MESSY DATA AND A MORE REAL-WORLD FUNCTION MACHINE

Scott A. Sinex
Professor Emeritus of Physical Sciences & Engineering
Prince George's Community College
Largo, MD 20774
sinexsa@retiree.pgcc.edu

Abstract

Authentic, "messy data" contain variability that is derived from a multitude of sources. We will examine sources and tweak the function machine to handle multiple variables and errors, a real-world scientific approach dealing with uncertainty. "Messy data" has become a topic in science education and has appeared in the literature. The topic is essentially *rediscovering error*, both random and systematic, which has been around since measurement appeared. Authentic big data (large data sets) that contains scatter needs more consideration in mathematics and the sciences. Here data from NASA, NOAA, and others with respect to climate change will be used to illustrate "messy" big data via Google Sheets.

Introduction

In this paper, an approach to introduce students to large (big) and scattered (messy) data sets will be considered for use in both mathematics and science classes. Most science laboratory experiments in first-year courses are designed to yield really good results, barring the sloppiness of some students. Some of the modeling activities discussed later were designed to yield really good-fit linear models, so that students did not decide on another possibility. Curvature in data and nonlinear systems came later. Most of these activities yield a good fit with a small amount of scatter. These activities typically include a spreadsheet simulation that can address both systematic and random errors. Students can numerically simulate data with more scatter to see how model results (slope, y-intercept, and r-square) are influenced. So, exposing students to larger data sets with more scatter to analyze would be a logical next step. A comparison of first-year chemistry laboratory results vs. results for a tide gauge and the full-blossom dates of cherry blossoms in Japan are given in Figure 1. So, how about the technology to use?

In a recent paper by Rosenberg et al. (2022), they surveyed 330 teachers to see what digital technologies were used by students to analyze and interpret data. The top five technologies are given in Table 1 below. …and the winner is Google Sheets! The spreadsheet is a versatile tool, especially for collaboration, that can handle big data. Plus,

spreadsheet simulations are glass-box simulations where the calculations are visible and Sheets is accessible to all. O'Reilly et al. (2017) introduced Project EDDIE which uses spreadsheets with large data sets . O'Reilly et al. (2022) discuss pedagogical ways to deal with large data sets and ramp-up computer programming and fancy visualizations; however, they recommend starting simple with spreadsheets.

