Preliminary analysis of crowdsourced sound data with FOSS

Nicolas Roelandt, **P. Aumond**, **L. Moisan**

FOSDEM 2023

Press P to access notes

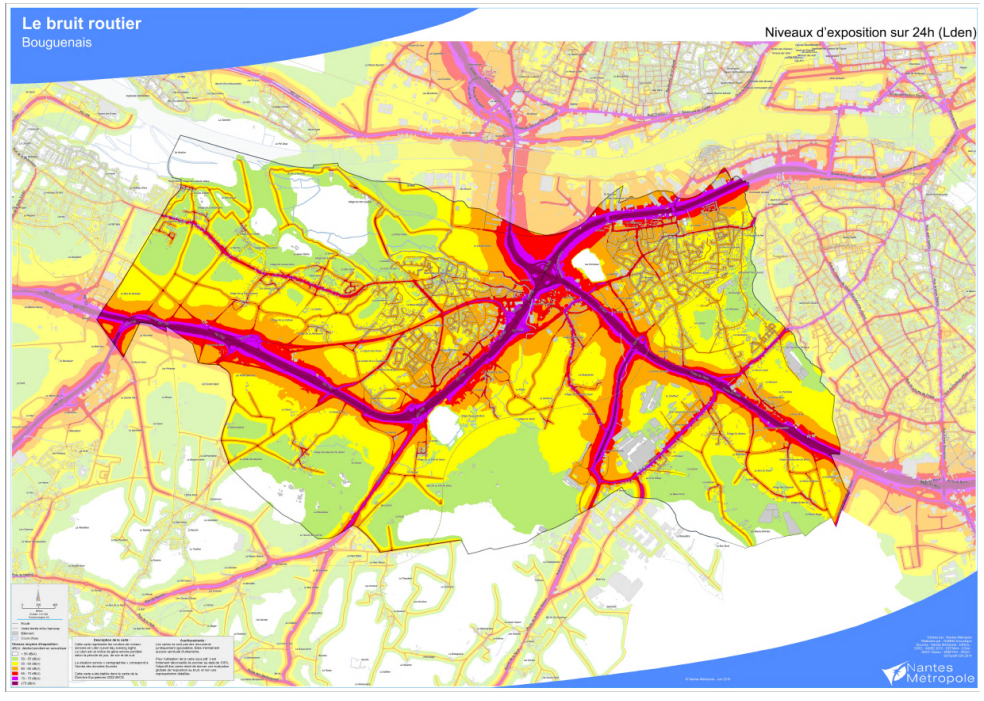
Introduction

Traffic noise is a major health concern :

- 1 million healthy life years (DALYs) lost each year in Western Europe due to traffic noise WHO 2011
- social cost of noise in France estimated at 147 billion euros per year ADEME 2021

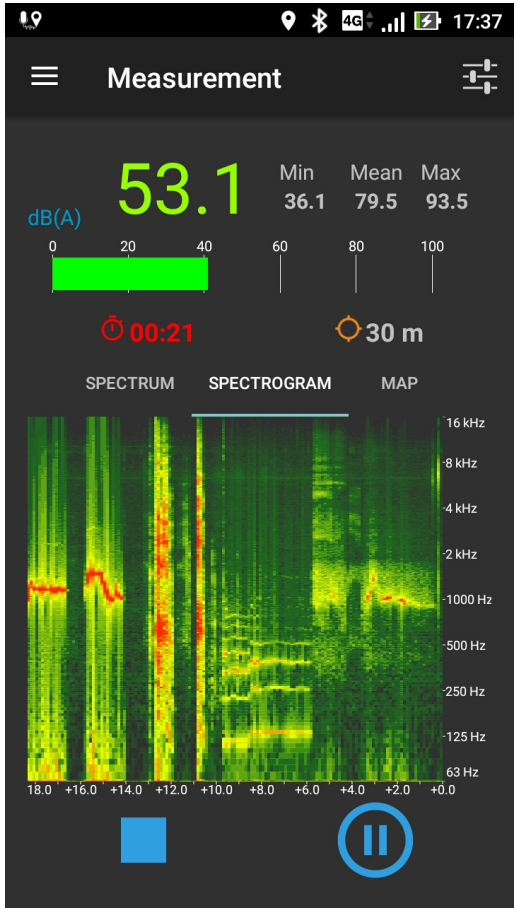
How to find problematic areas ?

