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 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.

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

Deleting missings is not usually an option.

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 variable-wise.

Correlations can be calculated pairwise.

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

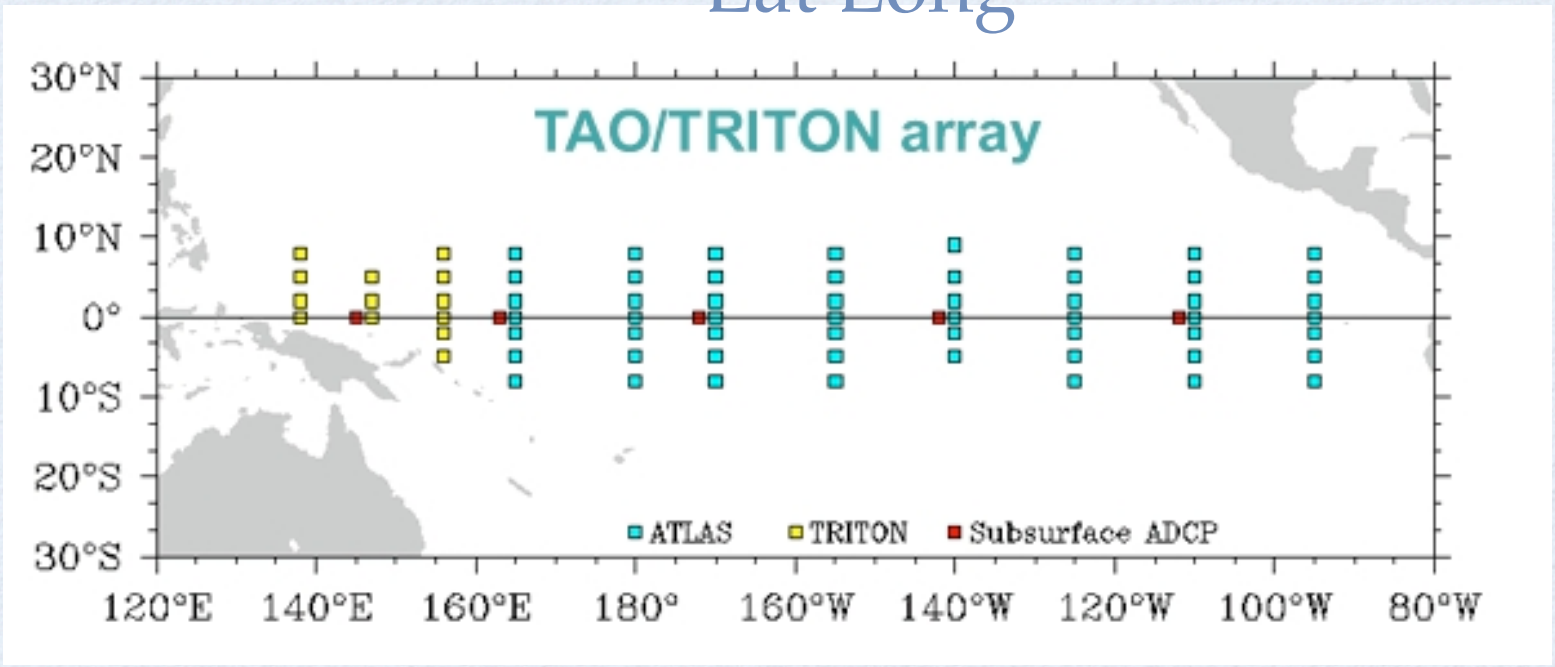
EXAMPLE

Tropical Atmosphere-Ocean Array

Number of cases: 736

Number of variables: 8

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



OVERVIEW

1997 El Nino

1993 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

