NTTS 2019: Transparency and reproducibility of models and algorithms: examples from the UN Global Platform

Keywords: algorithms, transparency, reproducibility, open algorithms.

1. Introduction

In the last two decades the scientific community has been hit by the reproducibility crisis. In fields such as behavioural sciences and medical trials, researchers that have aimed to reproduce classic experiments have found that, in numerous occasions, the results could not be reproduced. These 'false positive' results have had an impact on policy and further research. For example, the policies taken by national governments on the wake of the 2008 financial crisis were influenced by a paper by Reinhart & Rogoff [1], which was found to contain calculation errors. This has led to more researchers openly publishing their data and methodology. Other fields such as medicine and pharmacology have reacted similarly: the registration of clinical trials has become common practice to avoid reporting bias, and reducing the chances of the positive or negative effects of drugs being indirectly manipulated [2].

Given that public administrations, organisations and citizens depend on evidence from the statistical community to take informed decisions, we argue that the field of official statistics is not immune to the reproducibility crisis. Demand for more open and transparent statistics, data, and methodologies is only going to increase in an era in which communication is 24/7 and access to vast quantities of information is accessible to all citizens. Similarly, with this demand comes an opportunity: the openness and transparency of the official statistics industry can generate a greater degree of trust than we currently hold.

We face three immediate threats to our trust: a) a media environment in which unchecked facts are widespread in social media ('fake news'), b) a data science and technology evolution towards large, unstructured datasets and less transparent algorithms and models, c) the 'closeness' of many statistical datasets due to personal or commercial concerns. In this session we focus on how to tackle the latter two by the provision of more open and reproducible algorithms and datasets.

2. Algorithm reproducibility: current and ideal

Two of the main drivers of the UN Global Platform are the provision of trusted methods and trusted data. The methods we provide (statistical methods, machine learning models, AI models, and utility algorithms) have to prove to consumers (for example, academics or other statistical offices) that they do what they are expected to do, they do so securely, and have associated guidance on how to best use them. The trust on methods, algorithms and models depends on a number of factors, such as: a) openness of the code-base, b) openness of the data (and training data), c) documentation of the logic of the algorithm, d) reproducibility of the algorithm by other researchers, e) available use cases and examples.

We argue that the provision of a description of the methodology and availability of the code are the minimum expected from national statistical organisations, but that efforts should be made to provide fully reproducible algorithms. For example, as a minimum, the NSO should provide a published (and approved) documentation of the method used and publish the production code on GitHub (or another open source code repository). Ideally, the NSO should also provide a working version of the method with an execution environment that users can interact with, that is provisioned with either open or synthetic data for immediate testing of the algorithm.

3. Examples from the UN Global Platform

In the UN Global Platform we are focusing on enhancing trust in methods. This is done by increasing transparency, portability and reproducibility. Our initial practices have been to: a) provision a public endpoint and execution environment for each algorithm, so that researchers can apply it to data they are familiar with [3], b) synthetic datasets to explore the workings of the algorithm / model, and c) notebooks with demonstrators of the algorithm in publicly available environments so that citizen users can understand the workings and implications of the algorithm.

3.1. Urban Forests: API endpoints

The Urban Forests projects takes a project initially undertaken by the UK's Data Science Campus and moves many of the algorithms onto the UN Global Platform. It is composed of a number of algorithms for various tasks.

1. Sampling OpenStreetMap highways. This can take a query of either a bounding box, a city or region of city, or a street name. The method will then query the overpass API and return way_ids. The method calculates evenly spaced coordinates along each way_id with a direction given for each point. This effectively creates a set of sampling points alongside a street or road. These points are returned as a list, which can then be easily plotted on a map, allowing a user to check our sampling method.
2. Image downloader. Using the coordinates generated from the previous algorithm, this uses Google's streetview API to download images. These are downloaded into a user's data store in the methods service. The user can go through and check these images are what someone would expect to see.
3. Image segmentation. The algorithm used for segmentation is PSPnet, which was implemented in the Urban Forests project by the Data Science Campus. The network architecture and performance has been well documented[4]. While the understanding of how this algorithm segments the images is likely to be limited, in Urban Forests it serves the purpose of identifying vegetation. By making the segmentation algorithm available over API, its function can be queried more effectively.
4. Composite image creation. This takes the original image and merges it with the segmented image, giving each pixel a colour based on the class that the segmentation algorithm has assigned it.

The end result is a pipeline where a user can make a simple query, such as their street name and a folder of composite segmented images is returned in the users hosted storage. We have two ways to use this pipeline. To aid in transparency we have a notebook that clearly documents the transformation and flow of results from one algorithm to the input of the next algorithm. The algorithms themselves are all open on the methods service. Alternatively, the entire pipeline can be run with a single function call; this function houses the entire pipeline, abstracting away the more technical detail, allowing the focus to be on the results.

3.2. FEWS: API and synthetic dataset

The UN Global Working Group on Big Data Task Team on Scanner Data has also produced an implementation of FEWS in the UN Global Platform Methods Service. This method, which stands for Fixed Effects with a Window Splice, is used to estimate inflation from a heterogeneous set of priced items, for which product information is missing [5]. Although implementations of FEWS are available on GitHub, the implementation we provide facilitates reproducibility and enhances transparency by:

1. Providing an implementation on the Methods Service. The implementation is container based and stores the code and dependencies used in the FEWS algorithm. This ensures

the same working implementation is always used, with the associated packages and libraries being ‘frozen’.
2. Providing an API endpoint in the methods service that allows FEWS to be executed. A dataset which comprises items and their prices can be formatted into JSON and submitted to the API for estimation. This allows users to execute the method.
3. Statistics New Zealand provided a synthetic dataset that allows users to test the method without requiring them to provide commercial data.
4. Available notebook. A notebook that demonstrates how the method works and allows users to test the results and assumptions is also available.

4. Discussion

The previous examples showcase how to use good design principles and a choice of widely available tools to maximise transparency and reproducibility in official statistics. In the coming years it will be necessary for the statistical industry to consider how to make their solutions and algorithms more portable and replicable, for example, by the use of containers. Similarly, when facing the statistical users outside our immediate industry, it will become important to provision more algorithms and digital services that are aimed at citizen users and academic researchers.

References

[1] Carmen Reinhart and Kenneth Rogoff, "Growth in a Time of Debt", *American Economic Review* (2010), **100** (2): 573–78.
[2] US National Library of Medicine, 2018. ClinicalTrials.Gov. [Online] Available at https://www.clinicaltrials.gov. [Accessed 15 October 2018]
[3] UN Global Platform, 2018. UN Global Platform Home Page. [Online] Available at https://www.officialstatistics.org. [Accessed 15 October 2018]
[4] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, Jiaya Jia. “Pyramid Scene Parsing Network”. arXiv:1612.01105 (2016).
[5] Frances Krsinich, “The FEWS Index: Fixed Effects with a Window Splice”, *Journal of Official Statistics* (2016), **32** (2): 375-404.