Data Quality Assessment for Northampton Pub Fest

Introduction

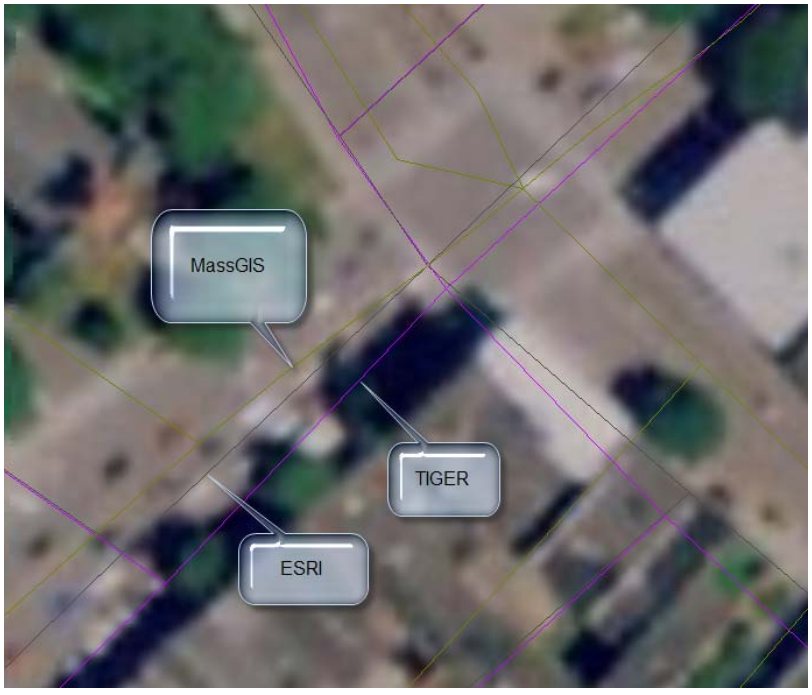
As a member of the Economic Development Council in the city of Northampton, Massachusetts, I have an idea to hold a Northampton Pub Festival to celebrate the many fine and quirky drinking establishments that are located in downtown "Noho". While I expect that locals will turn out for this event, my hope is to attract tourists to Northampton, some of whom may not be very familiar with the locations and characteristics of the city's bars. I have decided that a map that could be posted online and printed on flyers would be useful to guide visitors, and I use any chance I can get to play with ArcGIS. I want my map to allow users to find where the bars are in town.

Purpose

This memo is an analysis of the data that is available for this mapping project, how accurate the data is, and what uses are appropriate for the data.

Road centerline data

Road centerline data was obtained from MassGIS, the Census TIGER data set, and the StreetMap USA data set published by ESRI. When we look at the intersection of Bridge Street and Pleasant St., we can see the three roadlines do not agree spatially. Roadlines converge fairly well at the intersection, but to the southwest we see the road centerlines diverge. The light green MassGIS line is the closest to the orthophoto, while the gray ESRI line is 25 feet away, and the magenta TIGER line is another 25 feet away. For this project, maps will be displayed at the neighborhood level, so I would want to use MassGIS road centerlines. Unfortunately, the MassGIS road lines do not contain address ranges, so it cannot be used for geocoding.



According to its metadata, the MassGIS roads layer is based on tracing orthophotos at a 1:5,000 scale. According to the National Map Accuracy Standard, paper maps of this scale can be interpreted as

