Lithuanian Journal of Statistics *2012, vol. 51, No 1, pp. 5-16*
Lietuvos statistikos darbai *2012, 51 t., Nr. 1, 5–16 p.*
www.statisticsjournal.lt

DESIGN AND ESTIMATION OF SURVEYS TO MEASURE DATA QUALITY ASPECTS OF ADMINISTRATIVE DATA

James J. Brown[1], Oksana Honchar[2]

[1] Southampton Statistical Sciences Research Institute, University of Southampton
Address: University Road, Highfield, Southampton SO17 1BJ, UK. E-mail: jjb1@soton.ac.uk
[2] National Academy of Statistics, Accounting & Audit
Address: Pidgirna str., 1, Kyiv 04107, Ukraine. E-mail: ohonchar@list.ru

Received: June 2012 Revised: September 2012 Published: November 2012

Abstract. National Statistics Institutes (NSIs) have been increasingly seeking to replace or enhance traditional survey-based data sources with administrative data sources; with the aim to improve overall quality in the absence of a definitive register of the population. The Beyond 2011 Census Programme in England and Wales is an example of looking to replace a traditional census with administrative data collected for another purpose by a different organisation, when there is no definitive register as a starting point. There are also similar projects across NSIs within the area of business surveys looking to use administrative sources to reduce cost and burden. In this paper we start with considering all aspects of a quality framework for administrative data and then focus on the elements relevant to data quality such as accuracy and coherence. We fit these concepts into the framework for total survey error highlighting the components an NSI needs to measure to produce estimates based on the administrative data. We then explore the use of both dependent and independent quality surveys to adjust the administrative data for 'measurement' and 'coverage' aspects to improve the quality of estimates produced from the administrative data.

Keywords: Quality Surveys, Administrative Data, Estimation.

1. Introduction

A National Statistics Institute (NSI) is typically interested in generating statistical information [11] relating to some aspect of a population. This statistical information may be simple summaries of the data for individual units (means, totals, ratios) or more sophisticated statistical information such as regression parameters, index numbers, and underlying trends. When an NSI designs a data collection exercise to measure some aspects of a population it will typically consider balancing (see Fig. 1) the following three areas; direct financial costs to collect and process the data to create the required statistical information, burden (time and financial) on the units providing the data, quality of the statistical information (estimates) produced at the end.

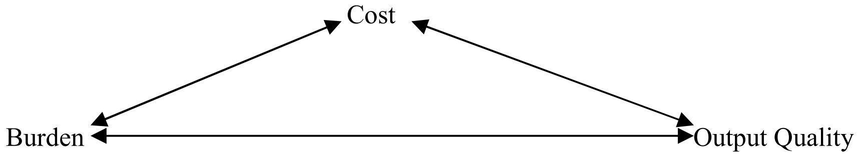


Fig. 1. Three-Way Trade-Off in Data Production

Typically an NSI would consider overall quality in terms of whether the statistical information produced from the data is 'fit for purpose'. In other words, can we produce relevant and coherent estimates in a timely and punctual manner with a level of accuracy appropriate for their use (output quality in Fig. 1) while balancing against the cost and burden? Increasingly, to help with the balance we want to utilise administrative data. An example is the 'Beyond 2011 Programme[1]' being undertaken by the Office for National Statistics (ONS) to explore the possibility of replacing a traditional census in England and Wales with a system based around administrative data sources. Unlike census or surveys, by definition the collection of such data is NOT under the direct control of the NSI with implications for controlling and measuring output quality. However, in terms of the trade-off there is at least a perception that

[1] http://www.ons.gov.uk/ons/about-ons/what-we-do/programmes---projects/beyond-2011/index.html

administrative data use can reduce financial cost[2] and respondent burden[3] as the data is already being collected for another purpose. Therefore, in this paper we consider the issues with understanding and measuring output quality when statistical information is generated by an NSI using administrative data. For our purposes we consider administrative data to be:

- collected for the purpose of managing and maintaining a system (healthcare, tax, education, …) typically unrelated to the purpose the NSI wish to use it for,
- collected independently of the NSI,
- collected from all units **participating** within the system (no sampling).

In Section Two we discuss aspects of output quality and utilise existing frameworks for survey data to fit our definition of administrative data highlighting where the errors occur that impact on output quality. In Section Three we present two examples relating to the Beyond 2011 Programme that demonstrate the potential uses of administrative data and how their quality issues fit within our frameworks. In Section Four we propose approaches to quality measurement and adjustment using survey approaches conducted both independent of and dependent on the administrative data. Finally, in Section Five we finish with some conclusions.

2. A quality framework and its various dimensions

There already exist several 'Quality Frameworks' with well-defined dimensions of output quality. We take as our starting point the Eurostat quality components [6], which are embedded within the three hyper-dimensions of [5], and consider them in the context of the output quality of statistical information generated from administrative data.

Relevance

- Is the administrative data able to generate statistical information on the aspects of the population relevant to the users? There may be a lot of administrative data available but it may not contain the variables needed to generate the statistical information required by users.

Accuracy

- Does the statistical information estimated from the administrative system have a small error in relation to the true underlying value in the population of interest? Administrative data count those that are entitled and then participate in the system, with participation often near 100% due to legal requirements. This means that accuracy is purely in terms of coverage and measurement errors leading to errors after aggregation (see [14]) rather than random sampling error. However, coverage and response can be considered as phases of sampling for the purposes of estimation.