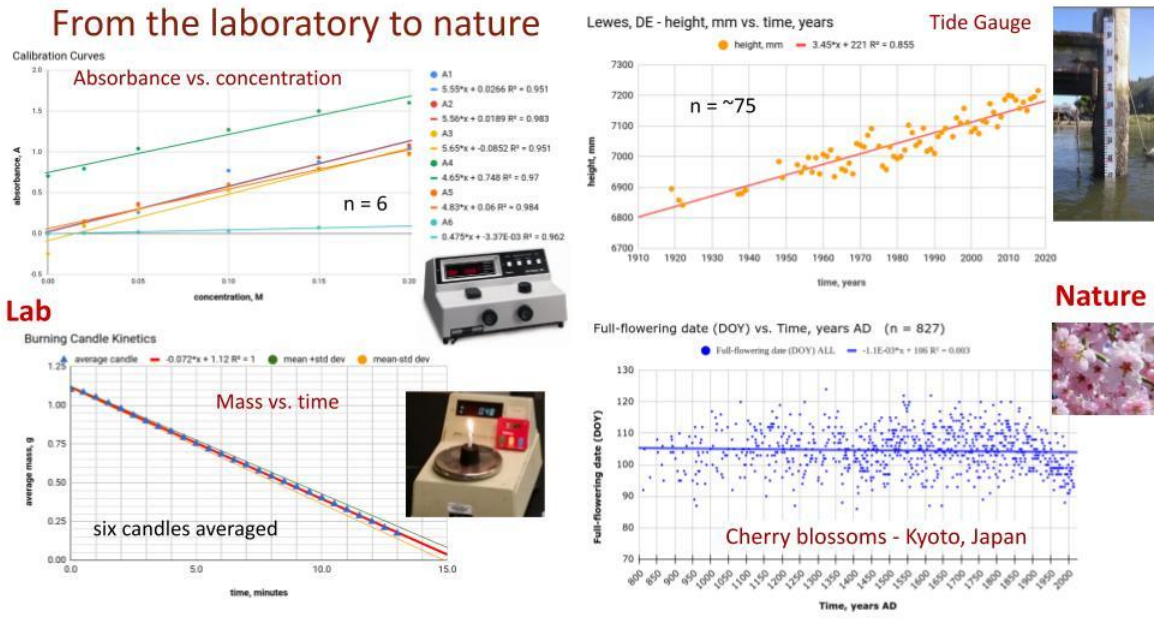

Figure 1 - Student lab data for two experiments, Tide gauge from psmsl.org, and Cherry blossoms from Kyoto, Japan from atmenv.envi.osakafu-u.ac.jp/aono/aono-e/

Table 1 - Digital technologies used by students to analyze and interpret data (n = 330 teachers)*

Tools/Resources - Top Five	Overall	Elementary School	Middle School	High School
Google Sheets	82.7%	81.2%	78.1%	85.8%
Calculator (not for graphing)**	72.1	75.0	71.9	75.0
Microsoft Excel	42.7	34.4	30.7	50.0
Graphing Calculator	28.8	18.8	17.5	38.4
Desmos	15.8	25.0	21.9	15.5

*Data from Rosenberg et al. (2022) **supports hand calculations, not a big data tool!

What is Messy Data?

Authentic, "messy data" contain variability that is derived from a multitude of sources (modified from Schultheis & Kjelvik, 2020):

- natural variation in nature (it's a multivariable world!)
  - random error or noise/scatter in data (can be much larger than from the lab!)
  - variables beyond control (may even be unknown variables!)
- chance occurrences during research such as sampling, sample preparation, measurements in the field and/or laboratory (random error)
- human errors that are *unbeknownst* to investigator (induce bias - systematic error)
  - miscalibration such as glassware, balances (mass), instrument calibration
  - contamination (or loss) especially for trace or low-level analysis
  - sampling bias (was a representative sample collected)
  - blunders? (think: novice vs. expert!)
    - outliers (need to use a statistical test to confirm removal)
    - missing data (missing data is just missing data unless it introduces a sample bias) - students have a large concern for this!
    - using a statistical test and NOT meeting the assumptions of the test
    - data formatting problems
    - miscalculations

Blunders are common in introductory science labs! Are blunders unfortunate or careless human errors (sloppy)? Give students a ruler where the zero is not at the end of the ruler and see if they correct for this fact! Novice learners are confronted with new measurements and need time to develop skills. Dealing with errors is a mindset we need to develop and reinforce.

Figure 2 illustrates how scatter in data from random error increases from developing a

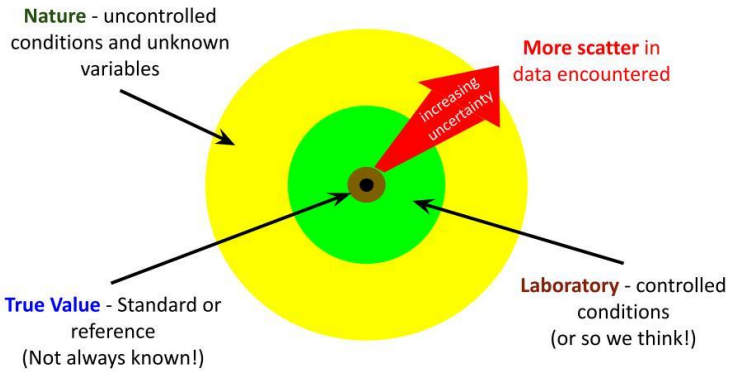

Figure 2 - Scatter in data

standard reference material, the general laboratory, and finally to measurements in nature (field), such as water quality or satellite measurements. Handling variation or scatter in data is something we need to start in middle school, see Hunter-Thomson (2022) for ways of dealing with variability.

Over the years, a number of simple investigations, mainly with simple manipulatives, have been developed that address both random and systematic errors as given in Table 2. It is very easy to add random scatter to data in a spreadsheet using the RANDBETWEEN function, see Sinex (2013, 2016) for instructions. Sometimes students are instructed to make the errors and sometimes the errors are addressed by spreadsheet simulations.

