

University of Montana

## ScholarWorks at University of Montana

---

Graduate Student Theses, Dissertations, &  
Professional Papers

Graduate School

---

2004

### Robust variogram estimation via rank correlation and median absolute deviation correlation

Isaac C. Grenfell  
*The University of Montana*

Follow this and additional works at: <https://scholarworks.umt.edu/etd>

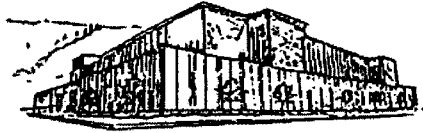
**Let us know how access to this document benefits you.**

---

#### Recommended Citation

Grenfell, Isaac C., "Robust variogram estimation via rank correlation and median absolute deviation correlation" (2004). *Graduate Student Theses, Dissertations, & Professional Papers*. 8326.  
<https://scholarworks.umt.edu/etd/8326>

This Thesis is brought to you for free and open access by the Graduate School at ScholarWorks at University of Montana. It has been accepted for inclusion in Graduate Student Theses, Dissertations, & Professional Papers by an authorized administrator of ScholarWorks at University of Montana. For more information, please contact [scholarworks@mso.umt.edu](mailto:scholarworks@mso.umt.edu).



Maureen and Mike  
MANSFIELD LIBRARY

The University of  
**Montana**

---

Permission is granted by the author to reproduce this material in its entirety, provided that this material is used for scholarly purposes and is properly cited in published works and reports.

**\*\*Please check "Yes" or "No" and provide signature\*\***

Yes, I grant permission

\_\_\_\_\_

No, I do not grant permission

\_\_\_\_\_

Author's Signature: \_\_\_\_\_

*[Handwritten signature]*

Date: \_\_\_\_\_

*5/18/04*

Any copying for commercial purposes or financial gain may be undertaken only with the author's explicit consent.

---



ROBUST VARIOGRAM ESTIMATION VIA RANK  
CORRELATION AND MEDIAN ABSOLUTE DEVIATION  
CORRELATION

By

Isaac C Grenfell

B.A. University of Montana, 2002

SUBMITTED IN PARTIAL FULFILLMENT OF THE  
REQUIREMENTS FOR THE DEGREE OF  
MASTER OF ARTS

AT

DEPARTMENT OF MATHEMATICAL SCIENCES  
THE UNIVERSITY OF MONTANA

MISSOULA, MT

MAY 2004

  
Chairperson

  
Dean, Graduate School

26 May 2004  
Date

UMI Number: EP39127

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



UMI EP39127

Published by ProQuest LLC (2013). Copyright in the Dissertation held by the Author.

Microform Edition © ProQuest LLC.

All rights reserved. This work is protected against unauthorized copying under Title 17, United States Code




ProQuest LLC.  
789 East Eisenhower Parkway  
P.O. Box 1346  
Ann Arbor, MI 48106 - 1346

# Abstract

Grenfell, Isaac C. M.A., May 2004

Mathematics

Robust Variogram Estimation via Rank Correlation and Median Absolute Deviation Correlation

Director: Rudy Gideon 

Variogram estimation via classical methods is highly sensitive to the presence of outliers in the data, as well as deviation from normality in the sampling population. To address this problem, we propose two estimation methods: one relying on the Greatest Deviation Correlation Coefficient  $r_{gd}$ , and the second based on estimates of the scale parameters and covariance through median absolute deviation (*MAD*). The effectiveness of these two methods is examined with computer simulation, and the methods are compared with both classical variogram estimation, and a separate robust method.

# Acknowledgements

I would like to thank Rudy Gideon, my supervisor, for providing support and motivation in my research endeavor. I offer thanks to Jon Graham for introducing me to the field of spatial statistics and teaching statistics in a rigorous manner. I also thank Doug Dalenberg and Brian Steele for reading my paper and offering their thoughts.

I thank Pat Andrews and Mark Finney for their support and ideas, as well as being examples of practical scientists.

I also thank all the professors at the University of Montana for exposing me to their work and challenging me.

Finally, I thank my parents for making me, Phil for loving my mom, and Richard, Chip, Brent, Tiffany, Jocelyn, and Vic for their friendship.

Missoula, MT  
May 7, 2004

Isaac C Grenfell

# Table of Contents

<b>Abstract</b>	<b>ii</b>
<b>Acknowledgements</b>	<b>iii</b>
<b>Table of Contents</b>	<b>iv</b>
<b>List of Figures</b>	<b>v</b>
<b>Introduction</b>	<b>1</b>
<b>Variogram Estimation</b>	<b>2</b>
0.1 The Variogram . . . . .	2
0.2 Greatest Deviation Correlation Coefficient . . . . .	3
0.2.1 Introduction . . . . .	3
0.2.2 Method of Estimation . . . . .	4
0.3 Median Absolute Deviation . . . . .	5
0.3.1 Introduction . . . . .	5
0.3.2 Estimation with <i>MAD</i> . . . . .	5
0.4 Simulation . . . . .	8
0.5 Results and Discussion . . . . .	9
0.6 Conclusion . . . . .	12
<b>Bibliography</b>	<b>16</b>



