Open Software - Restricted Data: The Suicide/Climate Case

Ivan Hanigan[1], David Fisher[2], Steven McEachern[3]

Restricted Data is Safe

Restrictions on access to confidential health data have increased recently. Enabling safe access to data and analytic software is needed to address the **Replicability Crisis** (Peng 2011). We present an environment for analysing restricted data using open software. The system is described in Figure 1 and a Case Study of the historical association of suicides with climate (and extrapolation of this under climate change/adaptation scenarios) in the bottom half of the poster. These tools allow users to access restricted data; protects confidentiality and allows use of open software for reproducibility (King 1995).

Restrictive IT Environments

Previous solutions to this challenge make access so restricted that usability is compromised. We aimed to build a collection of tools for the conduct of many types of health and social science research. The starting point for users is the data catalogue, which provides for finding data available from the store of Restricted data and Less Restricted data for approved use. Once data are discovered, the researcher has capacity to manipulate the datasets on the secure server. The PostgreSQL database integrates and Geoserver visualises, while statistical tools are available in the R-studio server browser.

Server-Client Architecture

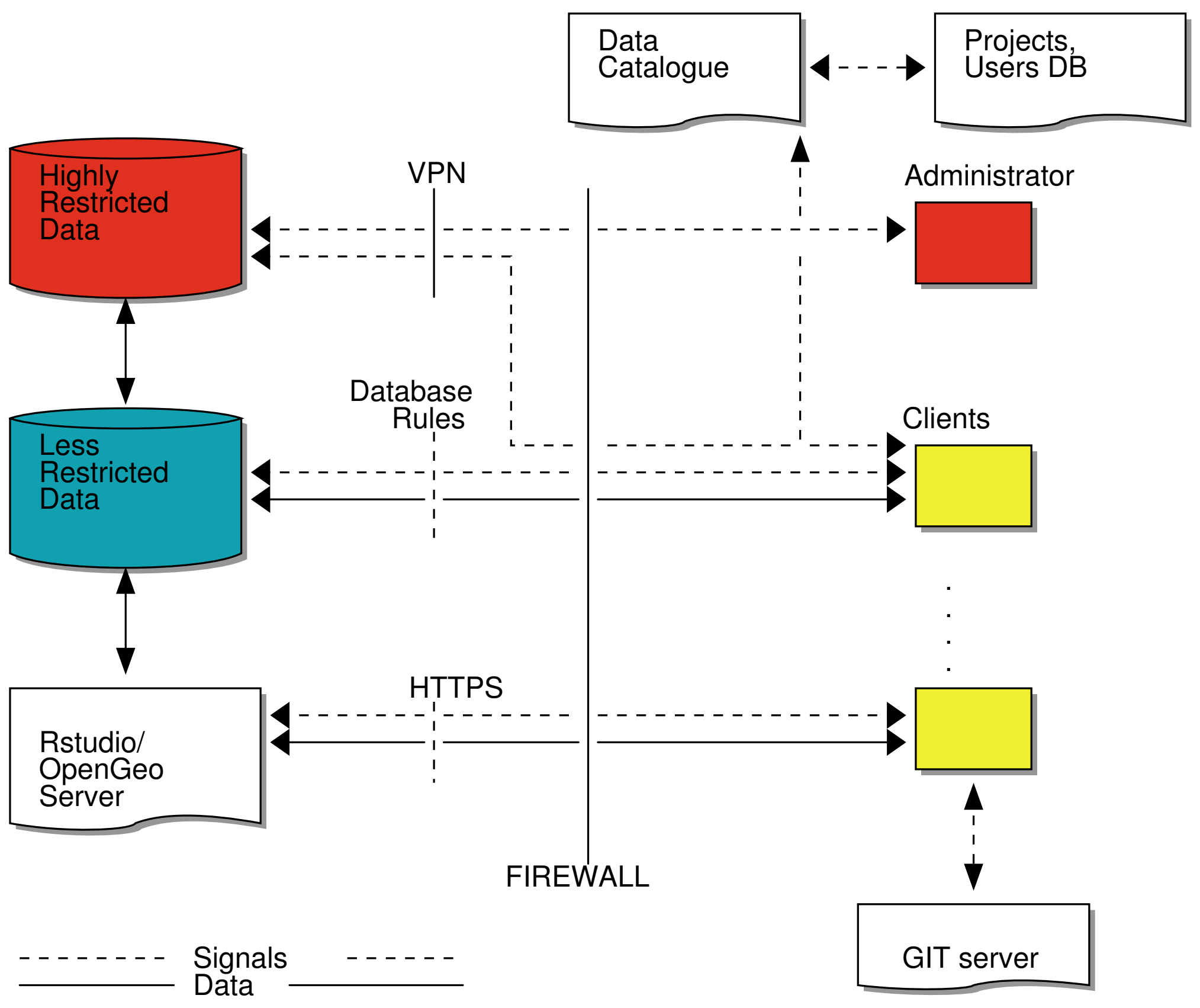

Figure: 1. System Design

The Stack

Hardware:

- National Research Cloud *http://www.nectar.org.au/research-cloud/*
- Centos 6.4 *www.centos.org/*

Database (The Brawn):

- PostgreSQL 9.2 *http://www.postgresql.org/*
- PostGIS 2.0 *http://postgis.refractions.net/*

Analysis (The Brains):

- R language for statistical computing *http://www.r-project.org/*
- Rstudio server *www.rstudio.com/*
- OpenGeo Suite *http://opengeo.org/*

Information Management:

- Projects,UsersDB Oracle XE APEX *www.oracle.com*
- Data Catalogue *http://assda.anu.edu.au/ddiindex.html*

The Client Side:

- The Kepler Project *https://kepler-project.org/*
- pgAdmin *www.pgadmin.org/*
- Git Version Control and GitHub *https://github.com/*

Reproducibility

Such analytical tools will enhance the ability of adaptive management practitioners to assess the potential influence of adaptations. The use of the system shows the ease with which multiple data sources (some restricted) can be analysed in a secure way using open software. This will build capacity to answer complex research questions and compare multiple climate change scenarios or adaptation assumptions; achieving simultaneous vision of potential future outcomes from different standpoints.

