Analyzing Spatial Patterns and Stochastic Interpolation of Cobalt Values from Vancouver Island Using Simple Kriging Model

Introduction

This following study is intended to gain further perspectives on the statistical methods of geochemical prospecting. Analyzing the spatial distribution of element concentration values and recognizing unusual features in the geographical context for potential mineralization is a major concern in geochemical prospecting (Howarth, 111). The study is based on sample data of cobalt values from Vancouver Island in British Columbia collected by the Geological Survey of Canada. This investigation attempts to analyze the pattern of spatial dependencies among the cobalt values and carry out stochastic interpolation of cobalt values in terms of a simple kriging model in order to identify potential cobalt deposits.

The Vancouver sample data extends over the area at the northern tip of the island shown in the Figure 1.1a and the area outlined in red denotes the full extent of the data site. For purposes of analysis, a smaller set of 286 sample sites was selected, as shown by the dots in Figure 1.1b.

Figure 1.1a. Vancouver Sample Area

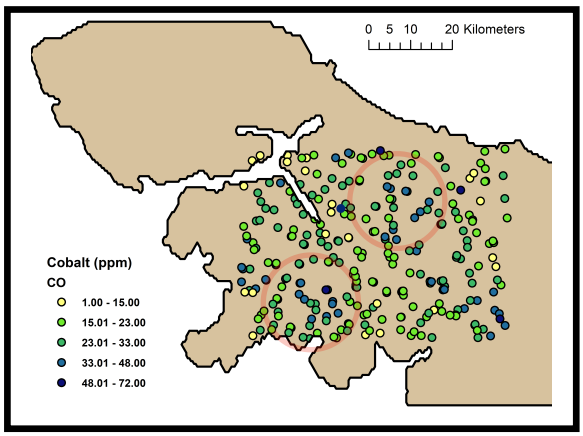


Figure 1.1b. Vancouver Sample Area

We explore distributions of patterns of cobalt elements here because cobalt (Co) is a potentially critical mineral increasingly used in magnets and rechargeable batteries. The disposition of the cobalt sample points in this study varies based on the nature of investigation and stream sediment or lake sediment sampling is dependent on drainage pattern (Howarth, 136). We can find the curvilinear patterns of the sample points. As with many geochemical surveys, samples are here taken mainly along stream beds and lake shores, where minerals deposits are more likely to be found (Smith, 33). In this sense, we want to use kriging, which is a statistical method of interpolation, to prospect cobalt values across our sample area. According to the general principles of geochemical prospecting, mineral deposits represent anomalous concentration of specific elements and mineral deposits has a central zone where valuable elements surrounding the zone decrease in concentration and we measure this concentration in parts per million (ppm)

or parts per billion (ppb) (Horsnail, 2). As shown by the red circles in Figure 1.1b, we can see that there are potential "central zones" of cobalt deposits that need further investigation.

Geochemical mapping have been executed in British Columbia to display and investigate concentrations of chemical elements in the past. One such study is R.E.Lett's paper that describes a reconnaissance scale regional geochemical survey carried out over the Bowser Lake in Northwestern British Columbia (Lett, 61). The survey produced new multi-element geochemical data and found abundant gold grains in two creeks, which reflected results of the Iskut-Palmiere prospect. The survey also found spatial dependencies of gold grain counts as gold grain counts are higher at sites close to the Eskay Creek mine site and lower at areas far from the mine site. Drawing on this paper, our study will explore the spatial dependencies of cobalt values at the Vancouver Sample Area and perform geochemical prospecting on the cobalt values.

Methodologies and Results

Since the mapped cobalt values exhibits strong similarities between neighboring values (as shown in Figure 1.1b), we can expect to find a substantial range of spatial dependence in this cobalt data. However, we can also see that the covariance-stationarity assumption of Isotropy, which states that similarities between concentration levels at different locations depend only on the distance between them, is questionable for this data. We can see diagonal "waves" of high and low cobalt values rippling through the sample area and we want to explore the spatial patterns of cobalt values in the variogram estimation procedure to follow.

Analysis of spatial dependencies is first carried out using empirical variogram estimation in Matlab. The key difference between continuous spatial data of cobalt values and point patterns is that there is a meaningful value at every location in the study area. Since our spatial variable in this case tend to exhibit some degree of continuity over space, we expect these variables to exhibit similar values at locations close together in space. In order to identify spatial trends in the spatial stochastic process, we expect that for sites that are sufficiently close together, the associated spatial residuals not captured by global trend will tend to exhibit statistical dependence, which is measured by the covariance of these spatial residuals. In particular, positive dependencies among spatial residuals will tend to be reflected by positive covariance among these residuals.

Before proceeding, it is important to emphasize the notion of spatial stationarity to model covariances among spatial random effects in cobalt values. According to the Spatial Random Effects Theorem, for any random vector of multi-location effects comprised of a sum of individual random factors with zero means and covariance matrices, if the distributions of these random factors are "not too different", and the dependencies among these random factors are "not too strong", then the distribution of multi-location effects is approximately multi-normal. In that sense, our aim is to specify the unknown covariance matrix for the random effects in cobalt values and model these unobserved dependencies with general spatial dependencies that should be common to all these covariance structures.

A spatial stochastic process, $\{Y(s): s \in R\}$, is said to be covariance stationary if and only if the following two conditions hold for all $s_1, s_2, v_1, v_2 \in R$:

(i) $E[Y(s_1)] = E[Y(s_2)]$, ii) $\| s_1 - s_2 \| = \| v_1 - v_2 \| \Rightarrow \text{cov}[Y(s_1), Y(s_2)] = \text{cov}[Y(v_1), Y(v_2)]$

The covariogram and its normalized form, the correlogram, are by far the most intuitive methods for summarizing the structure of spatial dependencies in a covariance stationary process (Smith, 1). Here, we will use C(h) to represent the common covariance value such that $\text{cov}[Y(s),Y(v)] = C(h)$ and h represents distances. Since the covariance values, C(h) are unique for each distance value, h, the function, C, of these distances is designated as the covariogram for a given covariance stationary process. However, as covariograms have particular units and are difficult to interpret, we analyze dependencies between random variables in terms of (dimensionless) correlation coefficients, which is a normalized form of the covariogram called the correlogram for a covariance stationary process. The correlation between any Y(s) and Y(v) is estimated by:

$$\rho[Y(s),Y(v)] = \frac{\text{cov}[Y(s),Y(v)]}{\sqrt{\text{var}[Y(s)}\sqrt{\text{var}[Y(v)}} = \frac{C(h)}{\sqrt{C(0)}\sqrt{C(0)}} = \frac{C(h)}{C(0)}$$