Timeliness and Punctuality

- In business statistics the available administrative data (say annual accounts) may not be timely enough for the purposes of quarterly National Accounts and other economic reporting. Conversely, in social statistics administrative data on the population may be timelier than say a census, which can take up to two years to produce the full set of results. Punctuality may be impacted on as the NSI does not directly control the collection of the core data, any delay in the supply of that data to the NSI may impact on the punctuality of outputs.

Accessibility and Clarity

- Accessibility to the core administrative data can be a serious barrier to its use by an NSI but this is not the issue referred to here. We are concerned with the accessibility and clarity of the statistical information produced from the data to the users of the information. Therefore, there is no reason why these should differ whether administrative data or a survey provided the input. However, when using administrative data, an NSI may find it more difficult to provide the appropriate meta-data for users as they did not control the primary collection of the data.

Comparability

- If using administrative data are definitions stable over time? Changes within the administrative system may damage comparability of statistical information.

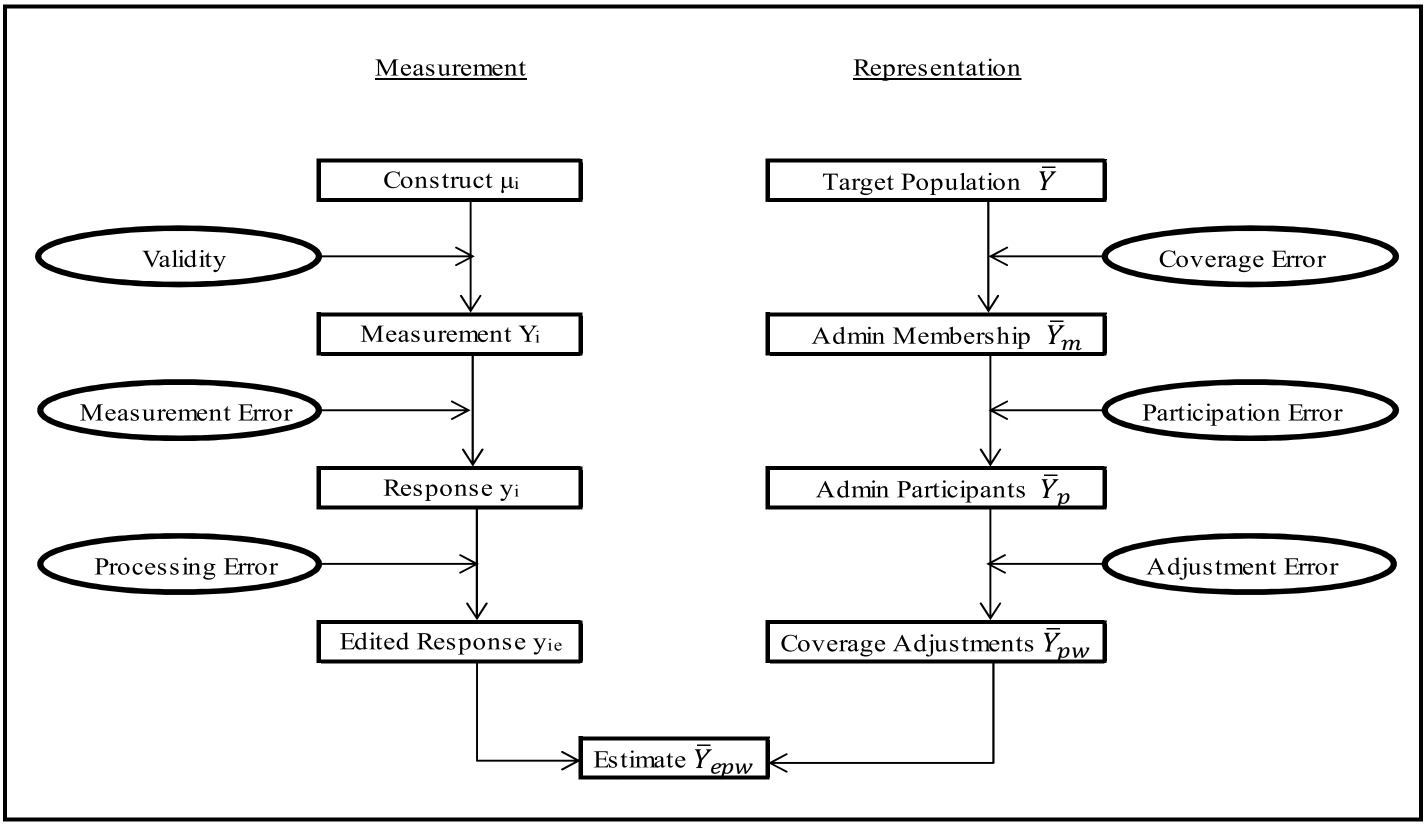
[2] Utilising the data will not be free as there will still be processing costs.

[3] If the respondent is already providing the information to an administrative system then providing the same data when responding to a census/survey is clearly adding to the burden.

Coherence

- Are common definitions applied across difference sources so that the statistical information produced provides a coherent picture of various aspects of the population? For example, we may have data on income by ethnicity from one source and health status by ethnicity from another. If we want to show a coherent picture of the relationship between ethnicity and poor health with ethnicity and poverty then ethnicity must be measured in a consistent way across the two sources.

We can link these dimensions of output quality to the idea of data quality from the perspective of the life cycle for extracting statistical information from an administrative system. Fig. 2 is based on the approach for surveys outlined in [8] but put into the context of an administrative system. We can see that the quality cycle for administrative data is very similar but without the sampling error aspect. Validity, measurement error, and processing error all link to the dimensions of accessibility and clarity, comparability, and coherence. Is the statistical information generated from the administrative data actually measuring the right concept? Coverage and participation link to our dimension of accuracy while adjustment is about what we do with the data when generating statistical information to improve output quality across all the dimensions. Complex adjustments requiring additional data collection may impact on timeliness and punctuality.