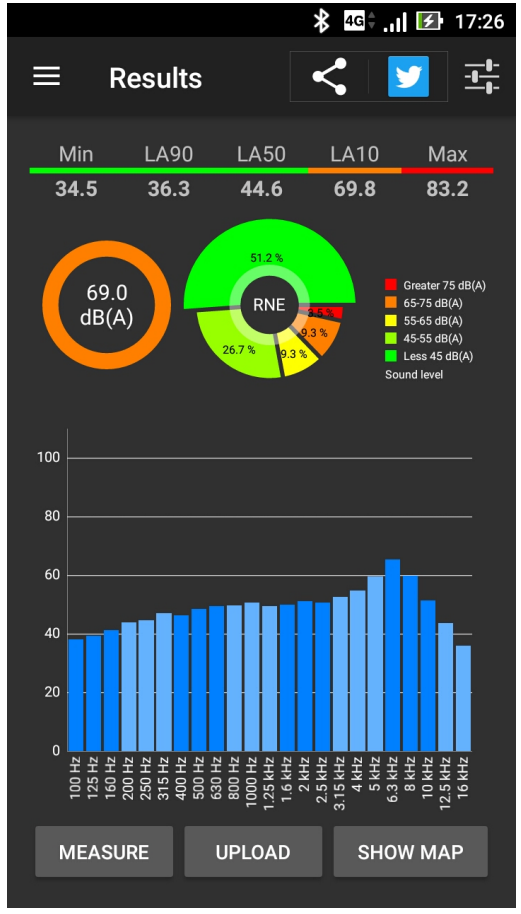
- Direct measure on the whole area is not possible
- Traditional way is simulation from traffic counts (air, rail, road) and infrastructure



Map generated with NoiseModelling

UMRAE proposal : Capture sound environment with a smartphone app.

