Longitudinal data analysis in R: General introduction

Filippo Biscarini

(Biostatistician, bioinformatician and quantitative geneticist) **CNR**, Milan (Italy)

The instructors

- Dr. Filippo Biscarini (CNR, Milan - Italy)
- Dr. Andreia J. Amaral (UÉvora, Évora - Portugal)

Filippo in one slide

- **Roma** (*born*)
- **Perugia** (*MSc degree*)
- **Cork, ICBF** (*Web-design & Database*)
- **Cremona, ANAFI** (*Quantitative Genetics*)
- **Guelph, CGIL** (*Visiting Scientist*)
- **Wageningen, WUR** (*PhD*)
- **Göttingen University** (*post-doctoral researcher*)
- **Lodi, PTP** (*'omics in animals, plants, humans*)
- **Milan - CNR** (*tenured researcher*)
- **Cardiff University** (*biostatistician*)
- **Milan - CNR** (*senior researcher*)
- **Bruxelles - ERC** (*seconded national expert*)
- **Milan - CNR** (*senior researcher*)

Andreia in one slide

- **Cascais, Portugal** (*born*)
- **Technical University, Lisbon** (*BSc degree (5 years)*)
- **INGA (***Common Agriculture Policy*)
- **ULisbon - Faculty of Sciences (***MSc degree*)
- **Wageningen, WUR** (*PhD*)
- **ULisbon - Faculty of Medicine** (*post-doctoral researcher*)
- **ULisbon - Faculty of Veterinary Medicine (**senior researcher**)**
- **UÉvora**(*Professor Animal Breeding and Genetics*)

Overview of the course

Day 1

- Longitudinal data: examples and challenges
  - Lab 1: First encounter with longitudinal data
- The basic experimental setting: treatments and timepoints
  - Lab 2: Treatments and timepoints in R
- The classical statistical perspective
  - Lab 3: Models to analyse data with repeated records over time (multiple time points) and space (multiple locations) in R
- Lecture 4: Difference-in-differences (diff-in-diff)
  - Lab 4: diff-in-diff in R

Overview of the course

Day 2

- Censored data and survival analysis
  - Lab 5: Survival analysis in R
- Cross-validation: simple and with spatial, temporal (or other) data structure
  - Lab 6: Cross-validation strategies in R
- Time series, autocorrelations and forecasting
  - Lab 7: Time series and forecasting in R

Overview of the course

Day 3

- Linear Mixed Models - Introduction
  - Fitting mixed model in R
  - Graphical representation of the model
    - Lab 8: Linear Mixed Models in R
  - Time and group as random effects
    - Lab 9: Testing for the effects of variables in R
    - Lab 10: Group effect and Interaction between time and group in R
    - Lab 11: Parametric curves and prediction of random effects in R

Overview of the course

Day 4

- Model diagnostics: a primer
    - Lab 12: Strategies for model diagnostics in R
- Generalized Estimating Equations
    - Lab 13 : Within-group correlation structure in R
- Generalized linear mixed-effects models
    - Lab 14: Discrete versus continuous data
- Epidemiological modelling of infectious diseases
    - Lab15: Temporal analysis
      Spatial-temporal analysis
      Detection of outbreaks

Overview of the course

timetable

repo

final quiz on Thursday

breaks: **long break at 16:30/17:00 (30 min.)**, each day (shorter breaks in between on a case-per-case basis)

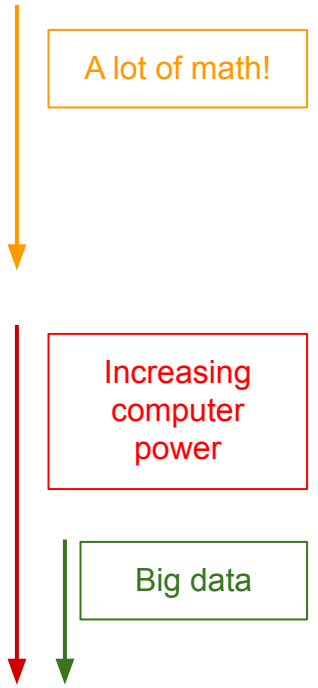
1st edition of this course

Caveat: at times a little skewed towards

- genetics / breeding
- predictions

It's been a long way to modern statistics

- 1870's: Francis Galton: linear regression
- ~1900: Karl Pearson: correlation
- 1925: Ronald Fisher's "*Statistical Methods for Research Workers*" (he later regretted the 0.05 p-value threshold) ⟶ **frequentist statistics**
- **Bayesian** resurgence: 1980s ⟶ **MCMC** (1986: Gibbs sampling by Geman & Geman)
- Non-parametric statistics & resampling methods
- The **statistical** (machine) **learning** paradigm



Types of longitudinal data

1. treatments and timepoints
2. repeated records
3. censored date
4. time series data

Types of longitudinal data

1. treatments and timepoints
2. repeated records
3. censored data
4. time series data

Q: can you think of other types of longitudinal data?

