Comparative Study of Geospatial Interpolation Methods using Quantitative Assessment

Manan Agarwal | February, 2022

Abstract

Spatial interpolation is commonly used to estimate physical data in a continuous domain where data is only available at a few distinct points. Different interpolation methods perform differently, depending on the properties and dynamics governing the physical phenomena underlying the data at hand. This paper aims to compare the performance of these interpolation methods over some actual data sets. Three different interpolation methods (IDW, Cubic Spline and Ordinary Kriging) are compared over three environmental datasets (ground-water, wind-speed, and rainfall).

Introduction

Spatial data is vital to various disciplines – it is necessary to make informed decisions ranging from environmental management to renewable plant locations. Unfortunately, the collection of data can largely only occur at a few, discrete points, through installed elements/gauges, requiring the use of interpolation if the relevant quantity has to be estimated elsewhere.

Therefore, many researchers have investigated and compared various interpolation methods previously. For instance, (Caruso et al, 1996) examines the performance of methods as a reflection of the spatial correlation of the datasets (environmental or mathematically generated) while (Ajvazi et al, 2019) applies methods to geographical elevation.

In this paper, we assess the performance of three spatial interpolation methods (IDW, Cubic Spline and Ordinary Kriging). Each of these methods, discussed in more detail in the Methods section, were applied to three different datasets derived from environmental studies:

ground water, wind speed and rainfall. The points in these datasets are irregularly spaced, an accurate reflection of real-world applications of interpolation.

Methods

The chosen interpolation methods are Inverse Distance Weighting (IDW), 2D Cubic Spline (CS) and Ordinary Kriging (OK).

1. IDW

Under this method, each interpolated value for an unknown point is essentially a weighted average of the values of sample points in the data, wherein the weight is inversely proportional to the distance between the unknown points and sample points. Thus, this method gives greater weight to the points near the unknown point. For spatial function f,

$$f(x,y) = \frac{\sum_{j=1\ldots N} \frac{v_j}{d_j}}{\sum_{j=1\ldots N} \frac{1}{d_j}},$$

where d_j is the Euclidean distance between the unknown point and sample point j, v_j is the value at sample point j, and N is the number of sample points. While neat, it also provides appropriate estimations for a large gamut of situations. Unfortunately, its biggest limitation is in never being able to provide estimates outside the range of values at the sample points. It is also particularly sensitive to the weighting, as is apparent, which can be a drawback depending on the mechanics of the measured variable.

2. CS

Spline predicts values on unknown points by creating piecewise polynomial surfaces on each patch based on the sample points. A cubic spline takes into account the function value and up to two derivatives to ensure smooth interpolated surfaces, with minimum curvature. The interested reader may find that (Agrapart et al, 2020) offers greater insight.

3. OK

A geostatistical method, Ordinary Kriging is similar to IDW in that it accounts for Euclidean distance between the unknown point and sample points, but by generating a spatial correlation between sampled points. It overcomes a limitation of IDW in that it accounts for clustering of certain sample points, preventing overweighting of said clusters. There is a wide range of choices available for the variogram, whose choice can be guided by the data itself. The interested reader may find that (Caruso et al, 1996) offers greater insight.

Data

The following three data-sets were used for our analysis:

1. Windspeed: Average wind speed at 50 locations near Vijaypura, Karnataka over FY 2020-21

   Courtesy: Greenko Energies Pvt. Ltd.

2. Groundwater: Average depth to water level in 34 districts of Maharashtra, May 2019

   Courtesy: Central Ground Water Board (http://cgwb.gov.in/documents/2019/MAHARASHTRA.pdf)

3. Annual total rainfall in 31 districts of Maharashtra, 2010

   Courtesy: Indian Meteorological Department and India Water Portal (compilation)

   (https://www.indiawaterportal.org/articles/district-wise-monthly-rainfall-data-2004-2010-list-raingauge-stations-india-meteorological)

In each of the three datasets, we have transformed the data so as to be amenable to interpolation, by converting the longitudes and latitudes into Euclidean space coordinates. The ‘depth to water level’ data is derived from a frequency table.

