Multivariate Analysis of the North Atlantic Ocean - the problem of missing values

Peter Challenor
Richenda Houseago-Stokes

Southampton Oceanography Centre
UNIVERSITY OF SOUTHAMPTON AND NATURAL ENVIRONMENT RESEARCH COUNCIL

Sea Surface Temperature (SST)

- Meteorologists and oceanographers are interested in the interaction between the atmosphere and ocean
- The ocean influences the atmosphere through the sea surface temperature
- (the atmosphere's influence on the ocean is more complex)
- SST is measured from ships, buoys and satellites

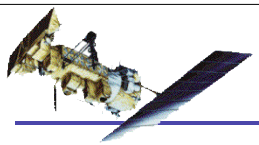
Southampton Oceanography Centre
UNIVERSITY OF SOUTHAMPTON AND NATURAL ENVIRONMENT RESEARCH COUNCIL

Sea surface temperature

- From satellite SST we can identify and monitor surface disturbances that cross entire ocean basins, track ocean eddies and map ocean fronts
- It can also reveal striking features such as ‘storms’ in the upper ocean, known as eddies.
  - These are typically ~100 km wide and carry large amounts of energy around the globe. They play an important role in ocean circulation and climate.
- Space-borne infra-red sensors estimate SST by measuring heat radiation from the ocean surface.

SST

- SST is measured using a infrared radiometer
- Spectral bands used are near the peak of surface emission - the peak ones aren’t used due to atmospheric effects
- It is measured by:
  - taking the intensity of radiation at top of atmosphere
  - removing the atmospheric contribution
  - Resulting in the brightness temperature at the surface
- Brightness temp is approximately equal to the SST
- Typical SST products:
  - NOAA MCSST global
  - ATSR ASST



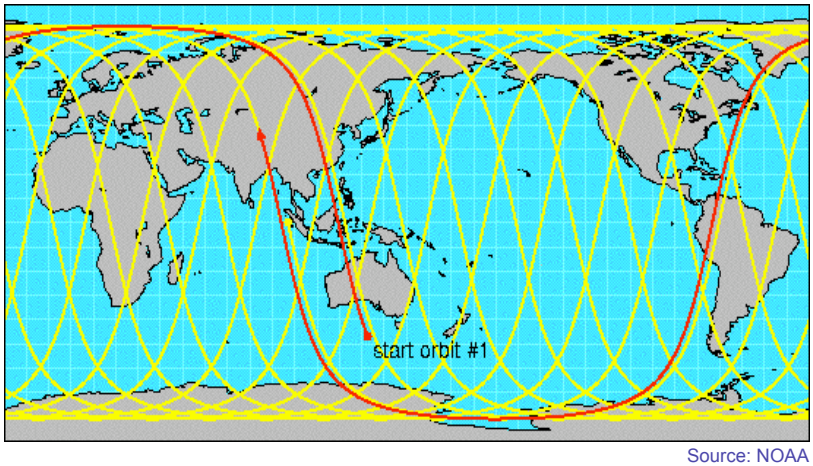
AVHRR

- The AVHRR instruments are onboard the NOAA-7, -9, -11 and -14 satellites
- The radiometer uses 5 detectors that collect different bands of radiation wavelengths

0.63 μm	Daytime cloud and surface mapping
0.86 μm	Land-water boundaries
3.74 μm	Night cloud mapping, sea surface
10.80 μm	Night cloud mapping, sea surface
11.50 μm	Sea surface temperature

- Data from both the ascending pass (daytime) and descending pass (nighttime) are available with a spatial resolution of about 5.7 pixels per degree of longitude and latitude.

