Administrative Data Linking: Enriching Administrative Data with Surveys

Tom Emery[1]

[1] *NIDI, Den Haag, Netherlands; emery@nidi.nl*

Abstract
In social survey research there is a great deal of interest in enriching survey data with administrative data sources such as income, employment or even criminal records. However, the added value of such data linking to administrative data sources is rarely considered. Using data from the GGS in Sweden as an example, this paper outlines ways in which administrative data can be enriched by linking to a social survey. First, the survey process provides an opportunity to attain consent from respondents to link data from normally dispersed administrative records (e.g. employment and birth records). In so doing, social surveys provide the key of consent for complex data linking and the subsequent ability to use administrative data to answer pressing social questions. Second, social surveys provide an opportunity to validate the data collection processes within administrative data collections. Third, the social surveys collect data themselves which is wholly absent from administrative records but is nonetheless of interest to both administrative data holders and social scientists. For example, the GGS contains data on the distribution of household work which, when taken in conjunction with administrative data provides key insights into gender roles throughout society. The analysis presented within the paper examines these three advantages and the degree to which they are evident in the case of the Swedish GGS. Given Sweden's strong administrative data tradition, it represents an example of how social survey data can supplement even a highly developed social statistics system.

Keywords: (1-5 words), survey data, administrative data, linking, validation

1. Introduction

The role of social surveys within social statistics and population registries is changing. The increasing abundance of administrative data has raised questions about the need for and role of social surveys within modern data infrastructures (Brown, 2013). Social surveys are costly to administer and are increasingly affected by low response rates (Davern, 2013). As national data infrastructures become increasingly sophisticated it becomes increasingly tempting to view administrative data sources such as population registries, tax records and health records as alternative source of data, especially when they can be linked together.

This paper examines the existing relationship between social surveys and administrative data and argues that the social survey should continue to play a central role within national data strategies and the design of data infrastructures. First, the survey process provides an opportunity to attain consent from respondents to link data from normally dispersed administrative records (e.g. employment and birth records). In so doing, social surveys provide the key of consent for complex data linking and the subsequent ability to use administrative data to answer pressing social questions. Second, social surveys provide an opportunity to validate the data collection processes within administrative data collections. Third, the social surveys collect data themselves which is wholly absent from administrative records but is nonetheless of interest to both administrative data holders and social scientists.

To examine these issues in practical terms, we draw on the experience of the Swedish Generations and Gender Survey (GGS) which is part of the International Generations and Gender Programme (www.ggp-i.org). The Swedish GGS was fielded in 2013 with 9,688 respondents (response rate = 54.7%). Interviews were conducted via CATI and a follow up postal questionnaire was used for additional modules. The Swedish GGS successfully linked the sample with a wide range of administrative records which were then integrated into fieldwork process and the publically available dataset for use by social scientists in conjunction with data from 18 other countries.

2. Accessing Administrative Data for Research

2.1. Limited Access in Administrative Data

The potential of administrative data to be used in scientific research is significant and substantial. However, there are many obstacles to the practical realization of this potential. The biggest issue with using administrative data in social research is access. Social scientists are required to go through access procedures that are usually extensive and typically involve the vetting of their research project by the data owners to ensure that the research project fits within the scope of appropriate use. Once this administrative barrier is overcome, the practical obstacles of data access are also encountered. Researchers are often required to access data on site in a controlled setting in which they are limited in the analysis that can be conducted. For example, it can be difficult to receive permission to link the data to comparable sources from other jurisdictions or enrich the data with variables from higher units (i.e. municipality level variables).

From the social scientist's perspective, these obstacles become increasingly restrictive as the data becomes of increasing scientific value. This is due to social scientists being primarily interested in the association between variables which often need to be drawn from multiple records. With each additional linkage however, the anonymity of individuals is increasingly compromised and this is matched by an increase in the restrictions placed on the researcher. Administrative data therefore not only has more stringent access conditions than other forms of data but these restrictions actually increase as the scientific value of the research increases. For example, examining employment records through administrative data records is of great scientific value but examining employment records relationship with health records is where the bulk of administrative data's potential lies. Acquiring access for such research is more difficult than accessing health and employment records independently.

