Dangers of Automated Data Analysis

William J. Leonard

Department of Physics & Astronomy and Scientific Reasoning Research Institute,
University of Massachusetts, Amherst, MA 01003-4525; wjleonard@phast.umass.edu

Recently, I was reading an advertisement for a computer-based lab tool that is used to measure the acceleration of a "picket fence" as it drops through an infrared (IR) photogate.[1] The apparatus takes position and time data and sends them to the computer. From this data, provided software computes velocity and acceleration data, and generates plots of position, velocity, and acceleration versus time. Table I summarizes sample data presented in the ad (hereafter referred to as "the SensorNet data").

What struck me as odd is that the time data have five significant digits, yet the individual accelerations fluctuate between 9.4 m/s^2 and 10.8 m/s^2, with all but one of them over 10.2 m/s^2. How can time data that precise yield such bad acceleration results? I decided to analyze the SensorNet data myself.

There are three ways to compute the acceleration using a given data set: (1) using a quadratic fit to the position-vs-time data; (2) using a linear fit to velocity-vs-time data; and (3) averaging individual acceleration data. I applied each of these methods to the SensorNet data in Table I, and got the results shown in Table II. Also included in Table II are estimates of the statistical uncertainties in the calculations of the acceleration.

Table I. Data table as shown in SensorNet advertisement.

Time (s)	Distance (m)	Velocity (m/s)	Acceleration (m/s^2)
0.01125	0.0		
0.03305	0.05	2.2936	
0.05305	0.1	2.5	10.32
0.07161	0.15	2.694	10.453
0.0891	0.2	2.8588	9.423
0.10561	0.25	3.0285	10.279
0.12125	0.3	3.1969	10.767
0.13615	0.35	3.3557	10.658

Table II. Three calculations of acceleration of picket fence using SensorNet data in Table I.

Method	Acceleration	Statistical uncertainty (estimates only)
quadratic best-fit to the position-vs-time data	9.87 m/s^2	0.07 m/s^2
slope of velocity vs time	10.22 m/s^2	0.14 m/s^2
average of individual accelerations	10.3 m/s^2	0.5 m/s^2

The quadratic fit to the data is both accurate and precise (<1% deviation from the expected value), which means that SensorNet's position and time data are fine. However, the slope of velocity versus time is about 3σ away from the expected value, and the average of the individual acceleration data is both inaccurate and imprecise. Position and time data that are as precise as that shown in the ad should yield better results than this.

The reason for the poor results stems from the way the velocity data are plotted and the way the acceleration data are computed. When a picket fence is dropped through an IR photogate, the displacement intervals are constant, so the time intervals are not. The average velocities are graphed as a function of the *endpoints* of the time intervals for which they are calculated. Although this method gives correct results when the time intervals are all the same, in this case it skews the slope calculation and severely alters the individual acceleration estimates.

When the acceleration is constant, the average velocity over any time interval is equal to the instantaneous velocity at the *midpoint* of the time interval, as shown graphically in Fig. 1. The displacement of the object between t_1 and t_2 is the area below velocity versus time during the same time interval. The area of the shaded trapezoid is equal to the area of a rectangle of height $v_{ave,12}$ and width $t_2 - t_1$. Because the acceleration is constant, $v_{ave,12}$ is equal to the instantaneous velocity at $t_{mid,12}$.

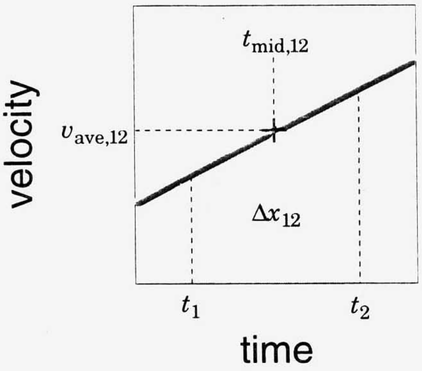

Fig. 1. Sketch of velocity vs time for an object experiencing constant acceleration. Shaded region between times t_1 and t_2 is displacement of object. Average velocity during the time interval is equal to the instantaneous velocity at midpoint of the time interval.