Polar orbit track

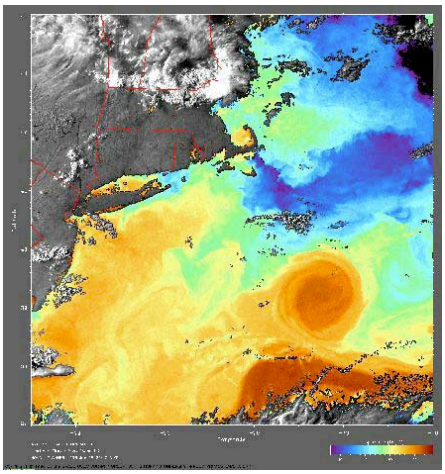


Source: NOAA

Cloud

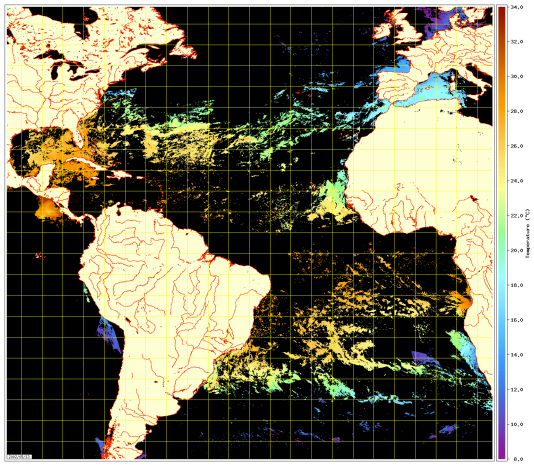
- Infrared radiation does not penetrate cloud
- Therefore where it is cloudy we get no data
- This means we have intermittent data

SST image



West Atlantic 15th June 1996 (NOAA)

Cloud contamination

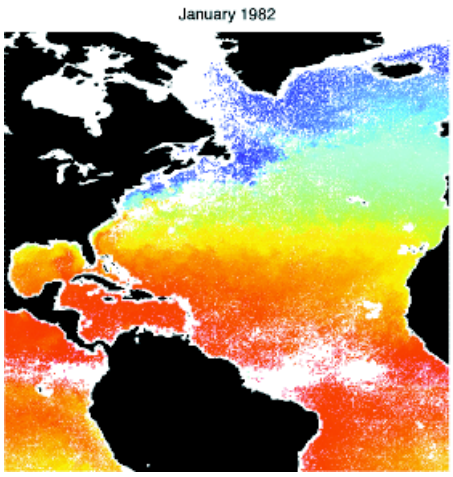


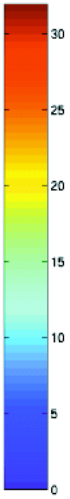
AVHRR image 13th May 2002 (NOAA)

MCSST Data

- Weekly averaged MCSST data are available from 11 November 1981 to 7 February 2001
- Data from January 1984 to December 2000 are being used - problem due to aerosols from the eruption of El Chichon
- Daytime and nighttime data is available.
- Areas, particularly north of 60˚N, have few values due to extensive cloud cover and sea ice.
- MCSST data are produced at 18km resolution we average to 1°

North Atlantic SST time series





Southampton Oceanography Centre
UNIVERSITY OF SOUTHAMPTON AND NATURAL ENVIRONMENT RESEARCH COUNCIL

PCA with missing values

- Conventional principal component analysis cannot be done if there are missing values
- Two possible approaches:
  - Interpolate the data before doing the analysis
  - Use a modified version of PCA that can cope with missing values

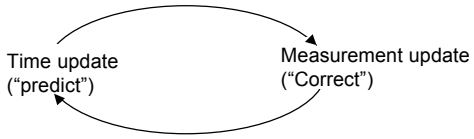
Southampton Oceanography Centre
UNIVERSITY OF SOUTHAMPTON AND NATURAL ENVIRONMENT RESEARCH COUNCIL

Interpolating the data

- We could interpolate using ‘spatial statistics’ (kriging)
- In fact we use a Kalman Filter instead
  - The Kalman Filter consists of two steps
  - A time update and a measurement update.

Update Equations

- The time update equations project forward in time the current state and the error covariance estimate to give an *a priori* estimate for the next time step.
- The measurement update equations are responsible for the feedback, i.e. incorporating a new measurement into the *a priori* estimate.
- The time update equations are the predictor equations with the measurement update equations as the corrector equations.
- The final estimation algorithm resembles that of a predictor corrector algorithm for solving numerical problems



Kalman Gain

- The data set was interpolated using a Kalman Filter
- The first task is to calculate the Kalman gain :

$$K = H^T R^{-1} (H^T R^{-1} H + S_a^{-1})^{-1}$$

- S_a the priori covariance includes latitude, longitude and time components.

$$[S_a]_{ij} = a_0^2 \, a_1^{|\phi_i - \phi_j|} \, a_2^{|t_1 - t_2|}$$

- R is the noise term (= $\sigma^2 I$).
- H is the measurement model (=I).

Incorporating the Measurement

- The next step is to use the observations, y, and to generate an *a posteriori* state estimate by incorporating the measurement

$$\hat{\underline{x}} = \underline{x}_a + K(\underline{y} - H\,\underline{x}_a)$$

  - $\underline{x}_a$= 0 at first then $\underline{x}_{t-1}$ after.
  - y is the observed values.

