The Alan Turing Institute

R (or Python) for Open and Reproducible Science

Andrew Mitchell

My background

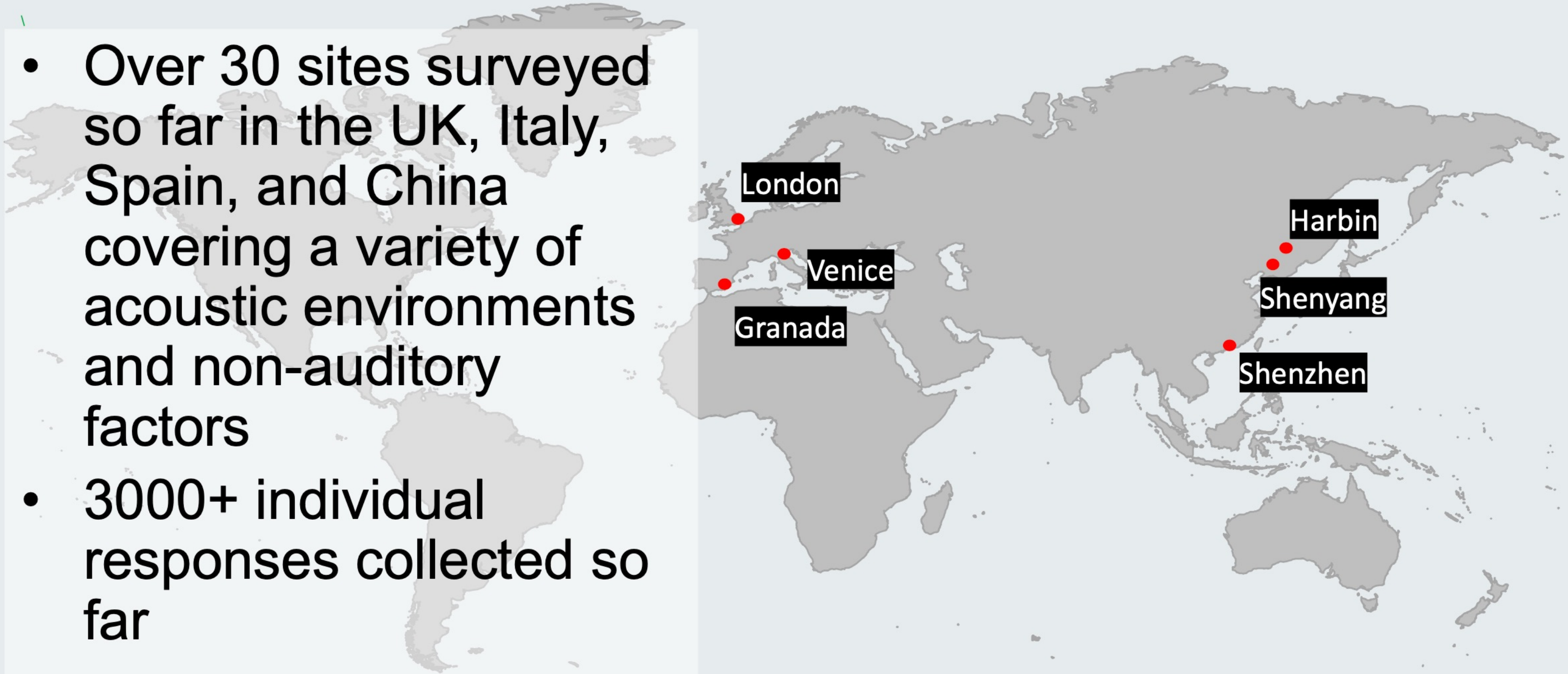
Research Fellow, University College London

My research – Soundscape Indices (SSID)

- Soundscape attempts to describe urban sound environments in terms of **how they are perceived**
- We describe soundscapes in terms of their pleasantness and eventfulness, telling us if they are vibrant, or calm, or chaotic, etc.
- SSID is a project to make this approach practical, through a model which can predict these perceptions based on physical inputs

