S1. Details on the EM algorithm

The EM algorithm consists of two steps: the E-step and the M-step. During the E-step, based on the parameter values at the current iteration we find the conditional expectation of the likelihood function in respect to the latent variable ($\eta$ and $\xi$ in our case) given the data. Then at the M-step, we find the parameter values that maximize the conditional expectation and we take the maximizers as the values for the next iteration. In our case, the E-step would cause much problem since without Gaussian assumption the conditional expectations are not available in closed form. Motivated by Sengupta & Cressie (2013), we use Laplace approximation (LA) to approximate the conditional expectations, which are based on second-order Taylor-series expansion.

We start with the complete-data log-likelihood function, $L_c$, for the unknown parameters given the data $Z$ and the latent variables $\eta$ and $\xi$. $L_c$ is equal to the joint distribution of $Z$, $\eta$, and $\xi$ given the parameters $\theta \equiv \{\beta, K, \sigma_{\xi_1}^2, \sigma_{\xi_2}^2\}$ seen as a function of $\theta$, which can be written as

$$L_c(\theta | Z, \eta, \xi) = \log(Z_1 | \beta_1, \eta, \xi_1) + \log(Z_2 | \beta_2, \eta, \xi_2) + \log(\eta | K) + \log(\xi | \sigma_{\xi_1}^2, \sigma_{\xi_2}^2)$$

$$= \text{const.} + \left\{ \sum_{k=1}^{2} \sum_{i=1}^{n_k} \left[ Z_k(A_{ki}) h_1^{(k)}(Y_k(A_{ki})) - h_2^{(k)}(Y_k(A_{ki})) \right] \cdot \frac{1}{r_k} \right\}$$

$$- \frac{1}{2} \log |K| - \frac{1}{2} \text{trace}(\eta\eta^TK^{-1})$$

$$- \sum_{k=1}^{2} \left[ \frac{n_k}{2} \log \sigma_{\xi_k}^2 \right] - \frac{1}{2} \text{trace}(\xi\xi^T(VE)^{-1}),$$

(1)
The E-step is taking the conditional expectation of $L_c$ with respect to the latent variables, $\eta$ and $\xi$ given the parameter values. Assume that at $l$-th iteration of the EM algorithm we obtain $\hat{\theta}^{(l)}$ for $\theta$, then the conditional expectation of $L_c$ can be written as:

$$Q(\theta, \theta^{(l)}) \equiv E\left( L_c(\theta | Z, \eta, \xi) | \theta^{(l)} \right)$$

$$= \text{const.} + \frac{2}{\tau_k^2} \sum_{k=1}^{2} \left\{ \frac{n_k}{2} \log |K| - \frac{1}{2} \text{tr} \left[ E(\eta\eta^T | Z, \theta^{(l)}) K^{-1} \right] \right\} - \frac{1}{2} \sum_{k=1}^{2} \left\{ \frac{n_k}{2} \log \sigma_k^2 \right\} - \frac{1}{2} \text{tr} \left[ E(\xi\xi^T | Z, \theta^{(l)}) (VE)^{-1} \right]$$

Since the posterior distribution of $\eta$ and $\xi$ are not available in closed form, the conditional expectations cannot be evaluated directly. One common approach to tackle this problem is to implement Monte Carlo integration by simulating the posterior distribution for $\eta$ and $\xi$ given the data during each iteration of EM algorithm. So the EM algorithm can be very slow to converge when the datasets are large. Alternatively, we approximate the conditional expectations by using Laplace approximation, which is based on second-order Taylor-series expansion of the logarithm of the integrands at the posterior mode of $\eta$ and $\xi$ for each iteration.

To obtain the mode at the $l$-th iteration, $(\eta^{(l)}, \xi^{(l)})$, Sengupta & Cressie (2013) use a coordinate-wise ascent method for the Poisson GLM and canonical log link, which maximizes alternatively with respect to $\eta$, and then with respect to $\xi$, until convergence. In this study, we find the mode for $\eta$ and then for $\xi$ successively by using Newton-raphson method. The algorithm is detailed in Section S1.1.

Applying a second-order Taylor-series approximation, we can approximate the posterior distribution of $[\eta, \xi | Z, \theta^{(l)}]$ with a Gaussian distribution with its mean equal to the posterior mode and the variance equal to the inverse of the negative Hessian matrix evaluated at the posterior mode. More details can be find in Section S1.2, where the resulting distribution is approximately a multivariate Gaussian and the mean and variance are shown as:

$$E\left( \eta \bigg| Z, \theta^{(l)} \right) = \eta^{(l)}$$

$$\text{var} \left( \eta \bigg| Z, \theta^{(l)} \right) = \begin{bmatrix} 1 & \text{cov}(\eta^{(l)}, \xi^{(l)}) \\ \text{cov}(\eta^{(l)}, \xi^{(l)}) & \text{var}(\xi^{(l)}) \end{bmatrix}^{-1}$$

where $J(u_0, v_0) = -\frac{\partial^2}{\partial u_0 \partial v_0} (L_c(\theta^{(l)} | Z, u, v))_{|u=u_0, v=v_0}$ for generic variables $u$ and $v$. The inverse of the matrix can be obtained using a block-matrix-inversion formula:

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} (A - BD^{-1}C)^{-1} & -(A - BD^{-1}C)^{-1}BD^{-1} \\ -(D - CA^{-1}B)^{-1}CA & (D - CA^{-1}B)^{-1} \end{pmatrix}$$

