Intermediate Spatial Data Science Lab Report

Title: Lab 3 - Part 2: Temperature Interpolation
Notice: Dr. Bryan Runck
Author: Andrew Arlt
Date: 11/26/2024

Project Repository: [andrewarlt/GIS5571/Lab3/]
Google Drive Link: n/a
Time Spent: 5 hours

Abstract

Interpolation methods can affect the results and appearance of the raster outputs. Temperature data should be thought of as continuous data sets between sample locations (stations). Methods like kriging use more complex relationships and autocorrelation, while spline is more simplistic and uses smooth best-fit curves. This project aims to examine three different interpolation methods on their ability to predict temperature values between the NDAWN stations using 30-day temperature means.

Problem Statement

The problem aimed to compare three different interpolation methods (IDW, kriging, and spline) using a 30-day collection of NDAWN temperature data from all stations. This function requires a few steps:

1. Build a fully functional real-time data visualization and analysis workflow
2. Compare and contrast three types of interpolation (IDW, kriging, and other → spline)

Maps for the three interpolation methods are shared within the results section and a comparison is described in the discussion and conclusions section.

Figure 1. Table showing the major steps applied to the data layers.

#	Requirement	Defined As	(Spatial) Data	Attribute Data	Dataset	Preparation
1	**NDAWN API**	Download data via an API	.CSV Files (temp data)	Station data (lat, long) [station name, date, temp, etc.]	NDAWN to “30-Day.csv”	API request for .csv file
2	**30-Day.csv**	Create DF, modify/ remove empty data, calculate means	.CSV	Station data (lat, long) [station name, date, temp, etc.], mean temp	to “Mean 30-Day.df”	Use .csv to make df, remove extras, calculate means
3	**Mean 30-Day.df**	Convert to SEDF and Reproject, convert to shapefile	.DF	Station data (lat, long) [station name, date, temp, etc.], mean temp	to “Mean 30-Day.sedf”--> “Mean 30-Day.shp”	Convert df to sedf, reproject to 4326, convert to shapefile
4	**Mean 30-Day.shp**	Interpolate using IDW: Power 2, Variable, 12 Points	Interpolation Layer IDW .tif	Station Data, Interpolation Values	to “30-Day IDW.tif”	Interpolation using IDW parameters
5	**Mean 30-Day.shp**	Interpolate using Kriging: Ordinary, Gaussian, Lag 0.021011, 12 points	Interpolation Layer Kriging.tif	Station Data, Interpolation Values	to “30-Day Kriging.tif”	Interpolation using Kriging parameters
6	**Mean 30-Day.shp**	Interpolate using IDW: Regular, Weight 0.1, 12 Pts	Interpolation Layer Spline .tif	Station Data, Interpolation Values	to “30-Day Spline.tif”	Interpolation using Spline parameters

Input Data

Data was obtained through API interactions with the NDAWN server. NDAWN data was requested for all NDAWN stations for the past 30-days, using the maximum temperature, minimum temperature, and average temperature values.