In our study, estimation of correlogram presents certain difficulties so we use variogram to display the variability between data points as a function of distance. The first step of our empirical variogram estimation is to aggregate point pairs of cobalt values with similar distances and hence estimate the estimator of the variogram value at only a small number of representative distances for each aggregate. We first partition distances into intervals, called bins, and take the average distance in each bin to be the appropriate representative distances, called lag distances. If N_k denotes the set of distance pairs, (s_i,s_j), in bin k, [with the size (number of pairs) in N_k denoted by $|N_k|$], and if the distance between each such pair is denoted by h_k, then the lag distance h_k, for bin k is defined to be:

$$h_k = \frac{1}{|N_k|}\sum_{(s_i,s_j)\in N_k} h_{ij}$$

The empirical variogram is usually informative in terms of the possible shapes of the true variogram. To examine possible maximum distances for the desired variogram, we first observe (by an application of the measurement tool in ARCMAP) that a neighborhood of 20,000 meters around typical sites appears to be large enough to contain most positive dependencies with other sites. Then we choose to fit a spherical variogram, which is a widely used variogram model. It is defined for all $h \geq 0$ by:

$$\gamma(h;r,s,a) = \begin{cases} 0 & , \ h = 0 \\ a + (s-a)\left(\frac{3h}{2r} - \frac{h^3}{2r^3}\right) & , \ 0 < h \leq r \\ s & , \ h > r \end{cases}$$

where r denotes the maximum range of positive spatial dependencies (designated simply as the range of the variogram), s corresponds to the sill of the variogram, and a corresponds to the nugget.

The choice of the maximum lag distance also involves some implicit restrictions as it cannot be greater than half of the maximum pairwise distance. We observed (again by using the measurement tool) that half of the maximum pairwise distance in this case is about 40,000 meters. Therefore, we first used a reasonable maximum distance of 20,000 meters and then used a maximum distance of 40,000 meters to construct two spherical variograms. As shown in Figure 1.2a, the blue dots are the empirical variogram points maximum distance of 20,000 meters and the estimated spherical variogram is shown in red, which rise toward a “sill” at about 17,010

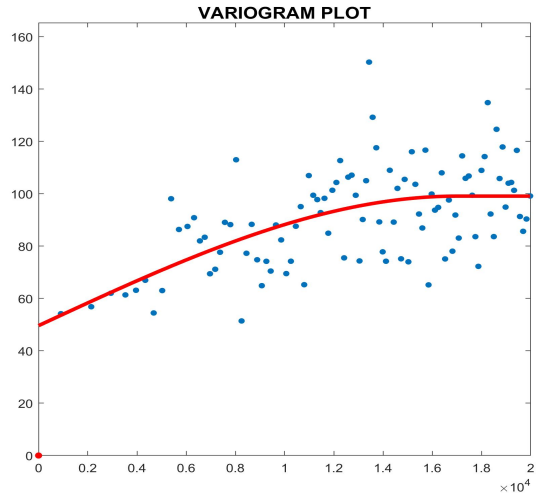


Figure 1.2a: Fitted Spherical Variogram

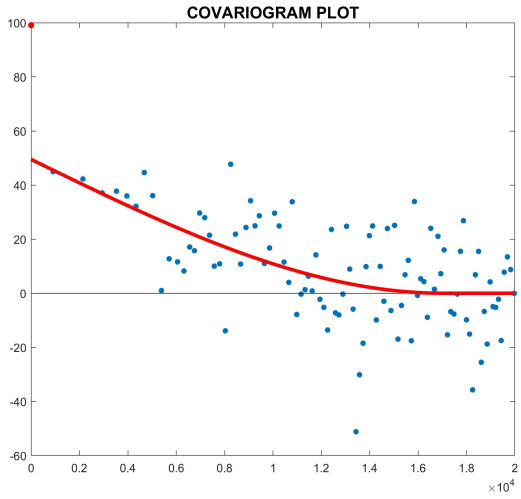


Figure 1.2b: Derived Spherical Covariogram

meters according to parameter estimate results. Figure 1.2b is the derived spherical covariogram corresponding to the empirical variogram in Figure 1.2a. It is clear that 17,010 meters is about the distance at which covariance (and hence correlation) first falls to zero. It denotes the distance beyond which there is estimated to be no statistical correlation between cobalt values. Turning to the other estimated parameters, note first from Figure 1.2a that the sill is about 99, which is seen to be the estimated variance of individual cobalt values (i.e., the estimated covariance at "zero distance"). Similarly, as shown in Figure 1.2b, the nugget is about 50, which is seen to be that part of the individual variance that not related to spatial dependence among neighbors. Since in this case the relative nugget effect is 0.51 (=50/99), the underlying process exhibits some degree spatial dependence.

Then we fit a spherical variogram with maximum distance of 40,000 meters as shown in Figure 1.3a, which still rise toward a "sill" at about 13,767 meters. In this case, the sill is about 92 according to the parameter estimate and the nugget is about 49. Therefore, the relative nugget effect is 0.53 (=49/92), so the underlying process exhibits some degree spatial dependence.

