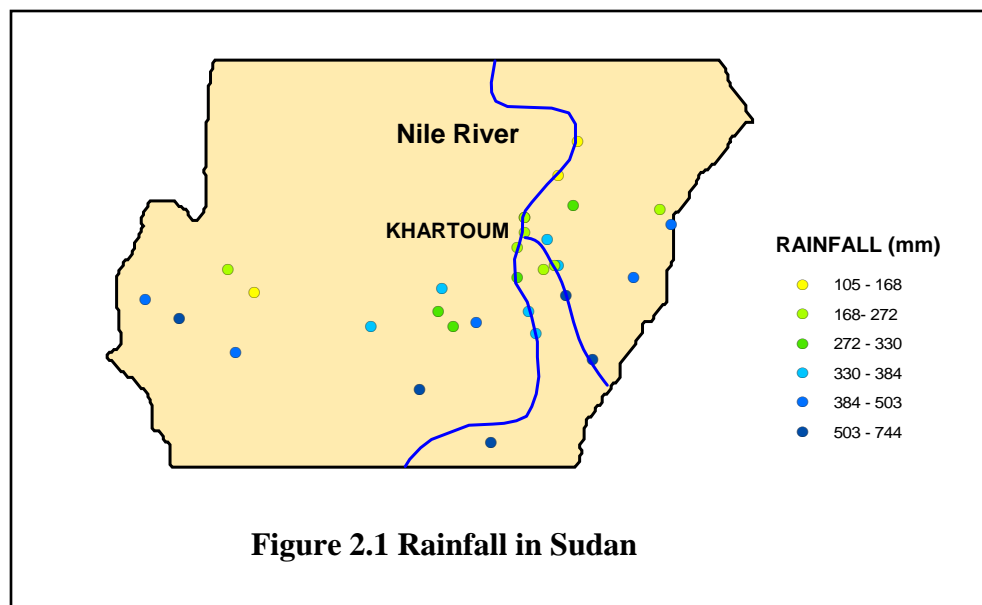


2. Examples of Continuous Spatial Data

As with point patterns, it is useful to consider a number of explicit examples of continuous spatial data that will serve to motivate the types of analyses to follow. Each of these examples is a case study in Chapter 5 of [BG], and the data for each example has been reconstructed in ARCMAP.

2.1 Rainfall in the Sudan

Among the most common examples of continuous spatial data are environmental variables such as temperature and rainfall, which can in principle be measured at each location in space. The present example involves rainfall levels in central Sudan during 1942, and can be found in the ARCMAP file, `arcview\Projects\Sudan\Sudan.mxd`. The Sudan population in 1942 was largely along the Nile River, as shown in Figure 2.1 below. The largest city, Khartoum, is at the fork of the Nile (White Nile to the west and Blue Nile to the east). There is also a central band of cities extending to the west.¹ Northern Sudan is largely desert with very few population centers. Hence it should be clear that the information provided by rainfall measurements in the $n = 31$ towns shown in the Figure will yield a somewhat limited picture of overall rainfall patterns in Sudan.



This implies that one must be careful in trying to predict temperatures outside this band of cities. For example, suppose that one tries a simple “smoother” like *Inverse Distance Weighting (IDW)* in ARCMAP (Spatial Analyst extension) [See Section 5.1 below for additional examples of “smoothers”]. Here, if the above rainfall data in each city,

¹ The population concentrations to the west are partly explained by higher elevations (with cooler climate) and secondary river systems providing water.

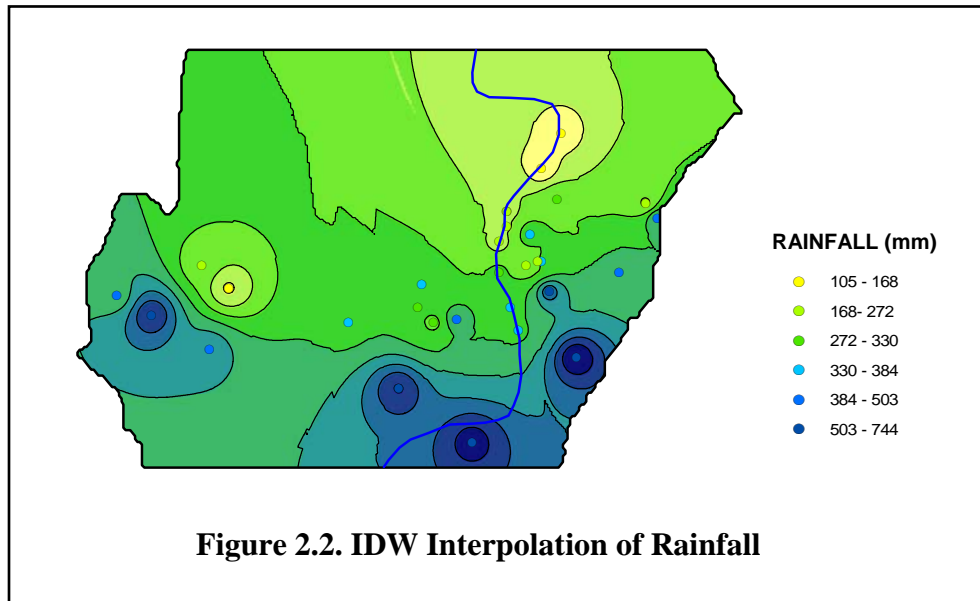
$i = 1, \dots, n$, is denoted by $y(s_i)$ the predicted value, $\hat{y}(s)$, at a point, $s \in R$, is given by a function of the form:

