To: Appalachian Basin Geothermal Play Fairway Analysis Group

From: Calvin Whealton and Jery Stedinger

Date: February 19, 2015

Subject: Outlier Identification Procedure

The Appalachian Basin Geothermal Play Fairway Analysis (AB-GPFA) must determine which algorithm should be used to identify outliers in the geospatial datasets. Outliers pose a problem for non-robust regression schemes because they would have high squared residuals. Many regression techniques seek to minimize the squared residuals, so an outlier can have undue influence on the results of the analysis.

This memo outlines the recommended outlier detection algorithm. Appendix 1 outlines the previous work on outlier algorithms for the NY and PA geothermal dataset. Appendix 2 illustrates the sensitivity of the final results to algorithm parameters over a reasonable range of values. Appendix 3 provides Monte Carlo type I error rates for different distributions type I error rates when the distribution parameters are known.

Outliers can be defined as "an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data" (Barnett and Lewis, 1994, p. 7). The following terms are defined below for use in the memo:

• *Global*: relating to the whole dataset, irrespective of location
• *Local*: relating to a subset of the data defined by a spatial relationship (e.g. 25 closest observations to the nearest point, points within 16 km of a point, etc.)
• *Sparse*: areas where there's insufficient data to evaluate a point to see if it is an outlier (e.g. only 4 local points if criterion is at least 25 local points)

The asymmetric boxplot algorithm used by Aguirre (2014) calculates upper and lower bounds from the sample quartiles, as defined in equations 1 and 2.

$$B_{lower} = Q_{0.25} - k(Q_{0.5} - Q_{0.25}) \quad [1]$$
$$B_{upper} = Q_{0.75} + k(Q_{0.75} - Q_{0.5}) \quad [2]$$

where

$Q_{0.25}$ is the lower quartile,
$Q_{0.75}$ is the upper quartile,
$Q_{0.5}$ is the median,
B_{lower} is the lower bound,
B_{upper} is the upper bound, and
k is a constant (standard value of 3).

Points outside the bounds are considered outliers. Aguirre (2014) calculates bounds both globally and locally, and only removed points which were both local and global outliers. She performed the global outlier test first, given the calculated values of B_{lower} and B_{upper} computed for the entire region. This greatly reduced the number of times that the local outlier computation was required.

We recommend that GPFA-AB group apply this asymmetric boxplot rule *only locally* with a value of the constant, $k = 3$. Additionally, the definition of local should be changed to the nearest 25 points provided that the points are within 16 km of the point being tested. If there are not 25 points within 16 km, then no outlier test is performed. Requiring points to be both local and global outliers will bias cold areas to be warmer and warm areas to be colder. Also, using 25 points allows a reasonable comparison with expected identification rates for several null distributions (see appendix 3). Using the Cornell NY and PA dataset of 8919 observations (Cornell University 2014) with Harrison corrected gradient and the recommended algorithm parameters, 6.8% (607 observations) were in sparse areas (fewer than 25 points within 16 km); 7.1% of the total dataset (629 observations) were removed as outliers (see appendix 2).

Encl.:

Appendix 1: Summary of Outlier Algorithms Used at Cornell
Appendix 2: Sensitivity Analysis of Recommended Algorithm
Appendix 3: Type I Error Rates
References

Appendix 1:
Summary of Outlier Algorithms Used at Cornell

Work at Cornell University has used outlier detection algorithms to remove potentially rogue observations before spatial regression. Rogue observations could have a high squared residual value, which can allow rogue observations to unduly influence the fit of non-robust spatial interpolation techniques.

Aguirre et al. (2013) broke the outlier analysis into global and local identification steps. For global analysis they considered a boxplot and asymmetric boxplot rules along with several other algorithms. Both of the boxplot-based algorithms use the quartiles of the data. The asymmetric boxplot rule was chosen because the global data seemed skewed and the asymmetric boxplot rule was robust to asymmetric data. They investigated several algorithms for local outlier analysis as well, but chose a method where the data was gridded in 16 km by 16 km blocks (other block sizes were also tested). In this early version of local outlier detection, local outliers were identified as more than three standard deviations from the block mean. The block standard deviation and the block mean were calculated for each block and only applied to points within that block**.** Although not explicitly mentioned, they do discuss that blocks with fewer than 20 points were not effective.

