KU LEUVEN

fwo

Data preprocessing for (dyadic) longitudinal data

Jordan Revol
21/04/2022

Introduction

Why it matters?

- ≈80% of your time ➔ Organization saves time
- Insight on data quality:
  - Issues to encounter
  - Generalization issues
  - Model possibilities
  - Potential bias
  - Variability
  - etc.

- Failure to clean / repair = inaccurate results and models

Implications

- Pre-registration implications (see template from Kirtley et al., 2021)
- Its part of the process: don’t rush into statistical modeling
- Inaccurate results and models: wrong conclusions
- Data cleaning/preprocessing is partly model-dependant

  ➔ Know the model you will use
- Sensitive analysis is possible

The notebook: What can you find?

- Introduction to useful packages/functions
- Preprocessing methods
- Introduction to common issues (time interval assumption, night break, etc.)
- Visualization: insight on data quality and time series
- Examples of functions (copy/paste)
- Structure:
  - Block1: General matters
  - Block2: Time series
  - Block3: Dyadic specification
- Note: data uncleaned

Summary of the notebook

1. Introduction
2. Import data
3. Useful packages
   a) Bases functions
   b) tidyverse
   c) data.table
   d) Ggplot2
   e) Time: bases and lubridate
4. **Block1:** General matters
   a) Types of variables
   b) First insights
   c) Look at NA
   d) Specific cases
   e) Descriptive
   f) Response frequency and Compliance
5. **Block2:** Time matters
   a) Data and variables structures
   b) Handle date format
   c) Sampling matters (apps and participant)
   d) When missing beeps are not recorded
   e) Night break
   f) Compute scores
   g) Check variable (histogram, etc.)
   h) Time series visualization
   i) Time series issues
6. **Block3:** Dyadic specification
   a) Sampling matters (apps and participant)
   b) Response frequency and Compliance
   c) Compute score
   d) Time series visualization
   e) Final data management for models
7. Outro

Framework

- Be organized! Find your own way
- How I work, two files:
  - R Scripts: solves issues, general to specific ones
    1. Import data
    2. Clean data
    3. Export data
  - Notebook : explore, find issues and keep tracks
    1. Import data from R scripts
    2. Note issues / changes in a list
    3. Explore with functions and plots

Visual Studio Code

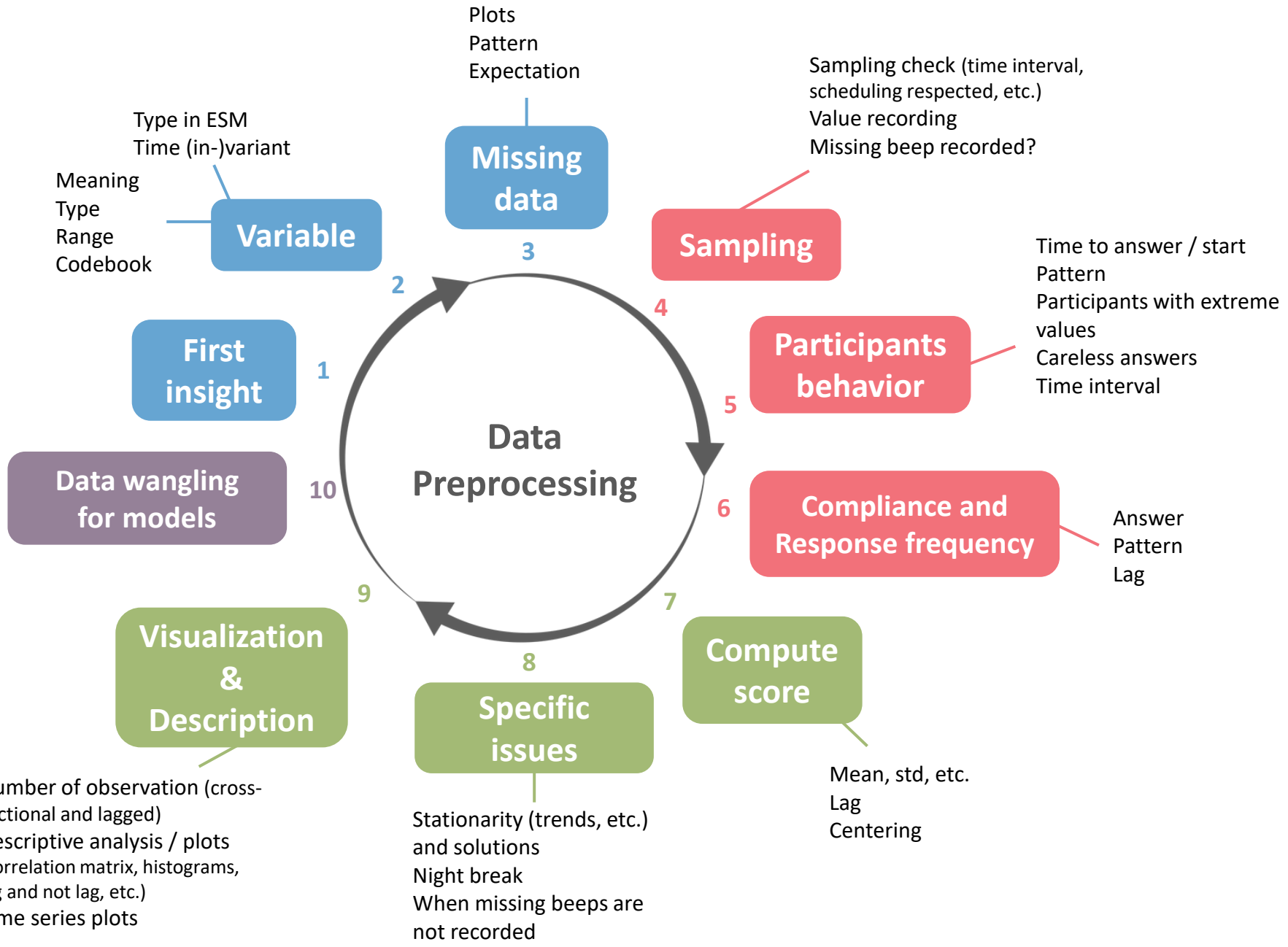
- Note issues / modifications / observations: in notebook or elsewhere
- Others: “mini-guide”, google (/duckduckgo) and cheat sheets

Useful packages / functions

- Base: ifelse(), which()
- Tidyverse packages
  - dplyr: select(), filter(), case_when(), group_by(), summarize(), join functions
  - tidyr: nest(), unite(), complete(), expand_grid(), gather(), spread()
  - stringr: manipulate characters
  - lubridate: manipulate dates
- data.table package
- Plots:
  - ggplot2
  - plotly: ggplotly()