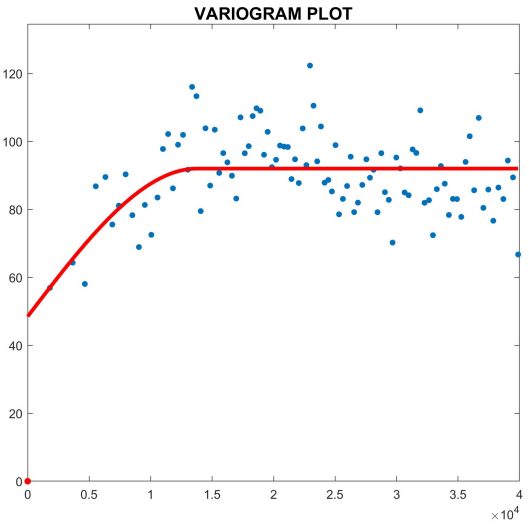


Figure 1.3a: Fitted Spherical Variogram

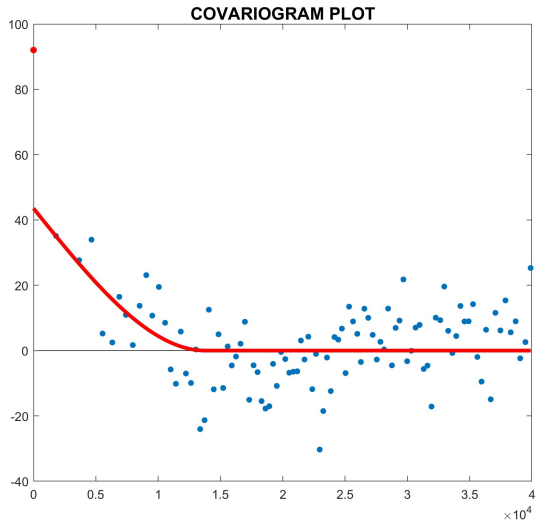


Figure 1.3b: Derived Spherical Covariogram

While we compare these two derived covariograms in Figure 1.2b and Figure 1.3b, it is important to note that the vertical (squared difference) scales for these two figures are the same, but the horizontal distance scales are now different. While the segment of Figure 1.3b up to 20,000 meters is qualitatively similar to Figure 1.2b, the bins and corresponding lag distances are not the same as in Figure 1.2b. Given this scale difference, we can see that the covariogram in Figure 1.3b shows a rise starting at about 20,000 meters, which can be interpreted to mean that pairs of y-values (cobalt measurements) separated by less than 20,000 meters tend to be more similar (positively correlated) than those separated by slightly larger distances. By again using the measurement tool in ARCMAP, it can be seen that the spacing of successive waves is about 20,000 meters. So it does appear that this effect is being reflected in the empirical variogram and derived spherical covariogram.

In the next step, we will use the variogram estimate obtained for the 20,000-meter max distance and simple kriging model to carry out a stochastic interpolation of cobalt values. Kriging model is a spatial prediction models that predict values based on local information. Simple kriging assumes that underlying stochastic process itself is entirely known and that the spatial trend is constant. We treat the observed data as a finite sample from a spatial stochastic process $\{Y(s): s \in R\}$ and we assume that some appropriate subset of sample locations, $S(s_0) \subseteq \{s_i : i = 1,..,n\}$ has been chosen for prediction. To determine a prediction $\hat{y}(s_0)$ based on sample data, we determine the prediction as a function of the random variables, $\{Y(s_1),..,Y(s_{n_0})\}$ associated with the observed data. We hypothesize that $\hat{Y}(s_0)$ can be represented as some linear combination of these random variables:

$$\hat{Y}(s_0) = \sum_{i=1}^{n_0} \lambda_{0i} Y(s_i)$$

where the weights λ_{0i} are yet to be determined and this hypothesis is referred to as the linear prediction hypothesis. We will use simple kriging model by derive optimal prediction weights in our analysis.

If we want to predict the cobalt value, $Y(s_0)$, at some location, then since $\mu(s_0) = \mu$ is already known, we see from the identity $Y(s_0) = \mu + \varepsilon(s_0)$ that it suffices to predict the associated error with optimal weights:

$$\hat{\varepsilon}(s_0) = \sum_{i=1}^{n_0} \lambda_{i0} \varepsilon(s_i)$$

We first krige this data in MATLAB and use the estimated parameter values to predict cobalt value at a point location (615000, 586500) using a kriging bandwidth of 5000 meters. According to the results of the variogram estimate obtained for the 20,000-meter max distance, the estimated range is 17,009, the estimated sill is 99, and the estimated nugget is 50. After we run the prediction using simple kriging, we get the predicted cobalt value as 17.73 and the standard error of prediction at the point is 8.04. Then we construct a 95% prediction interval for Cobalt value at this location, which is (4.44, 31.02). This means that there is a 95% probability that the cobalt value will be contained within this prediction interval.

Then we will use the spherical variogram parameters for 20000-meter max distance case to krige the cobalt data using the Geostatistical Analyst extension in ARCMAP. Before Kriging the Cobalt Data, we first calculate the mean Cobalt value using the attribute table and the mean value is 25.27. Then we use Geostatistical Wizard and get a mean value of 25.27 as well.

