Lab Report

Title: *NDAWN Real-time ETL and Interpolation*
Notice: Dr. Bryan Runck
Author: Jake Ford
Date: 11/30/2022

Project Repository: https://github.com/ThisFord/GIS5571-arc1.git
Time Spent: 30+

Abstract

The goal of this lab is to build an ETL data pipeline with pyhton and jupyter notebooks in ArcGIS Pro, gather real-time weather data, and perform data analysis through multiple interpolation methods. We then compare and contrast the four methods used in the resulting script to determine the most suitable method for interpolating temperatures from the source data. Methods compared are Inverse Distance Weighting (IDW), Spline, Ordinary Kriging and Empirical Bayesian Kriging (EBK). The data used for interpolation is temperature maximum, minimum and average normals gathered from North Dakota State University's online NDAWN weather network API. The data is programmatically downloaded for the most recent 30 days and transformed for use with jupyter notebooks and arcpy. The ArcGIS Pro notebook environment is used to perform the four interpolation methods and assess accuracy of the models. Results are critically compared and the best suited interpolation technique for the dataset is identified as Empirical Bayesian Kriging.

Problem Statement

Interpolating data from samples with a model is a core GIS data analysis methodology. In order to explore the different methods of interpolation, this project builds a fully functional real-time data visualization and analysis workflow to gather weather station data for the most recent 30 days from NDSU's NDAWN network. This data is used to compare and contrast four types of interpolation models on minimum and maximum temperatures across the region.

Table 1. Requirements

#	Requirement	Defined As	Prep
1	ArcGIS Pro	Software for geospatial processing	
2	Jupyter Notebook in ArcPro	Python programming interface in Esri's ArcPro software	
3	Lucid Chart	Model sketching online software	Sketch out workflows
4	Model Builder in Arc Pro	For building Geoprocessing workflows in Arc Pro and visualizing the process	Create and execute workflows

Input Data

The data used in the project, gathered through the ETL pipeline, is the most recent 30 days of temperature normal data from NDSU's NDAWN weather API. The data is comprised of temperature readings from all reporting stations in NDAWN's network, spanning North Dakota and parts of Montana and Minnesota. The API has been set to collect data from the most recent 30 days from the time the script is run, this feature provides the "real-time" interpolation effect, as new interpolations are constructed from new data each time the script is executed.

Table 2. Data

#	Title	Purpose in Analysis	Link to Source
1	NDAWN previous 30-day Normals Temperature data	Raw input dataset for sample locations and temperature interpolation method comparisons	NDAWN

Methods

The first step was to create an ETL pipeline to gather the most recent 30-day temperature data. This involved structuring an undated query to NDAWN's API using the python requests.get function. Removing the date parameters from the 30-day request forces the API to fill the information in on the NDAWN side, which defaults to the present day. This creates the real time effect, where each time the script is executed, the most recent data for the past 30 days is gathered to be used for the subsequent interpolations. The data is formatted for compatibility and saved to the disk as a csv. (Tobin, n.d.)

The csv is cast as a pandas dataframe for further in-script manipulation and the csv of points in imported into the project as a point feature class. The station locations are visualized categorically by average temperature. These points are used as reference for further mapping. The data is filtered using python and pandas to get the average minimum and maximum temperatures for the last 30 days for each station, which are used in the interpolation comparison. Each method, IDW, Ordinary Kriging, EBK and Spline is used on both the minimum and maximum temperature data and the results are compared in linked map displays with the minimum interpolated values on top, and the maximum values on the bottom.