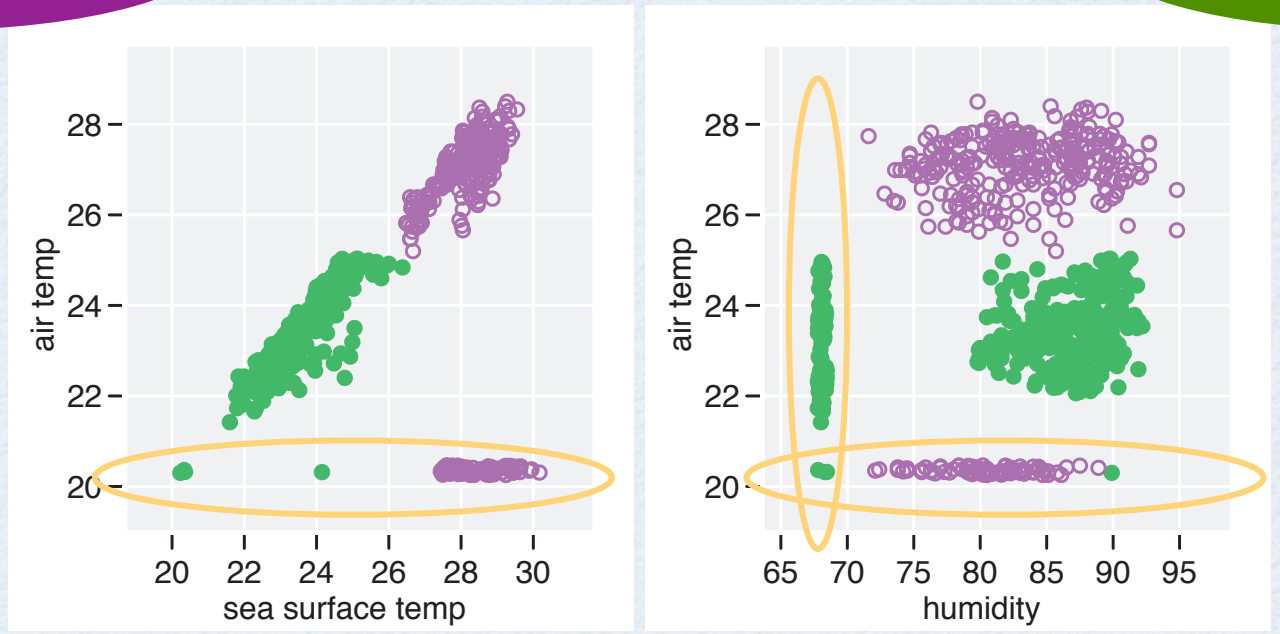
R package: norm

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

1997 El Nino

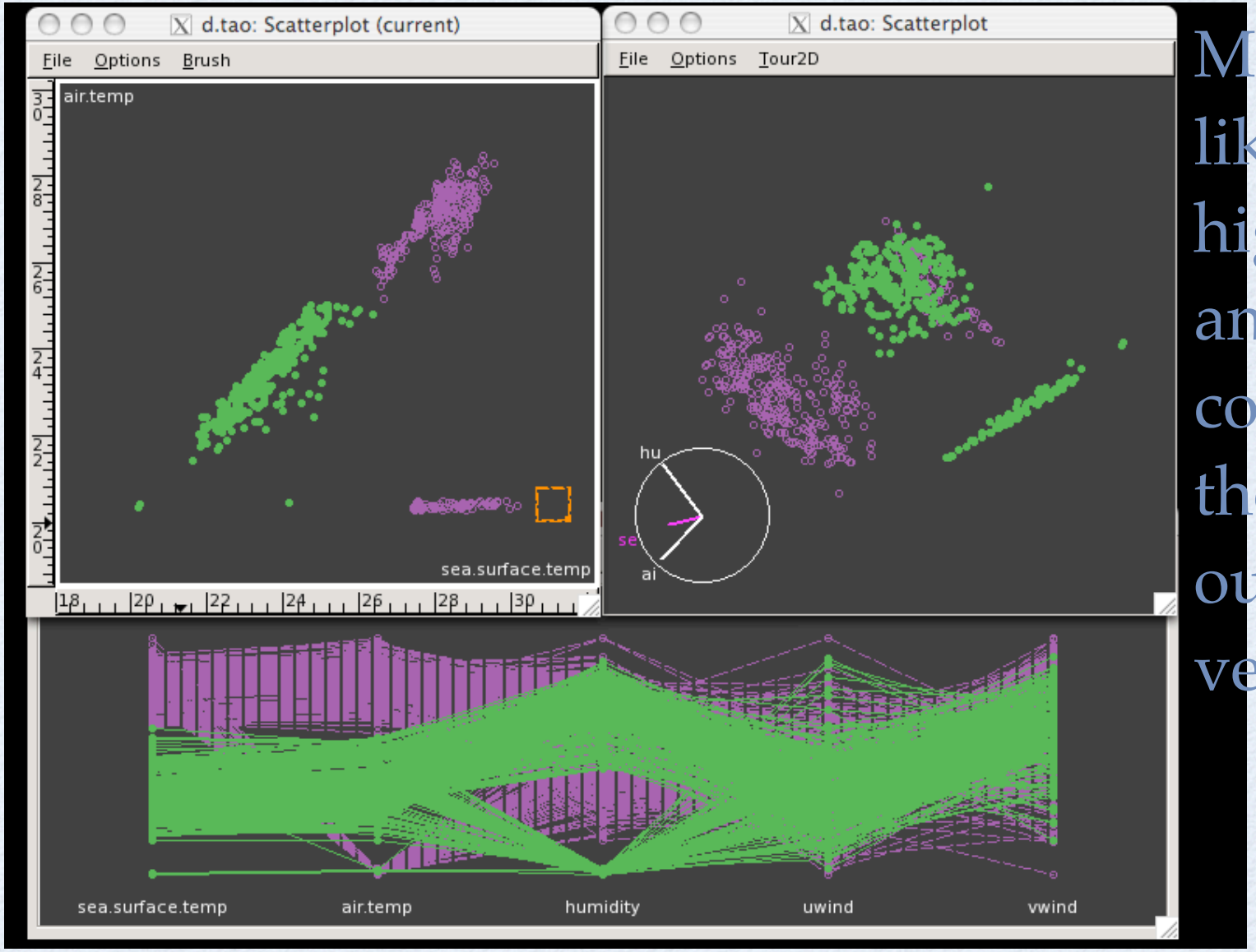
1993 Normal



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

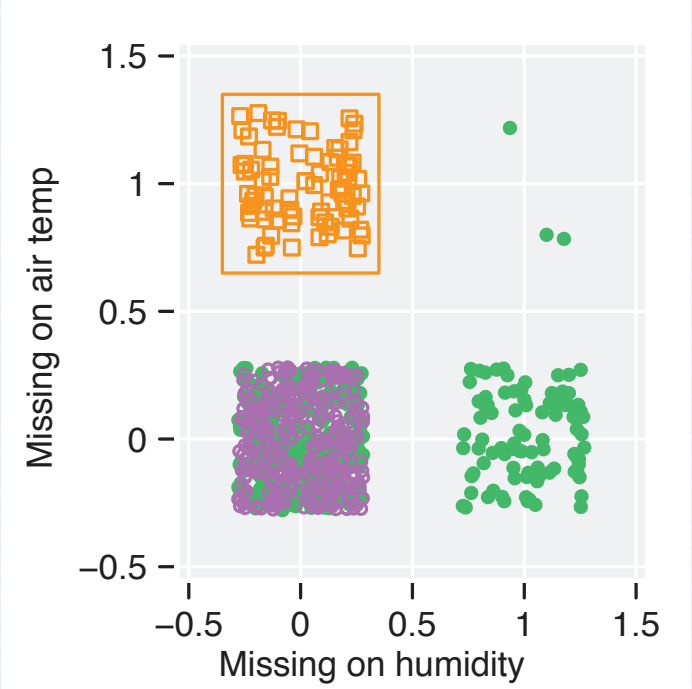
Missings on humidity only occur in 1997.

