WHAT TO DO WHEN SOME VALUES ARE MISSING

Statistics 407
ISU

OUTLINE

- Terminology
- Issues of missingness for multivariate data
- Plotting missings, and describing the distributions of missing vs not missing
- Imputation methods

BACKGROUND TERMS

- MCAR: probability that a value is missing does _____ on any other observed or unobserved value.
- MAR: probability that a value is missing _____ only on the _____ variables.
- MNAR: the reason for missing values depends on some _____ information - very difficult analysis.

EXAMPLE

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:

____ of the numbers

____ of variables

____ of samples

SUMMARY STATISTICS

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

Means can be calculated _____.

Correlations can be calculated _____.

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

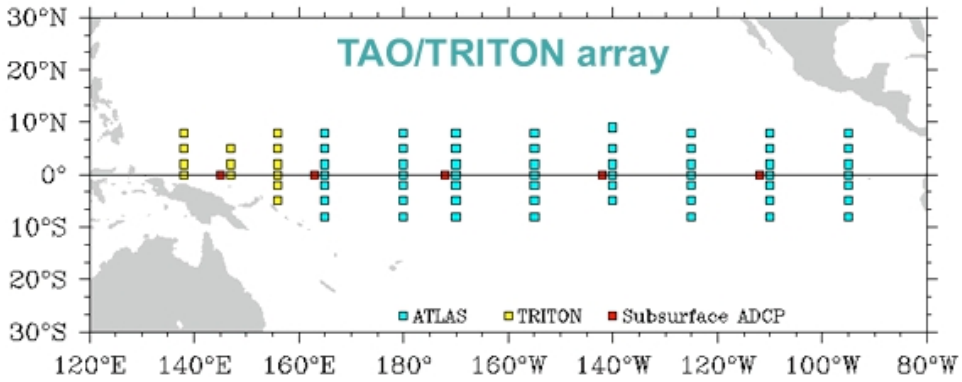
EXAMPLE

Tropical Atmosphere-Ocean Array

Number of cases: 736

Number of variables: 8

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



OVERVIEW

1993 El Nino

1997 Normal

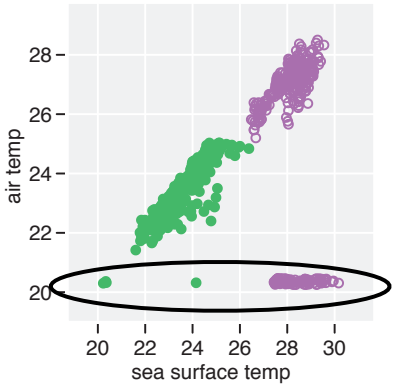
Variable	Number of missing values	
	1993	1997
sea surface temp	3	0
air temp	4	77
humidity	93	0
uwind	0	0
vwind	0	0