$$(2.1.1) \quad \hat{y}(s) = \sum_{i=1}^{n(s)} w_i(s) y(s_i)$$

where $n(s)$ is some specified number of points in $\{s_i : i = 1, \dots, n\}$ that are closest to s , and where the *inverse distance weights* have the form,

$$(2.1.2) \quad w_i(s) = \frac{d(s, s_i)^{-\alpha}}{\sum_{j=1}^{n(s)} d(s, s_j)^{-\alpha}}$$

for some exponent, α (which is typically either $\alpha = 1$ or $\alpha = 2$).² An interpolation of the rainfall data above is shown in Figure 2.2 below, for the default values, $n(s) = 12$ and $\alpha = 2$ in Spatial Analyst (**Interpolate to Raster** → **Inverse Distance Weighted**).³

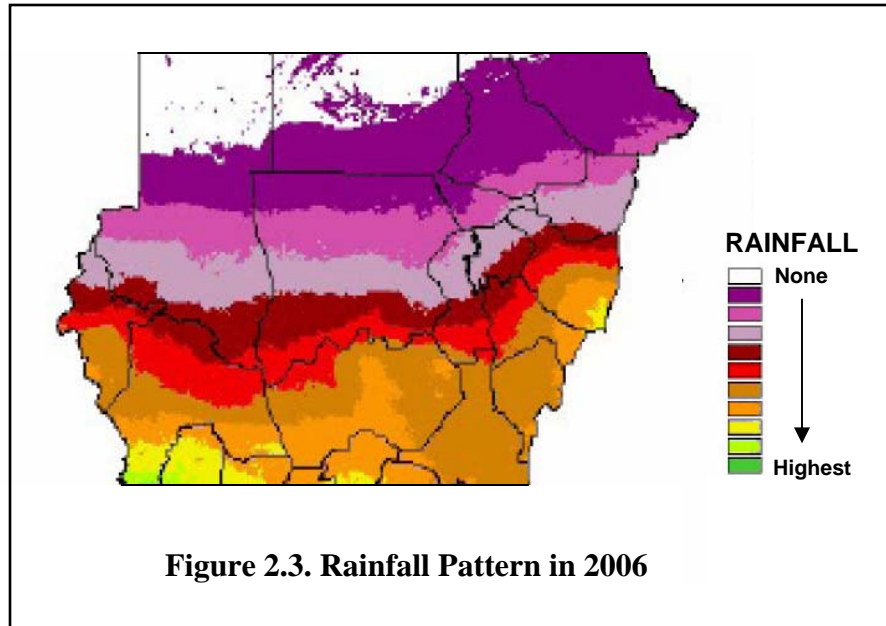


This is an “exact” interpolator in the sense that every *data* point, s_i , is assigned exactly the measured value, $\hat{y}(s_i) = y(s_i)$. But in spite of this, it should be evident that this interpolation exhibits considerably more variation in rainfall than is actually present. In particular, one can see that there are small “peaks” around the highest values and small “pits” around the lowest values. Mathematically, this is a clear example of what is called “overfitting”, i.e., finding a sufficiently curvilinear surface that it passes exactly through every data point.

² See also Johnston et al. (2001, p.114).

³ The results for IDW in the Geostatistical Analyst extension of ARCMAP are essentially identical.

For sake of comparison, a recent detailed map of rainfall in the same area for the six-month period from March to August in 2006 is shown in Figure 2.3 below.⁴ Since these are not yearly rainfall totals, the legend is only shown in ordinal terms. Moreover, while there is a considerable difference in dates, it is not unreasonable to suppose that the *overall pattern* of rainfall in 1942 was quite similar to that shown in the figure.