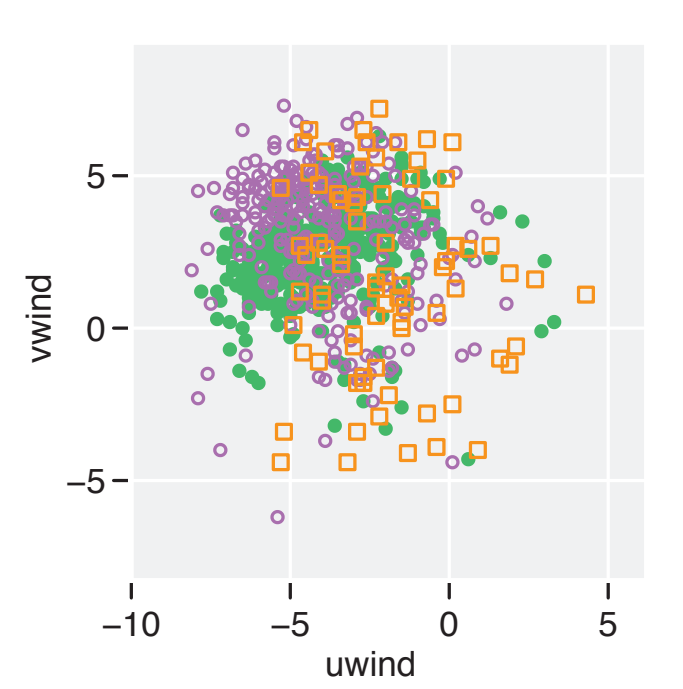
LIMITATION



Missings look like clusters in high-d plots, and in parallel coordinates they look like outliers at the very bottom.

Tracking Missings Using the Shadow Matrix





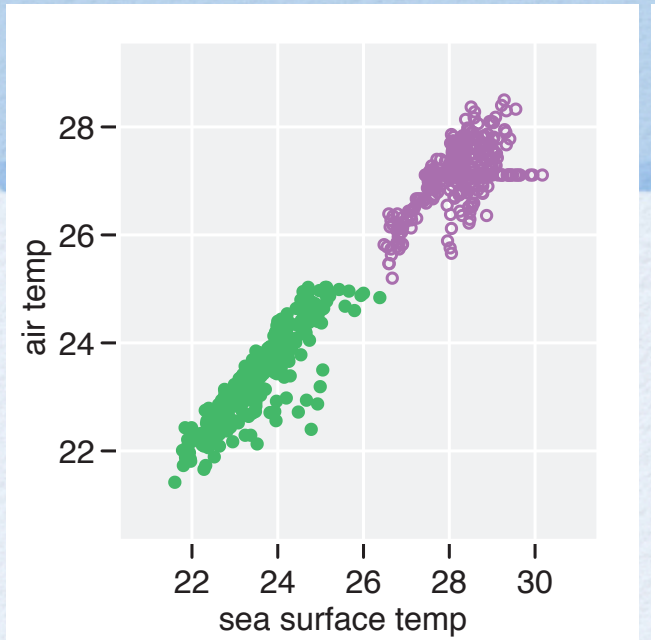
Missings on air temp have higher values on uwind than non-missings.

