The present and future status of open government data

STAT 448, Assignment 1-2, Fernando Cagua, March 2017

In order to fulfill their functions, governments tend to collect large amounts of data that are high in volume, variety, veracity, and arguably velocity. Hence, government data as a whole seems to be a good candidate to receive the "big data" label. Nevertheless, the government data has the potential to be big data for another, perhaps more important, criterium—value.

Few organizations have the potential that government has to use their data to transform society. It has been shown that the value of government datasets can be even larger when they are open (being open means that they can be freely used and shared by anyone). This is so because the benefits of open data are not only for the civil society, but also for the government itself and range from transparency improvements, to citizen empowerment, passing by public service improvement and indirect effects in the wider economy (Ubaldi 2013). For example, the potential global value of open data has been estimated to be \$3 trillion (Chui, Farrell, and Jackson 2013), and 100,000 direct jobs would be created by 2020 in Europe alone (Carrara et al. 2015).

Despite the clear advantages, a large proportion of countries still have a limited or inexistent public open data infrastructure. While many civic initiatives attempt to ameliorate this issue by sharing independently collected data, they are usually unable to substitute the value of government datasets. Large initiatives like the Global Partnership for Sustainable Development Data—which unite governments and NGOs with the aim of fostering the applications of open government data to achieve the 2030 UN sustainable development goals–are trying to remedy the situation. In a prime example, the Open Government Partnership, a multi-country alliance in which governments commit to be "more open, accountable, and responsive to citizens", has grown from eight countries in 2011 to 75 in 2016 (Open Government Partnership 2016).

More and more governments are committing to open public data, certainly important progress. However, monitoring the *actual* progress to which government data is being made open is more challenging. Currently, three indexes attempt to quantify this progress. The Open Data Inventory includes 173 countries and focus on gaps on the official statistics provided by the national statistics offices (Open Data Watch 2016). The Open Data Barometer includes 92 countries and relies on expert assessment which inform on the readiness, implementation, and impact of open data (World Wide Web Foundation 2015). Finally, the Open Data Index includes 122 countries and relies on peer-reviewed crowdsourced data from individuals and citizen initiatives. It goes beyond national statistics and evaluates the status of open data on dimensions like legislation, GIS, pollutants, land ownership, etc (Open Knowledge Foundation 2017).

Although these indices provide important insight into the current status of open government data, they do not provide all the information required to analyze its evolution because many governments embraced open data policies before the indices were developed. Understanding the macro factors that drive the implementation of open government data, and predicting future trends, would be an valuable asset to design strategies to speed up the adoption of the "data revolution" across countries.

Here, I do precisely that. Specifically, I use the date in which a country opened an open government data portal (such as data.govt.nz) as a proxy for a country support for open data. Although the opening date is not able to accurately measure the quantity and quality of public data, I found it to be highly correlated with both the Open Data Index and the Open Data Barometer (Spearman correlation coefficient of 0.62 and

0.57, respectively). Open data portals are a good indication of the progress of open data because–by making datasets discoverable and managing metadata–they have the potential to accelerate the creation of value (Attard et al. 2015).

To obtain the web address of the open data portals were open I curated an automated search that returned the 10 first results of a Google Search in an english locale for the string "`Open Data + [country]`" for each of the 193 United Nations member states. I then obtained an approximate opening date for the portal by automatically retrieving the date in which the site was first registered by the Wayback Machine, which keeps historical snapshots of billions of URLs over time. This data would be easily improved by performing searches in local languages. Code can be foun in github:efcaguab/open_data_history.

Using this methodology I found that the adoption of open government data is not homogeneous across regions (Figure 1). Europe seems to be at the vanguard of in terms of support for open government data portals. In particular there was a period of rapid growth between 2012 and 2014 when most West European countries launched their portals. In Asia and the Americas the largest growth occurred after 2014 and are currently on track to catch up with European countries. Growth in Africa and Oceania has been rather moderate and, with a few exceptions, governments are yet to embrace open data. Using the historical records also allow us to identify the pioneers of open data. The USA, UK, Norway, Australia, and New Zealand, all launched their open government data portals in the earliest dates, setting up the example to other countries to follow.
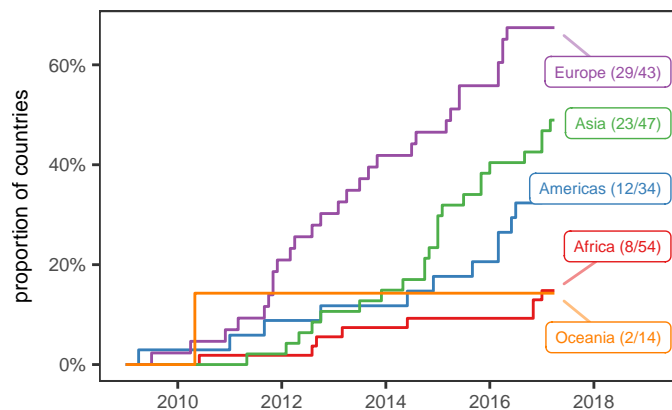

Figure 1: Proportion of countries for which national open government data portals exist.