A few examples from literature

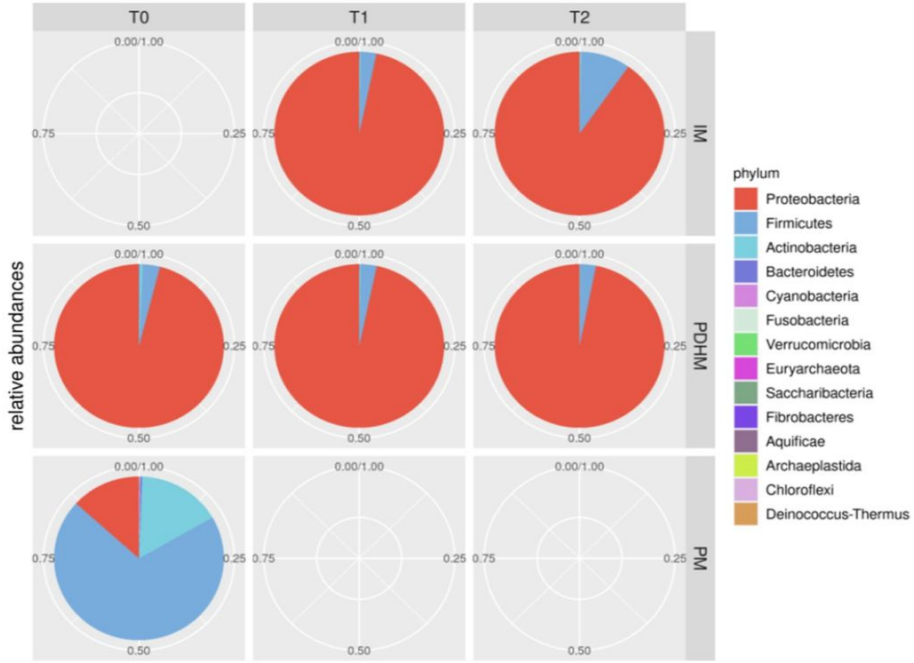


Fig. 3 Pie-charts of phylum relative abundances in the PDHM, PM and IM samples. Pie charts showing the distribution of the dominant bacterial phyla in the PDHM, PM and IM samples. The numbers around the pie-charts indicate the percentage of abundance. *PM* preterm milk samples, *PDHM* pasteurized donor human milk samples, *IM* inoculated milk samples. T0: baseline (before inoculum); T1: 2 h after inoculation; T2: 4 h after inoculation