The motivation to use these datasets is as follows:

1. Wind speed may be needed to find ideal locations for installation of windmills such that efficiency is maximised. More importantly, accurate predictions of power output can be made informing business-oriented decisions.
2. Watershed management requires groundwater depth estimation to ensure that there is a suitable level of groundwater throughout, which is necessary for irrigation (agricultural) purposes as groundwater feeds rivers and streams.
3. Watershed management also requires interpolation of rainfall data to build hydrological models which then support forecasting, critical to industries like agriculture.

Error metrics

Illustrated below are a few metrics through which performance is measured. This paper primarily considers relative mean error (RME) in analysing the results. This is because the values within 2 of the 3 datasets have high standard deviation which RME accounts for, by taking relative error.

1. Mean Absolute Error = $\overline{|z_i - \widehat{z}_i|}$;
2. Relative Mean Error = $\overline{\frac{|z_i - \widehat{z}_i|}{z_i}}$;
3. Root Mean Square Error = $\left(\overline{(z_i - \widehat{z}_i)^2}\right)^{1/2}$;

where the horizontal parenthesis above z_i denotes predicted values.

Methodology

While testing for each dataset, the points therein have been randomly split into two subsets A and B (in differing ratios) with points in A being used as sample points to interpolate, and their accuracy is checked (using RME as the metric) by comparing the interpolated values to the actual ones for all points in B. This is known as ***crossvalidation*** (Caruso et al, 1998).

For OK, where there is a choice of variogram, we have used the spherical variogram as that yielded the best performance.

A trial consists of one instance of dividing the given dataset randomly into A and B (as described above), interpolating and computing RME for each point in B and then averaging the RME over all points in B. For each method and for each dataset, we do a thousand such trials, measure RME for each of the thousand trials and report the average across these trials in the results below.

Results

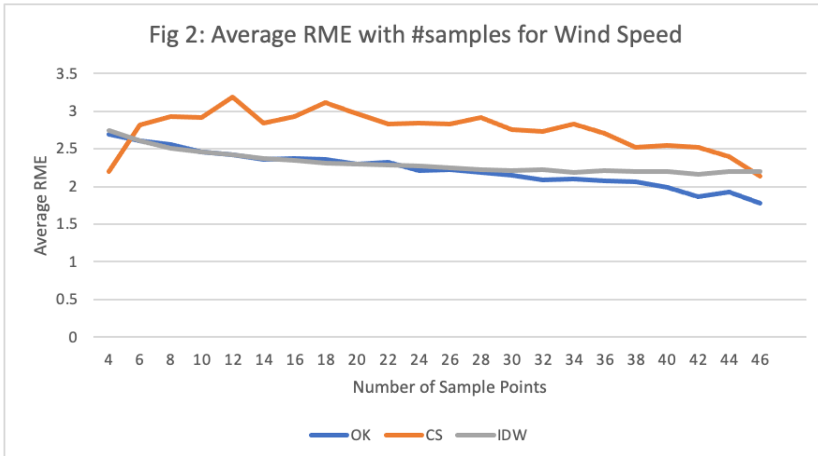
The summary results are tabulated in Figure 1 below.

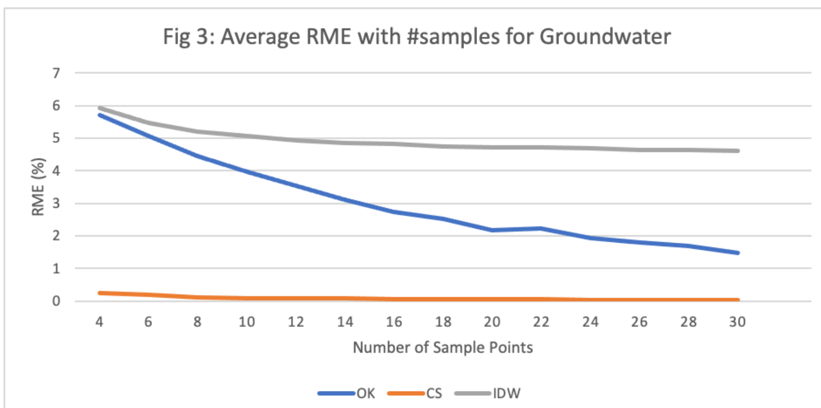
Fig 1: Average %RME produced per method per dataset

