Proceeding of the 10th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences
Florianopolis-SC, Brazil, July 10-13, 2012.

# On the interpolation algorithm ranking

*Carlos López-Vázquez*

LatinGEO Lab, SGM+Universidad ORT del Uruguay
carlos.lopez@ieee.org

## Abstract

*Interpolation of data gathered at a finite number of locations is an everyday issue with spatial data. The choice of the best interpolation algorithm has been a topic of interest for a long time. Typical papers take a single dataset, a single set of data points, and a handful of algorithms. They report results of considering a subset A of the data points, application of each algorithm to the complement of A, and evaluating the MAD/RMSE over such points. The lower the better, so a ranking among methods (without confidence level) can be derived based upon it. We believe that the best interpolation algorithm should consider not merely the function value at some designated points, but also the spectral properties of the original field. We have used a metric named ESAM for that. Using a sample of N = 500, 2500 and 5000 irregularly distributed points taken from a reference DEM, we applied a number of interpolation methods and create a ranking among them using MAD, RMSE and ESAM as the figure of merit. ESAM ranking does not agree with the others. In addition, in this paper we will show how to build a ranking with a confidence level.*

**Keywords**: Interpolation, RMSE, MAD, ESAM, ranking

## 1. Introduction

Any interpolant is an empirical function estimated with data from scattered locations which honours the data values. They are widely used in almost any field (geosciences, mathematics, physics, etc.) for predictive and visualization purposes. Since the empirical functions are not unique (unless extra requirements are imposed, like a prescribed function subspace) there is a permanent interest in assessing and selecting the best one.

There exist a number of options, and the user might have difficulties to choose among them; the selection of the best interpolation algorithm has been a topic of interest for a long time (Goodin *et al.*, 1979; Franke, 1982; Lam, 1983; Dubois 1997; Caruso and Quarta, 1998; Li and Revesz, 2004, and others). Typical papers take a single dataset, a single set of data points, and a handful of algorithms. They report results of considering a subset A of the data points, use them to estimate the empirical function, apply it to the complement of A, and evaluating some metric over discrepancies observed at such points. There is one value of the metric for each interpolation method. Subset A might be kept fixed in the experiment, or might be systematically build by excluding one point at a time (process known as

Cross Validation, CV, to be described below). There exist a number of possible metrics: mere square root of the sum of squares (RMSE), mean absolute deviation (MAD), etc. Irrespective of the choice, the smaller metric the better, and thus a ranking among methods can be derived based upon its value.

## 2. Traditional metrics

There have been attempts to have a well established testbed to compare a large number of methods (Dubois 1997). In their experiment they coordinated a blind experiment where daily rainfall values at 100 locations were used to predict the outcome at 367 other locations. After the estimate was received, the participants evaluate their own merit using RMSE and a rank could be derived. However, no confidence level was attached to the result.

(Kidner *et al.*, 1999) compared six synthetic DEMs against a number of local polynomial functions, concluding that the higher order ones over perform the linear ones. For each of the test DEM, they produced a rank among the available functions, also based upon the RMSE. Later, the final winner was declared after summing their ranks. No confidence levels were reported, as well as enough detailed information to repeat the experiment.

(Falivene *et al.*, 2010) described an experiment where the figure of merit (metric) is the mean absolute deviation of a vector of cross-validated (CV) values. Its j-th element can be calculated by hiding the j-th data value, use the N-1 others for estimate it with each algorithm and calculate the difference with respect to the hidden known value. Ideally all vector elements should be zero; a good result will be that all values are small. For convenience, they emphasize that some methods consider just a small number of neighbours in the vicinity of the known value, and ignore the others; in some methods such number might be regarded as a parameter. They claim that appropriate metrics of success is based upon the absolute deviation, using its mean (MAD) and its standard deviation (SD), but considered jointly. They reported that in the surveyed literature no consistent result regarding ranking using MAD metric could be derived: method A outperforms method B in some papers and the opposite happens in others. According to the authors neglecting the results in terms of SD is the reason behind such inconsistence. They found that SD decreases while increasing the number of neighbours considered in the calculations (which might be as large as the total number of available points minus one, and as low as one, as it happens for the nearest neighbour method). The authors claim that there exists a trade off between the minimum MAD and the SD; increasing the number of neighbours decrease MAD but increases SD. The conclusion is that the best method (with its optimal parameters) should minimize at the same time MAD and SD. In their calculation the authors use all the available data with a jacknife approach.