Based on the Survey Life Cycle; [8], p.48.

Fig. 2. Life Cycle to Generate Statistical Information from Administrative Data

On measurement side of Fig. 2, we still have the underlying concept we wish to measure for each variable within the data. The administrative system will have a measurement tool to collect the variable and that may not line-up well with the concept (validity). With an administrative system the NSI will typically not be 'in control' of defining the measurement tool and a concept like ethnicity may have a poorly defined measurement tool. Using the tool to collect the data will typically involve some interaction with the participants[4] of the administrative system. This will lead to measurement error but we may expect the extent to vary across variables depending on their importance to the

[4] Participants may be users of the system (patients in a healthcare system) or those actually involved with the administration of the system (teachers in a school system).

administrative system. Basic variables such as age and sex of patients in a healthcare system can be so important that it would be reasonable to assume they are recorded almost perfectly. Less effort may be given to collecting secondary attributes like ethnicity leading to increased measurement error. These different sources of measurement error are shown in detail in Fig. 3, adapted from Figures 2.2 and 2.3 in [11]. We also highlight the fact that the NSI will usually have little or no control over the processes generating the measurement error.

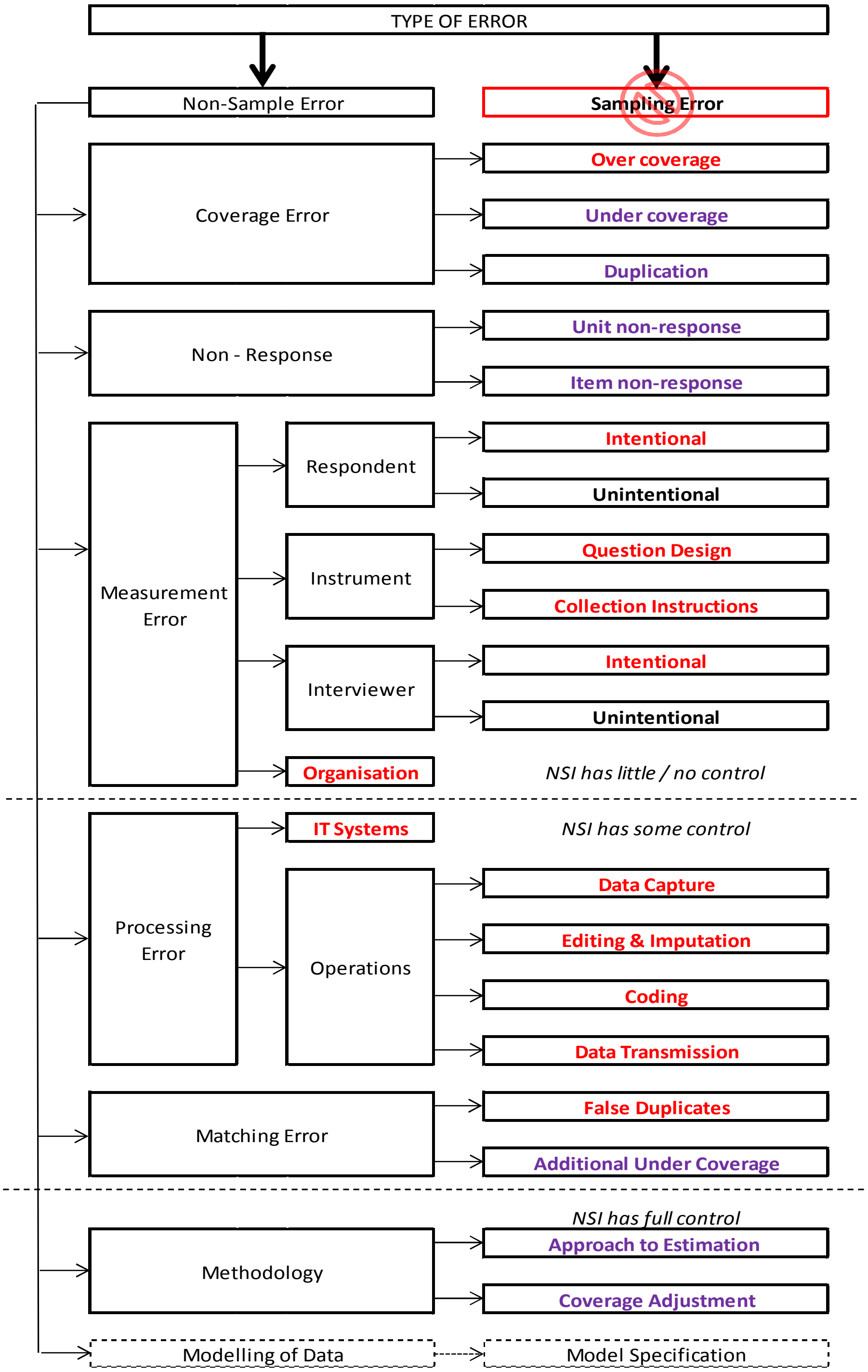


Fig. 3. Errors that Impact on the Output Quality of Statistical Information

The collected data is then input into the administrative system leading to the possibility of processing errors. In many cases the data will be directly entered into the system at the point of collection but not always. As with measurement errors we would expect the effort made to minimise such errors will depend on the importance of the variable to the administrative system. In Fig. 3 the various components of processing error are shown. We also highlight that some processing will be within the control of the NSI, for example it may well apply editing and imputation to the raw data supplied by the administrative system, but the core processing that generates the basic database will not be within the control of the NSI. This core processing will include data validation of entered data that may result in editing and imputation outside the control of the NSI.