Table 2 - Simple Manipulative Investigations for Error Analysis

Investigation	Measurement	Manipulative	Errors* & Science Process	Reference
Stacking Cookies	Height, ruler	Any sandwich cookie, Oreo cookies for a variety of models	Ruler Error (systematic) Induce errors with poker chips!	Sinex (2012) Sinex et al. (2007)
Stacking Bricks	Height, meter sticks	Real bricks or Lego bricks (rulers)	Not including mortar (revise model) Brick wall sim	Sinex (2017)
Unstacking Coins	Height, ruler or mass by balance	Pennies	Verify with US Mint (judge accuracy) X- & Y-intercepts	Sinex (2018)
Stacking Styrofoam Cups	Height, ruler	Styrofoam cups	Cup Stacking sim Y-intercept Inverse function use	Sinex (2008)
Stacking Nested Cubes	Height, ruler	Nested cubes	Quadratic model Derive model parameters	Sinex (2015)
Tumors Volume	Capilar, balance, & graduated cylinders	M&M Peanut Candies	Comparison with calculated volume from mass using density (y = x plot)	Sinex & Chambers (2019)
Mass of a Bolt	Scale or Balance (to 0.01g)	Nuts and bolts	Extrapolation to find bolt mass	Sinex et al. (2011)

*All activities include spreadsheet simulations of random error plus more.

Simulations allow for a wider range to be numerically investigated. For a discussion of introducing error analysis, see Sinex (2005) for defining error types, and for the pedagogy see Sinex (2013). For the die-hards, see Taylor (1997) for error analysis to the max.

Dealing with Real-world Messy Big Data

Let's examine some real-world messy big data for New York City's The Battery tide gauge on the southern tip of Manhattan Island (green arrow on map) given in Figure 3. Notice a good linear fit for 149 data points and the scatter. The slope of the regression line yields the rate of relative change in sea level, which is rising at +2.91 mm/yr or 29.1 cm/century. For a thorough discussion of using PSMSL tide gauge data to determine relative sea level change, see Sinex (2022a). This is a nice transition from laboratory to real-world data.

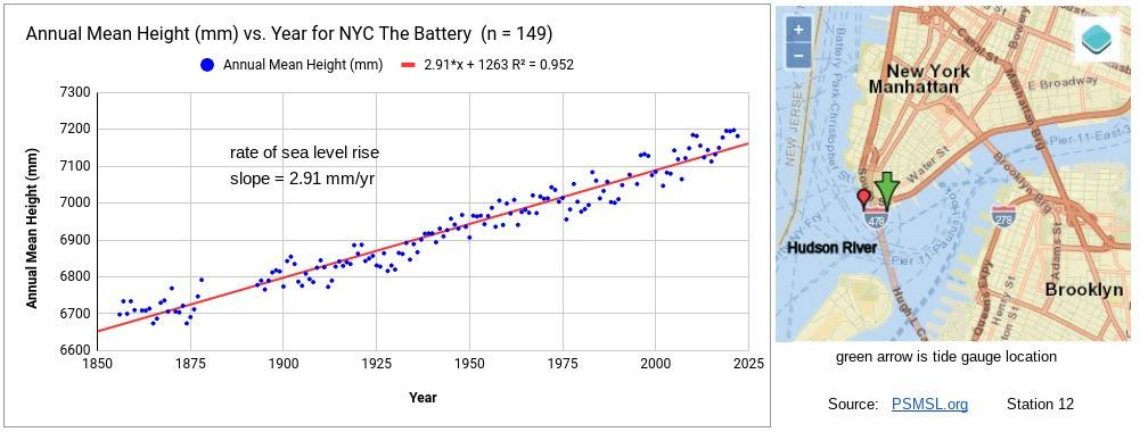

Figure 3 - NYC's The Battery Tide Gauge with a missing data gap (1879-1892)

Now let's examine a large data set with considerable scatter to see if we can extract some useful information from it. If you ever wondered why you can see the new moon phase, then you have observed Earthshine, sunlight reflected off the Earth back to the Moon. The reflected light from Earth is the Earth's albedo. Goode et al. (2022) present a summary of twenty years of Earthshine measurements.