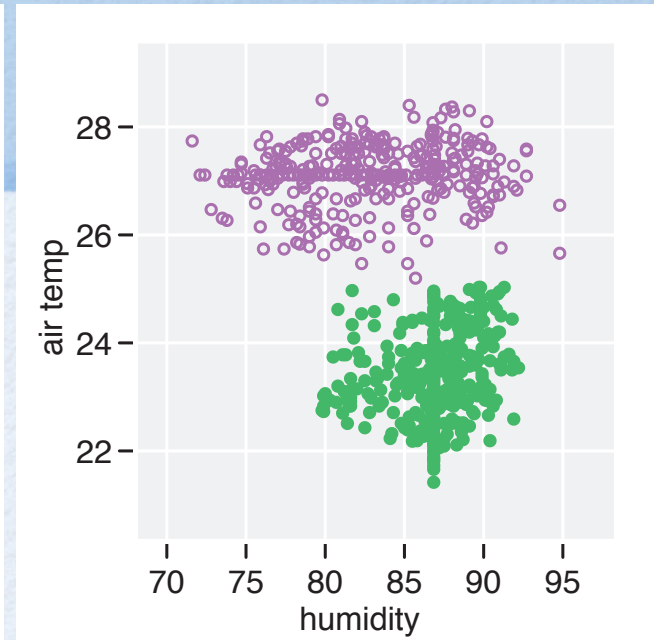
MISSING STRUCTURE

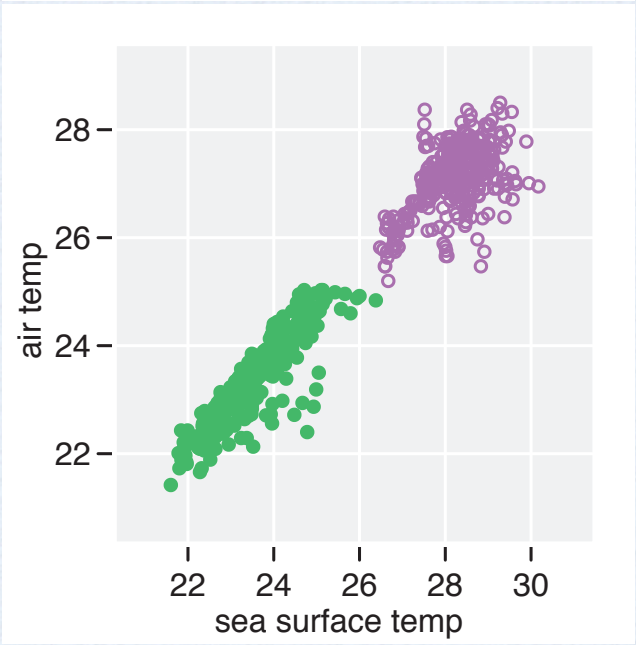
Missing values are NOT MCAR!

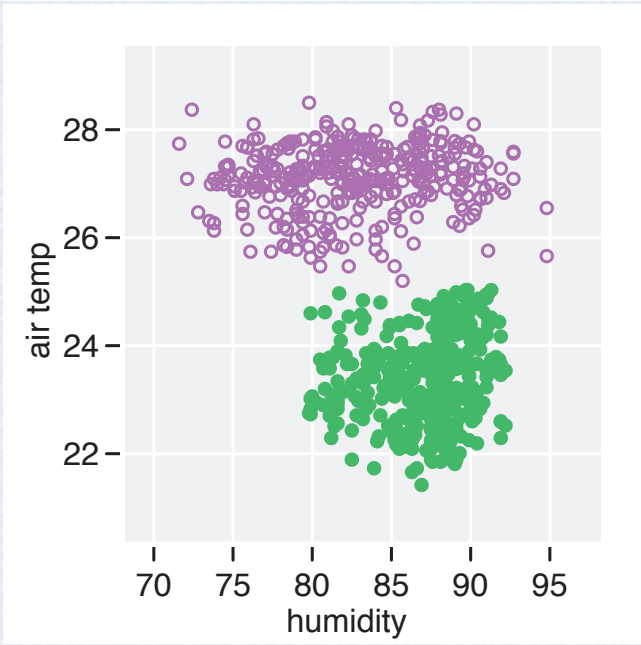
Imputation will need to use dependence of missing and not missing.