Here rainfall levels are seen to be qualitatively similar to Figure 2.2 in the sense that rainfall is heavier in the south than in the north. But it is equally clear that the actual variation in Figure 2.3 is much *smoother* than in Figure 2.2. More generally, without severe changes in elevation (as was seen for the California case in the Example Assignment) it is natural to expect that variations in rainfall levels will be gradual.

This motivates a very different approach to interpolating the data in Figure 2.1. Rather than focusing on the specific values at each of these 31 towns, suppose we concentrate on the *spatial trend* in rainfall, corresponding to $\mu(\cdot)$ in expression (1.2.1) above. Without further information, one can attempt to fit trends as a simple function of location coordinates, $s = (s_1, s_2)$. Given the prior knowledge that rainfall trends tend to be smooth, the most natural specification to start with is the smoothest possible (non-constant) function, namely a linear function of (s_1, s_2) :

$$(2.1.3) \quad Y(s) = \mu(s) + \varepsilon(s) = \beta_0 + \beta_1 s_1 + \beta_2 s_2 + \varepsilon(s)$$

This can of course be fitted by a linear regression, using the above data $[y(s_i), s_{1i}, s_{2i}]$ for the $i = 1, \dots, 31$ towns above. This data was imported to JMPIN as **Sudan.jmp**, and the

⁴ This figure is a portion of the map of rainfall in <http://www.unjlc.org/sudan/bulletin/bulletin80/view>.

1942 rainfall data (**R-42**) was regressed on the town coordinates (**X,Y**). The estimates ($\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$) were then imported to MATLAB in the workspace, **sudan.mat**. Here a grid, **G**, of points covering the Sudan area was constructed using **grid_form.m** (as in Section 4.8.2 of Part I) and the predicted value, $\hat{y}_g = \hat{\beta}_0 + \hat{\beta}_1 s_{g1} + \hat{\beta}_2 s_{g2}$, at each grid point, g , was calculated. These results were then imported to **Sudan.mxd** in ARCMAP and were interpolated using the *spline interpolator* in Spatial Analyst (**Interpolate to Raster → Spline**).⁵ The results of this procedure are shown in Figure 2.4 below:

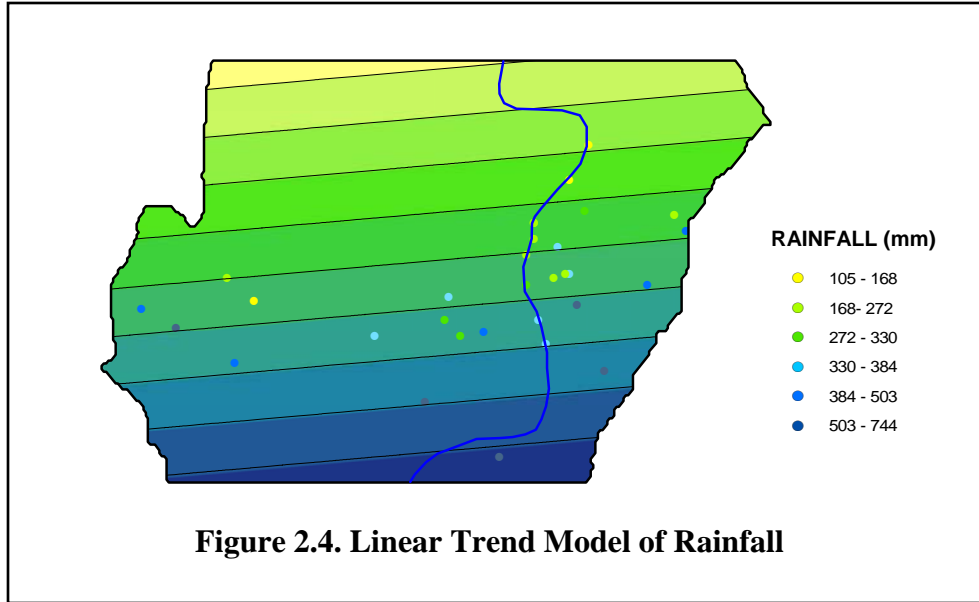


Figure 2.4. Linear Trend Model of Rainfall

A visual comparison of Figure 2.4 with Figure 2.3 shows that this simple linear trend model is qualitatively much more in agreement with actual rainfall patterns than the IDW fit in Figure 2.2.⁶ The results of this linear regression are shown in Table 2.1 below.

