Journal of Business Research xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Journal of Business Research

Constructing spatiotemporal poverty indices from big data☆

Christopher Njuguna [a,*], Patrick McSharry [a,b]

[a] ICT Center of Excellence, Carnegie Mellon University, Kigali, Rwanda
[b] Smith School of Enterprise and the Environment, University of Oxford, UK

ARTICLE INFO

Available online xxxx

Keywords:
Call detail record (CDR)
Poverty index
Machine learning
Big data
Socioeconomic level
Rwanda

ABSTRACT

Big data offers the potential to calculate timely estimates of the socioeconomic development of a region. Mobile telephone activity provides an enormous wealth of information that can be utilized alongside household surveys. Estimates of poverty and wealth rely on the calculation of features from call detail records (CDRs), however, mobile network operators are reluctant to provide access to CDRs due to commercial and privacy concerns. As a compromise, this study shows that a sparse CDR dataset combined with other publicly available datasets based on satellite imagery can yield competitive results. In particular, a model is built using two CDR-based features, mobile ownership per capita and call volume per phone, combined with normalized satellite nightlight data and population density, to estimate the multi-dimensional poverty index (MPI) at the sector level in Rwanda. This model accurately estimates the MPI for sectors in Rwanda that contain mobile phone cell towers (cross-validated correlation of 0.88).

1. Introduction

According to the United Nations (UN), by 2015, about 12% of the world's population or approximately 800 million people were considered extremely poor (United Nations, 2015). While this constituted a significant decline in extreme poverty from a global estimate of 36% (approximately 1.9 billion) in 1990 and an over-achievement of the millennium development goal (MDG) target to halve the global population that is extremely poor by 2015, the number of global poor still remains a major challenge. The UN General Assembly in adopting the agenda for sustainable development stated that ending poverty in all its forms and dimensions, including extreme poverty, is "the greatest global challenge and an indispensable requirement for sustainable development" (Transforming Our World: The 2030 Agenda for Sustainable Development, 2015). This is reflected in the prime sustainable development goal, SDG1, which pledges to "end poverty in all its forms everywhere" (Transforming Our World: The 2030 Agenda for Sustainable Development, 2015) and sets its first target as "to eliminate extreme poverty for all people everywhere by 2030" (Transforming Our World: The 2030 Agenda for Sustainable Development, 2015). Poverty, in order to be eliminated, must first be defined and measured. This is done using poverty indices which fall into two main categories: monetary poverty indices and non-monetary poverty indices (Bourguignon & Chakravarty, 2003). Monetary indices are based on monetary income measures in local or global currency and include the different national poverty lines as well as the well-known World Bank $2 a day and $1.25 a day poverty indices. Meanwhile, non-monetary indices rely on proxies for wealth and estimate poverty intensity by how much one is deprived of the listed proxies. One such non-monetary index is the multi-dimensional poverty index (MPI) developed at the Oxford University (Oxford Poverty and Human Development Initiative, 2013). The UN currently measures extreme poverty using the $1.25 a day monetary poverty index (Transforming Our World: The 2030 Agenda for Sustainable Development, 2015).

Poverty indices are among the key figures and indicators that are used to influence policy and, thus, need to be based on data that are representative of the population. Traditionally, the census and survey are two tools that have been commonly used to obtain data but, while they have been proven over the years to be accurate and reliable, the great monetary cost, time and effort of carrying them out means that they can only be undertaken periodically - censuses are typically implemented in ten-year cycles and surveys in a smaller multiple of years, typically three to five years (Blumenstock, Cadamuro, & On, 2015). For example, in Rwanda, the National Institute of Statistics of Rwanda (NISR) undertakes the Rwanda Population and Housing Census (RPHC) every ten years while the Rwanda Economic and Living

☆ The authors would like to thank Patrick Nyirishema the Director General of Rwanda Utilities Regulatory Authority (RURA) for partnering with the Carnegie Mellon University on this project. They would also like to thank Francis Ngabo, the then Director of the Frequency Monitoring Division of RURA, Georges Kwizera, Protais Kanyankore and the staff of RURA in general. The authors are grateful to Rajiv Ranjan, Governance Unit of the United Nations Development Program (UNDP) and ICT Advisor to the National Institute of Statistics of Rwanda (NISR), and Tom Bundervoet, Senior Economist with the World Bank for their invaluable help and support. The authors acknowledge Chuma Vuningoma and the staff of MTN, Rwanda for their support and assistance with CDR acquisition.

* Corresponding author.
E-mail addresses: chris.njuguna@gmail.com (C. Njuguna), patrick@mcsharry.net (P. McSharry).

http://dx.doi.org/10.1016/j.jbusres.2016.08.005
0148-2963/

Table 1
Censuses and surveys in Rwanda as of Dec. 2015.

Name	Frequency (Years)	Last Year	Households sampled
Population and Housing Census (RPHC4)	10	2012	Whole population
Housing and Living Conditions Survey (EICV4)	3	2013/14	14,419
Demographic and Health Survey (DHS5)	3	2014/15	12,793

Conditions Survey (EICV) and the Rwanda Demographic and Health Survey (RDHS) are carried out every three years (National Institute of Statistics of Rwanda (NISR) & M. of F. and E. P. (MINECOFIN), 2014) (See Table 1.). Thus, the census approach to data collection presents a one-off, detailed snapshot of the socioeconomic status of the country but with these snapshots decision-makers can only assess the impact of their policies after three to ten years. Furthermore, decisions made using census data are essentially made using old data since many factors may have changed by the time they are analyzed and disseminated. Worse still, in some developing countries, especially in Sub-Saharan Africa, no official census or survey has been undertaken in decades leading to a "dearth of reliable statistics" (Letouzé, 2014). Needless to say, any statistics for such countries cannot be said to be accurate, a situation that the World Bank has termed Africa's "statistical tragedy" (Letouzé, 2014). This lack of data and, where data do exist, their low update frequency, prompted a global search for alternatives to these traditional tools.

