Why data quality matters

now more than ever

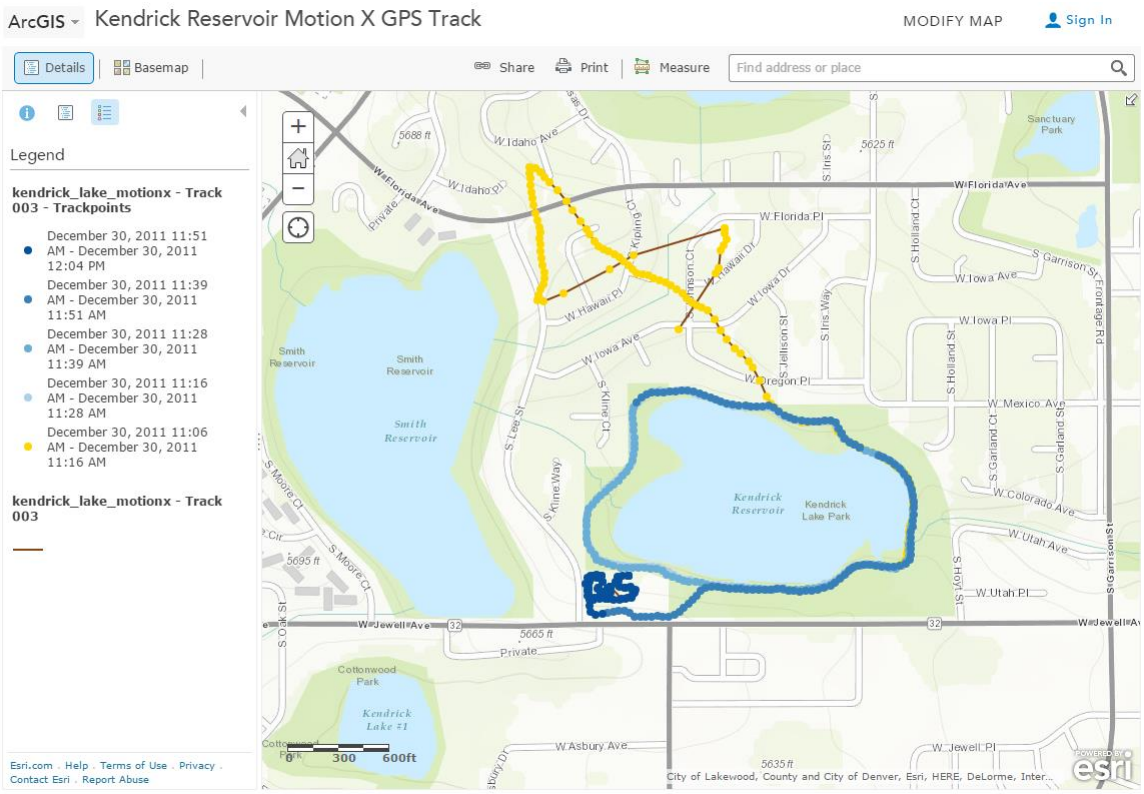
August 19, 2015
Share
by Joseph J. Kerski

Not long ago, obtaining data for a GIS-based project was an arduous task. Because great time and effort was involved with either creating your own data or obtaining data that someone else created, you had to think carefully about the quality of the data that would go into your project. While it can still be cumbersome to obtain data at specific scales for specific areas, cloud-based data services, crowdsourced maps and databases and real-time streaming make it easy for anyone to obtain vast amounts of data in a short amount of time. In an environment where so much data is available, is data quality still of concern? I believe that yes, data quality does matter. In fact, because it is so easy to obtain data nowadays, and with the advent of crowdsourcing and cloud-based GIS, I submit that data quality considerations actually matter now more than ever. Consider the following three examples that focus on criticizing, analyzing and scaling your data.

Be critical of data — even when it's your own!

Thanks to mobile technologies, anyone can create spatial data, even from a smartphone, and upload it into the GIS cloud for anyone to use. This has led to incredibly useful collaborations such as OpenStreetMap, but this ease of data creation means that caution must be employed more than ever before.

