Missing values

Evaluate the nature and the distribution of the gaps, to see whether a remedy must be applied before further analysis.

How do you deal with missing values?

- How bad is the problem?
  - How many are there?
  - How are they distributed?
  - How are they correlated?
- How do you fix the problem?
  - Discard variables or case
  - Imputation

Types of missingness

- MCAR: probability that a value is missing does not depend on any other observed or unobserved value.
- MAR: probability that a value is missing depends only on the observed variables.
- MNAR: the reason for missing values depends on some unseen or unobserved information - very difficult analysis.

How many?
How distributed?

Case	X_1	X_2	X_3	X_4	X_5
1	NA	20	1.8	6.4	−0.8
2	0.3	NA	1.6	5.3	−0.5
3	0.2	23	1.4	6.0	NA
4	0.5	21	1.5	NA	−0.3
5	0.1	21	NA	6.4	−0.5
6	0.4	22	1.6	5.6	−0.8
7	0.3	19	1.3	5.9	−0.4
8	0.5	20	1.5	6.1	−0.3
9	0.3	22	1.6	6.3	−0.5
10	0.4	21	1.4	5.9	−0.2

Missing:
10% of the numbers
100% of variables
50% of samples

Shadow Matrix

Case	X_1	X_2	X_3	X_4	X_5
1	NA	20	1.8	6.4	−0.8
2	0.3	NA	1.6	5.3	−0.5
3	0.2	23	1.4	6.0	NA
4	0.5	21	1.5	NA	−0.3
5	0.1	21	NA	6.4	−0.5
6	0.4	22	1.6	5.6	−0.8
7	0.3	19	1.3	5.9	−0.4
8	0.5	20	1.5	6.1	−0.3
9	0.3	22	1.6	6.3	−0.5
10	0.4	21	1.4	5.9	−0.2

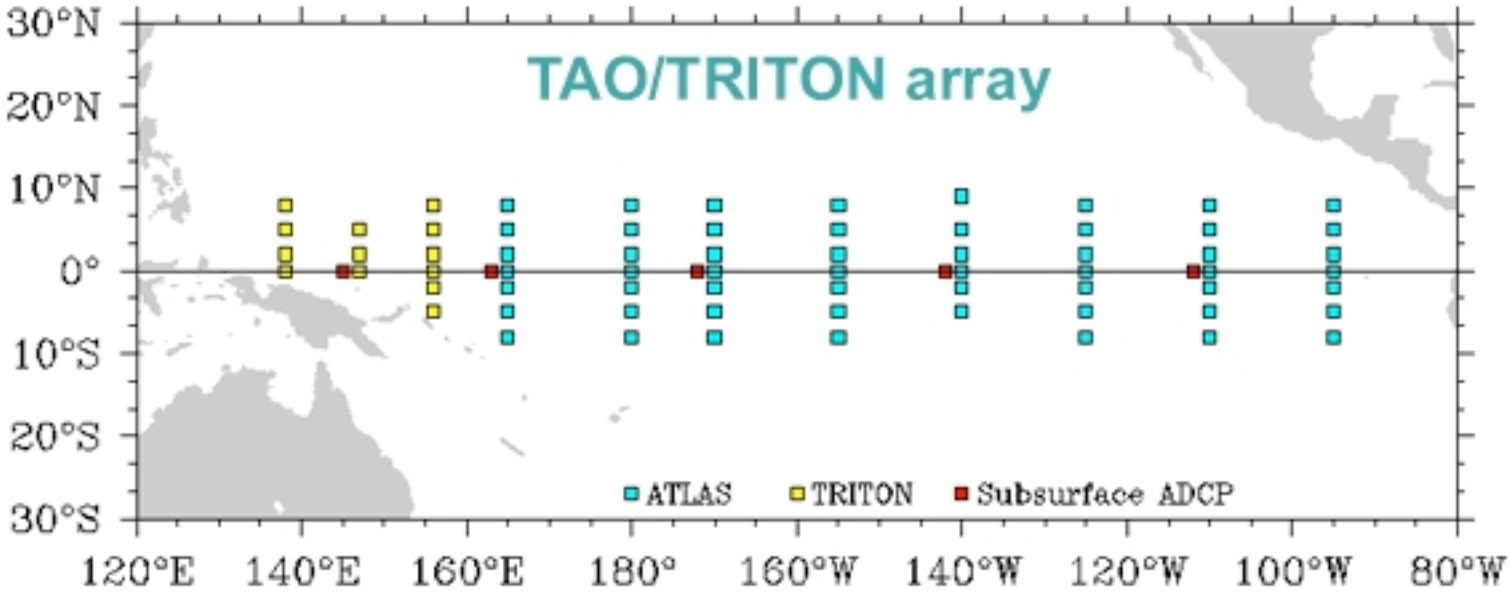
Case	X_1	X_2	X_3	X_4	X_5
1	1	0	0	0	0
2	0	1	0	0	0
3	0	0	0	0	1
4	0	0	0	1	0
5	0	0	1	0	0
6	0	0	0	0	0
7	0	0	0	0	0
8	0	0	0	0	0
9	0	0	0	0	0
10	0	0	0	0	0

