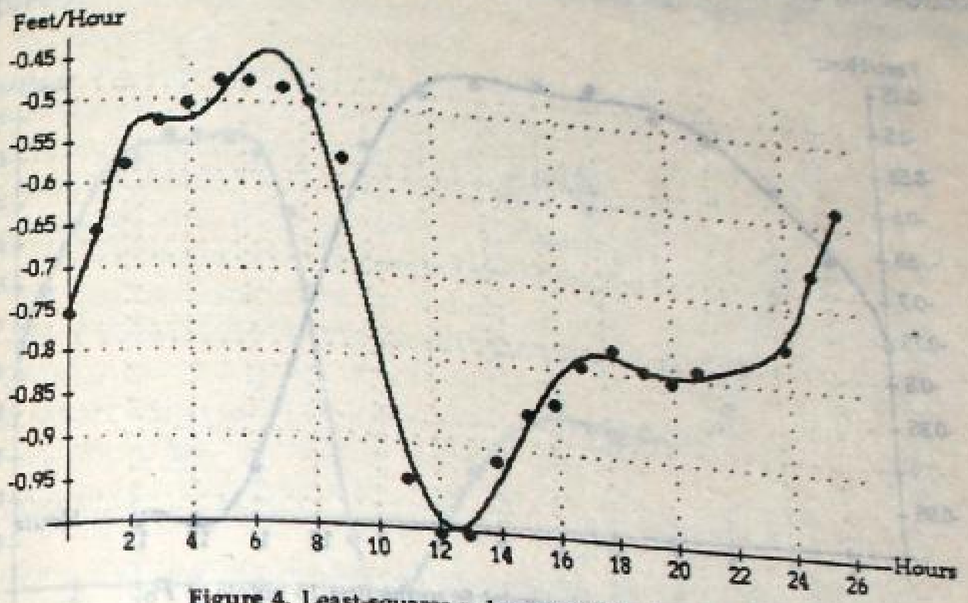

**Figure 4.** Least-squares polynomial fit to data set $P_0$.

The fit appears to be relatively good (except for the extra wiggle between hours 2 and 5), but we desire a better one. In order to improve the fit, using the same method of least squares, we change the way we look at the data. Noticing that the first part of the graph (before the pump first comes on) appears to be very smooth and therefore conducive to a polynomial fit, we split off that part of the graph for a separate fit. The remaining points are used to fit a second polynomial. This approach will be explored in depth in the following section, since it is the heart of our attack.

## Splitting the Data

The best way to model these data is to split them up and fit a curve to each subset. Our approach is to find a fit to the first six points that has a lower sum of squares for error, while smoothing the data, than the polynomials obtained doing a least-squares fit to the entire data set. We add additional points consecutively from $P_0$ until the polynomial fits fail to retain low error with good visual conformance to the slope estimates. (Interestingly enough, sometimes the presence of additional points helps to reduce the sum of squares for error by "persuading" the polynomial to adjust to data trends.) This procedure is done with the last six points as well, working in the opposite direction.

We found that the best two-piece fit to $P_0$ is obtained by fitting a polynomial $P_1$ of degree eight to the first 13 points and a polynomial $P_2$ of degree nine to the last 14 points. The first-half fit yielded an SSE of 0.00107; the second-half, 0.00165 (see Figures 5 and 6).