Aguirre (2014) continued to conduct global outlier analysis with the asymmetric boxplot rule. The final algorithm used the asymmetric boxplot rule for both local and global analysis. This allowed for a more robust local outlier detection algorithm because the standard deviation method used in her previous work was not robust. Only observations that were both local and global outliers and were in a box (32 km by 32 km) with 25 points were removed as outliers.

Aguirre's analyses are conservative because only points that are unusual both locally and globally are removed as outliers. If there is little signal (spatial trend) in the data then this is reasonable. Testing globally and then only testing global outliers to see if they are local outliers might reduce computation time, but the computational savings would be in the order of minutes. Given that the global outlier bounds differ by a factor of approximately 3, it seems that there could be signal of spatial variability in the data. If this were the case it would be best to use only a local analysis. Otherwise, "cold" areas will be biased warm because their coldest points will be both local and global outliers. Similarly, "warm" areas will be biased cold. Completing a local analysis would be robust to signal in the data, provided the local region is small enough. In the Cornell dataset about one quarter of the data comes from a single county in New York (Cornell University 2014). Given the large proportion of the data from a single county and the small variability within that one county, the global outlier test bounds could easily have been biased by this one county with Aguirre's methods. Choosing points that were spatially representative of the area in our dataset might have been more robust for determining global outlier bounds.

Appendix 2
Sensitivity Analysis of Recommended Algorithm

In the algorithm that we recommend there are essentially two parameters: the points used for a local test and the maximum radius one at which one can take points. To test the sensitivity of the algorithm to these two parameters we ran the algorithm on the Cornell dataset of 8,919 observations (Cornell University 2014). The variable tested was a gradient based on the Harrison correction (negative and past peak values used) and a uniform surface temperature of 9 °C. Figure 2.1 displays the results. Locations were projected from WGS84 into UTM 18N.

Figure 2.1 shows that for a large number of points and a small maximum radius, very little of the dataset can be evaluated as outliers (bottom right). As the radius increases and the points criterion decreases a greater fraction of the dataset can be tested as outliers. The increase in the percentage of data considered outliers grows from the bottom right because more of the dataset can be tested. However, eventually the increase will stop because 2.3% of the data is considered as outliers in a global test (when the local area is large the test converges to the global test). For instance, when the point's criterion was 1,000 and the radius criterion was 200 km, only 4.4% of the data was considered outliers. Note in the upper right hand portion of the graph the proportion of data removed as outliers is approximately twice what one would expect for normal data (6-8% versus 3% for normal, see table 3.1). The percentage of outliers identified is closer to what one would expect from a fairly fat-tailed kurtotic Student t distribution.

Based on these results it seems reasonable to choose the points criterion as 25 and the radius criterion as 16 km. This will be close to the parameters used by Aguirre (2014), except in her algorithm the grid spacing was 32 km. When the 25 points and 16 km radius were applied to the test dataset, this left 6.8% (607 observations) in sparse areas. In total, 7.1% of the whole dataset (629 observations) were removed as outliers. It is likely that some of the data in sparse areas would be omitted for other reasons, including not enough points in the county or the points are outside our area of interest.