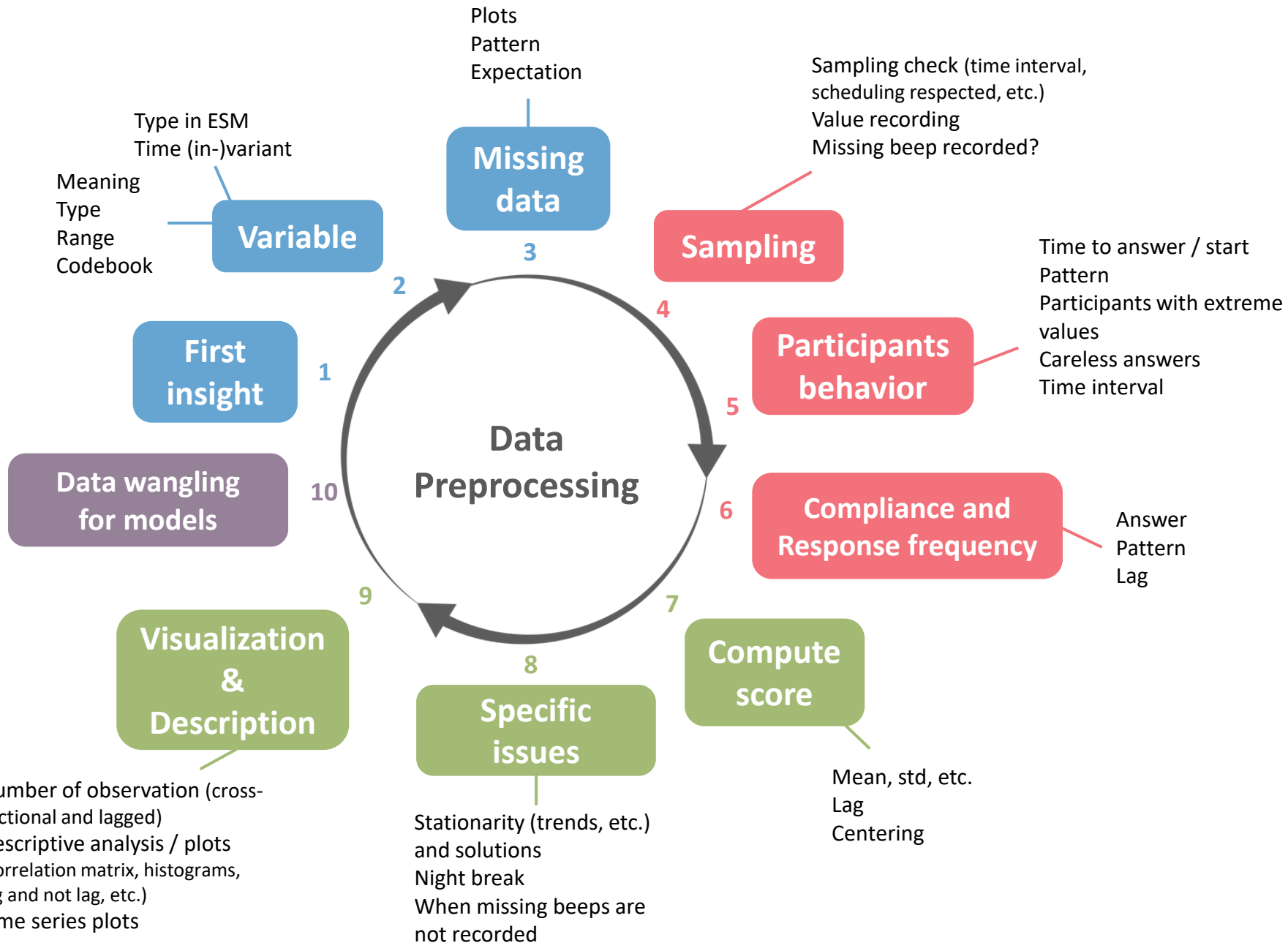
Schema

- Not linear
- Adapt:
  - Data structure (overall / pp / dyads)
  - Model-dependant
- Not exhaustive



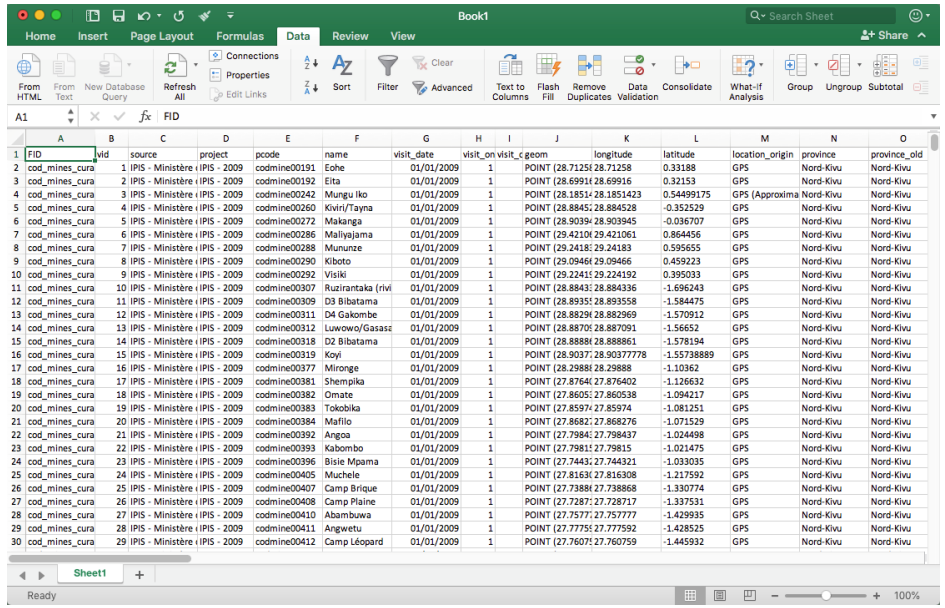
Preprocessing your data

More in the notebook


Data Preprocessing
1
First insight
2
Variable
Type in ESM
Time (in-)variant
Meaning
Type
Range
Codebook
3
Missing data
Plots
Pattern
Expectation
4
Sampling
Sampling check (time interval, scheduling respected, etc.)
Value recording
Missing beep recorded?
5
Participants behavior
Time to answer / start
Pattern
Participants with extreme values
Careless answers
Time interval
6
Compliance and Response frequency
Answer
Pattern
Lag
7
Compute score
Mean, std, etc.
Lag
Centering
8
Specific issues
Stationarity (trends, etc.) and solutions
Night break
When missing beeps are not recorded
9
Visualization & Description
Number of observation (cross-sectional and lagged)
Descriptive analysis / plots (correlation matrix, histograms, lag and not lag, etc.)
Time series plots
10
Data wangling for models