In addition to this problematic association between scientific value and access, there are some areas of social science which are highly problematic with administrative data. Specifically, any research requiring a comparative perspective is problematic. Given that access is stringently controlled, it becomes a severe practical challenge to simultaneously analyze two sources of administrative data from two separate countries. Given that most comparative

research relies on having data from a considerable number of countries simultaneously, it becomes unclear how such research can be conducted using administrative data.

There have been projects which have aimed to tackle some of these issues. For example, IPUMS based at the University of Minnesota has brought together census records from a large number of countries and controls the exact data that researchers can access, thus preventing disclosure of individual's identities. However, the range of variables available within IPUMS is currently highly limited and focused on basic demographics (e.g. gender, age, marital status etc). A more ambitious attempt to provide a platform for comparative research with administrative data exists in the form of Data without Boundaries[1]. This project brought together administrative data sources from several countries to a single site which made joint analysis possible. Yet the number of countries for which data is available is currently limited to just 3 countries and researchers are still required to access the data from controlled sites.

Access to administrative data is restricted for valid and highly important reasons but these obstacles to data access run counter to the open science agenda[2]. Social science data infrastructures serve a vast community of researchers. The European Social Survey for example has 92,536 users as of April 2016[3]. Administrative data infrastructures using existing access procedures are not able to carry such capacity. These users include students who are learning quantitative methods and the basic practices of social research. They also include researchers from poorer areas where science is under resourced and the transnational access provided by the likes of DwB is not feasible. These two user groups are highly disadvantaged in terms of accessing administrative data sources. Given these access issues, it would appear

[1] http://www.dwbproject.org/

[2] http://ec.europa.eu/research/openscience

[3] http://www.europeansocialsurvey.org/about/user_statistics.html

that administrative data is limited in its ability to fully replace survey data as the primary source of data for the social sciences.

2.2. Potential Role played by Social Surveys

By contrast to administrative data sources, social surveys have progressed towards an increasingly open approach to access that involves low costs and high usage. Both large scale data infrastructures and small scale research projects are encouraged to deposit data collected in accessible archives that are open to general research community. This serves to increase the scientific returns on the initial investment in data collection and increase transparency and subsequently the quality of science emanating from the data collected. Data infrastructures in particular have been tasked with providing as free and as open access as possible (Spichtinger, 2014). This is driven by a desire to maximize the returns on the substantial investments in data collection and act as a supply-side reform to boosting scientific output. The lower costs to conducting science, the more science can be conducted.

The social survey solution to this has been to take advantage of its respondent contact in order to circumvent concerns over data security. Europe's largest surveys generally assure users that the data will not include any direct identifiers and that users of the data will not actively look to identify respondents but stop short of ensuring anonymity. Effectively, social surveys simply assure respondents that the data will only be used for research. However once such assurance has been granted social surveys generally do not seek to eliminate individuals who represent an identifiable or near identifiable characteristics. For example, if a respondent is of an ethnic minority, has a large number of children and works in a small industry it is reasonable to expect them to be identifiable. Such information is highly prevalent in many social surveys but statistical analysis of the risk of identification is rarely if ever conducted by these social surveys (Mulder, 2014). This is because the phrasing of respondent agreements merely prohibits such identification rather than making it impossible. This is only made possible through the explicit consent of the respondent during the survey process.

This strategy offers administrative data a similar route through which to simultaneously increase access whilst ensuring data security. By requesting an individual's permission to use

their personal data for research purposes, it becomes possible to provide administrative data with reduced access controls and circumvent many of the issues identified earlier. One way to acquire such permission is through the social survey process itself. By acquiring permission for data linking during the survey process, the survey becomes a pivot within the administrative data environment and enables the scientific potential of administrative data to be more fully realized.

Within the Swedish GGS consenting to participation in the survey included the linkage with data records by Statistics Sweden who conducted the data collection. This meant that participation in the survey was dependent on consenting to record linkage. Regarding data quality, the response rate is therefore critical and in Sweden this was 54.7% which is comparable to the GGP average response rate of 55.7% (CITE). However, this does indicate significant loss through non-response relative to the direct use of administrative data. This is of course a considerable drawback to the use of a survey based approach over direct use of administrative data. As with most socials surveys, re-identification risk assessments were not conducted.

