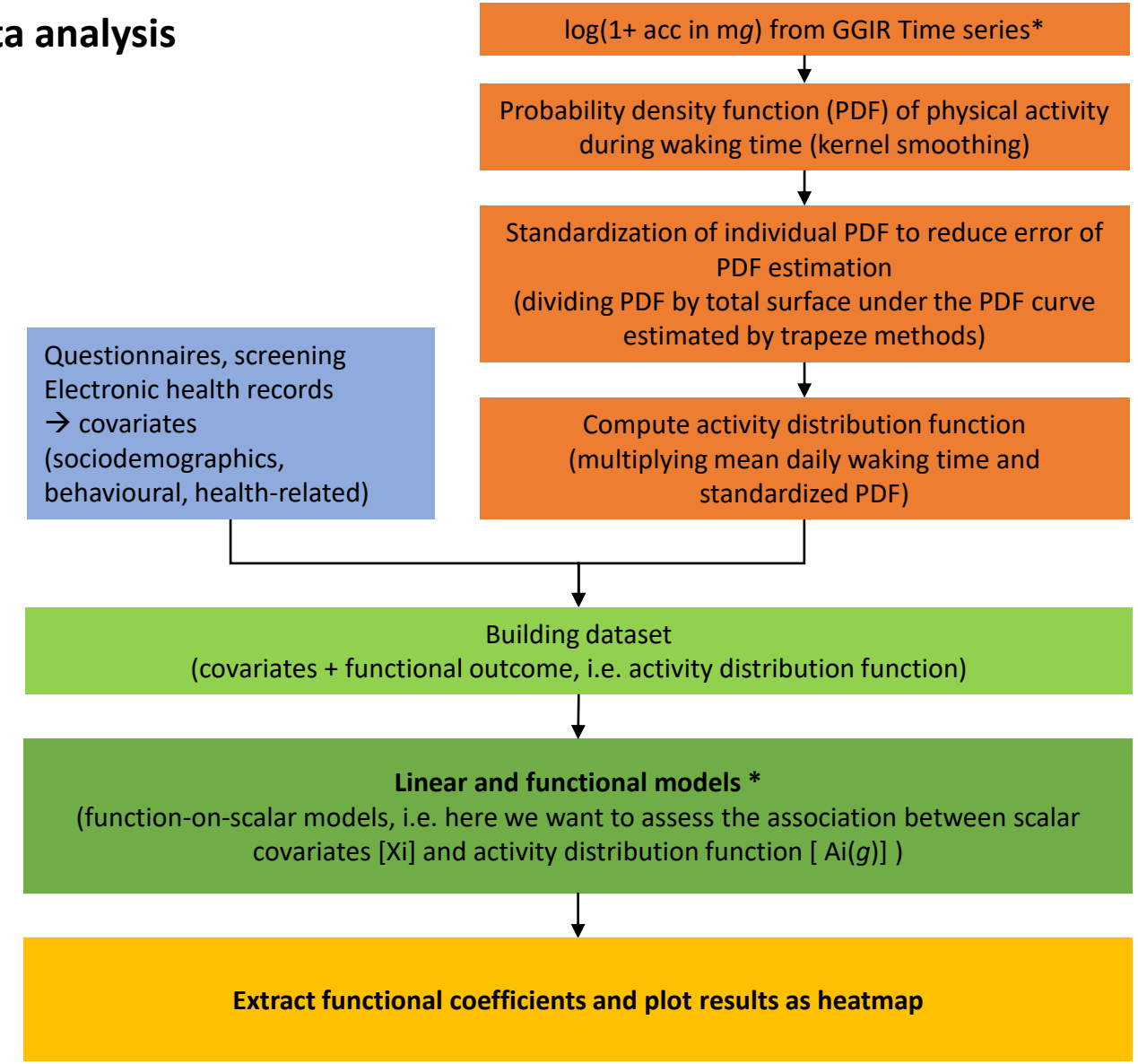
Workflow – functional data analysis



Corresponding script
00_1_KERNEL_DENSITY

Corresponding script
00_2_STANDARDIZED_DENSITY

Corresponding script
00_3_ACTIVITY_DISTRIBUTION

Corresponding script
***00_POPULATION_p3-7-11**:*
building the covariates datasets for each phase