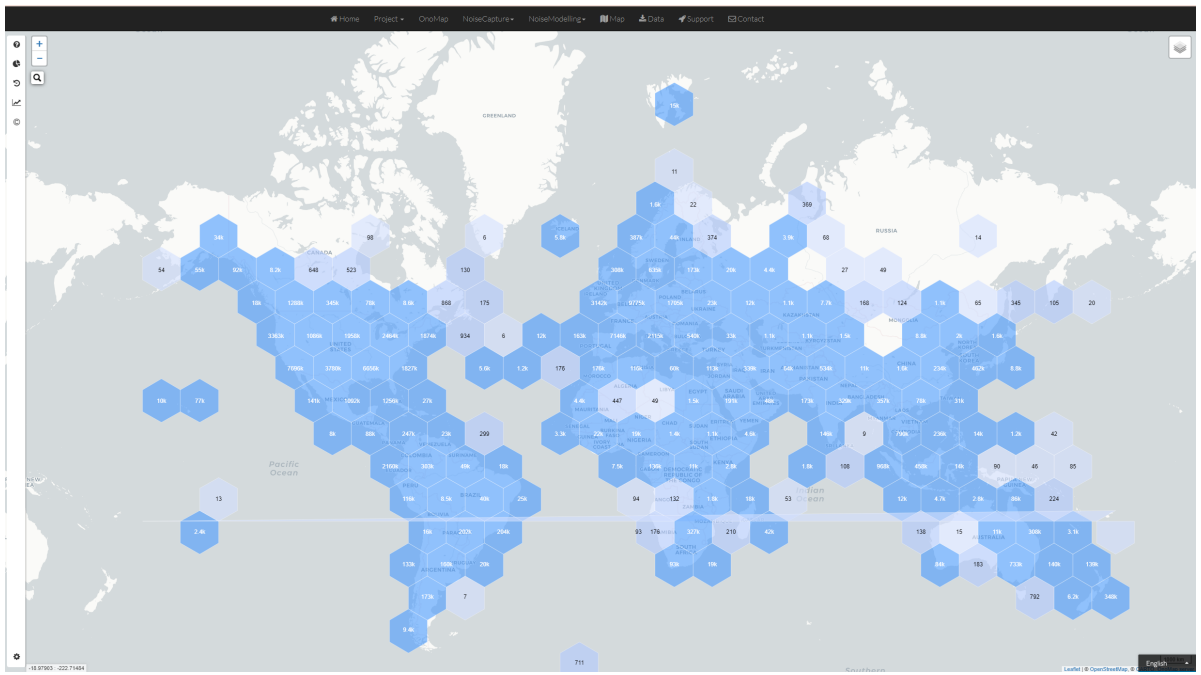




NoiseCapture is available on F-Droid

NoisePlanet Project

- NoiseModelling: generate noise maps from Open Source geodata
- NoiseCapture : measure and share sound environment
- OnoMap : Spatial Data Infrastructure
- Community maps



What can we do with the data collected by the app ?

NoiseCapture dataset

- 3 years data extraction (2017-2020, still collecting)
- 260 000 tracks worldwide
- sound spectrum, tags and gps localization
- ODC Open Database License v1.0

data.univ-gustave-eiffel.fr/dataset.xhtml?persistentId=doi:10.25578/J5DG3W

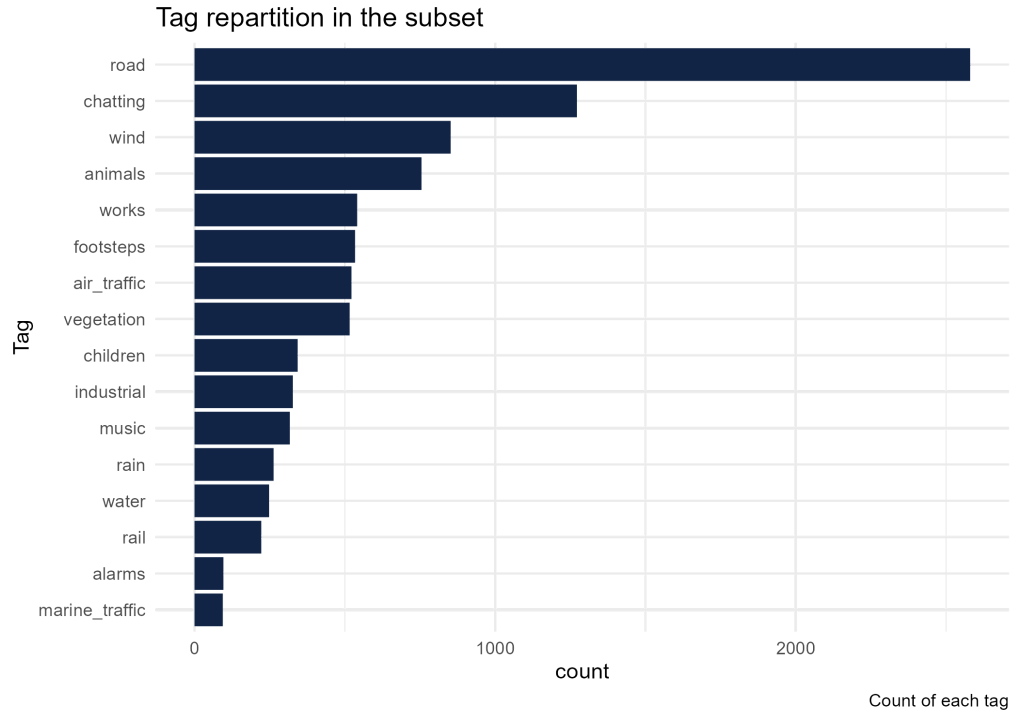
How to characterize of the user environment with the collected data ?

