Statistics for analyzing and modeling precipitation isotope ratios in IsoMAP

The IsoMAP uses the multiple linear regression and geostatistical methods to analyze isotope data. Suppose the response variable $Y(s)$ and the vector of independent variable $x(s)$ are observed at a finite number of sites $s_1, \cdots, s_n$. The multiple linear regression method assumes $Y(s)$ and $Y(s')$ are independent but the geostatistical method assume they are not. In the following, we will first present the statistical methodology and then provide a real example to numerical interpret it.

1 Multiple Linear Regression

The multiple regression model provides the estimates of parameters and their ANOVA table. The multiple linear regression approach is to view the data as arising from independent observations, in which the statistical model is

$$Y(s) = x'(s)\beta + \epsilon(s), \tag{1}$$

where β is the vector of unknown parameters and $\epsilon(s)$ is an identically independent distributed (iid) error term with

$$\epsilon(s) \sim^{iid} N(0, \sigma^2).$$

In matrix notation, Model (1) can be expressed

$$Y = X\beta + \epsilon,$$

where $Y = (Y(s_1), \cdots, Y(s_n))'$, $X = (x(s_1), \cdots, x(s_n))'$ and $\epsilon = (\epsilon(s_1), \cdots, \epsilon(s_n))'$. It assumes

$$\epsilon \sim N(0, \sigma^2 I)$$

so that $Y \sim N(X\beta, \sigma^2 I)$. The unknown parameters β and σ^2 can be estimated by the least square (LS) method:

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{2}$$

and

$$\hat{\sigma}^2 = \frac{1}{n-p} Y'[I - X(X'X)^{-1}X']Y, \tag{3}$$

where p is the column dimension of X. The variance-covariance matrix of $\hat{\beta}$ is computed by

$$Cov(\hat{\beta}) = \hat{\sigma}^2 (X'X)^{-1}. \tag{4}$$

Let $\hat{\beta}_j$ be the j-th component of $\hat{\beta}$ and $\hat{\sigma}_{jk}$ be the (j,k) entry of $Cov(\hat{\beta})$. Then, its variance

$$V(\hat{\beta}_j) = \hat{\sigma}_{jj}$$

and its standard error is

$$std(\hat{\beta}) = \sqrt{\hat{\sigma}_{jj}}.$$

The t-value of $\hat{\beta}_j$ is

$$t(\hat{\beta}_j) = \frac{\hat{\beta}_j}{std(\hat{\beta}_j)}.$$

Let α be the significance level. If we test

$$H_0 : \beta_j = 0 \leftrightarrow H_1 : \beta_j \neq 0,$$

then we claim $\beta_j \neq 0$

$$|t(\hat{\beta}_j)| > t_{\alpha/2,n-p},$$

where $t_{\alpha/2,n-p}$ is the upper $\alpha/2$ quantile of t_{n-p} distribution. The p-value of the test is given by

$$P\{|t_{n-p}| > |t(\hat{\beta}_j)|\}.$$

The fitted value of the response variable at a general site s is

$$\hat{Y}(s) = x'(s)\hat{\beta}.$$

The sum of square of model (SSM) is defined by

$$SSM = \sum_{i=1}^{n} [\hat{Y}(s_i) - \bar{Y}]^2,$$

the sum of square of error (SSE) is defined by

$$SSE = \sum_{i=1}^{n} [Y(s_i) - \hat{Y}(s_i)]^2,$$

and the sum of square of total (SST) is defined by

$$SST = \sum_{i=1}^{n} [Y(s_i) - \bar{Y}]^2,$$

where

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y(s_i).$$

Then,

$$SSM = SST - SSE.$$

The R-square of the model is defined by

$$R^2 = \frac{SSM}{SST}.$$

It is known that $0 \leq R^2 \leq 1$. If R^2 is close to 1, then the fit is good; otherwise, the fit is bad. Usually, the fit of a model is considered good if $R^2 > 0.2$.