Following the processing stage there may also be a record linkage exercise to increase the utility of the final database by combining information across two or more administrative sources at the unit level. Such an exercise will generally be under the control of the NSI, but in the absence of a unique identifier across sources will result in matching errors that will certainly impact on the output quality when the combined data are aggregated. As discussed in [14], the result will be missed matches causing duplicated units in the final database if the two seemingly unrelated records are left in. Alternatively, if the un-matched records are discarded due to having an incomplete set of variables then additional under-coverage is created in the final database.

On the representation side of Fig. 2, we will still start with defining the target population we wish to measure each variable on and then a population level quantity (mean, total, ratio) defined in terms of the individuals' records within the target population. As with surveys where there is often an error in coverage between the target population and the frame, so with administrative data there will be a difference between our target population and those defined as members of the administrative system. We will also get unit non-response in the survey from refusals and non-contacts while with administrative data not all members will participate, although levels of participation will likely be considerably higher than response rates achieved in a survey.

In Fig. 3 we break-down these coverage errors into different components highlighting that the NSI will usually have little control over these issues within the administrative system. In some specific cases over-coverage can be directly corrected by the NSI once it receives the data provided the records contain the information to identify the over-coverage with respect to the NSI's population of interest. Membership rules can result in both over-coverage and under-coverage at the macro-level in the same way that a sample frame can contain out-of-scope units as well as miss units belonging to the target population. Records that are erroneous with respect to our target population exist within the administrative systems while others are excluded that do belong to the target population. The way members participate within the system will also have an impact. When an individual migrates but maintains membership this will create over-coverage in one geographic area with a corresponding under-coverage in another, which will net-out at the macro-level, until participation leads to the record's location being updated. This may occasionally also create duplication, with a new record being created in the new location without removing the over-coverage from the incorrect location. This will not net-out at the population level but the duplication can be found because the real record exists within the system. Participation will also result in additional erroneous over-coverage at the macro-level from records relating to emigration and death remaining in the administrative system once membership has finished. It will also contribute to macro-level under-coverage as there will be a delay in new members such as immigrants joining the system. We will discuss this further in Section 3 with the use of examples and this distinction between macro and micro over-coverage will be important for measurement in Section 4.

3. Examples

In this section we present some examples of administrative data that NSIs in the UK would like to use with respect to the Beyond 2011 Programme. The research is investigating the potential of administrative data to replace the traditional census post-2011 in a situation where there is not a central population register. Therefore, the NSI wishes to utilise a variety of administrative sources that it does not have control over.

A) Pupil Level School Census Data (England)

All State schools in England complete a census of their pupils with key characteristics[5]. This takes place each term and is combined over time to track pupils through the school system and monitor school-performance in national tests. Each pupil is issued with a unique identifier when they enter the school system and the link to performance ensures that schools work effectively to complete the census returns.

[5] More details on the data and how to access it can be found at http://www.bristol.ac.uk/cmpo/plug/

Coverage Errors and Non-Response: It is in a schools interest to ensure it makes a correct return so we can consider that coverage errors and non-response within the system from participation (micro-level) are minimal; near 100% coverage of the State School system with up-to-date locations. However, in terms of the population of children in England there are coverage problems with respect to home-school children and those attending private schools. This creates macro-level under-coverage due to membership of the administrative system. In addition, at a local level parents do 'play the system' by having a residence within a school's catchment area and registering that as the address of the child; while in fact living in the family home in a different area. This effectively creates micro-level coverage issues if trying to create counts of the local population of children that net-out at higher levels of aggregation.

Measurement Errors: Unlike a survey the data are not collected in a systematic way with clear rules and definitions being universally applied. Each school makes a standard return but the way they choose to complete the return varies from an individual secretary filling out the form to detailed information being collected from the parents of each pupil and then collated by the school. For key variables such as age and sex this differential approach to measurement is not problematic. We expect the system to have very accurate measurement of these variables, especially age, as it relates to where a child should be within the system. However, a variable like ethnicity has more measurement issues. Firstly, it is a less clearly defined concept and a person must self-define their membership to a particularly category given the definition. Therefore, without clear rules it is difficult for pupils to be consistently assigned across schools. Secondly, this lack of control over who provides the information on individual children creates more uncertainty. Therefore, we have issues with both the respondent and the instrument.

B) National Health Service Registration (England and whole UK)

Within the UK each individual is required to register with a local doctor (GP) as the main way of accessing the National Health Service (NHS). Each GP Practice holds a list of patients and payment from the NHS to the Practice is driven partially by this list of patients, and to some extent their health-related characteristics. Each person is issued with an NHS Number at birth (or when they first register as an international in-migrant) although historically individuals can be issued with a second number or be on the database twice with the same number.

Coverage Errors and Non-Response: It is in a GP's interest to maximise the size of their list of patients. However, unlike the school example where families will immediately register their children with a local school resulting in de-registration from the previous school, there is not the same motivation amongst patients to re-register when they move. Therefore, at any point in time a GP's list will contain genuine patients; macro-level over-coverage from patients that have died or emigrated (over-coverage due to erroneous membership); micro-level over-coverage from patients that have migrated internally to outside of the GP's area (over-coverage linked to under-coverage and potential duplication). This means that when considering the population of an area based on the GP lists it will have over-coverage due to delayed re-registrations by internal out-migrants, under-coverage due to delayed first registration by international immigrants and internal in-migrants, over-coverage due to delayed de-registrations for deaths and international emigrants. There is the additional problem that the population definition for inclusion on a GP list is not clear. This interacts with the timing of first registration by international immigrants resulting in additional under and over coverage with respect to a standard usual resident definition from the membership rules of the system.