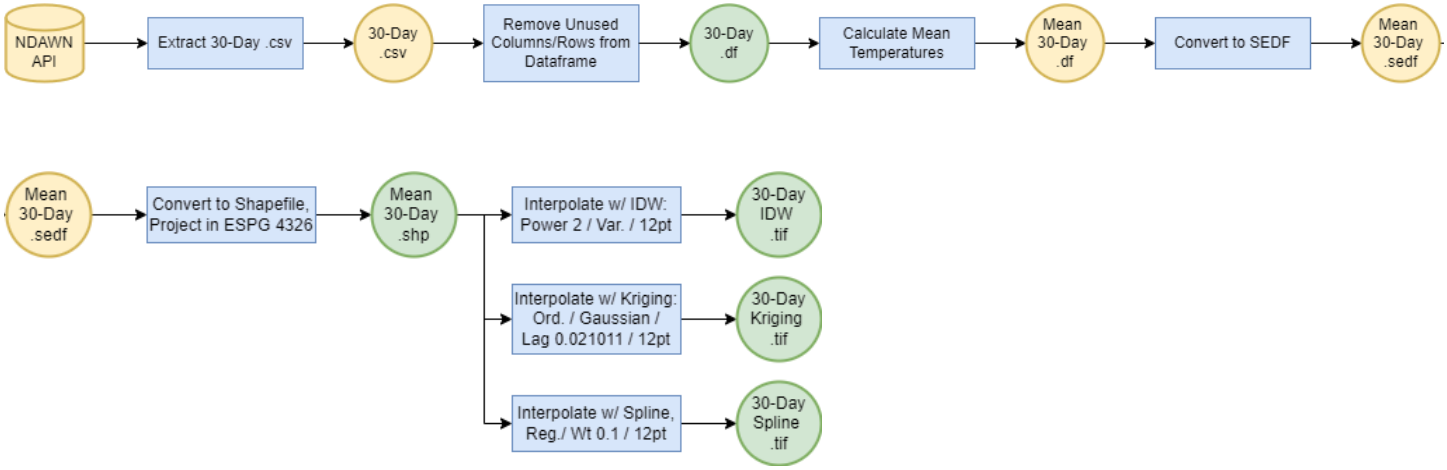
Figure 2. Data layer(s) used to perform the described processes.

#	Title	Purpose in Analysis	Link to Source
1	**NDAWN Temperature Data**	Used to calculate the mean 30-day temperatures for each station, which was interpolated across all NDAWN stations	NDAWN

Methods

Figure 3 shows the ETL used to extract the 30-day temperature data from all of the NDAWN stations and then set up a 30 day average that can be used for comparing interpolation techniques. NDAWN data was received as a .csv file which included extra header rows and columns with no data. The .csv file was converted to a dataframe in order to modify the data table. Each station had 30 average temperature values which were averaged for a single mean average temperature. The modified dataframe was converted to a spatially enabled data frame (SEDF) to encode the spatial location data.

Figure 3. Data flow diagram showing the API interactions used in this ETL.



The SEDF was converted to a shapefile and the file was reprojected to ensure that the data was in the EPSG 4326 (WGS 1984). The mean average temperature data for each station was used to interpolate mean average temperature for the entire NDAWN region, spanning all of North Dakota, far eastern Montana, and the western edge of Minnesota (Figure 4).

Three different generic interpolation methods were compared: inverse distance weighting (IDW), kriging, and spline. To better compare the interpolation methods, each method used a 12 point setup and (ESRI) default settings were retained in the remainder of the interpolation parameters. Raster files (.tif) were made for each of the interpolation methods.

Results

The ETL was able to successfully download, modify, and interpolate the NDAWN temperature data for a 30 day time period. Data points were able to properly project across the NDAWN station area (Figure 4).

