A BRIEF NOTE ON REGIONALIZATION

H.L. Gray and Gene Herrin

Technical Report No. SMU/DS/TR/230
Department of Statistical Science, DARPA Contract

August 1989

Research sponsored by DARPA
Contract F19628-88-K-0042

A BRIEF NOTE ON REGIONALIZATION

by

H. L. Gray and Gene Herrin

I. We assume a region R composed of regions R_1 and R_2, i.e. $R = R_1 \cup R_2$, R_1, $R_2 \neq \phi$. Using the notation $mb_{ji} \in R_j$ to denote the occurrence of the ith event in R_j, let

$$mb_{ji} = A_j + B_j W_i + \epsilon_{ji}, \quad mb_{ji} \in R_j, \quad j = 1,2$$

and let

$$mb_i = A + BW_i + \epsilon_i,$$

where mb_i is equally likely to be from R_1 as R_2. Further assume $\epsilon_{ji} \sim N(0,\sigma^2)$ and let $X_i = mb_i - BW_i$. Then

$$E[X] = \tfrac{1}{2}(A_1 + A_2) = A$$

and (1)

$$\sigma_X^2 = \sigma^2 + d^2, \quad \text{where } 2d = A_2 - A_1 .$$

Without loss in generality, we assume $A_2 \geq A_1$.

<u>NOTE</u>: The distribution of X is unimodal unless

$$|A_2 - A_1| > 2\sigma.$$

Now the distribution of X is not normal but is a mixture of normals. Treating X as normal can lead to significant errors if d is large. However we will now show that if $2d \simeq \sigma$ or $2d = \sigma_X$ (which is approximately the case in the Shagon region) then X can be treated as normal for all practical purposes. This can be shown as follows. Let Z_α be the 100(1-α) percentile point of a N(0,1) distribution. And we have

$$P\left[\frac{X-A}{\sigma_X} \leq Z_\alpha \mid X \in R \right] = 1 - \beta . \quad (2)$$

If X is normal $\alpha = \beta$. The question here is how much does α differ from β. But, again assuming $P[X \in R_1] = \frac{1}{2}$, we have

$$P\left[\frac{X-A}{\sigma_X} \le Z_\alpha \mid X \in R\right] = \frac{1}{2}P\left[\frac{X-A}{\sigma_X} \le Z_\alpha \mid X \in R_1\right] + \frac{1}{2}P\left[\frac{X-A}{\sigma_X} \le Z_\alpha \mid X \in R_2\right]$$

$$= \frac{1}{2}P\left[\frac{X-A_1}{\sigma} \le \frac{Z_\alpha \sigma_X + d}{\sigma} \mid X \in R_1\right] \tag{3}$$

$$+ \frac{1}{2}P\left[\frac{X-A_2}{\sigma} \le \frac{Z_\alpha \sigma_X - d}{\sigma} \mid X \in R_2\right].$$

Now suppose $2d = \sigma$. Then from (3)

$$P\left[\frac{X-A}{\sigma_X} \le Z_\alpha \mid X \in R\right] = \frac{1}{2}P\left[\frac{X-A_1}{\sigma} \le \frac{\sqrt{5}\,Z_\alpha + 1}{2} \mid X \in R_1\right]$$

$$+ \frac{1}{2}P\left[\frac{X-A_2}{\sigma} \le \frac{\sqrt{5}\,Z_\alpha - 1}{2} \mid X \in R_2\right],$$

or

$$P\left[\frac{X-A}{\sigma_X} \le Z_\alpha \mid X \in R\right] = \frac{1}{2}\left[\Phi\left(\frac{\sqrt{5}\,Z_\alpha + 1}{2}\right) + \Phi\left(\frac{\sqrt{5}\,Z_\alpha - 1}{2}\right)\right] = 1 - \beta, \tag{4}$$

where Φ is the standard normal distribution function. Equation 4 can now be used to find β. Table 1 below compares α with β when $2d = \sigma$. It is very clear that in the tails of the distribution there is no practical loss in treating X as normal when the A_i are this close together.