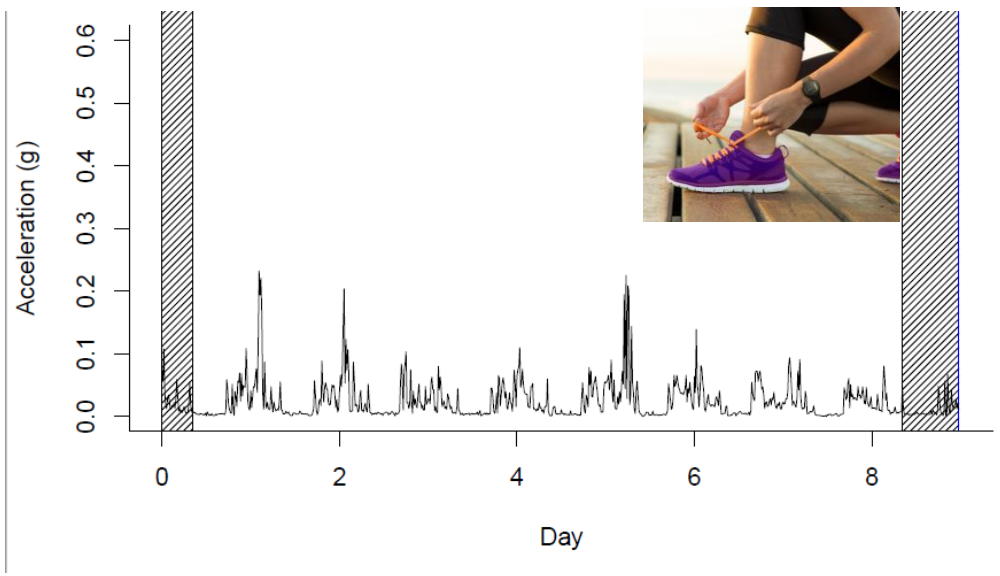
Corresponding scripts
Starting with 01_DATA

Corresponding scripts
- ***MODELS.R** (full list of models)*
- ***MODELS_quick.R** (only those used in the plots for the article)*

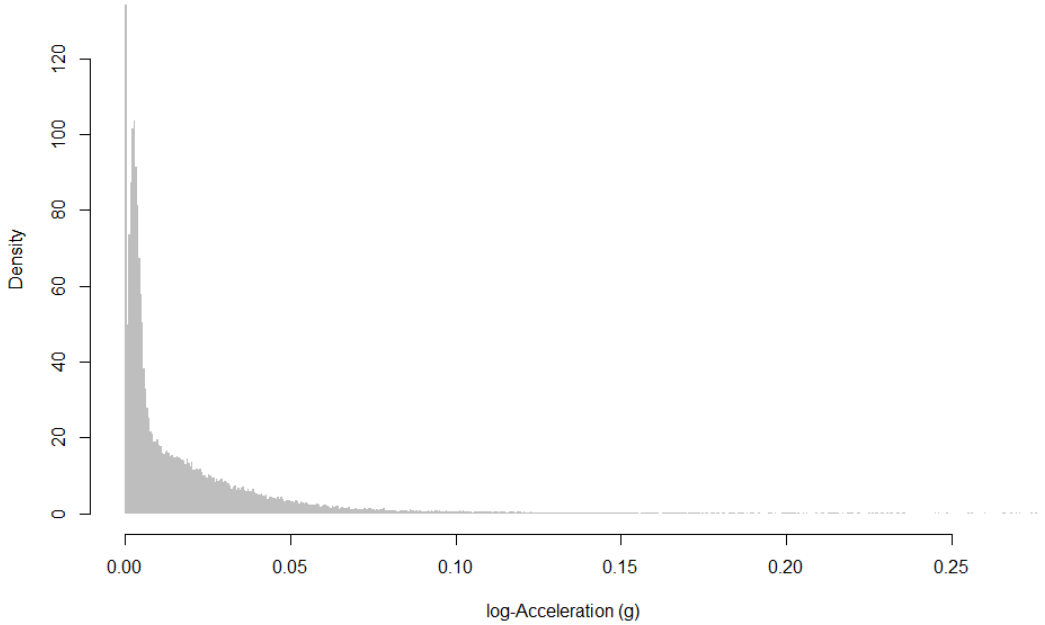
Corresponding script
04_TABLES_FIGURES.R