Case study sites

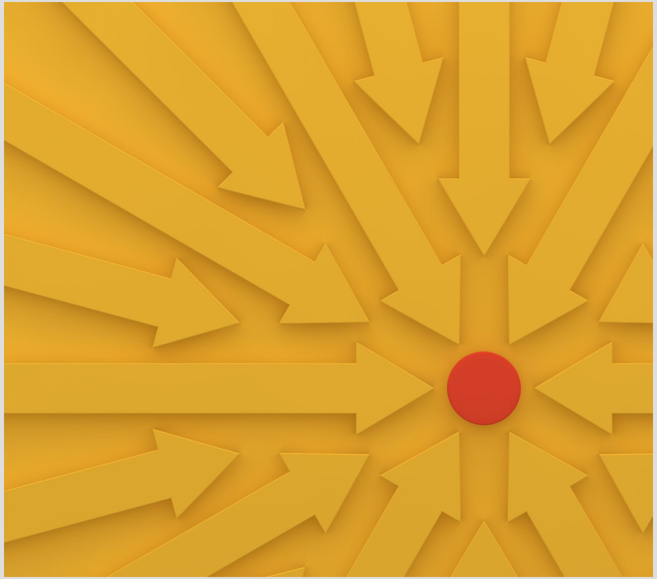
- Over 30 sites surveyed so far in the UK, Italy, Spain, and China covering a variety of acoustic environments and non-auditory factors
- 3000+ individual responses collected so far



Goals

To show how to make your R more open and reproducible and why you should do it

- NOT to teach you how to write code
- To introduce tools and workflows to improve your work
- To give real examples from my own work



Part 1: Open Science

- What is Open and Reproducible Science?
- How does R (or Python) help?