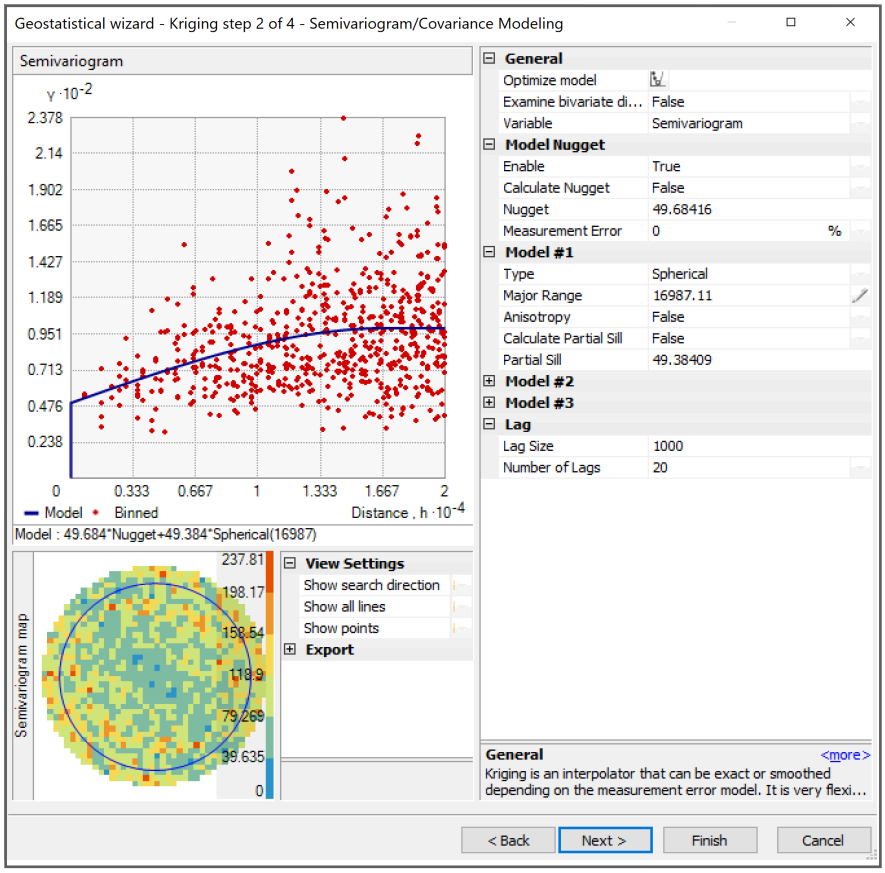


Figure 1.4 Spherical Variogram Fit in Geostatistical Analyst

As shown in Figure 1.4, by setting the number of lags to 1,000 and choosing a constant bin size of 20 meters (as seen in the Lag window in the lower right of Figure 1.4), we will obtain a maximum distance of exactly 20,000 meters (as seen on the distance axis of the variogram plot). The fitted spherical variogram is shown by the blue curve in Figure 1.4, and the empirical variogram is shown by red dots. Now we compare the Major Range and Nugget values in Geostatistical Analyst with the Range and Nugget computed in MATLAB. A comparison of the parameter estimates using both MATLAB and GA in this example (Figure 1.5 below) show that in spite of the differences above, they are qualitatively very similar.

	MATLAB	GA
Range	**17,009.8**	**16,987.1**
Sill	**99.1**	**99.1**
Nugget	**49.6**	**49.7**

Figure 1.5 Parameter Estimates

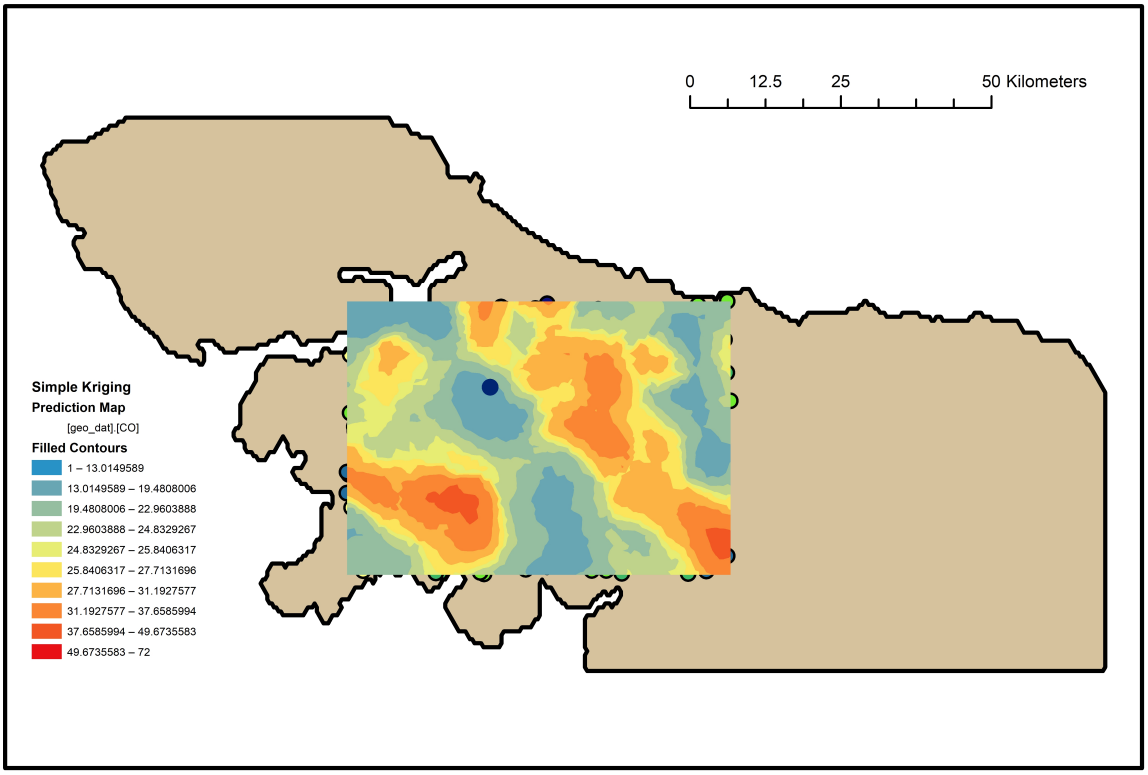


Figure 1.6 Simple Kriging

As shown in Figure 1.6, spatial interpolation of cobalt values is executed using simple kriging model in Geostatistical Analyst. The deep blue point is the point at (615000, 586500) and we want to examine the cobalt values near this point and get the mean of these cobalt values. We first choose a sample of five neighboring kriged cobalt value, which are 17.6, 17.9, 18.5, 18.7, 16.2, and calculate the average of these values, which is 17.78 (=(17.6+17.9+18.5+18.7+16.2)/5). Then we compare it with the our predicted cobalt value (17.73) in Matlab and find that they are very similar. This means that the cobalt value predicted by Geostatistical Analyst using simple kriging is almost the same as the cobalt value predicted in MATLAB.

Then we want to select six sites surrounding the point at (615000, 586500) and calculate the mean value of these six cobalt values as shown in Figure 1.7. From the results in ARCMAP, the mean cobalt value from these six sites is 16.5, which is very similar to the predicted value we calculated in MATLAB and the mean value of our previous sample of five neighboring kriged cobalt values. Therefore, the kriged value seems reasonable in relation to these selected data points.