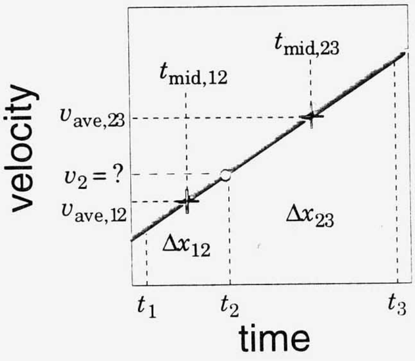

Fig. 2. Sketch of velocity vs time for an object experiencing constant acceleration. Velocity at time t_2 is an interpolation of velocities "measured" at $t_{mid,12}$ and $t_{mid,23}$.

If we re-plot the average velocity versus the *midpoint* in time, we get a best-fit slope of 9.88 ± 0.13 m/s^2 (compared with 10.22 ± 0.14 m/s^2 with the SensorNet analysis). Using the midpoints in time to re-compute individual accelerations, we find that the new average is 9.89 ± 0.37 m/s^2 (compared with 10.3 ± 0.5 m/s^2). Note that these new results are consistent with the quadratic fit to the original position-time data.

Although the midpoint method works fine, it is very inconvenient, because now students have one set of time data for the position measurements, a new set of time data computed for the velocity estimates, and a second set of new time data for acceleration. Is there a way to compute velocity and acceleration data for the same instants of time that the position data were taken? There is, and it is best described using another sketch of velocity versus time, as shown in Fig. 2.

We assume that the acceleration is constant. We assume also that we have a set of position-vs-time data, three of which are taken at times t_1, t_2, and t_3. The goal is to compute the velocity at t_2 using only position and time data.

The average velocity during any time interval is equal to the instantaneous velocity at the midpoint of the interval, which means that we know the velocity just before and just after time t_2. The velocity at t_2 is now determined by interpolating between these two instantaneous velocities. After simplification the result is:

$$v_2 = \frac{v_{ave,12}\,\Delta t_{23} + v_{ave,23}\,\Delta t_{12}}{\Delta t_{13}} \qquad (1)$$

where $v_{ave,12} = \Delta x_{12}/\Delta t_{12}$ and $v_{ave,23} = \Delta x_{23}/\Delta t_{23}$.

Note that v_2 is computed completely using position-time data, and it is valid whether the time intervals are equal or not.[2] With n position-vs-time data, we can use this relationship to make $n - 2$ estimates of the velocity. Using this method and the SensorNet data, the slope of velocity versus time yields a best-fit acceleration of 9.86 ± 0.13 m/s^2 (compared with 10.22 ± 0.14 m/s^2 using endpoints in time).

Note that when the time intervals are equal, Eq. (1) reduces to:

$$v_2 = v_{ave,13} = \frac{\Delta x_{13}}{\Delta t_{13}} = \frac{x_3 - x_1}{t_3 - t_1}$$

(for equal time intervals) (2)

because t_2 is the midpoint in time between t_1 and t_3.

A simpler relationship emerges for the acceleration. Because the acceleration is constant, an estimate of the acceleration at time t_2 is the slope of the line

Table III. Comparison of valid methods for analyzing position-time data in Table I.

Method	Acceleration of the picket fence	Statistical uncertainty (estimates only)
quadratic best-fit to the position-vs-time data	9.87 m/s^2	0.07 m/s^2
slope of velocity vs time (midpoint method)	9.88 m/s^2	0.13 m/s^2
(interpolation method)	9.86 m/s^2	0.13 m/s^2
average of individual accelerations	9.89 m/s^2	0.37 m/s^2

"Dangers of Automated Data Analysis" VOL. 35, APRIL 1997 THE PHYSICS TEACHER 221

connecting the points ($v_{ave,12}$, $t_{mid,12}$) and ($v_{ave,23}$, $t_{mid,23}$) in Fig. 2. The result is:

$$a_2 = \frac{v_{ave,23} - v_{ave,12}}{t_{mid,23} - t_{mid,12}}$$

$$= 2\frac{v_{ave,23} - v_{ave,12}}{t_3 - t_1} \qquad (3)$$

Like Eq. (1), Eq. (3) is purely a function of the original position-time data. Note also that because the acceleration is constant, the set of accelerations computed using Eq. (3) is identical to the set computed using the midpoint method.