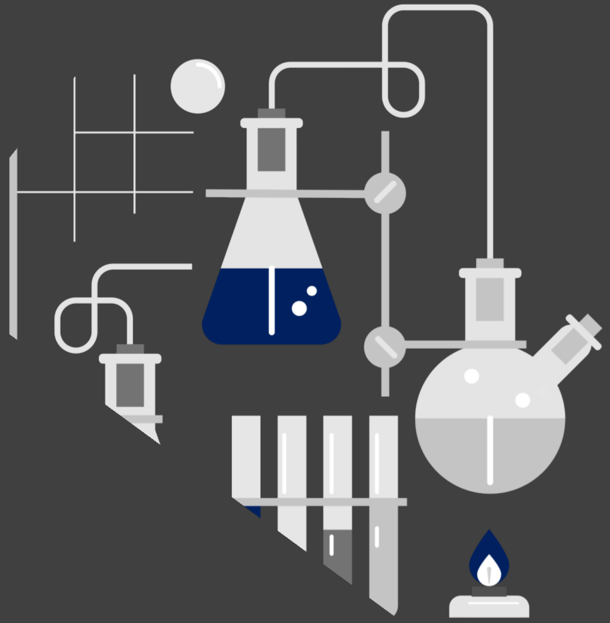
Part 2: Practical Python

- Data Processing
- Data Analysis
- Interactive Code
- Sharing and Collaborating

Part 3: Examples

- Soundscapy
- Sharing Data
- Sharing Code
- Making it accessible

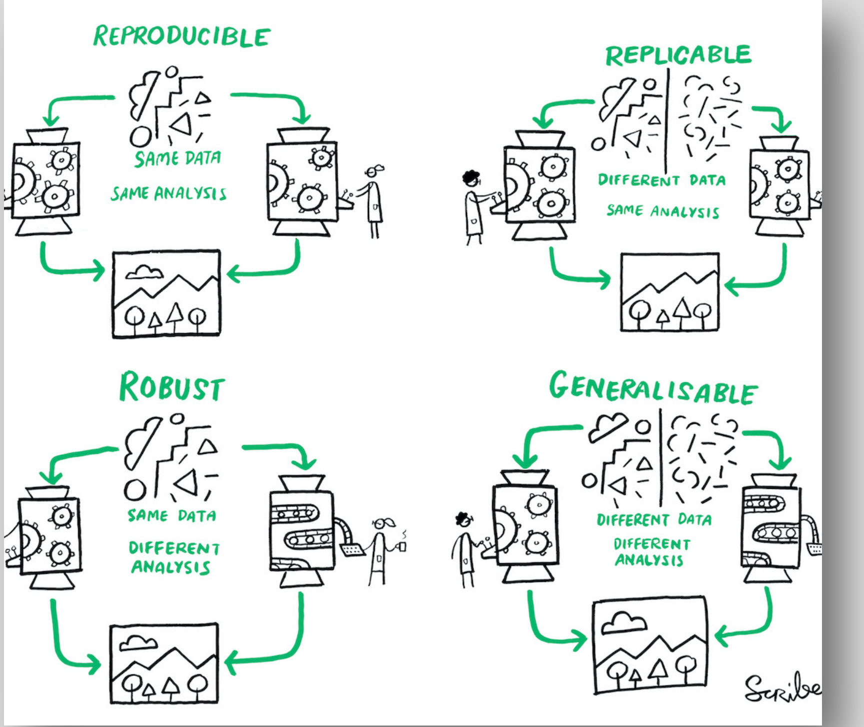
Part 1 – Open Science



What is Open and Reproducible Science?

Reproducible

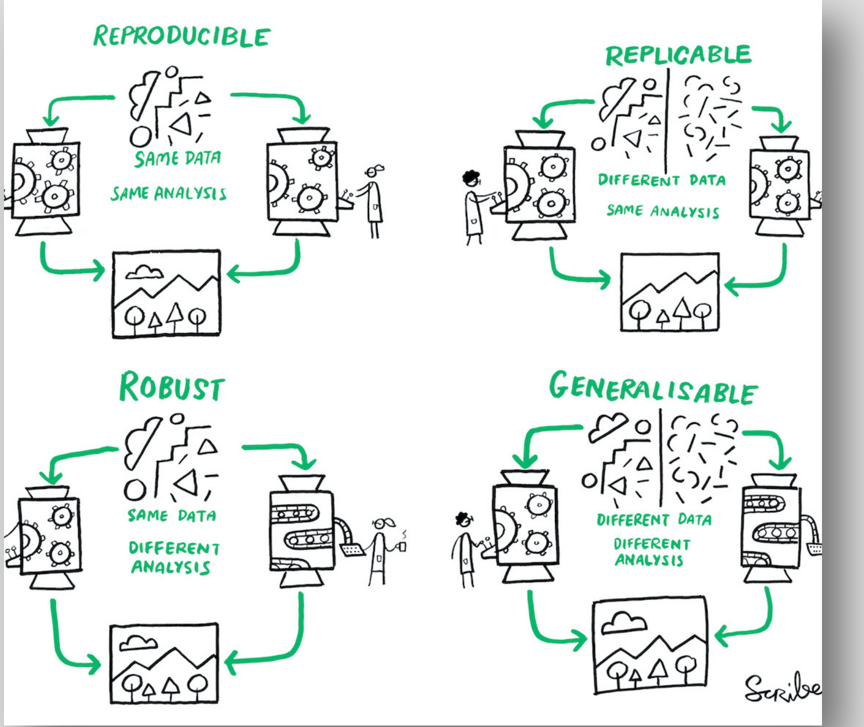
Authors provide all the necessary data and the computer codes to run the analysis again, re-creating the results



The Turing Way project illustration by Scriberia. Used under a CC-BY 4.0 licence. DOI: 10.5281/zenodo.3332807.¶

Replicable

A study that arrives at the same scientific findings as another study, collecting new data (possibly with different methods) and completing new analyses.



