

Notes on Error Assessment Using EOFs

Gerald R. North
December 7, 2004

1 Introduction

In statistical parlance *estimation* means setting up *estimators* that are useful in inferring the magnitude of some parameter of the population based upon a limited sample of actual data. We always want to know the degree of bias and the error variance associated with the particular estimator chosen. Different estimators may have no bias but quite different error variances. Often we will want the unbiased estimator with the least error variance, but sometimes we will choose an estimator with a known bias but even less error variance. The choice inevitably depends on the problem being addressed. In our most recent notes, for example, we had one realization of data and we tried to infer the strength of an embedded signal. In this section of the notes we will look at a few further examples of estimation using estimators proportional to our EOF amplitudes. The beauty of this approach is that the estimators will then be statistically independent and therefore can be combined easily into optimal estimators.

2 Estimating an Area Average

In this case we are concerned with a problem such as the global average surface temperature. We want to assess the amount of error associated with a finite number of gauges on the the spherical surface (Shen et al., 1994). There are several gauge configurations that have been used in these estimates in the past and we want to know the error structures. We imagine temporal averages over some time interval τ defined as:

$$\bar{T}_\tau = \frac{1}{4\pi} \int_{4\pi} T_\tau(\hat{\mathbf{r}}, t) d\Omega \quad (1)$$

where

$$T_\tau(\hat{\mathbf{r}}, t) = \frac{1}{\tau} \int_{t-\tau/2}^{t+\tau/2} T(\hat{\mathbf{r}}, t') dt' \quad (2)$$

Departures of the τ -average from the ensemble average are due to natural variability. Such a departure is called an anomaly:

$$\Delta\bar{T}_\tau(\hat{\mathbf{r}}, t) = T_\tau(\hat{\mathbf{r}}, t) - \langle T_\tau(\hat{\mathbf{r}}, t) \rangle \quad (3)$$

Similarly we can define an anomaly of the global average temperature:

$$\Delta\bar{T}_\tau(t) \equiv \bar{T}_\tau(t) - \langle \bar{T}_\tau(t) \rangle \quad (4)$$

By definition

$$\langle \Delta\bar{T}_\tau(t) \rangle = 0 \quad (5)$$

In what follows we deal with the anomaly field and its global average. To keep the notation simple we drop the Δ .

The global average temperature anomaly may be estimated from the data streams collected from a given network of N_{net} stations $\{\hat{\mathbf{r}}_s, s = 1, 2, \dots, N_{net}\}$ by the estimator

$$\hat{\bar{T}}_\tau(t) \equiv \sum_{s=1}^{N_{net}} w_s T_\tau(\hat{\mathbf{r}}_s, t) \quad (6)$$

where w_s is the weight assigned to the s th station. If we assume that $\langle T_\tau(\hat{\mathbf{r}}, t) \rangle$ hardly changes in space, then the no-bias constraint on the weights is

$$\sum_{s=1}^{N_{net}} w_s = 1 \quad (7)$$

We may write the estimator of the global average anomaly into an integral form

$$\hat{\bar{T}}_\tau(t) = \frac{1}{4\pi} \int_{4\pi} w_{net}(\hat{\mathbf{r}}) T_\tau(\hat{\mathbf{r}}, t) d\Omega \quad (8)$$

where

$$w_{net}(\hat{\mathbf{r}}) \equiv 4\pi \sum_{s=1}^{N_{net}} w_s \delta(\hat{\mathbf{r}} - \hat{\mathbf{r}}_s) \quad (9)$$

We may now form the mean squared error:

$$\epsilon^2 = \langle (\bar{T}_\tau - \hat{\bar{T}}_\tau)^2 \rangle \quad (10)$$

$$= \left\langle \left(\frac{1}{4\pi} \int_{4\pi} d\Omega [1 - w_{net}(\hat{\mathbf{r}})] T_\tau(\hat{\mathbf{r}}, t) \right)^2 \right\rangle \quad (11)$$

where ϵ^2 is a function of the weights w_s .

