Plan Overview

A Data Management Plan created using DMPonline

Title: Population Reconstruction by age, sex and education attainment

Creator: Felipe Sanchez

Principal Investigator: Felipe Sanchez Segura

Affiliation: University of Manchester

Funder: Economic and Social Research Council (ESRC)

Template: ESRC Template Customised By: University of Manchester

ORCID iD: 0000-0002-3272-5829

Project abstract:

This project reconstructs the population of Colombia from 1998 to 2018 using administrative registers, censuses, and Life Quality Surveys. The data are disaggregated annually by age (in 5-year groups), sex, and education level (based on ISCED 2011).

ID: 180091

Start date: 01-09-2021

End date: 29-07-2025

Last modified: 23-06-2025

Grant number / URL: ES/P000665/1.

Population Reconstruction by age, sex and education attainment

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?

- No – only institution involved

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

- Re-use existing data (please list below)

Colombian censuses of 1993, 2005 and 2018. Life Quality Surveys from 1998 to 2018.

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

- P Drive (postgraduate researchers and students only)

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?

- No

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

- 11 - 20 years

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)?

- Anonymised personal data

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

- Store data on University of Manchester approved and securely backed up servers or computers

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

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

- Yes, and innovative technologies will collect or process personal data (please list the innovative technologies below)

Manuscript in preparation about the methodology. But data needs as an input of other project.

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

Myself

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

2025-06-12

Assessment of existing data

Provide an explanation of the existing data sources that will be used by the research project, with references

1. Population and Housing Censuses (1993, 2005, and 2018)

The project relies on data from the **1993, 2005, and 2018 Colombian Population and Housing Censuses** , conducted by the National Administrative Department of Statistics (DANE). These censuses provide nationally representative data on the age and sex structure of the population and levels of educational attainment. The 1993 census serves as a historical benchmark for pre-2000 estimation, while the 2005 and 2018 censuses inform the mid- and end-points of the reconstruction.

Reference: DANE (1993, 2005, 2018). *Censo General de Población y Vivienda*. Bogotá: Departamento Administrativo Nacional de Estadística (DANE).

https://www.dane.gov.co

2. Life Quality Surveys (Encuesta de Calidad de Vida – ECV)

The **Life Quality Survey (ECV)** is a nationally representative annual household survey carried out by DANE. Data from the years 2008 to 2018 are used to estimate the distribution of educational attainment across age and sex groups. The ECV captures self-reported education, which is harmonised and classified according to ISCED 2011 standards.

Reference: DANE (2008–2018). *Encuesta de Calidad de Vida (ECV)*. Bogotá: DANE.
https://microdatos.dane.gov.co

Provide an analysis of the gaps identified between the currently available and required data for the research

There is a significant data gap in Colombia regarding population estimates disaggregated simultaneously by age, sex, and educational attainment for the period 1998 to 2018. Official population data are available for: • Age and sex annually via DANE projections, but not linked to education. • Age, sex, and education only in census years (1993, 2005, and 2018). • Educational distributions can be partially inferred from Life Quality Surveys (ECV) since 2008, but these are sample-based, not comprehensive administrative counts. As such, no official or complete dataset exists that provides annual joint distributions of Colombia's population by age, sex, and education for the full period. This limits the ability to conduct demographic or socioeconomic analyses—such as fertility by education level—that require consistent time series data. The gap is addressed in this project through statistical reconstruction using harmonised census data and interpolated education distributions derived from survey data. This allows for consistent annual estimates suitable for research and policy modelling.

Information on new data

Provide information on the data that will be produced or accessed by the research project

This project will produce a harmonised, annual dataset of the Colombian population from 1998 to 2018, disaggregated by: • Year (calendar years 1998–2018) • Sex (male, female) • Age (5-year age groups, from 0–4 to 100+) • Educational attainment, harmonised to ISCED 2011 categories: • ISCED 0: No education • ISCED 1: Primary • ISCED 2–3: Secondary • ISCED 4+: Post-secondary/tertiary Each row in the dataset corresponds to a population estimate for a unique combination of year × age × sex × education level. The final data format is long (tidy) format.

Quality assurance of data

Describe the procedures for quality assurance that will be carried out on the data collected at the time of data collection, data entry, digitisation and data checking.

The dataset produced in this research project is based entirely on secondary data sources, primarily population and housing censuses and household surveys conducted by the Colombian National Administrative Department of Statistics (DANE). As such, initial quality assurance at the point of data collection is the responsibility of DANE. These procedures include rigorous protocols for census enumeration (in 1993, 2005, and 2018) and probabilistic sampling, questionnaire design, and interviewer training for the Life Quality Surveys (ECV) between 2008 and 2018. DANE applies standard validation, editing, and imputation rules, which are documented in their technical reports and metadata releases.