For example, let's look at a map that I created using MotionX-GPS on an iPhone and mapped using ArcGIS Online,that follows my track around Kendrick Reservoir in Colorado. This map was symbolized at the time of GPS collection, from yellow to gradually darker blue dots as time passed.



Note the components of the track to the northwest of the reservoir. These pieces were generated when the smartphone was just turned on and the track first began, indicated by their yellow color. These segments and track points cut across the terrain, not following city streets or sidewalks — erroneously. Change the base map to a satellite image. Cutting across lots would not have been possible on foot given obstructing fences and houses. When I first turned on the smartphone, not many GPS satellites were in view, but as I kept walking, the phone recorded a greater number of GPS satellites, and as the number of satellites increased and an increasing number of Wi-Fi hotspots and cell phone towers could be triangulated against, the positional accuracy improved until the track points more closely represented my true position.

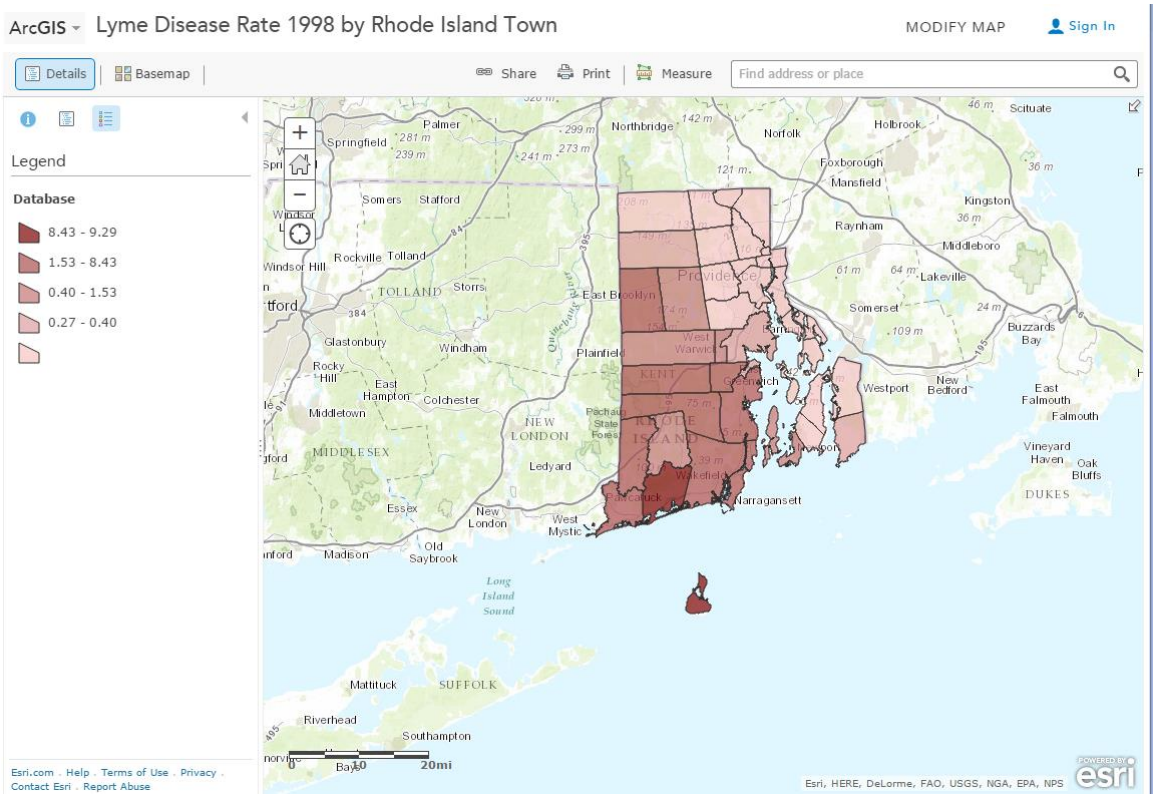
Use the distance tool in ArcGIS Online to answer the following: How far were the farthest erroneous points from the lake? Although it depends on where you begin to measure, some of the farthest erroneous points were 600 meters from the lake. By selecting each dot to access the date and time each track point was collected, it can be

determined that the error lasted about 10 minutes. Using the same selection tool, the time when the application correctly began to follow my walk around the lake can be determined as 11:12 a.m.