Measurement Errors: The NHS system contains very few details relating to an individual beyond health related characteristics. For example, ethnicity is not consistently or uniformly recorded and its collection is not required. However, given the importance of age and sex, and the fact that the majority of individuals within the system enter as a result of a birth, it is reasonable to assume that these two key characteristics would be measured near perfectly for those on a GP list.

4. Quality survey design

The sampling error in Fig. 3 is perhaps unique amongst the errors associated with statistical information in that it can be quantified (through standard error calculations) just with the achieved sample data, provided a sensible approach is taken to estimation. The unintentional errors during data collection (in **black**) will generally be random and so also feed into the sampling error. Of course, pure sampling error is not a source of error when working with administrative data, although the random element in the measurement errors will still exist. The other sources of error (in **red** and **purple**) will generally have some systematic (bias) component, as well as a random component especially as unintentional errors are mixed with intentional errors. While adjustments can be made in estimation based on *sensible* assumptions with the expectation of reducing the impact of such errors;[6] the actual level of the error cannot be quantified without additional information. This will often be in the form of a Quality Survey[7] and can either be dependent or independent of the original data collection, in our case the administrative data system.

4.1) Dependent Quality Survey (DQS) for Measurement Errors

In Fig. 2 we distinguish errors that will impact on the quality of the statistical information in terms of measurement and representation. In Fig. 3, those processes that result in measurement issues are shown in **red**. We have grouped over-coverage with measurement error, although they are clearly generated by different processes. It is convenient to think of over-coverage as a form of measurement error, gaining a response when no data should be returned, when it comes to estimating from a quality survey. Therefore, measurement issues relate to the impact of the data we have collected on the quality of the statistical information produced, not data that is missing from the collection.

It is accepted (see [2]) that matching to administrative data for quality checking of surveys is not ideal due to measurement issues with the administrative data. Instead, [2] refers to the gold standard as a re-interview with reconciliation. This type of Quality Survey is by definition a dependent survey design as those re-interviewed depend on the original respondents to the survey. With highly trained interviewers we can seek to impose the correct concept and then seek to reduce measurement error (and subsequent processing error) to close to zero. Often the interviewer will repeat the questionnaire with the respondent and then reconcile any differences with the original answers; hence the dependent nature of the design. This will find erroneous respondents as the reconciliation will confirm the individual is not part of the target population.

When linking a DQS with a standard population survey, there will be the issue of burden on those respondents sampled for the DQS but not an issue with confidentiality as they gave the original information as part of the survey, which you are checking. However, it is less clear that it would be acceptable, or even legal within confidentiality frameworks, for a DQS interviewer to reconcile discrepancies with say an individual's medical record unless the sole purpose of the DQS was to audit the quality of a GP's Patient List. In other words, an NSI seeking to create statistical information from another organisation's administrative data will find it difficult, and in the UK context illegal, to conduct a full DQS with reconciliation in the field on the administrative data.

4.2) Independent Quality Survey (IQS) for Representation Errors

In Fig. 3, those processes that result in errors of representation are shown in **purple**. Representation errors relate to the impact of the data we **failed to collect** on the quality of the statistical information produced, and the subsequent processes done to alleviate the impact of the missing data. Within the field of population census there is a long history of such surveys, with the specific aim of coverage assessment, a recent example being the 2011 Census Coverage Survey for England and Wales outlined in [4]. They are conducted independently of the original census and aim to measure those missed by the census. Therefore, by definition they cannot depend on the responses obtained by the original data collection. The independence also has advantages in estimation as it allows for dual-system estimation to be applied adjusting for those *missed by both* the census and the survey ([15]; [10]). Additional external information can be included to relax the independence assumption ([16]; [3]); although finding such external information is by its very nature difficult.

This is an area where administrative data is at an advantage as undertaking an IQS,and applying such approaches to estimation, becomes more plausible given that the original collection of the data was done for an independent purpose. You also do not have the legal and acceptability issues of a DQS as any matching takes place after the data have been collected independent of the administrative system. However, unlike a coverage survey for a census, where both are under-taken with the same concepts and measurement tools, it is not clear that an IQS will measure the population in a similar way to the administrative data system that is having its coverage assessed. In fact, as the IQS data collection will be controlled by the NSI it should at least start from the desired concepts even if there are some issues with the measurement processes in the field. Therefore, we would expect measurement differences between the

[6] A classic example would be re-weighting the achieved data via calibration with the aim of reducing non-response bias.

[7] We do not consider a follow-up survey of non-respondents as a true Quality Survey. Its role is to reduce the impact of non-response bias in the original survey rather than quantify the actual bias and correct for it.

two collections, and without a further survey to reconcile differences (but see the discussion on the acceptability of this in the previous section) it is difficult to argue that one is always superior to the other with respect to measurement errors.