We can expand the formula for the mean squared error and we find:

$$\epsilon^2 = \int_{4\pi} d\Omega' \int_{4\pi} d\Omega'' \left[\frac{1}{(4\pi)^2} - \frac{2}{4\pi} \sum_{s=1}^{N_{net}} w_s \delta(\hat{\mathbf{r}}' - \hat{\mathbf{r}}_s) \right] \quad (12)$$

$$+ \sum_{s,s'=1}^{N_{net}} w_s w_{s'} \delta(\hat{\mathbf{r}}' - \hat{\mathbf{r}}_s) \delta(\hat{\mathbf{r}}'' - \hat{\mathbf{r}}_{s'}) \left] \rho_\tau(\hat{\mathbf{r}}', \hat{\mathbf{r}}'') \quad (13)$$

where

$$\rho_\tau(\hat{\mathbf{r}}', \hat{\mathbf{r}}'') \equiv \langle T_\tau(\hat{\mathbf{r}}', t) T_\tau(\hat{\mathbf{r}}'', t) \rangle \quad (14)$$

2.1 Minimizing the Mean Square Error

Our next task is to find the set of weights w_s which minimize ϵ^2 , subject to the constraint (7). This is accomplished by using the method of Lagrange multipliers: We simply extremize:

$$J[w] = \epsilon^2[w] - 2\Lambda \left[\sum_{s=1}^{N_{net}} w_s - 1 \right], \quad (15)$$

where 2Λ is a Lagrange multiplier. We continue by taking

$$\frac{\partial J}{\partial w_s} = 0, \quad s = 1, \dots, N_{net} \quad (16)$$

$$\frac{\partial J}{\partial \Lambda} = 0 \quad (17)$$

Inserting the expression for ϵ^2 results in

$$\sum_{s'=1}^{N_{net}} w_s \rho_\tau(\hat{\mathbf{r}}_s, \hat{\mathbf{r}}_{s'}) - \Lambda = \frac{1}{4\pi} \int_{4\pi} \rho(\hat{\mathbf{r}}, \hat{\mathbf{r}}_s) d\Omega, \quad (18)$$

$$i = 1, \dots, N_{net} \quad (19)$$

$$\sum_{s=1}^{N_{net}} w_s = 1 \quad (20)$$

These last are a set of $N_{net} + 1$ linear equations for the w_s and Λ .

At last we get to the EOFs which are the eigenfunctions of the kernel $\rho_\tau(\hat{\mathbf{r}}', \hat{\mathbf{r}}'')$:

$$\int_{4\pi} \rho_\tau(\hat{\mathbf{r}}, \hat{\mathbf{r}}') \psi_n(\hat{\mathbf{r}}') = \lambda_n \psi_n(\hat{\mathbf{r}}) d\Omega \quad (21)$$

We can decompose the kernel as

$$\rho_\tau(\hat{\mathbf{r}}, \hat{\mathbf{r}}') = \sum_{n=1}^{\infty} \lambda_n \psi_n(\hat{\mathbf{r}}) \psi_n(\hat{\mathbf{r}}') \quad (22)$$

The set of linear equations become

$$\sum_{s=1}^{N_{net}} \alpha_{ss'} w_{s'} - \Lambda = \beta_s \quad (23)$$

where

$$\alpha_{ss'} = \sum_{n=1}^{\infty} \lambda_n \psi_n(\hat{\mathbf{r}}_s) \psi_n(\hat{\mathbf{r}}_{s'}) \quad (24)$$

$$\beta_s = \frac{1}{4\pi} \sum_{n=1}^{\infty} \lambda_n \psi_n(\hat{\mathbf{r}}_s) \bar{\psi}_n(\hat{\mathbf{r}}_{s'}) \quad (25)$$

and

$$\bar{\psi}_n \equiv \frac{1}{4\pi} \int_{4\pi} \psi_n(\hat{\mathbf{r}}) d\Omega \quad (26)$$

The problem has been reduced to the algebraic one of solving the $N_{net} + 1$ equations.