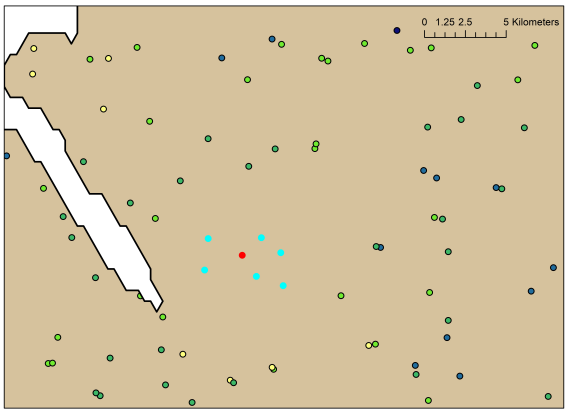


Figure 1.7 Six Selected Sites

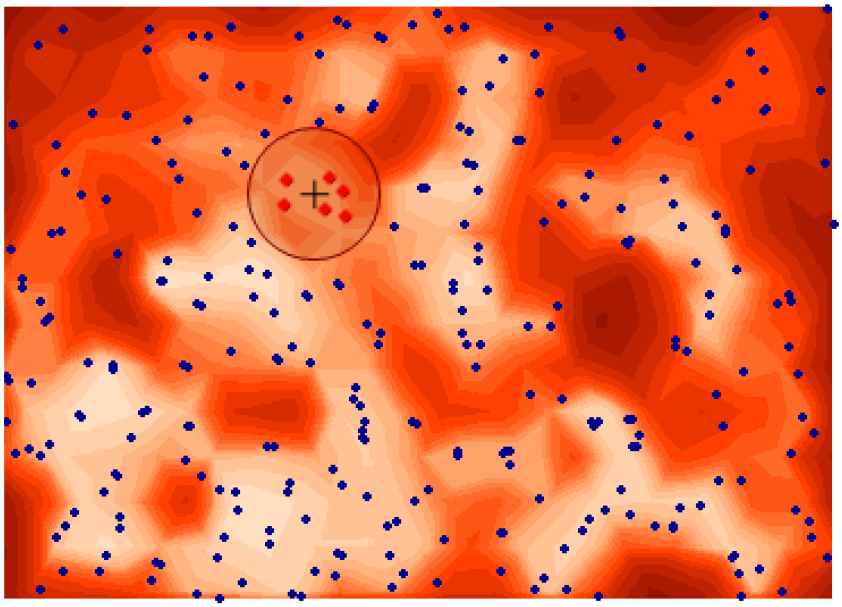


Figure 1.8 Standard Error of Prediction Value

Then we will again examine the prediction-error value at point (615000, 586500) using Geostatistical Analyst. The six neighbors of the point that have been used in the Kriging prediction are shown in the red circle in Figure 1.8. The standard error of predicted value at point (615000, 586500) is 8.04, which is exact the same as the prediction-error value we get in MATLAB. Areas in red have higher standard errors of predicted values than areas in orange.

Finally, we investigate the simple kriging prediction map in Figure 1.6 and see that it shows spatial dependencies of predicted cobalt values. As there are three spots that have very high predicted cobalt values, predicted cobalt values close to these three spots are higher and predicted cobalt values far from these three spots are lower. We also compare the simple kriging prediction map in Figure 1.6 and the standard error of predicted value in Figure 1.8 and see that they both show diagonal "waves" of high and low predicted cobalt values and standard errors of predicted values. This corresponds with our expectation that an assumption of covariance stationarity might be an over-simplification of this spatial data pattern. These waves in standard error of predicted value are roughly parallel to the Pacific coastline, and would seem to reflect the history of continental drift in this region.

Discussion

We first analyzed the spatial dependencies of cobalt values from Vancouver Island using fitted spherical variograms and derived spherical covariograms. We arrived at results that pairs of cobalt values separated by less than 20,000 meters tend to be more similar (positively correlated) than those separated by slightly larger distances. We then run simple kriging model in both MATLAB and ARCMAP to predict cobalt values across the sample area and arrived at very similar prediction results at a specific point. Our findings in the spatial interpolation of cobalt values correspond with R.E.Lett's paper that sites close to the "central zone" of cobalt values

have higher predicted values and sites far from the “central zone” have lower predicted cobalt values (Lett, 65).

Although the calculated relative nugget effect shows the underlying process exhibits relatively little spatial dependence, the relative nugget effect at a maximum distance of 20,000 meters is smaller than at a maximum distance of 40,000 meters. This means that there is relatively more spatial dependencies when pairs of cobalt values were separated by less than 20,000 meters. From the simple kriging prediction map, we can also conclude that while the “wave” effect is being reflected in the spatial interpolation and an assumption of covariance stationarity might be an over-simplification of this spatial data pattern, our predicted cobalt values do show “central zones” where valuable elements surrounding the zone decrease in concentration.

References

R.E.Lett, P.W.B.Friske and Jackaman, Wayne. “National Geochemical Reconnaissance Program in Northwestern British Columbia: Bowser Lake (NTS 104A) Regional Geochemical Survey.” (2005).

R. F. Horsnail, "Geochemical prospecting", in AccessScience@McGraw-Hill, http://www.accessscience.com, DOI 10.1036/1097-8542.285700, last modified: March 29, 2001.

Webb, J. S., and R. J. Howarth. “Regional Geochemical Mapping.” Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, vol. 288, no. 1026, 1979, pp. 81–93. JSTOR, www.jstor.org/stable/2398727. Accessed 30 Mar. 2020.

Smith, Tony E. “Part II. Continuous Spatial Data Analysis .” *Variograms*, www.seas.upenn.edu/~ese502/NOTEBOOK/Part_II/4_Variograms.pdf.

Fitting a Trend Surface to Predict Mean Annual Temperature in South America

Introduction

This following study is intended to broaden the understanding of fitting a trend surface to the mean annual temperatures in South America. Trend surface analysis is the most widely used global surface-fitting procedure where the mapped temperature are approximated by a polynomial expansion of the geographic coordinates of the points, and the coefficients of the polynomial function are found by the method of least squares, ensuring that the sum of the squared deviations from the trend surface is a minimum. This study attempts to use trend surface analysis to understand the attributes that influence the spatial pattern of mean temperatures in South America.

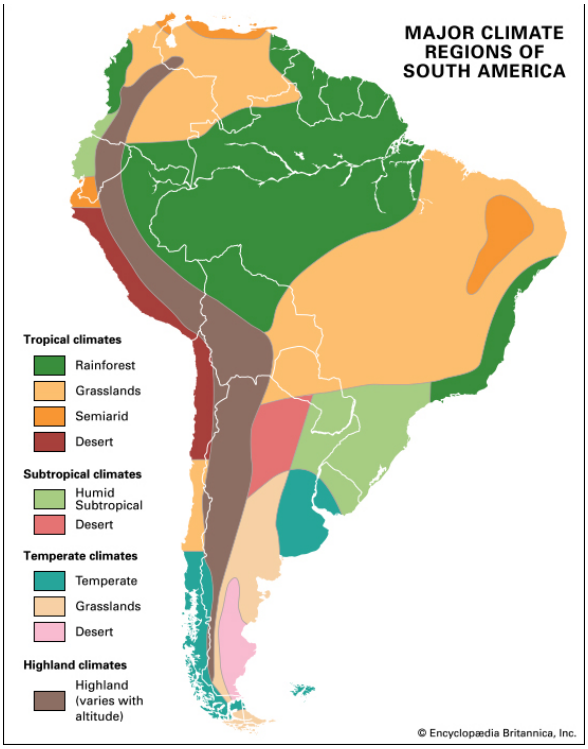
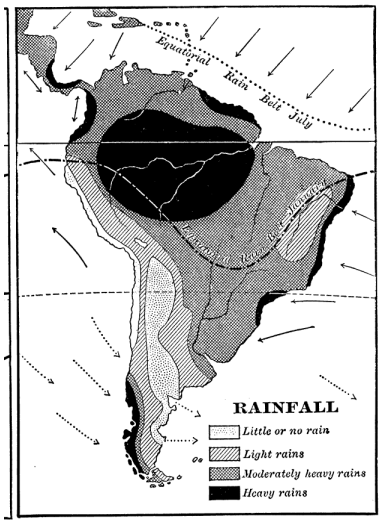