R package: norm

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

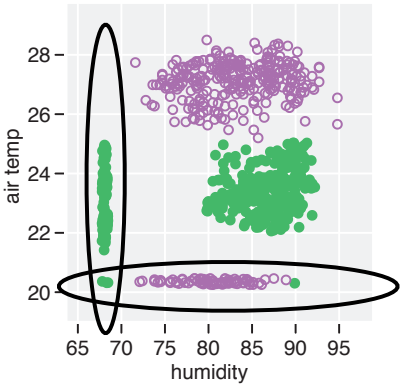
USING THE MARGINS

1993 El Nino



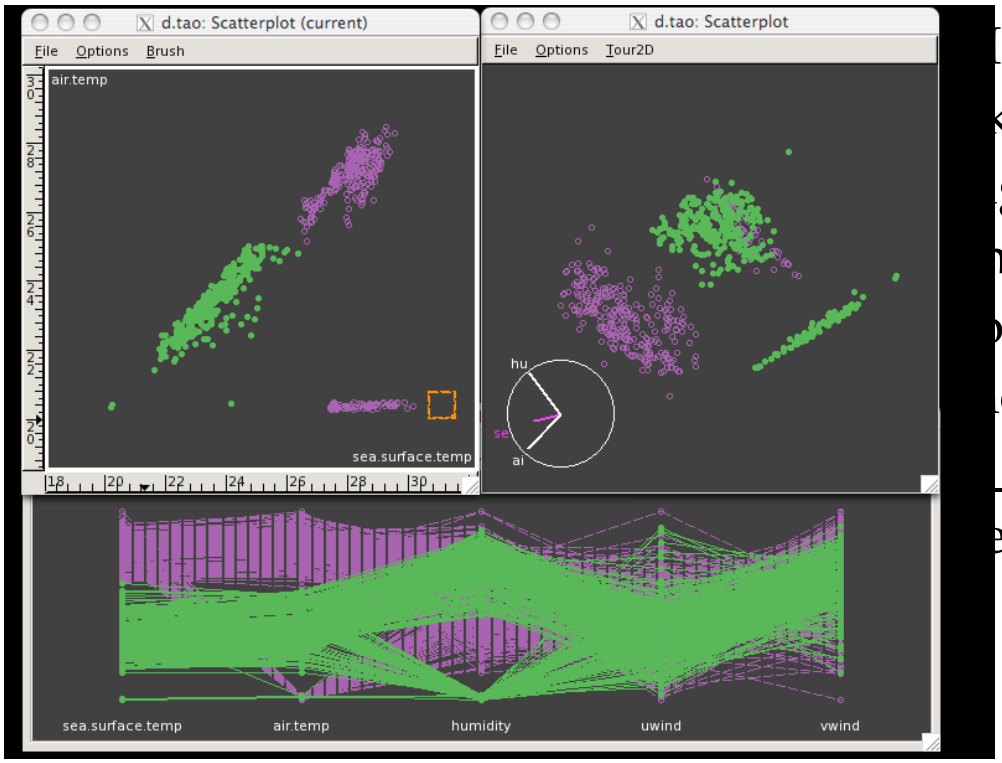
Association between temperatures. Years separated. More missings on _____ than _____.

1997 Normal



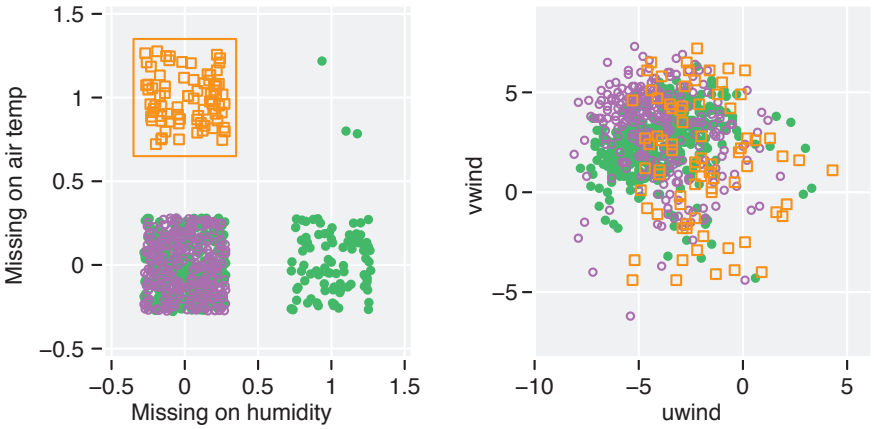
Missings on _____ only occur in 1997.

LIMITATION



issings look
ke _____ in
gh-d plots,
nd in parallel
oordinates
ey look like
_____ at the
ery bottom.

TRACKING MISSINGS USING THE SHADOW MATRIX



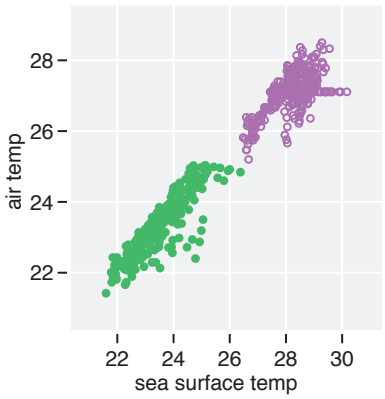
Missings on air temp have _____ values on uwind than non-missings.

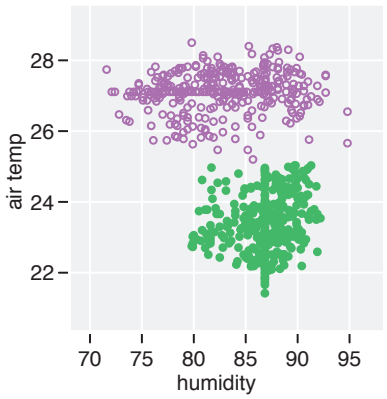
MISSING STRUCTURE

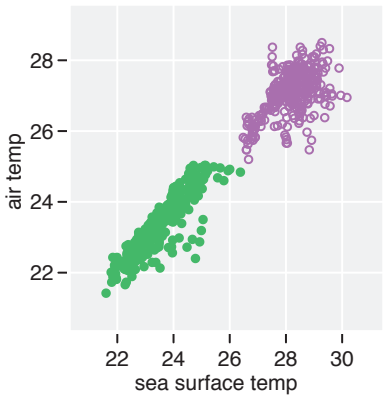
Missing values are _____!