Shen et al. (1994) proceed by expressing the ψ_n as a series of spherical harmonics, etc.

The actual mean squared error when the optimal weights are used can be expressed as:

$$\epsilon_M^2 = \sum_{n=1}^{N_{modes}} \lambda_n \left[\bar{\psi}_n - \sum_s^{N_{net}} w_s^{opt} \psi_n(\hat{\mathbf{r}}_s) \right]^2 \quad (27)$$

2.2 Time Series of Estimates

Next we show some actual estimates based on optimal versus uniform average weighting of the station data.

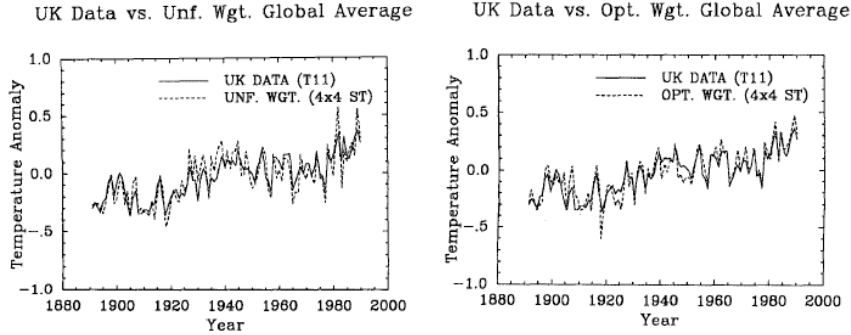


Figure 1: Left Panel: Time series of actual measurements (solid line) and estimates based on uniform weighting for a 16 gauge network uniformly distributed over the globe. Right Panel: Same as LP except that the weighting of the gauge data is optimal.

We return to (6) and feed real data into the estimator using uniform weighting ($w_s = \frac{1}{N_{net}}$) and optimal weighting. The Figures tell the story: there is considerable improvement when optimal weighting is used, but also the errors are very small even with uniform weighting for the A-K Network.

2.3 Statistically Independent Estimators

Now we return to our estimate of the anomaly based on real data:

$$\hat{T}_\tau(t) \equiv \sum_{s=1}^{N_{net}} w_s T_\tau^{data}(\hat{\mathbf{r}}_s, t) \quad (28)$$

If we expand

$$T_\tau^{data}(\hat{\mathbf{r}}, t) = \sum_{n=1}^N T_{\tau,n}^{data} \psi_n(\hat{\mathbf{r}}) \quad (29)$$

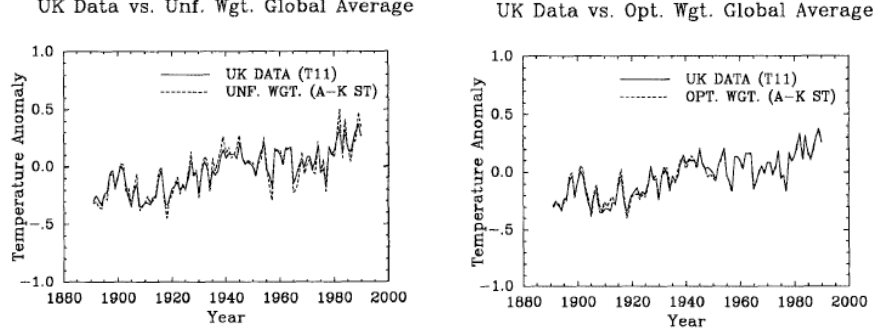


Figure 2: Same as previous Figure except for the 63 Station Angel-Korshover Network

Then we have:

$$\hat{T}_\tau(t) \equiv \sum_{s=1}^{N_{net}} \sum_{n=1}^{N_{modes}} w_s T_{\tau,n}^{data} \psi_n(\hat{\mathbf{r}}_s) \quad (30)$$

Note that if we consider an ensemble of data streams the $T_{\tau,n}^{data}$ are independent from one value of n to another. In other words we have a set of N statistically independent estimators of the error.