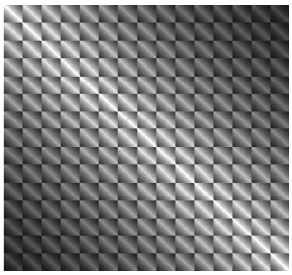
- These equations operate on the linear version of the problem, and operates sequentially in time or space. In this case the interpolation is spatial.

Step by step guide

- Initially we use the previous time step as an estimate for the missing data, $\underline{x}_a$
- As simplification, the measurement model, H, is taken as the identity matrix.
- A local Kalman filter is then used on a local spatial grid (15 by 15 degrees) as a method of 'spatial interpolation'. The weighting of the surrounding values uses the equation:

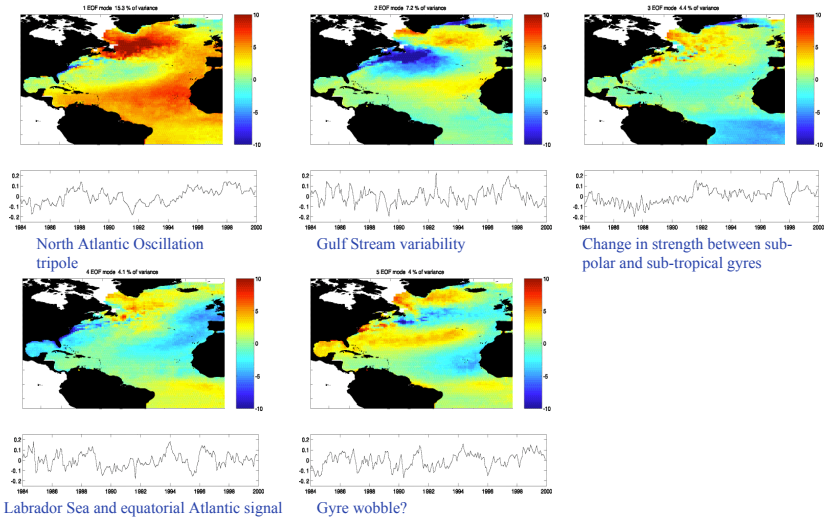
$$\left[S_a\right]_{ij} = a_0^{\,2}\, a_1^{|\square_i - \square_j|}\, a_2^{|\square_1 - \square_2|}$$

- The generated value is used to 'replace' the missing value.



S_a the priori covariance for the 15 by 15˚ grid

North Atlantic modes of variability



North Atlantic Oscillation tripole

Gulf Stream variability

Change in strength between sub-polar and sub-tropical gyres

Labrador Sea and equatorial Atlantic signal

Gyre wobble?

Probabilistic Principal Components

- Assume that the data we have is a linear combination of some 'hidden' or latent variables.
- Then

$$\mathbf{t} = \mathbf{W}\mathbf{x} + \boldsymbol{\mu} + \boldsymbol{\varepsilon}$$

- Where t is the observed data (n x 1), x is the latent variables (p x 1), W is a matrix (n x p), $\boldsymbol{\mu}$ is the mean of t and $\boldsymbol{\varepsilon}$ is an error term

PPCA-2

- The **x** are iid N(**0**,**I**)
- $\boldsymbol{\varepsilon}$~N(0,$\boldsymbol{\Psi}$)
- Then **t**~N($\boldsymbol{\mu}$, **WW**^{t}+$\boldsymbol{\Psi}$)
- If $\boldsymbol{\Psi}$ is diagonal then this is Factor Analysis
- If we put $\boldsymbol{\Psi}=\sigma^2\mathbf{I}$ then this becomes equivalent to Principal Components
- So PPCA enables us to calculate the first p principal components ((n-p)σ^2 gives the variance not explained)

The EM algorithm (Roweis)

- Expectation step

$$X = (C^T C)^{-1} C^T T$$

- Maximisation step

$$C^{new} = TX^T (XX^T)^{-1}$$

where:

$$C = W^T W$$

The EM algorithm for Missing Data

- e-step $X = (C^T C)^{-1} C^T Y$

$$Y = CX$$

- m-step $C^{new} = YX^T (XX^T)^{-1}$

- **Y** is a matrix of the observed data *(pxn)*
- **X** is the matrix of the unknown states *(kxn)*
- **C** is the estimated rotated covariance matrix *(pxk)*

These equations are iterated until they converge.

