Efficacy of Interpolation Methods on Geomagnetic Field Measurements

Ryan S. Johnson*

University of Illinois Urbana-Champaign

Abstract

This project began with a large set of measurements of the strength of the Earth's magnetic field at various points in space taken by a satellite. This dataset had gaps ranging in size from a single dropped measurement to large swaths of space in which no measurements were taken. The goal of this project was to assess how accurately these gaps could be filled in using computationally cheap interpolation methods. Specifically, methods of interpolation involving radial basis functions were employed. Approximately 10,000 random data points were selected from the dataset. Then, this subset was randomly split into a training group (with probability 0.9) and a testing group (with probability 0.1). SciPy RBF interpolators with *linear*, *cubic*, and *thin plate* methods were trained on the training set and tested against the test set. Respectively, these methods had relative errors of approximately 28.5%, 17.4%, and 21.0%. When the data was not randomized (i.e., a cluster of 10,000 closely spaced data points), the relative errors dropped to 1.59%, 1.06%, and 1.11%. These results suggest that for small gaps in the data, cheaper interpolation methods perform relatively well when interpolating the data. However, when larger gaps are present, it is likely that more costly methods will need to be used in order to more accurately predict the values of missing datapoints.

Keywords: Geomagnetic field, satellite, radial basis functions, interpolation.

1 Introduction

This project began with a large set of measurements of the strength of the Earth's magnetic field at various points in space taken by a satellite. This dataset contained over 2GB of raw data about the satellite's position and the total magnetic flux measured at that position. In total, 2,609,922 of these datapoints were useable, meaning they had valid readings. A visualization of these points can be found in figure 1. Figure 2 has also been provided as it more clearly shows what the data looks like.

The spatial resolution at which measurements were taken varies from region to region, but of particular concern are the regions for which there are no measurements at all. Because remeasuring the magnetic flux for these regions is potentially costly, or even impossible, it would be valuable if interpolation methods could be used to estimate the values in these regions. However, it is not immediately clear what methods would be best to use. For the regions in space where a single datapoint is dropped, the data would be dense enough for a reasonable approximation to be made without much trouble at all. In many of these cases, even naïve methods like nearest neighbor would likely suffice in generating acceptable approximations. The larger gaps, however, pose a greater challenge.

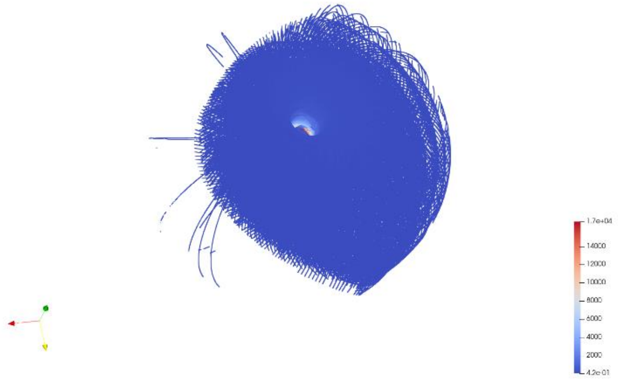

Figure 1: First 2,609,922 points of non-empty measurements.

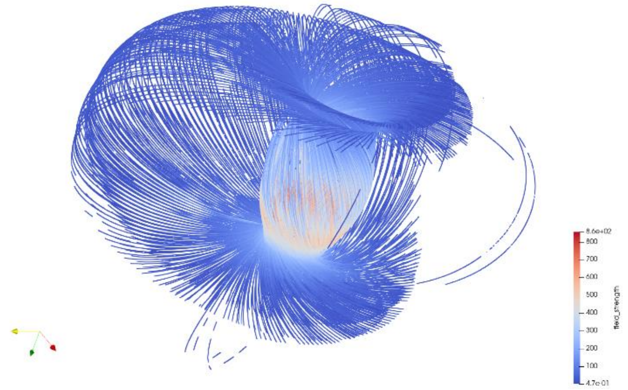

Figure 2: First 1,000,000 points of non-empty measurements.

In lower dimensional problem spaces, a variety of interpolation and approximation methods exist. The issue with this dataset is that it is a function over a 3D space with scalar values for outputs. This is yet another problem in the way of finding cheap methods of approximating missing data, as the increase in the number of dimensions increases the complexity of any model applied to the data. Ultimately, various radial basis functions were decided upon as the interpolation method.

2 Methods

In total, 2,609,922 datapoints were read from the dataset. Of these points, 10,000 were selected for training and testing. These test/train points were selected in two ways: in order and at random. Note that when saying data has been randomized, this means that the datapoints selected were random. No random values/points were generated, and all points came from the original dataset,

* rsj2@illinois.edu

whether they were randomly chosen, or simply chosen in order. This was done to simulate dense data with some missing datapoints (best case) and sparse data with many missing datapoints (worst case) respectively.

Then, each of the 10,000 points were randomly put into either the testing group or the training group. They were put into the testing group with probability 0.1 and into the training group with probability 0.9. See figure 4 for visualizations of the randomized and non-randomized testing and training groups. Additionally, see figure 3 for more statistics on the datasets themselves.