# List of Figures

1	The 75th Quantiles of the T distributions . . . . .	7
2	Cases 1 ( $\epsilon = 0, \sigma = 1$ ) and 2 ( $\epsilon = 0.10, \sigma = 5$ ) . . . . .	9
3	Cases 3 ( $\epsilon = 0.20, \sigma = 5$ ) and 4 ( $\epsilon = 0.30, \sigma = 5$ ) . . . . .	10
4	Cases 5 ( $\epsilon = 0.10, \sigma = 10$ ) and 6( $\epsilon = 0.10, \sigma = 20$ ) . . . . .	11

# Introduction

When conducting analysis of spatial data, a vital tool for both understanding the structure of spatial autocorrelation and the determination of kriging weights is the variogram. The most commonly used technique to estimate the variogram relies on classical statistical methods, which are very sensitive to the presence of outliers or influential observations in the data set, as well as to departure from the assumption of normality of the data. Either problem can have an ill effect on kriging weights, and therefore, kriging predictions. With real data, say from satellite imagery or field observations, it is not uncommon to have in excess of 10-15% of the observations as outliers. If an observation can be determined to be an outlier with certainty, it is best to omit the record. If it is in doubt as to whether or not an observation is an outlier, one shouldn't remove an observation without justification. This is why we seek a method that still yields meaningful results whether or not outliers are present.

# Variogram Estimation

## 0.1 The Variogram

The objective of the variogram is to measure the association of a response variable at different distances, and possibly directions. The model of the variogram assumes intrinsic stationarity. That is, for a study region  $D$ ,  $E(Z_i) = E(Z_j)$ , and  $\text{Var}(Z_i - Z_j) = 2\gamma(\mathbf{h})$  for all points  $i, j \in D$ . The variogram function  $2\gamma(\mathbf{h})$  is a function of the lag distance  $\mathbf{h}$  that separates the points. Of interest in analyzing the variogram are three parameters: the range, sill, and nugget. The range is the lag distance at which spatial independence occurs. The sill is the value the variogram takes beyond the range. The nugget is the value the variogram takes at lag zero. We typically look at the semivariogram, which is just one half the variogram and is often referred to as the variogram.

Our goal is to estimate the function  $\gamma(\mathbf{h}) = \frac{1}{2} \text{Var}(Z_i - Z_j)$  for all points  $i$  and  $j$  separated by a distance vector  $\mathbf{h}$ . The classical estimate of the variogram (referred to as Matheron's method in other texts) is just:

$$\hat{\gamma}_c(\mathbf{h}) = \frac{1}{2N(\mathbf{h})} \sum_{(i,j)|h_{i,j}=\mathbf{h}} (z_i - z_j)^2 \quad (0.1.1)$$

We do not consider the method proposed by Cressie and Hawkins [1], as its robustness has been questioned [2]. Now, if we expand the variance function, we get:

$$\begin{aligned}
\gamma(\mathbf{h}) &= \frac{1}{2} \text{Var}(Z_i - Z_j) \\
&= \frac{1}{2}(\sigma_{z_i}^2 + \sigma_{z_j}^2 - 2 \text{Cov}(z_i, z_j)) \\
&= \frac{1}{2}(\sigma_{z_i}^2 + \sigma_{z_j}^2 - 2\sigma_{z_i}\sigma_{z_j}\rho).
\end{aligned} \tag{0.1.2}$$

where  $\sigma_{z_i}$  and  $\sigma_{z_j}$  are the standard deviations for  $Z_i$  and  $Z_j$ , and  $\rho$  is the correlation coefficient between them. In order to obtain an alternative estimate of  $\gamma(\mathbf{h})$ , we can simply obtain estimates of  $\sigma_{z_i}$ ,  $\sigma_{z_j}$ , and  $\rho$ , and use those estimates to estimate  $\gamma(\mathbf{h})$  by computing:

$$\hat{\gamma} = \frac{1}{2}(\hat{\sigma}_{z_i}^2 + \hat{\sigma}_{z_j}^2 - 2\hat{\sigma}_{z_i}\hat{\sigma}_{z_j}\hat{\rho}) \tag{0.1.3}$$

## 0.2 Greatest Deviation Correlation Coefficient

### 0.2.1 Introduction

In this section we develop a method to estimate the variogram using  $r_{gd}$ . Gideon and Hollister [3] introduced a correlation coefficient derived from the ranks of a data set, rather than the data themselves. This correlation coefficient is based on the idea of greatest deviation, where we consider only the ranks of the data and use the distribution of those ranks to obtain a correlation coefficient, which we call  $r_{gd}$ . It has several desirable characteristics. First, if we have normally distributed data,  $r_{gd}$  usually gives similar estimates to those obtained by classical techniques. If, however, we are working with a symmetric distribution such as a Student's T or a Cauchy distribution, rather than a Gaussian distribution, a good estimate of the scale parameter is attainable through  $r_{gd}$ , whereas classical methods can break down as the distribution becomes heavier in the tails. Moreover, if the data contains outliers or extreme observations, classical methods tend to give excessive weight to those observations, while methods based on  $r_{gd}$  tend to be more resistant to the influence

of those observations. All of these characteristics make  $r_{gd}$  a desirable tool to estimate the variogram.