The Turing Way project illustration by Scriberia. Used under a CC-BY 4.0 licence. DOI: 10.5281/zenodo.3332807.¶

Open Research

Aims to transform research by making it more reproducible, transparent, reusable, collaborative, accountable, and accessible to society

- Be publicly available
- Be reusable
- Be transparent

From *The Turing Way,* “Open Research” https://the-turing-way.netlify.app/reproducible-research/open.html

Open Research

Open Data

Open Source Software

Open Access

Open Notebooks

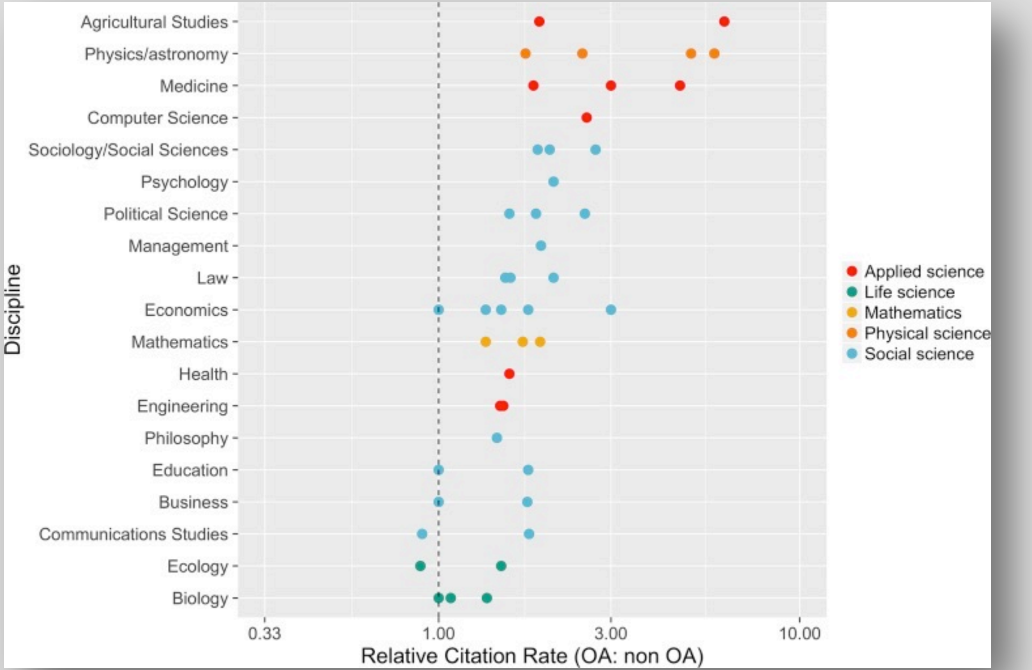
From *The Turing Way,* “Open Research” https://the-turing-way.netlify.app/reproducible-research/open.html

Five selfish reasons to work reproducibly

1. Reproducibility helps to avoid disaster
2. Reproducibility makes it easier to write papers
3. Reproducibility helps reviewers see it your way
4. Reproducibility enables continuity of your work
5. Reproducibility helps to build your reputation

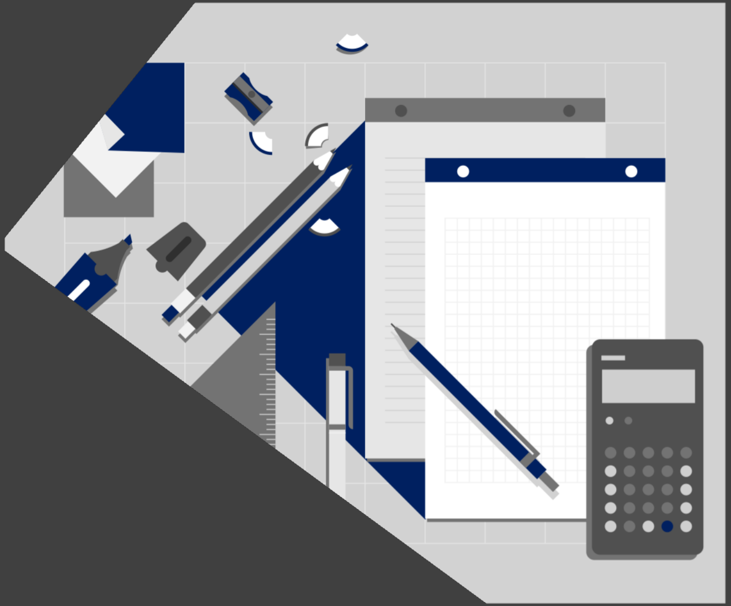
Markowetz, F. Five selfish reasons to work reproducibly. Genome Biol 16, 274 (2015). https://doi.org/10.1186/s13059-015-0850-7

Open Access research gets cited more



The relative citation rate (OA: non-OA) in 19 fields of research. This rate is defined as the mean citation rate of OA articles divided by the mean citation rate of non-OA articles. Multiple points for the same discipline indicate different estimates from the same study or estimates from several studies.

