Is Facebook the future of the national census?

Every ten years, governments around the world conduct their national census: a count of the resident population. This is no easy task. The 2011 Chinese census, for instance, was a vast operation relying on six million enumerators taking census questionnaires from door to door. Knowledge of a population's composition is crucial for government planning, for example in terms of investing resources in elderly care. In some countries, it is the basis for dividing parliamentary seats over electoral districts.

Historically, census counts have been relevant to the development of modern nation states because they link populations to a territory. They have also been key to the governance of colonial populations by, among other things, introducing and reinforcing hierarchies of ethnicity, race and class.

Today, national censuses are still potential sites of exclusion, as shown by US Census Bureau plans to include a question about citizenship in 2020. Advocacy groups, social scientists and others fear that current anti-immigration politics will cause a decline in response rates among non-citizen immigrants. This could affect the uptake of their interests in funding decisions and Congress representation of states with higher immigration rates.

These real effects demonstrate that the manner in which censuses and other population counts are carried out is important. This something we should bear in mind with the increasing application of big data. Companies, governments and academia regularly suggest using big data produced by technology corporations such as Facebook, telecom companies or privatised public services to complement census data. With Facebook allegedly working on a map of everyone in the world, what will this mean for censuses as conducted by statistical agencies?

The evolving census

Research I conducted as part of the ARITHMUS project suggests that national census methods in the EU are changing partly in response to corporate, third sector and government criticisms. Criticisms include their high cost, low frequency, publication lag, limited geographic detail and limited breakdown of population classifications into sub-categories.

These problems are recognised by National Statistical Institutes (NSIs) and are being addressed through European harmonisation and innovation programmes that started more than two decades ago. As a result, many EU NSIs using door-to-door questionnaires as their

main data source are planning to replace this method with digital population registers and other administrative registers. Some countries, such as the UK, are introducing online questionnaires. In addition, regulations are in place to ensure censuses are carried out in more geographic detail and at more frequent intervals.

Furthermore, various external organisations are now publishing population counts that can be compared with, complement or challenge census counts. In 2017 for example, Facebook was criticised by advertisers for reporting a UK audience between 18 and 25 years old that exceeded the census count by two million people. Responding to these allegations, a Facebook representative told The Guardian that their methods, partially based on advert clicks, "are not designed to match population or census estimates".

The example illustrates some issues with big data counts, in this case based on social media. In contrast to survey data consisting of answers to pre-set questions, social media data is produced through the use of apps and platforms. They may not correspond to what censuses measure. Additional problems are that big data collections contain more "noise" (double clicks in the above example) and under-represent parts of the population (older people not using social media). Door-to-door surveys, however, also lack coverage of particular groups and need to take into account response bias.

Most NSIs in our study do not foresee taking up big data as a source for *producing* censuses in the near future. But many – among them Statistics Netherlands, the UK Office for National Statistics, and Statistics Estonia – are undertaking experiments. An example is the use of mobile phone location data to establish the "daytime population" of specific geographic areas.

Invisible data collection

Such supplementary data can be useful. Knowing the daytime population of a place, for example, is important information for town planners, traffic, services, and so on, because it provides information about where people spend time during the day. More fundamentally, not only residents are included, but also visitors and internationally mobile workers. Population counts based on data produced by technology companies, telecom operators and public services can therefore provide opportunities to rethink basic definitions of what makes up a population.

But we should also be wary because the big data methods trialled by NSIs and external organisations lack a key characteristic of the census. Government questionnaires allow individuals and civic organisations to negotiate demographic categories, as was the case with the addition of a "Hispanic" category in the 1980 US census. Questionnaires allow people to use the census to negotiate the way in which they are categorised – by refusing to tick any of the available options, for example, or suggesting new ones. Although indirectly, register-based censuses also enable this, as was the case when Dutch transgender advocacy groups succeeded in relaxing the requirements for sex and name changes.

Most big data, by contrast, is collected and processed invisibly. Companies offer little insight into their methods and categories, as concluded by researchers using pre-categorised Facebook data to estimate migrant numbers. Because such research often focuses on migrants and other "hard-to-reach" groups, the results can increase the potential for surveillance and

exclusion of vulnerable groups. The resulting methods also provides limited possibilities for these groups to negotiate their categorical status.

So the census may not disappear as a consequence of competition by Facebook. But it is important to remember that increasing use of big data by statistical agencies and other organisations can affect how people are able to influence how they are counted, categorised and governed. To maintain census methods that promote input from the population, it's key that we practice transparency about data collection and keep categorisation procedures open. This is relevant of census counts and for counts produced by external organisations.

These are not only legal and technical issues, but opportunities to engage people as participants in the definition of themselves and their population.