* For more information on basic use of refund R package to fit functional models please refer to https://cran.r-project.org/web/packages/refund/refund.pdf

Computation of the activity distribution

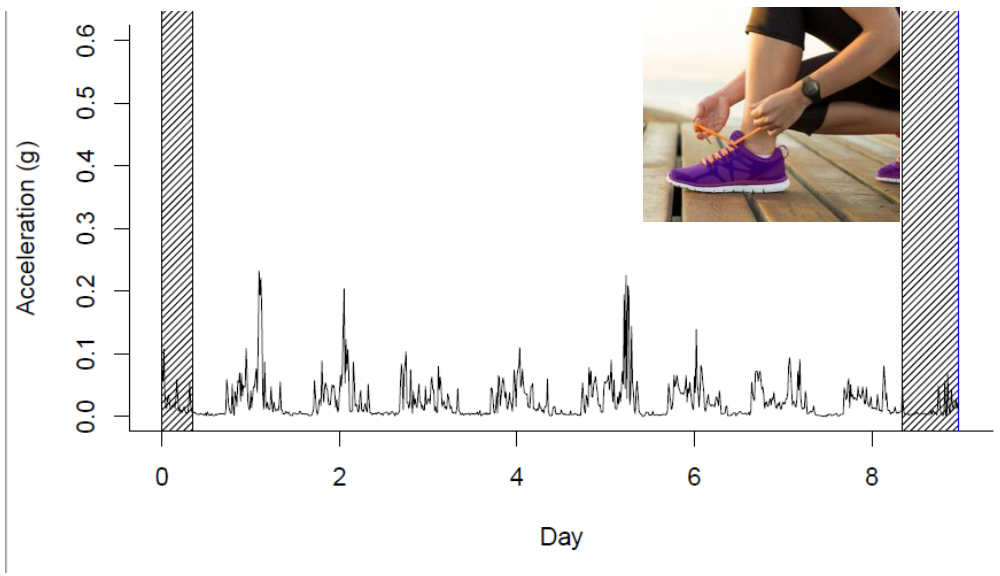


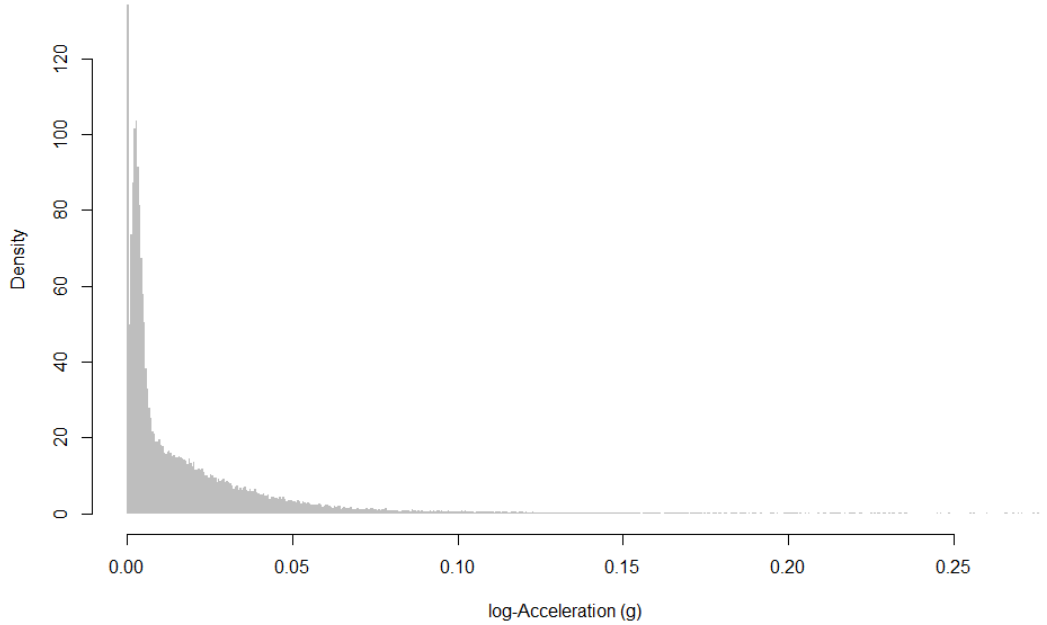
Raw accelerometer data

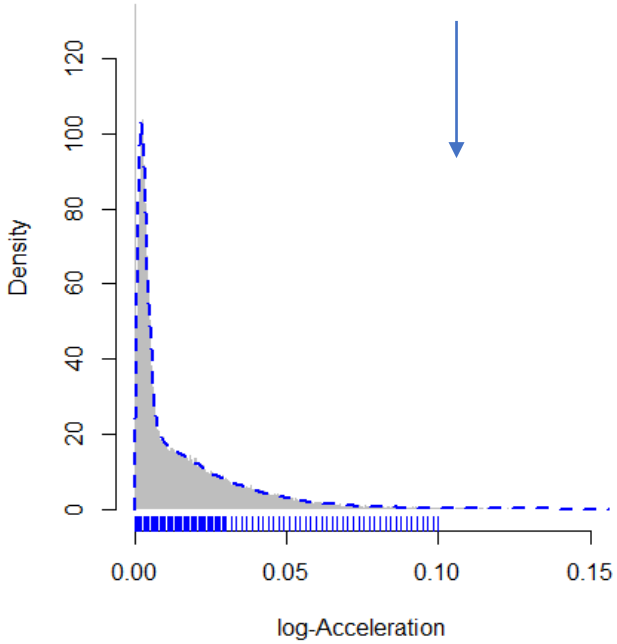


Histogram: approximate representation of the distribution of intensities on the acceleration range (discontinous)

Computation of the activity distribution







Probability density function

- Continuous
- Area under the curve = **proportion of the time** spent in a given range of intensities

Computation of the activity distribution

Corresponding script
00_1_KERNEL_DENSITY

1. Probability density function (PDF) of physical activity during waking time

The challenge is to characterise the PDF of each individual using the kernel smoothing method.
We use the kde() function of the ks* R package (https://cran.r-project.org/web/packages/ks/ks.pdf).
Two parameters need to be predefined before estimating the function:

- Knowing that very low values are over-represented compared to the highest values (participants accumulate a lot of time in sedentary behaviour, i.e. in the lowest intensities of the activity spectrum, and very little time in moderate-to-vigorous activity, i.e. in the highest intensities), we estimate the PDF at **150 points distributed in a logarithmic way** (so that there are more points in the low intensities than in the high intensities
- The bandwidth parameter must also be defined. By default the function will select the value that will minimise the estimation error of the density to be estimated for each participant (one bandwidth value per participant). We decided to re-estimate the density by taking the **median value of the band over the whole population** in order to have a common value for the whole population.

* The density() function in the basic R package also uses the kernel method and the bandwidth is defined by Silverman's Rule of Thumb method. But we cannot give it any points on which to calculate the function, hence our choice to use kde()

Corresponding script
00_2_STANDARDIZED_DENSITY

2. Standardization of individual PDF to reduce error of PDF estimation

For 99.4% of the participants, total area under the PDF curve was not equal to 1. To correct the estimation of PDF, we divided it by the total area under the PDF curve. The area under the curve was computed by trapeze methods, using the cumytrapz() function of the pracma R package (https://www.rdocumentation.org/packages/pracma/versions/1.9.9/topics/trapz).

```
# Area computation
surf_y <- cumtrapz(x, y)

# Standardization
# y_2 <- y/surf_y

# -> test if (cumtrapz(x, y_2) == 1)
```