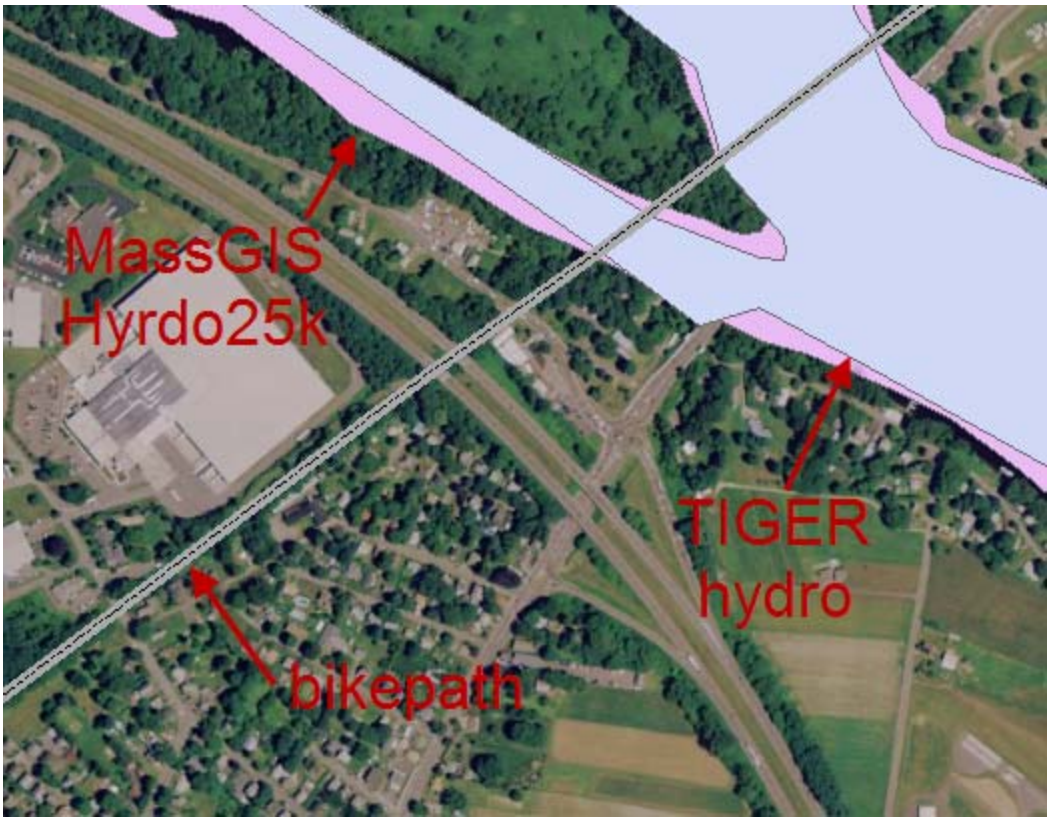
having an accuracy of about +/- 15 feet. Since the MassGIS linework was made directly from orthophoto images, +/- 15 feet is a good estimate of its accuracy.

The Topologically Integrated Geographic Encoding and Referencing System (TIGER) linework published in 2000 is shown here as magenta lines. The TIGER linework is very useful dataset that includes address ranges for the entire country. Unfortunately, this linework was produced at a small scale, and the spatial uncertainty is fairly large. TIGER was made using USGS Digital Line Graphs and 1:100,000 scale topographic maps. According to the US National Map Accuracy Standard, we should expect an uncertainty to these lines of at least +/- 167 feet. Areas that were traced from digital or paper maps would have an even lower accuracy. We would not want to display the TIGER lines, but they are still useful to us because they allow us to geocode data from a variety of sources.

ESRI publishes a version of the TIGER streets that, according to the metadata provided by ESRI, has been enhanced for improved accuracy and use for performing network analysis. Although StreetMap USA is privately produced, and so cost the city some money, it includes the geocoding potential that the TIGER lines have, while also being updated more often than the TIGER – the last time in 2002. ESRI recommends that StreetMap USA be displayed at scales no larger than 1:50,000. According to the US National Map Accuracy Standards, this would correspond to an accuracy of +/- 83 feet.