Figure 4 plots the Goode et al. data with 20-40 days of measurements per year. The top graph shows the 550 measurements taken over 1998-2017. The prominent gap of missing data was the result of renovation of the telescope. Ask students, "Is there a trend in the data?" The middle graph has a linear regression line added to the data. The r-square may be significant with 550 data points; however, the fit is not very convincing.

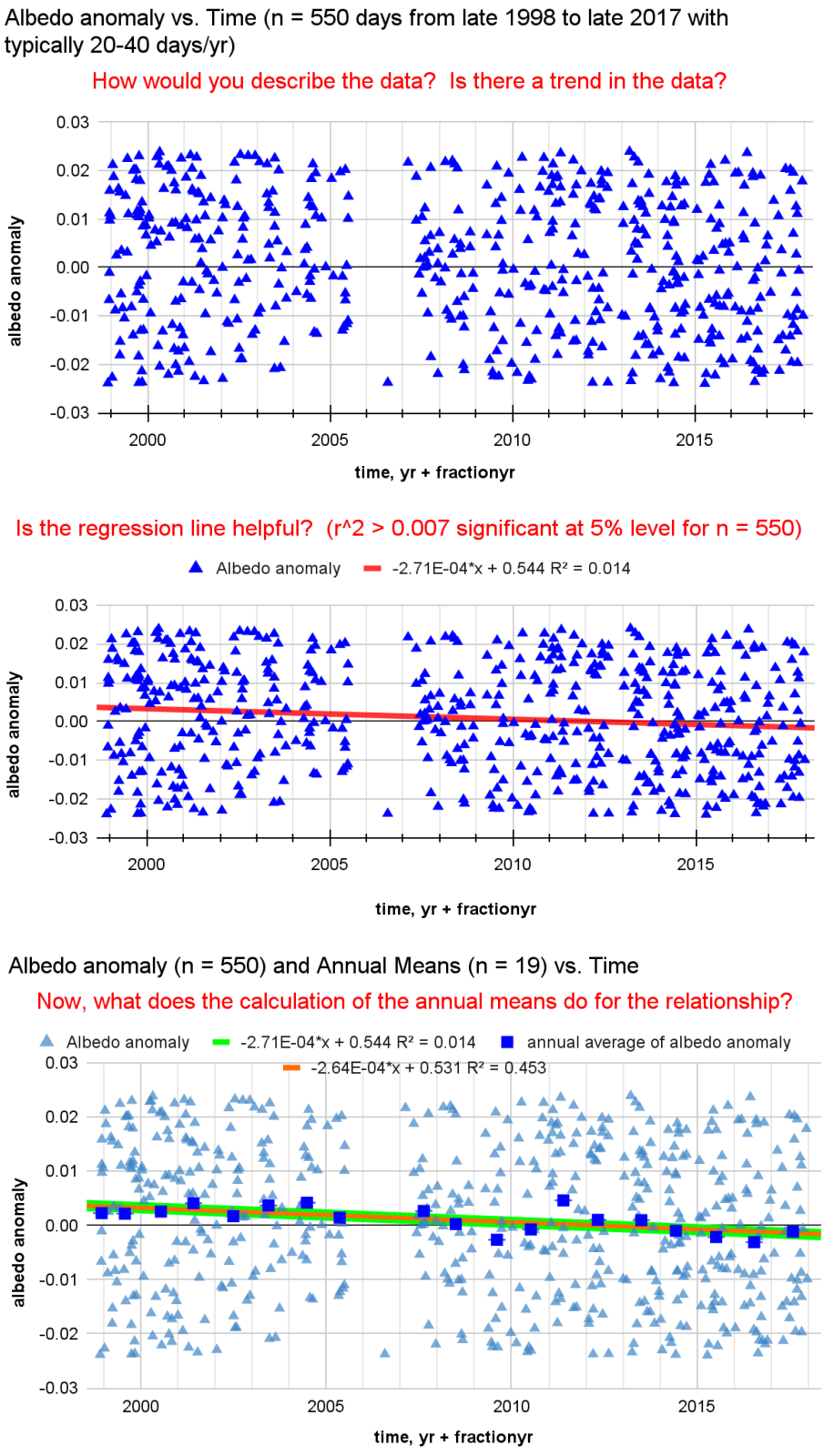

Figure 4 - Earthshine (albedo anomaly) data from Goode et al. (2022)

On the bottom graph on Figure 4, the data for each year (20-40 days) has been averaged to produce an annual mean. These data points are overlain on the daily points and fit with a linear regression line, which for 19 data points has a much higher r-square value and is significant at the 5% level. The data set shows that the Earth's albedo has decreased over the time period measured by Goode et al. (2022). The NASA CERES data confirms the albedo decrease.