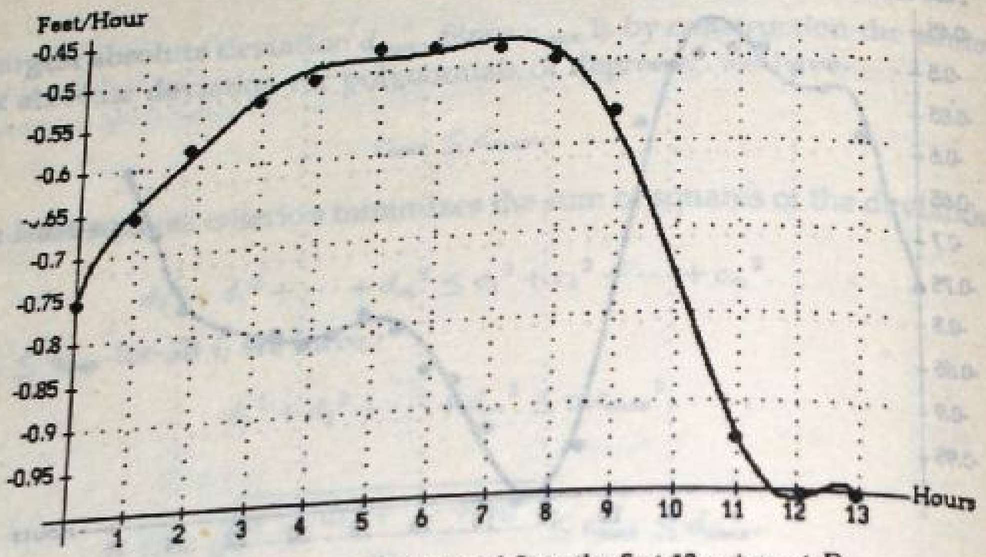

Figure 5. Least-squares polynomial fit to the first 13 points of $P_0$.

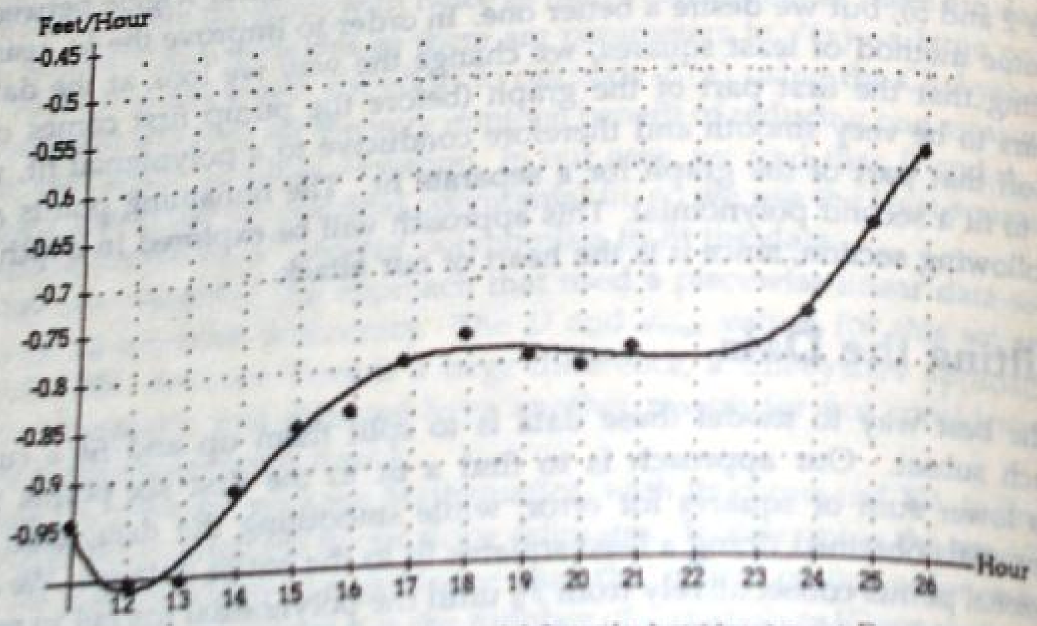

Figure 6. Least-squares polynomial fit to the last 14 points of $P_0$.

Combining these two curves gives us our second approximation for the water flow. This fit has a total SSE of 0.00267 (the error is actually slightly less than this, because the curves overlap), as opposed to our first single-polynomial approximation, which has an error of 0.0117. The fits are patched together at time 11.52 hours, the point of intersection with the smallest difference in slope. The combination of these two curves we label $f_1$ and depict in Figure 7.