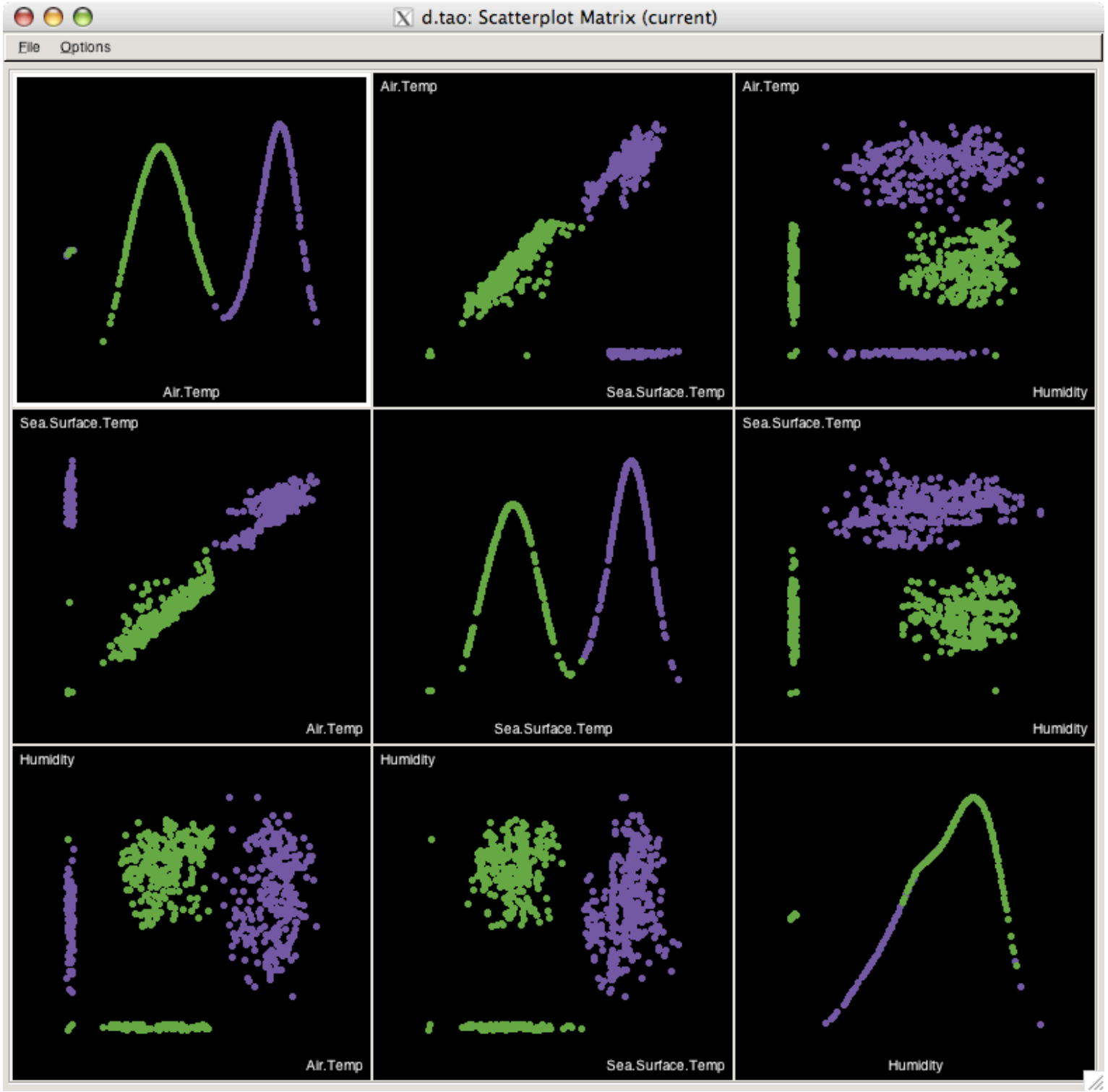
Example

Tropical Atmosphere-Ocean Array

Number of cases: 736
Number of variables: 8

Sea Surface Temp, Air Temp, Humidity, UWind, VWind + Year, Lat Long




d.tao: Scatterplot Matrix (current)
File
Options
Air.Temp
Air.Temp
Sea.Surface.Temp
Air.Temp
Humidity
Sea.Surface.Temp
Air.Temp