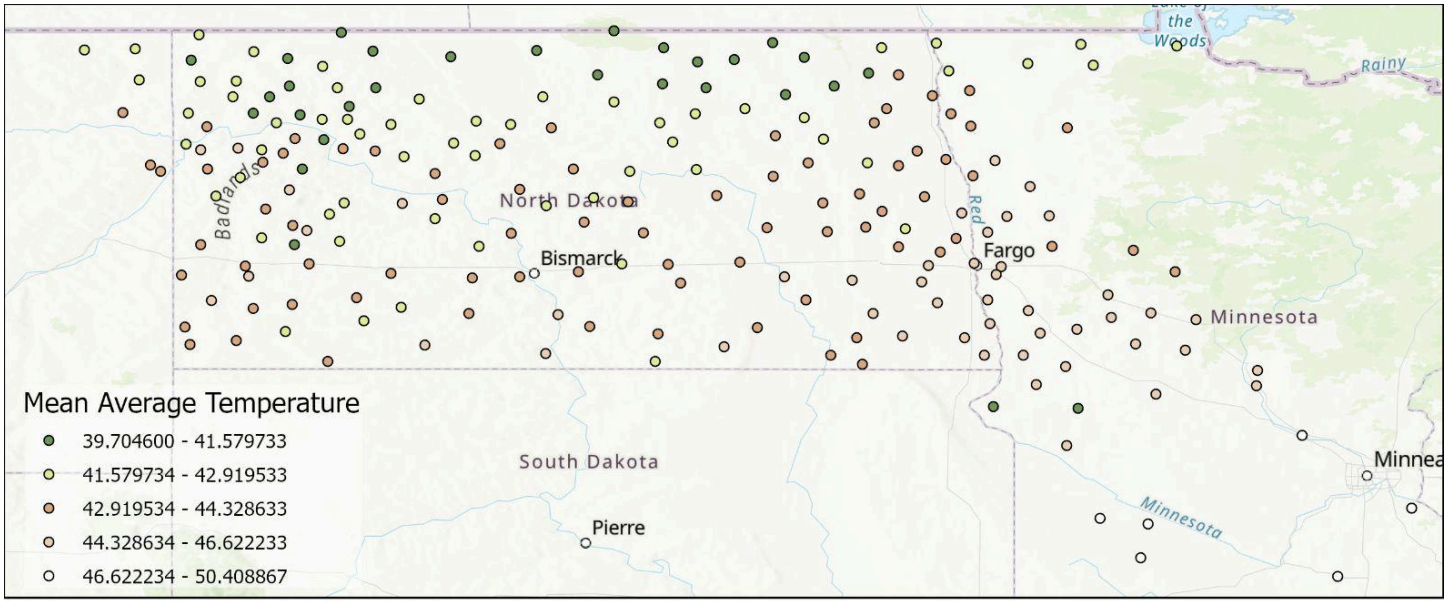


Figure 4. Map showing the distribution of mean average temperatures across all NDAWN stations.

Interpolation outputs are shown in the figures below (Figure 5, Figure 6, Figure 7). Each method created a raster output that represents mean average temperatures between all NDAWN stations. Color ramps are generally consistent across all of the interpolation and station maps, with green representing the coldest temperatures, and pink representing the warmest temperatures.