How it works

- Open a web browser / log on to the catalogue / find data
- In the web browser log on to Rstudio server
- Connect to database / query datasets / join / subset / transform
- Get data to your Rstudio server workspace and analyse
- Commit all code to GitHub, download resulting dataset and reports

Summary

This system:

- Assists rigorous data management practices
- Enables multiple data sources (some restricted) to be analysed
- Storage of data is secure
- Analytic code is made available as open software
- Enhances reproducibility

A Case Study

Exposure/Response function

Historical association between Suicide and Climate Variables were established in a Poisson time-series model (Hanigan et al 2012) using:

- Restricted Health and Drought data and
- Less Restricted Population data

(Colours refer to data storage and access rules shown in Figure 1).

$$\begin{aligned} log(O_{ijk}) \; = \; & s(ExposureVariable) + OtherExplanators \\ & + AgeGroup_i + Sex_j \\ & + SpatialZone_k \\ & + sin(Time \times 2 \times \pi) + cos(Time \times 2 \times \pi) \\ & + Trend \\ & + offset(log(Pop_{ijk})) \end{aligned}$$

Where:
O_{ijk} = Outcome (counts) by Age$_i$, Sex$_j$ and SpatialZone$_k$
ExposureVariable = Data with Restrictive Intellectual Property (IP)
OtherExplanators = Other Less Restricted Explanatory variables
s( ) = penalized regression splines
$SpatialZone_k$ = Less Restricted data representing the $SpatialZone_k$
Trend = Longterm smooth trend(s)
Pop_{ijk} = interpolated Census populations, by time in each group

Climate Change Scenarios

We can use methods like Bambrick et al 2008 to estimate Climate Change Health Impacts:

$$Y_{ijk} = \sum_{lm} (e^{(\beta_{ijk} \times X_{lm})} - 1) \times BaselineRate_{jkl} \times Population_{jklm}$$

Where:
β_{ijk} = the ExposureVariable coefficient for zone$_i$, age$_j$ and sex$_k$
X_{lm} = Projected Future ExposureVariables with Restrictive IP
BaselineRate$_{jkl}$ = avgDeathsPerTime/avgPopPerTime in age$_j$, sex$_k$ and zone$_l$
Population$_{jklm}$ = projected populations by age$_j$, sex$_k$, zone$_l$ and time$_m$ (With Less Restrictions)

Suicide and Temperature

The association of suicide and maximum temperature anomalies is shown in Figure 2. This can be used to estimate future climate impacts using future climate scenarios (Data with Restrictive IP) and population at risk (Less Restricted data).