Recall that the Sherman-Morrison-Woodbury formula:

$$(A + BDC)^{-1} = A^{-1} - A^{-1}B(D^{-1} + CA^{-1}B)^{-1}CA^{-1}$$
We use this formula in the block-matrix inversion formula to obtain the following equivalent formula:
\[
\begin{pmatrix}
A & B \\
C & D
\end{pmatrix}^{-1} = \begin{pmatrix}
(A - BD^{-1}C)^{-1} & -(A - BD^{-1}C)^{-1}BD^{-1} \\
-D^{-1}CA(A - BD^{-1}C)^{-1} & D^{-1} + D^{-1}CA(A - BD^{-1}C)^{-1}BD^{-1}
\end{pmatrix}
\tag{7}
\]
Then we use (7) to invert the matrix in (4) and we have the following results:
\[
\text{var}(\eta|Z, \theta^{[l]}) = \left( J(\hat{\eta}^{[l]}, \hat{\eta}^{[l]}) - J(\hat{\eta}^{[l]}, \hat{\xi}^{[l]})J(\hat{\xi}^{[l]}, \hat{\eta}^{[l]})^{-1}J(\hat{\xi}^{[l]}, \hat{\eta}^{[l]}) \right)^{-1}
\]
\[
\text{var}(\xi|Z, \theta^{[l]}) = J(\hat{\xi}^{[l]}, \hat{\xi}^{[l]})^{-1} + J(\hat{\xi}^{[l]}, \hat{\eta}^{[l]})^{-1}J(\hat{\xi}^{[l]}, \hat{\eta}^{[l]}) \times \left( J(\hat{\eta}^{[l]}, \hat{\eta}^{[l]}) - J(\hat{\eta}^{[l]}, \hat{\xi}^{[l]})J(\hat{\xi}^{[l]}, \hat{\xi}^{[l]})^{-1}J(\hat{\xi}^{[l]}, \hat{\eta}^{[l]}) \right)^{-1} \times J(\hat{\eta}^{[l]}, \hat{\xi}^{[l]})J(\hat{\xi}^{[l]}, \hat{\xi}^{[l]})^{-1}.
\tag{8}
\]
Note that all we need to invert is the fixed-rank \(r \times r\) matrix, \(J(\hat{\eta}^{[l]}, \hat{\eta}^{[l]})\), and the \(N \times N\) diagonal (or sparse if there is overlap between the satellite footprints) matrix, which provide us significant computational advantage and allow us to obtain the expression for \(E(\eta\eta^T|Z, \theta^{[l]})\) and \(E(\xi\xi^T|Z, \theta^{[l]})\) in (2) as follows:
\[
E(\eta\eta^T|Z, \theta^{[l]}) = \text{var}(\eta|Z, \theta^{[l]}) + E(\eta|Z, \theta^{[l]}E(\eta|Z, \theta^{[l]})^T
\]
\[
E(\xi\xi^T|Z, \theta^{[l]}) = \text{var}(\xi|Z, \theta^{[l]}) + E(\xi|Z, \theta^{[l]}E(\xi|Z, \theta^{[l]})^T
\tag{9}
\]
where the terms on the right-hand side are evaluated using (3) and (4).

Again, we still apply a second-order Taylor-series expansion to approximate the remaining terms in (2) and we have the following results:
\[
E \left( h_k^{(l)} (Y_k^{[l]}(A_{ki})) | Z, \theta^{[l]} \right) \approx h_k^{(l)}(Y_k^{[l]}(A_{ki})) + \frac{1}{2} h_k^{(l)}(Y_k^{[l]}(A_{ki})) \times \left\{ S(A_{ki})^T \text{var}(\eta|Z, \theta^{[l]})S(A_{ki}) + 2S(A_{ki})^T \text{cov}(\eta, \xi|Z, \theta^{[l]})e(A_{ki}) + e(A_{ki})^T \text{var}(\xi|Z, \theta^{[l]})e(A_{ki}) \right\}
\tag{10}
\]
where \(Y_k^{[l]}(A_{ki})\) is the value of \(Y_k(A_{ki})\) evaluated at \((\eta^{[l]}, \xi^{[l]})\), and \(e(A_{ki})\) is a vector of length \(N\) whose elements is the value returned from the indicator function \(I(A = A_{ki})\).

Now we perform the M-step by maximizing the conditional expectation \(Q(\theta, \theta^{[l]})\) with respect to each of the parameters in \(\theta\). By setting the first derivatives of \(Q(\theta, \theta^{[l]})\) with respect to \(K\), \(\sigma^2_{\xi_1}\), and \(\sigma^2_{\xi_2}\) equal to zero, and solving the resulting equations, we obtain the following updating equations:
\[
\sigma^2_{\xi_1}^{[l+1]} = \frac{1}{N_1} \text{trace} \left( E(\xi_1 \xi_1^T|Z, \theta^{[l]}) : E_1^{-1} \right)
\]
\[
\sigma^2_{\xi_2}^{[l+1]} = \frac{1}{N_2} \text{trace} \left( E(\xi_2 \xi_2^T|Z, \theta^{[l]}) : E_1^{-1} \right)
\]
\[
K^{[l+1]} = E(\eta|Z, \theta^{[l]})E(\eta|Z, \theta^{[l]})^T + \text{var}(\eta|Z, \theta^{[l]}).
\tag{11}
\]

Since the solution for the linear coefficient \(\beta\) are not straightforward, we adopt a Newton-Raphson update at each M-step as follows:
\[
\beta^{[l+1]} = \beta^{[l]} - \left[ \frac{\partial}{\partial \beta} R(\theta) \right]^{-1} \frac{\partial R(\theta)}{\partial \theta^{[l]}},
\tag{12}
\]
where \(R(\theta)\) denotes the score function obtained by taking the partial derivative of \(Q(\theta, \theta^{[l]})\) with respect to \(\beta\). More details are in Section S1.3.
S1.1. The coordinate-wise ascent algorithm

Coordinate-wise ascent is based on the idea that maximization of a multivariate function can be achieved by maximize the object function with respect to a single parameter at a time. In another word, the complex multivariate optimization problem can be solved by solving more simpler univariate optimization problem iteratively. Convergence of coordinate-wise ascent algorithm requires the objective function to be strictly concave and differentiable, with respect to the \((\eta, \xi)\). From Section S1.2, it is easily to see that the Hessian matrix of \(L_c(\theta^0|Z, \delta)\) is negative definite; therefore, the coordinate-wise ascent method will converge to \(\eta^*\) and \(\xi^*\).