How does R (or Python) help?



Levels of Code

Function

Script

Module

Package (or library)

Reproducible by default

- The biggest advantage is reproducibility, both for yourself and for others
- Transparent reproducibility

Iteration of Analysis

- By working in code, we can iterate and improve our analysis, without starting from scratch
- Suggestions from reviewers can be (more) easily integrated

Collaboration

- Code is easily shared between research partners
- Using notebooks can make the process even easier
- Collaborative development and improvement is at the heart of open source software more broadly

Python

- General Purpose
- Very flexible
- Popular outside Academia
- Learning curve is smooth
- Better extension to machine learning

Both

- Notebooks
- Lots of open-source libraries
- Readable

R

- Stats Focused
- Very popular in Academia
- Simpler to install
- Easy to start with, can get very difficult for advanced work

Interactive Code:

Quarto / Jupyter Notebooks

Quarto/Rmarkdown/Jupyter Notebooks

Analyse

- Break up code development into blocks
- Iterate your analysis strategy
- View results inline
- Easily switch out datasets while keeping the same analysis

Collaborate

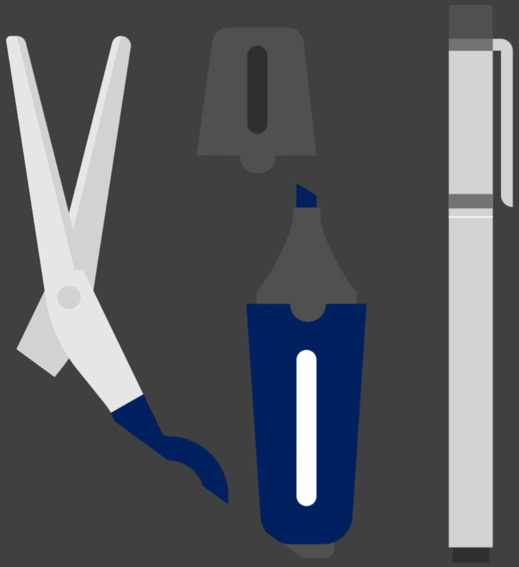
- Combine rich markdown text and inline code
- Share preliminary results
- Can provide context and results to collaborators
- Share either dynamic or static versions