Figure 2.1a Climates in South America

As shown in Figure 2.1a, there are four major types of climates in South America, which are tropical climates, subtropical climates, temperate climates, highland climates. According to Robert C. Eidt in his article *The Climatology of South America*, he pointed out that among the various external factors, the astronomical location of a place is the most important in determining major climatic characteristics. In this regard, because the width of South America, which extends along 45 degree of longitude, is greatest in the equatorial latitudes, a tropical condition dominates over half the land area. The narrowing of the land at these middle and high latitudes also permits a greater maritime influence so that summer and winter temperature extremes are attenuate (Eidt, 54). Norberto O. Garcia also stated in his paper *South American climatology* that the climate of South America is a consequence of its geographical position with respect to latitude and ocean current activity. This characteristic of South American Climate is also evident in Figure 2.1a as tropical and subtropical climates are at higher latitudes whereas temperate and highland climates are at lower latitudes. As we further investigate other attributes that may affect South American climate in Figure 2.2a and Figure 2.2b, it is important to note that the wind direction in Northern part of South America is southwest and northwest whereas the wind direction in Southern part of South America is southeast. This difference in prevailing winds may be associated with differences in Ocean current activity and also influence the climate in South America.

In view of these previous work related to the analysis of South America Climate, we would like to fit regression models to regress temperature on coordinate variables that are in kilometers,



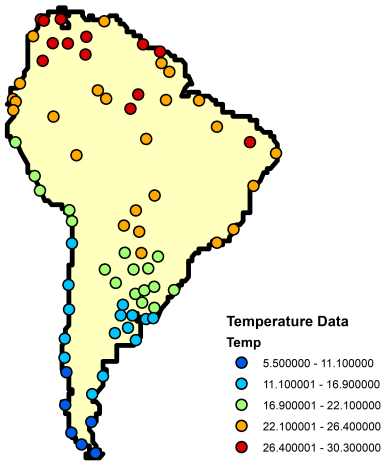


Figure 2.2a Prevailing Wind and Rainfall **Figure 2.2b Temperature Data Points**

which denote distances above and below the equator. We would first fit a generalized linear regression model and then fit a geo-regression with a quadratic spatial trend function to regress temperature data points in South America on distances to equator. The goal in fitting a geo-regression model is to remove the spatial dependencies among residuals and we will check our results by the nearest-neighbor regression procedure and provide useful prediction for the temperature in South America.

Methodologies and Results

Regression analysis to fit spatial models to the South America temperature data is first done by fitting a general linear regression (GLS) model. In general, spatial trends that vary smoothly over space tend to be well approximated (locally) by such polynomial functions. However, as covariance structure of the residuals may vary in different regression models, our primary interest in GLS models is to allow covariance structures to reflect spatially dependent random effects. We will first explore a spatial regression model with 76 data points {Si = (xi, yi): i=1,2,…76}. Here we fit temperature Yi by a linear function of these points,

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 y_i + u_i, i = 1,...,76$$

Then we examine the residuals of this regression using the Residual by Predicted Plot in JMP, which is shown in Figure 2.3.

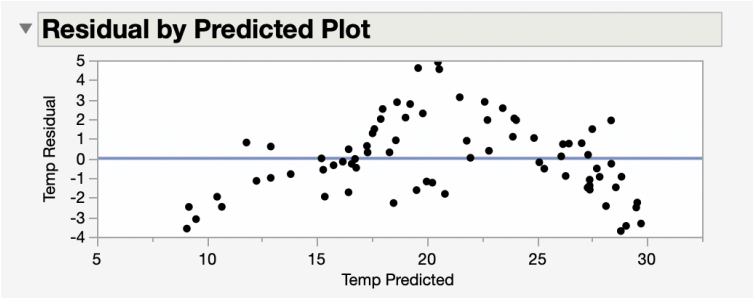


Figure 2.3 Residual by Predicted Plot

From the residual plot in Figure 2.3, it is clear that the residuals in the general linear regression model are not independent and there are possible spatial dependencies among residuals. In that sense, we will further explore possible spatial autocorrelation among the residuals by carrying out a nearest-neighbor regression on the residuals. We first construct nearest neighbor residual variable by using the residuals closest to each residual. As shown in Figure 2.4a, by regressing these residuals on their nearest-neighbor residuals, we can see that there indeed exist significant spatial dependency among the unobserved residuals, which is exemplified by the linear relationship. This significant spatial dependency is also shown by the p-value in the parameter estimates of nn_Res_1. Here we have a small p-value of less than 0.0001 (as shown in Figure 2.4b), which indicates strong evidence against the null hypothesis, which states there is no relationship between residuals and nearest neighbor residuals. As we reject the null hypothesis, we have strong evidence to state that there is significant spatial dependencies among the residuals of the GLS model.