This basis of linkage consent enabled the fieldwork to pre-load administrative records into the survey process itself, enabling respondents the opportunity to correct the data where they deemed necessary. As Sweden is renowned for its well-integrated registers, this is regarded as non-controversial and common practice. Yet the added value of this implicit survey consent shifts the nature of the resulting data as the consent form refers to the distribution of data outside Sweden on the condition that direct identifiers are removed, does not mention anonymity and allows for the transfer of micro data files to researchers both inside and outside Sweden[4]. So whilst it is common for registers to be used for research purposes in Sweden, the added value of this process is that it enables the register data to be incorporated into an open access framework which even Nordic registers struggle to adapt to.

[4] Further information available via http://www.suda.su.se/ggs/Home_English.html

3. Data Validation

In addition to the increased access that is afforded by linking administrative data to survey data and the additional variables that such linkages bring, the survey data process also provides an opportunity to validate the data collected by administrative processes. Even though administrative data are often viewed as a gold standard, particularly on demographic issues such as births, deaths and marriages, there are a large number of reasons why administrative data could be erroneous. These include delays, misreporting, processing errors or incomplete records. By pre-loading administrative data into the survey process it is possible to effectively validate administrative data against the survey responses. Given that the majority of administrative data is process driven rather than collected for scientific purposes it is conceivable that survey responses and administrative data do not match.

In the example of the Swedish GGS, respondents were asked to validate their information on a number of indicators including their birth dates, partnership status, number of children, educational level, employment history and income. For demographic events there was a high degree of coherence between the survey responses and the population register. This is to be expected given that the process of registering births, deaths and partnerships is robust. However, on issues such as educational status there were far greater number of discrepancies reported. Of the 9,403 individuals within the sample (97.1% of all respondents) for which an educational status was extracted from the registers, 1,729 (18.3%) reported that it was incorrect. This indicates significant discrepancies in survey responses and administrative data records. Some of these would have been due to a lag between registry extraction in October 2011 and the fieldwork dates of late 2012-13 but it is unlikely to account for all of these discrepancies. It is also possible that these discrepancies occur for specific sub-groups such as migrants who attained a certain educational status outside of Sweden or who attained qualifications that do not fit within the conventions of the Swedish registries.

The further analysis of such issues is necessary in order to understand the limitations of both survey and administrative data and improve the quality of both. This in itself is vital if the potential of administrative data is to be realized.

4. The Added Value of Social Surveys

The primary logic for continuing social surveys in spite of the increasing availability of administrative data sources is that social surveys capture information that is complimentary to administrative data sources. As process driven data, administrative data is excellent at capturing behavior and outcomes. However, it is not capable of capturing a wide range of variables that are of interest not only to social scientists but also to administrative data owners. These include attitudes, norms, values and psychological concepts among others. In the Swedish GGS, demographic indicators such as individual's fertility history, partnership history, employment history, educational history and income can all be taken from registers, reducing the fieldwork time, reducing respondent burden and increasing data quality. However, many of the crucial variables in the conceptual design of the GGS lie outside of the scope of administrative data. These include gender attitudes, the distribution of household tasks, reported relationship quality or individuals plans for future family development. It is from these variables that the scientific advances of the GGS are achieved. The interaction between these concepts and the outcomes and behaviours identifiable within administrative data are of crucial importance and its certainly not feasible to supplant the social survey. Instead the survey should be seen as an enhancement of the administrative data.

In addition to this both administrative and data sources are limited in their ability to cover all social concepts required for cutting edge scientific research. To achieve such cutting edge research it is necessary to further link survey and administrative data with social media data, biometric data or commercial data. Such linking is harder to achieve in the administrative data environment given the aforementioned constraints of explicit consent and privacy that are apparent when using administrative data. Therefore, national data strategies and administrative data providers should be looking to engage with and build upon links with social surveys in order to enhance and strengthen their capacity so as to support cutting edge social scientific research.

5. Conclusions

The aim of this paper was to discuss the role of social surveys in the exploitation and scientific use of administrative data. The example of the Swedish GGS illustrates how the integration of administrative and survey data cannot only improve the survey data but also how the survey can improve the accessibility, quality and scientific relevance of administrative data sources. Given this administrative data sources should be looking to cooperate intensively with social surveys in order to fully realise the potential of administrative data sources.