This simple example points to the serious consequences that could result from using data without being critical of its source, spatial accuracy, precision, lineage, date, collection scale, methods of collection and other considerations. Therefore, be critical of the data — even when it's your own!

Misleading data: Mapping Lyme disease

My colleague, Lyn Malone, and I have taught workshops using Lyme disease case counts from 1992 to 1998 by town in the state of Rhode Island. Most recently, we started with an Excel spreadsheet and used Esri Maps for Office to map and publish the data to ArcGIS Online. The results are here.



Rhode Island Towns Lyme Disease Rate – 1998.

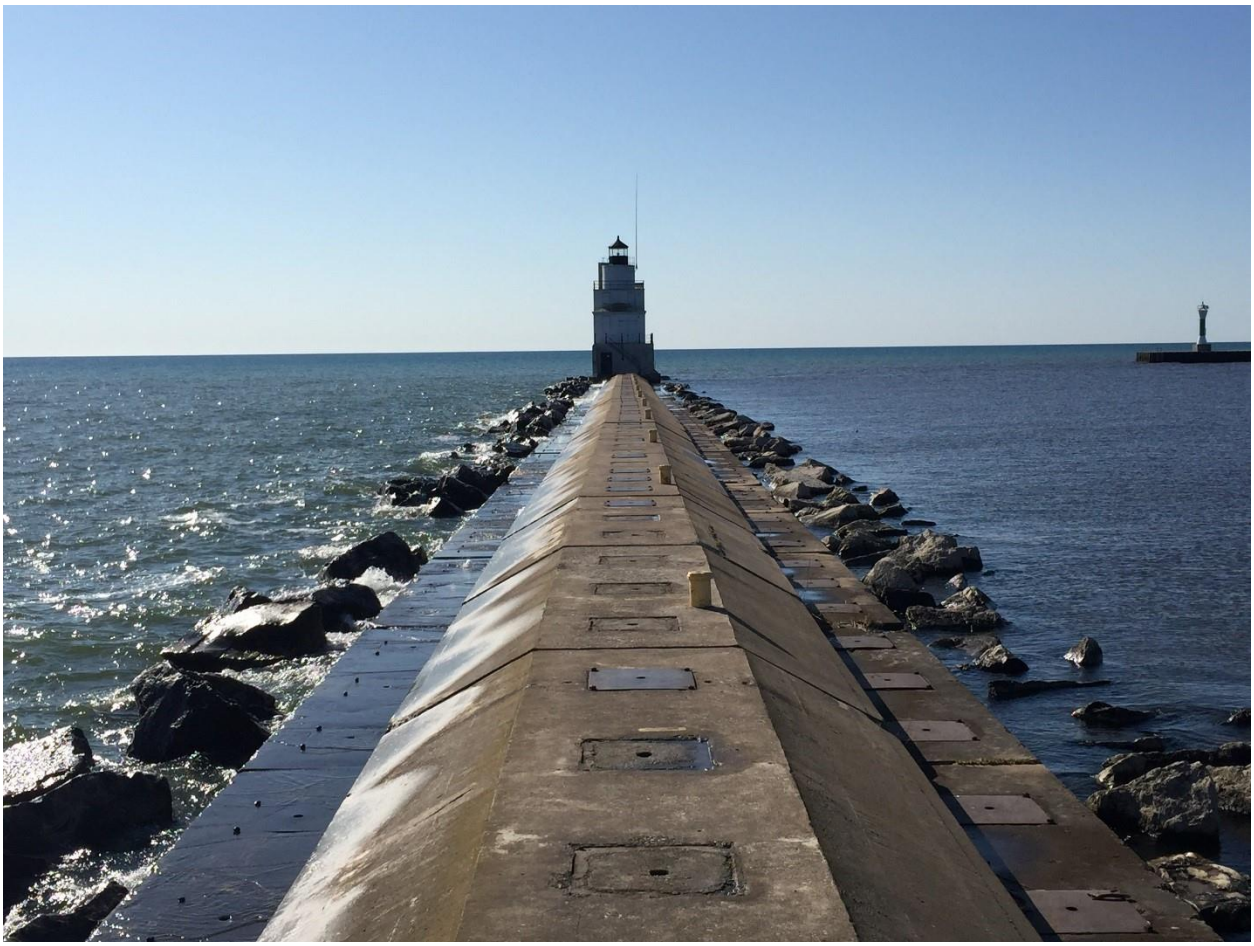
After one of the workshops, we sought to update the data with information from 1999 to the present, so we contacted the Rhode Island Department of Health. They not only provided the data, they also provided valuable information about the data. The Public Health staff told us that Lyme disease surveillance is time and resource intensive. During the 1980s and 1990s, as funding and human resource capacity allowed, the state ramped up surveillance activities — including robust outreach to healthcare providers. Prioritizing Lyme surveillance allowed the state to obtain detailed clinical information for a large number of cases and classify them appropriately. The decrease observed in the 2004-2005 case counts was due to personnel changes and a shift in strategy for Lyme surveillance. Resource and priority changes reduced their active provider follow up. As a result, in the years since 2004, the state has been reporting fewer cases than in the past. They believe this decrease in cases is a result of changes to surveillance activities and not to a change in the incidence of disease in Rhode Island.