### 0.2.2 Method of Estimation

Let  $(X, Y)$  be bivariate normal random variables. Consider a random sample  $\{(x_i, y_i)\}_{i=1}^n$ . Sort the data by the  $x_i$ 's and let  $p_i$  be the rank of the corresponding  $y$  value. Let  $I(A)$  be the indicator function of the event  $(A)$  ( $I(A) = 1$  if  $A$  occurs,  $I(A) = 0$  if  $A$  if not), and define  $d_i$  and  $d_i^c$  in the following way:

$$d_i = \sum_{j=1}^i I(i < p_j) \quad (0.2.1)$$

$$d_i^c = \sum_{j=1}^i I(i < n + 1 - p_j) \quad (0.2.2)$$

We are now ready to define  $r_{gd}$  as:

$$r_{gd} = \frac{\max_{i=1\dots N}(d_i^c) - \max_{i=1\dots N}(d_i)}{\lfloor \frac{N}{2} \rfloor} \quad (0.2.3)$$

From Gideon [3], we get an estimate of  $\rho$  through the following transformation:

$$\hat{\rho}_{gd} = \sin\left(\frac{\pi}{2} r_{gd}\right) \quad (0.2.4)$$

Now, if  $(X, Y)$  has an elliptical probability density function, we can obtain estimates of  $\sigma_x$  and  $\sigma_y$  in the following way: let  $\{k_i\}_{i=1}^n$  be the quantiles of the standard normal distribution, and let  $y_{(i)}$  be the ordered observations on  $Y$ . While there is an assumption of normality intrinsic in this method, it is not sensitive to violation of this assumption when the true distribution is symmetric [4]. From Sheng [5], our estimate of scale is the value of  $s$  that solves this equation:

$$r_{gd}(k_i, y_{(i)}) - sk_i = 0 \quad (0.2.5)$$

Using (0.2.4) together with (0.2.5), we now have a method to estimate  $\gamma(\cdot)$ ,  $\mathbf{h}$  which we will call  $\hat{\gamma}_{gd}$ :

$$\hat{\gamma}_{gd}(h) = \frac{1}{2}(s_{z_i}^2 + s_{z_j}^2 - 2s_{z_i}s_{z_j}\hat{\rho}_{gd}) \quad (0.2.6)$$

## 0.3 Median Absolute Deviation

### 0.3.1 Introduction

The *MAD* estimate has an intuitive appeal. It uses the median of the absolute value of the deviations from the median as an estimate of scale for a sample, and we can also obtain a direct estimate of covariance. The procedure follows from Gideon [6]. In order to achieve unbiasedness, we must rescale our estimate of scale by the 75th quantile of the hypothesized distribution. This appears to have the drawback of imposing a distributional assumption on the data, however if we look at what the actual values of these quantiles are for the *T* distributions, they converge quite quickly to that of the standard normal 75th quantile of 0.6745.

### 0.3.2 Estimation with *MAD*

Consider again the bivariate normal random variables  $(X, Y)$ , with random sample  $\{(x_i, y_i)\}_{i=1}^n$ . We define  $MAD_x$  as  $\text{med}(|x_i - \text{med}(x_i)|)$ , and similarly for  $MAD_y$ . Using the *MAD* function, we can obtain a direct estimate of the covariance function. First, we need some results for random variables. Let  $Z_x = \frac{X - \mu_x}{\sigma_x}$  and  $Z_y = \frac{Y - \mu_y}{\sigma_y}$ . Since  $X$  and  $Y$  are normally distributed,  $P(|Z_x| \leq 0.6745) = 0.5000$ . Thus,  $MAD_{Z_x} = \text{med}|Z_x| = 0.6745$ . This implies that  $\frac{MAD_x}{0.6745}$  and  $\frac{MAD_y}{0.6745}$  will be unbiased estimates for  $\sigma_x$  and  $\sigma_y$  [6]. Note that,  $\text{Var}(Z_x + Z_y) = 2(1 + \rho)$ , and  $\text{Var}(Z_x - Z_y) = 2(1 - \rho)$ . This implies that

$$\text{med}|Z_x + Z_y| = 0.6745\sigma_{Z_x+Z_y} = 0.6745(2(1 + \rho))^{1/2} \quad (0.3.1)$$

$$\text{med } |Z_x - Z_y| = 0.6745\sigma_{Z_x+Z_y} = 0.6745(2(1 - \rho))^{1/2} \quad (0.3.2)$$

Now, define

$$T^+ = X - \text{med } X + (Y - \text{med } Y) \quad (0.3.3)$$

$$T^- = X - \text{med } X - (Y - \text{med } Y) \quad (0.3.4)$$

and let  $t^+$  and  $t^-$  be the corresponding estimates. Suppose  $(X, Y)$  is bivariate normal with parameters  $\sigma_x$ ,  $\sigma_y$ , and  $\rho$ . *Claim.* The covariance function  $\text{Cov}(X, Y)$  can be represented as:

$$\text{Cov}(X, Y) = \frac{\text{med}^2(|T^+|) - \text{med}^2(|T^-|)}{4(0.6745)^2} \quad (0.3.5)$$