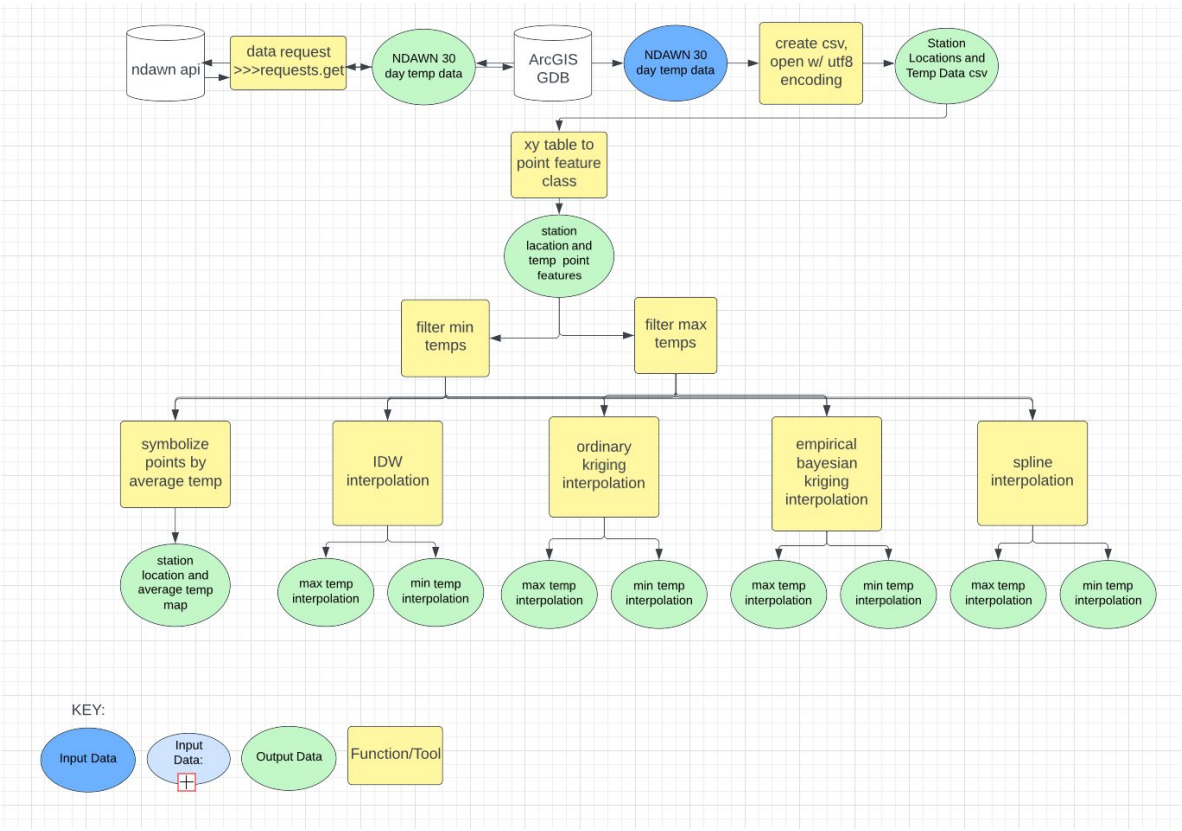

Figure 1. Data flow diagram for wrangling, mapping and interpolating NDAWN data.

Results

The real time data interpolation pipeline succeeded in gathering data and creating interpolated temperature maps using four different interpolation methods: Inverse Distance Weighted, Spline, Ordinary Kriging and Empirical Bayesian Kriging. The models used maximum and minimum observed temperature data to produce predicted values at unknown points with varying levels of accuracy. IDW produced spotty results, with high degrees of variation around sample points (figure 2), suggesting that the data is too sparsely distributed across the study for IDW to predict accurately. The spline method smoothed some of these variations out, but artifacts of spatial autocorrelation remain as apparent hot and cold spots where data is sampled and in regions where it is under sampled (figure 3). The kriging methods performed better, correcting for the distribution of points with autocorrelated data, providing similar results with EBK producing the smoothest interpolation with the smallest RMSE, (figure 4, 5, 6).(*An Introduction to Interpolation Methods—ArcMap | Documentation*, n.d.)