2 possibilities :

- from the sound spectrum (ongoing analysis)
- from the *tags* defined by the contributor

Database and subset

- 260 422 tracks
- 124 363 with tags
- 50280 not indoor or tests
- 47 412 duration > 5 s
- 11 492 in France



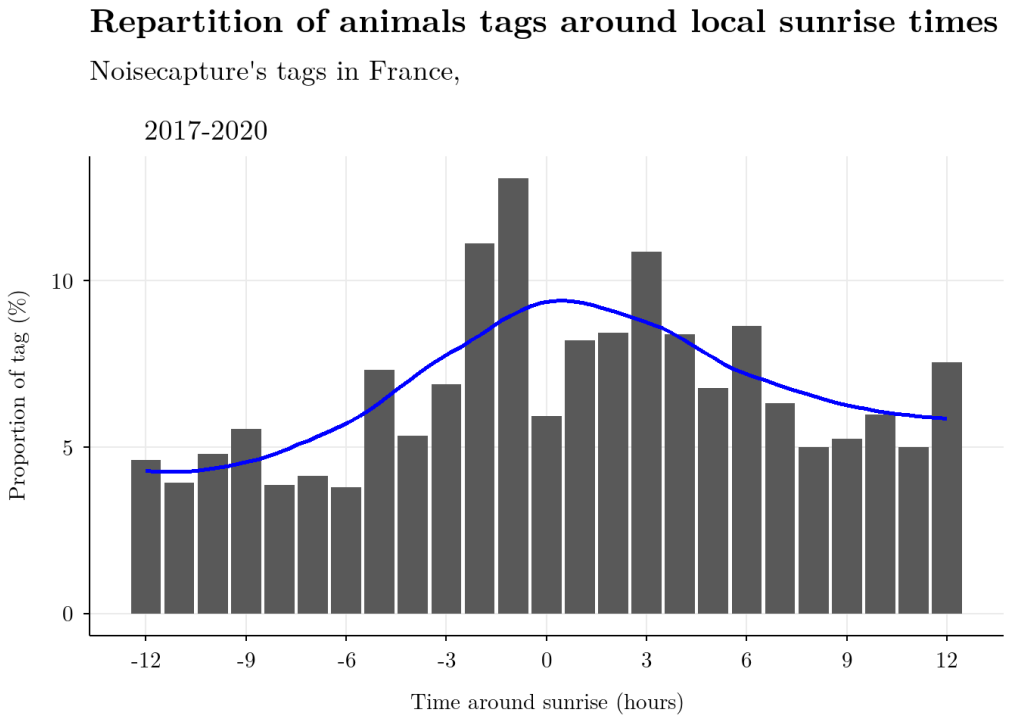
Toolkit

A quite simple one:

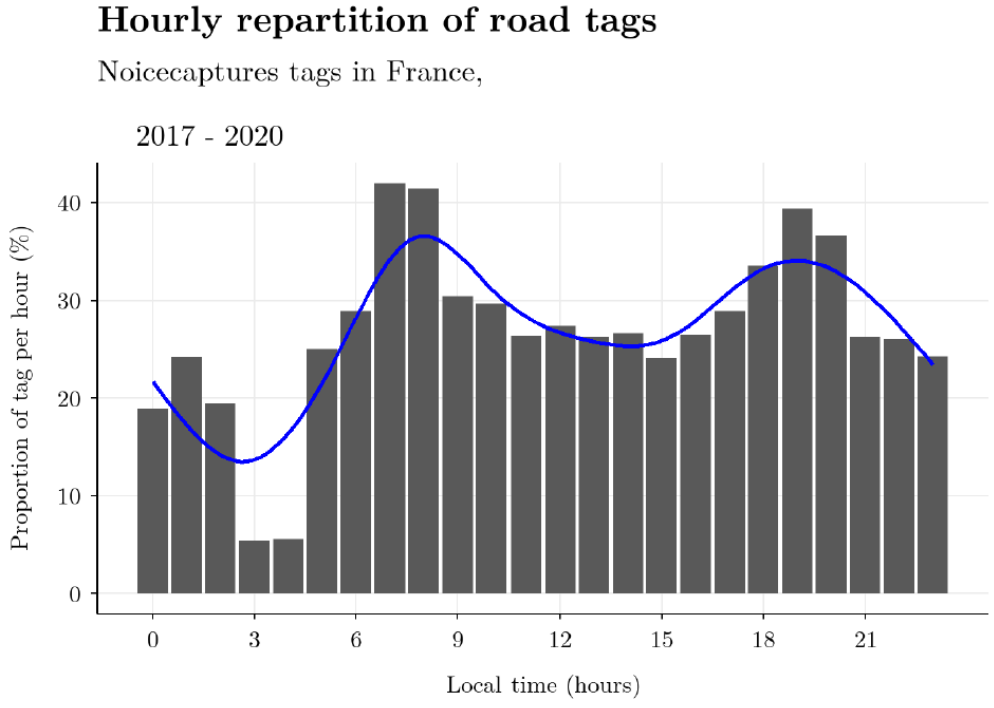
- PostgreSQL/PostGIS
- R
- Lots of R packages : Tidyverse, sf, geojsonsf, stats, suncalc…
- Dependencies : Pandoc, Markdown, Reveal.js, Proj, GEOS, GDAL, etc…

What do we found in the dataset ?