Sea.Surface.Temp
Sea.Surface.Temp
Humidity
Humidity
Air.Temp
Humidity
Sea.Surface.Temp
Humidity

1993 Normal

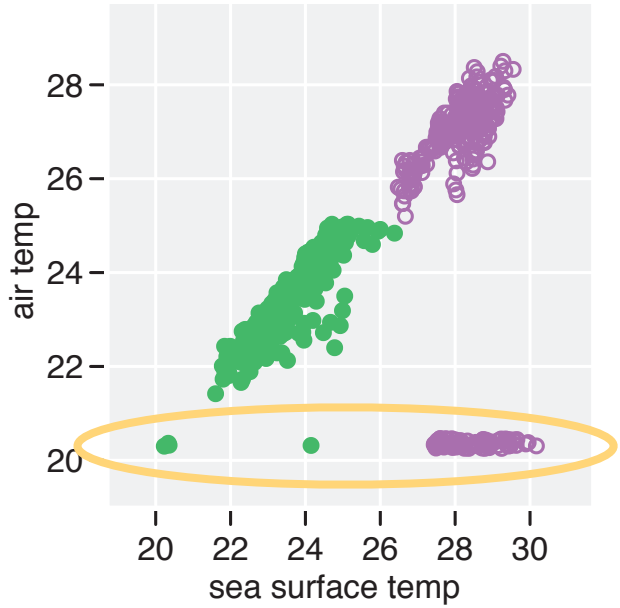
Overview

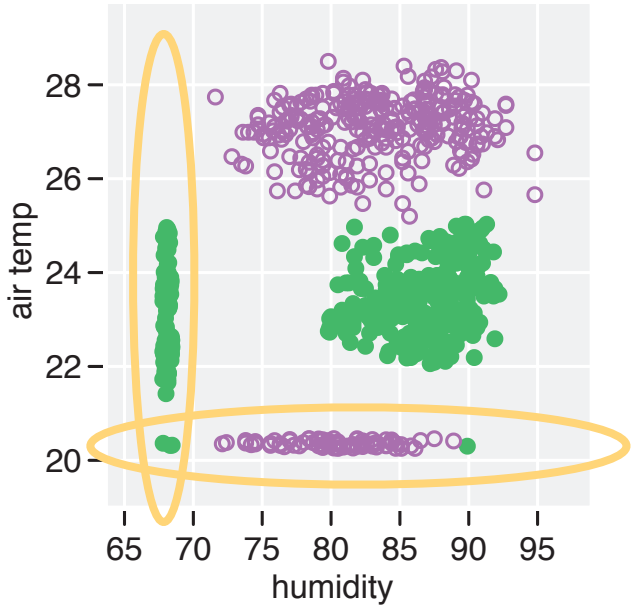
1997 El Nino

Variable	Number of missing values	
	1993	1997
sea surface temp	3	0
air temp	4	77
humidity	93	0
uwind	0	0
vwind	0	0