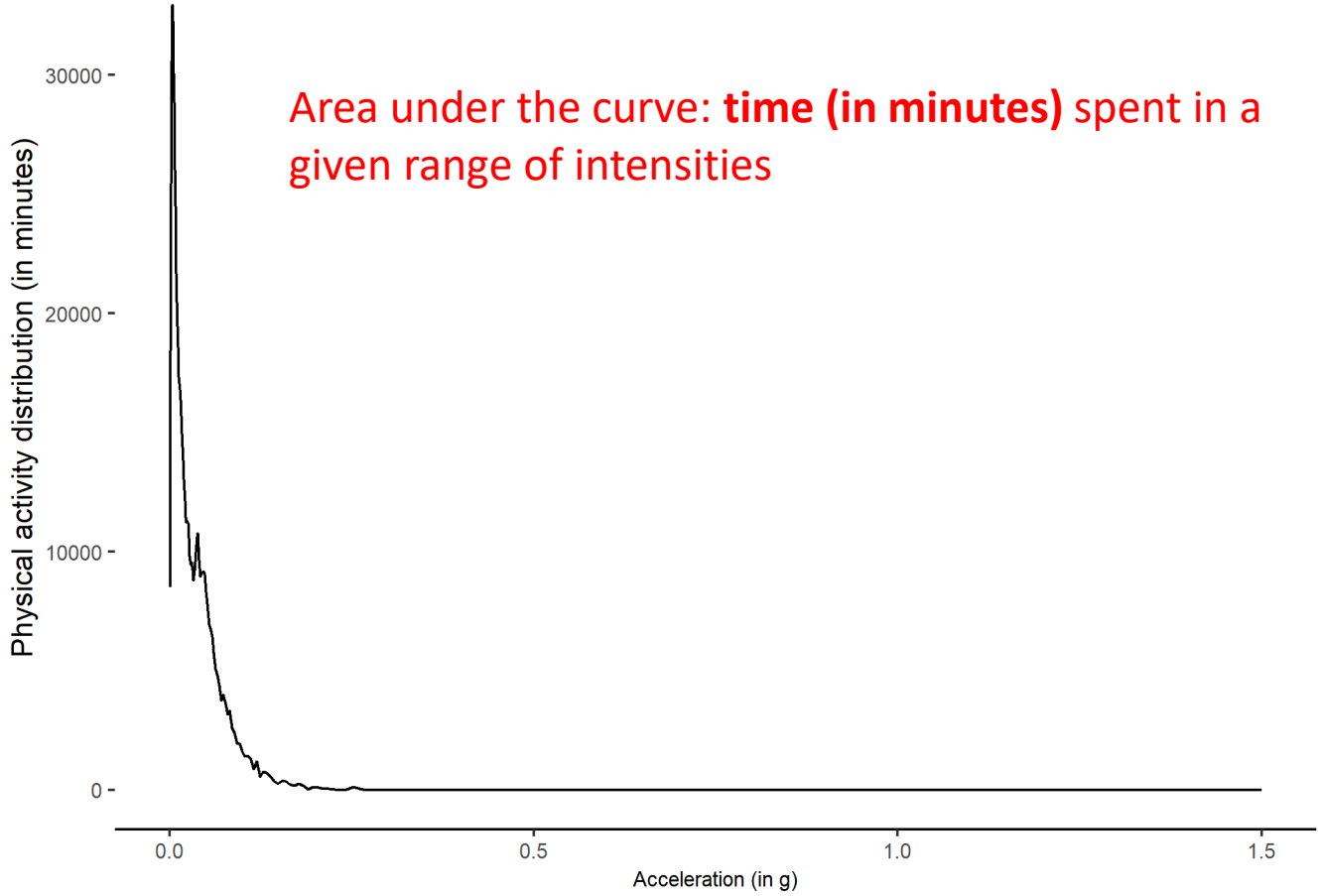
Computation of the activity distribution

Corresponding script
00_3_ACTIVITY_DISTRIBUTION

3. Compute activity distribution function

Standardized PDF is contrained (always positive and total area under the curve = 1). To ease interpretation of the coefficients of the functional models we multiply the PDF by waking time. This will allow us to interpret coefficients of the function-on-scalar model as variations in time rather than proportion of time.

→ Estimation of the activity distribution function
= waking time * density at each value of acceleration



Area under the curve: **time (in minutes)** spent in a given range of intensities

Computation of the activity distribution

In fact, for each participant, we have 150 observations from the activity distribution function

i	f_i(x_j)	x_j
1	f_1(x_1)	x_1
1	f_1(x_2)	x_2
…	…	…
1	f_1(x_150)	x_150

→ **Functional data**