The coordinate-wise ascent method is outline below:

1. We start with an initial value for \(\eta^0\) and \(\xi^0\).
2. At \(t + 1\)-th iteration, we obtain \(\eta^{t+1}\) that maximize \(L_c(\theta^0|Z, \eta^t, \xi^t)\) from Newton-raphson method.
3. Successively we obtain \(\xi^{t+1}\) that maximize \(L_c(\theta^0|Z, \eta^{t+1}, \xi)\) using Newton-raphson method.
4. Repeat Step 2 and 3 until convergence is reached.

Newton-raplhson method requires the first partial derivative and the second partial derivative of \(L_c(\theta^0|Z, \delta)\) with respect to \(\eta\) and \(\xi\), which is shown below:

\[
\frac{\partial}{\partial \eta} \left( L_c(\theta^0|Z, \eta, \xi) \right) = \left( S_1^T D_1 \begin{bmatrix} 0 & S_2^T D_2 \end{bmatrix} \right) - K^{-1} \eta,
\]

\[
\frac{\partial^2}{\partial \eta \partial \eta^T} \left( L_c(\theta^0|Z, \eta, \xi) \right) = \left( S_1^T D_1 S_1 \begin{bmatrix} 0 & S_2^T D_2 S_2 \end{bmatrix} \right) - K^{-1},
\]

\[
\frac{\partial}{\partial \xi} \left( L_c(\theta^0|Z, \eta, \xi) \right) = \begin{bmatrix} D_1 & 0 \\ 0 & D_2 \end{bmatrix} - (CV)^{-1} \xi,
\]

\[
\frac{\partial^2}{\partial \xi \partial \xi^T} \left( L_c(\theta^0|Z, \eta, \xi) \right) = \begin{bmatrix} D_1^2 & 0 \\ 0 & D_2^2 \end{bmatrix} - (CV)^{-1},
\]

where \(D_k \equiv diag\{(Z_k(s_{ki})h_1^{(k)\prime}(Y_k(s_{ki})) - h_2^{(k)}(Y_k(s_{ki})))/\tau_k^2\}\), and \(D_k' \equiv diag\{(Z_k(s_{ki})h_1^{(k)\prime}(Y_k(s_{ki})) - h_2^{(k)\prime}(Y_k(s_{ki})))/\tau_k^2\}\), for \(i = 1, \ldots, n_k\).

S1.2. Approximations involved in the EM algorithm’s E-step

We derive Laplace approximation to the density \([\eta, \xi|Z, \theta^0]\) where \(\theta^0\) is the parameter value at \(l\)-th iteration. We consider \(\eta\) and \(\xi\) jointly by stacking them to form an \(m(m = r + N)\)-dimensional vector, \(\delta \equiv (\eta^T, \xi^T)^T\) and let \(\delta^0\) be the mode that maximize the complete-data log-likelihood, \(L_c(\theta^0|Z, \delta)\). Thus, the density of \(\delta|Z, \theta^0\) is given by:

\[
p(\delta|Z, \theta^0) \propto \exp(L_c(\theta^0|Z, \delta))
\]

Now apply a second-order Taylor approximation to \(L_c(\theta^0|bZ, \delta)\), centered at \(\delta^0\):

\[
L_c(\theta^0|Z, \delta) = L_c(\theta^0|Z, \delta^0) + \frac{1}{2}(\delta - \delta^0)^T \left[ \frac{\partial^2}{\partial \delta \partial \delta^T} \left( L_c(\theta^0|Z, \delta) \right) \right]_{\delta = \delta^0} (\delta - \delta^0) + \text{higher-order terms}
\]

\[
\approx L_c(\theta^0|Z, \delta^0) - \frac{1}{2}(\delta - \delta^0)^T Q_L A(\delta^0, \theta^0|Z)(\delta - \delta^0),
\]
where \( Q_{LA}(\delta^{[i]}, \theta^{[i]}|Z) \equiv -\left[ \frac{\partial^2}{\partial \delta \partial \delta} (L_c(\theta^{[i]}|bZ, \delta)) \right]_{\delta=\delta^{[i]}} \) is the negative Hessian matrix of \( L_c(\theta^{[i]}|bZ, \delta) \) evaluated at \( \delta^{[i]} \). Notice that the first-order linear term is zero since the first-order derivative of \( L_c(\theta^{[i]}|bZ, \delta) \) with respect to \( \delta \), evaluated at \( \delta = \delta^{[i]} \) is zero given that \( \delta^{[i]} \) maximizes \( L_c(\theta^{[i]}|Z, \delta) \). Therefore, we can approximate the density of \([\eta, \xi|Z, \theta^{[i]}]\) as

\[
p(\delta|Z, \theta^{[i]}) \propto \exp(L_c(\theta^{[i]}|bZ, \delta)) \exp \left( -\frac{1}{2} (\delta - \delta^{[i]})^T Q_{LA}(\delta^{[i]}, \theta^{[i]}|Z) (\delta - \delta^{[i]}) \right).
\]