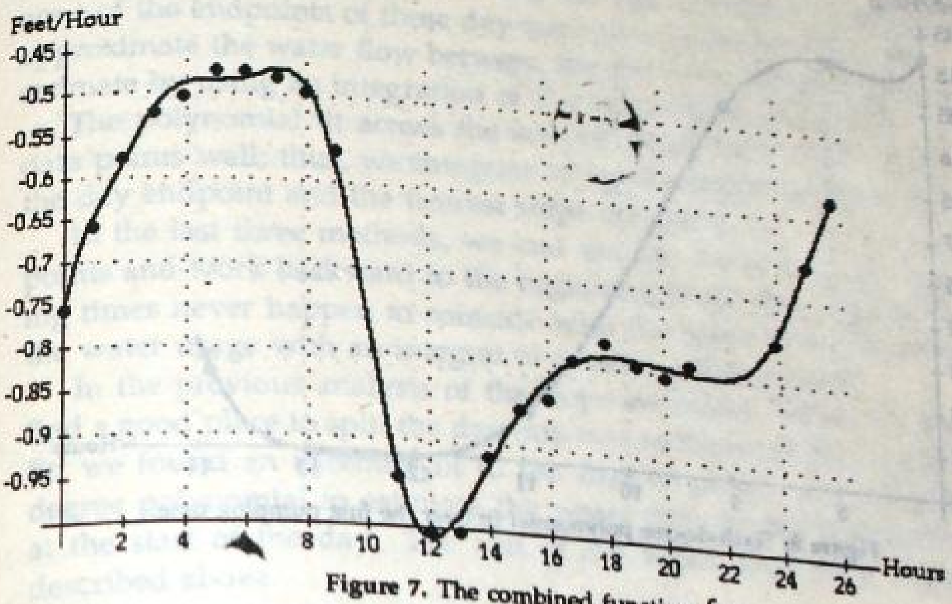

Figure 7. The combined function $f_1$.

## Estimating Water Usage

The only reason we can't figure out the water usage directly from the data, by simply subtracting the ending water level from the beginning one, is that the pump refills the tank twice during the day—causing a sudden increase in the water level, with the amount of pumped water being unknown (though it can be estimated). Besides, we have no data for the water level during pump activity. In order to find the total amount of water used, we could integrate the estimated water-flow function $f_1$ over a 24-hour period. But this is not the most accurate method, because $f_1$ is a fit to *all* of the data and may not accurately represent the points at or around the pumping times, even though it may well be a very good *overall* approximation.

What we do, then, is focus on determining an estimate for $f$ through the unknown periods, using the empirical data over the known intervals. For both instances of the pump turning on, we fit a curve to the two data points before pumping and to the two after, and then add additional data points on each side until we get the best visual fit (smooth, yet representative of trends in the data, including observed concavity) and the lowest SSE.

For the first pumping time, we fit a sixth-degree polynomial $p_1$ to the seven data points around the unobserved interval; the SSE for this fit is effectively zero ($1.42 \times 10^{-33}$). For the second pumping time, we fit a fourth-degree polynomial $p_2$ to the seven points around the interval, with an SSE of 0.0049. (See Figures 8 and 9 for the two fits.)