Algorithm for PPCA with missing values

- Start with an initial guess of the covariance matrix.
- Cycle through a series of steps, replacing the missing values.
- Then re-estimate the covariance matrix.
- Continue until the values converge.

How many EOF's to retain?

1) Look at the cumulative percentage variation explained.
2) Look at the size of the variance explained by each EOF - reject if less than one.
3) The scree plot.

However, none of these methods work for this technique.

For this study we have decided to retain 9 EOF's

Removing data to test the EM algorithm

Two methods are used to test the EM algorithm

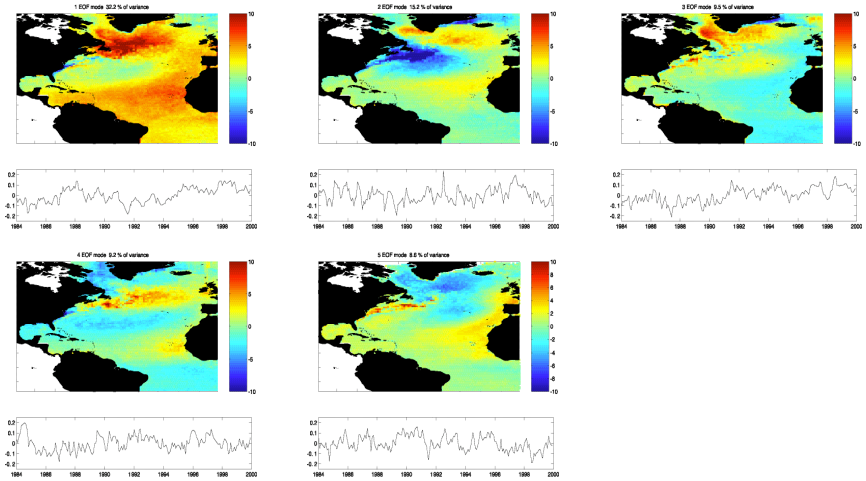
1. Data are removed randomly from the previously interpolated data set before the EM algorithm.

2. Data are removed from a block in the centre of the previously interpolated data set, randomly over time - to simulate large cloud bands.

Southampton Oceanography Centre
UNIVERSITY OF SOUTHAMPTON AND NATURAL ENVIRONMENT RESEARCH COUNCIL

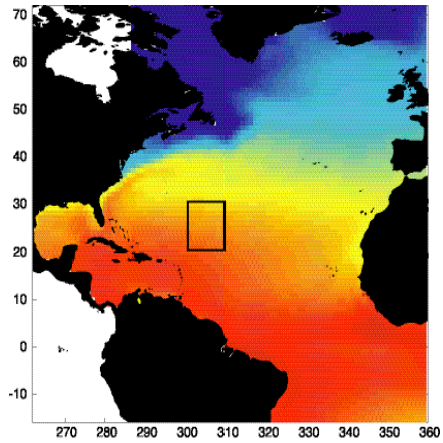
First 5 EOF's using the EM Algorithm

with 10% of data removed



Southampton Oceanography Centre
UNIVERSITY OF SOUTHAMPTON AND NATURAL ENVIRONMENT RESEARCH COUNCIL

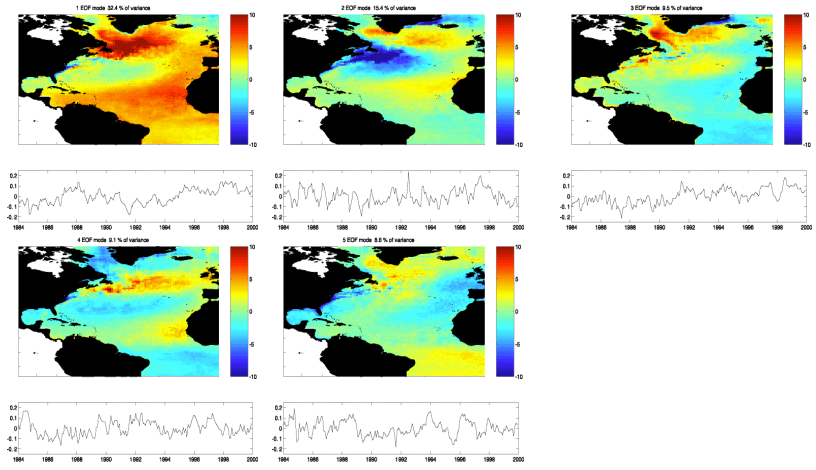
Removing data from a block



Southampton Oceanography Centre
UNIVERSITY OF SOUTHAMPTON AND NATURAL ENVIRONMENT RESEARCH COUNCIL