Publish

- Allow readers/reviewers to easily reproduce results
- Show the process

Part 3 – Examples

Soundscapy, CircE, ARAUS



How to incorporate these principles

Collaboration

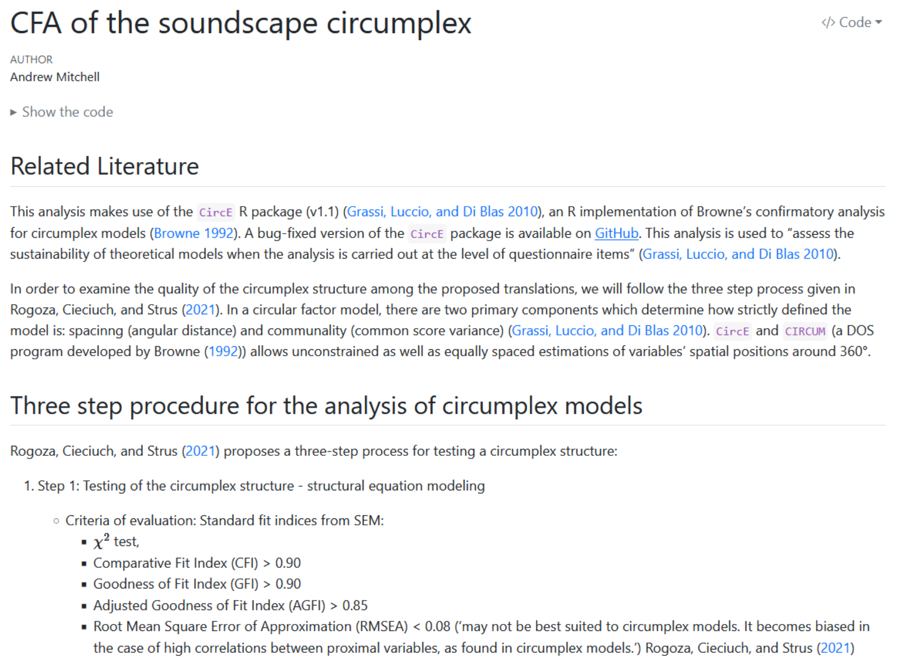
Share code and analysis side-by-side

Publishing with a paper
Reproducible code and data
Make sure others can run the code

Making a tool
Turn your research code into a package

Collaboration

Use a notebook to easily organize and communicate your analysis with collaborators



CFA of the soundscape circumplex

AUTHOR

Andrew Mitchell

Related Literature

This analysis makes use of the `CircE` R package (v1.1) (Grassi, Luccio, and Di Blas 2010), an R implementation of Browne's confirmatory analysis for circumplex models (Browne 1992). A bug-fixed version of the `CircE` package is available on GitHub. This analysis is used to "assess the sustainability of theoretical models when the analysis is carried out at the level of questionnaire items" (Grassi, Luccio, and Di Blas 2010).

In order to examine the quality of the circumplex structure among the proposed translations, we will follow the three step process given in Rogoza, Cieciuch, and Strus (2021). In a circular factor model, there are two primary components which determine how strictly defined the model is: spacinng (angular distance) and communality (common score variance) (Grassi, Luccio, and Di Blas 2010). `CircE` and `CIRCUM` (a DOS program developed by Browne (1992)) allows unconstrained as well as equally spaced estimations of variables' spatial positions around 360°.

Three step procedure for the analysis of circumplex models

Rogoza, Cieciuch, and Strus (2021) proposes a three-step process for testing a circumplex structure:

1. Step 1: Testing of the circumplex structure - structural equation modeling
    - Criteria of evaluation: Standard fit indices from SEM:
        - χ^2 test,
        - Comparative Fit Index (CFI) > 0.90
        - Goodness of Fit Index (GFI) > 0.90
        - Adjusted Goodness of Fit Index (AGFI) > 0.85
        - Root Mean Square Error of Approximation (RMSEA) < 0.08 ('may not be best suited to circumplex models. It becomes biased in the case of high correlations between proximal variables, as found in circumplex models.') Rogoza, Cieciuch, and Strus (2021)

