DMP title

Project Name Schoolprobit

Principal Investigator / **Researcher** merralja@mcmaster.ca

Project Data Contact paezha@mcmaster.ca

Description This project will use RAHB data, public and separate school GIS data, and other data such as census, to study the effect of schools on house prices in a hedonic regression framework.

Institution McMaster University

Data Collection

What types of data will you collect, create, link to, acquire and/or record?

Secondary data used will include:

1. RAHB home sales data in .csv spreadsheet format, with co-ordinates sufficient to generate attributed sf point files;

2. Statscan census data (downloaded via CHESS) in .csv format, and Statscan GIS shapefiles for CTs and DAs

3. EQAO data provided from MoE in .csv format, with .xls descriptor files

4. Spatial and other data from HWDSB and HWCDSB for school catchments and point locations

5. possible inclusion of surfaces generated by Chris Higgins in his previous work, representing pollution, road network accessibility or neighbourhood development types, in an ArcGIS format

6. background GIS data as necessary

7. verification notes on schools, written in .xls and .doc formats, with html links or source citations for the primary documents found online

The primary data generated will include:

1. data analysis and regression in R code

2. any intermediate or refined data produced by R code, saved as .Rdata in sf format, only in those instances where we have reproduction rights

What file formats will your data be collected in? Will these formats allow for data re-use, sharing and long-term access to the data?

Statscan - .csv. Is an open format, and can be reconstructed from the ground up.

GIS - ArcGIS file format. Presently supported in R package sf. The sf data format allows for reconstruction from the ground up, as it's essentially as .csv with geometry.

EQAO - data comes in .csv format, with .xls data descriptors. This dataset comes with restrictions from the Ministry of Education, and cannot be shared.

RAHB - is provided in .xlsx format. Can be converted to .csv if needed. .xlsx format is supported by open programs like LibreOffice. This data set is proprietary and shall not be shared, or insecurely stored, so our RAHB dataset will not be available for others in any case.

notes - are generally being saved in Word 97-2003 .doc format. This format is supported by open programs like LibreOffice.

R code - will be generated in RStudio.

What conventions and procedures will you use to structure, name and version-control your files to help you and others better understand how your data are organized?

The file structure will follow the R package format as presented in Session 7 off the lectures for this class. The /raw folder will contain all data that is not shareable; it will be divided into subdirectories for each primary data source, divided further into subdirectories for as-received and processed, and divided further into years. The R package documentation will have sufficient information to understand data organization.

Intermediate versions of school catchment data are unnecessary to save locally; Github version control will take care of this.

R code may need to be divided into subprograms, e.g. for initial data preparation and for statistical analysis. Again, Github can handle version control.

Documentation and Metadata

What documentation will be needed for the data to be read and interpreted correctly in the future?

Code documentation and data structure documentation will be handled within an R package, as described in Session 7 of the lectures.

Calculations will be performed in R, and that code will be well-commented. ArcGIS data comes with its own metadata in ArcGIS, including co-ordinate system.

In particular, documentation will be provided on who created the data and when, how and where the data originate, who to contact to access the non-shared data, explanation of data coding and analysis performed, notes on quality and accuracy, expanation of restrictions on the use of the data, and details of who has worked on the project and performed each task.

Note that the secondary data is for the most part unshareable, and that which is shareable (i.e. Statscan) is also easily accessible to other researchers as is; instead of sharing this data, I will simply provide information on who to contact or how to access this data.

How will you make sure that documentation is created or captured consistently throughout your project?

Documentation will need to be updated at regular intervals as analysis continues.

If you are using a metadata standard and/or tools to document and describe your data, please list here.

Data will be described in R code, so that it can be found independently by the reader

and converted to the proper format.

Storage and Backup

What are the anticipated storage requirements for your project, in terms of storage space (in megabytes, gigabytes, terabytes, etc.) and the length of time you will be storing it?