4.3) Combined Approach

As we have already stated, administrative data cannot, in general, be seen as a gold-standard measurement (i.e. pupil ethnicity on the school census) while an IQS will have its measurement controlled by the NSI but will not be without some measurement error. However, we can argue that a limited set of variables, typically those directly related to the system the administrative data supports, will be of very high quality (i.e. age and sex of pupils). Therefore, a linkage between an IQS and the administrative data post the survey collection potentially allows for some adjustment in both directions. We can argue that for these key variables the administrative data can reassure us that the IQS has low measurement error. Then we can use the IQS to at least improvement the quality of other variables on the administrative data arguing that it has low measurement error for the key variables and is under the direct control of the NSI for the measurement of other variables and concepts. In addition, as the IQS is conducted by the NSI there will be information on the design of the data collection tool, and its implementation in the field, to support further any claims of low measurement error. It would also be possible to conduct an additional small DQS on the IQS respondents to at least qualitatively assess any residual measurement error in the full IQS,

As already discussed in Section 4.2, linkage between an IQS and the administrative data can estimate coverage errors and we are now proposing that it can at least improve data quality with respect to measurement errors. An IQS can also measure and adjust for duplication (Fig. 3) as it will find the correct location and linkage can identify the duplicate. Likewise, it can find individuals that belong to the target population but are over-coverage in one area and under-coverage in another. Such an approach for dealing with micro-level over-coverage is discussed in [12] for the 2011 Census Coverage Survey in England and Wales. However, an IQS cannot estimate directly the impact of macro-level over-coverage caused by the continued erroneous inclusion of ex-members or out-of-scope members within the administrative data. As the IQS is based on an independent sample from the population of interest it cannot *by definition* find units that remain in the administrative system but no longer exist in the population of interest.

In many countries, universal coverage by the death registration system offers the potential for administrative data to be cleaned of this type of erroneous individual if they remain in the system after death. However, de-registration of international emigrants is an issue, even for countries with good population registers. It can be possible to limit the impact of emigrants by applying rules to the administrative data, for example removing individuals from a GP list who have not had an interaction within a certain period. Of course, how to define such rules is problematic. Matching across different administrative systems may help; an individual with a tax record and a health record for the same address is less likely to be an emigrant than someone with no recent activity on their health record and no associated tax record. But again, this relies on applying rules that will be difficult to test.

In the US, the E-sample ([9]; [10]) is carried-out to assess whether a census return is erroneous once it has been identified as not matching to the P-sample (the independent coverage survey). Ideally, we would like to avoid that second survey process. If the IQS is using an address list as a frame that is also the basis of the administrative data system, with computer assisted interviewing the computer could be pre-loaded with the basic characteristics of the individuals registered to an address. This would allow for an independent collection of data for the IQS with the opportunity for the interviewer to then reconcile with the respondent those individuals on the administrative data but not identified by the IQS. Therefore, an estimate of those erroneously included on the administrative data would be possible. Such an approach could work but would be subject to the confidentiality and acceptability issues faced by a traditional DQS when operating to support the use of administrative data. The survey respondent will wonder why the interviewer is questioning them about individuals they have not included in the household and there is an ethical issue regarding asking someone about another individual that they have not offered information on, such as a previous resident of an address. Another way might be to use one of the indirect approaches that is often used to measure mortality for countries with poor registration data using surveys (see for example [7]). In this case we would ask if the survey respondent had a close relative that had emigrated in some recent period of time generating an indirect sample of international emigrants. We could then search the administrative data to check for de-registration and therefore estimate the level of erroneous inclusion. The sample would need to be adjusted for multiplicity, requiring additional questions on the IQS, but this may be more acceptable to respondents, as well as fitting within legal frameworks, than the interviewer having access to the administrative data of the respondents.

4.4) Estimation

In this section we explore using the ideas of the previous sections to estimate quality from a single IQS. We assume that as discussed in Section 4.3 we have, at least qualitatively, confirmed the IQS has low measurement error by comparing to variables in the administrative system, such as age and sex, which we expect to be measured with near zero error by the administrative system. Further, deaths have been removed by linkage to death registration and emigrants have been removed by applying rules on activity or through reconciliation with other administrative sources. Therefore, we concentrate developing an estimation strategy that can adjust for all other coverage errors (under-coverage, duplication, over-coverage due to wrong location) and deal with measurement errors assuming the IQS has lower measurement error.

Let $Y_i^{(c)}$ be the survey (true) response for individual i taking the value 1 if they belong to category c = 1, …, C of variable Y and 0 otherwise. Therefore, a standard estimate of the population total for category c, based on the sample of respondents s, would be $\hat{Y}^{(c)} = \sum_{i \in s} w_i Y_i^{(c)}$ where w_i is the sampling weight associated with individual i. The overall total of the population is then estimated by summing $\hat{Y}^{(c)}$ across the categories.

Clearly, we expect that the survey will have non-response so let us assume we have administrative data where $X_i^{(c)}$ records individual i belonging to the same category c (i.e. sex with categories male or female and no measurement difference between the survey and administrative data) with the known total on the administrative data given by $X^{(c)}$. We do not expect the administrative data to have full coverage of the population[8] but by using the survey respondents s_m that match to the administrative data we can define a calibration weight $w^{(c)} = \frac{X^{(c)}}{\sum_{i \in s_m} w_i X_i^{(c)}}$ (just a g-weight for GREG or in this case a post-stratification adjustment) to adjust for the survey non-response in category c. For simplicity we have assumed a single adjustment factor but this will often be embedded within estimation strata, either from the sample design or a post-stratification exercise, with the factor defined by membership of the estimation strata. In the context of A from Section 3, this factor will adjust the survey for its non-response with respect to State-School children. However, we also need to apply the calibration to the survey respondents s_n that do not match to the administrative data (representing those children within the private school system) so that our new estimator for $Y^{(c)}$ is given by

$$\tilde{Y}^{(c)} = \sum_{s_m} \tilde{w}_i^{(c)} Y_i^{(c)} + \sum_{s_n} \tilde{w}_i^{(c)} Y_i^{(c)} \qquad (1)$$

where $\tilde{w}_i^{(c)} = w_i \times w^{(c)}$. This is essentially the form of the PREG estimator, developed in [1], which is used by the Australian Bureau of Statistics (ABS) to adjust for non-response in the census as an auxiliary variable for the post-enumeration survey that estimates the true resident population. The 'P' stands for prediction because the second term in (1) is predicting the non-response adjustment for those missed from the administrative data based on the adjustment made for those covered by the administrative data. This is now working like a dual-system estimator ([13]). It is assuming that response to the administrative data is independent of response to the survey and that the survey response rate is constant across individuals within an estimation stratum[9] for the PREG. In the context of A in Section 3 this implies that non-response in a survey of resident children is independent of whether the child attends a state or private school.