From Mallardi et al. 2021

A few examples from literature

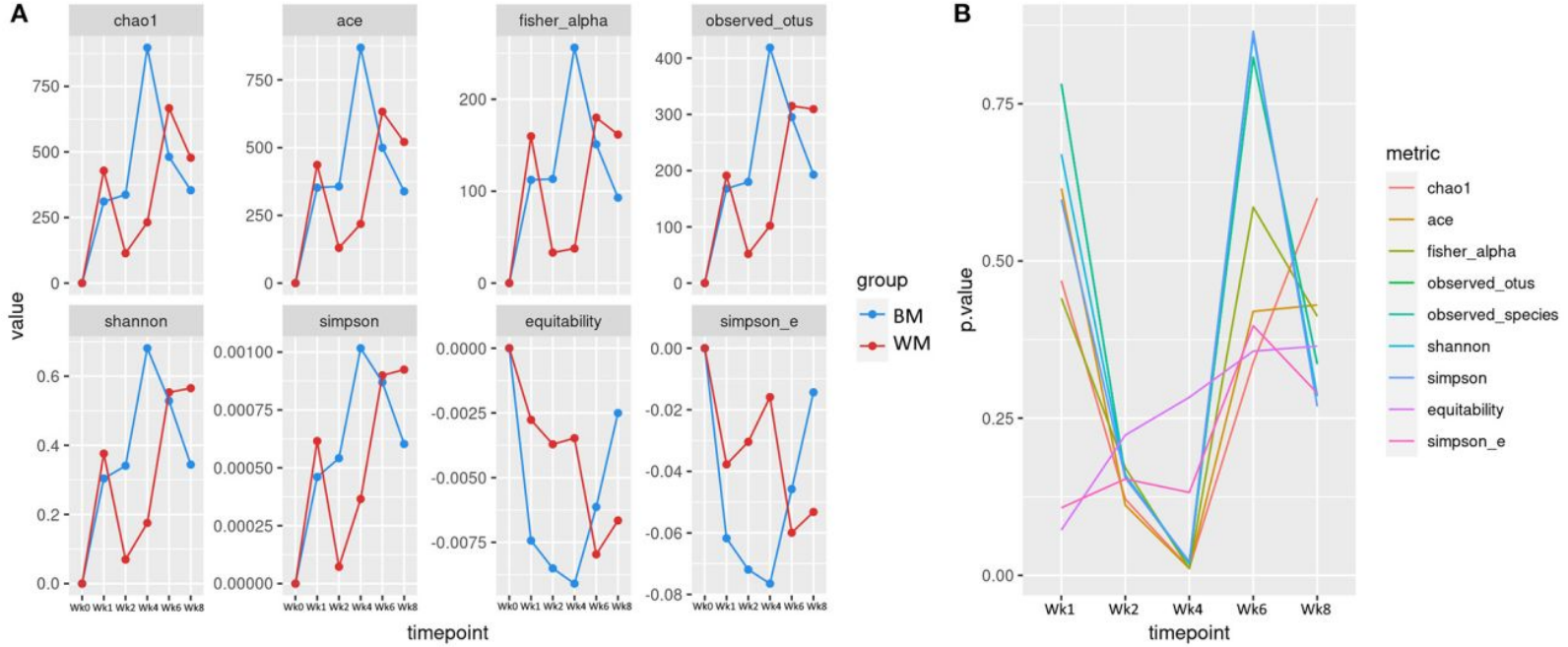


FIGURE 2 | (A) Alpha diversity indices at the different trial time points for the two groups: bulk tank milk (BM) and waste milk (WM). Index values are indicated as differences from baseline (Wk0 = 0). **(B)** Statistical significance of alpha diversity indices at the various time points.