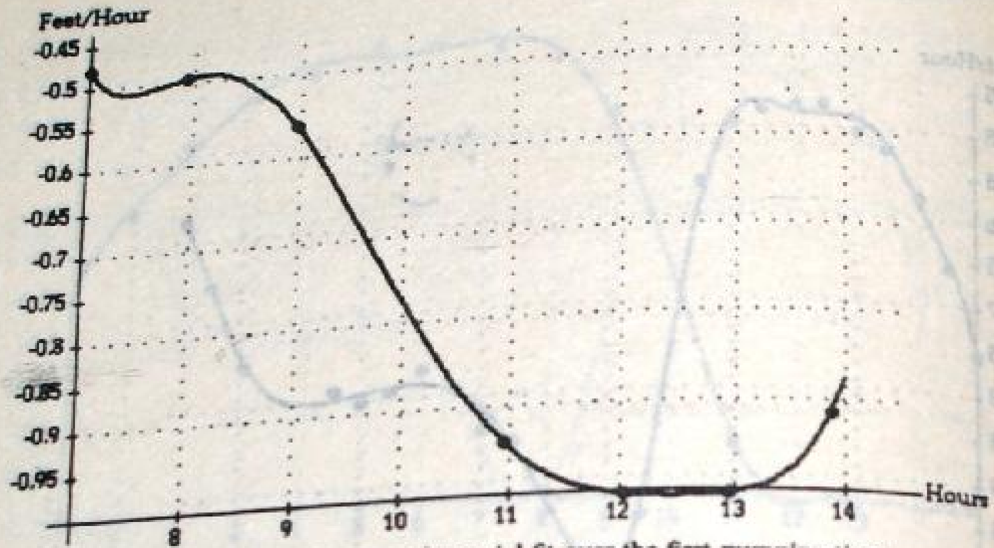

**Figure 8.** Sixth-degree polynomial fit over the first pumping time.

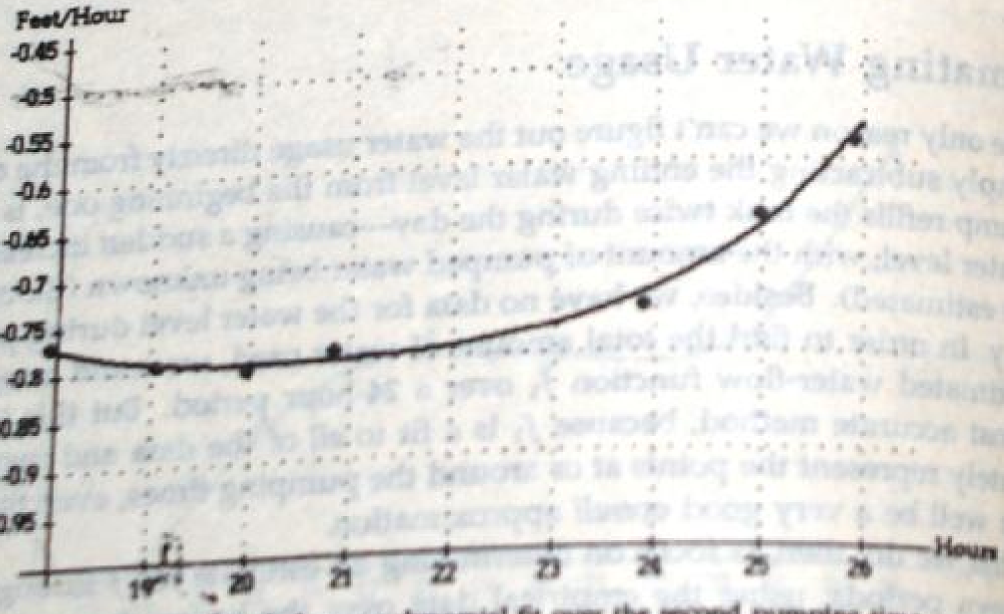

**Figure 9.** Fourth-degree polynomial fit over the second pumping time.

We can now determine water usage. First we determine the amount of water used over the periods of pump inactivity by subtracting the water-level measurements at the boundaries of pump inactivity. Then we utilize the two polynomial fits $p_1$ and $p_2$ to estimate the water usage during the pumping times. The sum of these values estimates the water usage for the entire observation period.

Unfortunately, the data span a period of 25 hr, 54 min, and 30 sec; what we need is the water level over a normalized time period of 24 hours (= 86,400 seconds). To approximate the normalized usage, we use six different start and stop times and average the results.

In the first three methods, we start at each of the first three data points and go to a point near the end of the data set that is exactly 24 hours later. Since none of the endpoints of these day-estimates coincides with a data point, we approximate the water flow between the day endpoint and the nearest slope estimate by using an integration of the appropriate polynomial fit.

The polynomial fit across the second pump time also fits the last three data points well; thus, we integrate to approximate the water flow between the day endpoint and the nearest slope estimate.