No. of missings on a case	1993		1997	
	No. of cases	%	No. of cases	%
3	2	0.5	0	0
2	2	0.5	0	0
1	90	24.5	77	20.9
0	274	74.5	291	79.1

Using the margins



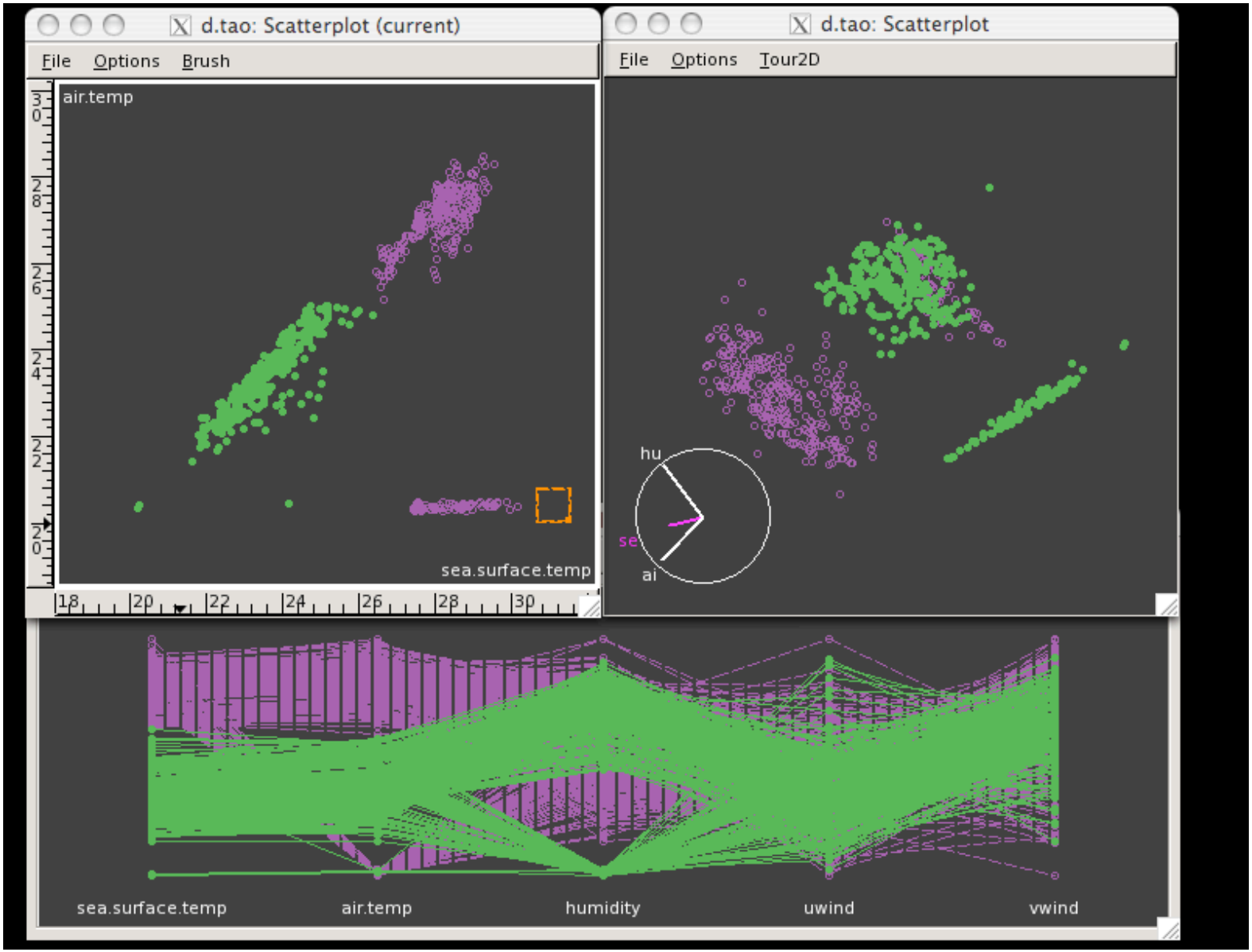


Correlation between temperatures. Years separated. More missings on air temp than sea surface temp.