TABLE 1

α	β
.100	.101
.050	.050
.025	.025

If d is slightly larger the approximation by a single normal is still valid, although not quite as good. For example, if $2d = \sigma_X$, then rather than Equation (4) we obtain

$$P\left[\frac{X-A}{\sigma_X} \leq Z_\alpha \mid X \in R\right] = \frac{1}{2}\Phi\left(\frac{2Z_\alpha + 1}{\sqrt{3}}\right) + \frac{1}{2}\Phi\left(\frac{2Z_\alpha + 1}{\sqrt{3}}\right) = 1 - \beta ,$$

Table 2 below compares α and β for the case $2d = \sigma_X$.

TABLE 4

α	β
.100	.102
.050	.050
.025	.024
.010	.009

It is worth noting that the above results assume that X is equally likely to come from R_1 as R_2. If this is not the case then

$$A \neq \frac{A_1 + A_2}{2}.$$

In fact if we assume $X \in A_1$ with probability p_1 and $X \in A_2$ with probability p_2, then $p_1 + p_2 = 1$ and

$$E[X] = A = p_1 A_1 + p_2 A_2$$

and (5)

$$\sigma_X^2 = \sigma^2 + 4p_1p_2d^2,$$

$2d = A_2 - A_1$.

Note that not only does $A \neq \frac{A_1 + A_2}{2}$ but $\sigma_X^2 \neq \sigma^2 + d^2$.

The implication of (5) is that care must be taken in formulating the assumptions when data is best modeled as regionalized. It is interesting to note that although the mean and variance are strongly effected by the p_i, the approximation of the mixture distribution as a single normal is very robust to the p_i and actually depends primarily on the magnitude of d. That is, following our previous steps it is easy to show in general that

$$1-\beta = P\left[\frac{X-A}{\sigma_X} \leq Z_\alpha \mid X \in R\right] = p_1\Phi(\sqrt{1+p_1p_2}\, Z_\alpha + p_2) + p_2\Phi(\sqrt{1+p_1p_2}\, Z_\alpha - p_1) \qquad (6)$$

Equation 6 can be used to calculate β. Table 3 shows that treating X as Normal when $2d = \sigma$

introduces only a small error. The result is entirely similar if $2d = \sigma_X$.*

TABLE 3

p_2	α	β	p_2	α	β
.1	.100	.111	.7	.100	.099
.1	.050	.052	.7	.050	.047
.1	.025	.027	.7	.025	.022
.3	.100	.102	.9	.100	.098
.3	.050	.052	.9	.050	.048
.3	.025	.027	.9	.025	.024

II. <u>Normalized mb and normalized mLg</u>

We assume the basic relationship in I, where A_1, A_2 and A are known, at least to a stated precision. Now we assume that there exist constants δ, δ_1 and β, β_1, β_2 such that

$$mLg_i = \delta + \beta \log Y_i + e_i \quad \text{over } R,$$

$$mLg_{ji} = \delta_j + \beta_j \log Y_i + e_{ji} \text{ over } R_j, \quad (j=1,2). \tag{7}$$

However (following the same reasoning as Russell) since Y_i is unknown the δ's and β's cannot be estimated directly. Now by (1) and (7) (we restrict ourselves for now to R but the procedure is the same over R_j)

$$\frac{mb_i - A}{B} = W_i + \frac{\epsilon_i}{B}, \quad \frac{mLg_i - \delta}{\beta} = W_i + \frac{e_i}{\beta}. \tag{8}$$

Therefore

$$\begin{aligned} mLg_i &= \frac{\beta}{B} mb_i + \delta - \frac{\beta}{B} A + e_i - \frac{\beta e_i}{B} \\ &= C + Dmb_i + e_i', \end{aligned} \tag{9}$$

---

*If $2d=\sigma_X$, then $1-\beta = p_1\Phi\left(\frac{Z_\alpha - p_1}{\sqrt{1-p_1p_2}}\right) + p_2\Phi\left(\frac{Z_\alpha + p_2}{\sqrt{1-p_1p_2}}\right)$.