Table 1: Format of parameter estimates of the regression method

Variable	Estimate	Std	t-value	p-value
β_0	$\hat{\beta}_0$	$std(\beta_0)$	$\hat{\beta}_0/std(\hat{\beta}_0)$	$P\{\lvert t_{n-p}\rvert > \hat{\beta}_0/std(\hat{\beta}_0)\}$
β_1	$\hat{\beta}_1$	$std(\beta_1)$	$\hat{\beta}_1/std(\hat{\beta}_1)$	$P\{\lvert t_{n-p}\rvert > \hat{\beta}_1/std(\hat{\beta}_1)\}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
β_{p-1}	$\hat{\beta}_{p-}$	$std(\beta_{p-})$	$\hat{\beta}_{p-1}/std(\hat{\beta}_{p-1})$	$P\{\lvert t_{n-p}\rvert > \hat{\beta}_{p-1}/std(\hat{\beta}_{p-1})\}$

Table 2: Format of ANOVA of the regression method

Variable	Degree of Freedom	Sun of Suqare	Mean of Square	F-value	P-value
β_1	df_1	SS_1	$MS_1 = SS_1/df_1$	$F_1 = MS_1/MSE$	$P\{F_{df_1,n-p} > F_1\}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
β_{p-1}	df_{p-1}	SS_{p-1}	$MS_{p-1} = SS_{p-1}/df_{p-1}$	$F_{p-1} = MS_{p-1}/MSE$	$P\{F_{df_{p-1},n-p} > F_{p-1}\}$
Error	$n-p$	SSE	$MSE = SSE/(n-p)$		
Total	$n-1$	SST			

The IsoMAP reports the type I ANOVA by the stepwise method. To compute the ANOVA value of j-variable, it computes the SSM_1 of the first model

$$Y(s) = \beta_0 + \beta_1 x_1(s) + \cdots + \beta_{j-1} x_{j-1}(x) + N(0, \sigma^2)$$

and SSM_2 the second model

$$Y(s) = \beta_0 + \beta_1 x_1(s) + \cdots + \beta_{j-1} x_{j-1}(x) + \beta_j x_j(s) + N(0, \sigma^2).$$

Then, the ANOVA value of β_j is

$$SS_j = SSM_2 - SSM_1 = SSE_1 - SSE_2.$$

Let df_j be the degrees of freedom of j-th varaible, then its p-value is

$$P\{F_{df_j,n-p} > F_j\}$$

where

$$F_j = \frac{SS_j/df_j}{SSE/(n-p)} = \frac{MS_j}{MSE}, j = 1, \cdots, p-1.$$

If the p-value is less than the significance level, then β_j is significantly different from 0.

To summarize, we display the format of parameter estimates in Table 1 and the format of ANOVA in Table 2

2 Geostatistical Model

The approach taken here is to view the data as arising from geological space where each observation corresponds to a spatial location. From this perspective, it might be appropriate to call our method a geostatistical method [2]. The statistical model describes the relationship between the response variable and the independent variables according to a geostatistical model in which $\epsilon(s)$ in Equation (1) is assumed a spatially correlated error term. The error term $\epsilon(\mathbf{s})$ are normally distributed with zero mean and a certain covariance function. The covariance function is assumed stationary and isotropic, which has the form of

$$Cov(\epsilon(s), \epsilon(s+h)) = \sigma^2 \rho(u),\ u = \|h\|, \tag{5}$$

where $\sigma^2 = V[\epsilon(s)]$ is the variances of the Gaussian process and $\rho(u)$ is the correlation function. A legitimate correlation function must be positive-definite. In practice, this is usually ensured by working within one of several standard classes of parametric models. Overall, one can generally choose the well-known Matérn correlation function [9] given by

