Hand D.J. (2018) Statistical challenges of administrative and transaction data (with discussion). *Journal of the Royal Statistical Society*, *Series A*, **181**, 555-605.

Statistical challenges of administrative and transaction data

David J. Hand
Imperial College London
and
Winton Capital Management
{d.j.hand@imperial.ac.uk}

Abstract:

Administrative data are becoming increasingly important. They are typically the side-effect of some operational exercise, and are often seen as having significant advantages over alternative sources of data. While it is true that such data do have merits, statisticians should approach the analysis of such data with the same cautious and critical eye they approach the analysis of data from any other source. This paper identifies a number of statistical challenges, with the aim of stimulating debate about and improving the analysis of administrative data, and encouraging methodology researchers to explore some of the important statistical problems which arise with such data.

Keywords:

Operational data; management data; data quality; big data; repurposed data

1. Introduction

Administrative data are data generated during the course of some operation, and then retained in a database. They are becoming increasingly important as the potential for discovery from such data sources is being recognised and as alternative data sources become more costly or difficult to use (e.g. because of declining response rates in surveys). In the main, this means that the analysis of administrative data is *secondary* - the data are being *repurposed* - although, as explained below, this is not always the case. The existence of large, often administrative, data sets, offering potential for secondary analysis, was one of the primary drivers behind the development of data mining technology (Hand *et al*, 2000) as well as the modern rise of interest in "big data". But the analysis of administrative data presents new statistical challenges. This can be seen by a cursory examination of the examples in most basic statistics texts, which will almost all involve "random samples": administrative data are, by definition, typically not random samples. The aim of this paper is to explore these statistical challenges and to stimulate discussion. The hope is that it will help to focus attention on what is needed for valid and accurate analysis of administrative data. The need is illustrated by the comment made by Wallgren and Wallgren (2014, p3) on the closely related topic of analysing data from statistical registers: "Although register-based statistics are a common form of statistics used for official statistics and business reports, no well-established theory in the field exists. There are no recognised terms or principles, which makes the development of register-based statistics and register-statistical methodology all the more difficult. As a consequence, *ad hoc methods are used instead of methods based on a generally accepted theory*." It is hoped that this paper might serve as a framework to stimulate discussion about what "generally accepted theory" might be taught for the analysis of administrative data.

There are many definitions of statistics. This is because the discipline has a number of aspects, including the study of methods for collecting, presenting, interpreting, and analysing data, but also because it involves expertise in coping with uncertainty and chance. My own definition (Hand, 2008)

tries to capture this diversity: *statistics is the technology of extracting meaning from data and of handling uncertainty*.

There are fewer definitions of administrative data. The OECD (OECD, 2016) defines administrative data as having the following features:

- the agent that supplies the data to the statistical agency and the unit to which the data relate are *usually* different: in contrast to most statistical surveys;
- the data were originally collected for a definite non-statistical purpose that might affect the treatment of the source unit;
- complete coverage of the target population is the aim;
- control of the methods by which the administrative data are collected and processed rests with the administrative agency.

The definition continues by saying that "In most cases it is normal to accept (and expect) that the administrative agency will be a government unit that is responsible for implementing an administrative regulation." That leads to a rather narrower definition than is taken in this paper. For example, it excludes corporate use of administrative data, describing the workforce, products, processes, and so on, as well as narrowly restricting the uses to which such data are put. Instead, while accepting that the features described above do characterise administrative data, I shall follow Norbotten (2010), and simply distinguish between statistical data and administrative data. Statistical data are collected primarily for statistical purposes – for example, to summarise in order to shed light on the system generating the data, or to make predictions. In contrast, administrative data are initially collected for some administrative purpose – to run an organisation, such as a company, government, charity, school, hospital, and so on. Running the organisation might require ongoing operational analysis of the data but, once collected and stored, the data can later be analysed to shed light on what has happened, to help predict what might happen in the future, and to evaluate systems and their performance. That is, the data can later be subjected to statistical analysis. Often statistical data consist of mere samples from the universe of possible values which could have been obtained, and these will have been collected by surveys or experiments for example. In contrast, administrative data will ideally consist of data on all of the cases, records, or transactions in some population. This leads to something of a conceptual distinction: sample data are used to obtain *estimates* of a population *parameter*. In contrast, administrative data are *summarised* to obtain a descriptive *feature* of the population.

Transaction data are an important kind of administrative data concerned with *events*, typically with sequences of events. Usually the prime operational purpose of collecting the data is to inform the transaction (e.g. to decide how much to charge a supermarket customer or to decide how much tax someone should pay), but once collected the data can be retained in a database and analysed to improve understanding of the organisation's operations.