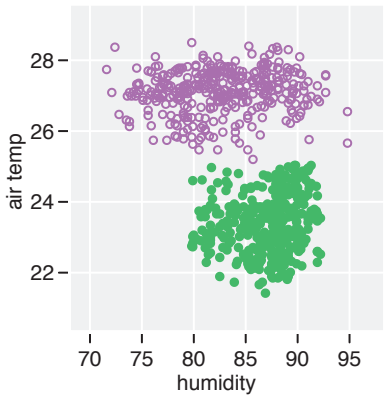
Imputation will need to use _____ of missing and not missing.

IMPUTING MISSINGS









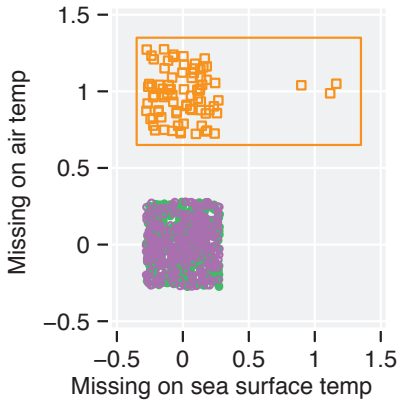
_____ for each year

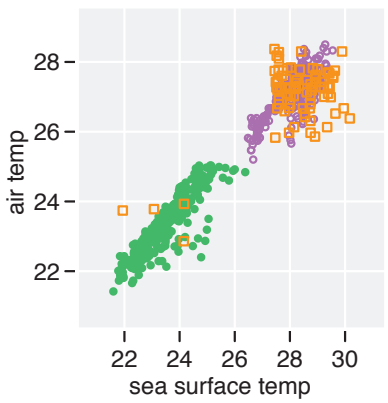
_____

from each year

What do you notice?

USING THE SHADOW MATRIX

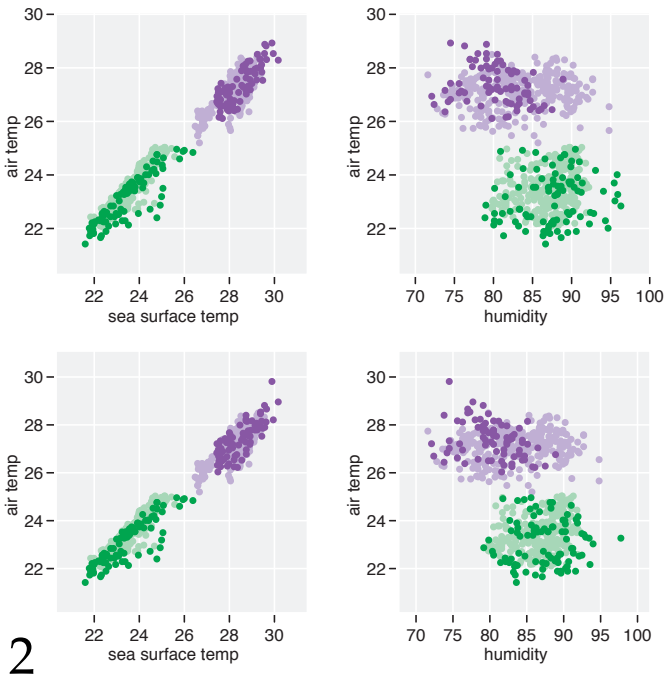




Imputed values which disappeared can be revealed by _____ on the shadow matrix.

MULTIPLE IMPUTATION

1



2

3, 4, 5, ...

Missing values are imputed by simulating from a _____ _____, having mean vector and variance-covariance matrix equal to the sample quantities. Sampling _____ times allows for estimating statistics for the missing values.

SUMMARY

- _____ missings: by variable, by case
- _____ plots of missings, in the margins
- _____ summary statistics using as much data as possible.
- Determine _____ of missings: MAR, MCAR, MNAR
- Decide on a good way to _____ missings, as simple as possible with out affecting results.