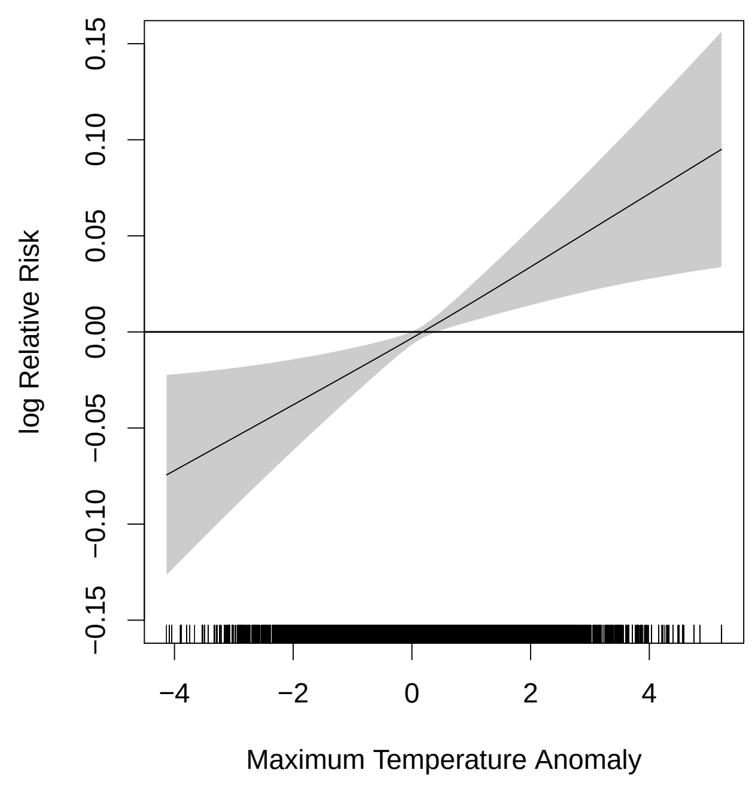

Figure: 2. Suicide/Temperature

Conclusions

This system:

- Enables data analysis in a safe environment
- Allows comparison of multiple climate scenarios and assumptions
- Demonstrated with a Climate/Health Impact Assessment
- And this is Reproducible

References

Roger D. Peng.
Reproducible research in computational science.
Science, 334(6060), December 2011.

Gary King.
Replication, replication.
Political Science and Politics, 28(3), September 1995.

Ivan C. Hanigan, Colin D. Butler, Phillip N. Kokic, and Michael F. Hutchinson.
Suicide and drought in New South Wales, Australia, 1970-2007.
Proceedings of the National Academy of Sciences, 109(35), August 2012.

Hilary J. Bambrick, Keith B.G. Dear, Rosalie E. Woodruff, Ivan C. Hanigan, and Anthony J. McMichael.
The impacts of climate change on three health outcomes: temperature-related mortality and hospitalisations, salmonellosis and other bacterial gastroenteritis, and population at risk from dengue.
Technical report, Garnaut Climate Change Review, Canberra, 2008.

Acknowledgements

- [1] National Centre for Epidemiology and Population Health, ANU
- [2] Information Technology Services, ANU
- [3] Australian Data Archive, ANU

Australian National University

ands AUSTRALIAN NATIONAL DATA SERVICE

Australian Government
Department of Industry Innovation, Science, Research and Tertiary Education

This project is supported by the Australian National Data Service through the National Collaborative Research Infrastructure Strategy Program and the Education Investment Fund (EIF) Super Science Initiative.

More information from ivan.hanigan@gmail.com or at http://opensoftware-restricteddata.github.io