Publishing with a paper

Data: The ISD on Zenodo

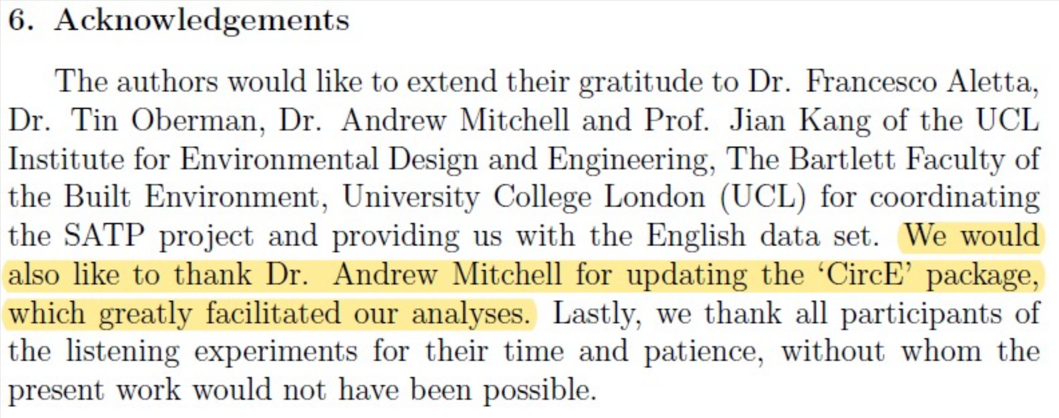
Code: Fully reproducible analysis

ARAUS Dataset analysis

Binder

Publishing a tool

Making code available on Github

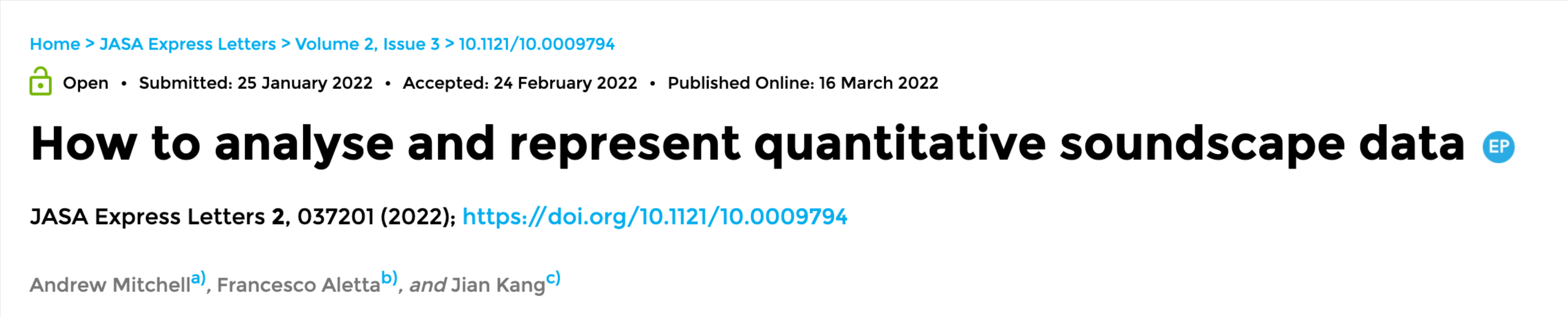


6. Acknowledgements

The authors would like to extend their gratitude to Dr. Francesco Aletta, Dr. Tin Oberman, Dr. Andrew Mitchell and Prof. Jian Kang of the UCL Institute for Environmental Design and Engineering, The Bartlett Faculty of the Built Environment, University College London (UCL) for coordinating the SATP project and providing us with the English data set. We would also like to thank Dr. Andrew Mitchell for updating the ‘CircE’ package, which greatly facilitated our analyses. Lastly, we thank all participants of the listening experiments for their time and patience, without whom the present work would not have been possible.

Publishing as a package

Soundscapy



Home > JASA Express Letters > Volume 2, Issue 3 > 10.1121/10.0009794

Open • Submitted: 25 January 2022 • Accepted: 24 February 2022 • Published Online: 16 March 2022

How to analyse and represent quantitative soundscape data

JASA Express Letters **2**, 037201 (2022); https://doi.org/10.1121/10.0009794

Andrew Mitchell[a], Francesco Aletta[b], *and* Jian Kang[c]

Sharing Data:	Zenodo
Sharing Notebooks:	Binder
Transition from SPSS to R:	StatsNotebook
Collaboration:	Github

Guidance and Tutorials

- The Turing Way
- Awesome Reproducible Research
- Research Software Engineering course
- Learning Statistics with R

Thank you for your attention!

The code used for demonstration was based on:

Mitchell, A., Aletta, F., & Kang, J. How to Analyse and Represent Quantitative Soundscape Data. *JASA Express Letters.* 2021. https://doi.org/10.1121/10.0009794

All of the data used is openly available at:

Mitchell, A., *et al.* The International Soundscape Database: An integrated multimedia database of urban soundscape surveys – questionnaires with acoustical and contextual information. *Zenodo [data set].* doi*:* 10.5281/zenodo.5654747

For more on me and my work, visit:

Website: https://andrew-mitchell.netlify.app/

And my podcast: https://www.justnoisepod.com/