

# Identifying Outliers In Multivariate Spatial Data: An Undergraduate's Perspective

Anthony Franklin, Eric B. Howington, and Keshav Jagannathan  
Department of Mathematics and Statistics, Coastal Carolina University

## Introduction:

Our analysis of the NASA data focuses on using statistical tools accessible to undergraduates to detect unusual data observations. We created a robust multivariate regression model to simultaneously model six response variables. The matrix of residuals from the robust model was treated as a multivariate data set and explored for outliers and atypical observations. We limited our region of interest to the Southeastern United States.

## NASA Atmospheric Data:

Predictor Variables:  
Year & Month  
Latitude & Longitude  
Elevation (in meters)

Response Variables: All responses are monthly means.  
Surface Temperature (Kelvin)  
Near-Surface Air Temperature (Kelvin)  
Ozone Abundance (Dobson units)  
Low Cloud Coverage (percent)  
Medium Cloud Coverage (percent)  
High Cloud Coverage (percent)

\*Pressure was also present in the original data, but not used in the analysis because of the limited variability in the pressure measurements.

## Methodology:

After formatting the data, we used MCD regression to fit a robust multivariate regression model using Ozone, Air Temperature, Surface Temperature, Cloud Low, Cloud Mid, and Cloud High as response variables. We wanted both spatial and temporal effects to be present in the model. We chose Latitude, Longitude, Latitude\*Longitude interaction, Elevation, and Elevation<sup>2</sup> as "spatial" predictor variables. We used year and indicator variables for months as our "temporal" predictor variables.

We then used robust distances to analyze the matrix of residual vectors for outlying observations. Any observation with a robust distance beyond a cut-off point was classified as an outlier.

Finally, we deleted the outlying observations from the dataset and analyzed the clean data using regular multivariate regression.

## MCD Regression:

Minimum Covariance Determinant regression is a robust method for multivariate regression based on robust estimation of the joint location and scatter matrix of the explanatory and response variables (Rousseeuw et al 2004).

The multivariate regression model is given as  $\mathbf{y} = \mathbf{B}'\mathbf{x} + \alpha + \epsilon$

Denote the location of the joint  $(\mathbf{x}, \mathbf{y})$  variables by  $\boldsymbol{\mu}$  and their scatter matrix by  $\boldsymbol{\Sigma}$ , then partition them as follows:

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sum_{xx} & \sum_{xy} \\ \sum_{yx} & \sum_{yy} \end{pmatrix}$$

Johnson and Wichern (2002) show that the least squares estimates of the model parameters can be written as

$$\hat{\mathbf{B}} = \hat{\boldsymbol{\Sigma}}_{xx}^{-1} \hat{\boldsymbol{\Sigma}}_{xy} \quad \text{and} \quad \hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\mu}}_y - \hat{\mathbf{B}}' \hat{\boldsymbol{\mu}}_x$$

MCD regression replaces the usual estimates of the mean vector and covariance matrix with robust estimates derived from the Minimum Covariance Determinant estimates of location and scatter, producing robust estimates of the regression coefficients.

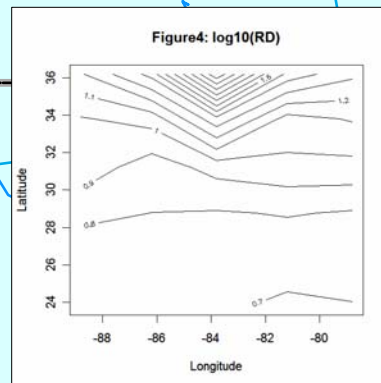
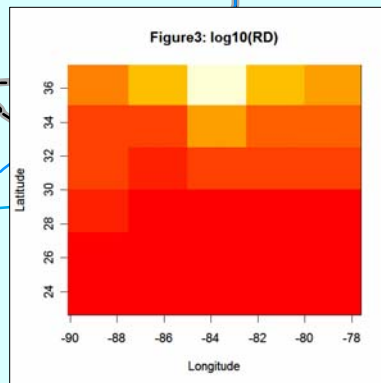
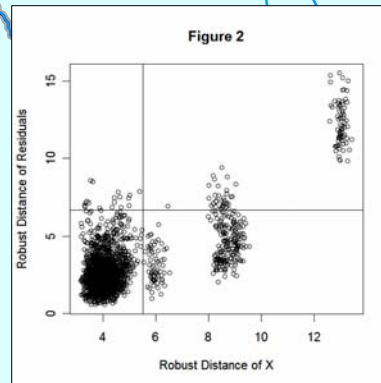
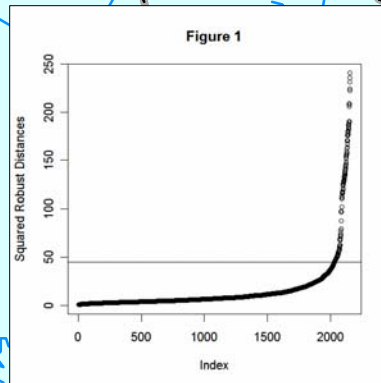
## Identifying Outliers:

We treat the matrix of residuals from the robust regression as its own multivariate dataset and analyze it using robust distances. Again, we use the MCD to obtain robust estimates of multivariate location and scatter. Then we define the squared robust distances as

$$RD_i^2 = (\mathbf{e}_i - \boldsymbol{\mu}_{MCD})' \boldsymbol{\Sigma}_{MCD} (\mathbf{e}_i - \boldsymbol{\mu}_{MCD})$$

where  $\boldsymbol{\mu}_{MCD}$  is the robust estimate of location and  $\boldsymbol{\Sigma}_{MCD}$  is the robust estimate of scatter (Rousseeuw and Van Zomeren 1990).

If  $RD_i^2 > 45$ , then the observation is classified as an outlier and flagged for further consideration. Note: This cut-off was obtained by subjective visual inspection of figure 1 below. An objective cut-off value could be used, as noted in Hardin and Rocke (2005).



## Results:

There were 124 observations identified as outliers by our procedure. These observations were removed from the dataset and a regular multivariate regression analysis was conducted.

It is apparent from figure 2 that there were numerous observations outlying in both X and Y "directions." Figures 3 and 4 indicate that a majority of the outlying observations were all from the same area. In fact, 87 of the observations that were identified as outlying were from the central and east Tennessee regions. More than likely, this is an indication of model lack-of-fit for this particular area of the southeast.

## Conclusions:

**Surface Temperature:** There were few significant spatial or temporal factors. Elevation was a significant effect, along with seasonal variation.

**Air Temperature:** Air temperature was highly affected by spatial factors. All spatial variables considered were significant. In addition, year had a statistically significant positive coefficient, giving some support to the theory of global warming.

**Ozone:** There were spatial and temporal trends in the ozone measurements. Latitude, longitude, and elevation were all significant in the final model. Ozone had a slightly increasing trend over the study years for the southeastern US.

**Low Cloud Coverage:** This variable was highly influenced by spatial, including elevation, and temporal factors. Along with seasonal effects, the southeast experienced a decreasing trend in low cloud coverage over the six years of the study.

**Mid Cloud Coverage:** This variable was not significantly influenced by any spatial predictors except elevation. Mid cloud coverage experienced a significant decreasing trend over the years of the study as well.

**High Cloud Coverage:** This variables appeared to vary significantly over both latitudes and longitudes. Again, there was a decreasing trend in high cloud coverage over time.

## Discussion:

Clearly, a much more sophisticated analysis, incorporating specialized spatial and time-series analysis methodology is desirable.

However, an undergraduate armed with knowledge of regression, model-building, and basic multivariate analysis can produce an interesting and insightful analysis of quite complex datasets.

## References:

- Hardin, J. and Rocke, D. M. (2005), "The Distribution of Robust Distances," *Journal of Computational and Graphical Statistics*, 14, 928-946.
- Johnson, R. A. and Wichern, D. W. (2002), *Applied Multivariate Statistical Analysis* (5<sup>th</sup> ed.), Upper Saddle River, NJ: Prentice-Hall.
- Rousseeuw, P. J., Van Driessen, K., Van Aelst, S., and Agullo, J. (2004), "Robust Multivariate Regression," *Technometrics*, 46, 293-305.
- Rousseeuw, P. J. and Van Zomeren, B. C. (1990), "Unmasking Multivariate Outliers and Leverage Points," *Journal of the American Statistical Association*, 85, 633-639.

## Acknowledgements:

The authors thank Coastal Carolina University for the Academic Enhancement Grant that partially funded this project.

## Bios:

**Anthony Franklin** is a senior majoring in mathematics and minoring in statistics at Coastal Carolina University. The Simpsonville, South Carolina native plays football for CCU and works as a student minister at Live Oak Church in Myrtle Beach, South Carolina.

**Eric B. Howington** is an assistant professor in the Department of Mathematics and Statistics at Coastal Carolina University.

**Keshav Jagannathan** is an assistant professor in the Department of Mathematics and Statistics at Coastal Carolina University.

Gulf of