In the last three methods, we end the day on each of the last three data points and work backward to the beginning of the day. Again, the day starting times never happen to coincide with the given data, so we approximate the water usage with an integral of another polynomial fit.

In the previous analysis of the slope estimates, when we were trying to find a good place to split the day into two sections for the double-polynomial fit, we found an excellent fit to the first 10 points; we integrate this fifth-degree polynomial to estimate the water flow to the nearest slope estimate at the start of the day. The rest of the water level data is determined as described above.

**Table 1** summarizes the six methods and their estimates of water usage. The approximate water usage for this day, for this town, is 330,000 gallons.

According to Leeden et al. [1990], the average water use per person is 105 gallons a day. By dividing this into the water usage (in gallons per day), we can estimate the population of this small town to be about 3,150 people.

**Table 1.**

Estimates of water usage for each of the six methods.

| Time interval (times in seconds) | Water usage (gals/day) |
|---|---|
| [0, 86,400] | 329,629 |
| [3,316, 89,716] | 329,659 |
| [6,625, 93,035] | 330,113 |
| [6,687, 93,270] | 330,032 |
| [3,553, 89,953] | 329,885 |
| [−432, 85,968] | 329,910 |
| Average | 329,871 |

## Error Analysis

In both analyses of the slope estimates, we created two worst-case data scenarios; specifically, when the water level of the first measurement is actually 0.5% lower than the measurement, then the water level of the second measurement is actually 0.5% higher, the third is actually 0.5% lower, etc., in an oscillatory pattern. The two scenarios consist of starting with the first measurement being high, then oscillating; and starting with the first measurement being low, then oscillating.

## Piecewise-Linear Slope Estimates

Alas, the water-level data points are so close together (in time) that slight errors in measurements lead to uniformly large percentage errors in the slope estimates, averaging 46.3% (see Figure 10).

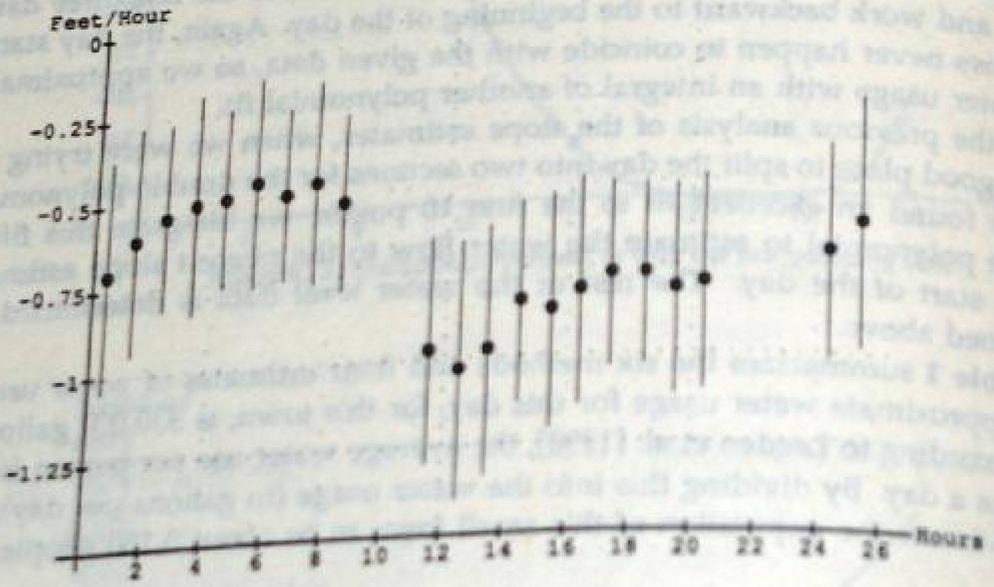

Figure 10. Maximum possible error in the piecewise-linear slope estimates.