Importation and first insight

- LOOK AT YOUR DATA!
- str()
- head() and tail()
- Add some randomness: use sample()
- skim() from skimr package



Variables types

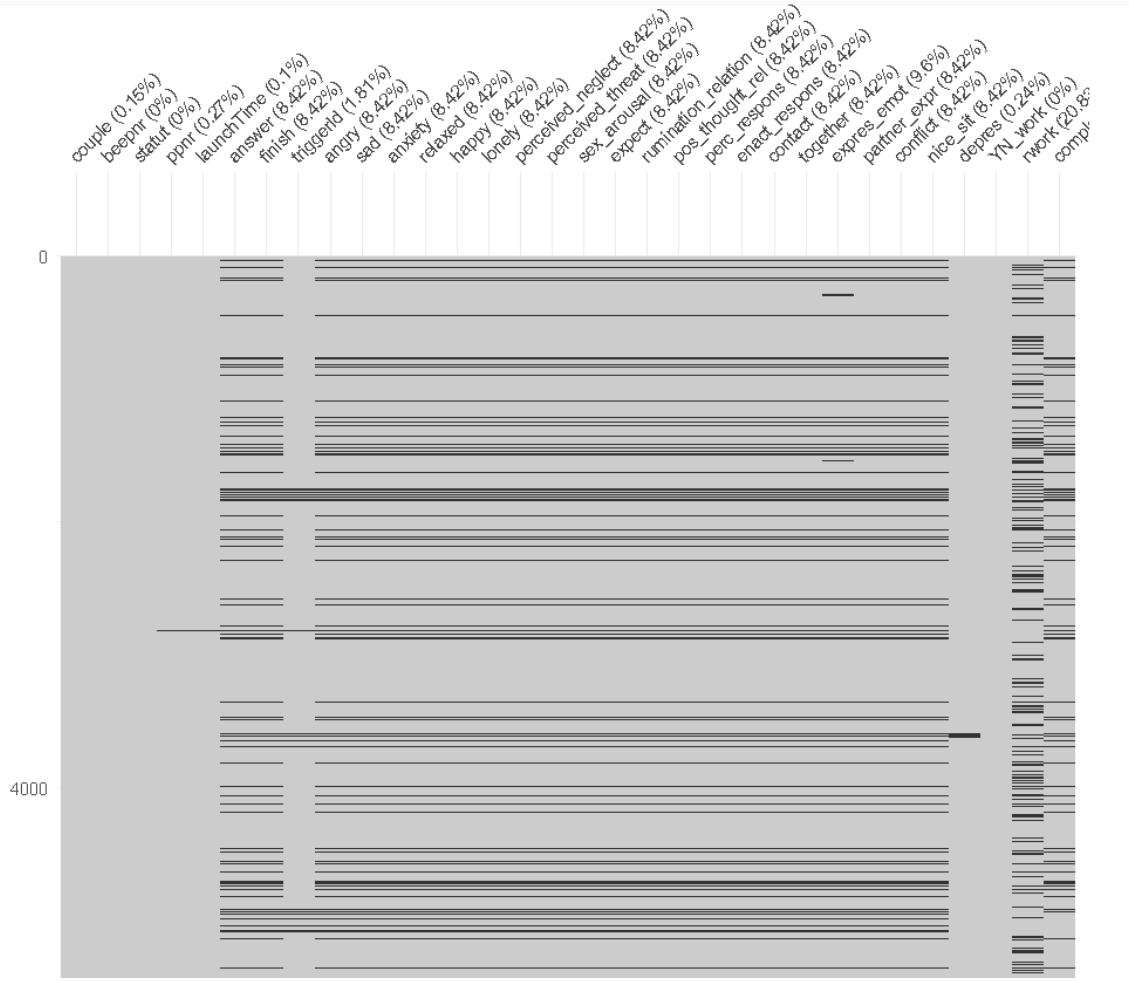
- Variables:
  - Meaning of the values
  - Type: character, integer, factor, etc.
  - Range and distribution
- Type of variable in ESM context:
  - **Subject identifier variables:** Participants numbers, Distinguishable / Undistinguishable dyads
  - **Design variables:** day/beep number, experimental condition, answer time
  - **Time-variant variables** (e.g., positive / negative affects)
  - **Time-invariant variables** (e.g., depression score)
- Expectation according to the type of data (e.g., missing values expectations)
- **Always take in account the structure of the data !**

Numb	Variable	Label	Type ESM	Measurement level	Range	Code	Items	Observation
1	couple	Couple id	Subject identifier	Nominal	1-300	-	-	
2	beepnb	Beep number	Design variable	Ordinal	1-100	-	-	-
3	dep	Depression score	Time-invariant	Continuous	0-50	-	Mean of items	Note recorded for some particpants
4	angry	Affect: angry	Time-varying	Continuous	0-100	-	"angry"	
5	contact	Couple contact	Time-varying	Ordinal	0-3	0: no contact 1: less than 1h 2: between 1h and 6h 3: spend the day	"Rate the time spend with your partner today	
...	...	...	...	...	...	...	...	...

Codebook is important!