$$\rho_\theta(u) = \frac{\theta_1}{2^{\theta_3-1}\Gamma(\theta_3)}(\frac{u}{\theta_2})^{\theta_3} K_{\theta_3}(\frac{u}{\theta_2}),\ \theta = (\theta_1, \theta_2, \theta_3),\ u > 0, \tag{6}$$

where $0 \leq \theta_1 \leq 1$, $\theta_2 > 0$ and $K_{\theta_3}(\cdot)$ is the modified Bessel function. The range parameter θ_2 controls the rate of decay of $\rho_\theta(u)$ between observations as distance increases. The smoothness parameter θ_3 controls the behavior of the smoothness of $\rho_\theta(u)$. The Matérn class includes the exponential correlation function when $\theta_3 = 0.5$. The correlation function also includes the nugget effect, where the nugget effect is present if $\theta_1 < 1$. In addition, θ_1 also controls the magnitude of spatial dependence in which a geostatistical model reduces to a multiple regression model if $\theta_1 = 0$. The unknown parameter β in Equation (6) along with the correlation function $\rho_\theta(u)$ reflects the spatial distribution of the response variation $Y(s)$ (i.e. δ^{18}O in this article).

In matrix notation, Models (1) and (5) can be expressed as

$$Y = X\beta + \epsilon$$

where

$$R_\theta = Corr(Z) = \begin{pmatrix} 1 & \rho_\theta(d_{12}) & \cdots & \rho_\theta(d_{1n}) \\ \rho_\theta(d_{21}) & 1 & \cdots & \rho_\theta(d_{2n}) \\ \vdots & \vdots & \ddots & \vdots \\ \rho_\theta(d_{n1}) & \rho_\theta(d_{2n}) & \cdots & 1 \end{pmatrix} \tag{7}$$

where d_{ij} is the distance between sites s_i and s_j.

We use the maximum likelihood method to estimate β and θ, which estimates β and θ by maximizing the following loglikelihood function

$$\ell(\beta, \sigma^2, \theta) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2}\log|\det(R_\theta)| - \frac{1}{2\sigma^2}(Y - X\beta)^t R_\theta^{-1}(Y - X\beta). \tag{8}$$

Because it is difficult to maximize $\ell(\beta, \sigma^2, \theta)$ with respect to β, σ^2 and θ simultaneously, we use the profile likelihood method to first compute the MLE of θ and then compute the MLE of β, σ^2. We briefly introduce our algorithm below.

If θ is given, then the (conditional) MLE of β can be derived by the generalized least square (GLS) method:

$$\hat{\beta}_\theta = (X' R_\theta^{-1} X)^{-1} X' R_\theta^{-1} Y \tag{9}$$

and

$$\hat{\sigma}_\theta^2 = \frac{1}{n} Y' M_\theta Y, \tag{10}$$

where

$$M_\theta = R_\theta^{-1} - R_\theta^{-1} X (X' R_\theta^{-1} X)^{-1} X' R_\theta^{-1}.$$

Put (9) and (10) into (8). We have the profile loglikelihood function

$$\ell_P(\theta) = -\frac{n}{2}[1 + \log(\frac{2\pi}{n})] - \frac{1}{2}\log|\det(R_\theta)| - \frac{n}{2}\log(Y^t M_\theta Y). \tag{11}$$

The profile maximum likelihood estimate (PMLE) of θ can be derived by using the Newton-Raphson algorithm. Because the smoothness parameter θ_3 is sensitive in the algorithm, we compute the PMLE of θ_1 and θ_2 conditioning on θ_3 and then find the best θ_3 by a one-dimensional optimization method (e.g. the Golden Section Algorithm). If $\hat{\theta}$ is derived, then the MLE of β and σ^2 can be derived by using (9) and (10).