*Proof.* Let  $\rho^* = \frac{\text{med}^2(|T^+|) - \text{med}^2(|T^-|)}{4(0.6745)^2}$ . We now show that  $\rho^* = \text{Cov}(X, Y)$ .

$$\rho^* = \frac{\text{med}^2(|T^+|) - \text{med}^2(|T^-|)}{4(0.6745)^2}$$

Using (0.3.3) and (0.3.4), this can be represented as:

$$\rho^* = \frac{\text{med}^2(|X - \text{med } X + Y - \text{med } Y|)}{4(0.6745)^2} - \frac{\text{med}^2(|X - \text{med } X - Y + \text{med } Y|)}{4(0.6745)^2}$$

Since  $\mu_x = \text{med}(X)$ ,  $Z_x\sigma_x = X - \text{med } X$ :

$$\rho^* = \frac{1}{4(0.6745)^2} (\text{med}^2(|Z_x\sigma_x + Z_y\sigma_y|) - \text{med}^2(|Z_x\sigma_x - Z_y\sigma_y|))$$

Using (0.3.1) and (0.3.2),

$$\begin{aligned} \rho^* &= \frac{1}{4(0.6745)^2} \left( \left( .6745 \sqrt{\text{Var}(Z_x\sigma_x + Z_y\sigma_y)} \right)^2 - \left( .6745 \sqrt{\text{Var}(Z_x\sigma_x - Z_y\sigma_y)} \right)^2 \right) \\ &= \frac{1}{4} \left( \left( \sqrt{\sigma_x^2 + \sigma_y^2 + 2 \text{Cov}(Z_x\sigma_x, Z_y\sigma_y)} \right)^2 - \left( \sqrt{\sigma_x^2 + \sigma_y^2 - 2 \text{Cov}(Z_x\sigma_x, Z_y\sigma_y)} \right)^2 \right) \\ &= \frac{1}{4} (4\sigma_x\sigma_y \text{Cov}(Z_x, Z_y)) = \sigma_x\sigma_y\rho \\ &= \text{Cov}(X, Y). \end{aligned}$$

□

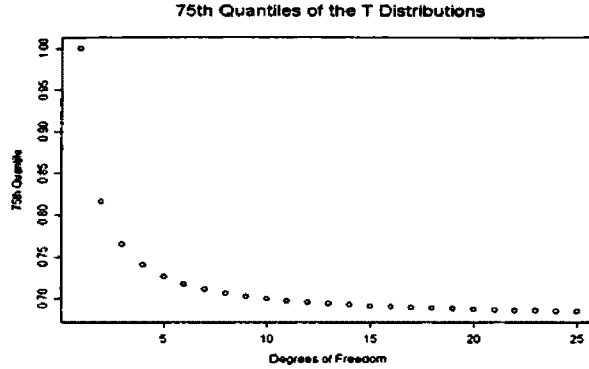


Figure 1: The 75th Quantiles of the T distributions

If  $(X, Y)$  is not normal, but still symmetric, our estimates may be biased. To see this, consider the family of T distributions. In figure 1, we see that, using 0.6745 as an estimate of the 75th quantile, we will not be far off if the degrees of freedom exceed 3 or 4. Also, the bias we observe is far smaller than if we simply took the sample standard deviation via classical methods, since the heavy-tailed distributions will produce more extreme observations. In general,  $MAD$  tends to be more robust given the presence of such observations than the sample standard deviation function.

We can now use the following sample analog as an estimate of covariance:

$$\widehat{\text{Cov}}(x, y) = \frac{\text{med}^2(|t^+|) - \text{med}^2(|t^-|)}{4(0.6745)^2} \quad (0.3.6)$$

Combining this with our estimates of scale, we get the following  $MAD$  estimate of the variogram  $\gamma(\mathbf{h})$ , which we will denote  $\hat{\gamma}_{mad}(\mathbf{h})$ :

$$\hat{\gamma}_{mad}(\mathbf{h}) = \frac{1}{2} \left( \left( \frac{MAD_{z_i}}{0.6745} \right)^2 + \left( \frac{MAD_{z_j}}{0.6745} \right)^2 - 2\widehat{\text{Cov}}(z_i, z_j) \right) \quad (0.3.7)$$



## 0.4 Simulation

To evaluate the performance of these variogram estimation methods, we perform a procedure similar to that done by Genton [2]. First, we create a spherical variogram function and use the statistical program S-plus to generate a set of 1-dimensional spatial data of 400 observations with a correlation structure imposed by this variogram. For our simulations, the variogram has a nugget of 1, a sill of 5, and a range of 7 (these values were selected arbitrarily). We then use the methods to estimate this clean data, with no imposed noise. Next, we begin taking sub-samples of this data and replace it with  $100 * \epsilon\%$  of the observations with observations from an independent  $N(0, \sigma^2)$  distribution. We perform this for the following values of  $\sigma$  and  $\epsilon$ :

Simulation Conditions		
<i>Case</i>	$\epsilon$	$\sigma$
I	0	1
II	0.1	5
III	0.2	5
IV	0.3	5
V	0.1	10
VI	0.1	20

Once the variograms for these different cases are obtained, we use a nonlinear-least squares method from S-plus to get estimates of the sill range and nugget. This process was repeated 40 times, yielding a table of the mean and median of the difference between the estimates and parameters via the different methods (measuring the bias of the estimates), as well as the mean squared error (MSE) and median absolute error (MAE) of the estimates (measuring the variability of the estimates).