Well known temporal sound source dynamics



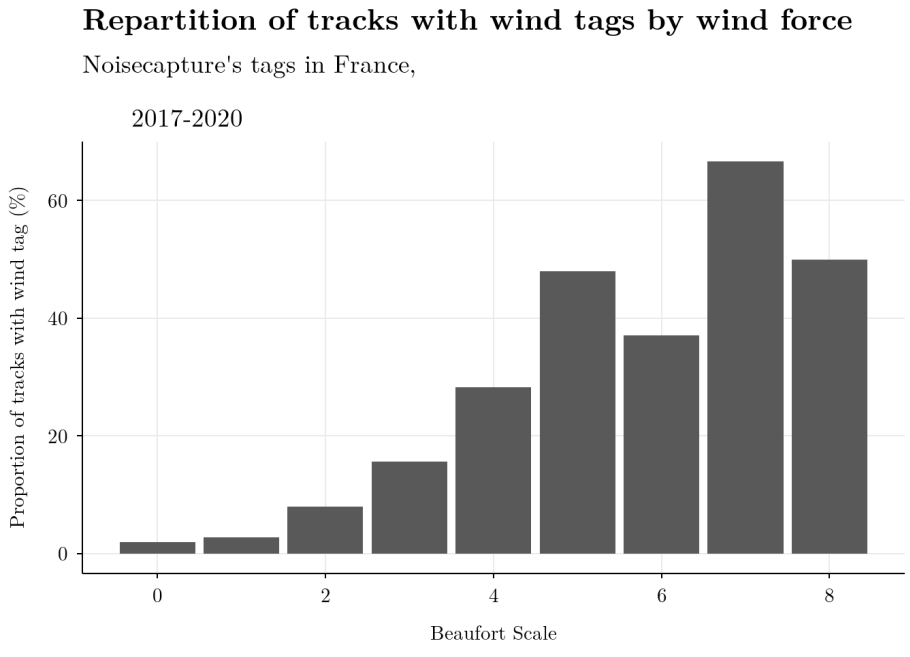
Bird songs at dawn

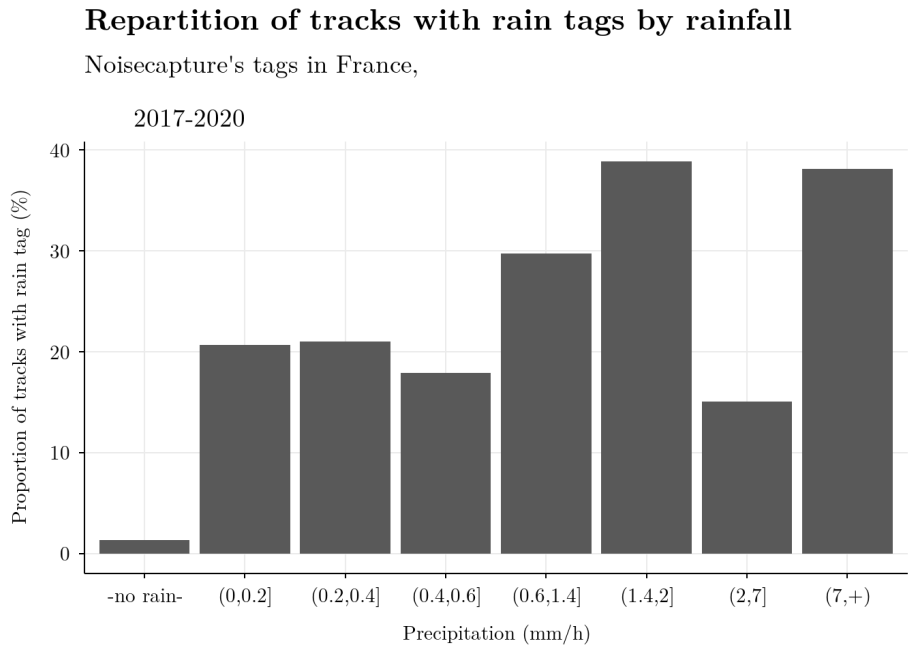


Commuters traffic noise

What do we found in the dataset ?

Physical events





r(7) = .93 ($p < 0.01$) between `wind` tag proportion and the measured wind force

r(6) = 0.68 ($p < 0.1$) between the `rain` tag proportion and the measured rain fall

Reproducible Science is an issue

Good

- Data available
- Source code available (SQL scripts and R notebooks)
- Setup available

Bad

- Some notebooks needs work on reproducibility (and code factoring)
- Information on software environment is too scarce (and hard to reuse)

Reproducible Science is an issue

Some avenues of investigation

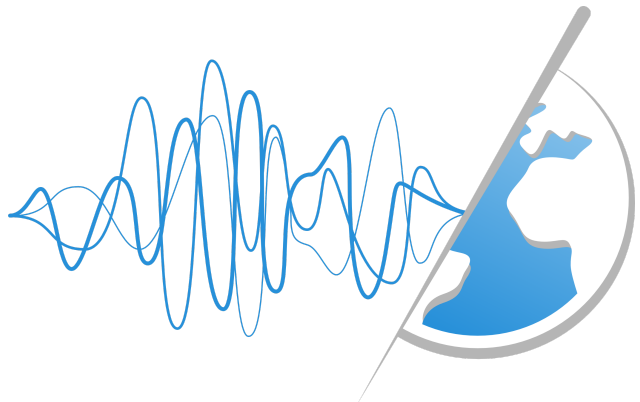
- R package Renv
- Docker
- Guix

Conclusion

Conclusion

- Crowdsourced data can be useful for science
- This dataset is usable
- FOSS are key for Reproducible Science
- Reproducible Science is hard to achieve
- Notebooks **are not enough**

data.univ-gustave-eiffel.fr/dataset.xhtml?persistentId=doi:10.25578/J5DG3W



noise-planet.org

Thanks!

Nicolas Roelandt - Univ. Gustave Eiffel

nicolas.roelandt@univ-eiffel.fr @NRoelandt@sciences.re

This presentation : https://s.42l.fr/FOSDEM2023-LASSO

Access to code source : github.com/Universite-Gustave-Eiffel/lasso-data-analysis

Detailed articles and notebooks : universite-gustave-eiffel.github.io/lasso-data-analysis/articles/

Slides created via the R packages xaringan and gadenbuie/xaringanthemer