So far we have assumed that measurement in the survey and administrative data is consistent for those respondents s_m that match. In the context of A in Section 3 we would expect that to hold for child characteristics such as age and sex

[8] For example, A discussed in Section 3 only covers resident children in the State-School-System while the survey may be attempting to cover all resident children.

[9] Equation (1) can be embedded within a set of estimation strata within which we would expect the survey response rate to be close to homogeneous.

but be more problematic when dealing with a concept like ethnicity. However, suppose that due to measurement errors[10] individual i belongs to category k = 1, …, K of variable X (say ethnicity) on the administrative data where we expect that $X_i^{(k)}$ will often agree with $Y_i^{(c)}$ as measured by the survey. Given that the survey is measuring the 'correct' underlying variable and category for individual i, and that measurement errors in the administrative data are independent of whether an individual responds to the survey; using the matched survey respondents s_m we can estimate the transition probability between the true category and the category on the administrative data as

$$\hat{p}_{k|c} = \frac{\sum_{s_m} w_i\left(Y_i^{(c)} \times X_i^k\right)}{\sum_{s_m} w_i Y_i^{(c)}} \quad . \tag{2}$$

Our ability to estimate these transition probabilities well will depend on the extent to which the **red** measurement errors in Fig. 3 also include random noise from unintentional errors.

We now define a new calibration weight for the non-response in the survey as $w^{(k)} = \dfrac{X^{(k)}}{\sum\limits_{i\in s_m} w_i X_i^{(k)}}$ and re-write equation (1) as

$$\tilde{Y}^{(c)} = \sum_{s_m} w_i Y_i^{(c)}\left(\sum_{k=1}^{K} w^{(k)} X_i^{(k)}\right) + \sum_{s_n} w_i Y_i^{(c)}\left(\sum_{k=1}^{K} w^{(k)} \hat{p}_{k|c}\right). \tag{3}$$

The first term in (3) adjusts the survey for non-response based on the calibration to the administrative data category. This is just like using an auxiliary variable X in small area estimation that is correlated with the variable of interest Y. The second term in (3) then has to predict, using the transition probabilities given in (2), the administrative data category for the respondents not covered by the administrative data and then apply the appropriate calibration constraint to adjust for survey non-response. In the context of A in Section 3, this is equivalent to predicting how the administrative data would record ethnicity for private school pupils and then applying the state school adjustment for the survey non-response given the relationship between true ethnicity (measured by the survey) and the administrative system's measurement of ethnicity.

The performance of the estimator in (3) to correct for errors will depend on the extent to which they are systematic and random. When there is a strong systematic component to both the **purple** and **red** errors in Fig. 3, an IQS will do well at estimating the nature of the errors. However, when there is a strong variable component in the coverage errors then utilising the administrative data as auxiliary information to calibrate the survey will not perform as well and likewise a strong variable component in the measurement errors weakens the estimation of (2) and therefore impacts of the variability coming from the second term in (3).

A key coverage issue for administrative data is over-coverage. This can be either recording an individual in the wrong place (delayed re-registration by internal migrants on the NHS registers) or the completely erroneous inclusion of an individual (delayed de-registration of deaths and international emigrants from the NHS registers). In the context of counting the usual resident population using the census in Australia, which is a person present base, the former are those that are in Australia but not at their usual residence on Census Night while the latter are those individuals visiting Australia on Census Night. As discussed earlier, this latter type of over-coverage cannot be dealt with by a simple independent survey[11] but the former can be.

Let us define an additional variable $Z_i^{(k)} = 0, 1, 2, \ldots$ that is the number of times individual i is recorded on the administrative data in category k. We now define the calibration weight as $w^{(k)} = \dfrac{Z^{(k)}}{\sum\limits_{i\in s_m} w_i Z_i^{(k)}}$ to allow for the multiple

[10] We have discussed the ways that measurement errors can occur within the administrative data.

[11] The Census in Australia removes these individuals based on residency questions in the data – with administrative data some can be removed based on (lack off) activity but it is not straightforward.

returns in the administrative data but then apply this weight within the same framework as (2) and (3). Therefore, we are reflecting the over-coverage and the measurement error. This general framework can also cope with the situation where an individual is counted twice on the administrative data but in two different categories, such as when the categories are geographic areas and an individual is double-counted by the administrative system at two different locations. While not elaborated here, in [1] the whole approach is then embedded within a GREG-type framework to create weights that: utilise several X variables as calibration constraints, cope with the measurement issues when X and Y should match, apply to any Y that is correlated with the X's.