For specific p_i and α the results are about the same as Table 3.

where

$$C = \delta - \frac{\beta}{B}A, \quad D = \frac{\beta}{B}, \qquad e_i' = e_i - \frac{\beta}{B}\epsilon_i \, . \tag{10}$$

Since mb_i and mLg_i are observable, least squares estimates can be obtained for C and D. That is

$$\hat{C} = \overline{mLg} - \hat{D}\,\overline{mb}, \qquad \hat{D} = \frac{\Sigma(mLg_i - \overline{mLg})(mb_i - \overline{mb})}{\Sigma(mb_i - \overline{mb})^2} \, . \tag{11}$$

From (10) we have

$$\hat{\beta} = B\hat{D}, \quad \hat{\delta} = \hat{C} + \hat{\beta}\,\frac{A}{B} \, . \tag{12}$$

Now let

$$mb_i^* = \frac{mb_i - A}{B} \quad \text{and } mLg_i^* = \frac{mLg_i - \hat{\delta}}{\hat{\beta}} \, . \tag{13}$$

Then

$$mb_i^* = W_i + \frac{\epsilon_i}{B} \quad \text{and } mLg_i^* = (W_i + \frac{e_i}{\beta})\,\frac{\beta}{\hat{\beta}} + \frac{\delta - \hat{\delta}}{\hat{\beta}} \, .$$

Note that if $\beta = \hat{\beta}$, then $E[\hat{\sigma}] = \delta$ and

$$mLg_i^* = W_i + \frac{e_i}{\beta} + \frac{\delta - \hat{\delta}}{\beta} = W_i + \frac{e_i^*}{\beta}, \quad e_i^* = e_i + \delta - \hat{\delta} \, ,$$

and hence $E[mLg_i^*] = W_i$. Of course $\hat{\beta} \neq \beta$ and therefore $E[mLg_i^*] \neq Y_i$. However, as has been pointed out in the past, errors in $\hat{\beta}$ when Y_i is in the neighborhood of 150 are not very significant.

<u>Correlation over the Partitioned Sites when $\hat{\beta} \equiv \beta$.</u>

Following our previous notation, let

$$mb_{ji}^* = \frac{mb_{ji} - A_j}{B} \quad \text{and } mLg_{ji}^* = \frac{mLg_{ji} - \hat{\delta}_j}{\hat{\beta}_j} \, , \quad j = 1, 2 \, , \tag{14}$$

where we assume β, $\hat{\beta}_j$, and A_j are known and $\hat{\delta}_j$ is determined by (12) with the data restricted to R_j.

Then

$$\mathrm{Cov}\left\{(mb_i^*-w_i),(mLg_i^*-w_i)|R\right\} = \tfrac{1}{2}\mathrm{Cov}\{(mb_i^*-w_i),(mLg_i^*-w_i)|R_1\}$$

$$+ \tfrac{1}{2}\,\mathrm{Cov}\{mb_i^* - w_i),(mLg_i^*-w_i)|R_2\}$$

$$= \tfrac{1}{2}\{\mathrm{Cov}[(mb_{1i}^*-w_i),(mLg_{1i}^*-w_i)] + \mathrm{Cov}[(mb_{2i}^*-w_i),(mLg_{2i}^*-w_i)]\}. \quad (15)$$

Now let

$$\sigma_{mb^*}(R_j) = \sigma_1 \quad \text{and } \sigma_{mLg^*}(R_j) = \sigma_2, \quad j = 1,2.$$

Then from (15), denoting the correlation between mg* and mLg* by ρ_{12}^*,

$$\rho_{12}^*(R) = \frac{\sigma_1\sigma_2}{2\sigma_{mb^*}(R)\sigma_{mLg^*}(R)}\left[\rho_{12}^*(R_1) + \rho_{12}^*(R_2)\right] . \quad (16)$$