(FGDC, 1998) issued the National Standard for Spatial Data Accuracy (NSSDA hereinafter), providing a procedure to evaluate a metric of success based upon RMSE. The procedure could be readily adapted for our purposes: the control points are splitted into a subset A and its complement. The former might be chosen by hand or at random, and the latter should have at least 20 points well distributed in the domain. They will take the role of "ground truth" (without being so, as correctly pointed by Bater and Coops, 2009). The interpolation procedure only considers

data from subset A, and the accuracy of the interpolated surface w.r.t. ground truth could be evaluated by building a vector of discrepancies of at least 20 elements. Its norm (i.e. RMSE), its MAD, or any other metric could be used. (Bater and Coops, 2009) used this procedure to over 1.3e6 points (which is their case is 3% of the available set). Notice that, with the CV method the error vector used for the metric is unique, while every random choice of the 20 points reserved for NSSDA calculation will provide a different value, giving room to perform a simulation. This difference will be crucial for our proposal.

(Bater and Coops, 2009) reported an experiment where LIDAR elevation data were pre-processed to locate bare soil points, and afterwards they were interpolated to estimate a DEM. The goal was to calculate tree height through the difference between the raw LIDAR elevation data (measuring top of canopy as well as bare soil) and the DEM. Their metric of success was the MAD and MSE calculated over a set of control points. By considering other sources of information (like vegetation index, slope, etc.) they were able to enrich the deterministic DEM with a companion uncertainty surface. They performed a simulation with just 48 events, and ranked the results according to the mean of every relevant metric.

As a final conclusion, we could state that:

a) most of the rankings are derived from a single dataset, using all available data at once with a deterministic procedure. This includes as a particular case the CV case. Others used a stochastic approach, by select a random subset of the available information to estimate the interpolant, and the rest to evaluate its merits.
b) when a simulation has been performed, the ranking is the result of analyzing the mean value of the relevant metric (RMSE, MAD, etc.)
c) the relevant metric might be problem dependent

## 3. A metric which considers spectral properties

It has been argued that some interpolation methods adds artifacts to the surface, which do not appears in the real field. In the image processing community it is commonplace to analyze such topic not in the euclidean space but in the spectral one. For simplicity, we will assume that our ground truth is arranged in a regular grid, and that all interpolation methods will also estimate values at pixel locations.

Our assumption is that the best interpolation method will not only fit data at control points, but also will produce either a 2D periodogram or a 2D Power Spectrum Density (PSD) which resembles the expected one. This need not to be a perfect fit; we could use a standard metric (like the ESAM one, described by Chen *et al.*, 2008) to measure its discrepancy, or resort to some other home-made visual aid. The ESAM metric compares images; we will use to pairs of PSD. If applied to images U and V with corresponding pixels $u_i$, $v_i$, the expression is

$$ESAM(U,V) = \arccos\left(\frac{2\sum_i u_i v_i}{\sum_i u_i^2 + \sum_i v_i^2}\right) \qquad (1)$$

which is exactly zero for U=V. Our images U and V will be the interpolated DEM and the ground truth, both minus their own spatial average. The interpolated DEM is build as the sum of the low accuracy DEM (completely known in the domain)

plus an interpolation of the difference against ground truth but only known at N control points.

## 4. Proposed ranking procedure, now with confidence level

As mentioned before, the stochastic approach as used in the reported literature take first the average results over the different events, and afterwards sorts the methods according to a "greater-than" rule. As a fundamental disadvantage, this procedure hides the fact that some method might be better in most of the events, but fail miserably in a handful of others degrading its average performance. For our purposes, we are willing to select at the top of the ranking the method which *performs the best in most of the events*.