## Parabolic Slope Estimates

The parabolic fit to the water level data points provides a significantly better fit for data not adjacent to endpoints (the start or finish of the observations or just before or after pumping times). These points fared an average percentage error of 6.32%. However, the endpoints suffered an average percentage error of 91.3% (the measurement error often caused the concavity of the parabola to flip, having a severe effect on the endpoints). Fortunately, however, in fitting polynomials across these points, our use of several of the nearby non-endpoint estimates allows us to have confidence in the fits (see Figure 11).

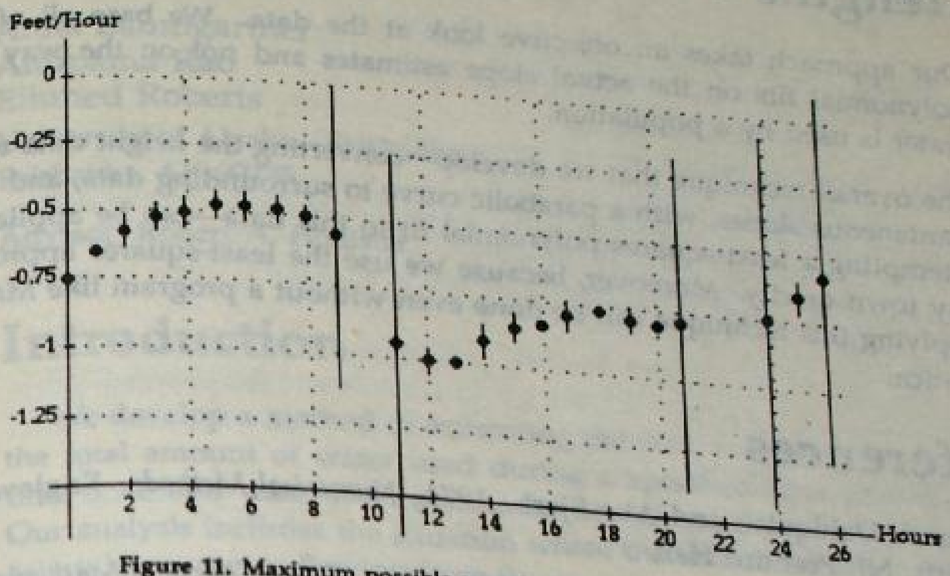

**Figure 11.** Maximum possible error in the parabolic slope estimate.

## Water Usage

Taking into consideration the given measurement error of 0.5% for the water level as given by the problem, we find an upper bound on our approximation for water usage of 357,000 gallons and a lower bound of 319,900 gallons.

## Weaknesses

- Determining where to split the data involves some subjectivity. However, because the final divisions decided on were chosen because they represented the fits with the smallest SSE, this subjectivity is very minimal.
- Because of the delicate nature of our polynomial fits, small errors in the data could change the curve to such a degree as to change sections that were increasing to be level or even decreasing.
- Leeden et al. [1990] provides a table showing "Typical Urban Water Use By a Family of Four." The question is, how can they do that, given the variability of water use from, say, spring to fall? A table like this is exactly what it says—typical—and is determined from data for the previous year's water use—a lot of data. A weakness of our model is that a generalization beyond a few weeks is not feasible. This is because we have only one day of data!

## Strengths

- Our approach takes an objective look at the data. We base all of our polynomial fits on the actual slope estimates and not on the way that water is used by a population.
- The overall technique that we develop—converting the height data to instantaneous slopes, with a parabolic curve to surrounding data, and then attempting a least-squares polynomial fit to this data—can be applied to any town or city. Moreover, because we use the least-squares approach, applying this technique can be done even without a program like Mathematica.

## References

Germund, Dahlquist, and Ake Bjork. 1974. *Numerical Methods*. Englewood Cliffs, NJ: Prentice Hall.

Giordano, Frank R., and Maurice D. Weir. 1985. *A First Course in Mathematical Modeling*. Monterey, CA: Brooks/Cole.

van der Leeden, Frits, Fred L. Troise, and David Keith Todd. 1990. *Water Encyclopedia*. 2nd ed. Chelsea, MI: Lewis Publishers.

Maron, Melvin J. 1982. *Numerical Analysis*. New York: Macmillan.