From Penati et al. 2021

A few examples from literature

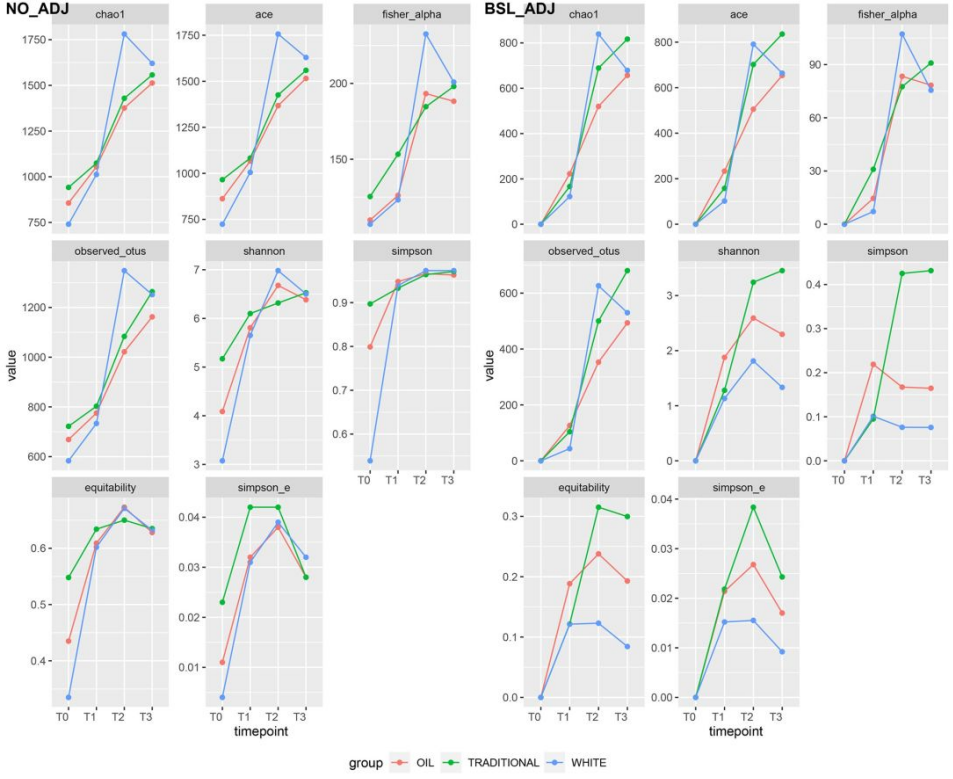


Fig 5. Alpha diversity. Average alpha diversity indices per group over timepoints. Non-adjusted (left) and baseline-adjusted (right) values.

From Cremonesi et al. 2022

A few examples from literature

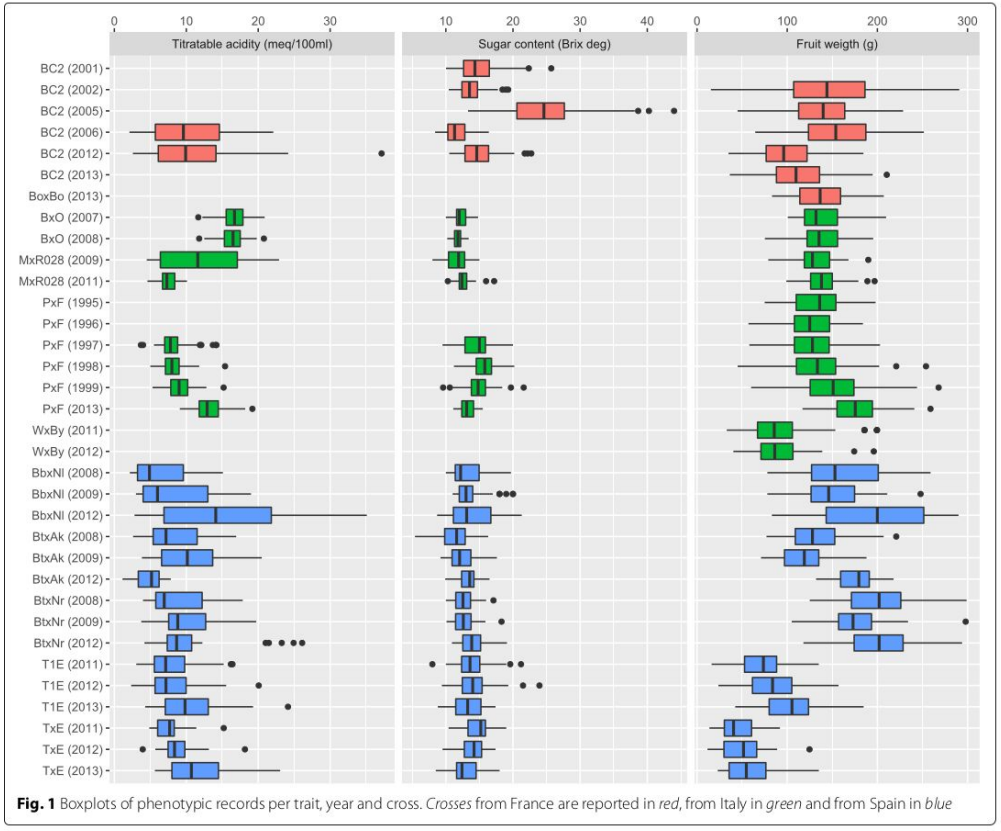


Fig. 1 Boxplots of phenotypic records per trait, year and cross. *Crosses* from France are reported in *red*, from Italy in *green* and from Spain in *blue*