IMPUTING MISSINGS







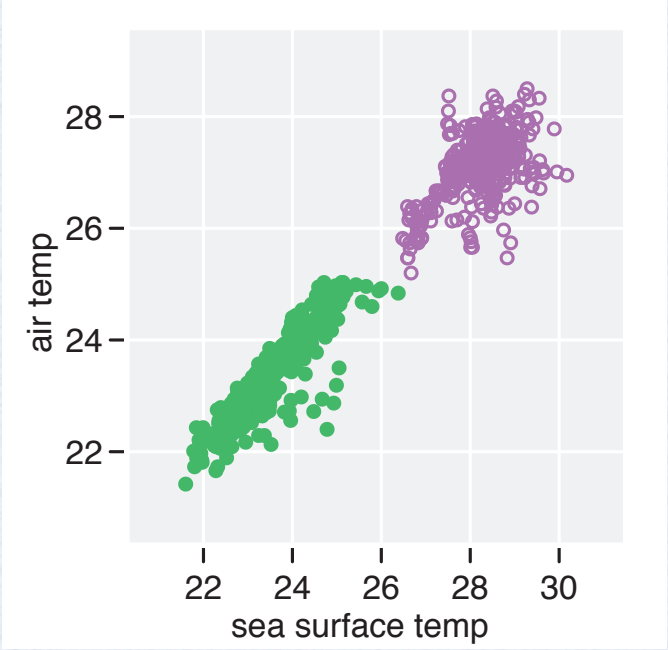


Means for each year

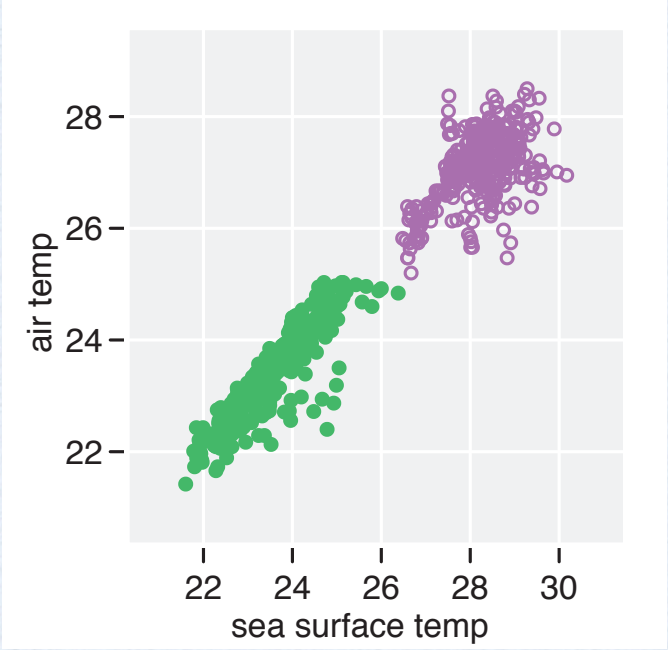
Random values from each year

What do you notice?