Following acquisition, no manual data entry is conducted in this project. Instead, raw datasets are accessed via official repositories and imported using reproducible code written in R. File integrity is checked using metadata-provided hashes (where available), and all scripts make use of structured import functions such as readxl, readr, and haven. These routines eliminate risks associated with

manual processing and ensure transparency in the data pipeline. Version control is maintained using Git to track all transformations.
During data processing, several quality assurance procedures are implemented to ensure consistency and correctness. These include the detection and removal of logically implausible records—such as education reported for children under five—and the harmonisation of educational attainment into internationally comparable ISCED 2011 categories. Cross-tabulations and frequency checks are used to monitor the integrity of key variables such as age, sex, and education. Interpolation between census years is conducted using spline methods, and intermediate results are assessed through convergence diagnostics and graphical validation.
Final quality checks include comparisons between reconstructed population aggregates and official national totals published by DANE, as well as the visual inspection of trends through population pyramids and time-series plots. These graphical outputs are used to identify outliers, abrupt changes, or structural inconsistencies. Additionally, peer review from academic supervisors is integrated into the validation process prior to data publication.

Backup and security of data

Describe the data security and backup procedures you will adopt to ensure the data and metadata are securely stored during the lifetime of the project.

All data processing and storage during the project will comply with the University of Manchester’s Research Data Management (RDM) policy and IT security guidelines. The working datasets, code, and metadata are stored on the University-managed OneDrive for Business cloud platform, which provides encrypted, access-controlled storage with daily automated backups.

Management and curation of data

Outline your plans for preparing, organising and documenting data.

The dataset will be prepared using a fully reproducible workflow in the R programming language. Raw data from the Colombian censuses and Life Quality Surveys (ECV) will be imported and harmonised using scripted procedures, ensuring transparency and replicability. Data will be organised in a long (tidy) format, where each row represents a unique combination of year, age group, sex, and educational attainment. Variables will be clearly named using consistent conventions and labelled using the labelled package in R to support future export to Stata or SPSS. A comprehensive README.md file will accompany the dataset, providing an overview of the project, the structure of the data files, and descriptions of all variables and classification schemes used (including ISCED mappings). In addition, a data dictionary will be included to define variable names, types, units, and permitted values. Processing scripts will be commented and version-controlled using Git. All files will be stored in a structured folder system with subdirectories for raw data, scripts, outputs, and documentation, following best practices in data organisation.

Difficulties in data sharing and measures to overcome these

Identify any potential obstacles to sharing your data, explain which and the possible measures you can apply to overcome these.

NA

Consent, anonymisation and strategies to enable further re-use of data

Make explicit mention of the planned procedures to handle consent for data sharing for data obtained from human participants, and/or how to anonymise data, to make sure that data can be made available and accessible for future scientific research.

NA

Copyright and intellectual property ownership

State who will own the copyright and IPR of any new data that you will generate.

The copyright and intellectual property rights (IPR) for the new dataset generated by this project will be retained by the researcher (myself), as permitted under the University of Manchester's intellectual property policy for postgraduate research students. The dataset is based on publicly available secondary sources and was independently developed by the researcher during the course of their doctoral studies. It will be made available under a Creative Commons Attribution 4.0 International (CC BY 4.0) license, allowing others to freely use, adapt, and redistribute the data with appropriate attribution. All associated code and documentation will be released under an open-source license (e.g. MIT or GPL), supporting transparency, reproducibility, and reuse in demographic and policy research.

Responsibilities

Outline responsibilities for data management within research teams at all partner institutions

This project is conducted as part of a doctoral research programme at the University of Manchester. As such, the responsibility for day-to-day data management including data cleaning, harmonisation, documentation, version control, and secure storage rests with the researcher.

Preparation of data for sharing and archiving

Are the plans for preparing and documenting data for sharing and archiving with the UK Data Service appropriate?

Yes, the plans for preparing and documenting the data are fully aligned with the standards and requirements of the UK Data Service. The dataset will be structured in a long (tidy) format, with clear variable naming and consistent coding across years, ensuring that it is machine-readable and suitable for reuse. Variables will be labelled using controlled vocabularies where applicable, such as ISCED 2011 for educational attainment, and the dataset will be accompanied by comprehensive documentation. This includes a README.md file describing the file structure, processing steps, and data sources, as well as a data dictionary providing definitions, formats, and permitted values for each variable.

Is there evidence that data will be well documented during research to provide highquality contextual information and/or structured metadata for secondary users?

Yes, the dataset will be thoroughly documented throughout the research process to ensure it is suitable for secondary use. Documentation will include both human-readable and machine-readable components. A structured README.md file will accompany the dataset, providing contextual information on the purpose of the data, data sources, processing methods, variable definitions, and known limitations.

Planned Research Outputs

Dataset - "Annual Population Reconstruction for Colombia by Age, Sex, and Education, 1998–2018"

Planned research output details

Title	DOI	Type	Release date	Access level	Repository(ies)	File size	License	Metadata standard(s)	May contain sensitive data?	May contain PII?
Annual Population Reconstruction for Colombia by A ...	10.48420/29309810 ...	Dataset	2025-06-12	Open	None specified	111,849	Creative Commons Attribution 4.0 International	None specified	No	No