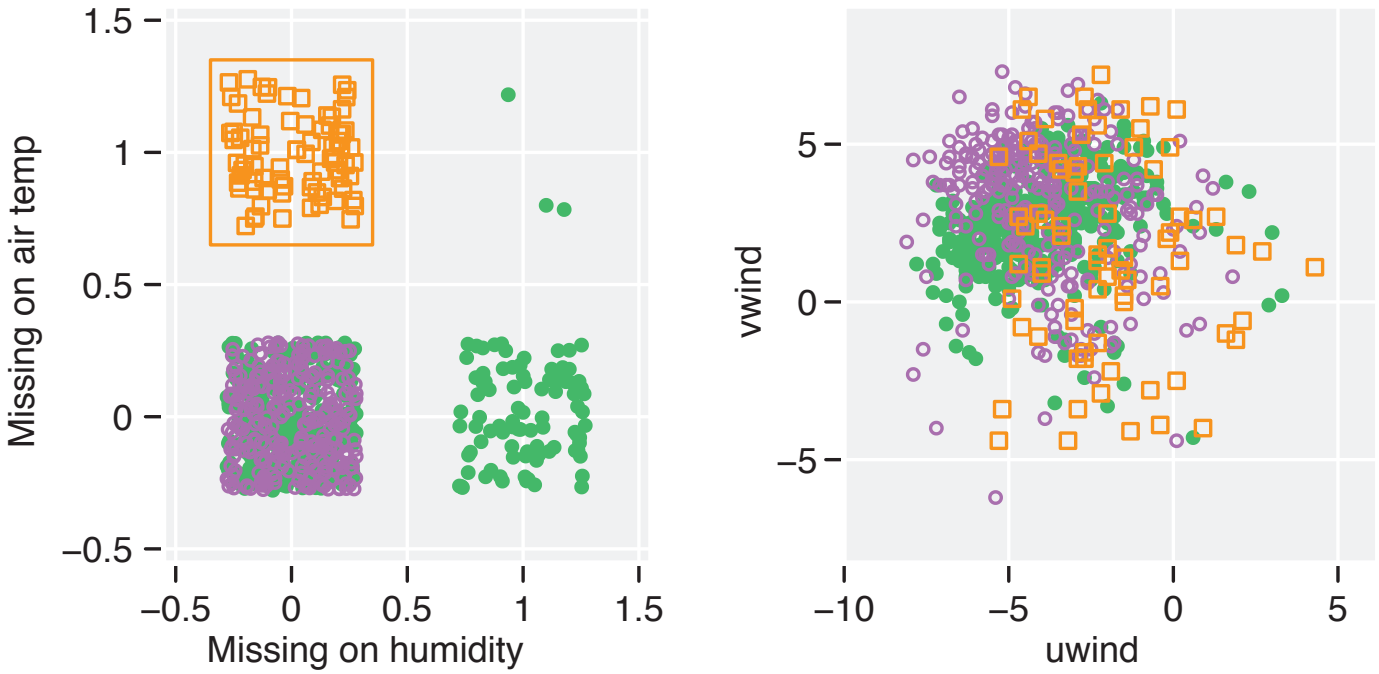
Missings on humidity only occur in 1997.