Term	Estimate	Std Error	t Ratio	Prob> t 	RSquare	0.59831
Intercept	12786.213	2031.626	6.29	<.0001	RSquare Adj	0.569618
X	7.1438789	5.934012	1.20	0.2387	Root Mean Square Error	1098.022
Y	-81.47974	12.89805	-6.32	<.0001	Mean of Response	3692.323

Table 2.1. Linear Regression Results

Notice in particular that the Y-coordinate (s_2 above) is very significant while the X-coordinate (s_1 above) is not. This indicates that most temperature variation is from north

⁵ See section ??? below for further discussion of spline interpolations.

⁶ It should be emphasized here that we have only used the “default” settings in the IDW interpolator to make a point about “over fitting”. One can in fact construct more reasonable IDW fits by using the many options available in the Geostatistical Analyst version of this interpolator.

to south, as is clear from Figures 2.3 and 2.4. However, the adjusted R-square shows that only about 57% of the variation in rainfall levels is being accounted by this linear trend model, so that there is still considerable room for improvement. With additional data about other key factors (such as elevations) one could of course do much better. But even without additional information, it is possible to consider more complex specifications of coordinate functions to obtain a better fit. As stressed above, there is always a danger of over fitting this data. But if adjusted R-square is used as a guide, then it is possible to seek better polynomial fits within the context of linear regression. To do so, it is natural to begin by examining the regression residuals, as shown in Figure 2.5 below.

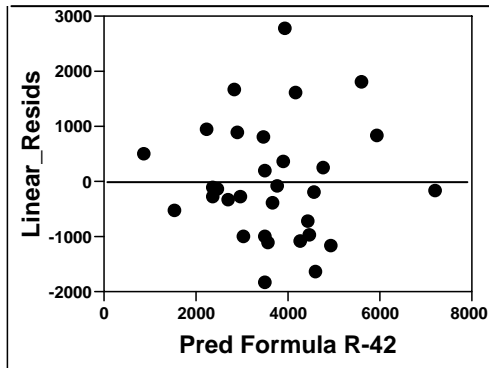


Figure 2.5. Residual Plot

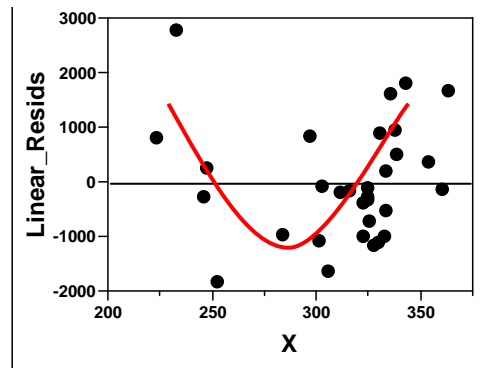


Figure 2.6. Residuals vs X

While these residuals show nothing out of the ordinary, a plot of the residuals against the X-coordinate is much more revealing. As seen in Figure 2.6 there appears to be a clear nonlinearity here, suggesting that perhaps a quadratic specification of X would yield a better fit than the linear specification in (2.1.3) above. This can also be seen by plotting the residuals spatially, as in Figure 2.7 below:

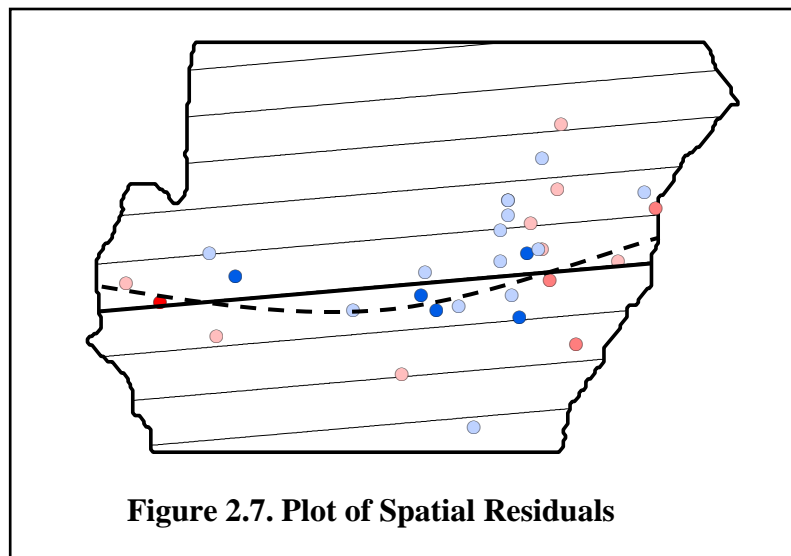


Figure 2.7. Plot of Spatial Residuals

If we focus on the heavy linear contour in the figure, then the residuals near the middle of this line are seen to be negative (blue), indicating that observed rainfall is *smaller* than predicted rainfall. Hence, recalling that higher rainfall values are to the south, these predictions could be reduced by pulling this contour line further south in the middle. Similarly, since the residuals near both ends of this line tend to be positive (red), a similar correction could be made by moving the ends north, yielding a *curved* contour such as the dashed curve shown in the figure.

Hence this visual analysis of spatial residuals again suggests that a quadratic specification of the X-coordinate should yield a better fit. Thus, as an alternative model, we now consider the following quadratic form:⁷

$$(2.1.4) \quad Y(s) = \beta_0 + \beta_1 s_1 + \beta_2 s_1^2 + \beta_3 s_2 + \varepsilon(s)$$

The results of this quadratic regression are shown in Table 2.2 below, and confirm that this new specification does indeed yield a significantly better overall fit, with adjusted R-square showing that an additional 10% of rainfall variation has been accounted for. In addition, it is clear that both the linear and quadratic terms in X are very significant, indicating that each is important.⁸

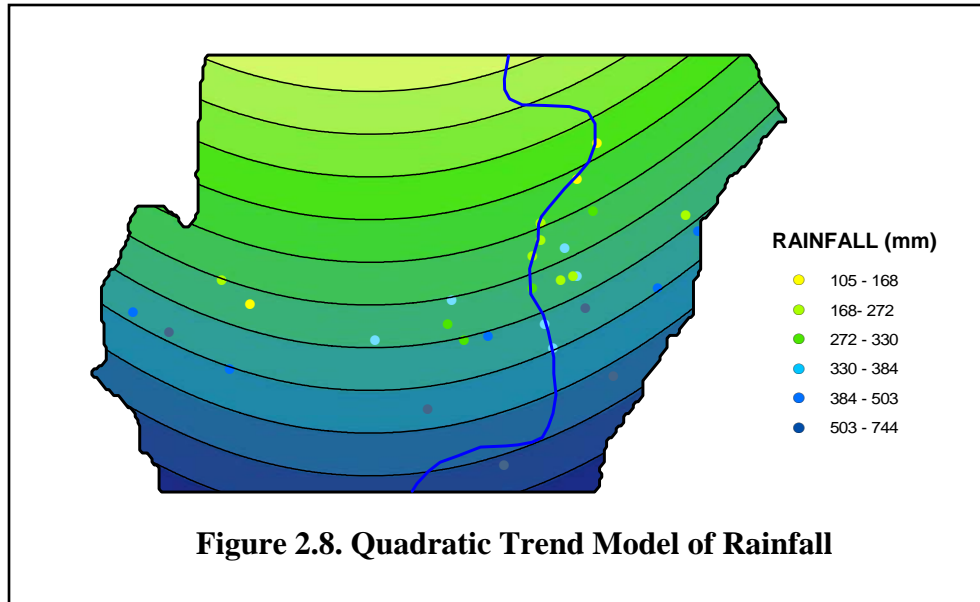
Term	Estimate	Std Error	t Ratio	Prob> t	RSquare	0.727274
Intercept	52409.813	11219.48	4.67	<.0001	RSquare Adj	0.696971
X	-258.7088	74.56896	-3.47	0.0018	Root Mean Square Error	921.3522
Y	-94.47108	11.41716	-8.27	<.0001	Mean of Response	3692.323
X^2	0.4573417	0.127993	3.57	0.0014		