The data sets for mean sea level determined by satellite altimetry (Sinex 2022b) and the data describing the causes of global sea level rise (Sinex 2023) are available in Google Sheets spreadsheets at Sea Level Change: Real-world Data Analysis and Mathematical Modeling Spreadsheet Projects.

Here are some links to data sets with large scatter (Table 3) that would be effects from climate change. They are all Google Sheets spreadsheets in various stages of development.

Table 3 - Big Data with Large Scatter

Topic with link	Source*	Comments
Earthshine	Goode et al. (2022)	Discussed above
Cherry Blossoms - Kyoto, Japan	Professor Aono at Osaka Prefecture University	Very large data set back to 9th century
Cherry Blossoms - Washington, DC	EPA	Large data set 1921-2022 for Tidal Basin
NYC Central Park Weather	NOAA NWS	Weather vs. climate - helps with distinction
US Wildfires	EPA	30 year record with large scatter
Permafrost in Alaska	EPA	40+ yrs of temperature data for 15 sites (New GSheets interactive map)
Examining the Global Surface Temperatures: What is in the pipeline for the future?	James Hansen at Columbia University	Hansen introduces the concept of hinge points in temperature records. This is an easy method to locate them.

*full references/links to original sources are included on each spreadsheet

Transforming the Function Machine for the Real-world

In mathematics, the function machine takes a number with standard deviation of zero and yields another number with standard deviation of zero. In science, the numbers are measurements with errors (standard deviations greater than zero), plus there are usually other variables with errors in the mix as well. We need a more real-world approach by being multivariable as well!

Can we adapt the function machine to handle real-world data and consider how it would be used in the sciences? The infographic below (Figure 5) illustrates and summarizes how this can be accomplished. In science, trying to find the function via mathematical modeling is slightly more involved than finding the rule. Plus, in science, we are typically performing measurements of the input and output, which are both subject to measurement error. Also, there may be other variables involved; hence, a multivariable function. Spreadsheet simulations can easily allow exploration and are glass-box simulations, all computations are visible! With little effort, students could be building spreadsheet simulations.