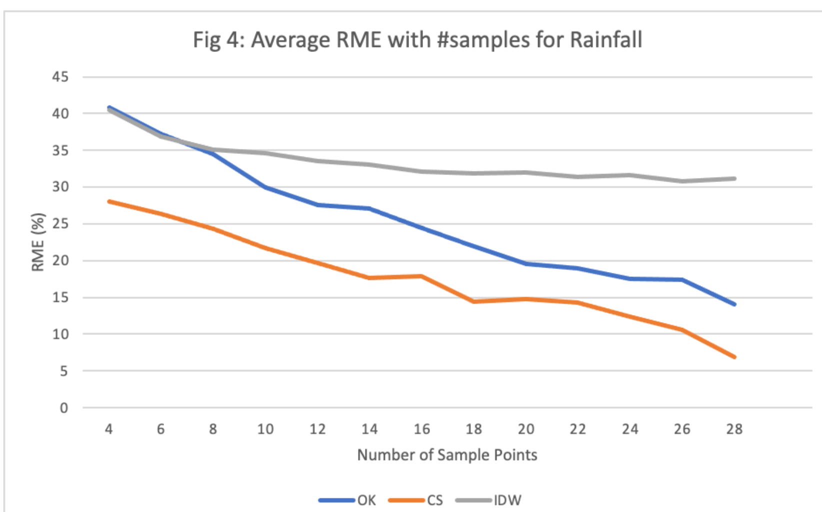
Methods	**Average RME* / %**		
	Wind [1] (30 sample, 20 unknown)	**Groundwater [2] (20 sample, 10 unknown)**	**Rainfall [3] (20 sample, 10 unknown)**
IDW	2.22	4.67	32.2
CS	2.82	0.0484	14.3
OK	2.13	2.31	22.2

** Rounded to 3 s.f.*

More insight can be derived by plotting how the number of sample points / unknown points used affects the average RME. Figures 2, 3 and 4 depict this for wind speed, groundwater and rainfall respectively.







Observations & Insights

Fig 2 shows that error levels are low for all three methods but OK and IDW are more accurate for interpolating wind speed as compared to CS. Accuracy improves marginally as more data points are included in the sample set, as expected, and OK is the most accurate method with higher sample points.

Fig 3 shows that CS is the most accurate method of interpolating groundwater data with negligible level of error and significantly better than IDW and OK. CS is quite accurate even with limited data points.

Fig 4 shows that rainfall has a weak spatial correlation with relatively (and even in absolute terms) high error levels across all methods but CS is better than OK and IDW. The accuracy improves, again across methods, with more data points, as expected.

A recurring trend is the divergence between the accuracies of OK and IDW as the number of samples increase: the performance of OK continues to improve, more so than IDW. This can be attributed to a supposed 'cluster' effect wherein IDW does not account for the spatial distribution of the points i.e. multiple points in a dense region or cluster, with similar values, will exert an overt influence on the prediction at the unknown point, a result of the importance given to the weighting function (mentioned in the Methods section). This notion is corroborated by the divergence being reduced for wind speed data. Since this data was measured by Greenko, a renewable energy company, to perform a survey for a prospective wind farm, the points are more evenly distributed, reducing the cluster effect.

