Plan Overview

A Data Management Plan created using DMPonline

Title: Creating synthetic data for social science projects: Machine-learning and Agent-based modelling go head to head

Creator: Jools Kasmire

Principal Investigator: Jools Kasmire

Affiliation: University of Manchester

Template: University of Manchester Generic Template

ORCID iD: 0000-0003-2684-6330

Project abstract:

Synthetic data can be generated and used in many ways. This project compares machine learning (ML) and agent-based modelling (ABM) as synthetic data generation methods for exploring human behaviour. Importantly, this research interprets ML as a ‘top-down’ generation method and ABM as ‘bottom-up’. Logically, the systems or problems researchers might study can also be classified as top-down or bottom-up. This suggests that ML and ABM are appropriate for studying different systems, but matching the system to the generation method may not be the most important consideration.

First, the distinction between top-down and bottom-up, and how to recognise each, may not be well known. Second, the different methods have different logics, require different inputs, and need different resources. Third, different generation methods incur different computing costs, estimated here by the time needed to generate the data.

This project sets out to create one ML and one ABM synthetic dataset. Both aim to mimic a relatively simplistic behaviour (daily counts of cycle commutes) within a simplistic human population (as captured by 2 censors on one cycle lane in Manchester, UK) over several months in 2026. The ML will be trained on and attempt to recreate the cycle lane censor data directly. The ABM will simulate commuters based on the cycle lane data, weather data, and local commuting surveys.

Both datasets are expected to match in many ways (e.g. size, columns, data type, etc.). However, the different methods are expected to produce measurable differences in descriptive statistics and in time needed to generate the datasets. By exploring the similarities and differences in data produced by the methods, as well as the resources needed to use each method, this project will help researchers get a better idea of when each method is appropriate for their research contexts.

ID: 209146

Start date: 06-08-2026

End date: 31-12-2026

Last modified: 06-08-2026

Copyright information:

The above plan creator(s) have agreed that others may use as much of the text of this plan as they would like in their own plans, and customise it as necessary. You do not need to credit the creator(s) as the source of the language used, but using any of the plan's text does not imply that the creator(s) endorse, or have any relationship to, your project or proposal

Creating synthetic data for social science projects: Machine-learning and Agent-based modelling go head to head

Manchester Data Management Outline

1. Will this project be reviewed by any of the following bodies (please select all that apply)?

- None of the above

2. Is The University of Manchester collaborating with other institutions on this project?

- Yes - Part of a collaboration and owning or handling data

This project uses data obtained from Transport for Greater Manchester (TfGM).

Permission to use the data, to cite the data in publications and to store the data in public repositories was granted.

3. What data will you use in this project (please select all that apply)?

- Re-use existing data (please list below)

The data use comes from TfGM sensors that track the number of cycles and motor vehicles that pass northbound and southbound one point on Oxford Road in Manchester.

The data was already being collected by the TfGM for their own purposes. Most of the data is also made publicly available through social media bots that report the cycle lane use each day.

4. Where will the data be stored and backed-up during the project lifetime?

- Other storage system (please list below)

The data is quite small so will be stored on the PI's own university-issued machine until the research is ready to be presented publicly and/or published.
At that point, the data and all the research processes (written in .ipynb and .nlogo files), documentation, analysis and images will be uploaded to a public GitHub repo hosted by the UK Data Service (https://github.com/ukdataservice).
The GitHub repo will be used to create a stable version that will be uploaded to Zenodo to create a DOI so the research can be documented properly to support the presentation/publication.

5. If you will be using Research Data Storage, how much storage will you require?

- Not applicable

6. Are you going to be receiving data from, or sharing data with an external third party?

- Yes

The response to this question combines the responses to questions 2 and 3.
2. This project uses data obtained from Transport for Greater Manchester (TfGM).
Permission to use the data, to cite the data in publications and to store the data in public repositories was granted.
3. The data use comes from TfGM sensors that track the number of cycles and motor vehicles that pass northbound and southbound one point on Oxford Road in Manchester.

The data was already being collected by the TfGM for their own purposes. Most of the data is also made publicly available through social media bots that report the cycle lane use each day.