From Biscarini et al. 2017

A few examples from literature

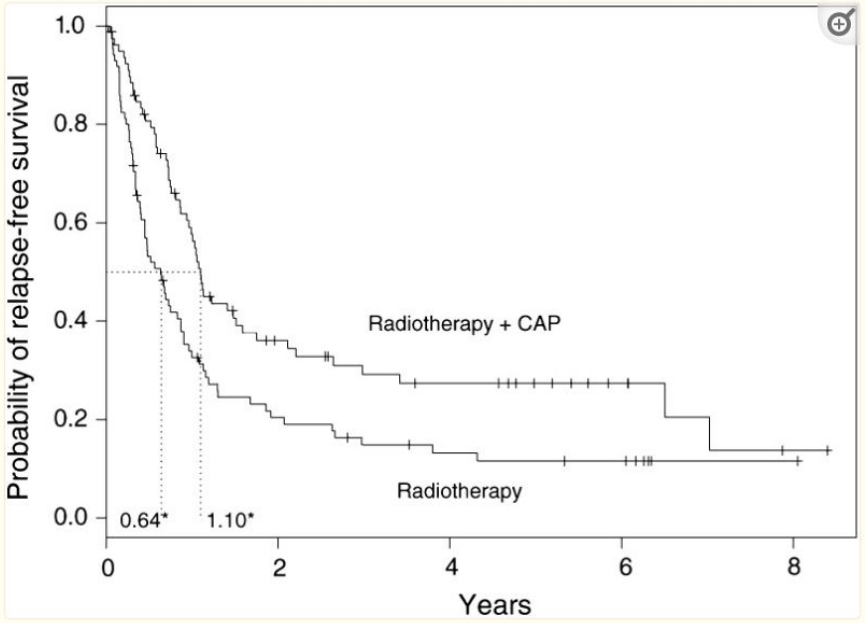


Figure 2

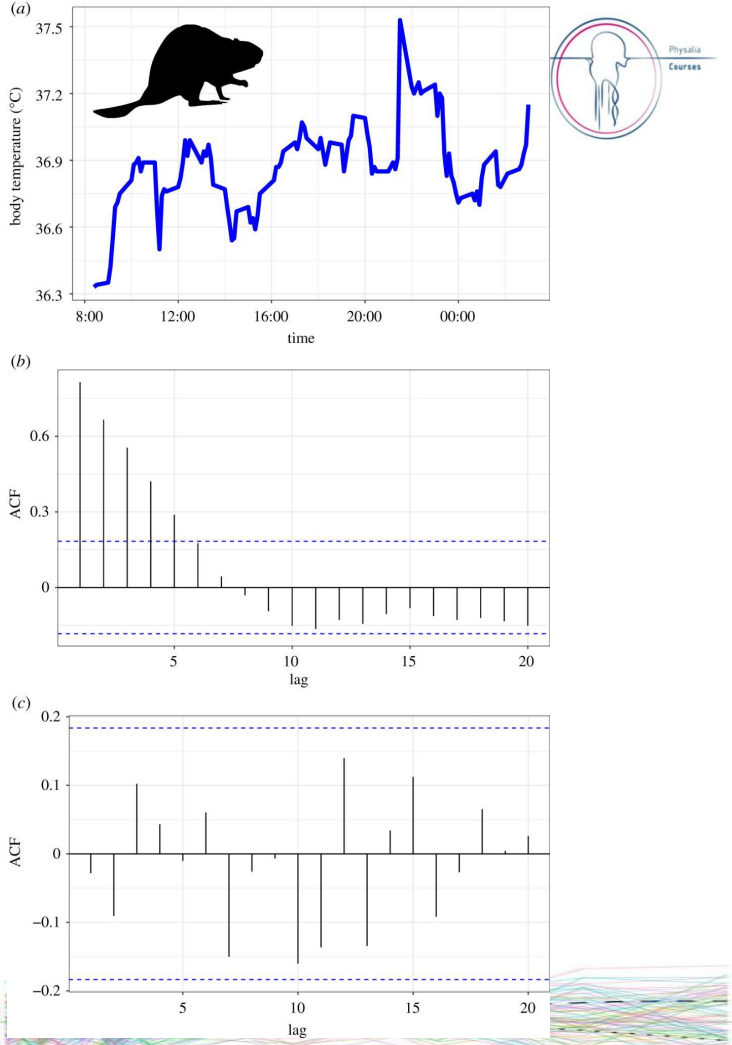
Relapse-free survival curves for the lung cancer trial. * Median relapse-free survival time for each arm, + censoring times, CAP=cytoxan, doxorubicin and platinum-based chemotherapy.

Clark et al 2003; https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2394262/

A few examples from literature

Time-series data on beaver body temperature:

- (*a*) Beaver body temperature data from a single individual recorded every 10 min over approximately 24 h
- (*b*) autocorrelation plot of the residuals; helps identify that there is still unmodelled autocorrelation in these data
- (*c*) Applying a more complex temporal autocorrelation model resolves these issues and produces a satisfactory autocorrelation plot



Harrison 2021; https://royalsocietypublishing.org/doi/10.1098/rstb.2020.0227