Thus \( p(\delta|Z, \theta^{[i]}) \) is approximately proportional to a Gaussian density with the first two moments as

\[
E(\delta|Z, \theta^{[i]}) = \delta^{[i]},
\]

\[
\text{var}(\delta|Z, \theta^{[i]}) = Q_{LA}(\delta^{[i]}, \theta^{[i]}|Z)^{-1}.
\]

The normalizing constant on the right-hand side of (15) can also be approximated as

\[
\int p(\delta|Z, \theta^{[i]}) d\delta = \exp \left( L_c(\theta^{[i]}|bZ, \delta) \right) (2\pi)^{m/2}|Q_{LA}(\delta^{[i]}, \theta^{[i]}|Z)^{-1/2}.
\]

Since \( \delta \equiv (\eta^T, \xi^T)^T \), it is clear to see that (16) and (17) are equivalent to (3) and (4).

Next, we derive the approximation for \( E \left( h_k^{(j)}(Y_k^{[i]}(A_k_i))|Z, \theta^{[i]} \right) \) for \( k = 1, 2 \) and \( j = 1, 2 \). Recall that \( Y_k^{[i]}(A_k_i) \equiv C(A_k_i) + X(A_k_i)^T \beta + S(A_k_i)^T \eta^{[i]} + \xi^{[i]}(A_k_i) \). Now we define a \( m = r + N \)-dimensional vector \( q(A_k_i) \) such that \( Y_k^{[i]}(A_k_i) \equiv C(A_k_i) + X(A_k_i)^T \beta + q(A_k_i)^T \delta^{[i]} \). We still apply a second-order Taylor-series expansion to \( h_k^{(j)}(C(A_k_i) + X(A_k_i)^T \beta + q(A_k_i)^T \delta^{[i]}), \) which yields:

\[
h_k^{(j)} \left( C(A_k_i) + X(A_k_i)^T \beta + q(A_k_i)^T \delta^{[i]} \right)
\]

\[
= h_k^{(j)} \left( C(A_k_i) + X(A_k_i)^T \beta + q(A_k_i)^T \delta^{[i]} \right)
\]

\[
+ (\delta - \delta^{[i]})^T \left( h_k^{(j)'} \left( C(A_k_i) + X(A_k_i)^T \beta + q(A_k_i)^T \delta^{[i]} \right) \times q(A_k_i) \right)
\]

\[
+ \frac{1}{2} (\delta - \delta^{[i]})^T \left( h_k^{(j)''} \left( C(A_k_i) + X(A_k_i)^T \beta + q(A_k_i)^T \delta^{[i]} \right) \times q(A_k_i)q(A_k_i)^T \right) (\delta - \delta^{[i]})
\]

where the \( m \)-dimensional vector \( h_k^{(j)'}(x_0) \equiv \frac{\partial}{\partial x} h_k^{(j)}(x)|_{x=x_0} \) and the \( m \times m \) matrix \( h_k^{(j)''}(x_0) \equiv \frac{\partial^2}{\partial x \partial x} h_k^{(j)}(x)|_{x=x_0} \).

Removing the higher-order terms and taking expectations on both side, we obtain approximately

\[
E \left( h_k^{(j)}(C(A_k_i) + X(A_k_i)^T \beta + q(A_k_i)^T \delta^{[i]}))|Z, \theta^{[i]} \right)
\]

\[
\approx h_k^{(j)} \left( C(A_k_i) + X(A_k_i)^T \beta + q(A_k_i)^T \delta^{[i]} \right)
\]

\[
+ E \left( (\delta - \delta^{[i]})|Z, \theta^{[i]} \right) \left( h_k^{(j)'} \left( C(A_k_i) + X(A_k_i)^T \beta + q(A_k_i)^T \delta^{[i]} \right) \right) \times q(A_k_i)
\]

\[
+ \frac{1}{2} \text{trace} \left\{ E \left( (\delta - \delta^{[i]})|Z, \theta^{[i]} \right) \left( h_k^{(j)''} \left( C(A_k_i) + X(A_k_i)^T \beta + q(A_k_i)^T \delta^{[i]} \right) \right) \times q(A_k_i)q(A_k_i)^T \right\}
\]

Recall that we approximate the density of \([\delta|Z, \theta^{[i]}] \) as a Gaussian density with mean \( \delta^{[i]} \) and variance \( Q_{LA}(\delta^{[i]}, \theta^{[i]}|Z)^{-1} \). Therefore, \( E((\delta - \delta^{[i]}))|Z, \theta^{[i]} \) equals to zero thus the second term on the right-hand side disappears. Then (20)
becomes

\[
\begin{align*}
\mathbb{E} \left( h^{(j)}_k (C(A_{k_i}) + X(A_{k_i})^T \beta + q(A_{k_i})^T \delta | Z, \theta^{[l]} \right) \\
\approx h^{(j)}_k \left( C(A_{k_i}) + X(A_{k_i})^T \beta + q(A_{k_i})^T \delta \right) \\
\quad + \frac{1}{2} \text{trace} \left\{ Q_{LA} (\delta^{[l]}, \theta^{[l]} | Z)^{-1} \times \left( h^{(j)_m}_k \left( C(A_{k_i}) + X(A_{k_i})^T \beta + q(A_{k_i})^T \delta \right) \right) \right\} \\
= h^{(j)}_k \left( C(A_{k_i}) + X(A_{k_i})^T \beta + q(A_{k_i})^T \delta \right) \\
\quad + \frac{1}{2} h^{(j)_m}_k \left( C(A_{k_i}) + X(A_{k_i})^T \beta + q(A_{k_i})^T \delta \right) \right) \times q(A_{k_i})^T Q_{LA} (\delta^{[l]}, \theta^{[l]} | Z)^{-1} q(A_{k_i}) \\
\end{align*}
\]