I used the word "operational" above. Occasionally one sees the terms "operational data", "management data", or "management information" used to describe data collected and analysed to guide the operation of a system. It is clear from the above that the data, once collected and placed into a database, are no different from administrative data. What differs is the way in which the data are being used – from immediate decisions to more considered analysis with longer term implications. Operational data becomes administrative data when they are stored and used for some purpose beyond the day-to-day operations of the organisation. In a sense, then, administrative data are *data exhaust*: that which is left over after the organisational machinery has used the data to drive itself forward.

Incidentally, in this paper, I will use the term "survey" to refer to data collection by sample survey, so that it is contrasted with administrative data, collected as a side-effect of an operational activity. This is different from, for example, Statistics Canada (2009), page 11, which uses the term survey "generically to cover any activity that collects or acquires statistical data", including the collection of data from administrative records.

At first glance, though we will see that appearances can be deceptive, administrative data appear to have a number of advantages compared to statistical data:

(i) Since the data have already been collected, no additional cost appears to be incurred in collecting them.

(ii) In a sense, one might reasonably expect that "all" the data are available. After all, a company will certainly process and can retain details of all its transactions.

(iii) The data might well be of high quality, since the effectiveness of the operation of the organisation depends on this.

(iv) The stored data will certainly be timely, and might be regarded as as up to date as it is possible to get, since they describe the organisation as it is, or at least as it was when the last change was made. This advantage is strikingly illustrated in the use of administrative data to derive estimates of population attributes at times intermediate between decadal censuses, and in essentially real-time estimates of price inflation.

(v) In a real sense administrative data often tell us what people *are* and what they *do*, not what they *say* they are and what they *claim* to do. One might thus argue that such data get us closer to social reality than survey data.

(vi) Administrative data may provide tighter definitions than alternative sources of data. Wallgren and Wallgren (2014, p33) give examples of data about income and children in families. Where the time restrictions on eliciting responses to a survey might mean one must simply ask "what is your yearly income before tax?", administrative data might, depending on the source of the data, specify whether this means "disposable income, taxable income, earned income or income including unearned income...".

Unfortunately, while all of those advantages of administrative data might apply in an ideal world, in practice things are typically not so straightforward. Regarding (i), effort will normally be required to extract the data, clean them, and possibly link them to other data sets. Moreover, while data may be free for the organisation which collected them, other organisations which wish to use these data may have to pay - and the cost must be balanced against that of data from alternative sources, such as surveys or administrative data collected by other organisations. Regarding (ii), data will usually enter a database via a complex social process – the sample of records in a database may not be representative of the population to which one wishes to make an inference. An operational database might not have a form which is convenient for statistical analysis exercises. In particular, different parts of an organisation might use different database systems - indeed, there is a great deal of current activity as organisations seek to put all of their data into a single data repository (a data warehouse, for example). The notion of "data = all" is discussed in Section 3. Regarding (iii), while sampling variation issues may not apply, administrative data will have other sources of uncertainty, and unfortunately these may be various and diverse, and not susceptible to resolution machinery as mathematically elegant or unified as sampling theory. More generally, while in principle one might expect administrative data to be of high quality, in practice all data sets, perhaps especially those

involving human beings, are susceptible to quality issues. A particular issue with administrative data sets arises from the very fact that they were not deliberately collected to answer the later statistical question being addressed. This means that the data may not be ideally suited to answer the question; there is often a compromise between cost and relevance. For example, a costly survey could be designed to answer the specific question whereas "free" administrative data might be only roughly suitable (but see Principle 8, clause 8.1, of the European Statistics Code of Practice (Eurostat, 2011), which says "When European Statistics are based on administrative data, the definitions and concepts used for administrative purposes are a good approximation to those required for statistical purposes."). Moreover, the definitions involved in administrative data are subject to changes for operational purposes which might impact the research questions that can reasonably be asked. It follows that time series of administrative data might exhibit discontinuities: an administrative database containing details of unemployment benefits might appear to be ideal to address questions of unemployment rates, but if the definition used in assessing benefits changed over time, then it might limit what can be done. Regarding (iv), while administrative data may be instantly available to the organisation collecting it, it may not be so to other organisations which wish to use it. Regarding (v), there are important kinds of administrative data which are not automatically accrued through some transaction, but are specifically sought and reported - for example, income tax data. And, finally, (vi) is not universally true: credit card transaction data contain considerable detail of the nature of the item purchased, but not necessarily to a level adequate for all potential analyses.

An example of the relative merits of administrative and statistical data is given by crime statistics in England and Wales. There are two main sources of such data, the Crime Survey for England and Wales (CSE&W) and Police Recorded Crime (PRC) (ONS, 2016). These can sometimes show trends going in opposite directions. The reason is that definitions – of crimes, of victims, of what data are collected – differ. With the administrative data (the PRC) we have to make use of the data we have, whereas with survey data (the CSE&W) we can decide what data are best collected to answer the questions we want to ask. Indeed, there has been extensive research on how best to formulate survey questions to elicit the information sought (see, e.g., Presser *et al*, 2004; Fowler, 1995).