Limitation

Missings look like clusters in high-d plots



Tracking missings using the shadow matrix



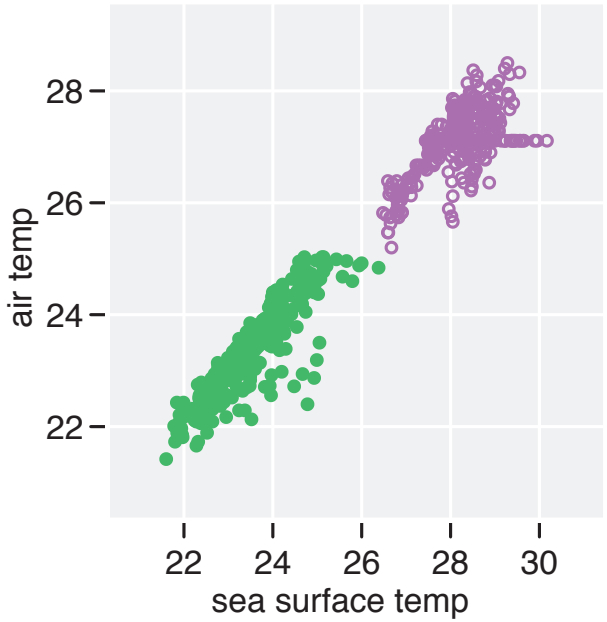
Missings on air temp have higher values on uwind than non-missings.

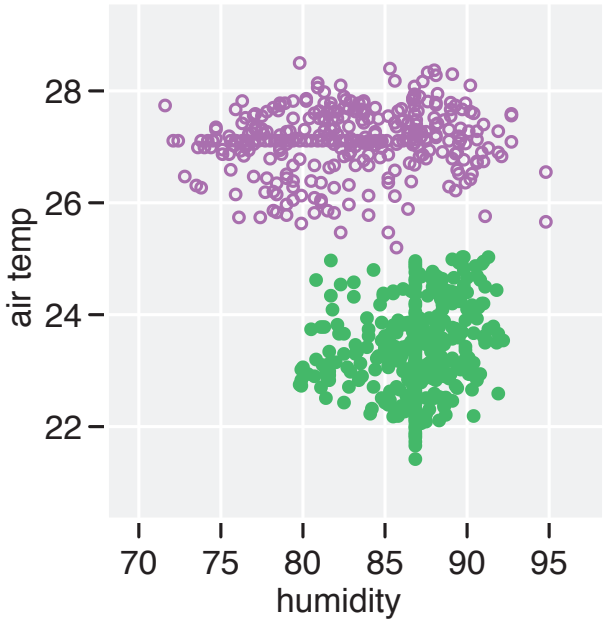
Missing Structure

Missing values are NOT MCAR!

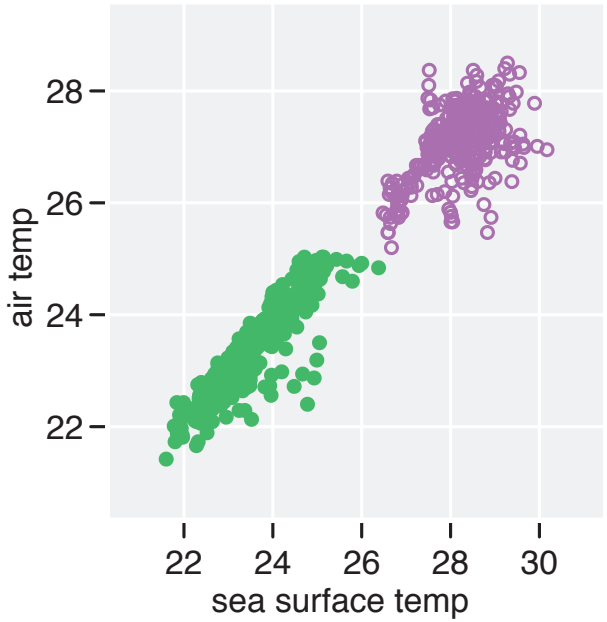
Imputation will need to use dependence of missing and not missing.