Although the expressions for v_2 and a_2 are too cumbersome for most high-school students using only calculators, they aren't too much for students using spreadsheets, and they are certainly not too difficult for software designers to use.

We now have three (or four, depending on how you count) valid methods of determining the acceleration of a picket fence falling through an IR beam, as summarized in Table III. How should teachers decide which is the preferred method? Consider first the quadratic fit to the position-time data. Although this method is the most precise, it is also the most abstract, and many students likely would not understand it. Further, it is impossible to visually distinguish between a parabola and other curved functions, so students would not be able to perceive that the raw position-time data correspond to constant acceleration. The method of averaging individual accelerations is fairly accurate, but it is much less precise than using the slope of velocity versus time. Another difficulty with averaging accelerations is that without an appreciation of uncertainty and without visible error bars, many students may find the claim unconvincing that the acceleration is constant. In my opinion, the preferred method is using a plot of velocity versus time. Its slope is relatively precise and relatively accurate. The primary advantage, however, is that the graph looks like it's a straight line, giving students a visual (rather than purely abstract) sense of what constant acceleration means.

Although this exercise is only a single example, there are many lessons we can learn from it:

- **Good raw data do not guarantee good science.** When precise data are mishandled, the results are likely to be skewed. If students do not understand what the measurements or the results represent, then they will not be able to assess if the results are appropriate. I doubt that most students could have discovered the source of the error in this software package even if the presence of the error were pointed out to them.
- **Having lots of significant digits doesn't mean that the results will be accurate.** Students are prone to confuse and interchange the meanings of *accuracy* and *precision*. The time data were precise, and the computations of velocity and acceleration were almost as precise. Yet the results were inaccurate because the data were improperly analyzed.
- **Results from "canned" software packages are only as good as the programmers.** A conceptual understanding of basic physics is missing from the SensorNet analysis. If programmers don't understand the mathematics or the physics involved in an experiment, then the manipulation of data may possibly be wrong.
- **Students need to understand how raw data are handled, and they should not put blind trust into the results of someone else's calculation.** In particular, I would strongly recommend that all students check the calculations done by any software package before using it to analyze their data. This is not beyond the abilities of students.

Automated data analysis relieves students of repetitive tasks, thereby saving students and teachers much time—time that can be used to learn other things. I do not believe, however, that students benefit by using data-analysis procedures without understanding how they work.

References

1. Acculab Products Group, *Sci. Teach.* **63**, 81 (September 1996).
2. Although it is not obvious from the derivation, this result is *identical* to doing an exact quadratic fit through the three position-time data and evaluating the resulting expression for the velocity at time t_2. Ordinarily, this would require solving three simultaneous equations for three parameters (usually x_0, v_{0x}, and a), plugging two of them (v_{0x} and a) into a general expression for $v_x(t)$, evaluating it at $t = t_2$, and simplifying it. Instead, using a sketch of velocity versus time, we can in one step write down an expression for v_2 and then simplify it.

et cetera...

Industry Is Down on Science Education

Apparently U.S. industry is more worried about the poor state of science education than are the educators, according to a recent survey.

> *The telephone poll...asked 600 public elementary school principals and human resource directors from a variety of companies for their evaluation of science education. The answers from the industry people were harsh.... About 60% said that...most young adults lack adequate science preparation for entry-level jobs in industry today...and 84% thought that science literacy—defined as the ability to understand newspaper articles about science—will soon be a requirement for all entry-level jobs.*
>
> *The principals were more optimistic than the [business] executives: More than 75% felt that public schools are successful at turning out students who know how to solve problems and think independently.*[1]

[1] C. Holden, *Science* **272**, 819 (May 10, 1996).

A^{2}B