Table 2.2. Quadratic Regression Results

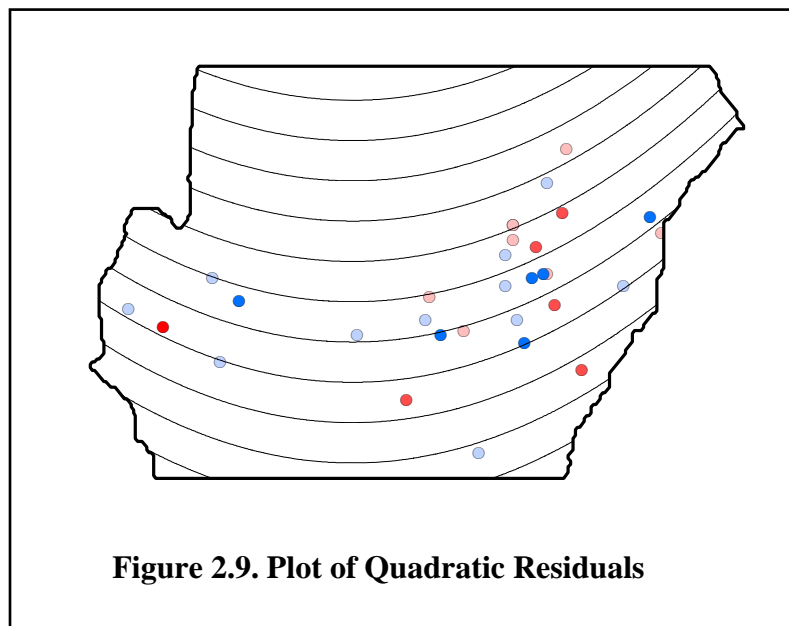
By employing exactly the same procedure outlined for the linear regression above, the results of this regression can be used to predict values on a grid and then interpolated in ARCMAP (again using a spline interpolator) to yield a plot similar to Figure 2.4 above. The results of this procedure are shown in Figure 2.8 below. Here a comparison of Figure 2.8 with the more accurate rain map from 2006 in Figure 2.3 shows that in spite of its mathematical simplicity, this quadratic trend surface gives a fairly reasonable picture of the overall pattern of rainfall in Sudan.

⁷ Here one can also start with a general quadratic form including terms for s_2^2 and $s_1 s_2$. But this more general regression shows that neither of these coefficients is significant.

⁸ It is of interest to notice that over short ranges, the variables X and X^2 are necessarily highly correlated. So the significance of both adds further confirmation to the appropriateness of this regression.

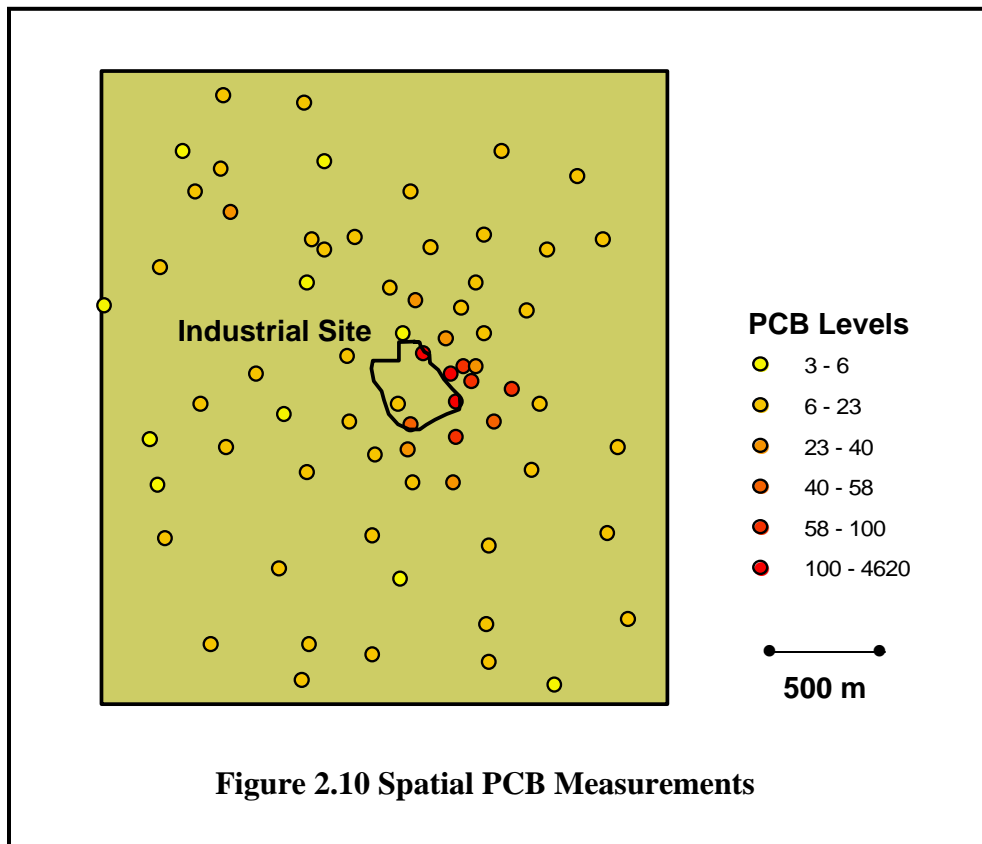


Finally a plot of the spatial residuals for this quadratic model, as in Figure 2.9 below, shows that much of the structure in the residuals for the linear model in Figure 2.7 has now been removed.