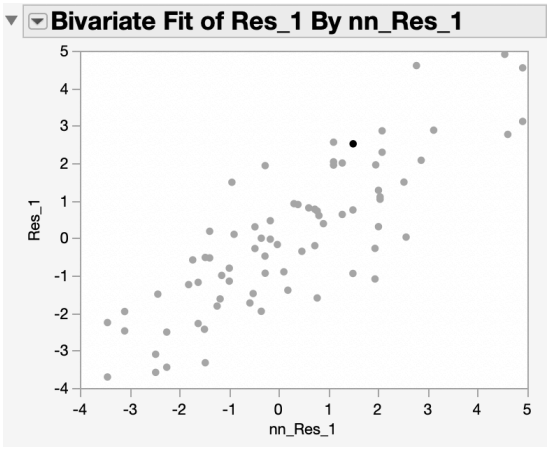


Figure 2.4a Nearest Neighbor Regression

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-0.156386	0.125283	-1.25	0.2159
nn_Res_1	0.8236498	0.063558	12.96	<.0001*

Figure 2.4b Parameter Estimates

Our observation that there exists significant spatial dependencies motivates us to perform an extended analysis using geo-regression to account for these dependencies. Notice in particular that the highest values tend to be in the north part of this continent, while the lowest values tend to be in the south corners. Thus we want to fit quadratic functions with the underlying coordinate variables, s=(x, y). This suggests that spatial trends in this data might be well fitted by a geo-regression with a quadratic spatial trend function of the form,

$$Y = \beta_0 + \beta_1 x + \beta_2 y + \beta_3 xy + \beta_4 x^2 + \beta_5 y^2 + \epsilon$$

Then we fit the quadratic regression in JMP and get the summary of fit in Figure 2.5a. According to p-values in parameter estimate, X and XX are not significant as their p-values are not small enough to reject the null hypothesis that there is no relationship between temperature and these two variables respectively. It is also important to note that the adjusted R-Square in this model is 0.9437. The adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases only if the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected by chance. As we find that X and XX are not

Summary of Fit

RSquare	0.947492
RSquare Adj	0.943741
Root Mean Square Error	1.454721
Mean of Response	21.01316
Observations (or Sum Wgts)	76

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	5	2673.0320	534.606	252.6242
Error	70	148.1349	2.116	Prob > F
C. Total	75	2821.1668		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	20.638954	4.944442	4.17	<.0001*
X	-0.001913	0.001441	-1.33	0.1886
Y	-0.001999	0.000713	-2.80	0.0065*
XY	-4.768e-7	9.918e-8	-4.81	<.0001*
XX	-1.483e-7	1.046e-7	-1.42	0.1605
YY	-1.846e-7	3.62e-8	-5.10	<.0001*

Figure 2.5a Full Quadratic Model

Summary of Fit

RSquare	0.945807
RSquare Adj	0.943549
Root Mean Square Error	1.457205
Mean of Response	21.01316
Observations (or Sum Wgts)	76

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	2668.2787	889.426	418.8597
Error	72	152.8882	2.123	Prob > F
C. Total	75	2821.1668		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	26.445443	0.256422	103.13	<.0001*
Y	-0.00223	0.000484	-4.61	<.0001*
XY	-5.046e-7	7.193e-8	-7.02	<.0001*
YY	-1.822e-7	3.605e-8	-5.05	<.0001*

Figure 2.5b Modified Quadratic Model

significant, we want to remove these two variables and rerun the quadratic regression with variables Y, XY, and YY. This time we get an adjusted R-Square of 0.9435, which is only slightly less than the adjusted R-Square we get from the full model. This means that the goodness of fit did not decrease as we remove X and XX from our quadratic regression model and thus help to justify the removal of X and XX.

Then we observe the parameter estimate of our modified quadratic model as shown in Figure 2.5b. As X and XX are not significant, we can say that longitude data is not significant in predicting the temperature in South America. Then we see that Y, XY and YY have very small p-values that means there are strong evidence that we can reject the null hypothesis that there is no relationship between temperature and the three variables respectively. We also perceive that variables Y, XY and YY have negative values of coefficient estimates and the absolute values of the coefficient estimates of XY and YY are larger than that of Y. As the positive and negative Y values denotes the distances above and below the equator, the negative values of coefficient estimate of Y, XY and YY means that as Y, XY, or YY gets larger, the temperature would get lower. In this case, most of the Y values are negative because most part of South America is below the equator.

Before we modify our quadratic regression model, Y is only weakly significant with a p-value of 0.0065. After we removed X and XX, Y is now very significant. This also tells us about the temperature trend that as the latitude gets far below the equator, the temperature would also get lower since the quadratic term of YY is positive if Y is negative. Then, we observe that X values are negative so XY is also positive. As Y gets far below the equator, the term XY would also be large and the predicted temperature would be low. Thus our result from the quadratic regression

model reflects what is observed in Figure 2.2b such that areas in the north are warmer than the areas in the south.

Then we also want to examine the Residual by Predicted Plot of this quadratic regression model. This is because we want to see whether this geo-regression has in fact removed the spatial dependencies among residuals. These residuals exhibit precisely the spatial covariance structure estimated by the geo-regression and we can see from Figure 2.6 that the residuals are almost random. This residual plots means that some of the spatial dependencies among residuals are removed with the quadratic regression model and we want to further examine the spatial autocorrelation by carrying out a nearest neighborhood regression on residuals as we did before.

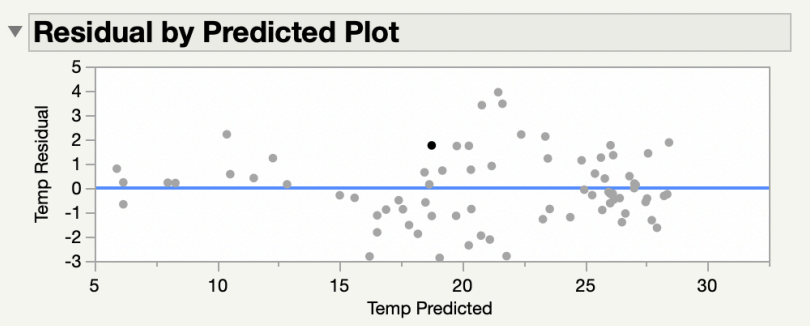


Figure 2.6 Residual by Predicted Plot

As shown in Figure 2.6, the predicted temperatures still show a little quadratic relationship with the with the residuals, so we may expect that there are slightly less spatial dependencies among the residuals than the previous GLS model. In light of the spatial dependencies, we examine the nearest neighborhood regression on residuals and the results is shown in Figure 2.7b. Here, we want to compare the nearest neighborhood regression of residuals of linear regression with the nearest neighborhood regression of residuals of quadratic regression as shown in Figure 2.7a and Figure 2.7b. It is obvious that there are still spatial dependencies in the residuals of the quadratic