7. How long do you intend to keep your data for after the end of your project (in years)?

- 21+ years

Zenodo is intended to be a long term research archival system, so the DOI created there will remain as long as Zenodo continues to function.
Similarly, the project repository on the UKDS GitHub is will last at least 5 years (current funding cycle) but is likely to remain functional as long as the UKDS continues to receive funding (20+ years so far).

Guidance for questions 8 to 13

Highly restricted information defined in the Information security classification, ownership and secure information handling SOP is information that requires enhanced security as unauthorised disclosure could cause significant harm to individuals or to the University and its ambitions in respect of its purpose, vision and values. This could be: information that is subject to export controls; valuable intellectual property; security sensitive material or research in key industrial fields at particular risk of being targeted by foreign states. See more examples of highly restricted information .

If you are using 'Very Sensitive' information as defined by the Information Security Classification, Ownerships and Secure Information Handling SOP, please consult the Information Governance Office for guidance.

Personal information, also known as personal data, relates to identifiable living individuals. Personal data is classed as special category personal data if it includes any of the following types of information about an identifiable living individual: racial or ethnic origin; political opinions; religious or similar philosophical beliefs; trade union membership; genetic data; biometric data; health data; sexual life; sexual orientation.

Please note that in line with data protection law (the UK General Data Protection

Regulation and Data Protection Act 2018), personal information should only be stored in an identifiable form for as long as is necessary for the project; it should be pseudonymised (partially de-identified) and/or anonymised (completely de—identified) as soon as practically possible. You must obtain the appropriate ethical approval in order to use identifiable personal data.

8. What type of information will you be processing (please select all that apply)?

- No confidential or personal data

The data used is simple sensor readings intended to count the number of travelers that that move past a point on Oxford Road either in a motor vehicle on the main traffic lane or on a bike in a bike lane. There is one sensor for northbound traffic and another for southbound traffic.

The main data set contains simple counts for the total travelers per lane per direction. A second data set contains the average travelers expected per 15 minute time block throughout a 24 hour period. This is averaged over the life of the data collected.

Neither data set collects any personal or identifiable information.

9. How do you plan to store, protect and ensure confidentiality of any highly restricted data or personal data (please select all that apply)?

- Not applicable

10. If you are storing personal information (including contact details) will you need to keep it beyond the end of the project?

- Not applicable

11. Will the participants' information (personal and/or sensitive) be shared with or accessed by anyone outside of the University of Manchester?

- Not applicable

12. If you will be sharing personal information outside of the University of Manchester will the individual or organisation you are sharing with be outside the EEA?

- Not applicable

13. Are you planning to use the personal information for future purposes such as research?

- No

Not applicable. No personal information is collected in this project and so there is no possibility of using any for future research.

14. Will this project use innovative technologies to collect or process data?

- No

The data was already collected by established technologies. This project processes it using established technologies (basic pre-processing in python, machine learning models in python and agent-based models in NetLogo).

This is not an innovative project in terms of technologies.

15. Who will act as the data custodian for this study, and so be responsible for the information involved?

J. Kasmire, the PI

16. Please provide the date on which this plan was last reviewed (dd/mm/yyyy).

2026-08-06

Project details

What is the purpose of your research project?

The purpose is to establish how synthetic data created by machine learning models compares to synthetic data created by agent-based models.

Both models will seek to create synthetic datasets that mimic the original data. How well the 2 synthetic datasets mimic the original data as well as how much time was needed to generate each synthetic data set will be compared.

The project aims to provide advice to researchers that want to create synthetic data but are unsure of which method is best suited to their needs.

What policies and guidelines on data management, data sharing, and data security are relevant to your research project?

The basic terms of use granted by TfGM when they provided the data apply. The terms and conditions of GitHub, Zenodo, UKDS and Royal Statistical Society (where I intend to present the research) apply.

Responsibilities and Resources

Who will be responsible for data management?

J. Kasmire, the PI

What resources will you require to deliver your plan?

No additional resources. This is a small project, with small datasets and small computational resource requirements. My university-issued laptop is sufficient.