The primary interpolation method used was radial basis function interpolation, with radial basis functions of linear, cubic, and thin plate. These were chosen because the conditioning of these methods was not significantly impacted by the large number of datapoints to interpolate, unlike other basis functions like the Gaussian radial basis functions. These interpolants were applied to the training points and then used to interpolate values for the testing points. The interpolated values were then compared with the actual values to determine the error of each interpolation method. Error was calculated by taking the 2-norm of the vector obtained from the difference between the calculated values at the testing points and the actual values at the testing points. Relative error was taken with respect to the 2-norm of the vector of actual values.

Dataset	Standard Deviation	Average	Max	Min	Count
All points	309.2812603	114.0879195	16711.19193	0.41921	2609922
Randomized Training Points	302.347603	113.6478386	9283.88255	0.91971	9017
Randomized Testing Points	417.8625152	130.9998485	10724.31445	1.21213	983
Ordered Training Points	161.5459589	177.5763652	781.61614	2.30867	8974
Ordered Testing Points	160.4778377	175.8604536	752.58246	2.40973	1026

Figure 3: Info on all sets of data.

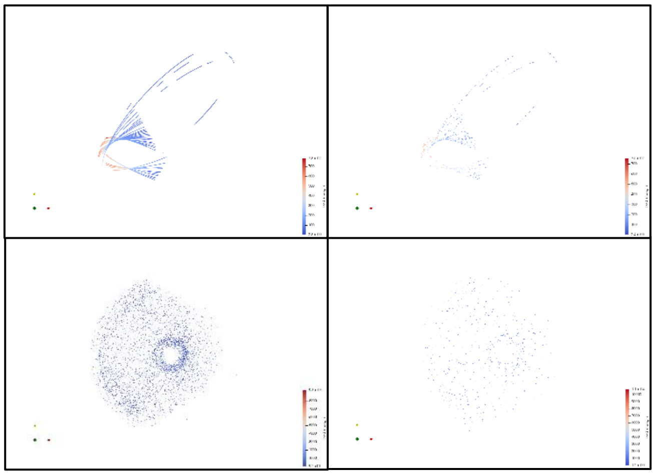

Figure 4: Non-randomized training data (top-left), non-randomized testing data (top right), randomized training data (bottom left), and randomized testing data (bottom right).

3 Results

All interpolation methods performed better on data that was not randomly chosen, and therefore more tightly clustered. The linear RBF had a relative error of ~1.6%, while both the cubic and thin plate RBF had relative errors of ~1.1%.

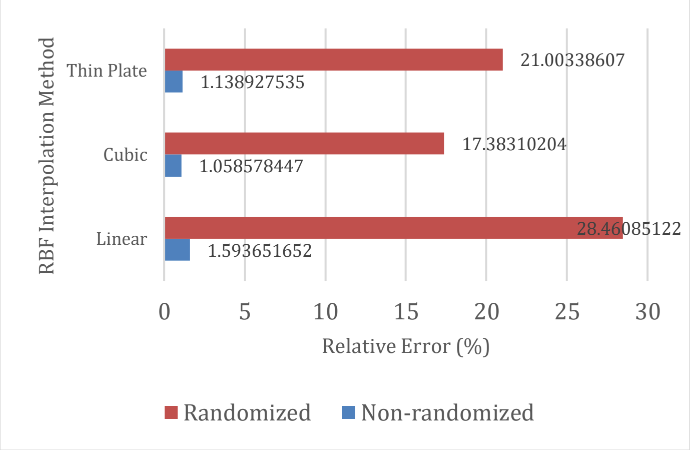

Table 1. Table 1

Figure 5: A graph of relative errors of the different methods applied to the randomized and non-randomized training and testing datasets.

However, when trained and tested over randomized data, all methods performed much worse. Linear RBF had the worst relative error at ~28.5%. The thin plate RBF performed better at ~21.0%, and the cubic RBF was best of all at ~17.4%.

4 Discussion

These results show that with tightly clustered/high resolution datapoints, one can expect to have a reasonably small error when using RBFs to interpolate the data. This is demonstrated by the small relative errors obtained when using ordered data. However, as soon as the data becomes sparse, the relative errors rise to unacceptable levels. This implies that simple interpolation methods may not be sufficient for filling in large gaps in data, as more context is needed for interpolating functions to have any hope of making a close approximation. Though there is still a chance that with enough context, it may be possible to interpolate large gaps in the dataset. More testing with deliberately designed gaps in the dataset, as well as a larger training set, would be required to determine this one way or another.

It should be noted that as the size of the random subset of data increases, one would expect that the relative errors should decrease. This is due to the fact that adding training datapoints will only increase the resolution of the data. However, due to hardware limitations, 10,000 points was the limit for this project. Trying to interpolate a significantly larger number of points caused the laptop used to run these tests to turn off without warning.

5 Conclusion

As previously discussed, it is unlikely that simple methods of interpolation can accurately predict the missing values of large gaps in datasets. It is not necessarily impossible, though, and more testing with faster, more capable hardware is required.

As for small gaps in datasets, so long as the data has a high resolution, cheap interpolation methods do a good job of predicting the correct value. This is in line with what one might expect.