SIA | SOCIAL INVESTMENT AGENCY ORANGA TANGATA

Investing in what works for better lives

Working principles for social investment analytics*

Workflow guidance and tips for data teams

Version 1: December 2017
SIA-2017-0267

New Zealand Government

Creative Commons Licence

Liability

While all care and diligence has been used in processing, analysing and extracting data and information in this publication, the Social Investment Agency gives no warranty it is error free and will not be liable for any loss or damage suffered by the use directly, or indirectly, of the information in this publication.

Citation

Social Investment Agency (2017). *Working Principles for Social Investment Analytics*. Wellington, New Zealand.

ISBN 978-0-9951022-4-8 (online)

Published in December 2017 by the Social Investment Agency
Social Investment Agency
Wellington, New Zealand

Intended audience

This guide is written for data analysts, statisticians, and developers looking to undertake an analytical project.

Resources, tools and guides

The SIA is developing a range of tools, products and guidance to enable agencies to develop their social investment approaches, and analyse and measure the impact and effectiveness of the services they're delivering.

** By 'social investment analytics' we mean analysis of anonymised, integrated data with the objective of measuring the impact of government services on people's lives.*

Turning data into insight

When social investment analysis is performed on integrated data it allows us to study people's journeys over time in order to learn what works, for whom, and at what cost. While this allows us to track people's journeys over time at the individual level, we do not report results for individuals, and we cannot see who these individuals are.

The analytic methodologies and 'big data' tools that support a social investment approach are powerful and innovative, but to be used effectively they require us to re-think the questions we ask of data. In the spirit of transparent, replicable, reusable and extendable work, the Social Investment Agency (SIA) is sharing what we've learned so that others can adopt these methodologies, if they are useful.

Naturally, demand for the insights we can produce is growing rapidly, and with growing demand comes the need to innovate and adapt to a fast moving environment.

Agencies have well established processes for reporting, however, when questions are asked that fall outside the scope of normal reporting, analytic projects must begin from scratch, and are usually undertaken as discrete pieces of work. This is an expensive and time consuming way of working, and it can cause duplication of effort across different projects and teams. This document provides an overview of the principles that we need to use when undertaking social investment analyses to prevent these issues and to enable a more efficient and effective way of working. This paper is intended for data analysts or statisticians looking to improve their workflow and processes.

Working principles for our data future

The SIA is focusing on moving from bespoke analytics projects in small teams to scaled, reuseable, replicable, automated and open analytics. This way of working will free up time to answer the questions that really matter, by doing the following:

- Create reusable analytics: Build with reuse in mind from the beginning. Reuse as much of your work as you can.
- Use version control to manage work effectively.
- Make your code open so that others can build on your achievements.
- Automate for efficiency so that you don't have to do the same work twice.

Create reusable analytics

Reuse is about moving away from bespoke analytic projects, to building our work in a reusable way, so we can enhance what we've done in the past and move us further ahead every time.

When beginning a new project, it's important to think about reuse from the start. It's not just about sharing code or documentation when you've finished (although that's important too). It means thinking about who else could use what you've done, whether that is someone else or your own team in the future, and what you could do now to make that possible.

Any process that is generic enough can be reused in a different context, if there is good documentation, and well-structured, logical code full of helpful comments. The Social Investment Analytical Layer and Social Investment Data Foundation are two examples of reusable work generated by the SIA. When we reuse things, we are able to make them better. Version control is essential to the transition from bespoke projects to reusable, systematic analytic work.

Use version control

The term 'version control' can apply to both the practice of managing changes to documents, and the software that enables this style of workflow. Version control is saving versions of files, rather than simply modifying one copy. This lets you revert back to older versions, easily undo changes, and maintain a documented record of your project development. This practice is essential for statisticians, data analysts and computer scientists, but it's not yet taught in universities.

The most commonly used version control systems for statistical analysis are Git and Mercurial. Both are decentralised version control systems which allow teams to work concurrently on their own version of a code project, and then integrate their contributions into a Master version in a controlled process. This facilitates teamwork, process and quality control, and easy sharing and adoption of reusable code. Open source version control systems are free to use, and beyond the initial investment in training, they generate no on-going expense whilst delivering innumerable benefits.

Make your code open

Being transparent and open is not only ethical, it's an opportunity to learn from one another and become better at what we do. When work is shared openly it saves others' time and reduces duplication of effort, while also creating the opportunity for improvement as others review and build upon it. Decentralised version control makes it possible for this to occur without any changes being made to the Master code without the authors' approval.

The SIA publishes all our code on our Github page, where it is available for anyone to download a version, and review, edit, and improve it. Although the Integrated Data Infrastructure data to which the code applies is not available without the appropriate clearance, the code is available for anybody to use and apply to their own work.

Automate for efficiency

Automation is about moving from repetitive work to replicable work. Much of the preparation that goes into an analytical project, such as cleaning data for analysis, is time-consuming, laborious and repeated time and time again on various projects. The cleaning of age and income bands, sex, region and other variables is a menial task that is repeated over and over again, at the expense of project timeframes and analysts' patience. As a general rule, once you've coded something twice, it's time to automate it.

The reality is that most statisticians aren't trained to automate their work, and learning to do so can be a challenge, fraught with additional obstacles such as IT and infrastructure restrictions. Building automation capability is a project in itself which can be eased by taking a staged approach.

Automating your work in stages will lay the ground work for more sophisticated automation in future, saving you time, energy and risk.

- On your first project, do it all manually to get comfortable with the data you are using.
- On the next project, try automating your data extraction.
- On the next, have a go at automating your variable cleaning.

For example, if you are creating a model then one task you will be familiar with is building a data set and checking all the variables for missing values or errors. When you check your variables for a model, you usually make some frequency tables or distribution plots in order to check that everything looks right. You can program the computer to do this so you don't need to manually check - the computer will be able to tell you if something is wrong.

By starting small and automating only the small components, you'll prepare yourself for the process of automating some of your decisions. Not only does automation save you time, it enables you to take your automated data extraction to your IT department, and ask them to put it into production. If your work is not automated, and requires manual checks of variables or model output, it can't be productionised. As demand for analytical skills grows, we need to be prepared to scale up workflow and make the best possible use of analysts' time and skills.

Support and infrastructure needed to work this way

The greatest barrier to changing how we work is the established structures and processes which are tried, tested and embedded in management, IT and policymaking. But as government's data and evidence-needs change, the conversation about how we can improve and streamline access to the insights needed for decision-making need to be extended beyond the data lab.

Data analysts and statisticians will need to learn some new skills in order to move away from discrete, bespoke analyses to systematic, productionised insights. They will need support to do this through the provision of time, systems and a mandate. We're going to need:

- Version control software.
- IT policies that permit the use of open-source tools.
- Powerful machines that have capacity to crunch the numbers efficiently.
- Data hosted on servers, instead of network drives.
- Permission, support and time to automate our work.
- Permission and support to share our code outside our own organisations.
- IT support to save datasets for reuse.

The best way to drive change is to demonstrate its value. As we begin to reuse, share, and automate our analyses, we'll be able to produce more, faster, and better quality insights.