Fabrice Murtin (OECD)

Using Big Data for Social Statistics: The Case of Well-being in the US

Topic 2 – Learning more from what we already know

Introduction

There is a growing interest in social sciences in going beyond the income-based approach of human development by using new measures of wellbeing (Stiglitz et al., 2009). In particular, GDP does not measure non-market social interactions such as friendship, family happiness, moral values or the sense of purpose in life. This motivates the recourse to subjective self-reported measures of well-being, such as Life Satisfaction (i.e. answers to the question: "All things considered, how satisfied are you with your life as a whole those days?"), which economists increasingly use as a direct measure of utility. Political leaders have embraced this move by calling for representative surveys of well-being to guide their policy, as illustrated by the Cameron's commission of well-being in UK. In spite of these achievements, subjective well-being measures still raise a number of challenges and concerns among economists.

Methods / Problem statement

First, they are not based on revealed behavior and choices, and are affected by the limits inherent to all self-reports (Deaton, 2013). Subjective well-being is also multidimensional (Kahneman and Deaton, 2010| Steptoe et al. 2014) and some aspects, especially pain or chronic disability are difficult to capture. Second, survey questions on subjective well-being are limited in coverage, time and space, which most of the time eliminates the possibility of measuring it with the appropriate business-cycle frequency and local level for policy decisions. Third, questions are sensitive to wording and ordering effects if the survey is changed (Deaton, 2013).

Results / Proposed solution

This paper contributes to this new research agenda by showing how Big Data can improve our understanding of the foundations of well-being. A major consequence of the accelerated digitization of social life is the traceability of social relations, embedded in large datascapes, such as Google, Facebook, Twitters or the Blogosphere. The quantification of those social traces is of considerable interest for social scientists. However, as stated by Lazer et al. (1999), while the capacity to collect and analyze massive amount of data has transformed the fields of physics and biology, such attempts have been much slower in social sciences. This paper illustrates the potential use of Big Data for both the measurement and the analysis of subjective well-being.

Conclusions

This paper demonstrates the capacity of search engine data to track and replicate the trends in subjective states that are traditionally captured by surveys and to elicit and identify the type of activities that predict subjective well-being. We construct robust predictors of subjective well-being measures in the United States using a very large amount of search engine data covering the years 2008-2013. We measure the life dimensions whose search intensity is robustly associated with self-reported well-being collected by the Gallup Healthways Wellbeing survey, such as life evaluation (Cantril ladder) or the percentage of people who declare that they have experienced happiness, stress or worry "during a lot of the day yesterday". With our composite categories, we predict a time-series that tightly fits the Gallup survey trends in subjective well-

being, and also behaves nicely out of sample. This method allows identifying the type of behaviors, activities and experiences that are associated with higher or lower subjective well-being. For all subjective well-being variables, material conditions are the most important family of predictors, followed by social factors and health/wellness categories. At the category level, we find that keywords related to job search, financial security, family life and leisure are the most important predictors of subjective well-being.