(21)

Remember that at the beginning we define \( q(A_{k_i}) \equiv (S(A_{k_i})^T, e(A_{k_i})^T)^T \) where \( e(A_{k_i}) \) is a vector of length \( N \) whose elements is the value returned from the indicator function \( I(A = A_{k_i}) \). Therefore, simple matrix multiplication yields

\[
\begin{align*}
\mathbb{E} \left( h^{(j)}_k (C(A_{k_i}) + X(A_{k_i})^T \beta + q(A_{k_i})^T \delta | Z, \theta^{[l]} \right) \\
\approx h^{(j)}_k \left( C(A_{k_i}) + X(A_{k_i})^T \beta + q(A_{k_i})^T \delta \right) \\
\quad + \frac{1}{2} h^{(j)_m}_k \left( C(A_{k_i}) + X(A_{k_i})^T \beta + q(A_{k_i})^T \delta \right) \right) \times \left( S(A_{k_i})^T \var(\eta | Z, \theta^{[l]}), S(A_{k_i}) + 2S(A_{k_i})^T \text{cov}(\eta, \xi | Z, \theta^{[l]}), e(A_{k_i}) + e(A_{k_i})^T \var(\xi | Z, \theta^{[l]}), e(A_{k_i}) \right) \\
\end{align*}
\]

(22)

Therefore, we obtain the approximations to the expectations involved in (2).

**S1.3. Derivation of Newton-Raphson update for \( \beta \)**

In this section, we derive the score functions \( R(\theta) \) and its first derivative with respect to \( \beta \) that involved in the Newton-Raphson update, assuming as many derivatives for \( h^{(j)}_k (\cdot) \) as necessary.

After substituting in the approximations to the required expectations, \( Q(\cdot, \cdot) \), given by (2), becomes

\[
Q(\theta, \theta^{[l]}) = \text{const.} + \sum_{k=1}^{2} \frac{1}{k} \sum_{i=1}^{n_k} \sum_{k_i=1}^{n_k} Z_k(A_{k_i}) \left\{ h^{(j)}_k (C(A_{k_i}) + X(A_{k_i})^T \beta + q(A_{k_i})^T \delta) \right\} \\
\quad + \frac{1}{2} h^{(j)_m}_k \left( C(A_{k_i}) + X(A_{k_i})^T \beta + q(A_{k_i})^T \delta \right) \right) \times q(A_{k_i})^T Q_{LA} (\delta^{[l]}, \theta^{[l]} | Z)^{-1} q(A_{k_i}) \\
\quad - \sum_{i=1}^{n_k} \left\{ h^{(j)}_k (C(A_{k_i}) + X(A_{k_i})^T \beta + q(A_{k_i})^T \delta) \right\} \\
\quad + \frac{1}{2} h^{(j)_m}_k \left( C(A_{k_i}) + X(A_{k_i})^T \beta + q(A_{k_i})^T \delta \right) \right) \times q(A_{k_i})^T Q_{LA} (\delta^{[l]}, \theta^{[l]} | Z)^{-1} q(A_{k_i}) \right\} \\
\quad - \frac{1}{2} \log |K| - \frac{1}{2} \text{trace} \left[ \mathbb{E}(\eta \eta^T | Z, \theta^{[l]}) K^{-1} \right] \\
\quad - \frac{1}{2} \sum_{k=1}^{2} \left[ \frac{n_k}{2} \log \sigma_k^2 \right] - \frac{1}{2} \text{trace} \left[ \mathbb{E}(\xi \xi^T | Z, \theta^{[l]}) (\mathbf{V} \mathbf{E})^{-1} \right]. \\
\end{align*}
\]

(23)

where \( q(A_{k_i}) \) and \( Q_{LA}(\delta^{[l]}, \theta^{[l]} | Z) \) are defined as the same as in Section S1.2. The approximations for the expectations, \( \mathbb{E}(\eta \eta^T | Z, \theta^{[l]}) \) and \( \mathbb{E}(\xi \xi^T | bZ, \theta^{[l]}) \) are given by (10) (which follows from Section S1.2).
Taking the partial derivative of $Q(\theta, \theta^{[l]})$ with respective to $\beta$, we obtain
\begin{align*}
R(\theta) &= \sum_{k=1}^{2} \frac{1}{2k} \left\{ \sum_{i=1}^{n_k} Z_k(A_{ki}) \left\{ h_k^{(1)^v}(C(A_{ki}) + X(A_{ki})^T \beta + q(A_{ki})^T \delta^{[l]})
    
    + \frac{1}{2} h_k^{-v}(C(A_{ki}) + X(A_{ki})^T \beta + q(A_{ki})^T \delta^{[l]}) \right\} \times (A_{ki}) \right\} \\
    
    &+ \frac{1}{2} h_k^{-v}(C(A_{ki}) + X(A_{ki})^T \beta + q(A_{ki})^T \delta^{[l]}) \right\} \times (A_{ki}) \right\} \\
    
    &+ \frac{1}{2} h_k^{(2)^v}(C(A_{ki}) + X(A_{ki})^T \beta + q(A_{ki})^T \delta^{[l]}) \right\} \times (A_{ki}) \right\} \\
    
