Plan Overview

A Data Management Plan created using DMPonline

Title: Real-Time Monitoring of Local Authority Socioeconomic Position and Deprivation Risk in the UK

Creator: Zak Plumridge

Principal Investigator: Robert Morris, Zak Plumridge

Data Manager: Robert Morris, Zak Plumridge

Project Administrator: Robert Morris, Zak Plumridge

Affiliation: University of Sussex

Template: Template for Wider Topics in Data Science

Project abstract:

Using passive data from internet connectivity and usage, energy consumption, transportation and social media, this five year project aims to build an interactive dashboard that provides a real-time probability of area level deprivation in the UK on behalf of the Ministry of Housing, Communities and Local Government.

Introducing higher velocity statistics into socioeconomical indicators can facilitate pro-active governmental policy making, improve policy targeting and evaluation and increase response times to economic shocks and crises.

ID: 199197

Start date: 01-06-2026

End date: 01-06-2031

Last modified: 30-04-2026

Real-Time Monitoring of Local Authority Socioeconomic Position and Deprivation Risk in the UK

Data Formats

Describe the data that you will use, including key contextual information, the types of data and its estimated volume. What file formats will be used for data collection and processing, and why are these formats appropriate?

This project will use a combination of open and licensed behavioural datasets to model deprivation risk at small-area level, initially in London and later it will scale to a national level. The core data will include Index of Multiple Deprivation, LSOA boundary and lookup files, Tfl usage and electricity consumption statistics. Where agreements permit, aggregated mobile activity and social media data will be included. Most data will be numerical, tabular and geospatial. Some metadata and methodological documentation will be textual.

Raw data will be collected from official repositories (IMD from UK Government, TfL entry and exit data, and energy consumption from Ofgem) it will be cleaned, standardised, then temporarily aligned and linked geographically through LSOA or local authority. The estimated volume for the initial stage is likely be tens of gigabytes, increasing if high frequency behavioural data is acquired. A structured folder system will separate the raw, process, scripts, models, outputs and documentation files.

Preferred formats will be CSV, TXT/MarkDown for README, PDF for documentation. GeoJSON for its spatial compatibility. These formats are widely supported, suitable for long term access and interoperable in R/Python. Metadata will include collection dates, geographic coverage, variable definitions, licence and quality notes.

The main risks to the data are inconsistent or changing geographical regions, missing or incomplete historical data and temporal misalignment between datasets. We will ensure that we mitigate this by maintaining all backup copies of deposited open data, versioning all source extracts and using geographic lookup tables to reconcile spatial units across datasets.

Data repository

Where will the data and outputs be stored and backed up during and after the project, and why are the chosen repositories or storage solutions appropriate?

Throughout the project, the data and documentation will be stored in secured University managed storage (e.g. cloud or research data storage services) with automated backups happening routinely to minimise the risk of data loss. Access will be shared only with members of the research team. All the code, scripts and version controlled analytical workflows will be maintained in a private GitHub repository during the development.

Once the project is complete, the datasets and associated documents will be uploaded to Sussex Figshare, the University of Sussex’s institutional repository. This platform allows for long term

preservation, enabling discoverability and reuse. Code and software outputs will be archived via GitHub.

These storage choices are appropriate because they align with FAIR (findability, accessibility, interoperability, reuse of digital assets) data principles. Sussex Figshare with its support for access controls and structured metadata makes it particularly suitable. GitHub supports collaborative development and version controls.

Restricted or licensed data that cannot legally be redistributed will not be deposited in open repositories. In this case, metadata describing the datasets and their conditions will be provided. Subject to third party agreements, access can be facilitated through controlled or secure data services.

Data quality assurance

Describe the procedures for quality assurance that will be carried out on the data during data collection, data entry, digitisation, processing, and data checking.

Quality assurance will be checked through the entire data lifecycle with these checks tailored to integrating heterogeneous, secondary datasets. As data is collected from all sources (such as IMD, Tfl, energy data) it will be assessed for provenance and their limitations using official documentation. Only reputable sources will be prioritised.

During processing, data will be checked for its consistency over time and ensuring that the LSOA codes align with the correct boundary year. Missing data will be systematically identified and will be resolved by exclusion or aggregation. If possible, data may be imputed but would be clearly flagged if this is the case and it would have to be of increased benefit to the dataset.

Prior to analysis, exploratory data analysis (EDA), outlier detection and cross source comparison will be performed for validation. Reproducible data processing will ensure that there are minimised manual errors.

Key risks include spatial boundary changes and inconsistent temporal granularity. This may affect the model validity and comparability across areas. Any limitations will be documented. Standardised lookup tables and temporal aggregation to common intervals will be utilised to mitigate the risks.

Data protection

What ethical, legal, and data protection considerations apply to your data, and how will these be addressed across the entire data lifecycle?

The project will primarily use non personal datasets. However, mobile activity data could possibly lead to indirect identification depending on its granularity and therefore this would be treated as potentially

sensitive.

All data handling will comply with the UK General Data Protection Regulation (GDPR) and the Data Protection Act 2018. Data providers will therefore be required to have legal authority to share its data and the usage of it will comply with their license agreements where applicable. Any third party data will only be used within the specified conditions whether that is research only use or no re-identification, for example.

Key risks include re-identification through collated datasets that offer greater detail and/or specificity. These risks will be mitigated through anonymisation prior to storage where necessary and no attempts will be made to deduce individual behaviour.

All procedures will align with the University’s ethical guidelines and data protection policies ensuring risks are minimised at each stage of the data lifecycle.

Data access

Who will be able to access the data and how will access be granted and controlled during and after the project?