2.2 Spatial Concentration of PCBs near Pontypool in Southern Wales

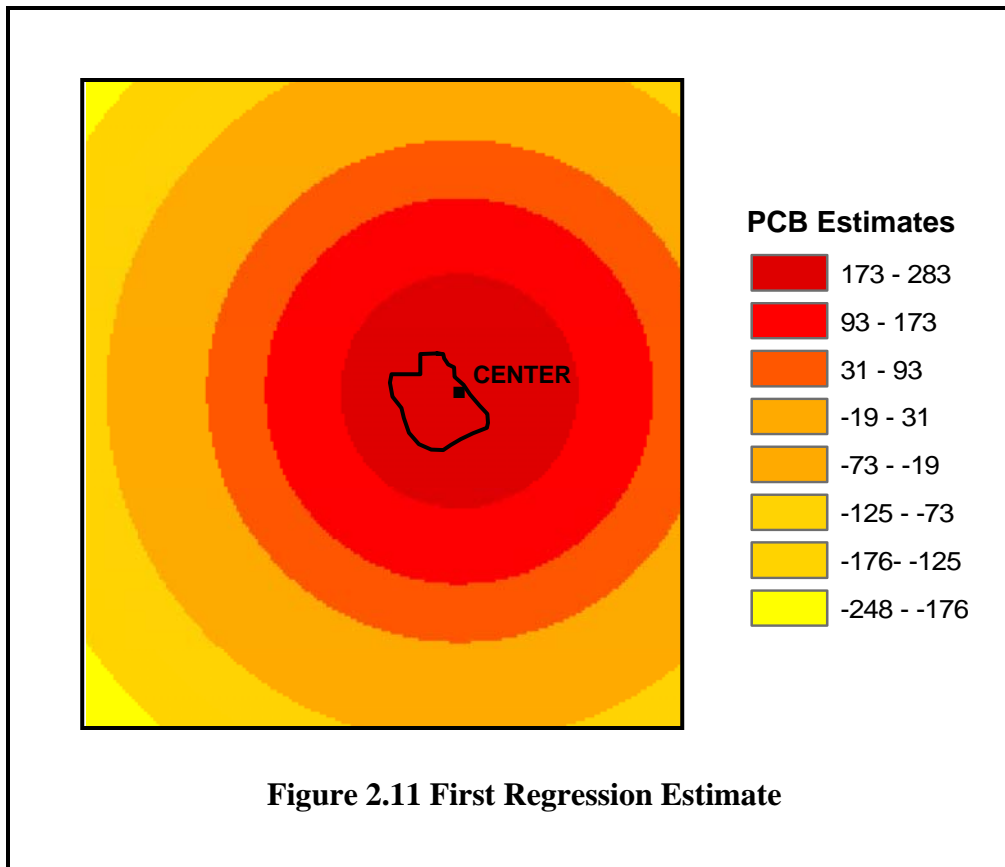
Among the most toxic industrial soil pollutants are the class of PCBs (polychlorinated biphenyls). The following data set from [BG] consists of 70 PCB soil measurements from the area surrounding an industrial site near the town of Pontypool, Wales, in 1991. The location and PCB levels for these 70 sites can be found in the JMPIN file, **Pcbs.jmp**. It is clear from Figure 2.10 below, that there is a significant concentration of PCB levels on the eastern edge of this site. The task here is to characterize the spatial pattern of variability in these levels surrounding the plant.