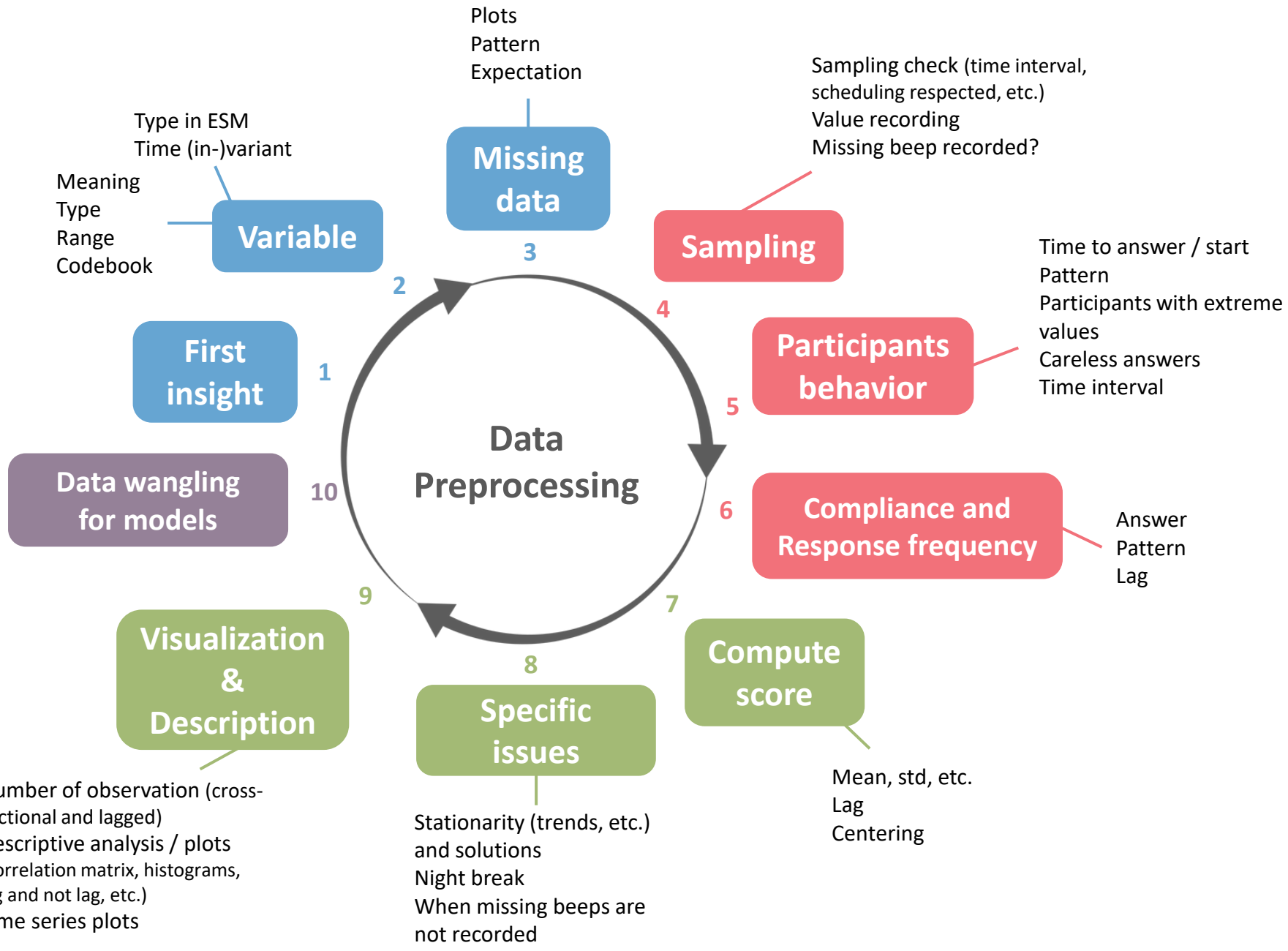
Introduction to missing data

- Hidden missing data: -999, 0, etc.
- Type of missing data
  - Missing completely at random (MCAR)
  - Missing at random (MAR)
  - Missing not at random (MNAR)
- Look for patterns / inconsistency
- Expectations in function of the type of variable
- Usefull packages: *visdat, naniar*



Short introduction to time

- Useful package: lubridate (see cheat sheet)
- Formats: Date, POSIXct, POSIXlt
- Time algebra (+, -)
- Be careful:
  - Origin
  - Time zone
  - Loosing format after transformation
- How to handle date format?
  1. Spread date (year, month, day, hour, etc.)
  2. Beeps number
  3. Continuous time variable


Data Preprocessing
1
First insight
2
Variable
Type in ESM
Time (in-)variant
Meaning
Type
Range
Codebook
3
Missing data
Plots
Pattern
Expectation
4
Sampling
Sampling check (time interval, scheduling respected, etc.)
Value recording
Missing beep recorded?
5
Participants behavior
Time to answer / start
Pattern
Participants with extreme values
Careless answers
Time interval
6
Compliance and Response frequency
Answer
Pattern
Lag
7
Compute score
Mean, std, etc.
Lag
Centering
8
Specific issues
Stationarity (trends, etc.) and solutions
Night break
When missing beeps are not recorded
9
Visualization & Description
Number of observation (cross-sectional and lagged)
Descriptive analysis / plots (correlation matrix, histograms, lag and not lag, etc.)
Time series plots
10
Data wangling for models