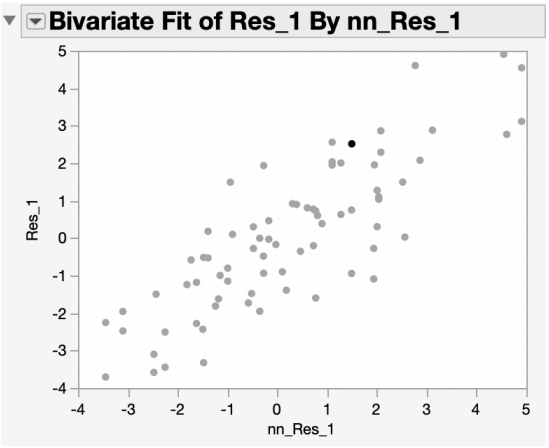


Figure 2.7a Linear Regression Residuals

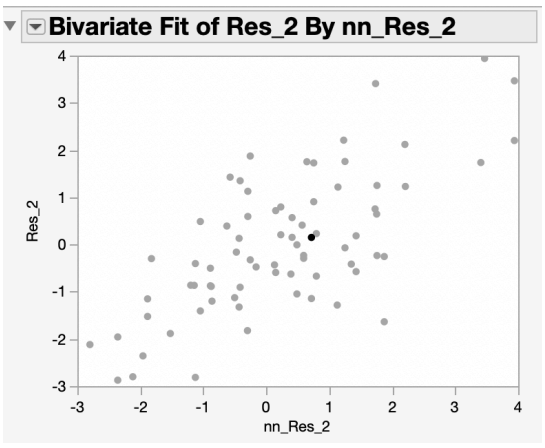


Figure 2.7b Quadratic Regression Residuals

regression, but the quadratic regression model is successful in removing some trace of spatial dependencies among residuals.

Finally, we want to construct spline interpolation using Spatial Analyst and display smoothed representations of these trend surfaces in ARCMAP. We first create an interpolation of the linear regression data for display using a raster spline in Spatial Analyst. Then we create a quadratic regression interpolation using a raster spline in Spatial Analyst. Our goal is to compare the quality of these two fits to the temperature data, both visually and quantitatively. As shown in Figure 2.8a and Figure 2.9b, we can see that the quadratic regression interpolation fits slightly better than the linear regression interpolation. In particular, by looking at the Northern tip of South America, as the contour lines for quadratic regression interpolation are curvilinear, we can better interpolate the temperatures with the quadratic regression model than the linear regression model.

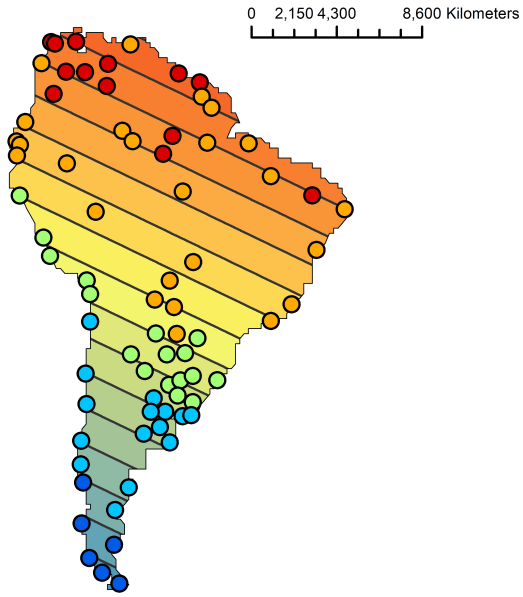


Figure 2.8a Linear Regression Interpolation

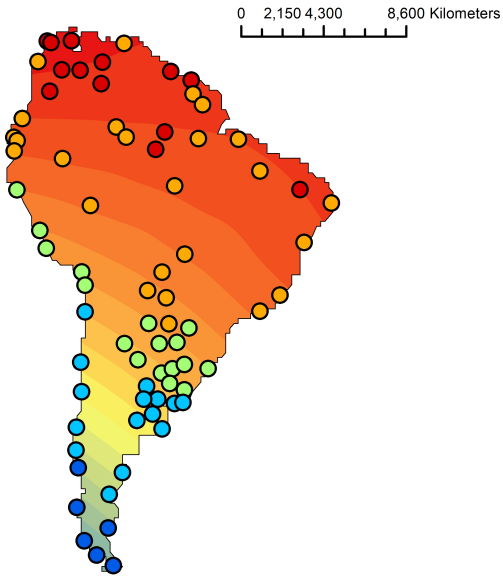


Figure 2.8b Quadratic Regression Interpolation

Now we want to make another comparison of the relative quality of the fits for these two regressions. We first find the city Georgetown, Guyana, whose FID in the original attribute table is 49. We then record that the mean annual temperature of Georgetown, Guyana is 27. Then, we use Identify tool to see that the approximate temperature value (Pixel value) predicted at P by the linear regression is 29.33. We also use Identify tool to see that the approximate temperature value (Pixel value) predicted at P by the quadratic regression is 27. As the approximate temperature value is more close to the mean annual temperature in the original data, it is obvious that the quadratic regression model performs better in predicting the mean annual temperature in South America.

Discussions

We construct two spatial prediction models to predict the mean annual temperature in South America by coordinate variables. The quadratic model performs much better than the linear model in predicting the temperature as we visually compare the two interpolations in ARCMAP and compare some specific temperature values with the original temperature data. This result points out the temperature in South America does not depends entirely on the location's longitude and latitude. Rather, it has a quadratic relationship with the latitude of the location, which may be caused by some other attributes such as the geographic characteristics or prevailing winds. This result corresponds with Norberto O. Garcia's study that he states the presence of the Andes range along the occidental coast modifies the climate of the western regions significantly. As we look at Figure 2.2b, we can indeed see some anomaly in mean annual temperature along the west coast, which may contribute to the quadratic relationship between the temperature and the latitude in South America.

Figure 2.9 Andes Range

References

Eidt R.C. (1969) The Climatology of South America. In: Fittkau E.J., Illies J., Klinge H., Schwabe G.H., Sioli H. (eds) Biogeography and Ecology in South America. Monographiæ Biologicæ, vol 18. Springer, Dordrecht.

García, Norberto O. "South American Climatology." *Quaternary International*, vol. 21, 1994, pp. 7–27., doi:10.1016/1040-6182(94)90018-3.