The introduction of digitized records, the growing use of mobile phones and other digital devices and the proliferation of remote sensing devices has led to an explosion in the amounts and variety of data available. This explosion of data has led to a phenomenon known as "big data" which is showing promise as an answer to the search for an alternative to the traditional data sourcing tools (Global Pulse, 2013). Big data, which due to its novelty still does not have a fixed definition, has been described in various ways. One common description sees big data as comprising data that have high volume, velocity and variety (Letouzé, 2014). Another definition, geared towards the use of big data in development, describes big data as data that have all or some of the following characteristics: they are digitally generated, passively produced (i.e. generated as a byproduct of everyday life), automatically collected and geographically or temporally trackable with the ability to be continually analyzed (Hilbert, 2013). This definition de-emphasizes the volume aspect of data arguing that the kind of data and their source are more relevant than their size. Thus, some regard the name "big data" as something of a misnomer (Letouzé, 2014).

Big data, also called business intelligence and analytics version 3.0 (BI&A 3.0), is a combination of data and the processes that analyze and convert these data into actionable insights and is considered to be the third stage in a business intelligence and analytics (BI&A) evolution over the years. This data evolution began with the emergence of business intelligence (BI&A1.0) which appeared in the 1990s and involves the statistical analysis of structured records stored in relational database management systems (RDBMSs). Business intelligence has been fully adopted into commercial systems today with features such as online analytical processing, reports, statistical analysis, data mining and prediction as examples. Business analytics (BI&A 2.0) appeared in the 2000s with the advent of the internet and the web (Chen, Chiang, & Storey, 2012) and builds on business intelligence by adding internet protocol (IP) and personal interaction information stored in server and transaction logs. Unstructured data such as pictures are also included under business analytics. These data allow for a deeper understanding of customer needs and business opportunities, are significantly larger in volume than business intelligence and are processed using text and web analytics. Some of the features of business analytics have been adopted into commercial systems and include web mining and social-network analysis. Finally, in the third phase of the evolution in the 2010s, big data (BI&A 3.0) appeared characterized by an exponential growth in the volumes of data. As mentioned earlier, this is brought about by the explosion in the number of phones and sensor-based internet-enabled devices. To illustrate this exponential growth, in 2012 estimates indicated that 90% of all the data in the world had been generated within the years 2010–2012, a period of slightly over two years (Letouzé, 2014). The storage, analysis and use of these data are still in the research phase and are not likely to be adopted by commercial systems in the near future (Chen et al., 2012).

Big data provides a "highly mobile, location-aware, person-centered and context-relevant" (Chen et al., 2012) dataset and its benefits are starting to be appreciated with the impetus for using big data in improving official statistics gaining momentum recently. The UN has recognized the positive impact of big data in projects and called for the increased use of big data to inform decision-making and to assist in working towards the achievement of the newly established global goals for sustainable development terming this trend the "data revolution for development". In line with this new direction, the UN Global Pulse was created. Describing itself as a "flagship innovation initiative of the United Nations Secretary-General on big data", the UN Global Pulse is based on the recognition that "digital data offers the opportunity to gain a better understanding of changes in human well-being and to get real-time feedback on how well policy responses are working" (Global Pulse, 2013). This activity within the UN demonstrates that big data is not only of academic interest but is already being considered for policymaking.

With all the promise that big data heralds some limitations must be considered. To start with, big data is not usually collected for analysis (Letouzé, 2014). This means that the sample of the data collected may not be representative of the population due to various biases. For example, data collected from digital platforms are likely to contain a selection bias since certain groups, such as the youth and the wealthy, are more likely to use these platforms than other demographics (Hilbert, 2013). Second, the power of big data is in the ability to store and process large amounts of data, however, due to cost and capacity this capability is less likely to be found in the communities which could actually benefit most from this analysis – poor countries. Big data, then, could end up benefiting wealthier nations rather than the developing world where these benefits are needed most and may actually result in a widening of the digital divide (Letouzé, 2014). A third limitation is in the disproportionate power given to certain actors in the big data ecosystem. An example would be a data analyst who might leave out a gender variable from a dataset or analysis causing inequalities in gender representation not to surface. The data analyst, then, must have good intentions and a good grasp of the data domain as well as be competent in order to offer a fair analysis (Taylor, Cowls, Schroeder, & Meyer, 2014). Finally, big data has limitations in access - most big data is not open and easily accessible (Hilbert, 2013). Due to privacy and commercial concerns, private companies want to mine their own data in the quest to find value from them and are afraid others might glean information that may hurt their competitiveness. Meanwhile, public institutions may not see opening data as being in their interest such as where performance- or corruption-related issues may be revealed. One of the ways to overcome these problems is for organizations to provide aggregated data which would prevent sensitive fields from being revealed. However, note that merging datasets, even aggregated ones with other datasets may still pose a privacy risk as linking and merging datasets can have the effect of revealing individual data (Taylor et al., 2014). Another way to overcome closed data is to involve different international actors with superior access to data and analytical capacity. Such actors can demonstrate the potential value of data by showcasing the benefits achieved in their national contexts, thereby, presenting compelling, practical cases for the sharing of data (Taylor et al., 2014). As big data grows in influence and usefulness, addressing the challenges of big data will continue to be a crucial subject area for researchers and practitioners ahead of mainstream acceptance.

Please cite this article as: Njuguna, C., & McSharry, P., Constructing spatiotemporal poverty indices from big data, *Journal of Business Research* (2016), http://dx.doi.org/10.1016/j.jbusres.2016.08.005