This is a known problem in medicine, usually applied for assessing the effect of a treatment to different individuals. Other equivalent formulation is a wine contest: each judge produces a ranking among the bottles, and the collective output is analyzed in order to find systematic differences. The null hypothesis is that they do not exist. The Friedman Test (described by Bewick *et al.*, 2007) assumes that there exist a table with enough rows (events) and as many columns as algorithms. Each row holds the rank among the methods for a particular event, according to the RMSE, MAD or whichever continuous or ordinal criteria. Given a confidence level, the test accepts or rejects the null hypothesis. In addition, it helps to state that some methods do not show statistically significant differences. The Friedman Test requires a minimum number of events (15) which we will show that can be easily achieved and a minimum number of methods (4) which some of the reported papers do not satisfy.

## 5. Data and methods

We claim that the best interpolation algorithm should not merely fit at much as possible the DEM height at some independent control points, but also should take into account the spectral properties of the field. To illustrate our point we devise an experiment where a reference DEM and a lower accuracy one are available. They are from the Aix-en-Provence region in the South of France, both of 12.42 km by 6.9 km, 30 m spacing and described by (Day and Muller, 1988). Using the latter and a sample of irregularly distributed points from the former, we applied a number of interpolation methods and create estimates of the reference DEM. Its accuracy can be evaluated using no less than 20 well separated control points (following FGDC, 1998) and estimating the RMSE/MAD of the differences between reference DEM and each interpolated DEM. Independently, we evaluate the 2D power spectrum density of the reference DEM as well as each of the interpolated ones. As a metric of success we used the ESAM one (Chen *et al.*, 2008) which belongs to [0,1]. Finally, we compared and discussed both rankings.

We will denote as *event* one particular choice of a sample of control points to be used by each interpolation method. In this experiment, control points are pixels so their spatial coordinates are discrete. For each event, we selected at random N=500, 2500 and 5000 non-duplicated pixels not coinciding with any of those used for the NSSDA evaluation. We applied six different interpolation algorithms, evaluate the

results over the 20 NSSDA pixel locations and quantify the accuracy as MAD and RSME following (FGDC, 1998). The PSD of the ground truth and each interpolated DEM were used to evaluate the ESAM metric. After collecting the results of all 250 events, we might proceed as usual by averaging each metric (i.e. RMSE for example) and ordering the methods according to the smallest values. Three possible rankings could be obtained, according to RMSE, MAD and ESAM. However, as described before, we could safely rank the methods with a 95% confidence levels using instead the Friedman test, producing a (surprisingly) different ranking.

## 5. Results

Results are summarized in Table 1. Following common practice, the average of 250 events of each metric is used to create a ranking among interpolation methods. For example, the conclusion is that when there exist a large number of control points, IDW_2 (Inverse Squared Distance Weighting) is the preferred option in terms of the ESAM criteria, followed by Nearest Neighbour (NEARESTN) and GRIDDATA_V4. However, if we rank each event, and afterwards apply the Friedman Test to 250 possible rankings, we can conclude with 95% confidence level that the best option is NEARESTN, followed by GRIDDATA_V4 and IDW_2. All differences between methods were statistically significative.

**Table 1:** Ranking among the six different interpolation algorithms with 95% confidence level as a function of the number of interpolation points. Results after 80, 120 and 250 events were almost the same, so they are not reported separately. Entries presented as "A/B" shows disagreements between the ranking according to Friedman test and the ordering based upon the mean values; otherwise a single value is shown.

| | MAD | | | RMSE | | | ESAM | | |
|---|---|---|---|---|---|---|---|---|---|
| | 500 | 2500 | 5000 | 500 | 2500 | 5000 | 500 | 2500 | 5000 |
| GRIDFIT_Smooth20 | 4 | 4 | 4/5 | 4 | 4 | 4/5 | 4 | 4 | 4/5 |
| GRIDFIT_Smooth05 | 5 | 5 | 5/4 | 5 | 5 | 5/4 | 5 | 5 | 5/6 |
| GRIDFIT_Laplacian | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6/4 |
| IDW_2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 3/1 |
| NEARESTN | 2 | 3 | 3/2 | 2/3 | 3 | 3/2 | 2 | 1 | 1/2 |
| GRIDDATA_V4 | 2 | 2 | 2/3 | 2/2 | 2 | 2/3 | 3 | 3 | 2/3 |

When considering MAD or RMSE, the IDW_2 consistently outperformed the others irrespective of the number of data points and the way we build the ranking. On the other hand, the second and third place depends on both; GRIDDATA_V4 is consistently the second while NEARESTN is very close for 500 points (there is a tie both for MAD and RMSE) but fall to the third place when more control points are available.