    &+ \frac{1}{2} h_k^{(2)^v}(C(A_{ki}) + X(A_{ki})^T \beta + q(A_{ki})^T \delta^{[l]}) \right\} \times (A_{ki}) \right\} \\
    
    &\times X(A_{ki})X(A_{ki})^T \right\}.
\end{align*}

Then we need further take the partial derivative of $R(\theta)$ with respect to $\beta$, which is given by:
\begin{align*}
\frac{\partial}{\partial \beta} R(\theta) &= \sum_{k=1}^{2} \frac{1}{2k} \left\{ \sum_{i=1}^{n_k} Z_k(A_{ki}) \left\{ h_k^{(1)^v}(C(A_{ki}) + X(A_{ki})^T \beta + q(A_{ki})^T \delta^{[l]})
    
    + \frac{1}{2} h_k^{-v}(C(A_{ki}) + X(A_{ki})^T \beta + q(A_{ki})^T \delta^{[l]}) \right\} \times (A_{ki}) \right\} \\
    
    &+ \frac{1}{2} h_k^{(2)^v}(C(A_{ki}) + X(A_{ki})^T \beta + q(A_{ki})^T \delta^{[l]}) \right\} \times (A_{ki}) \right\} \\
    
    &\times X(A_{ki})X(A_{ki})^T \right\}.
\end{align*}

Then by substituting (24) and (25) into (12), we obtain the value for $\beta$ at $l + 1$ iteration.

### S1.4. Initial values for the EM algorithm

In order to trigger the EM algorithm, we still need a set of initial values for the unknown parameters $\theta^{[0]}$. One may use the classical GLM estimate without considering the spatial dependence, $\hat{\beta}^{[0]}_{k;GLM}$, as the initial value for the trend coefficients. Then one may proceed by obtaining the detrended data, $\tilde{Z}_k(\cdot)$, as follows:
\begin{align*}
\tilde{Z}_k(A_{ki}) &\equiv g_k(Z(A_{ki}) + c) - C_k(A_{ki}) - X_k^T(A_{ki})\hat{\beta}^{[0]}_{k;GLM}, \quad i = 1, \ldots, n_k.
\end{align*}

where $c$ is a small user-specified constant that makes the link function defined within the range of the data.

Define $s^2_{\tilde{Z}_k} \equiv \frac{1}{n_k} \sum_{i=1}^{n_k} \tilde{Z}_k(A_{ki})^2$ as the empirical variance of $\tilde{Z}_k$, then the total variation can be simply captured through the trace operator, $\Sigma_{\tilde{Z}_k} = s^2_{\tilde{Z}_k} I_{n_k}$. We assign approximately 90% of this to the smooth small-scale variation and 10% to the fine-scale variation (Nguyen et al., 2014). That is, for the $k$-th process we select our initial values for $K_k$ and $\sigma^2_{\xi_k}$ to satisfy the following conditions:
\begin{align*}
S_k K_1^{[0]} S_k^T &\approx 0.9 \Sigma_{\tilde{Z}_k} \quad (27) \\
\sigma^2_{\xi_k} &\approx 0.1 \text{trace}(\Sigma_{\tilde{Z}_k})/\text{trace}(VE) \quad (28)
\end{align*}

Solving (27) by the Q-R decomposition $S_k = Q_k R_k$, we obtain the initial value for $K_k$ as
\begin{align*}
K_k^{[0]} = R_k^{-1} Q_k (0.9 \Sigma_{\tilde{Z}_k}) Q_k R_k^{-1} \quad (29)
\end{align*}

Then the initial value for the combined matrix $K$ may be obtained from the following block diagonal matrix,
\begin{align*}
K^{[0]} = \begin{pmatrix} K_1^{[0]} & 0 \\
0 & K_2^{[0]} \end{pmatrix} \quad (30)
\end{align*}
S2. Details of the empirical Bayesian inference

We divide the set of prediction locations into two groups, the locations having observation at the \(k\)-th instrument \(\mathcal{D}_{O_k}\), and the locations having no observation \(\mathcal{D}_{M_k}\). We define \(\hat{\xi}_O\) as the stacked vector of \(\hat{\xi}_1(\cdot)\) and \(\hat{\xi}_2(\cdot)\) at their corresponding observed locations and \(\hat{\xi}_M\) as the one at their corresponding unobserved locations. Then the posterior distribution \([\hat{\xi}_M|\mathbf{Z}, \eta, \hat{\xi}_O, \theta]\) can be shown as

\[
[\hat{\xi}_M|\mathbf{Z}, \eta, \hat{\xi}_O, \theta] = \frac{[\hat{\xi}_O, \hat{\xi}_M, \mathbf{Z}, \eta|\theta]}{[\mathbf{Z}|\eta, \hat{\xi}_O, \theta][\eta|\mathcal{K}][\hat{\xi}_O|\sigma^2_{\hat{\xi}_1}, \sigma^2_{\hat{\xi}_2}][\hat{\xi}_M|\sigma^2_{\hat{\xi}_1}, \sigma^2_{\hat{\xi}_2}]} \int [\mathbf{Z}|\eta, \hat{\xi}_O, \theta][\eta|\mathcal{K}][\hat{\xi}_O|\sigma^2_{\hat{\xi}_1}, \sigma^2_{\hat{\xi}_2}][\hat{\xi}_M|\sigma^2_{\hat{\xi}_1}, \sigma^2_{\hat{\xi}_2}]d\hat{\xi}_M,
\]

from which we note that \(\hat{\xi}_M\) is conditionally independent of \((\mathbf{Z}, \eta, \hat{\xi}_O)\). Therefore, for a new location \(s_0 \in \mathcal{P}\), we have