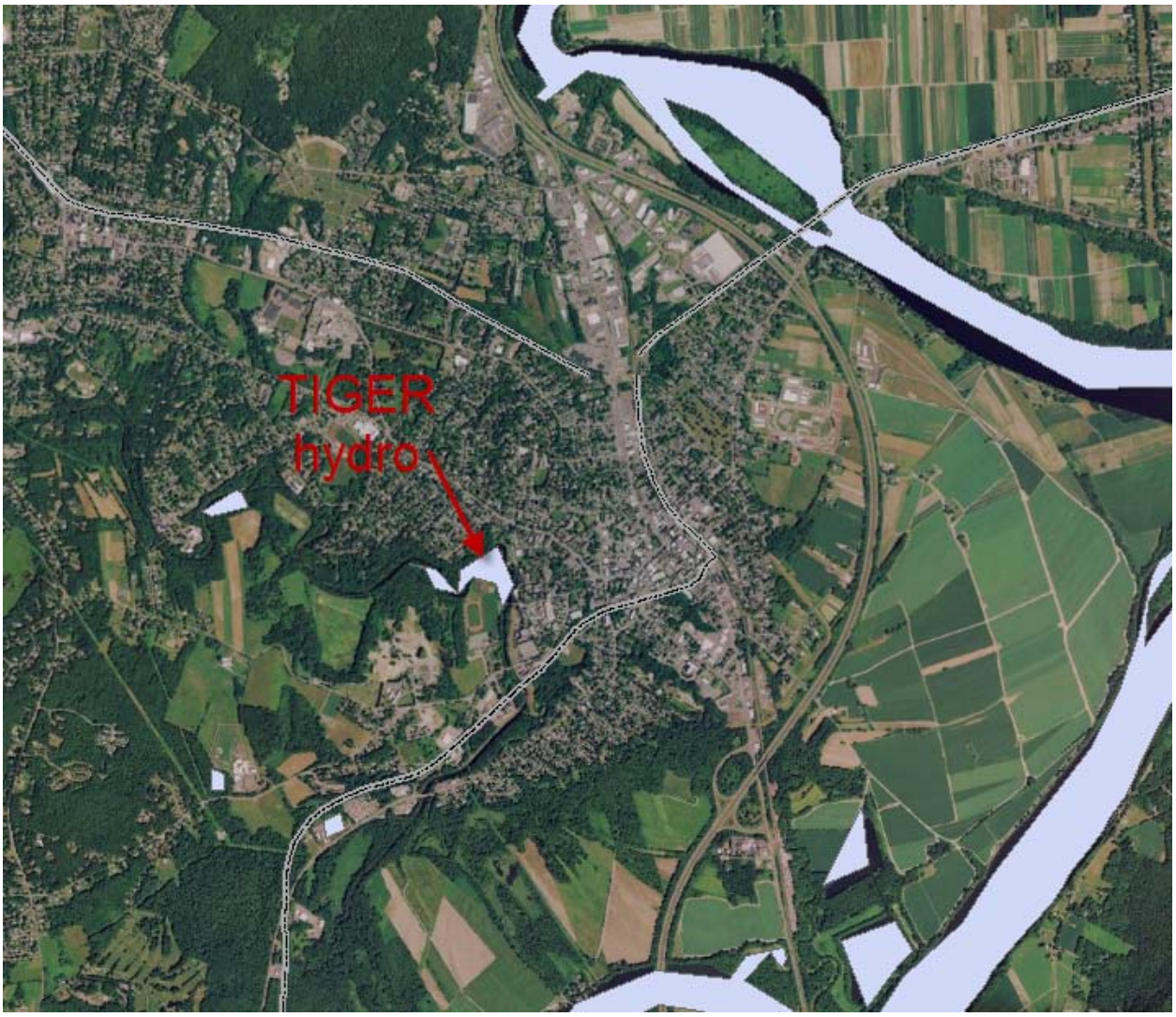
Hydrography data

Since this is a project showing local "watering holes", we also want to display the areas streams, rivers, ponds, and lakes. Beyond showing potential visitors places where they can cool off in the sticky summer, water features are valuable references, especially in an area defined by the valley carved by the Connecticut