The ranking by the ESAM criteria changes as the number of control points increases; NEARESTN appears to be the winner, improving with N, while IDW_2 degrades consistently.

## 3. Conclusion

Using the Friedman test, we could attach a confidence level to a ranking between interpolation methods. Our results show that common practice (rank the mean values of the Monte Carlo experiment) does not necessarily produce the same rank. In the experiment, disagreements were found for the case of 5000 control points irrespective of the metric considered.

Another contribution of this work is the consideration of spectral characteristics of the interpolated field. Methods which are the best with traditional metrics (RMSE, MAD) perform worse when considering spectral metrics, which suggest a tradeoff. The choice of which spectral metric (ESAM) to use might be a subject of future evaluation.

We believe that a rank with confidence level is a significant improvement over common practice, and that future comparison should take into consideration also spectral properties of the interpolated field.

## References

Bater, C. W. and Coops, N. C. (2009) "Evaluating error associated with lidar-derived DEM interpolation". Computers & Geosciences, Vol. 35, 289–300

Bewick, V.; Cheek, L. and Ball, J. (2004) "Statistics review 10: Further non-parametric methods". *Critical Care* Vol. 8, 3, 196-199

Caruso, C. and Quarta, F. (1998) "Interpolation Methods Comparison". *Computers Math. Applic.*, Vol. 35, 12, 109-126

Chen, S.; Su, H.; Zhang, R.; Tian, J. and Yang, L. (2008) "The Tradeoff Analysis for Remote Sensing Image Fusion using Expanded Spectral Angle Mapper", *Sensors*, 8, 520-528.

Day, T. and Muller, J.P. (1988) "Quality assessment of Digital Elevation Models produced by automatic stereo matchers from SPOT image pairs". *Proceedings of the 16th. International Congress of the International Society for Photogrammetry and Remote Sensing, Kyoto. Commission III*, pp. 148-159.

Dubois, G. (1997) "Spatial Interpolation Comparison 97: Foreword and Introduction". *Journal of Geographic Information and Decision Analysis*, Vol. 2, 2, 1-11

Falivene, O.; Cabrera, L.; Tolosana-Delgado, R. and Sáez, A. (2010) "Interpolation algorithm ranking using cross-validation and the role of smoothing effect. A coal zone example" *Computers & Geosciences* Vol. 36, 512–519

FGDC (1998) "Geospatial Positioning Accuracy Standards Part 3: National Standard for Spatial Data Accuracy', *Federal Geographic Data Committee, FGDC-STD-007.3-1998*, 28 pp

Franke, R. (1982) "Scattered Data Interpolation: Tests of Some Methods". *Mathematics of Computation*, Vol. 38, 157, 181-200

Goodin, W. R.; Mc.Rae, G. J. and Seinfeld, J. H., (1979). "Comparison of Interpolation Methods for Sparse Data: Application to Wind and Concentration Fields". *Journal of Applied Meteorology*, Vol. 18, 761-771

Kidner, D.; Dorey, M. and Smith, D. (1999) "What's the point? Interpolation and extrapolation with a regular grid DEM". *Geocomputation 99*. Accessed 2011-08-21 from www.geocomputation.org/1999/082/gc_082.htm

Lam, N. S. N., (1983) "Spatial Interpolation Methods: A Review". *The American Cartographer*, Vol. 10, 2, 129-149

Li, L. and Revesz, P. (2004) "Interpolation methods for spatio-temporal geographic data". *Computers, Environment and Urban Systems*, Vol. 28, 201–227