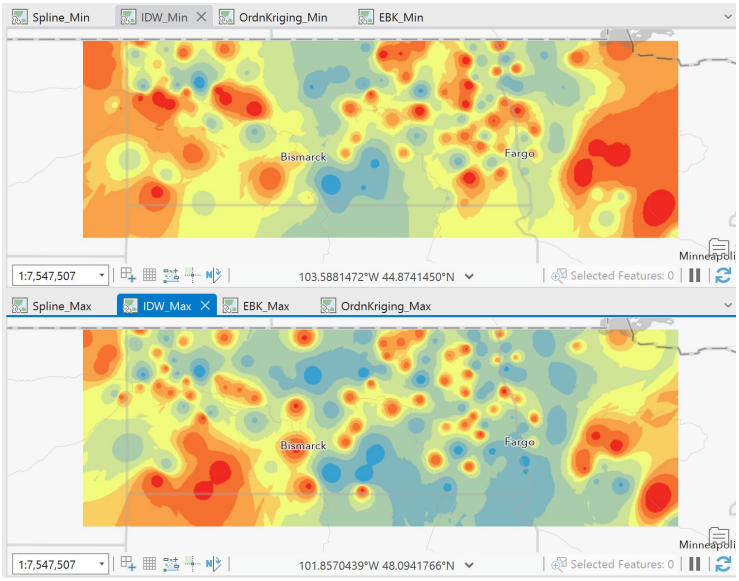

Figure 2: Inverse Distance Weighted method of interpolation, min temps on top, max on bottom.

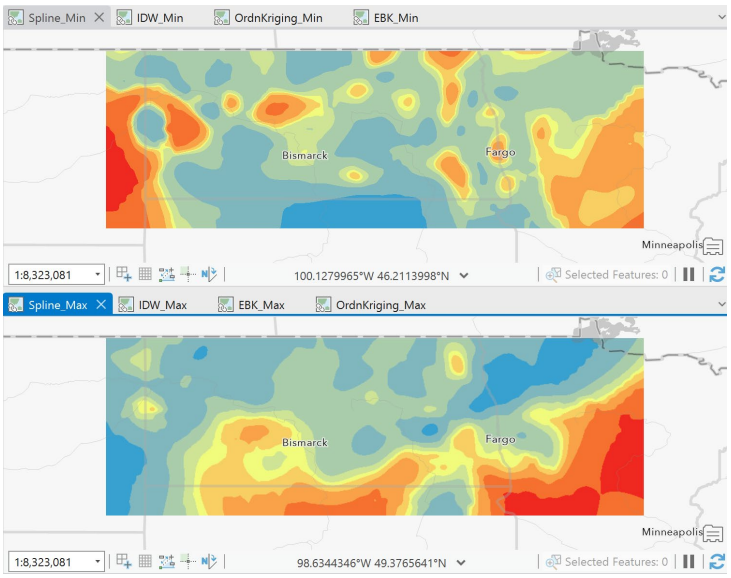

Figure 3: Spline method of interpolation, min temps on top, max on bottom.

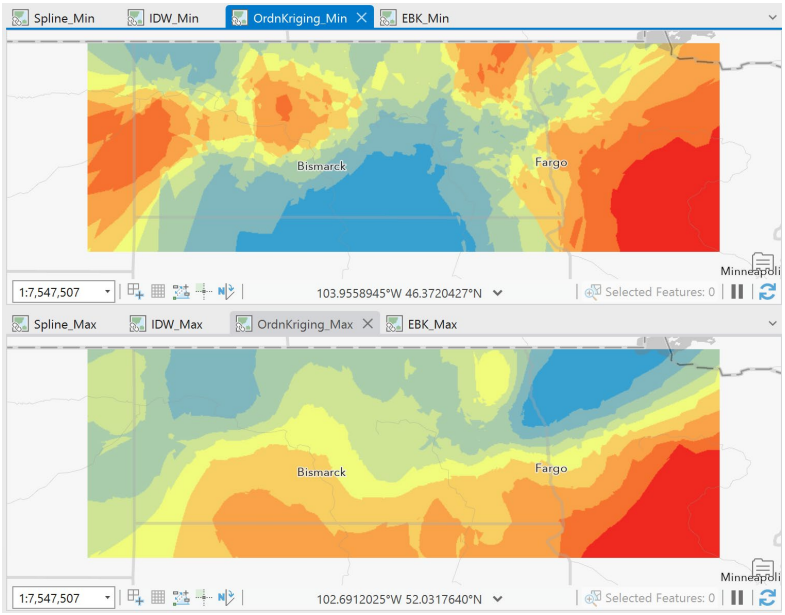

Figure 4: Ordinary Kriging method of interpolation, min temps on top, max on bottom.