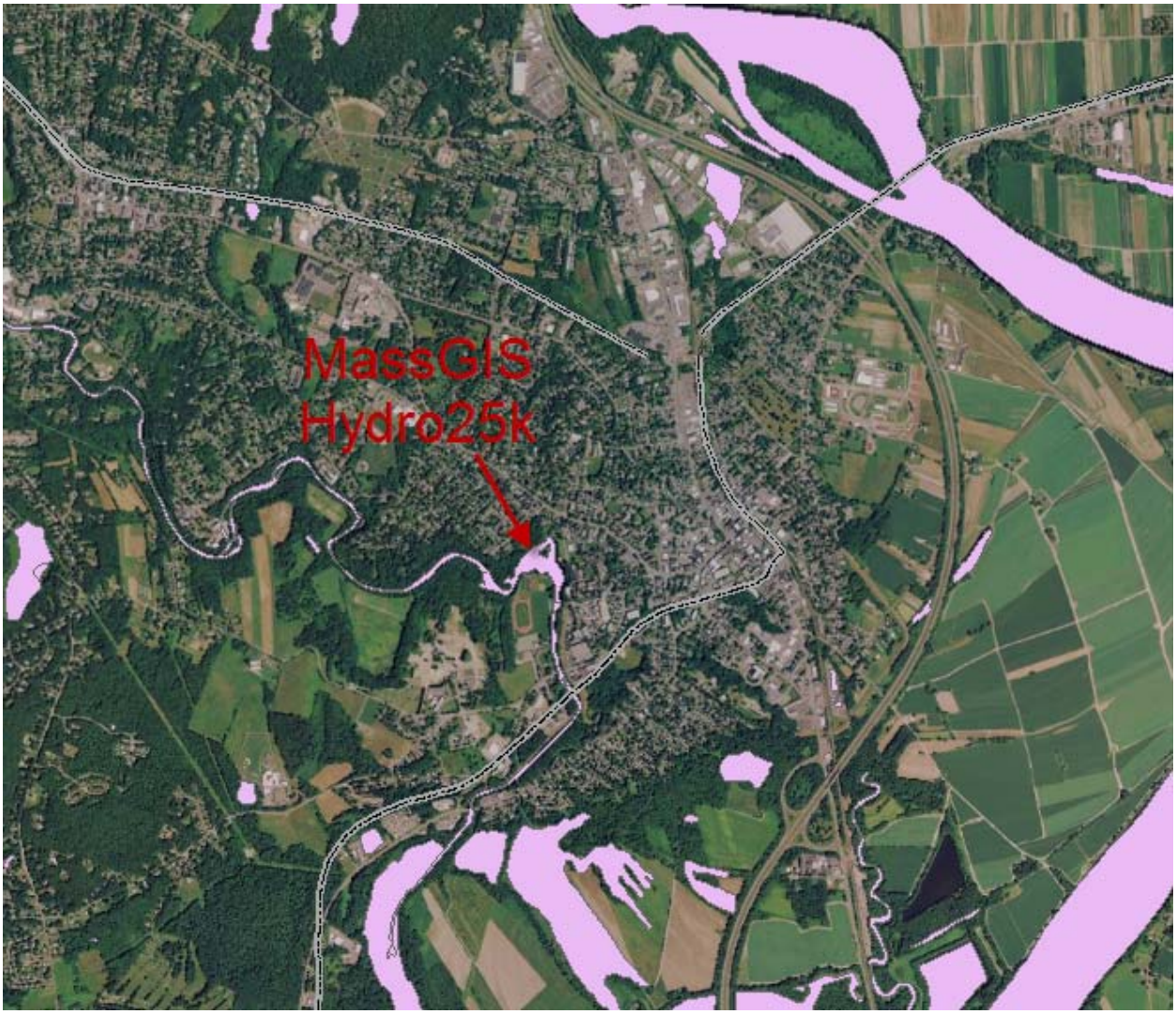


River. While we do not expect map users to need exact measurements of hydro features on their maps, we do want to present as attractive a map as possible. This will mean that we want hydro that is drawn at a large enough scale that a low resolution will make features look unrealistically "blocky".

On the above map, I am comparing the Census TIGER hydro layer with the Hydro25k layer produced by MassGIS. Both of these layers show the Connecticut River at the road crossing via the Coolidge Bridge and the Norwuttuck Rail Trail crossing. If a map showing different ways to approach town was desired at the scale shown (about 1:8,000), we would certainly prefer the MassGIS layer. Even zoomed out to about 1:35,000 (see below) we see that small streams are not captured by the TIGER hydro layer, which like the roads, was mapped from 1:100,000 maps.