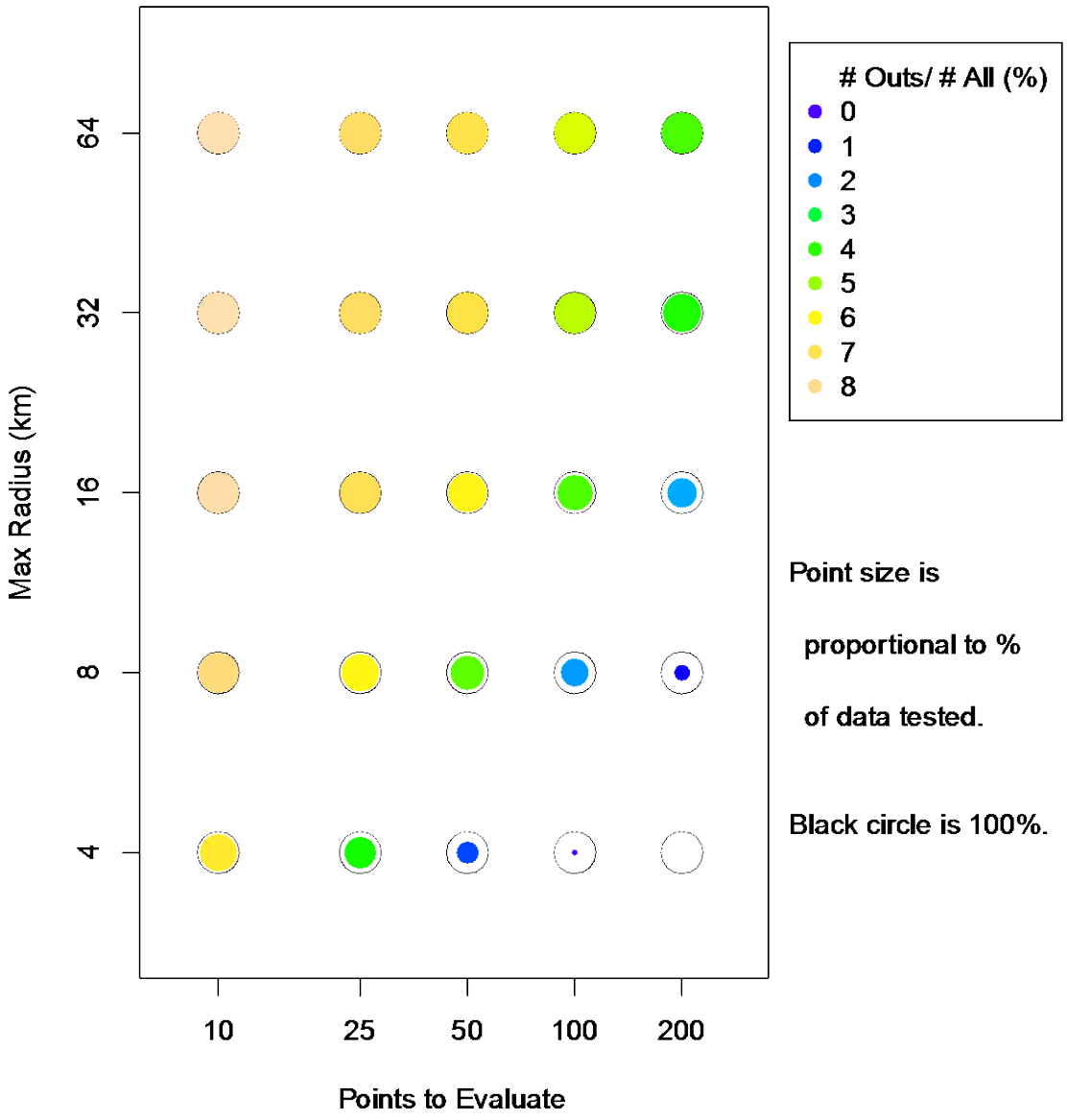

Figure 2.1: Plot showing the impact of the required number of points to evaluate a local outlier (horizontal axis) and the maximum radius at which points can be taken (vertical axis) on the percentage of the data set tested for local outliers (size of symbols) and the percentage of points considered outliers (color of symbol). The percentage of points considered outliers is relative to the number of points in the original dataset. The black circles represent 100% of the data being tested by the outlier algorithm. The recommended algorithm uses the 25 closest points within a maximum radius of 16 km.

Appendix 3
Type I Error Rates

The equations 1 and 2 given in the memo depend on k, which is a constant multiplied by the median-to-upper quartile or lower quartile-to-median range. Typically, one chooses an outlier criterion based on a specified type I error. Type I error is the probability that one incorrectly rejects the null hypothesis. In this example the null hypothesis is that the data is distributed according to the distribution listed. The type I error will be identifying a point as an outlier even though it is drawn from the specified distribution. Type I error rates in table 3.1 were calculated based on Monte Carlo calculation from 100,000 replicates of sample size 25.