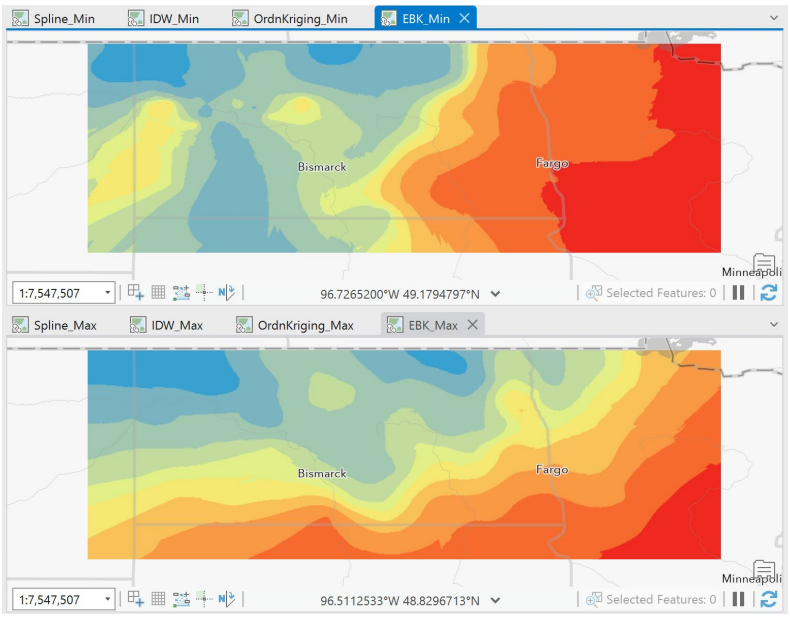

Figure 5: Empirical Bayesian Kriging method of interpolation, min temps on top, max on bottom.

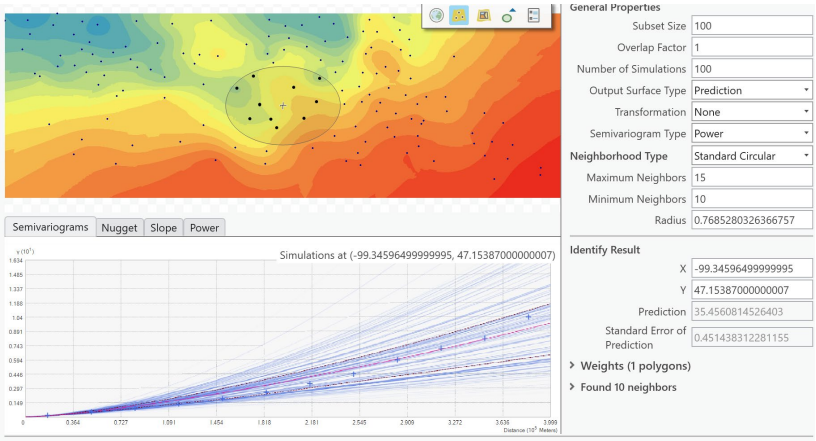

Figure 6: Semivariograms showing the results of multiple simulations and the line of best fit with EBK method on Max temp Averages.

Results Verification

The goal of this lab is to compare and contrast several techniques used in interpolating spatial data. Adhering to the standards of Multi Criteria Decision Analysis, the best approach is the one that satisfies the criteria of the user. In this project's case, the stated goal is to produce an ETL that allows for a compare and contrast analysis of several interpolation techniques, which has been accomplished.

Additionally, we test the accuracy of each of the models, with the Leave One Out method of Cross Validation accuracy assessment; where we perform the interpolation with one data point withheld and compare the resulting predicted value at that location with the actual value, iterating the process over all samples. The difference between the values is then used to compute a root mean square error. This process is done automatically in ArcGIS for every geostatistical interpolation method. The EBK method below shows very accurate interpolations, with an RMSE of about .5, or a half a degree difference on average between predicted and observed values, (figure 7.) With low error and a smoothed surface of

variation, EBK is very suitable for temperature modeling at this scale.(Loubier, 2007; *Using Cross Validation to Assess Interpolation Results—ArcGIS Pro | Documentation*, n.d.)