When $\hat{\theta}$, $\hat{\sigma}^2$ and $\hat{\beta}$ are derived, the interpolation of $Y(s)$ at an unobserved site s_0 can be derived by using the universal kriging method. Let $x_0 = x(s_0)$ be the vector of independent variables at site s_0 and

$$c_0 = Corr(Y(s_0), Y) = (\rho_{\hat{\theta}}(d_{01}), \cdots, \rho_{\hat{\theta}}(d_{0n})),$$

where d_{0i} is the distance between sites s_0 and s_i. The universal kriging method interpolates $Y(s_0)$ be $Y^*(s_0)$ with the conditional expected value as

$$Y^*(s_0) = E[Y(s_0)|Y] = x_0'\hat{\beta} + c_0' R_{\hat{\theta}}^{-1}(Y - X\hat{\beta}). \tag{12}$$

The variability of the universal kriging interpolation is given by the mean squared prediction error (MSPE), which is

$$\begin{aligned} MSPE[Y^*(s_0)] =& E[Y(s_0) - Y^*(s_0)]^2 \\ =& \sigma^2[1 + x_0'(X' R_{\hat{\theta}}^{-1} X)^{-1} x_0 - c_0' R_{\hat{\theta}}^{-1} X (X' R_{\hat{\theta}}^{-1} X)^{-1} X' R_{\hat{\theta}}^{-1} c_0]. \end{aligned} \tag{13}$$

The detail of the derivation of Equations (12) and (13) can be found in [10].

The aim of model selection is to find the best linear function $x'(s)\beta$ in Equation (1). The best linear function $x'(s)\beta$ is selected from a collection of candidates. The AIC of a specific model is

$$AIC = -2\ell(\hat{\beta}, \hat{\sigma}^2, \hat{\theta}) + 2k$$

where $\hat{\beta}$, $\hat{\sigma}^2$, $\hat{\theta}$ are the MLE of β, σ^2 and θ, and k is the number of parameters under a specific model. Let $Y^*_{(i)}(s_i)$ be the kriging interpolation of $Y(s_i)$ if it is excluded from the dataset. Then, the CV of a model is given by

$$CV = \frac{1}{n}\sum_{i=1}^{n}[Y(s_i) - Y^*_{(i)}(s_i)]^2.$$

The best model has the lowest AIC or CV value.

3 Moran's I Test for Spatial Dependence

A number of permutation testing methods have been proposed to test for spatial dependence. Almost all of them need a well-defined measure of the closeness (or weight) between two units. We choose Moran's I [8] in IsoMAP because it is the most popular one.

Let $\hat{\epsilon}_i$ be the residual of the regression model at i and w_{ij} be the measure of the closeness between units i and j. Moran's I is defined as

$$I = \frac{1}{S_0 b_2}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} w_{ij}\hat{\epsilon}_i\hat{\epsilon}_j, \tag{14}$$

where $S_0 = \sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} w_{ij}$ and $b_k = \sum_{i=1}^{n}\hat{\epsilon}_i^k/n$. Moran's I statistic usually ranges between -1 and 1 even though its absolute value could be over 1 in extreme cases. With a coefficient close to -1, Moran's I indicates neighborhood dissimilarity; with a coefficient close to 1, Moran's I indicates neighborhood similarity. When the coefficient of Moran's I is close to 0, it indicates spatial randomness or independence [1].

The p-values of Moran's I is computed under random permutation test schemes. A random permutation test generally calculates the moments of the test statistic under every possible arrangement of the data. These arrangements would be used to generate the distribution of the test statistic under the null hypothesis. A related approach uses many Monte Carlo rearrangements of the data rather than enumeration of all of the possible arrangements [3]. If the number of Monte Carlo rearrangements is large and each arrangement has equal probability in each Monte Carlo replicate, then the exact test and Monte Carlo permutation test will have similar results, and the Monte Carlo permutation test is asymptotically equivalent to the exact permutation test ([4] 185-187). Note that for a general n, there are $n!$ possible permutation arrangements in the exact permutation test. It is generally impossible to obtain the exact moments of a test statistic under the permutation test scheme. However, since the numerator of Moran's I is in quadratic form and their denominators are permutation invariant, the exact expressions of their moments under random permutation test scheme are available and have been included in the many textbooks (e.g. see Cliff and Ord [1]). In the following, we denote $E_R(\cdot)$ and $V_R(\cdot)$ as the expected value and variance of a statistic under the exact permutation test scheme. Formulae of the moments then are given below accordingly:

$$E_R(I) = -\frac{1}{n-1},$$

and

$$E_R(I^2) = \frac{S_1(nb_2^2 - b_4)}{S_0^2 b_2^2 (n-1)} + \frac{(S_2 - 2S_1)(2b_4 - nb_2^2)}{S_0^2 b_2^2 (n-1)(n-2)} + \frac{(S_0^2 - S_2 + S_1)(3nb_2^2 - 6b_4)}{S_0^2 b_2^2 (n-1)(n-2)(n-3)},$$

where $B_k = \sum_{i=1}^n z_i^k/n$, $S_1 = \sum_{i=1}^n \sum_{j=1, j\neq i}^n (w_{ij} + w_{ji})^2/2$ and $S_2 = \sum_{i=1}^n [\sum_{j=1, j\neq i}^n (w_{ij} + w_{ji})]^2$. The variances is

$$V_R(I) = E_R(I^2) - E_R^2(I),$$

Assume I is asymptotic normal. Then, it p-value is calculated by a two-sided z-test according to

$$2[1 - \Phi(|\frac{I - E_R(I)}{\sqrt{V_R(I)}}|)].$$

If the p-values are less than the significance level (e.g. 0.05), we conclude the significance of spatial dependence.

In the IsoMAP, we take z_i as the i-th residual of the regression model and choose w_{ij} by the k-nearest neighbor method. The k-nearest neighbor is defined by $w_{ij} = 1$ if the distance between s_i and s_j are among the least k distances between s_i and all the rest sites and $w_{ij} = 0$ otherwise.

4 An Example

The example is available at the homepage of IsoMAP with Key 5373, case name 'O18Global', and start time August 15, 2011. In this example, we selected data from 1980 to 1995 which included all the months from South Pole to North Pole. The final dataset contained 341 stations. We chose δ^{18}O as response variable, and elevation, absolute latitude, and the square of latitude as independent variables. Then, the statistical model was

$$\delta^{18}\mathrm{O}(s) = \beta_0 + \beta_1 e(s) + \beta_2 |l(s)| + \beta_3 l^2(s) + e(s) + \epsilon(s),$$

where $e(s)$ was the elevation and $l(s)$ was the latitude of site s.

We computed the spherical distance between sites s_i and s_j according to the formula

$$\begin{aligned} d_{ij} =& 2R_E \arcsin\{\frac{1}{2}[(\cos l_a(s_i)\cos l_o(s_i) - \cos l_a(s_j)\cos l_o(s_j))^2 \\ &+ (\cos l_a(s_i)\sin l_o(s_i) - \cos l_a(s_j)\sin l_o(s_j))^2 + (\sin l_a(s_i) - \sin l_a(s_j))^2]^{1/2}\}, \end{aligned} \tag{15}$$

where $R_E = 6378.1$km was the radius of the Earth. We used d_{ij} to define the Matérn correlation function $\rho_\theta(u)$ in Model (6) and the correlation matrix R_θ in Model (7).