Although informative, the regional bins do not allow us to understand the relationships between the adoption of open government data policies and potential covariates. I therefore constructed a model to determine how the gross domestic product per capita and the number of internet users (per 100 people) are related to the launch date of the data portals. Although many other factors might contribute to the progress of open data, I chose these two variables because they are likely to serve as a proxy for the capacity that countries have to implement and utilize open data. Specifically, I used a parametric survival regression model under the assumption that the launch date follows a Weibull distribution (Therneau and Grambsch 2000).

Although somewhat simplistic, this model already allows for some insightful information. First, albeit the two explanatory variables are moderately correlated, the proportion of internet users is a much stronger predictor of the portal launch date than the per-capita gross domestic product.

Second, by examining the model residuals and the launch date the model allows to identify open data champions. Countries that are "ahead of time" and launched open government data portals way before it

would have been expected given their levels of internet penetration and socioeconomic status. Specifically, the top five is composed by Burkina Faso, Ethiopia, Pakistan, Ghana, and Bangladesh.

Third, the model allows to forecast the expected date in which open government data portals could be implemented in the countries that have not yet done so (Figure 2). Under the current trajectory, the most likely outcome is that ~70% of the nations would embrace open government data portals by 2030. This includes all European countries and most countries in Asia, Oceania, and the Americas, but excludes a large proportion of central and west african countries, and poor and conflictive countries in Asia.
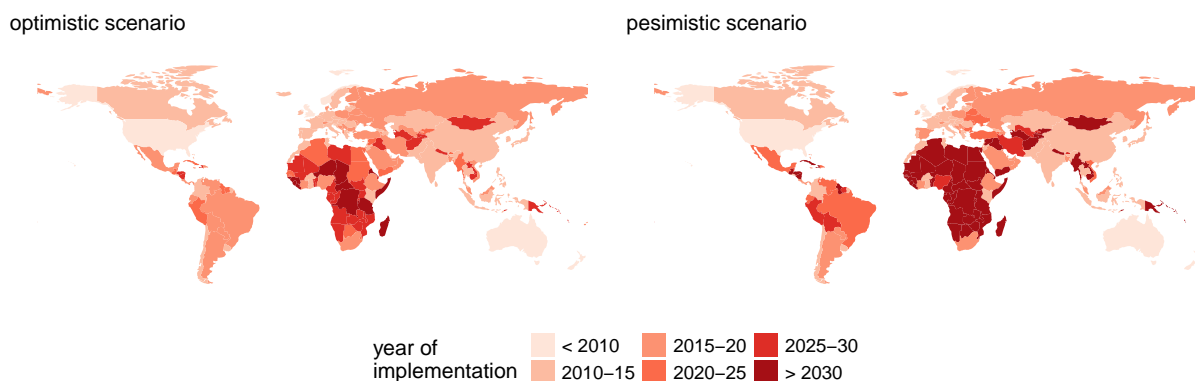

Figure 2: Predictions

A major goal of the major international organizations that promote open government data is to harness the big-data revolution to achieve the 17 goals of the 2030 Agenda for Sustainable Development. If indeed open government data can be used to "help end extreme poverty, combat inequality and injustice, and combat climate change", is value would be so large that it could, without a doubt, be called big-data. This analysis suggest that the adoption of open data could be accelerated by enabling all people to use the internet, which then can translate into creation of value when citizens demand and use public data. Lamentably, this analysis suggest that, unless this acceleration takes place, the countries that need it the most will also be the ones that are likely to miss it.

References

Attard, Judie, Fabrizio Orlandi, Simon Scerri, and Seren Auer. 2015. "A systematic review of open government data initiatives." *Government Information Quarterly* 32 (4). Elsevier Inc.: 399–418. doi:10.1016/j.giq.2015.07.006.

Carrara, Wendy, Wae San Chan, Sander Fischer, and Eva van Steenbergen. 2015. "Creating Value through Open Data: Study on the Impact of Re-use of Public Data Resources." European Comission - European Data Portal. doi:10.2759/328101.

Chui, Michael, Diana Farrell, and Kate Jackson. 2013. "How government can promote open data and help unleash over \$3 trillion in economic value." McKinsey Global Institute.

Open Data Watch. 2016. "The open data inventory 2016 methods report." Washington DC: Open Data Watch.

Open Government Partnership. 2016. "Annual report 2016: from commitment to action." Open Government Partnership. doi:10.1017/CBO9781107415324.004.

Open Knowledge Foundation. 2017. "Global Open Data Index." http://index.okfn.org/.

Therneau, Terry M., and Patricia M. Grambsch. 2000. *Modeling Survival Data: Extending the Cox Model.* Statistics for Biology and Health. New York, NY: Springer New York. doi:10.1007/978-1-4757-3294-8.

Ubaldi, Barbara. 2013. "Open government data: towards empirical analysis of open government initiatives." Paris: OECD Publishing. doi:10.1787/5k46bj4f03s7-en.

World Wide Web Foundation. 2015. "ODB global report: third edition." Open Data for Development Network.