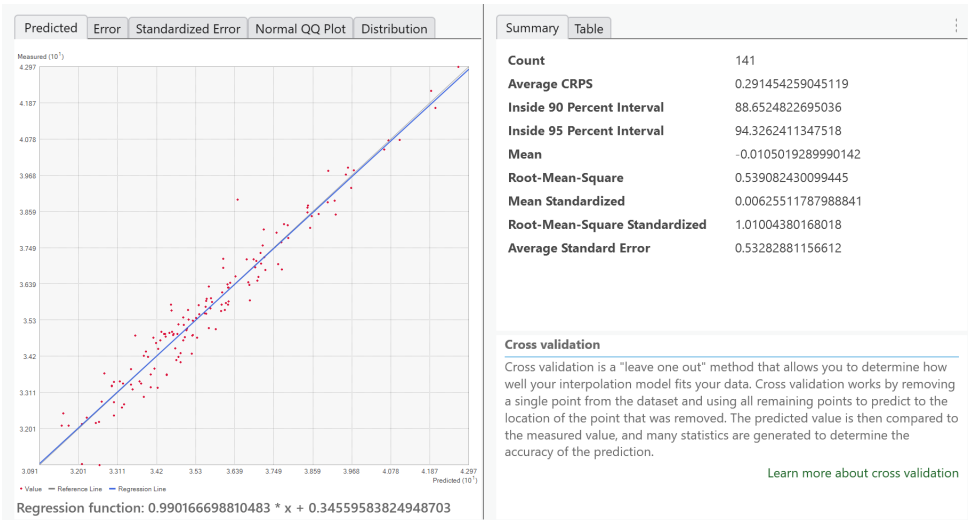

Figure 7: A RMSE of .53 shows that on average, the predicted max temp at a station was a half a degree off from the measured max temp, a pretty good fit for air temp modeling.

Discussion and Conclusion

There are many ways to interpolate weather data, all with their own best case use scenarios and various required inputs. The most successful method for the NDAWN weather data in this study proved to be Empirical Bayesian Kriging. Interpolation techniques like kriging and splining, while more computationally complex than deterministic methods like IDW and Nearest Neighbor, provide more accurate measurements with fewer covariate inputs, making more accurate predictive models easier to construct. Techniques like IDW are highly spatially dependent, which requires the model to consider spatially autocorrelated covariates like surface elevation, topography and landcover type to get very accurate measurements. Empirical Bayesian Kriging, in contrast, runs a series of simulations to automatically generate a best fit weights model that can account for spatial autocorrelation.(Jarvis & Stuart, 2001; Krivoruchko & Gribov, 2019; Loubier, 2007) The EBK model is both very accurate and relatively simple to execute due to the high level of automation and robust simulation process. This makes it more computationally intensive and susceptible to user error and data quality issues than some other models. For the purposes of this project, it remains the best fit with the smallest error and the simplicity of set up on the programing side. Of the four methods tested, the EBK method produced the smallest RMSE of about a half a degree of variance, which makes it the best performing model in this study.

References

An introduction to interpolation methods—ArcMap | Documentation. (n.d.). Retrieved November 30, 2022, from https://desktop.arcgis.com/en/arcmap/latest/extensions/geostatistical-analyst/an-introduction-to-interpolation-methods.htm

Jarvis, C. H., & Stuart, N. (2001). A Comparison among Strategies for Interpolating Maximum and Minimum Daily Air Temperatures. Part II: The Interaction between Number of Guiding Variables and the Type of Interpolation Method. *Journal of Applied Meteorology and Climatology*, *40*(6), 1075–1084. https://doi.org/10.1175/1520-0450(2001)040<1075:ACASFI>2.0.CO;2

Krivoruchko, K., & Gribov, A. (2019). Evaluation of empirical Bayesian kriging. *Spatial Statistics*, *32*, 100368. https://doi.org/10.1016/j.spasta.2019.100368

Loubier, J.-C. (2007). Optimizing the Interpolation of Temperatures by GIS: A Space Analysis Approach. In *Spatial Interpolation for Climate Data* (pp. 97–107). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470612262.ch8

Tobin, D. (n.d.). *The 6 Parts of ETL Data Pipeline Architecture*. Integrate.Io. Retrieved October 3, 2022, from https://www.integrate.io/blog/etl-architecture-building-blocks/

Using cross validation to assess interpolation results—ArcGIS Pro | Documentation. (n.d.). Retrieved November 30, 2022, from https://pro.arcgis.com/en/pro-app/latest/help/analysis/geostatistical-analyst/performing-cross-validation-and-validation.htm

Self-score
Fill out this rubric for yourself and include it in your lab report. The same rubric will be used to generate a grade in proportion to the points assigned in the syllabus to the assignment.

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**