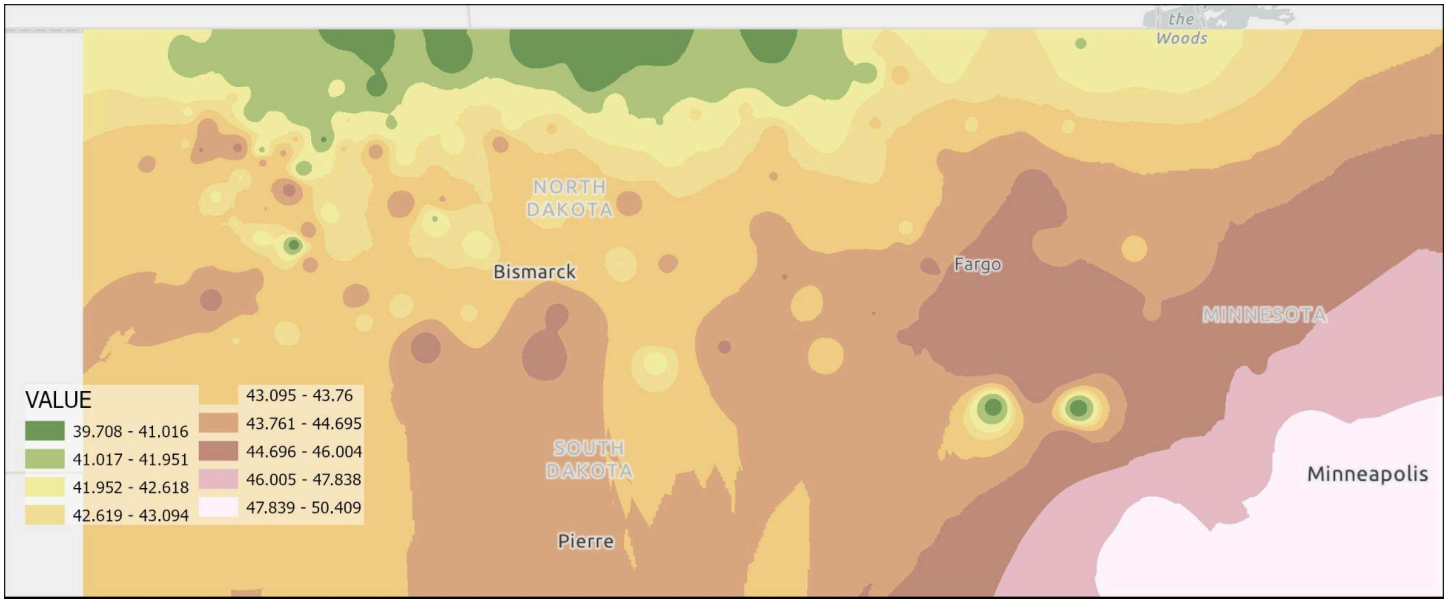


Figure 5. Map showing the interpolation of mean average temperatures using variable IDW (Power 2, 12 pt.).

IDW interpolation (Figure 5) shows temperature gradients that appear well aligned with the NDAWN station colors (consistently applied across all maps). Kriging interpolation (Figure 6) shows temperature gradients that appear more aligned with latitudes, and less varied by specific locations. Regular spline interpolation (Figure 7) shows temperature gradients that are more different from nearby station data and the other two methods. Spline interpolation resulted in the greatest distribution of interpolation temperature values.

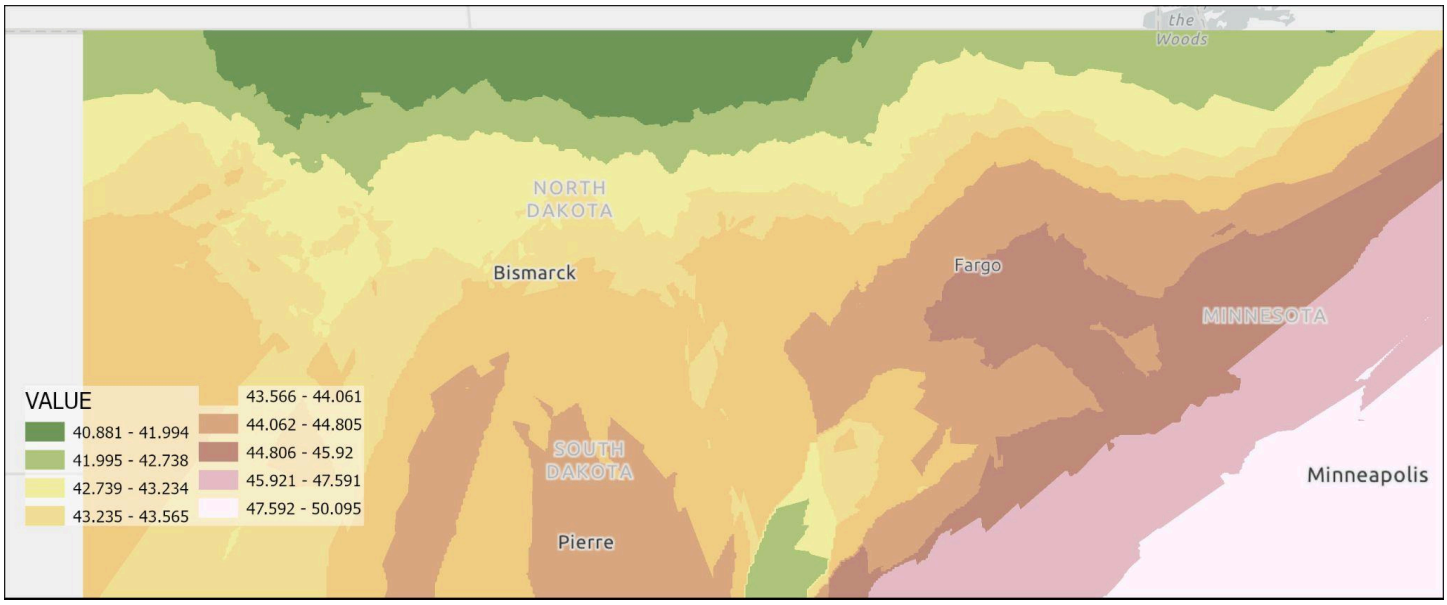


Figure 6. Map showing the interpolation of mean average temperatures using Gaussian Kriging (lag 0.021011, 12 point).