The ANOVA and parameter estimates are given by Tables 3 and 4. The estimate of the variance of the error term then is

$$\hat{\sigma}^2 = MSE = \frac{1851.66}{337} = 5.49454.$$

Therefore, the fitted multiple linear regression model was

$$\delta^{18}\mathrm{O}(s) = -5.689 - 0.00168 Elev(s) + 0.2053 |Lat(s)| - 0.00538 Lat^2(s) + \epsilon(s)$$

Table 3: ANOVA of the multiple regression model

Variable	df	SS	MS	F-value	p-value
Elevation (β_1)	1	58.0629	58.0629	10.5674	0.00126721
Abs(Latitude) (β_2)	1	4024.97	4024.97	732.539	0
Latitude Square (β_3)	1	449.111	449.111	81.7377	0
Error	337	1851.66	5.49454		
Total	340	6383.8			

Table 4: Parameter estimates of multiple regression model

Variable	Estimate	Standard Error	t-value	p-value
β_0	-5.68898	0.249172	-22.8315	0
β_1	-0.00168255	0.000106863	-15.7449	0
β_2	0.20534	0.0148348	13.8418	0
β_3	-0.00538409	0.000209837	-25.6584	0

with $\epsilon(s) \sim^{iid} N(0, 5.49454)$. This model can be used to interpolate δ^{18}O be the multiple regression method. The QQ-plot in the regression method was almost a straight line, which implies there was no significant variation of the normal assumption. The residual plot in the regression method showed randomness which implies no additional transformation is required for the response variable (e.g. by using the Box-Cox transformation).

We also fitted a geostatistical model, the parameter estimates and ANOVA table are displayed in Tables 5 and 6 respectively. The fitted correlation function was

$$\begin{aligned}\rho_\theta(u) =& \rho_{(0.8820,1092.53,1)}(u)\\ =& 0.8820(\frac{u}{1092.53})K_1(\frac{u}{1092.53})\end{aligned}$$

for $u > 0$. Because the p-value of Moran's I was almost 0, we recommend to use the geostatistical model instead of the regression model.

Table 5: ANOVA of the multiple regression model

Variable	df	SS	MS	F-value	p-value
Elevation (β_1)	1	1449.74	1449.74	191.415	0
Abs(Latitude) (β_2)	1	693.622	693.622	91.5813	0
Latitude Square (β_3)	1	55.9661	55.9661	7.3894	0.00690032
Error	337	2552.38	7.57384		
Total	340	1449.74			

Table 6: Parameter estimates of multiple regression model

Variable	Estimate	Standard Error	t-value	p-value
β_0	-4.43342	1.06554	-4.16071	$4.03135e-05$
β_1	0.00165697	0.000116952	-14.168	0
β_2	0.162598	0.0598151	2.71835	0.00690032
β_3	-0.00500147	0.000812748	-6.15379	$2.14676e-09$

References

[1] Cliff, A.D. and J.K. Ord. *Spatial Processes: Models And Applications*, Pion, London, 1981.

[2] Cressie, N. (1989), "Geostatistics", *The American Statistician*, **42**, 197-202.

[3] Dwass, M. Modified randomization tests for nonparametric hypothesis. *Annals of Mathematical Statistics*, 28:181-187, 1957.

[4] Good, P. *Permutation Tests*. Springer, New York, 2000.

[5] Hoeting, J.A., Davis, R.A., Merton, A.A. and Thompson, S.A. (2006), "Model Selection for Geostatistical Models", *Ecological Application*, **16**, 87-98.

[6] Huang, H.C. and Chen, Chen, C.S. (2007), "Optimal Geostatistical Model Selection," *Journal of American Statistical Association*, **102**, 1009-1024.

[7] Huang, H.C, Martinez, F., Mateu, J. and Montes, F. (2007), "Model Comparison and Selection for Stationary Space-Time Models", *Computational Statistics and Data Analysis*, **51**, 4577-4596.

[8] Moran, P.A.P. (1948), "The Interpretation of Statistical Maps," *Journal of the Royal Statistic Society Series B*, **10**, 243-251.

[9] Matérn, B. (1986). *Spatial Variation*, 2nd Edition (Lecture Notes in Statist.3 6). Springer, Berlin.

[10] Stein, A. and Corsten, L.C.A. (1991). Universal kriging and cokriging as regression procedure. *Biometrics*, **47**, 575-587.