## 0.5 Results and Discussion

For one set of simulations, we get the following plots, with the solid line representing the variogram function used to generate the data, and the points representing the estimate of the variogram at the given lag:

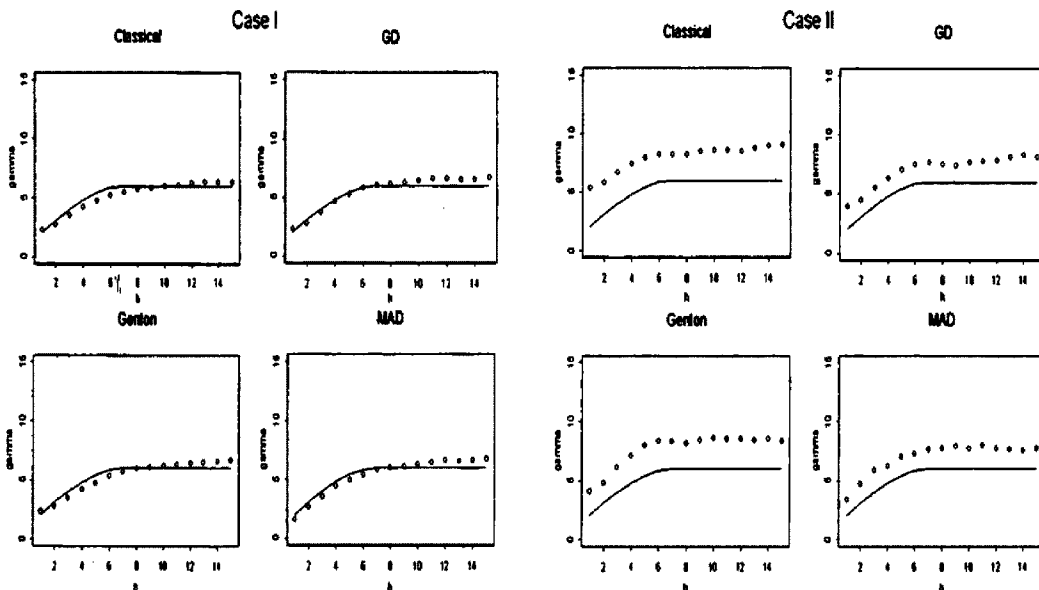


Figure 2: Cases 1 ( $\epsilon = 0, \sigma = 1$ ) and 2 ( $\epsilon = 0.10, \sigma = 5$ )

For the uncontaminated data (Case I), it is apparent that all four methods perform well. As a greater proportion of the observations are replaced with outliers, however, the classical method clearly becomes worse. Genton's method appears to be overestimating the sill, indicating that the presence of the outliers is inflating the estimate of the variability. Both the  $r_{gd}$  and MAD estimates appear to miss range somewhat, but appear to be better at portraying the overall shape of the variogram.

For cases V and VI, where there is more variability in the outliers, we see again that both  $r_{gd}$  and MAD methods appear to be a little better at estimating the underlying variogram than Genton's method. Clearly, the classical method performs horribly. Note the different

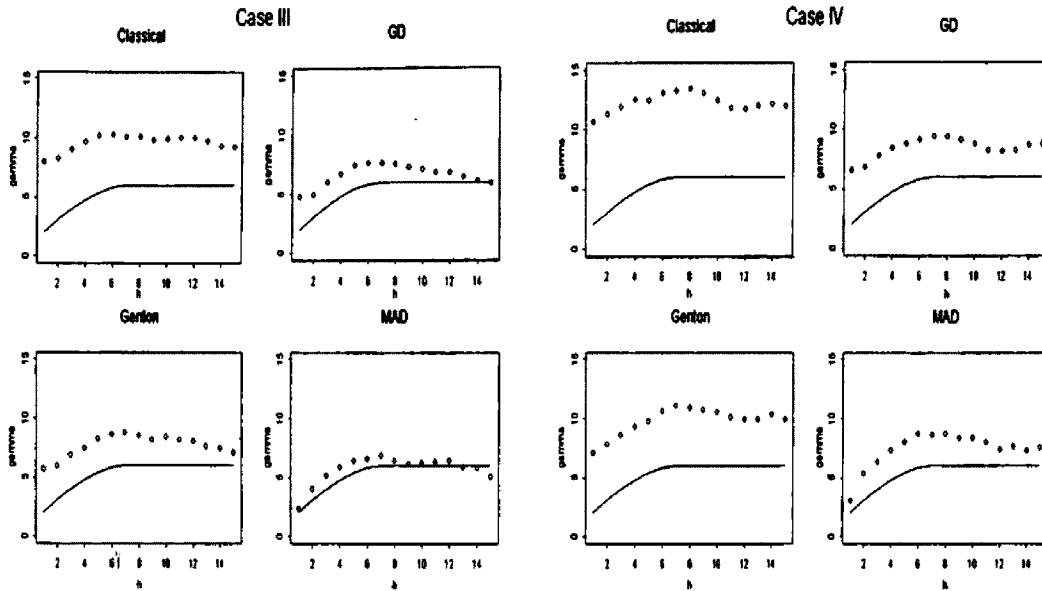


Figure 3: Cases 3 ( $\epsilon = 0.20, \sigma = 5$ ) and 4 ( $\epsilon = 0.30, \sigma = 5$ )

scale for cases V and VI.