Imputing missings

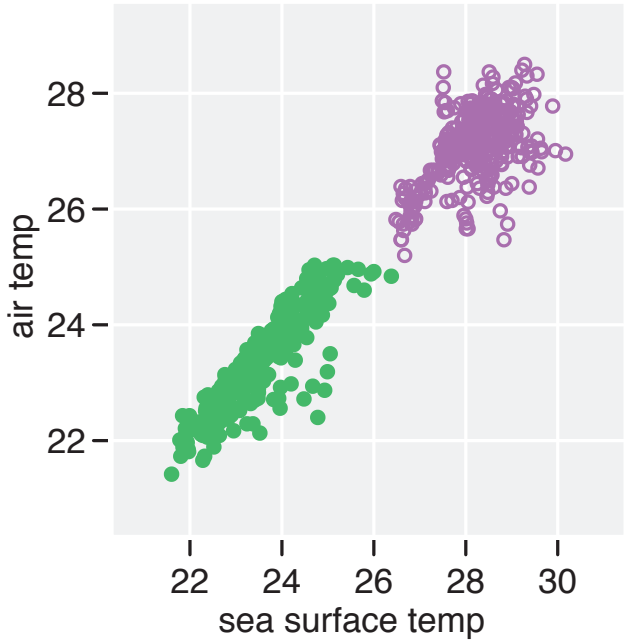




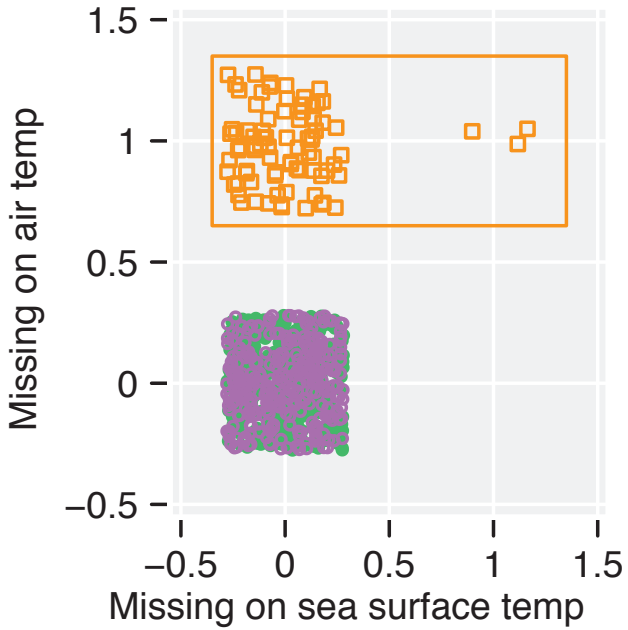
Means for each year

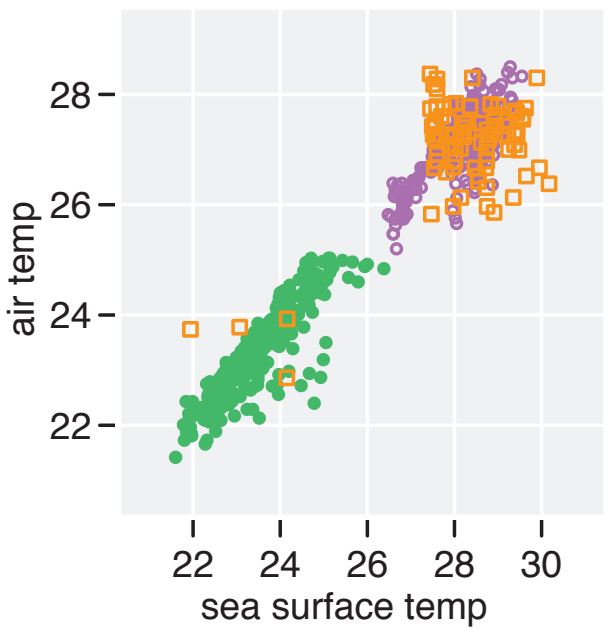


Random values from each year



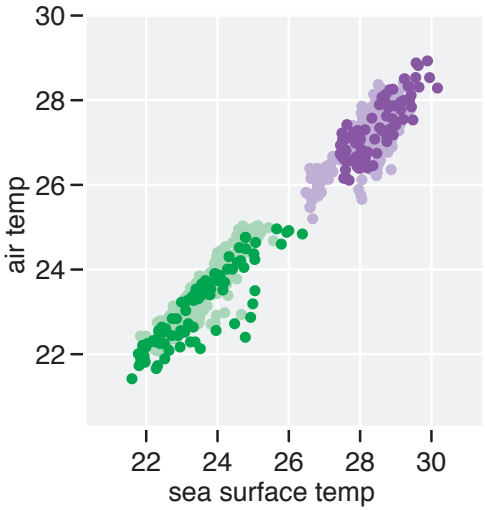
Imputed values which disappeared can be revealed by brushing on the shadow matrix.

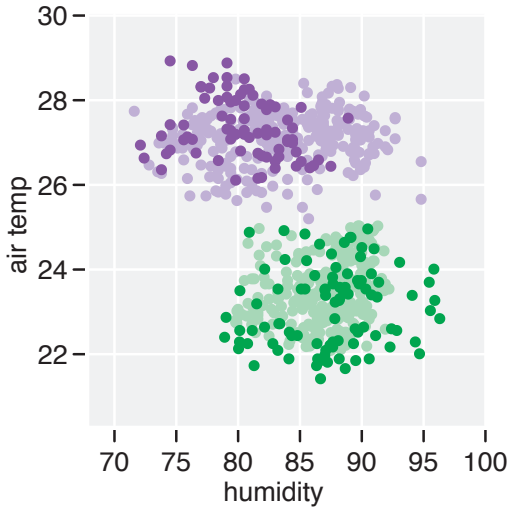


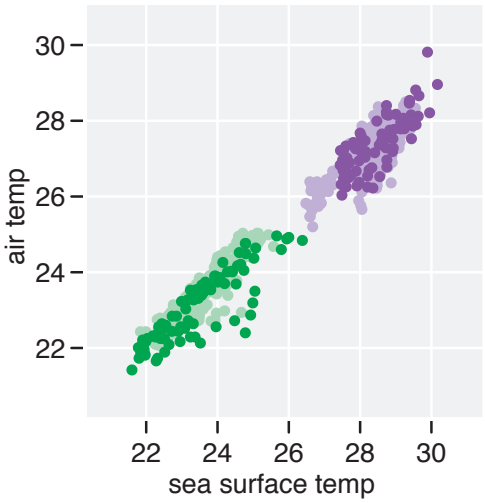


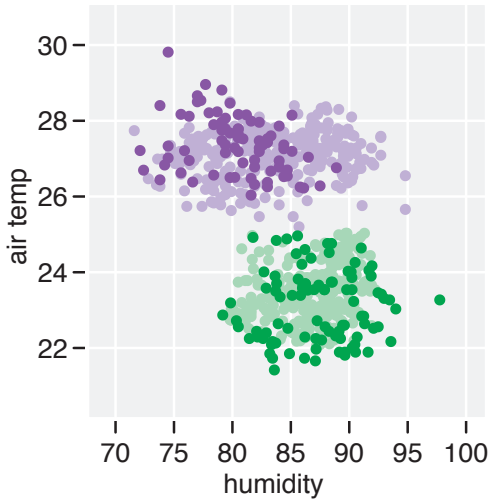
Imputed values which disappeared can be revealed by brushing on the shadow matrix.

Multiple imputation










```
> rngseed(1234567)
> theta.93 <- em.norm(
  d.tao.nm.93,
  showits=TRUE)
> d.tao.impute.93 <-
  imp.norm(d.tao.nm.93,
  theta.93, d.tao.93)
```

Conclusions

- Used the same plots for exploratory visualisation and to diagnose imputations
- Used both analytic and visual methods throughout

Timeline

20 mins	Toolbox
30 mins	Missing values
45 mins	Supervised Classification
Break	
45 mins	Unsupervised Classification
30 mins	Inference

Your turn

Describe the distribution of the wind and temperature variables conditional on the distribution of missing values in humidity, using brushing and the tour.