Using the Shadow Matrix

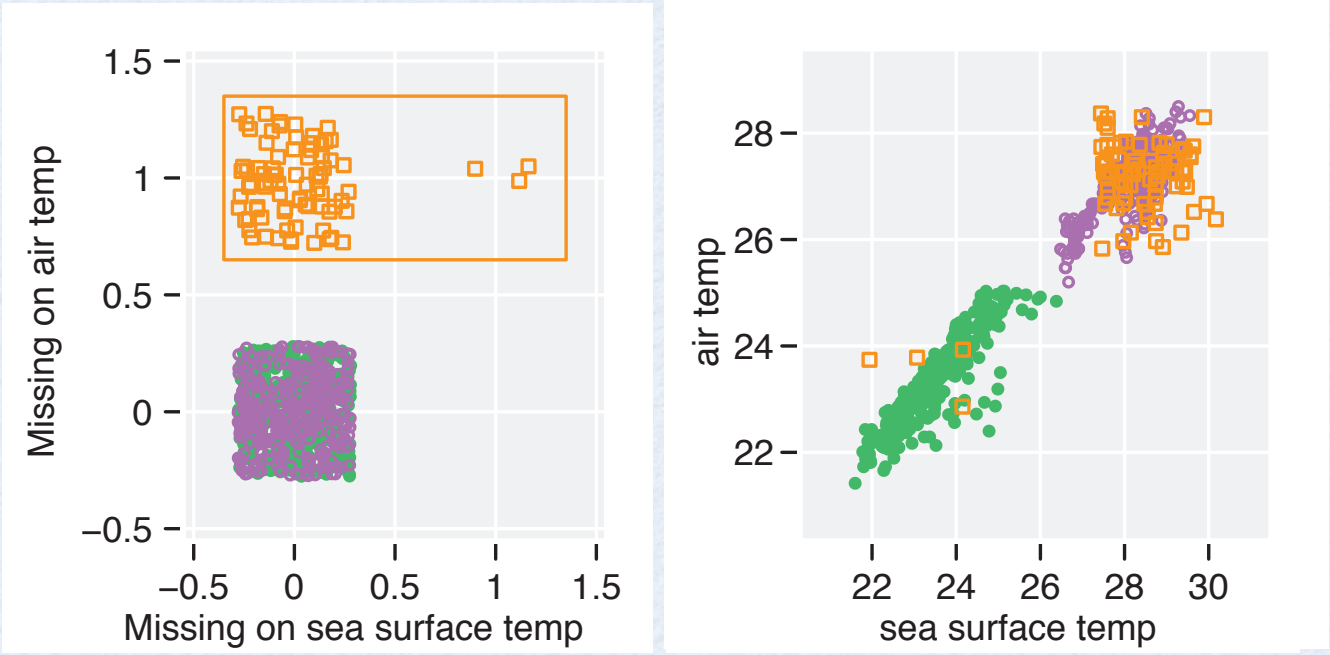


Using the shadow matrix



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

Using the Shadow Matrix



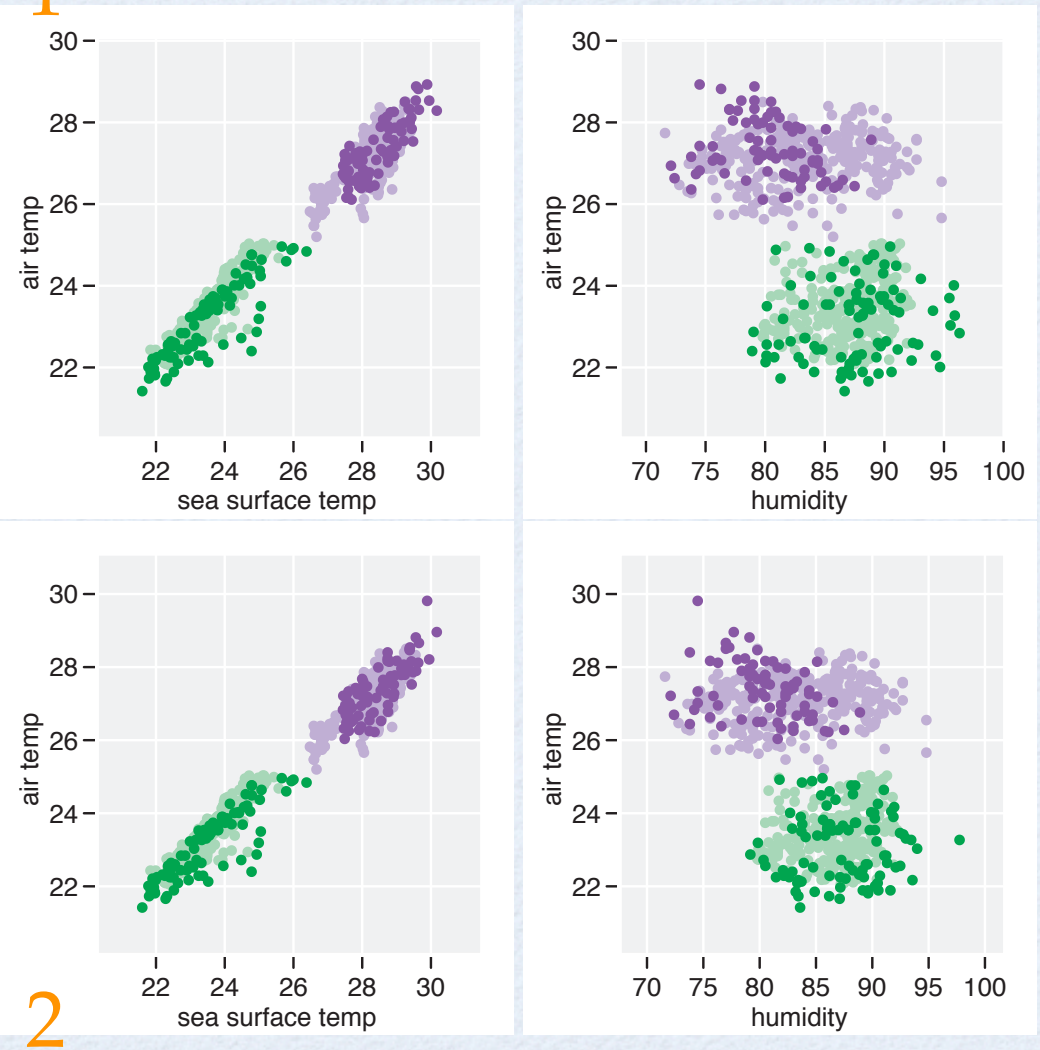
Imputed values which disappeared can be revealed by brushing on the shadow matrix.