The MassGIS Hydro25k layer (shown below) gives a much more accurate and visually pleasing picture of the water features around Northampton. We likely would want to eliminate certain classes of water features (such as marshes and wetlands) for the purposes of this map.


MassGIS
Hydro25k

Biking into Downtown Northampton

One map we want to provide for visitors is a guide to biking into downtown Northampton. The biketrails.arc layer shown above is available from MassGIS and a "bikeways" layer is available from the city of Northampton.

While the MassGIS biketrails layer shows only the bicycle and pedestrian path that runs through town, Northampton has mapped all roads that it has deemed as conducive to bike riding. It would seem that the NohoGIS bikeways layer is more accurate than the MassGIS layer. The screenshot above shows that the MassGIS biketrails layer seems to run down the active railroad tracks, while the actual path runs parallel nearby. Unfortunately, Northampton has not published any metadata to accompany its bikeways layer, so it is difficult to discuss its accuracy in a quantitative way. Also appreciated would be an explanation of what the criteria was for distinguishing a bikeway and when the bikeways were mapped. The MassGIS biketrails was mapped at a 1:50,000 scale in 2002. This layer is spatially accurate to about +/- 83 feet, and so displaying data at a neighborhood scale would not be advised. Since the metadata is missing for the Northampton biketrails, some ground-truthing would be required to finalize a bicycle access map.

Location of Northampton Bars

A list of bars (NAICS number 722410) in Northampton was obtained from the Reference USA Business Database. Inclusive of lounges, taverns, and other "drinking places", the data is compiled by a private company (referenceGROUP) that sells the information to users such as us. Although the data is expensive, the referenceGROUP provides a significant value by fact-checking business and residential databases. Unfortunately, Reference USA does not attach metadata to downloaded data, so we do not know when it was last checked.

We want to view this bar locations in our GIS, so I have geocoded the addresses given by Reference USA using the TIGER line files and the ESRI StreetMap USA line files. Also, bars are mapped using the longitude-latitude information given by Reference USA. The results showing a close-up of downtown are below:

Given my analysis of road locations shown above, I am not surprised to find that geocoding resulted in different locations for the bars. The green triangle showing the location of Pearl Street night club

according to the TIGER line data is approximately 40 feet from the location given by the StreetMap USA data. These geocoded locations are each about 60 feet away from the location for the Pearl Street Night Club given by the Lat-Long from Reference USA. This deviation in geocoded addresses is as good as we could expect from line files that themselves have uncertainties of 83 or 167 feet. For other bars, the geocoded and lat-long derived points cluster near the establishment, but Pearl Street presents a more difficult case for geocoding. The average difference between the actual location of Pearl Street and the calculated points is about 240 feet. The business is labeled as 10 Pearl Street, which is the correct address. Unfortunately, both TIGER and StreetMaps USA have an address range from 1-99 listed for this street. The Lat-Long point is located nearer to the erroneous geocoded points than to the actual location, suggesting that the Lat-Long data was created by Reference USA using a similar geocoding technique. Again, metadata on the Lat-Long information from Reference USA would be useful. On one hand we could modify the line files to reflect the correct 1-10 address range and get a more accurate result, but perhaps it is better to think of how to use geocoded data (and GIS) in a more appropriate manner.

A More Appropriate Use of this Data

Let's say that instead of using GIS and geocoding as a way to draw my map, I instead realize that GIS is more primarily an analysis tool. Many other graphics programs could be used to touch up aerial photographs showing the locations of the bars in Northampton. If instead I use GIS to help me figure out how many bars are within ¼-mile of the bike path, then it becomes a tool not just for creating graphics, but helping me as a town official for making better decisions on which bars to include in the event. By zooming out to about 1:14,000 I can view a map that is much better suited to the accuracy of the data I have.

At this scale I no longer care about the exact location of the bike trail, because my criteria of ¼-mile is a somewhat arbitrary figure and will not be very responsive to +/- 250 foot errors in my geocoded points. Regardless of whether Club Metro or the bike trail are in the exact right locations, I know that the bar is located more than ¼-mile from the bike trail, and therefore will not be included in the event.