Analysing the errors of CS, it consistently performs better than the other methods for both the groundwater and rainfall data, irrespective of the number of samples. Both these sets contain just a single datapoint from most districts in Maharashtra and so, are distributed over a state of >300,000 km^2 in area. In contrast, the wind speed dataset contains points within a 5 kilometer radius so the sampling density is greater. This amplifies the influence of any noise (which is inevitable in any measurement / observation), making it difficult for CS to 'smooth out' the curvature and increasing the scope for error. So, CS has the worst performance over wind speed data.

Finally, CS applied to groundwater data produces a miraculously accurate, almost anomalous result, deserving of special attention. The error is far lower than for any other dataset or method. Then, its accuracy must have to do with the particular measured variable and the specific interpolation technique. Turning to geology, owing to the phenomenon of groundwater flow, water flows "*from high to low hydraulic head*" (PennState, n.d.) which would cause gradients in water concentration to smooth out. This could possibly align with CS' tendency to minimise curvature in the elevation model, thereby reducing error substantially. Simply put, the interpolation method seems to be in sync with the underlying phenomenon.

Conclusions

Through the course of this paper, several interpolation methods have been investigated. It is evident that a blanket statement cannot be made regarding the performance of a technique – that one is better in all circumstances.

It is important to select the appropriate technique according to the data set, though it can be observed with some consistency that OK performs better than IDW for reasons discussed in the paper. Future investigations need not consider IDW beyond its use as a baseline, instead including other methods in the domain.

An important aspect of the use of CS is the sampling area at which data is available. In a related manner, the underlying ‘physics’ of the phenomenon can also inform the appropriateness of an interpolation method.

Finally, that accuracy improves with more data samples is known. More importantly, the number of samples taken can be set based on what error level is acceptable, thereby limiting costs.

References

1. Agrapart, Q., & Batailly, A. (2020). Cubic and bicubic spline interpolation in Python. [Other] *École Polytechnique de Montréal.* https://hal.archives-ouvertes.fr/hal-03017566v2.
2. Ajvazi, B., & Czimber, K. (2019). A comparative analysis of different DEM interpolation methods in GIS: case study of Rahovec, Kosovo. *Geodesy and Cartography*, *45*(1), 43-48. https://doi.org/10.3846/gac.2019.7921.
3. Arthur, M., & Saffer, D. (n.d.). *Hydraulic head and the direction of groundwater flow.* EARTH 111: Water: Science and Society. Retrieved October 2, 2021, from https://www.e-education.psu.edu/earth111/node/932.
4. Columbia University. (n.d.). *Kriging interpolation*. Columbia Mailman School of Public Health. Retrieved October 7, 2021, from https://www.publichealth.columbia.edu/research/population-health-methods/kriging-interpolation.
5. Caruso, C., & Quarta, F. (1996). Interpolation methods comparison. *Computers & Mathematics with Applications*, *35*(12), 109-126. https://doi.org/10.1016/S0898-1221(98)00101-1.
6. Kravchenko, A. N., & Bullock, D. G. (1999). A comparative study of interpolation methods for mapping soil properties. Agronomy Journal, 91, 393-400. https://doi.org/10.2134/agronj1999.00021962009100030007x.
7. Murphy, B.S., & Mueller S. PyKrige 1.6.1 [Source code] https://geostat-framework.readthedocs.io/projects/pykrige/en/stable/.
8. Virtanen et al. (2020) **SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python**. *Nature Methods*, 17(3), 261-272.
9. Wes McKinney. (2010). **Data Structures for Statistical Computing in Python**, Proceedings of the 9th Python in Science Conference. 51-56.
10. Wu, Y., & Hung, M. (2016).Comparison of Spatial Interpolation Techniques using Visualisation and Quantitative Assessment. *IntechOpen*. DOI: 10.5772/65996. Available from: https://www.intechopen.com/chapters/52704.

Acknowledgements

The technical support provided by Greenko Energies Pvt. Ltd. and open-access data provided by the Indian Meteorological Department (IMD) and Central Ground Water Board (CGWB) is greatly appreciated.