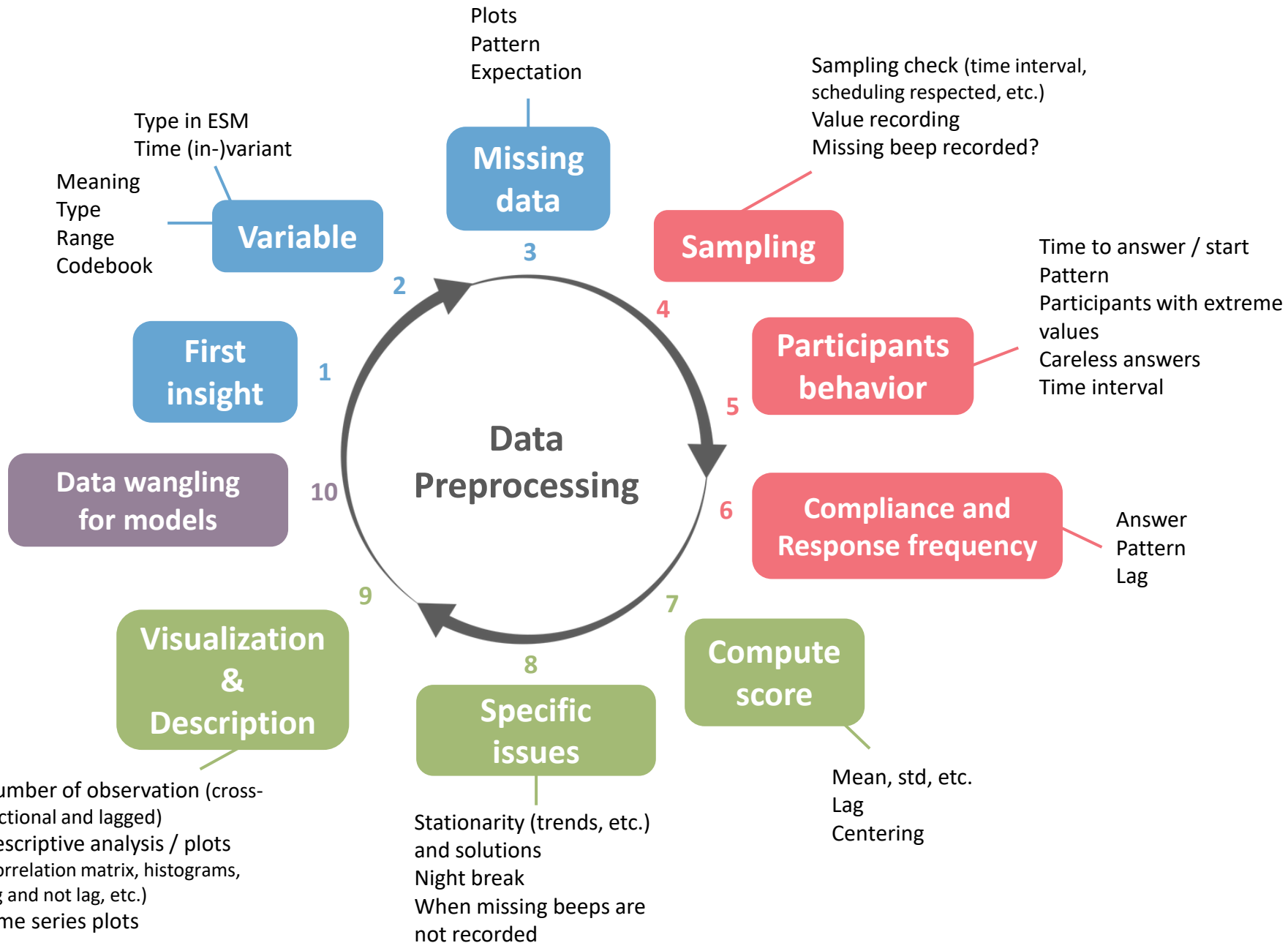
Response frequency and Compliance

- Definition
- Complete observation: to decide
- Multiple possibilities:
  - Overall
  - Per participant
  - Per dyad
  - Taking in account lag
- Questions: Enough data? Unbalanced between conditions? Some participants under/over-represented?

Visualization: design and participants time variables

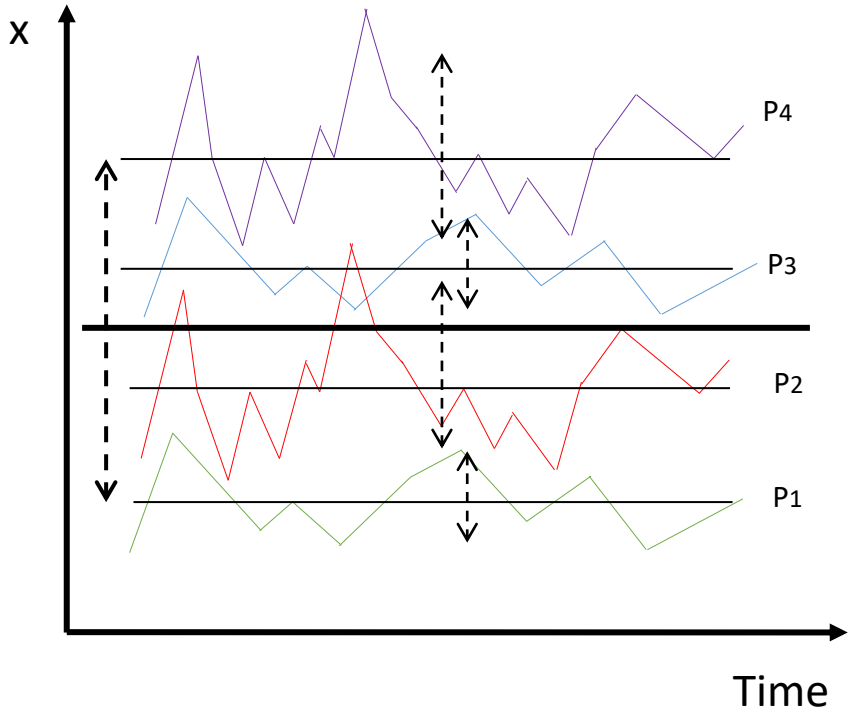
- Why is it important?
  - Checking possible issues
  - Descriptive insights about sampling
  - Future studies: hypothesis generation, methods, expectations
- Multiple point of view:
  Overall / per participant / per dyads
- Must be adapted to your study
- ggplot2 package

- **Insight on data collection procedure:**
  - Compare to expected sampling
  - How many beeps per participant
  - Time of beeps (filled / missed)
- **Insight on participant's behaviors:**
  - Time pp answer
  - Beeps that are missed
  - Time to launch survey
  - Time spend to answer
  - Time intervals between 2 answers
  - Respect sampling
  - Time interval between partner's beeps, answer and finished


Data Preprocessing
1
First insight
2
Variable
Type in ESM
Time (in-)variant
Meaning
Type
Range
Codebook
3
Missing data
Plots
Pattern
Expectation
4
Sampling
Sampling check (time interval, scheduling respected, etc.)
Value recording
Missing beep recorded?
5
Participants behavior
Time to answer / start
Pattern
Participants with extreme values
Careless answers
Time interval
6
Compliance and Response frequency
Answer
Pattern
Lag
7
Compute score
Mean, std, etc.
Lag
Centering
8
Specific issues
Stationarity (trends, etc.) and solutions
Night break
When missing beeps are not recorded
9
Visualization & Description
Number of observation (cross-sectional and lagged)
Descriptive analysis / plots (correlation matrix, histograms, lag and not lag, etc.)
Time series plots
10
Data wangling for models