In an attempt to quantify these observations, we look at the results for 40 repetitions of this process, using a non-linear least squares function built in to S-plus to fit the variograms. For each run, we estimate the sill, range, and nugget for the variogram using a nonlinear least squares method included in S-plus. For the 40 runs, we obtain a vector of the differences between the estimates and the true parameters for each of the variogram attributes. In the table below, we give the sample mean and median of these errors, indicating the bias in the estimation method. We also include the mean-squared error and median absolute error (MSE and MAE), indicating the variability in the estimates. For the first table, we also include two columns where we scale the bias and variability statistics by the true sill and range, to get a sense of the relative errors.

For the uncontaminated data, notice that all four methods perform roughly equally well. For range estimates, the *MAD* method appears to be the best of the four methods, except

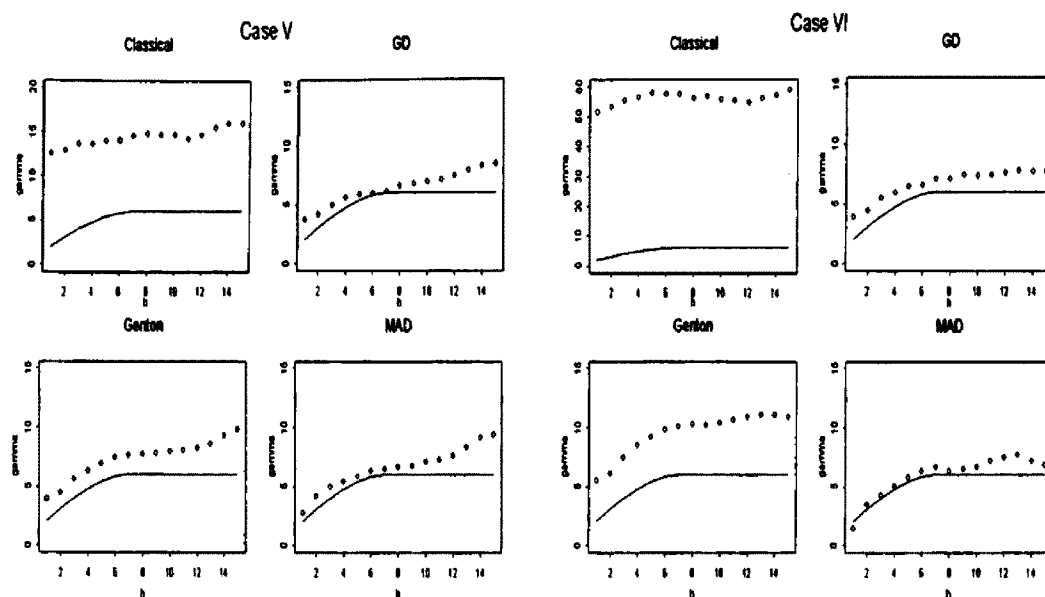


Figure 4: Cases 5 ( $\epsilon = 0.10, \sigma = 10$ ) and 6 ( $\epsilon = 0.10, \sigma = 20$ )

when considering MAE. In terms of its overall estimation quality, *MAD* is better in terms of mean bias and MSE, whereas  $r_{gd}$  is better under median bias and MAE. For Case II, *MAD* gives better overall results than the other methods under all criteria considered. It gives the most unbiased range estimates, but Genton's method has lower MSE and MAE. For Case III, *MAD* is better at overall estimation again. Here it outperform's Genton's method in range estimation as well.

In Case IV, again *MAD* is better at overall estimation. Genton's method appears to be better at range estimation here (except for MSE). But notice a pattern here - as more and more of the data is taken up by outliers, both Genton's method and the Classical method do worse at detecting the rise (the interval from lag 0 to the range), so the variogram estimate appears to be a pure nugget effect with these methods. In this sense, it may be better to consider the overall estimation quality rather than simply considering the range.

In Case V, *MAD* is again better than the other methods in the overall sense, and is