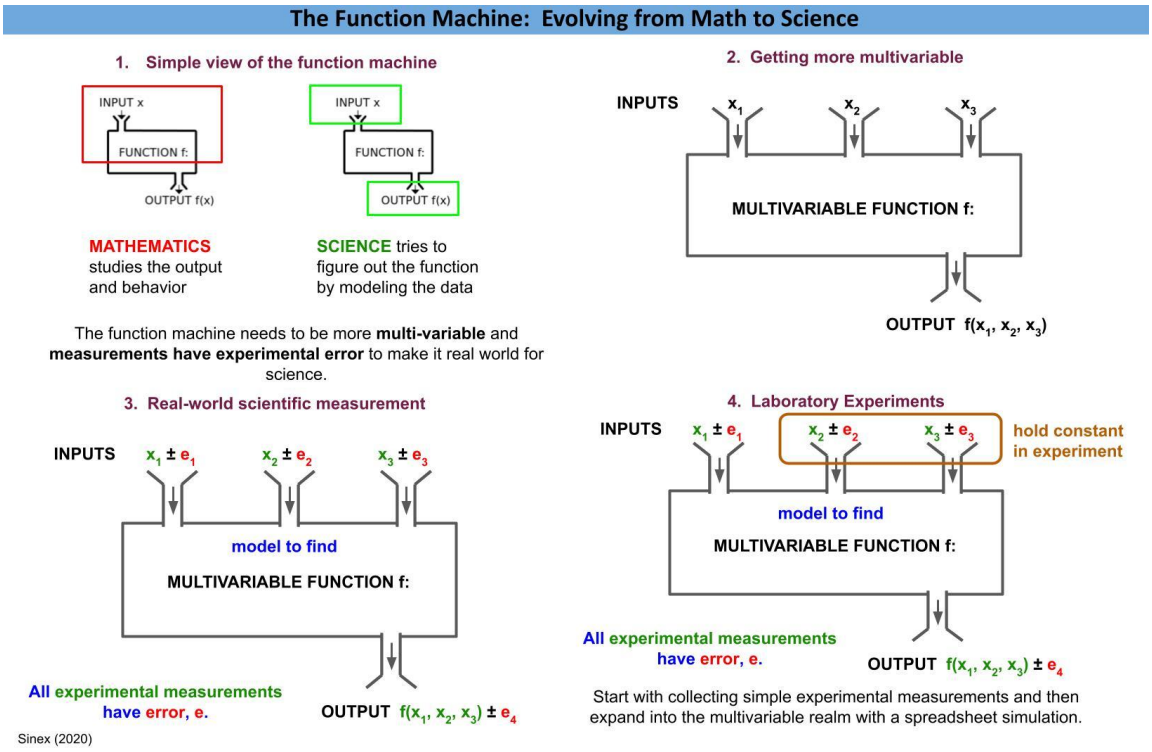

Figure 5 - The real-world function machine

Some Final Thoughts

Should we be teaching with big data sets that contain a large amount of scatter? Absolutely YES; however, we need to build up to this point. The investigative approach found in science, statistics, and mathematics, where we start with a question to investigate (collect data > develop mathematical model) can be extended by adding a more multivariable approach with error analysis that includes data with scatter. Spreadsheet simulations are an excellent tool to investigate the influence of other variables and errors. The modified function machine presented here is a great tool to use to get students to develop this more real-world approach. A really great approach, What-if spreadsheet math, uses numerical experimentation via spreadsheet simulations to let students play (explore)!

The science and mathematics communities need to be vigilant about how data science infringes upon the two disciplines. See editorials by Dorsey, C (2021) and Jiang, S et al. (2022) for background. Integrating data science into the science, statistics, and mathematics realm needs serious discussion by **all** the disciplines. Developing a clear definition of messy data would be a start.

Creating a "investifest" or "free for all" of multivariable, mixed qualitative and quantitative types, and no scaffolding is NOT the way to go, and does not consider causation. Starting with data and generating questions is backwards science, and large multicolumn data sets could easily cause cognitive overload. It is all too easy to create complex visualizations that novices are not ready to explore. Let's agree on the basics and how to integrate this into higher education.

If you want to generate your own big data set and examine variation by determining pi and pooling data, see Sinex and Chambers (2018). We explore accuracy, precision, and a little about quality control in Google Sheets, the number one choice of teachers (Rosenberg et al., 2022)!

Acknowledgements

The author wishes to thank Catherine Mejia Sinex, retired mathematics teacher, for providing invaluable comments on this paper.

References

Dorsey, C (2021) Teaching in a World of Messy Data, **The Science Teacher 88** (5), 8.

Goode, PR et al. (2022) Earth's Albedo 1998–2017 as Measured From Earthshine, **Geophysical Research Letters 48**, e2021GL094888.

Hunter-Thomson, K (2022) Why is Variability Worth the Teaching Challenge? (Data Literacy 101) **Science Scope 45** (3), 8-13.

Jiang, S et al. (2022) Data science education across the disciplines: Underexamined opportunities for K-12 innovation, **J. Educ. Technol. 53** (5), 1073-1079.

O'Reilly CM et al. (2017) Using Large Data Sets for Open Ended Inquiry in Undergraduate Science Classrooms, **BioScience 67** (12), 1052–1061.

O'Reilly CM et al. (2022) Pedagogy of teaching with large datasets: Designing and implementing effective data-based activities. **Biochem. Mol. Biol. Educ. 50** (5), 466–472**.**

Rosenberg, JM et al. (2022) Big data, big changes? The technologies and sources of data used in science classrooms, **J. Educ. Technol. 53** (5), 1179-1201.

Schultheis, EH & Kjelvik, MK (2020) Using Messy, Authentic Data to Promote Data Literacy and Reveal The Nature Of Science, **The American Biology Teacher 82** (7), 439–446.