Data Collection

What data will you collect or create?

I will create 2 synthetic datasets, one created via machine learning models (in python) and one by agent-based models (in NetLogo).

How will the data be collected or created?

I will create 2 synthetic datasets, one created via machine learning models (in python) and one by agent-based models (in NetLogo).

Documentation and Metadata

What documentation and metadata will accompany the data?

The data will be cited as coming from the website which was sent to me by TfGM.The data was downloaded through simple browser interface, producing excel files. The metadata for the files was included within the file (e.g. the top 7 lines of data included the date it was downloaded, the website it came from, the name of the sensor it drew data from, etc.)

The original files, as downloaded from the website, are included in the project files to be shared on GitHub with permission from TfGM.

The entire process of separating the metada from the main data, as well as pre-processing the main data to check for missing values, errors, etc. is documented in extensively commented jupyter code notebooks. These are also included in the project files to be shared on GitHub.

The synthetic data files created will be accompanied by metadata files that detail the research project,

the origin of the real data, the synthetic generation process and links to all the tools used. The GitHub repo will contain requirement.txt files, licensing files, README files and all of the typical supporting documents needed for reproducible research projects.

Ethics and Legal Compliance

How will you manage any ethical issues?

I used the UoM decision tool to determine if there are any ethical issues and was told that there was no need to apply for an ethics review. I have screenshot the decision tool outcome and this will be included in the project files to be uploaded to GitHub.

No particular ethics issues are expected as the project is entirely desk-based research with no human participants or personal information.

Nevertheless, the project will be reviewed thoroughly prior to making the GitHub repository public (a precursor to creating a DOI and presentation/publication) in case any have arisen.

How will you manage copyright and Intellectual Property Rights (IPR) issues?

No particular issues are expected. There is nothing particularly to copyright or protect that is not part of the typical academic presentation/publication processes.

Thus, all research, software and packages/libraries that contributed to or are used within the project will be cited in the references section of the presentation/publication and all contributors will be listed as co-authors or in the acknowledgements as appropriate.

Storage and backup

How will the data be stored and backed up?

The data will be stored on the PI's university-issued laptop and backed up on OneDrive until it is ready to be made public. It will then be copied to the UKDS GitHub repository to be publicly accessible. It will remain on both for the foreseeable future (e.g. until the PI no longer uses OneDrive, until the UKDS stops using GitHub, etc.)

How will you manage access and security?

Access to my university-issued laptop and the OneDrive cloud storage backup is managed by the university log on processes. I may share some documents with colleagues (e.g. may ask a colleague to read my draft for publication, at which point I would use the University OneDrive sharing system.

Just prior to publication, I will copy all the relevant data, code notebooks, model code, metada, support

documents, etc. to the UKDS GitHub repository. It will be publicly visible or downloadable, but the public will not be able to edit it unless a pull request is approved (not expected to happen).

Selection and Preservation

Which data should be retained, shared, and/or preserved?

All data is expected to be retain, shared and preserved. It is very small. This is a tiny project in terms of data volume.

What is the long-term preservation plan for the dataset?

The long-term preservation plan is expected to be the UKDS GitHub repository, Zenodo (as a long term archive for the published research project) and potentially the UKDS ReShare repository (if there is enough interest to motivate its inclusion there).

This is not a big project and in not expected to have particularly long term impacts. However, buy making it publicly accessible, it is expected to be a useful resource for those who want to teach themselves to create synthetic data in multiple ways.

Data Sharing

How will you share the data?

The long-term sharing plan is expected to be the UKDS GitHub repository, Zenodo (as a long term archive for the published research project) and potentially the UKDS ReShare repository (if there is enough interest to motivate its inclusion there).

This is not a big project and in not expected to have particularly long term impacts. However, buy making it publicly accessible, it is expected to be a useful resource for those who want to teach themselves to create synthetic data in multiple ways.

Are any restrictions on data sharing required?

No restrictions are expected. The original data was given freely by TfGM and the synthetic data, methodology documents, code, models and metadata are expected to be shared freely as a potentially useful resources.