The accumulated error can be studied as a function of the number of EOFs retained in the estimate. A very interesting aspect of the

formula (30) is that it can be further regrouped:

$$\hat{T}_\tau(t) = \sum_{n=1}^{N_{modes}} \mathcal{D}_n T_{\tau,n}^{data} \quad (31)$$

where

$$\mathcal{D}_n \equiv \sum_{s=1}^{N_{net}} w_s \psi_n(\hat{\mathbf{r}}_s) \quad (32)$$

is called the *design factor*. The regrouping in (31) shows that the sum over modes consists of two factors, \mathcal{D}_n which depends only on the observational design (projected on mode n) and the second factor which depends only on the realization of the data stream. In

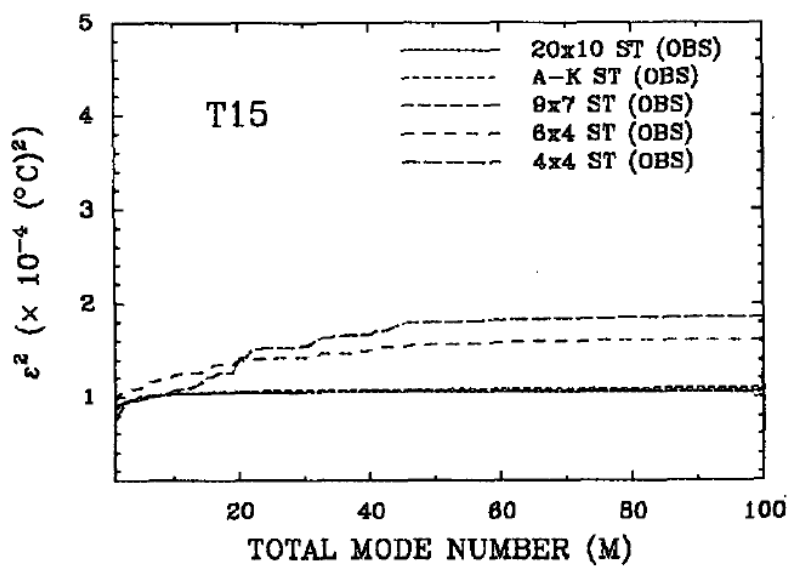


Figure 3: Graphs of accumulated error squared as a function of the number of EOF estimators retained for five different gauge configurations. Most important is the A-K Network, which shows that the errors level off below ten EOF modes. The indicator T15 is the truncation level of spherical harmonics used by Shen et al., (1994).

effect, the factor \mathcal{D}_n acts as a filter through which the data streams passes to finally obtain an estimate of the anomaly for that particular realization of the data.

A similar decomposition occurs in the mean squared error formula (13), which can be written

$$\epsilon^2 = \int_{4\pi} d\Omega' \int_{4\pi} d\Omega'' \mathcal{D}^{(2)}(\hat{\mathbf{r}}', \hat{\mathbf{r}}'') \rho(\hat{\mathbf{r}}', \hat{\mathbf{r}}'') \quad (33)$$

where $\mathcal{D}^{(2)}(\hat{\mathbf{r}}', \hat{\mathbf{r}}'')$ comes from (13). Again the integrand factors into two parts one of which depends only on the design and the other only on the structure of the random field being sampled. This formula can assist us in deciding how to assess the error for a given design. Once we know the design of the observing system ($\mathcal{D}^{(2)}(\hat{\mathbf{r}}', \hat{\mathbf{r}}'')$), we can find out what needs to be measured ($\rho(\hat{\mathbf{r}}', \hat{\mathbf{r}}'')$) in order to complete the assessment.