The project will not require storage beyond a couple gigabytes. RAHB and GIS data will be the largest part of storage requirements. After completion of the project, some use agreements require that the data be destroyed; secondary source data will not be shared with the package without explicit agreement from the owner.

How and where will your data be stored and backed up during your research project?

Primary work is being performed on JM's home computer; regular updates are being uploaded to Github via desktop client, whenever alterations to code or dataset are made. This allows for 2 code copies: 1 local and 1 remote. This is sufficient.

Data storage of cleaned and transformed data uses Github, with private directories; RAHB data cannot be stored on Github, but the primary copy resides with Chris Higgins. Other secondary data resides in original format with the outside agency that developed it (EQAO, Statscan), and thus a third copy exists in case of catastrophic data loss.

How will the research team and other collaborators access, modify, and contribute data throughout the project?

Github collaborator settings allow for access and contribution by the other team members, either by cloning, forking, or transfer of ownership.

Preservation

Where will you deposit your data for long-term preservation and access at the end of your research project?

The final code will be deposited on Github. Third-party secondary data will not be shared.

Documentation for verification and extension of HWDSB and HWCDSB point, catchment, and school data can be provided within the Github repository. I will also want to consult with the McMaster Map Library to see whether the original school data files, and/or my derived data files, may be added to the digital collection there: they might be able to help me negotiate this with the school boards.

Indicate how you will ensure your data is preservation ready. Consider preservation-friendly file formats, ensuring file integrity, anonymization and de-identification, inclusion of supporting documentation.

I am going to assume that if data is presently readable with an open-source program, it can be reconstructed in the future.

Otherwise, the majority of my data is unshareable and accessible (in theory) through

outside agencies. A copy of the HWDSB/HWCDSB data will be kept in a private directory on Github, since the school boards have no policy for archiving old catchment data.

Sharing and Reuse

What data will you be sharing and in what form? (e.g. raw, processed, analyzed, final).

The R code developed for the project, and the final report, will be shared; this obviously must include the graphs, charts and maps in the final report which are derived from the data. The data from outside agencies will not be shared due to restrictions.

Notes for development of the data, including the extensive notes on school openings and closures, will be shared in the form of long .doc documents on schools, with attribution.

Have you considered what type of end-user license to include with your data?

End-user license for the code has not yet been considered. I would like to identify a license that is explicitly academic-only, not just non-commercial; and that allows simply for download and viewing, not further application.

All secondary data being used comes with restrictions, so licenseability resides with the originating agency.

What steps will be taken to help the research community know that your data exists?

There is no primary dataset to share.

I will be considering liaising with the McMaster map library to deposit the school catchment data for later use by other researchers, if this can be done with little effort on my part.

Responsibilities and Resources

Identify who will be responsible for managing this project's data during and after the project and the major data management tasks for which they will be responsible.

I will be collecting data and programming code and managing the repository. After completion, the work will remain on Github, and will be available for the rest of the project team to clone or fork.

How will responsibilities for managing data activities be handled if substantive changes happen in the personnel overseeing the project's data, including a change of Principal Investigator?

Data activities for this thesis will cease and data will remain on Github. The supervisors have access to the Github repository and can inherit by forking if they wish.

What resources will you require to implement your data management plan? What

What resources will you require to implement your data management plan? What do you estimate the overall cost for data management to be?

This is a funding question and does not apply to theses.

Ethics and Legal Compliance

If your research project includes sensitive data, how will you ensure that it is securely managed and accessible only to approved members of the project?

The following are the only two sets of sensitive data:

1. RAHB data will be stored securely and not shared. The researchers have no authorization to share this data. Copies exist on private computers and can not be publicly posted.

2. EQAO data comes with conditions. It is stored in private repositories on Github.

There is no ethics approval required for this project.

If applicable, what strategies will you undertake to address secondary uses of sensitive data?

Not applicable.

How will you manage legal, ethical, and intellectual property issues?

These issues will be managed following whatever framework has been developed by McMaster University.

This document was generated by DMP Assistant (https://assistant.portagenetwork.ca)