5. Conclusions

In this paper we have discussed how data quality frameworks apply when an NSI uses administrative data to generate statistical information. We have presented some administrative data examples from the UK, where the Office for National Statistics is under-taking a research programme to explore replacing the traditional census collection with administrative data to generate census-like statistical information on the usual resident population. These examples highlight the measurement and coverage errors that occur with administrative data and we show how the approach used by the ABS to measure census coverage can be applied with an independent quality survey to create statistical information adjusting for (some of) the measurement and coverage errors within the administrative data.

One of the key assumptions here is that the independent quality survey has little or no measurement error on key variables. We proposed that this can be assessed qualitatively through comparison with variables we expect the administrative data to measure with high accuracy due to their importance to the system that the data relates to. Further work is needed to consider what could be done if this qualitative assessment called into question the assumption of little measurement error in the independent quality survey. Is it possible to make simultaneous adjustments to both data sources for measurement errors or is the only solution an additional dependent quality survey that gives adjustments to the IQS without making reference to the administrative source?

A second area is that of erroneous members on the administrative data that are not duplicates of actual members creating macro-level over-coverage. By definition, an independent quality survey cannot measure directly the impact of these erroneous records on data quality as they do not exist in the population to be sampled by the survey. However, rules can be applied to remove many of them from the administrative data prior to assessing coverage and measurement errors. Further work is needed to explore the idea of using multiplicity sampling within an independent quality survey to generate a sample of international emigrants to check against the administrative data for erroneous inclusions.

References

1. Bell, P., Clarke, C. and Whiting, J. (2007) An Estimating Equation Approach to Census Coverage Adjustment. *ABS Research Paper* **1351.0.55.019**.
2. Biemer, P. and Lyberg, L. (2003) *Introduction to Survey Quality*. Wiley: New York.
3. Brown, J., Abbott, O. and Diamond, I. (2006) Dependence in the 2001 one-number census project. *Journal of the Royal Statistical Society Series A*, **169**, 883-902.
4. Brown, J., Abbott, O. and Smith, P. A. (2011) Design of the 2001 and 2011 census coverage surveys for England and Wales. *Journal of the Royal Statistical Society Series A*, **174**, 881-906.
5. Daas, P., Ossen, S. and Tennekes, M. (2010) Determination of administrative data quality: recent results and new developments. European Conference on Quality in Official Statistics 2010, Helsinki, Finland. http://q2010.stat.fi/papersbig/
6. Ehling, M. and Körner, T. (2007) Handbook on Data Quality Assessment Methods and Tools. Eurostat: Wiesbaden.
7. Graham, W., Brass, W. and Snow, R. W. (1989) Estimating Maternal Mortality: The Sisterhood Method. *Studies in Family Planning*, **20**, 125-135.
8. Groves, R., Fowler, F., Couper, M., Lepkowski, J., Singer, E. and Tourangeau, R. (2009) *Survey Methodology (2nd edition)*. Wiley: New York.
9. Hogan, H. (1993) The 1990 Post-Enumeration Survey: Operations and Results. *Journal of the American Statistical Association*, **88**, 1047-1060.
10. Hogan, H. (2003) The Accuracy and Coverage Evaluation: Theory and Design. *Survey Methodology*, **29**, 129-138.
11. Honchar, O. (2011) Q*uality Assurance of Statistical Information: Methodology and Organization.* – Format: Kyiv.
12. Large, A., Brown, J., Abbott, O. and Taylor, A. (2011) Estimating and Correcting for Over-count in the 2011 Census. *Survey Methodology Bulletin*, **69**, 35-48.
13. Sekar, C. C. and Deming, W. E. (1949) On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association*, **44**, 101-115.

14. Wallgren, A. and Wallgren, B. (2007) *Register-Based Statistics: Administrative Data for Statistical Purposes*. Wiley: Chichester.

15. Wolter, K. M. (1986). Some Coverage Error Models for Census Data. *Journal of the American Statistical Association*, **81**, 338-346.

16. Wolter, K. (1990) Capture-Recapture Estimation in the Presence of a Known Sex Ratio. *Biometrics*, **50**, 1219-1221.

ADMINISTRACINIŲ DUOMENŲ KOKYBĖS MATAVIMO TYRIMŲ PLANAS IR VERTINIMO METODAI

James J. Brown, Oksana Honchar

Santrauka. Valstybinės statistikos įstaigos vis dažniau bando patikslinti tradiciniu (apklausos) būdu gautus statistinių tyrimų rezultatus arba pakeisti juos administraciniais duomenimis, kuriais siekiama pagerinti tyrimo kokybę, kai nėra gero populiacijos registro. Anglijos ir Velso gyventojų 2011 metų surašymo programa yra pavyzdys, kaip tradicinį surašymą siekiama pakeisti kitos organizacijos ir kitais tikslais surinktais administraciniais duomenimis, neturint gero registro. Statistikos įstaigos taip pat vykdo panašius projektus verslo statistikoje, bandydamos pasinaudoti administraciniais šaltiniais, kad sumažintų kaštus ir naštą respondentams. Šiame straipsnyje aptariami administracinių duomenų kokybės aspektai, susitelkiant ties tikslumo ir suderinamumo elementais. Įtraukiant šias koncepcijas į bendros tyrimo paklaidos rėmus, išryškinamos tos komponentės, kurios statistikos įstaigoje turi būti matuojamos, skaičiuojant įverčius iš administracinių duomenų. Nagrinėjamas tiek priklausomų, tiek nepriklausomų kokybės tyrimų naudojimas, siekiant įvertinti administracinių duomenų matavimo ir aprėpties aspektus bei pagerinti įverčių, gautų iš administracinių duomenų, kokybę.

Reikšminiai žodžiai: kokybės tyrimai, administraciniai duomenys, vertinimas.