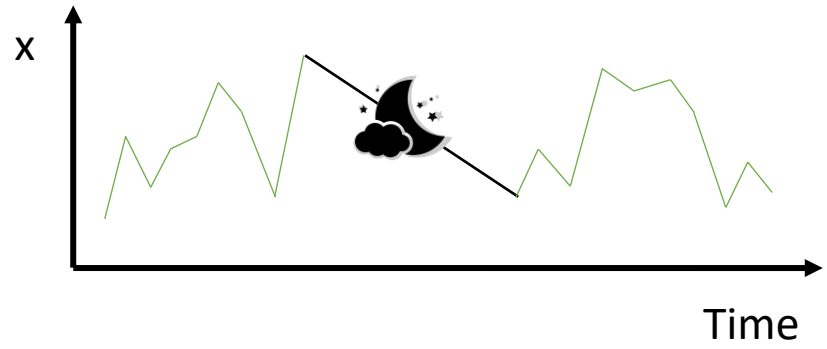
Compute Scores

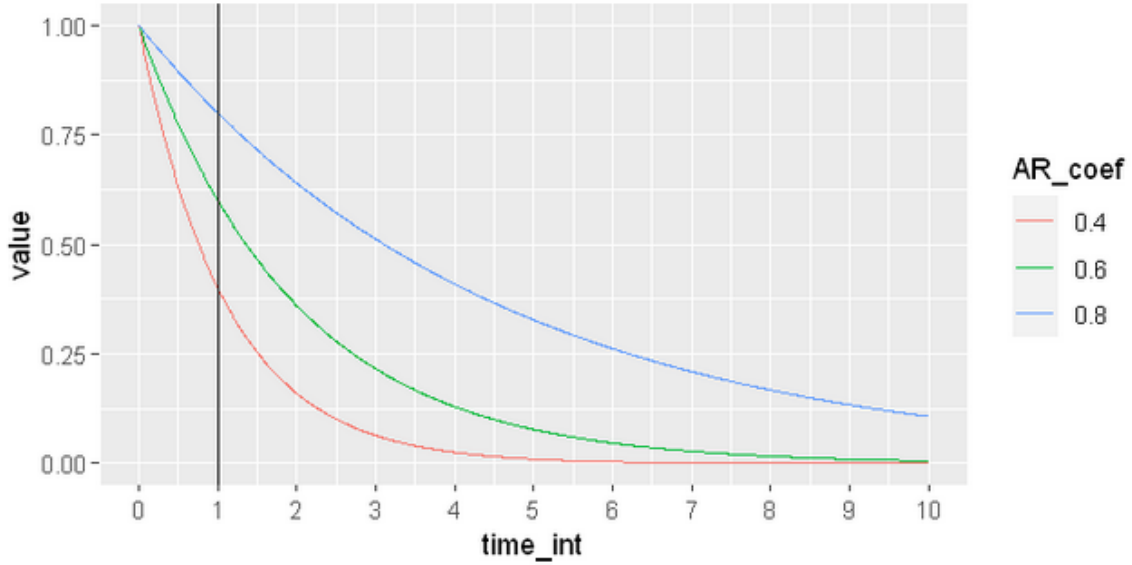
- Using summarize() and mutate() with group_by()
- Mean, standard deviation, etc.
- Centering:
  - Grand-mean centering
  - Person-mean centering
- Lag variable



Night break

- Time interval assumptions in discrete time models





Night break

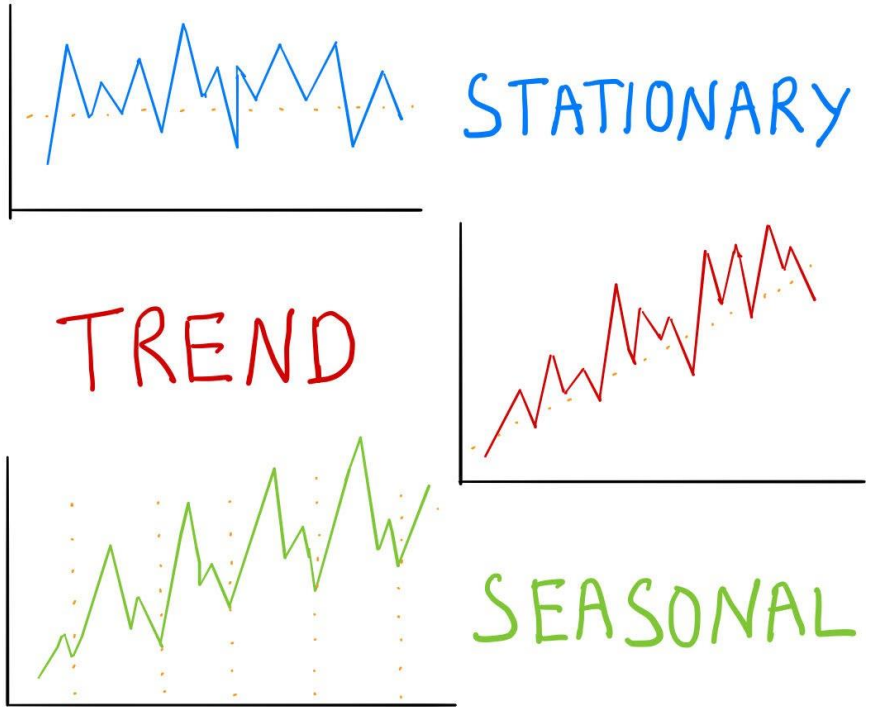
- Time interval assumptions in discrete time models
- Solutions in function of the model / software / modeling approach
- Solutions:
  1. Add NAs
  2. Add a dummy variable (end of the day)
  3. NA imputation
     ➔ Not recommended

Day	Hour	beepnr	Y	Y_lag
1	10	1	30	NA
1	15	2	34	30
1	20	3	59	34
1	24	4	NA	59
2	10	5	13	NA

Day	Hour	beepnr	Y	Y_lag	Night
1	10	1	30	NA	0
1	15	2	34	30	0
1	20	3	59	34	0
2	10	5	13	59	1