Therefore if $\rho_{12}^*(R_1) = \rho_{12}^*(R_2)$, then

$$\rho_{12}^*(R) = \frac{\sigma_1\sigma_2}{\sigma_{mb^*}(R)\sigma_{mLg^*}(R)}\,\rho_{12}^*(R_1) . \quad (17)$$

From (17), if also $\sigma_{mLg^*}(R) = \sigma_{mLg^*}(R_j)$, then

$$\rho_{12}^*(R) = \frac{\sigma_1}{\sigma_{mb^*}(R)}\,\rho_{12}^*(R_1) \quad (18)$$

or

$$\rho_{12}^*(R_2) = \rho_{12}^*(R_1) = \frac{\sigma_{mb^*}(R)}{\sigma_1}\,\rho_{12}^*(R)$$

$$= \frac{\sigma_{mb^*}(R)}{\sqrt{\sigma^2_{mb^*}(R)\text{-}d^2}}\,\rho_{12}^*(R)\,. \tag{19}$$

Note that if $2d = \sigma_{mb^*}(R)$, then

$$\rho_{12}^*(R_1) = \frac{2}{\sqrt{3}}\,\rho_{12}^*(R) \doteq 1.15\,\rho_{12}^*(R)\,. \tag{20}$$

But if $2d = \sigma_{mb*}(R_1)$, then

$$\rho_{12}^*(R_1) = \frac{\sqrt{5}}{2}\,\rho_{12}^*(R) \doteq 1.12\,\rho_{12}^*(R)\,. \tag{21}$$

Notice that (assuming the β_j are known) from Section I it follows that if $\sigma_{mLg}(R_j) = \sigma_1$, then $\sigma_{mLg}(R) = \sigma_{mLg}(R_j)$ if and only $E[mLg_{1i}] = E[mLg_{2i}] = E[mLg_i]$. Therefore, if we claim $\beta_j = \beta$, $\sigma_{mLg}(R) = \sigma_{mLg}(R_j)$ if and only if $\delta_j = \delta$. From (12) it is clear, that if independent estimates (independent of δ_j) of A_1 and A_2, can be obtained such as that obtained for A, then we can test the hypothesis that all of the reduction in variance of the difference in mb and mLg is due to a reduced variance in mb by testing the hypothesis that $\delta_1 = \delta_2$. Even if $\beta_j \neq \beta$ but we assumed the β_j are known, the hypothesis can be tested.

III. <u>The Unified estimate</u>

Let

$$u = \frac{mb^* + mLg^*}{2}\,, \quad \lambda^2 = Var[mLg^* \text{ - } mb^*].$$

Then

$$\sigma_u^2 = \tfrac{1}{4}(\sigma^2_{mb^*} + \sigma^2_{mLg^*} + 2\,(\sigma_{mb^*}\sigma_{mLg^*}\rho^*_{12})\,, \tag{22}$$

and

$$\lambda^2 = \sigma^2_{mb^*} + \sigma^2_{mLg^*} - 2\,\sigma_{mb^*}\sigma_{mLg^*}\rho^*_{12} \,. \tag{23}$$

Therefore

$$\sigma_{mLg^*} = \rho^*_{12}\sigma_{mb^*} + [\sigma^2_{mb^*}(\rho^{*2}_{12} - 1) + \lambda^2]^{\frac{1}{2}} \;. \tag{24}$$

If we are given σ_{mb^*}, λ, and ρ^*_{12}, we can calculate σ_{mLg^*} from (24) and hence obtain σ_u and the accompanying F number.