Case I							
<i>Method</i>	<i>Statistic</i>	<i>Sill</i>	<i>Rel. Sill</i>	<i>Range</i>	<i>Rel. Range</i>	<i>Nugget</i>	<i>Total</i>
<i>Classical</i>	<i>Mean Bias</i>	-0.480	-0.069	0.951	0.19	0.606	2.038
	<i>MSE</i>	1.120	0.16	3.797	0.759	0.399	5.317
	<i>Median Bias</i>	-0.494	-0.071	0.628	0.126	0.587	1.709
	<i>MAE</i>	1.098	0.157	1.573	0.315	0.871	3.542
<i>Genton</i>	<i>Mean Bias</i>	-0.414	-0.059	1.161	0.232	0.581	2.155
	<i>MSE</i>	1.173	0.168	5.447	1.089	0.383	7.003
	<i>Median Bias</i>	-0.497	-0.071	0.734	0.147	0.555	1.786
	<i>MAE</i>	1.401	0.2	1.716	0.343	0.823	3.939
<i>MAD</i>	<i>Mean Bias</i>	0.696	0.099	0.031	0.006	-0.487	1.215
	<i>MSE</i>	1.684	0.241	2.693	0.539	0.508	4.884
	<i>Median Bias</i>	0.627	0.09	-0.449	-0.09	-0.524	1.600
	<i>MAE</i>	1.171	0.167	1.747	0.349	0.834	3.753
<i>r<sub>gd</sub></i>	<i>Mean Bias</i>	-0.409	-0.058	1.138	0.228	0.606	2.152
	<i>MSE</i>	1.427	0.204	5.920	1.184	0.471	7.818
	<i>Median Bias</i>	-0.309	-0.044	0.614	0.123	0.556	1.478
	<i>MAE</i>	1.013	0.145	1.674	0.335	0.824	5.511

better at range estimation in all senses except for MSE, where Genton's method excels. Estimation with  $r_{gd}$  gives better overall results than Genton's method. In case VI, both *MAD* and Genton's method have a difficult time estimating the sill, so  $r_{gd}$  provides the best overall variogram estimates for this case. The *MAD* method outperforms the other methods at range estimation here.

It is worth noting that this is a small study, so the conclusions may be adversely affected by this. In order to rule out this possibility, it would be necessary to conduct a more extensive study with hundreds or thousands of simulations.

## 0.6 Conclusion

The method of variogram estimation proposed by Cressie and Hawkins and the Classical method have been shown to fail when outliers are present in the data set [2]. I have proposed two alternatives to the method presented by Genton, and tested them via computer

Case II					
<i>Method</i>	<i>Statistic</i>	<i>Sill</i>	<i>Range</i>	<i>Nugget</i>	<i>Total</i>
<i>Classical</i>	<i>Mean Bias</i>	-1.283	1.180	3.318	5.780
	<i>MSE</i>	2.284	10.501	11.354	24.139
	<i>Median Bias</i>	-1.327	0.920	3.230	5.477
	<i>MAE</i>	1.967	1.662	4.788	8.418
<i>Genton</i>	<i>Mean Bias</i>	-0.290	1.093	1.706	3.089
	<i>MSE</i>	0.934	4.208	3.030	8.173
	<i>Median Bias</i>	-0.312	0.960	1.730	3.002
	<i>MAE</i>	0.946	1.783	2.564	5.293
<i>MAD</i>	<i>Mean Bias</i>	0.713	0.181	0.018	0.913
	<i>MSE</i>	1.862	4.598	0.560	7.020
	<i>Median Bias</i>	0.952	-0.060	-0.096	1.109
	<i>MAE</i>	1.591	2.529	0.803	4.922
$r_{gd}$	<i>Mean Bias</i>	-0.712	1.214	1.734	3.660
	<i>MSE</i>	1.624	5.685	3.162	10.471
	<i>Median Bias</i>	-0.681	1.025	1.717	3.424
	<i>MAE</i>	1.303	2.217	2.547	6.066

simulation. In most of the cases, the *MAD* method performed better than that of *Genton*. When the outliers have high variability compared with the clean observations, the method using  $r_{gd}$  performs well also. In light of all this, my suggestion is to use the *MAD* method. The *Classical* method only works when no outliers are present, and  $r_{gd}$  only appears to be better in the worst case.

Case III					
<i>Method</i>	<i>Statistic</i>	<i>Sill</i>	<i>Range</i>	<i>Nugget</i>	<i>Total</i>
<i>Classical</i>	<i>Mean Bias</i>	-2.108	1.052	6.120	9.280
	<i>MSE</i>	4.850	5.250	37.544	47.643
	<i>Median Bias</i>	-2.145	0.534	5.951	8.629
	<i>MAE</i>	3.180	1.627	8.822	13.629
<i>Genton</i>	<i>Mean Bias</i>	-0.074	1.084	3.328	4.486
	<i>MSE</i>	0.618	5.534	11.411	17.563
	<i>Median Bias</i>	-0.116	0.831	3.246	4.194
	<i>MAE</i>	0.847	1.870	4.812	7.530
<i>MAD</i>	<i>Mean Bias</i>	1.625	-0.371	0.320	2.316
	<i>MSE</i>	3.774	2.481	0.586	6.842
	<i>Median Bias</i>	1.559	-0.582	0.234	2.375
	<i>MAE</i>	2.311	1.618	0.590	4.519
<i><math>\tau_{gd}</math></i>	<i>Mean Bias</i>	-0.872	1.305	3.246	5.423
	<i>MSE</i>	1.607	7.310	11.042	19.958
	<i>Median Bias</i>	-0.807	0.914	3.234	4.955
	<i>MAE</i>	1.228	2.205	4.794	8.227