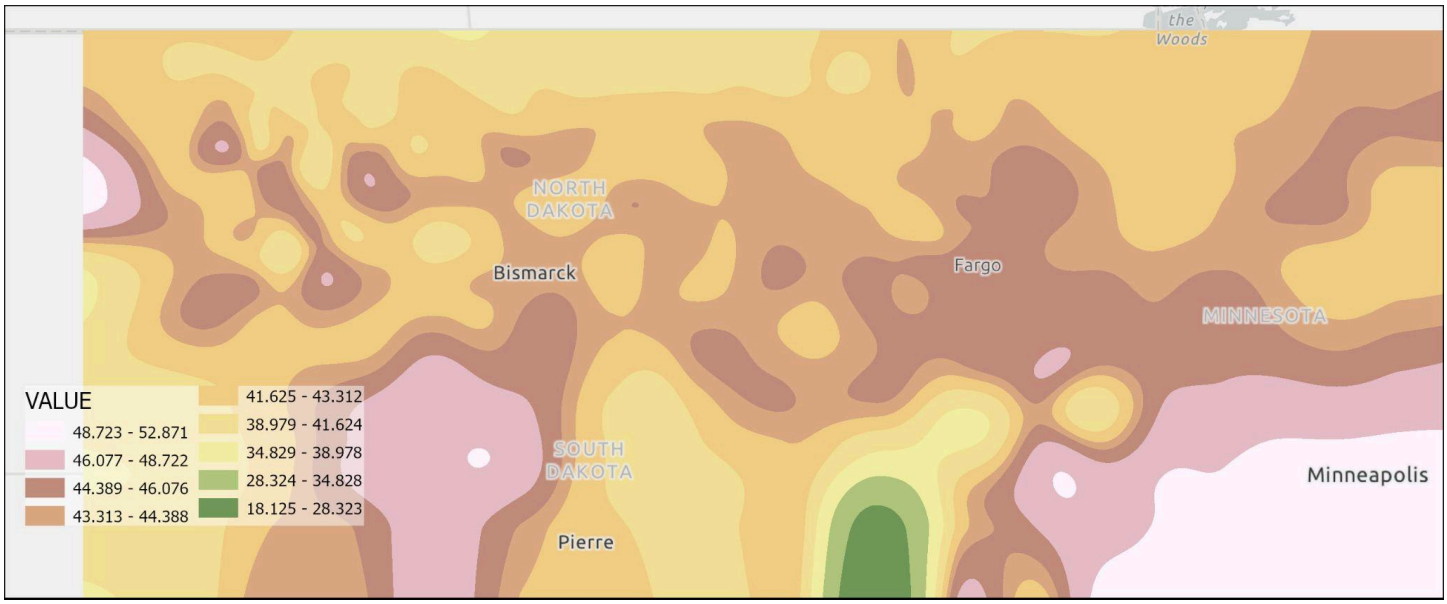


Figure 7. Map showing the interpolation of mean average temperatures using regular spline (wt. 0.1, 12 pt.).

The biggest variation between the interpolation methods that were applied can be seen around the stations south of Fargo, North Dakota. These stations had mean average temperatures around 39°F, while neighboring stations had values around 44°F. IDW resulted in bubbles around the stations, while both Kriging and spline methods resulted in more regional low areas, with the spline area having values at 18-28°F.

Results Verification

All of the stations were represented in the study, and each station had data from the past 30 days (at the time of the collection). Data was able to correctly be used for calculating a mean 30-day average temperature. Interpolation methods correctly projected a raster dataset for mean average temperatures across the region using IDW, kriging, and regular spline interpolation methods.

Discussion and Conclusion

Interpolation techniques vary in their algorithms that are used for calculating interpolated values between the input data. Interpolation can be either deterministic or stochastic, where deterministic outputs a specific value based on a specific algorithm while stochastic outputs utilize additional data and/or variograms to estimate a value based on the known dataset. Methods vary in their ability to present the user with error information, test spatial autocorrelation, present probability information, manage outlier information, and provide exact data values.