\[
\mathbb{E}(Y_k(s_0)|\mathbf{Z}, \theta) = C_k(s_0) + \mathbf{X}_k^T(s_0)\beta_k + \mathbf{S}_k(s_0)^T\mathbb{E}(\eta_k|\mathbf{Z}, \theta) + \mathbb{E}(\hat{\xi}_k(s_0)|\mathbf{Z}, \theta)/(s_0 \in \mathcal{D}_{O_k})
\]

where \(C_k(s_0), \mathbf{S}_k(s_0), \mathbf{X}_k(s_0),\) and \(\hat{\xi}_k(s_0)\) represent the values and vectors derived by evaluating the corresponding terms at the prediction location \(s_0\). The last term on the right-hand side has an indicator function which equal to 1 if \(s_0\) is observed at \(k\)-th instrument and 0 otherwise.

If \(s_0\) is not observed at the \(k\)th instrument, then

\[
\text{var}(Y_{pk}(s_0)|\mathbf{Z}, \theta) = \mathbf{S}_{pk}(s_0)^T\text{var}(\eta_k|\mathbf{Z}, \theta)\mathbf{S}_{pk}(s_0) + \sigma^2_{\hat{\xi}_k}(s_0).
\]

If \(s_0\) is observed at the \(k\)th datasets, then

\[
\text{var}(Y_{pk}(s_0)|\mathbf{Z}, \theta) = \mathbf{S}_k(s_0)^T\text{var}(\eta_k|\mathbf{Z}, \theta)\mathbf{S}_k(s_0) + \text{var}(\hat{\xi}_{pk}(s_0)|\mathbf{Z}, \theta) + 2\mathbf{S}_k(s_0)^T\text{cov}(\eta_k, \hat{\xi}_k(s_0)|\mathbf{Z}, \theta).
\]

S2.1. The MCMC algorithm

The empirical-Bayesian inference is conducted by substituting the estimates of unknown parameter \(\hat{\theta}\) into the predictive distribution \([\eta, \xi|\mathbf{Z}, \theta]\). If the predictive distribution is not available in closed form, then we use an MCMC algorithm to obtain samples from it. The MCMC algorithm is based on Gibbs sampler and incorporates Metropolis-Hastings algorithm where it is necessary (i.e. the full conditionals are not available in closed form).

The joint distribution, \([\mathbf{Z}, \eta, \xi|\theta]\), can be written as:

\[
[\mathbf{Z}, \eta, \xi|\theta] = [\mathbf{Z}|\eta, \xi, \theta] \times [\eta|\mathcal{K}] \times [\xi_1|\sigma^2_{\hat{\xi}_1}] \times [\xi_2|\sigma^2_{\hat{\xi}_2}].
\]

If \(Z_k(\cdot)\) follows Poisson distributions then the full conditionals can be written as:

\[
[\eta|\xi_1, \xi_2, \cdot] \propto \exp \left( \sum_{k=1}^{2} \sum_{i=1}^{n} Z_k(A_{ki})\mathbf{S}(A_{ki})^T \eta - e^{Y_k(A_{ki})} \right) \exp \left( -\frac{1}{2} \eta^T \mathbf{K}^{-1} \eta \right)
\]

\[
[\xi_1(A_{1i})|\eta, \xi_2] \propto \exp \left( \xi_1(A_{1i})Z_1(A_{1i}) - e^{Y_1(A_{1i})} \right) \cdot \exp \left( -\frac{\xi_1^2(A_{1i})}{2\sigma^2_{\xi_1}|D \cap A_{1i}|} \right)
\]

\[
[\xi_2(A_{2i})|\eta, \xi_{10}] \propto \exp \left( \xi_2(A_{2i})Z_2(A_{2i}) - e^{Y_2(A_{2i})} \right) \cdot \exp \left( -\frac{\xi_2^2(A_{2i})}{2\sigma^2_{\xi_2}|D \cap A_{2i}|} \right)
\]
For Poisson distribution, the full conditionals for $\eta$, $\xi_1$, and $\xi_2$ are not available in closed form. So we implement Metropolis-Hastings algorithm for all of the variables. The algorithm used to draw samples from the full conditionals, $[\eta^{[t+1]}|\xi_1^{[t]}, \xi_2^{[t]}, \cdot], [\xi_1^{[t+1]}|\eta^{[t]}, \xi_2^{[t]}, \cdot],$ and $[\xi_2^{[t+1]}|\eta^{[t+1]}, \xi_2^{[t+1]}, \cdot]$ is described below.

To sample from $[\eta^{[t+1]}|\xi_1^{[t]}, \xi_2^{[t]}, \cdot]$ at the $(t+1)$-th stage, we do

1. Draw a candidate $\eta^*$ from a proposal density, $\text{Gau}(\eta^{[t]}, c_\eta \Sigma_\eta^*)$, where $c_\eta$ is a tuning parameter that can be tailored to control the acceptance rate. The co-variance matrix $\Sigma_\eta^*$ can be chosen as an identity matrix or the negative inverse of the Hessian matrix evaluated at $\eta^{[t]}$.
2. Generate $U_1$ uniformly on $(0,1)$.
3. Compute the joint density of $\eta$ and all other unknowns fixed at their most recently sampled value, $l(\eta^{[t]}, \text{rest})$ and $l(\eta^*, \text{rest})$.
4. If $U_1 < \min\{l(\eta^*, \text{rest})\}$, accept the candidate and update $\eta^{[t+1]} = \eta^*$.

