Chapter 11
Big Data in Official Statistics

Kees Zeelenberg
Statistics Netherlands, The Netherlands

Barteld Braaksma
Statistics Netherlands, The Netherlands

ABSTRACT

Big data come in high volume, high velocity and high variety. Their high volume may lead to better accuracy and more details, their high velocity may lead to more frequent and more timely statistical estimates, and their high variety may give opportunities for statistics in new areas. But there are also many challenges: there are uncontrolled changes in sources that threaten continuity and comparability, and data that refer only indirectly to phenomena of statistical interest. Furthermore, big data may be highly volatile and selective: the coverage of the population to which they refer, may change from day to day, leading to inexplicable jumps in time-series. And very often, the individual observations in these big data sets lack variables that allow them to be linked to other datasets or population frames. This severely limits the possibilities for correction of selectivity and volatility. In this chapter, we describe and discuss opportunities for big data in official statistics.

INTRODUCTION

Big data come in high volume, high velocity and high variety; examples are web scraping, Twitter and Facebook messages, mobile-phone records, traffic-loop data, and banking transactions. This leads to opportunities for new statistics or redesign of existing statistics. The potential of big data for official statistics lies in the immense amount of information contained. For example, the sheer size might make it possible
Big Data in Official Statistics

to add more details to official statistics. Also, big data sources may cover areas of society for which official statistics do not yet exist. Their high volume may lead to better accuracy and more details, their high velocity may lead to more frequent and timelier statistical estimates, and their high variety may give rise to statistics in new areas.

There are various challenges with the use of big data in official statistics, such as legal, technological, financial, methodological, and privacy-related ones; see e.g., Struijs, Braaksma, and Daas (2014), UN-ECE (2013), and Vaccari (2016). This paper focuses on methodological challenges, in particular on the question of how official statistics may be produced from big data. Specifically, we look at the question: what are the best strategies for using big data in official statistics?

From a methodological perspective, big data also poses many challenges. Big data may be highly volatile and selective: the coverage of the population to which they refer may change from day to day, leading to inexplicable jumps in time-series. And very often, the individual observations in big data sets lack linking variables and so cannot be linked to other datasets or population frames. This severely limits the possibilities for correction of selectivity and volatility. The use of big data in official statistics therefore requires other approaches. We discuss two such approaches.

In the first place, we may accept big data just for what they are: an imperfect, yet very timely, indicator of developments in society. In a sense, this is what national statistical institutes (NSIs) often do: we collect data that have been assembled by the respondents and the reason why, and even just the fact that they have been assembled is very much the same reason why they are interesting for society and thus for an NSI to collect.

Secondly, we may extend this approach in a more formal way by modelling these data explicitly. In recent years, many new methods for dealing with big data have been developed by mathematical and applied statisticians.

The structure of this chapter is as follows. In section 2, we briefly describe big data sources and possible uses. In section 3, we look at statistics from big data as they are collected or assembled, i.e., as statistics in their own right; we describe several examples of big data as official statistics. In section 4, we discuss how models may be useful for creating information from big data sources, and under what conditions NSIs may be using models for creating official statistics; we also describe several examples of statistics derived through modelling big data. We look at how to create representative estimates and how to make the most of big data when this is difficult or impossible. We show how big data may be useful in solving several of the major challenges to official statistics, in particular the quality of national accounts (rate of growth of gross national product or GNP), the timeliness of official statistics,
Statistical Approach for Ranking OECD Countries Based on Composite GICSES Index and I-Distance Method
www.igi-global.com/chapter/statistical-approach-for-ranking-oecd-countries-based-on-composite-gicses-index-and-i-distance-method/165658?camid=4v1a

Artificial Intelligent Embedded Doctor (AIEDr.): A Prospect of Low Back Pain Diagnosis
www.igi-global.com/article/artificial-intelligent-embedded-doctor-aiedr/247457?camid=4v1a