OTHER APPROACHES

- **Model-based:** For eg, fit a regression model where the variable with missing values is the response, and all other variables are explanatory variables. Use the model to predict missing response values. Repeat for all variables with missings.
- **Nearest neighbors:** Find the closest cases to the case with a missing value, and average the values of these cases to impute the missing.

Multiple Imputation

1

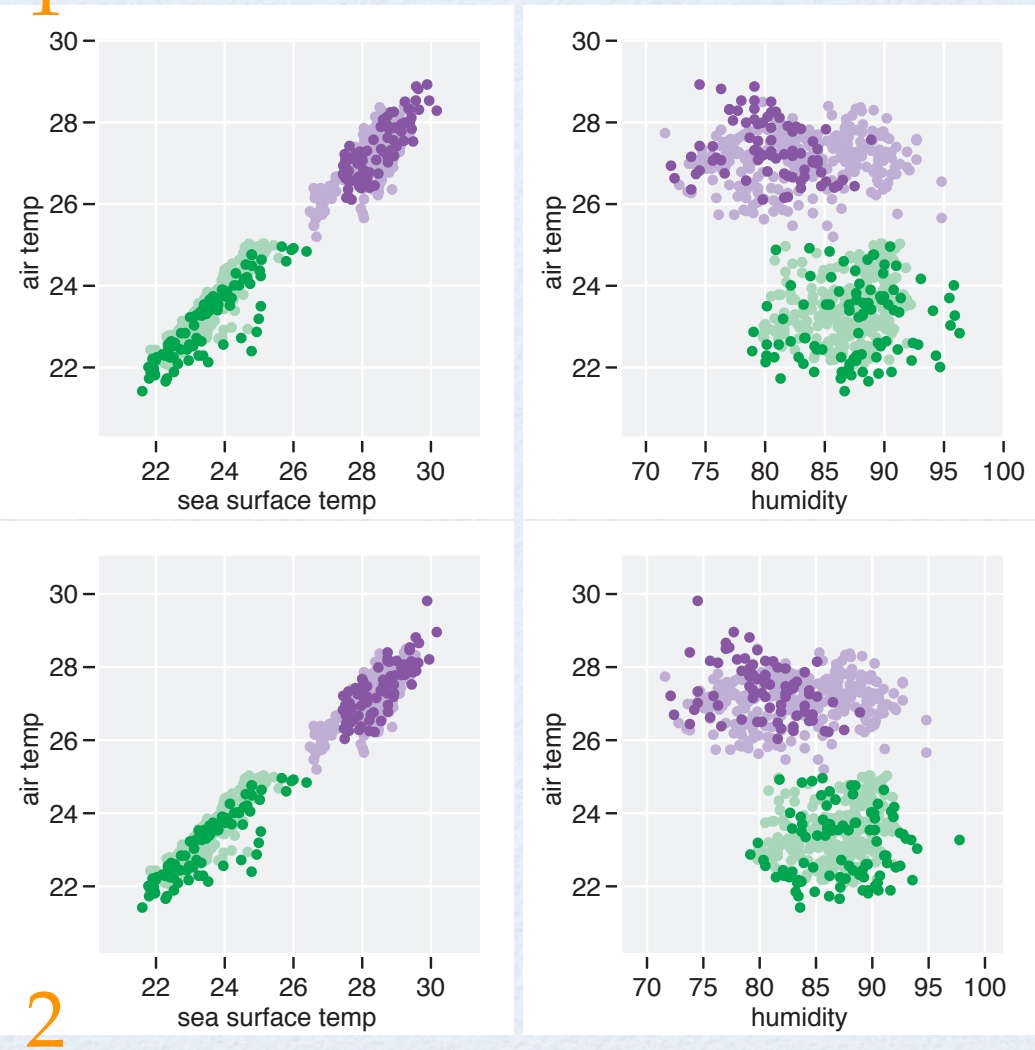


2

3, 4, 5, ...

Multiple Imputation

1



2

3, 4, 5, ...

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

SUMMARY

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