For the recommended value of $k = 3$, if the data is normally distributed one would expect to identify about 3% (see table 3.1) of the data as outliers. Thicker tailed distributions, such as the Student t, will have higher identification rates. Thin-tailed Beta(1,1) (uniform distribution) and Beta(2,2) have high type I error rates in table 3.1 compared to the values in table 3.2. For example, in table 3.1 the type I error for Beta(2,2) with $k = 1.5$ is 8.29%, but if the distribution parameters were known exactly the type I error would be 2.49% as given in table 3.2. Beta(1,1) shows large differences between the two cases.

Table 3.1. Type I error (%) for asymmetric boxplot test based on 100,000 replicates of sample size 25. Beta(1,1) is the uniform distribution. The argument for the Student t distribution is the shape parameter (also referred to as the degrees-of-freedom), which controls the thickness of the tails. Student t (∞) is the normal distribution. Upper and lower bounds used to define outliers are based on equations 1 and 2 in the memo.

				Distribution				
k	**Normal**	**Beta (1,1)**	**Beta (2,2)**	**Student t (2)**	**Student t (4)**	**Student t (6)**	**Student t (8)**	**Student t (10)**
1.0	19.76	10.79	15.84	25.25	22.63	21.71	21.22	20.94
1.5	12.36	4.67	8.29	19.20	15.95	14.78	14.19	13.82
2.0	7.86	2.14	4.26	14.97	11.44	10.24	9.62	9.23
2.5	5.00	1.06	2.42	12.00	8.36	7.21	6.61	6.27
3.0	3.26	0.56	1.36	9.81	6.32	5.16	4.64	4.34
3.5	2.16	0.31	0.80	8.10	4.79	3.77	3.31	3.06
4.0	1.45	0.18	0.49	6.84	3.68	2.79	2.40	2.18
4.5	1.00	0.11	0.32	5.83	2.91	2.11	1.77	1.57
5.0	0.70	0.06	0.21	5.02	2.32	1.61	1.32	1.17

Table 3.2. Type I error (%) for asymmetric boxplot test based on perfect knowledge of parameters (large sample). Beta(1,1) is the uniform distribution. The argument for the Student t distribution is the shape parameter (also referred to as the degrees-of-freedom), which controls the thickness of the tails. Student t (∞) is the normal distribution. Upper and lower bounds used to define outliers are based on equations 1 and 2 in the memo and the quantiles are calculated from the population distribution.

	Distribution							
k	**Normal**	**Beta (1,1)**	**Beta (2,2)**	**Student t (2)**	**Student t (4)**	**Student t (6)**	**Student t (8)**	**Student t (10)**
1.0	17.73	0.00	12.57	24.41	21.26	20.13	19.54	19.19
1.5	9.18	0.00	2.49	17.80	13.77	12.30	11.54	11.08
2.0	4.30	0.00	0.00	13.40	9.04	7.48	6.69	6.21
2.5	1.82	0.00	0.00	10.37	6.05	4.58	3.86	3.43
3.0	0.70	0.00	0.00	8.23	4.14	2.84	2.23	1.88
3.5	0.24	0.00	0.00	6.67	2.90	1.79	1.30	1.03
4.0	0.07	0.00	0.00	5.51	2.08	1.15	0.77	0.57
4.5	0.02	0.00	0.00	4.62	1.52	0.76	0.46	0.32
5.0	0.01	0.00	0.00	3.92	1.13	0.51	0.28	0.18

References

Aguirre, G. A., Stedinger, J. R. and Tester, J. W. (2013). Geothermal Resource Assessment: A Case Study of Spatial Variability and Uncertainty Analysis for the State of New York and Pennsylvania. *Procedings: 38th Workshop on Geothermal Reservoir Engineering*. Stanford University.

Aguirre, G. A. (2014). *Geothermal Resource Assessment: A Case Study of Spatial Variability and Uncertainty Analysis for the States of New York and Pennsylvania*. Master's Thesis. Environmental and Water Resources Systems Engineering, School of Civil and Environmental Engineering, Cornell University.

Barnett, V. and Lewis, T. (1994). *Outliers in Statistical Data.* 3rd ed. John Wiley and Sons: New York.

Cornell University (2014). Cornell University Heat Flow Database (NY and PA). Southern Methodist University Geothermal Laboratory. http://geothermal.smu.edu/static/DownloadFilesButtonPage.htm (Accessed 16 June 2014).