If this isn't the perfect example of "know your data", I don't know what is. If one didn't know that surveillance activities had changed, an erroneous conclusion about the spatial and temporal patterns of Lyme disease would surely have occurred — and often, this kind of information doesn't make it into standard metadata forms. This is also a reminder that contacting the data provider is often the most helpful way of obtaining the inside scoop on how the data was gathered, even though it sounds "so 20th century". And you can bet that we made sure this information was included in the metadata when we served this updated information.

Walking on water? Reflections on resolution and scale

I once gave presentations at the University of Wisconsin Milwaukee for GIS Day, and took the opportunity to get out onto the landscape. I walked on the Lake Michigan pier at Manitowoc, enjoying a stroll in the brisk wind to and from the lighthouse there, recording my track on my smartphone in an application called Runkeeper. When my track ended and was mapped, it appeared as though I had been walking on the water! Funny, but I don't recall even getting wet!

Walking on water? This is how Runkeeper mapped my track.



My view of Lake Michigan as I walked toward the lighthouse at Manitowoc.

It all comes down to paying close attention to your data and knowing its sources, which leads us to a larger discussion on the importance of scale and resolution in any project involving maps or GIS. In my case, even if I scrolled in to a larger scale, the pier did not appear on the Runkeeper application's base map. It does, however, appear on the base map in ArcGIS Online.

Most of the GIS literature understandably focuses on the success stories, but if you dig a bit, you can find examples where neglecting these important concepts have led not only to bad decisions, but have cost people their property and sometimes, even their lives. Today, while GIS tools allow us to instantly zoom to a very large scale, the data that you are examining might have been collected at a much

smaller scale. Remember, if you are making decisions at 1:10,000 scale and your base data was collected at 1:50,000 scale, you're treading on dangerous ground, or, one could say, you are “walking on water”!

Final thoughts

With great opportunity comes great responsibility. As never before, we have a vast array of data at our fingertips, with powerful and easy-to-use tools and models with which to analyze it. Don’t get lulled into complacency and use a map or data set just because it's so easily accessible or because the symbology looks sharp. Be sure to be critical of the data.

Remember that with the ability to publish your data in the cloud, embed your Web maps in Web pages, or build communication tools such as Storymaps around your data, thousands or millions of people could be looking at your maps. By checking your data sources, your map is more likely to be on firmer scientific ground and you are more likely to reduce any possible misinterpretation of your data.

Finally, practice what you preach about metadata. We all breathe a sigh of relief when the data we are seeking is well populated with metadata. But when you are publishing your own data, are you providing metadata so that others will breathe the same sigh of relief?

For more information on the topic of data quality and related data issues, see the book that Jill Clark and I wrote, entitled *The GIS Guide to Public Domain Data*, and the blog that we update weekly, Spatial Reserves.