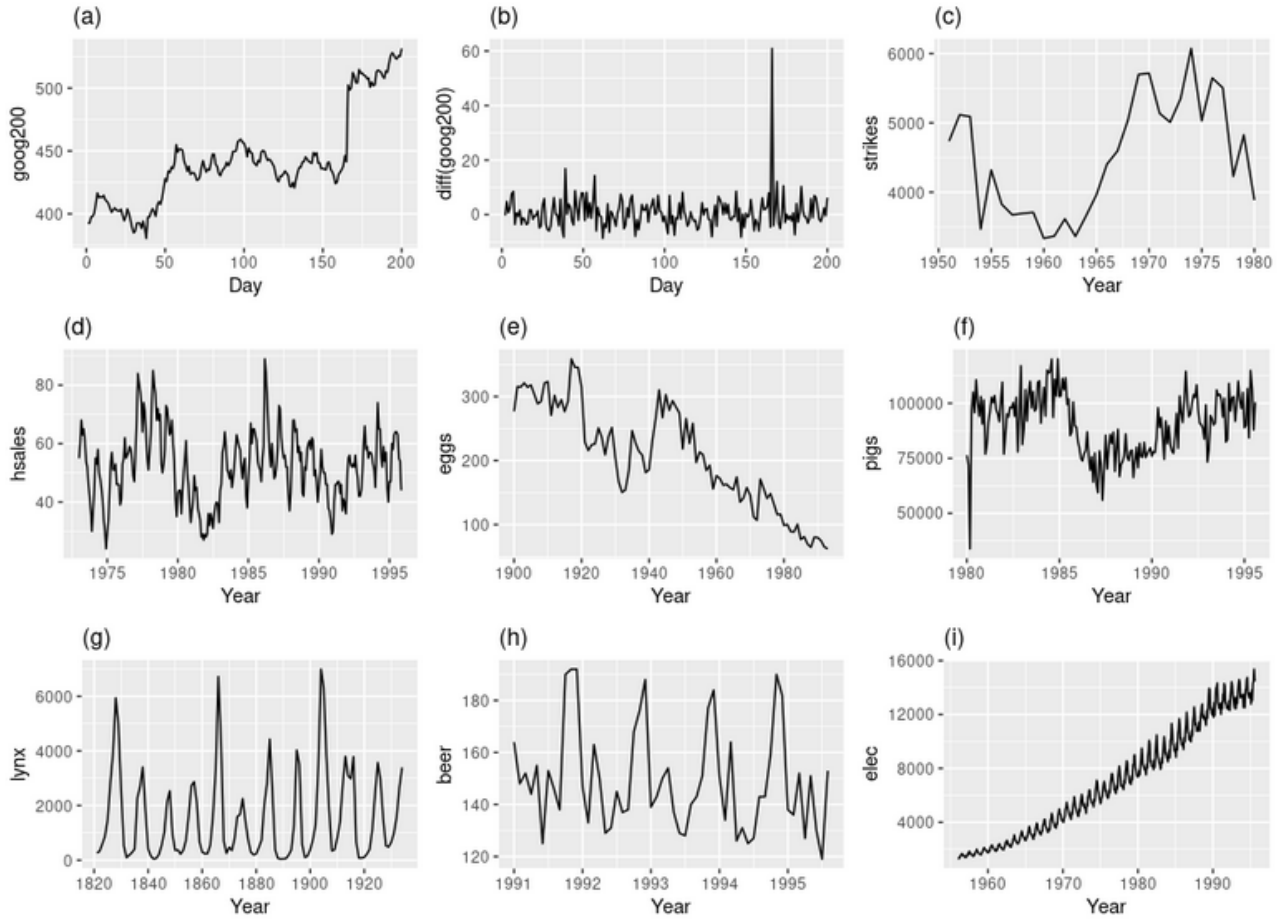
Visualization: Time series

- Structure of the data
- Stationary: trend inspection
- ggplotly() function (plotly package)

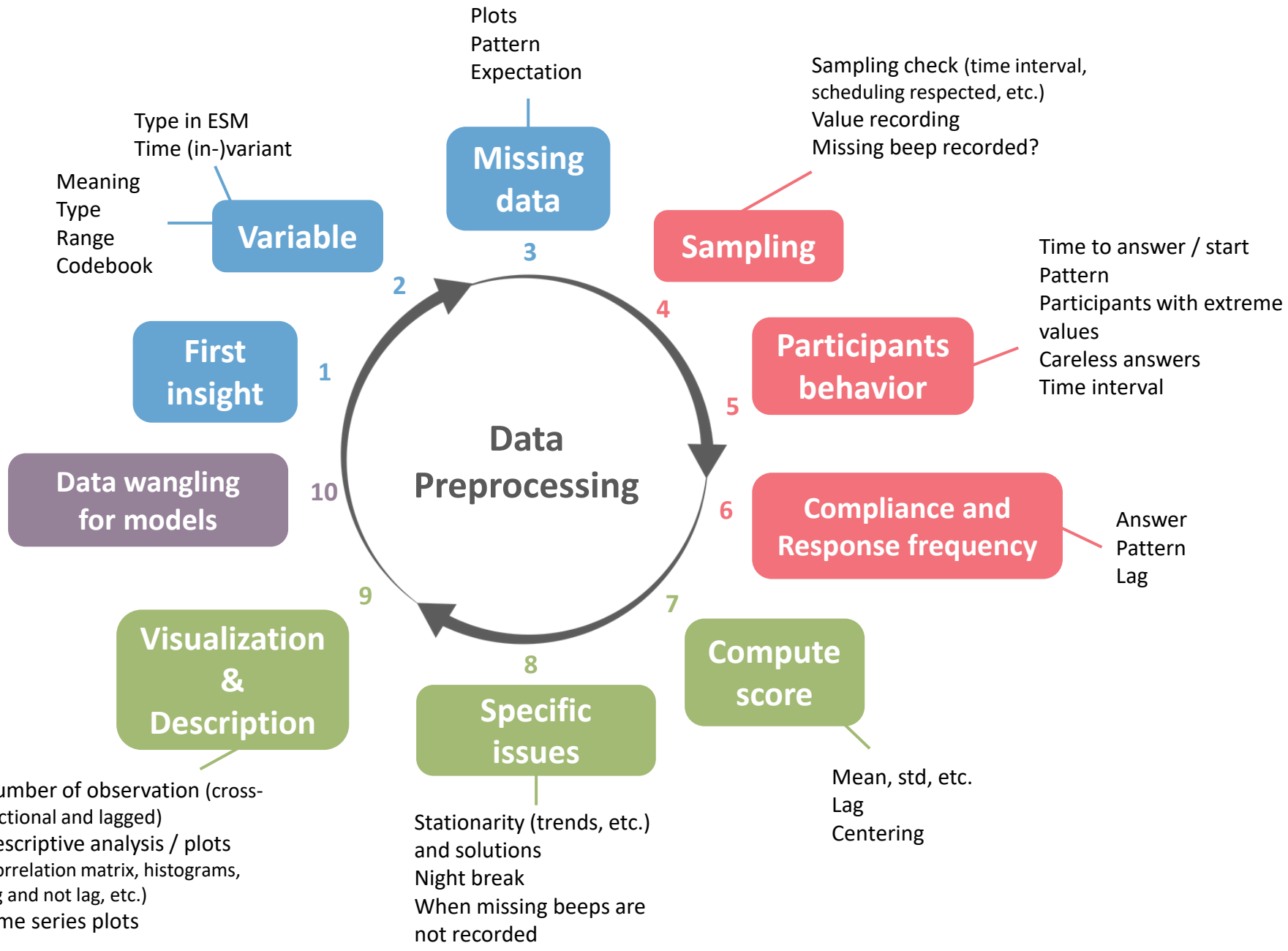


Visualization: Time series

- Structure of the data
- Stationary: trend inspection
- ggplotly() function
  (plotly package)