Case IV					
<i>Method</i>	<i>Statistic</i>	<i>Sill</i>	<i>Range</i>	<i>Nugget</i>	<i>Total</i>
<i>Classical</i>	<i>Mean Bias</i>	-2.321	1.899	8.318	12.538
	<i>MSE</i>	6.108	24.705	70.305	101.118
	<i>Median Bias</i>	-2.429	0.554	8.208	11.191
	<i>MAE</i>	3.601	2.686	12.169	18.457
<i>Genton</i>	<i>Mean Bias</i>	-0.461	0.785	5.292	6.538
	<i>MSE</i>	1.172	5.857	28.565	35.594
	<i>Median Bias</i>	-0.402	0.232	5.222	5.856
	<i>MAE</i>	1.054	1.665	7.742	10.461
<i>MAD</i>	<i>Mean Bias</i>	2.115	-1.222	0.718	4.055
	<i>MSE</i>	5.849	4.618	1.598	12.065
	<i>Median Bias</i>	2.075	-1.608	0.495	4.179
	<i>MAE</i>	3.076	2.857	0.942	6.875
<i><math>\tau_{gd}</math></i>	<i>Mean Bias</i>	-1.190	1.276	4.723	7.190
	<i>MSE</i>	2.520	8.532	23.120	34.171
	<i>Median Bias</i>	-1.239	0.739	4.587	6.565
	<i>MAE</i>	1.837	2.374	6.801	11.021

Case V					
<i>Method</i>	<i>Statistic</i>	<i>Sill</i>	<i>Range</i>	<i>Nugget</i>	<i>Total</i>
<i>Classical</i>	<i>Mean Bias</i>	-0.944	0.448	10.757	12.149
	<i>MSE</i>	2.432	4.152	119.872	126.456
	<i>Median Bias</i>	-1.222	0.239	10.668	12.129
	<i>MAE</i>	2.345	2.070	15.817	20.232
<i>Genton</i>	<i>Mean Bias</i>	1.263	1.102	2.227	4.591
	<i>MSE</i>	3.151	3.953	5.135	12.239
	<i>Median Bias</i>	1.177	1.151	2.181	4.508
	<i>MAE</i>	1.744	2.017	3.233	6.994
<i>MAD</i>	<i>Mean Bias</i>	1.749	0.373	-0.483	2.605
	<i>MSE</i>	4.879	4.381	0.896	10.156
	<i>Median Bias</i>	3.036	1.949	1.122	6.107
	<i>MAE</i>	3.036	1.949	1.122	6.107
<i>r<sub>gd</sub></i>	<i>Mean Bias</i>	-0.330	1.228	1.987	3.546
	<i>MSE</i>	1.389	5.676	4.140	11.204
	<i>Median Bias</i>	-0.343	1.120	1.980	3.443
	<i>MAE</i>	1.056	2.203	2.935	6.195

Case VI					
<i>Method</i>	<i>Statistic</i>	<i>Sill</i>	<i>Range</i>	<i>Nugget</i>	<i>Total</i>
<i>Classical</i>	<i>Mean Bias</i>	0.015	-0.544	40.420	40.978
	<i>MSE</i>	11.420	10.810	1693.190	1715.418
	<i>Median Bias</i>	0.152	-1.349	40.361	41.862
	<i>MAE</i>	3.372	4.161	59.839	67.371
<i>Genton</i>	<i>Mean Bias</i>	2.955	1.172	2.607	6.734
	<i>MSE</i>	11.183	3.971	7.033	22.187
	<i>Median Bias</i>	2.888	1.209	2.544	6.641
	<i>MAE</i>	4.282	2.015	3.771	10.068
<i>MAD</i>	<i>Mean Bias</i>	2.566	0.207	-1.078	3.852
	<i>MSE</i>	8.477	2.921	1.553	12.950
	<i>Median Bias</i>	2.634	-0.064	-1.050	3.748
	<i>MAE</i>	3.906	1.709	1.556	7.171
<i>r<sub>gd</sub></i>	<i>Mean Bias</i>	-0.135	1.191	2.126	3.452
	<i>MSE</i>	1.340	5.493	4.732	11.620
	<i>Median Bias</i>	-0.107	1.110	2.105	3.323
	<i>MAE</i>	1.231	2.254	3.121	6.605



# Bibliography

- [1] N. Cressie and D.M. Hawkins. "Robust Estimation of the Variogram", *Mathematical Geology*, vol. 12, no. 2, 1980.
- [2] Marc G. Genton. "Highly Robust Variogram Estimation", *Mathematical Geology*, vol. 30, no. 2, 1998.
- [3] Rudy Gideon and Robert Hollister. "A Rank Correlation Coefficient Resistant to Outliers", *Journal of the American Statistical Association*, vol. 82, no. 398, 1987.
- [4] Rudy Gideon. "Location and Scale Estimates With Correlation Coefficients", unpublished paper (URL: <http://www.math.umt.edu/gideon/locscale.pdf>).
- [5] HuaiQing Sheng. "Estimation in Generalized Linear Models and Time Series Models with Nonparametric Correlation Coefficients", Ph.D. dissertation, University of Montana, 2002.
- [6] Rudy Gideon. "A Generalized Interpretation of Pearson's  $r$ ", unpublished paper (URL:<http://www.math.umt.edu/gideon/Generalized%20CC.pdf>).