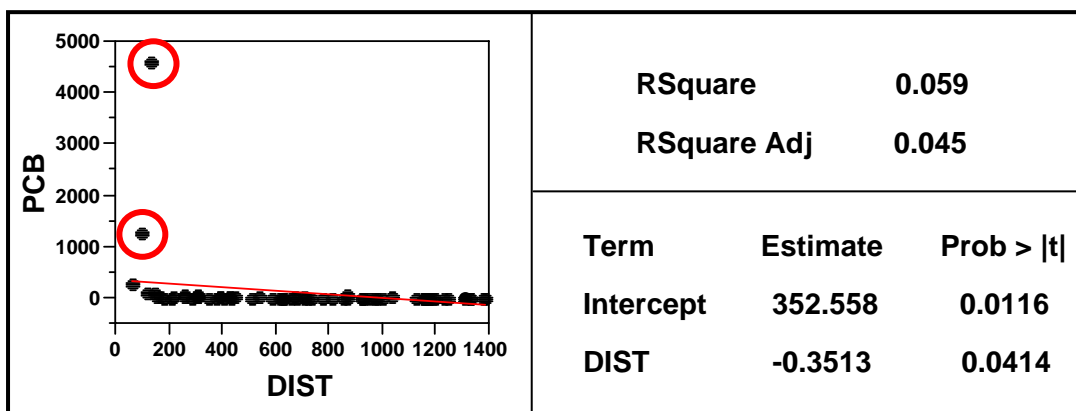
A visual inspection suggests that the concentration falls off with distance from this area of high concentration. To model this in a simple way, a representative location in this site, designated as the “Center” in Figure 2.11 below,⁹ was chosen and distance from this location to each measurement site was recorded (in the DIST column of **Pcbs.jmp**). Here the simplest possible model is to assume that these PCB levels fall off linearly with distance from this center. A plot of this regression is shown in Figure 2.11 below, and

⁹ The coordinates of this center location are given by $(x, y) = (330064, 198822)$.

look quite “reasonable” in terms of the concentric rings of decreasing PCB levels from this center point.



However an examination of the regression diagnostics in Figure 2.12 below tell a different story. Notice in particular that while distance is significant, the R-Square indicates that *less than 6%* of the variation in PCB levels is actually accounted for by such distances.



The reason for this is evident from an examination of the scatter plot on the left side of this figure, which reveals the presence of two dramatic *outliers*, circled in red. One could of course remove these outliers and produce a much better linear fit. But an examination of their distance shows that both are close to the center point in Figure 2.11, and hence are *extremely important* data points. So removing them would defeat the whole purpose of the analysis.

An alternative approach would be to attempt to transform the data to accommodate this extreme nonlinearity. One possibility would be to take logs of the variables. But even this is not sufficient in the present case. However a slight modification involving quadratic functions of logged variables works reasonably well. In particular, if we perform the following “translog” regression:¹⁰

$$(2.2.1) \quad \ln PCB_i = \beta_0 + \beta_1 \ln DIST_i + \beta_2 (\ln DIST_i)^2 + \varepsilon_i, \quad i = 1, \dots, n$$

then we obtain a vastly improved fit as well as more significant coefficients.¹¹ (Note that the positive coefficient on the quadratic term reflects the slight bowl shape seen in Figure 2.12 above.)

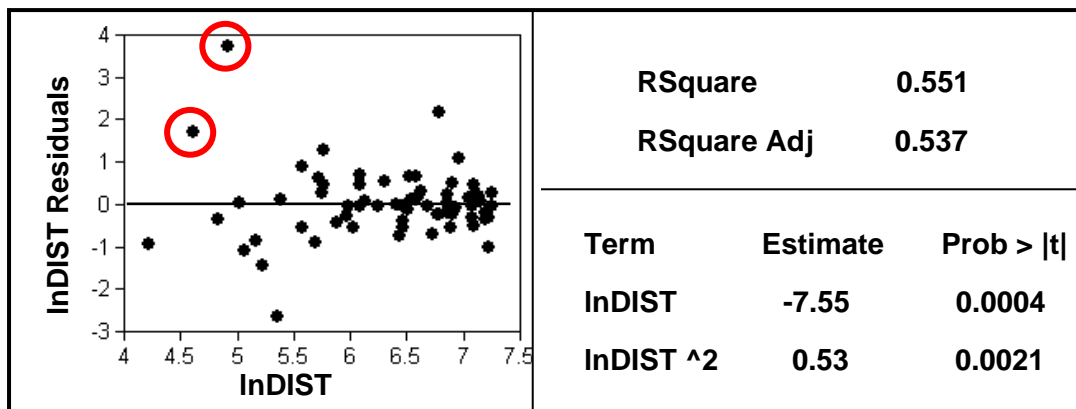


Figure 2.13. Transformed Residuals

Moreover, the two outliers (again shown by red circles in Figure 2.13) have been dramatically reduced by this data transformation. But while this transformed model of PCBs seems to capture the spatial distribution in a more reasonable way, we cannot draw sharp conclusions without an adequate *statistical model* of the residuals ($\varepsilon_i : i = 1, \dots, n$) in (2.2.1). This is the task to which we now turn.

¹⁰ This is closely related to the translog specifications of commodity production functions often used in economics. See for example <http://www.egwald.ca/economics/cesdatatranslog.php>.

¹¹ The estimated intercept term has been omitted to save space.