Access throughout the project will be restricted to the research team members and where required authorised supervisors. Access will be granted using the secure university managed storage systems such as OneDrive with authentication controls. The private GitHub repository will also be restricted to the research team.

Access to restricted data will only be available through a formal request process, requiring users to specify the purpose of use and agree to conditions such as non-commercial use and prohibition of re-identification attempts will be put in place. Data transfer may be limited to secure environments rather than direct download.

All users will have to agree to the terms of use and must ensure proper citation, ethical use and compliance set out in the licensing conditions. Downloading of open datasets will be permitted.

These measures allow access to open datasets whilst still ensuring compliance with ethical and legal obligations.

Data sharing and reproducibility

Which data arising from your project will be shared, where and when it will be shared, and what ethical, legal, commercial, or intellectual property considerations will influence decisions about sharing? What data, code, and documentation will be made available to support verification, reproducibility, and reuse of your research findings?

The project will aim to share as much data as possible in line with restrictions, licensing, or commercial

terms. The following items will be made openly available: non-sensitive cleaned datasets that have been created using publicly accessible data, deprivation risk factors at an appropriate level of aggregation, metadata, data dictionaries, codebooks, README files, documentation of the workflow, visualisation scripts and analysis code that has been committed to version control. Such information will be deposited in Sussex Figshare and code will be stored in GitHub and archived at release.

Third party restricted data, such as licensed behavioural data sets, will not be shared unless explicitly permitted to be shared under license conditions and when disclosure risks are sufficiently low. If the latter holds true, it would only include aggregated or anonymised forms of the dataset, along with an explanation as to the conditions of access. Metadata records for such data will also be made available.

To support reproducibility, the following items will be shared to be able to replicate the analysis including scripts used for data acquisition and cleaning, scripts for creating model-ready datasets, notebooks, package requirements, software versions, parameters used and workflow documentation. Shared data will be preserved in CSV format, and documentation provided for the variables, units, abbreviations used, time periods covered, and geographic linkages.

Data underpinning reported findings will be accessible either directly through repository deposit or indirectly through metadata and instructions where source restrictions apply. Limitations to reproducibility is that some licensed datasets may not be redistributable; this will be stated transparently in all outputs.

Data preservation

What will happen to the data at the end of the project? What arrangements will be put in place to document, preserve, and curate the data and outputs so that they remain accessible, understandable, and usable in the long term after the project ends?

At the end of the project, appraisal of the data and outputs will take place with respect to their long term research, education and policy use. Those that have lasting value will be preserved in the form of cleaned publicly sourced datasets, generated non-sensitive analytical tables, metadata, documentation, source code, model specifications and final outputs. These materials are valuable because they support future methodological work on deprivation monitoring and potential policy reuse.

Preserved data will be deposited into Sussex Figshare, ensuring long term accessibility. This data will be stored in open and preservation friendly formats such as CSV, JSON and PDF in order to mitigate risks of software obsolescence and ensure accessibility.

To ensure long term usability, comprehensive documentation will be included with all datasets such as README files, data dictionaries and workflow documentation (code notebooks and version histories).

Data will be retained for a minimum of 10 years in line with best practices for research. Key risks to preservation include data loss, repository dependency and format obsolescence which will be mitigated by the use of established repositories and adherence to open standards.

Intellectual property (IP)

Who will own the copyright and intellectual property rights of data, software, or materials arising from the project and what are the terms of use for the data you are re-using? How will these rights be managed to enable appropriate reuse, publication, and protection of any commercially or scientifically valuable outputs?

Intellectual property for this project will primarily be the processed datasets, data processing pipelines and frameworks for deprivation risk prediction. These outputs have value in both academic research and public policy applications, particularly for local authorities.

Copyright of newly created materials (e.g. cleaned datasets and code) will be held by the researchers, subject to University of Sussex policies. Whilst ownership of reused data will remain with the original providers, subject to their licensing agreements. Any third party data reuse will comply with the original licensing terms.

The project will aim to have an open approach to sharing. However, this will be balanced against intellectual property and licensing constraints. Where feasible, source code and non-confidential results will be published under open license agreements. Though some data such as derived variables or modeling techniques with possible business implications might need to be documented but not shared to protect their potential applications.

Overall, Intellectual Property will be managed to balance legal compliance and potential future value. The outputs would usually be released once the results have been obtained, ensuring that the academic benefits are achieved without limiting future use.

Use of Artificial Intelligence (AI) tools

Will artificial intelligence (AI) tools be used in the project? If so, which tools will be used, for what specific tasks, and what measures will be taken to document their use, manage outputs, and ensure transparency and reproducibility in line with institutional guidance on responsible research practice?

Artificial intelligence tools may be utilised in a selective and supportive capacity throughout the project. All outputs from the use of AI applications will be evaluated critically by the research team before implementation. Suggestions for coding will be tested and validated within the analytical pipeline and all textual outputs will be checked for authenticity, coherence and originality. AI tools will not be used to generate final results or conclusions without human verification.

Use of AI tools will be documented through the course of the research, this includes the type of tool used and the stage of the workflow in which it was applied. This ensures transparency and allows others to understand how outputs were produced.

Potential risks include bias in AI generated suggestions, lack of transparency in model outputs and the possibility of inaccurate or misleading information. Particular care will also be taken to avoid the inclusion of sensitive or potentially identifiable data despite the focus on aggregated datasets. These

risks would be mitigated through critical evaluation, reliance on authoritative sources, avoidance of sensitive data in external platforms and only anonymised data will be used in the process.

Use of AI will comply with the University’s guidance on academic integrity and responsible research practice. The research team will retain full responsibility for all outputs.