Sinex, SA (2005) Investigating Types of Errors, **Spreadsheets in Education 2** (1) 115-124.

Sinex, SA, Gage, BA, and Beck PJ (2007) Exploring Measurement Error with Cookies: A Real and Virtual Approach via Interactive Excel, **The AMATYC Review 29** (1) 46-53. (Excelet)
(If you want to experimentally explore random and systematic errors, it's cookies and poker chips plus spreadsheet simulations.)

Sinex, SA, Chambers, TJ, and Halpern, JB (2011) Mass, Measurement, Materials, and Mathematical Modeling: The Nuts and Bolts of Extrapolation, **Spreadsheets in Education 5** (1), Article 3.

Sinex, SA (2012) Cookies, Spreadsheets, and Modeling: Dynamic, Interactive, Visual Science and Math at Network Connections Conference in Pittsburgh, PA. (accessed May 2023)

Sinex, SA (2013) Investigating and Visualizing Measurement Error for Novice "STEM" Learners in *Proceedings of the Spring 2013 Mid-Atlantic Section Conference of the*

American Society of Engineering Education, pp. 222-234. (discusses using four simple manipulative experiments)

Sinex, SA (2015) Deriving a Non-linear Multivariable Model for Stacking Nested Cubes via Spreadsheet Simulations in *Electronic Proceedings of the ICTCM Conference*, Vol. **26**, pp. 298-307.

Sinex, SA (2016) The Mechanics of Inducing Error in a Spreadsheet (accessed May 2023)

Sinex, SA (2017) Multivariable Spreadsheet Modeling and Scientific Thinking via Stacking Bricks in *Electronic Proceedings of the ICTCM Conference*, Vol. **28**, pp. 381-390.

Sinex, SA (2018) Algebraic and Scientific Thinking via Spreadsheets: The Unstacking Coins Model in *Electronic Proceedings of the ICTCM Conference*, Vol. **29**, 14pp.

Sinex, SA and Chambers, TL (2018) Discovering Pi and its Measurement Variation: A Collaborative Cloud Activity, **Spreadsheets in Education 10** (3), Article 4.

Sinex, SA and Chambers, TL (2019) Volumetric Measurement of Tumors: Mathematical Models, Assumptions, and Errors in *Electronic Proceedings of the ICTCM Conference*, Vol. **30**, 13pp.

Sinex, SA (2022a) Relative Sea Level Change: A Global Problem in Your Backyard in *Electronic Proceedings of the ICTCM Conference*, Vol. **33**, 8pp.

Sinex, SA (2022b) Global Determination of Sea Level Height via Satellite Altimetry in *Electronic Proceedings of the ICTCM Conference*, Vol. **33**, 11pp.

Sinex, SA (2023) Exploring the Data for Delineating the Causes of Global Sea Level Rise in *Electronic Proceedings of the ICTCM Conference*, Vol. **34**, 20pp.

Taylor, JR (1997) **Introduction to Error Analysis**, 2nd ed., University Science Books, 327pp.

Classroom Resources

The Function Machine: Evolving from Math to Science Infographic - click here for pdf (accessed May 2023)

Rulers and Measurement Error
https://academic.pgcc.edu/~ssinex/excelets/Ruler_error.xls (accessed May 2023)

Project EDDIE (Environmental Data-Driven Inquiry and Exploration)
https://serc.carleton.edu/eddie (accessed May 2023)

Common Online Data Analysis Platform (CODAP) includes data sets
https://codap.concord.org/ (accessed May 2023)

Data Nuggets (bring real research and data into the classroom)
http://datanuggets.org/ (accessed May 2023)

Dealing with Scientific Data in Google Sheets: Data > Model > Simulation
https://sites.google.com/view/ssinex/home/dealing-with-data-in-gsheets (accessed May 2023)

Sea Level Change (and its causes from other global problems): Real-world Data Analysis and Mathematical Modeling Spreadsheet Projects
https://sites.google.com/view/ssinex/home/sea-level-change (accessed May 2023)

Data Pooling and On-line Collaboration
https://sites.google.com/site/datapoolinthecloud (accessed May 2023)