Sampling from $[\xi_1^{[t+1]}|\eta^{[t+1]}, \xi_2^{[t]}, \cdot], [\xi_2^{[t+1]}|\eta^{[t+1]}, \xi_2^{[t+1]}, \cdot]$ are similar but were conducted element-wisely. The dispersion of the proposal distributions, $\Sigma_\eta^*$ and $\Sigma_{\xi_1}^*$ are both specified as an identity matrix times a tuning parameter, $c_{\xi_1}$ and $c_{\xi_2}$, respectively.

### S3. Details of the simulation set-up

Recall that in order to control the variability of $Y$, we choose $K = hK^0$, where $h$ is chosen to preserve the total variation satisfying the following condition,

$$\text{trace}(\Sigma)/(2N) = 1 = (\text{trace}(hSK^0S^T) + \text{trace}(\sigma_{\xi_1}^2 I_N + \sigma_{\xi_2}^2 I_N))/(2N).$$

(39)

To specify $\sigma_{\xi_1}^2$ and $\sigma_{\xi_2}^2$, we define the Fine-scale Variation Proportion (FVP),

$$\text{FVP} = \frac{\text{trace}(\sigma_{\xi_1}^2 I_N + \sigma_{\xi_2}^2 I_N)}{\text{trace}(SK^0S^T) + \text{trace}(\sigma_{\xi_1}^2 I_N + \sigma_{\xi_2}^2 I_N)}$$

(40)

which equals to $\frac{\sigma_{\xi_1}^2 + \sigma_{\xi_2}^2}{2}$ from (39). By setting $\sigma_{\xi_1}^2 = 0.45$ and $\sigma_{\xi_2}^2 = 0.55$, the proportion of fine-scale variation to the total variability is held at 5%; thus, we obtain $h = 0.8667$.

To specify the spatial trend coefficients $\beta_1$ and $\beta_2$, we define the variation of the "signal", $V_s(k)$, as:

$$V_s(k) \equiv \frac{1}{N} \text{trace}(S_k K_k S_k^T + \sigma_{\xi_1}^2 I_N) + \frac{1}{N} \sum_{i=1}^N (X_k(s_i)^T \beta_k - \text{ave}_{s_i \in D}(X_k(s_i)^T \beta_k))^2$$

(41)

The parameter $\beta_1^{(1)}$ and $\beta_1^{(2)}$ are selected such that $V_s^{(1)}$ and $V_s^{(2)}$ were approximately 2 (Aldworth & Cressie, 1999, Section 3.2.4). Since the intercept $\beta_0^{(k)}$ is a free parameter that has no impact on $V_s^{(k)}$, we fix $\beta_0^{(1)} = 1$ and $\beta_0^{(2)} = 3$ and then obtain $\beta_1^{(1)} = 0.089$ and $\beta_1^{(2)} = 0.084$.

To investigate the predicative performance of our data fusion method, we select 100 locations within a $10 \times 10$ square block as missing locations, and we create two such blocks for $Z_1$ and only one of the blocks for $Z_2$. Figure S1 shows the visualization of the missing blocks in these two datasets.
Table S1. True parameter values and the average of the estimates over the 1,800 simulated pairs of datasets. The number of observations are fixed at 1,400 for $N_1$ and $N_2$ during each simulation repetition.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>True value</th>
<th>Average of the estimate</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0^{(1)}$</td>
<td>1.0</td>
<td>0.89</td>
<td>0.8097</td>
</tr>
<tr>
<td>$\beta_0^{(2)}$</td>
<td>3.0</td>
<td>3.28</td>
<td>0.8108</td>
</tr>
<tr>
<td>$\beta_1^{(1)}$</td>
<td>0.089</td>
<td>0.0978</td>
<td>0.0282</td>
</tr>
<tr>
<td>$\beta_1^{(2)}$</td>
<td>0.084</td>
<td>0.0871</td>
<td>0.0256</td>
</tr>
<tr>
<td>$\sigma_{\xi}^2$</td>
<td>0.45</td>
<td>0.4480</td>
<td>0.0267</td>
</tr>
<tr>
<td>$\sigma_{\xi}^2$</td>
<td>0.55</td>
<td>0.5777</td>
<td>0.0397</td>
</tr>
</tbody>
</table>

### S4. Performance of the EM algorithm

To assess the performance of the EM estimates, we simulate 1,800 pairs of datasets based on the true parameter values. We also assume that data at some locations are missing. For each pair of the simulated datasets, $\{(Z_1^{[l]}, Z_2^{[l]}) : l = 1, \ldots, 1800\}$, we randomly sample 200 locations from $\mathcal{D}$ as missing for both of two datasets. Then we use the EM algorithm described in Section S1 with the starting values proposed in Section S1.4 to estimate the unknown parameters at each simulation repetition.

We calculate the average of the parameter estimates along with the associated empirical root mean squared error (RMSE). Table S1 shows the summary of results, which is consistent with the true values.

Since $K$ is a $40 \times 40$ matrix, we calculate the elementwise mean of the EM estimates, $\{K_{EM}^{[l]} : l = 1, \ldots, 1800\}$ as:

$$
\text{ave}(K_{EM}) \equiv \frac{1}{L} \sum_{l=1}^{L} K_{EM}^{[l]}, \quad (42)
$$

where $L = 1,800$. Figure S2 shows the image plots of the true values of $K$ and the $\text{ave}(K_{EM})$ over the 1,800 simulation repetitions.
**Figure S2.** The left figure shows the true matrix $K$ and the right figure shows the elementwise average over the 1,800 EM estimates, $\text{ave}(K_{EM})$.

**References**