IDW is one of the most basic methods available in that it is deterministic and simply uses the input values and the weighted distances in the calculation for the unknown points. Figure 5 shows the output of the IDW method on the NDAWN dataset. This method uses a weighted average system that applies no additional assumptions to the data, and can create the bulls-eyes like those south of Fargo. IDW is simple to run, but appears to provide reasonable output data that makes sense as a generalized map. This method produced the most reliable appearing data output of the three maps.

Kriging is considered a more complex interpolation method since it uses additional data to make probabilistic valuations for each location. A variogram is developed based on the input data, and then used to help assign weights to the cells using neighborhood consideration. Spatial autocorrelation is also determined for each cell area, and can be applied locally or globally. Figure 6 shows a local kriging method applied to the NDAWN dataset. We can see that the existing dataset fits pretty well, though areas (like the region south of Fargo) do seem to be out of place based on the learned trends and variogram. While the result is consistent with the IDW method, we would need to run a cross-validation study to determine the overall performance of the method.

Spline is meant to work very well for smooth datasets since it uses a smooth trend-line to apply values. A smooth trend-line is developed using the existing dataset, and then distance is used to determine the proximity of the location to specific values on the graph. The value at the trend-line is used to predict the value of the cell. While temperature data is generally consistent, at more global scales (like that of the NDAWN data) it performs less well, since the stations represent large distances. As a result, we can see the effect of “outlier” data in the dataset in Figure 7. The output range is the largest of all three methods, with data produced at values as low as 18°F, which is well beyond the input range. The expected trend of low temperatures in the north and warmer temperatures in the south is also absent from this output, though the strange cold spot is still in the area south of Fargo.

Temperature data does not consist of a single method of interpolation that produces the best, or most accurate, data output. Depending on the conditions of the topography, the scale of the interpolation range, and the distances between the data stations, interpolation validity can vary significantly. Hofstra et al. (2008) tested interpolation methods for weather data around Europe. The research found that all models were successful with specific types of environments and conditions. They found that at broad (global) scales, differences were relatively minor, but that the distribution and spacing of the stations had the more significant impact on output reliability. A version of IDW (ADW2) was found to be the most successful with distant stations, much like the data in this lab activity. Hofstra et al. (2008) did conclude that a global kriging (GK) interpolation method was the most reliable overall, when considering all of the variables and cross-validation studies.

References

Hofstra, N., Haylock, M., New, M., Jones, P., & Frei, C. (2008). Comparison of six methods for the interpolation of daily, European climate data. *Journal of Geophysical Research: Atmospheres*, *113*(D21), 2008JD010100. https://doi.org/10.1029/2008JD010100

Classification trees of the interpolation methods offered in Geostatistical Analyst—ArcMap | Documentation. (n.d.). Retrieved November 25, 2024, from https://desktop.arcgis.com/en/arcmap/latest/extensions/geostatistical-analyst/classification-trees-of-the-interpolation-methods-offered-in-geostatistical-analyst.htm

Self-Score

Category	Description	Points Possible	Score
Structural Elements	All elements of a lab report are included **(2 points each)**: Title, Notice: Dr. Bryan Runck, Author, Project Repository, Date, Abstract, Problem Statement, Input Data w/ tables, Methods w/ Data, Flow Diagrams, Results, Results Verification, Discussion and Conclusion, References in common format, Self-score	28	**28**
Clarity of Content	Each element above is executed at a professional level so that someone can understand the goal, data, methods, results, and their validity and implications in a 5 minute reading at a cursory-level, and in a 30 minute meeting at a deep level **(12 points)**. There is a clear connection from data to results to discussion and conclusion **(12 points)**.	24	**24**
Reproducibility	Results are completely reproducible by someone with basic GIS training. There is no ambiguity in data flow or rationale for data operations. Every step is documented and justified.	28	**28**
Verification	Results are correct in that they have been verified in comparison to some standard. The standard is clearly stated **(10 points)**, the method of comparison is clearly stated **(5 points)**, and the result of verification is clearly stated **(5 points)**.	20	**20**
		100	**100**