First 5 EOF's using the EM Algorithm

with 10% of data removed from a block



Southampton Oceanography Centre
UNIVERSITY OF SOUTHAMPTON AND NATURAL ENVIRONMENT RESEARCH COUNCIL

EOF breakdown

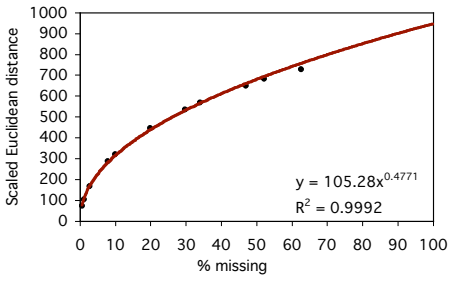
The first five EOF's appeared unaffected by the missing values, until 30% of the data was removed

The reconstructed data was compared to the original data using the scaled Euclidean distance

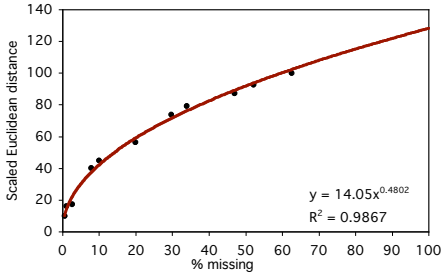
$$D_s = \sqrt{\sum_{i=1}^{N} \frac{(x_i - y_i)^2}{\sigma^2}}$$

Southampton Oceanography Centre
UNIVERSITY OF SOUTHAMPTON AND NATURAL ENVIRONMENT RESEARCH COUNCIL

EOF Breakdown



Random missing data

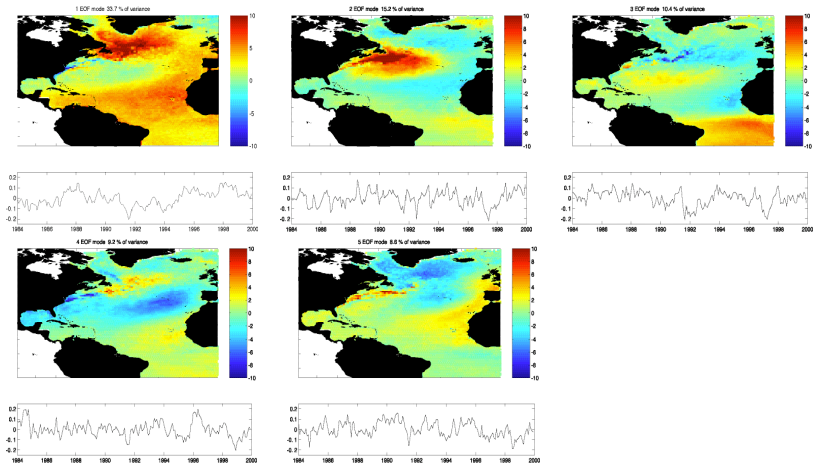


Block missing data

Southampton Oceanography Centre
UNIVERSITY OF SOUTHAMPTON AND NATURAL ENVIRONMENT RESEARCH COUNCIL

First 5 EOF's using the EM Algorithm

using uninterpolated data



Southampton Oceanography Centre
UNIVERSITY OF SOUTHAMPTON AND NATURAL ENVIRONMENT RESEARCH COUNCIL

PPCA Summary

- The EM algorithm has produced similar results to those produced through standard procedures.
- Calculation of the EOF's was quicker, as the full covariance matrix was not calculated.
- The replacement of missing values was computationally more efficient than other interpolation methods.

Southampton Oceanography Centre
UNIVERSITY OF SOUTHAMPTON AND NATURAL ENVIRONMENT RESEARCH COUNCIL

References

- Roweis, S., 1997: EM Algorithms for PCA and SPCA, *Neural Information Processing Systems 10 (NIPS'97)*, 626-632.
- Tipping, M.E. and C.M. Bishop, 1999: Probabilistic Principal component analysis. *Journal of the Royal Statistical Society, Series B*, 61(3), 611-622.
- Houseago-Stokes, R.E. and Challenor P.G. Using PPCA to estimate EOF's in the presence of missing values, Submitted to the Journal of Atmospheric and Oceanic Technology