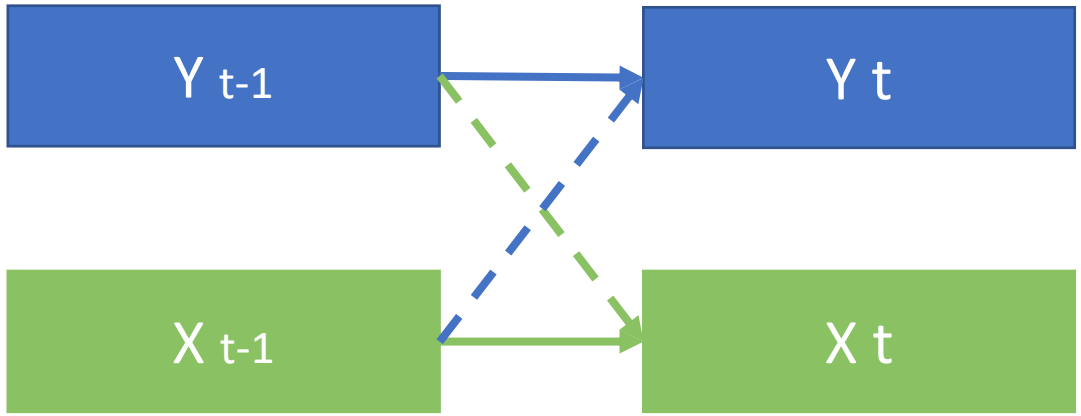


From Hyndman & Athanasopoulos, 2018


Data Preprocessing
1
First insight
2
Variable
Type in ESM
Time (in-)variant
Meaning
Type
Range
Codebook
3
Missing data
Plots
Pattern
Expectation
4
Sampling
Sampling check (time interval, scheduling respected, etc.)
Value recording
Missing beep recorded?
5
Participants behavior
Time to answer / start
Pattern
Participants with extreme values
Careless answers
Time interval
6
Compliance and Response frequency
Answer
Pattern
Lag
7
Compute score
Mean, std, etc.
Lag
Centering
8
Specific issues
Stationarity (trends, etc.) and solutions
Night break
When missing beeps are not recorded
9
Visualization & Description
Number of observation (cross-sectional and lagged)
Descriptive analysis / plots (correlation matrix, histograms, lag and not lag, etc.)
Time series plots
10
Data wangling for models

Model specification

- Have in mind the model to set up your dataset.
- Data format (e.g., wide, long) is model/software/package dependant
- For instance:
  - Using nlme package in R
  - L-APIM, a linear mixed-model
  - Long format with partner's value as predictor



dyad	pp	own	part	own_lag	part_lag

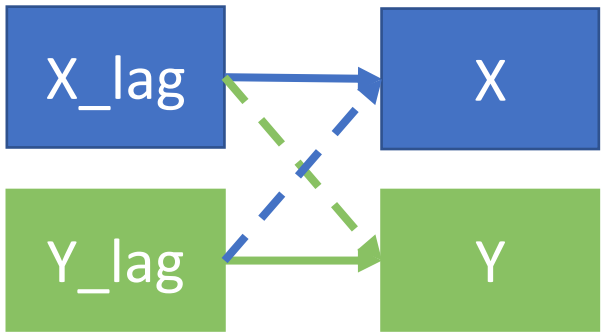
Final Data Management

1. Wide format with duplicated partner's rows side by side

dyad	time	pp	statut	Var1	…	Var1_lag	pp.part	var1.part	…	var1_lag.part
1	3	10	M	4	…	2	11	5	…	6
1	3	11	F	5	…	6	10	4	…	2
…	…	…	…	…	…	…	…	…	…	…

2. Select variables for the model

dyad	time	pp	statut	X	Y	X_lag	Y_lag



Previous vs. new results

```
Random effects:
 Formula: ~-1 + male + female | couple
 Structure: General positive-definite, Log-Cholesky parametri
         StdDev   Corr
male     76.42864 male
female   81.54977 0.988
Residual 50.62911

Correlation Structure: Compound symmetry
 Formula: ~1 | couple/beepnr
 Parameter estimate(s):
       Rho
-0.07160316
Variance function:
 Structure: Different standard deviations per stratum
 Formula: ~1 | statut
 Parameter estimates:
       f        m
1.000000 0.838761
Fixed effects: Y ~ -1 + male + male:Y_lag_pc + male:X_lag_pc
                   Value Std.Error   DF   t-value p-value
male             1.70047  7.699448 9648  0.220856  0.8252
female          56.16722  8.224939 9648  6.828892  0.0000
male:Y_lag_pc    0.30474  0.013564 9648 22.466586  0.0000
male:X_lag_pc    0.29583  0.012053 9648 24.543381  0.0000
Y_lag_pc:female -0.13527  0.014370 9648 -9.413613  0.0000
X_lag_pc:female  0.01891  0.016173 9648  1.168940  0.2425
```

```
Random effects:
 Formula: ~-1 + male + female | couple
 Structure: General positive-definite, Log-Cholesky parametrization
         StdDev    Corr
male      6.783301 male
female   11.975776 -0.275
Residual 16.255577

Correlation Structure: Compound symmetry
 Formula: ~1 | couple/beepnr
 Parameter estimate(s):
       Rho
-0.1227169
Variance function:
 Structure: Different standard deviations per stratum
 Formula: ~1 | statut
 Parameter estimates:
       f        m
1.0000000 0.7095191
Fixed effects: Y ~ -1 + male + male:Y_lag_pc + male:X_lag_pc + female
                   Value Std.Error   DF  t-value p-value
male             7.77471 0.7194548 9580 10.80639       0
female          62.80546 1.2576094 9580 49.94036       0
male:Y_lag_pc    0.26185 0.0141265 9580 18.53594       0
male:X_lag_pc   -0.04523 0.0096137 9580 -4.70432       0
Y_lag_pc:female  0.36777 0.0135852 9580 27.07156       0
X_lag_pc:female -0.10286 0.0199764 9580 -5.14903       0
```

Conclusion

Tips and Rules

- Look at your data
- Be careful of already preprocessed data (not done by yourself)
- Don't rush into statistical modeling
- Double check issues found
- Comment your code (for yourself and others)
- Have a backup & Never touch raw data
- Keep all the information but adapt the database to your statistical analysis

Conclusion

- Look at the notebook
- Cleaning and data quality inspection are IMPORTANT and TAKES TIME!
- Looks like an investigation: clues, suspects, crimes, etc.
- Visualization helps

Literature

- Viechtbauer, W. (2021). Chapter 8: Structuring, Checking, and Preparing the Data. In Dejonckheere, E., & Erbas, Y. (2021). *The open handbook of experience sampling methodology: A step-by-step guide to designing, conducting, and analyzing ESM studies (2nd ed.)*
- Kirtley, O. J., Lafit, G., Achterhof, R., Hiekkaranta, A. P., & Myin-Germeys, I. (2021). Making the Black Box Transparent: A Template and Tutorial for Registration of Studies Using Experience-Sampling Methods. *Advances in Methods and Practices in Psychological Science.* https://doi.org/10.1177/2515245920924686