The assumption that σ_{mb^*} and ρ^*_{12} are known can be reduced to <u>either</u> σ_{mb^*} or ρ^*_{12} is known by the added assumption that $\sigma_{mb^*} = \sigma_{mLg^*}$. In this case, it follows at once from (23) that

$$\sigma_{mb^*} = \frac{\lambda}{\sqrt{2(1-\rho^*_{12})}} \tag{25}$$

and, for each fixed ρ^*_{12},

$$(\max)\sigma^2_u = \frac{\sigma^2_{mb^*}}{2}(1 + \rho^*_{12}) = \frac{\lambda^2}{4}\,\frac{1 + \rho^*_{12}}{1 - \rho^*_{12}} \quad . \tag{26}$$

As a function of σ_{mb^*} and σ_{mLg^*}, i. e. for fixed ρ^*_{12}, σ^2_u is clearly a maximum when $\sigma_{mb^*} = \sigma_{mLg^*}$. Equation 26 gives an expression for max σ^2_u as a function of ρ^*_{12}. It could of course be just as easily given in terms of σ_{mb^*}. That is, the assumption $\sigma_{mb^*} = \sigma_{mLg^*}$ also easily yields, for each fixed ρ^*_{12} ,

$$(\max)\sigma^2_u = \sigma^2_{mb^*} - \frac{\lambda^2}{4} \quad . \tag{27}$$

Whether you use (26) or (27) depends on whether you transport σ_{mb^*} or ρ^*_{12}. If you transport <u>both</u> you get an estimate of σ^2_u instead of an estimated bound on σ^2_u .

<u>The Unified Estimate on Partitioned Site</u>

If A_1 and A_2 are given and δ_1 and δ_2 are estimated, from our previous remarks,

$$\sigma^2_{mb^*}(R_1) = \sigma^2_{mb^*}(R) - \left(\frac{A_2 - A_1}{2}\right)^2$$

$$\sigma^2_{mLg^*}(R_1) = \sigma^2_{mLg^*}(R) - \left(\frac{\hat{\delta}_2 - \hat{\delta}_1}{2}\right)^2$$

$$\rho^*_{12}(R_1) = \frac{\sigma^2_{mb^*}(R)\sigma^2_{mLg^*}(R)}{\sigma^2_{mb^*}(R_1)\sigma^2_{mLg^*}(R)} \rho^*_{12}(R).$$

Therefore all of the parameters can be translated to the subregion required to get $\sigma^2_u(R_i)$ and thus $F(R_j)$.

Estimated λ^2, $\beta = \hat{\beta}$

$$mLg_i = C + D\, mb_i + e'_i ,$$

$$e'_i = e_i - \frac{\beta}{B} \epsilon_i , \quad (e_i \sim mbg_i , \quad \epsilon_i \sim mb_i) .$$

Now if $\hat{\beta} = \beta$, $E[\hat{\delta} - \delta] = 0$, so that $E[mLg^*_i] = E[mb^*_i] = W_i$ and

$$\lambda^2 = Var(mLg^*_i - mb^*_i) = E[(mLg^*_i - mb^*_i)^2] .$$

So

$$\hat{\lambda}^2 = \frac{1}{n-1} \sum_1^n (mLg^*_i - mb^*_i)^2 .$$

More than one CORRTEX

If more than 1 CORRTEX is available or the historical data is valid, it is not necessary to assume σ_{mb^*} or ρ^*_{12} .

In that event, let

$$u_i = m_i - \tilde{w}_i , \quad i = 1,2,\ldots, k,$$

w_i = observed log yield. Then letting $\hat{A} = \overline{u}$, and

$$\hat{W}_{k+1} = \frac{m_{k+1} - A}{B} ,$$

it can be shown that

$$\sigma^2_{\hat{w}} = \frac{k+1}{k} \sigma^2_u .$$

σ^2_u can then be estimated from the sample variance of the u_i and a test can be obtained which does not depend on σ_{mb^*} or ρ^*_{12}.

If k>1 calibration shots are available, a bound on the variance of $\sigma^2_{\hat{w}}$ and hence a bound on the F-number for R and R_j could therefore be obtained without assuming ρ^*_{12} or σ_{mb^*} by using the methodology of Alwine, Gray and McCartor (1988).

1. Ralph W. Alwine, III, Henry L. Gray, Gary D. McCartor, Gregory L. Wilson, "Seismic Monitoring of A Threshold Test Ban Treaty (TTBT) Following Calibration of the Test Site with CORRTEX experiments, Mission Research Technical Report # MRC - R - 1139, (1988).