There are also other issues which arise with administrative data which are slightly different from those arising with survey data. An obvious one relates to privacy and confidentiality - discussed in more detail below. Since survey data are collected purely for the purpose of statistical analysis, data released for analysis would not normally retain identifying information: apart from its use in ensuring consistency and representativeness, the identifying information is not relevant to the use of the data. On the other hand such information is central to the initial (operational) purpose for which administrative data are collected. The (hoped for) comprehensiveness of administrative data increases the risk of reidentification - and perhaps the public concern.

We could go on, but it is clear that while administrative data do have merits, the statistician should approach such data sets with the same critical eye they approach any other data set.

It is worth noting that it is sometimes useful to distinguish between two kinds of administrative data. The first kind is that which is *necessarily* collected during the course of some operation. Credit card transaction data, for example, necessarily involves the recording of the amount spent, the currency, and the business where the transaction occurs, since these items of information are needed to run the credit card operation. The second kind is additional information which is not needed for an operation, but which is helpful for other reasons, and which is collected during the administrative process. The age and gender of a customer might fall into this category: a product might be bought by anyone, but it could well be useful to analyse the customer details later to inform new marketing strategies. In some sense this second kind of data lie between administrative and statistical: they are collected for statistical rather than operational purposes, but are collected during and as part of the

administrative process. There is an important lesson to be taken from this: benefits can be gained from involving the statisticians and data analysts in the data collection stage. This is not a knew lesson: we recall Ronald Fisher's comment that "to call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of."

This last point leads us into the modern world of so-called "big data". The term has no universally accepted definition, but we might define it as the result of some automatic data collection system. Indeed, I have argued elsewhere that the data revolution is not so much a consequence of the size of modern data sets and the ability to store them (big data) but rather of the fact that data are nowadays largely collected automatically without requiring explicit human effort. Examples of automatically collected data are everywhere, and include personal health data collected by wrist monitors, automated monitoring of tickets as people travel through a rail network, telemetry of engine functioning, recording of metadata of phone calls, and so on. Data arising from the so-called *Internet of Things* would clearly be of this type.

Given that so many official and economic statistics are based on administrative data, or on a combination of administrative and survey data, one might have expected there to exist a substantial literature in the leading methodological statistical journals describing the statistical challenges and how to overcome them. This appears not to be the case, with such journals carrying relatively few papers on the statistical challenges of administrative data (being mostly focused on the consequences of sampling theory). For example, a search of the *Journal of the American Statistical Association* for occurrences of the phrase "administrative data" yields 44 results. Obviously the *Journal of Official Statistics* is an exception, although even there most of the papers including the phrase "administrative data" in the title are concerned with particular applications. More generally, papers on the topic seem to be widely scattered, and often appear in the proceedings of conferences and workshops, or perhaps as reports of official exercises (e.g. from official statistics offices). Given this wide scattering, it is certain that important contributions have been omitted from this paper, and I welcome attention being drawn to them in the discussion.

One of the best discussions is the excellent and comprehensive introduction to register-based statistics by Wallgren and Wallgren (2014). In one sense, this has a much wider scope than this paper, including discussion of register structures and the creation of registers, but in another sense it is narrower, being focussed on official statistics and not including, for example, commercial or engineering applications of administrative data.

The series of conferences on New Techniques and Technologies for Statistics (e.g. NTTS, 2013; NTTS 2015), organised by the European Commission, often have items touching on administrative data challenges: the phrase "administrative data" appeared in the 2013 conference proceedings 220 times. To give the flavour of the breadth of topics covered, these proceedings included papers on state space models (Horn and Czaplewski, 2013), structural equation models (Scholtus and Bakker, 2013), business statistics, and other topics. Romanov and Gubman (2013) explored regression to the mean in survey responses to questions on income by using administrative (tax) data, and Lewis and Woods (2013) described some of the issues which must be tackled when using administrative data in the form of Value Added Tax and Company Accounts data, as the basis for business statistics. As well as problems of matching and cleaning administrative and survey data, they also discussed differences of timeliness and periodicity. Kloek and Vâju (2013) discussed integration of administrative data with other kinds of data. They characterised five different kinds of use: direct use at micro level, use as auxiliary information at micro level, use as auxiliary information at aggregate level, use as a source for the population frame, and use as circumstantial evidence. They also explored the distinction between administrative data for business and that for households. This,

of course, reflects the general point that data describing different kinds of entities might have different characteristics (e.g. more pronounced skewness for some variables for business data compared to household data). Ćetković *et al* (2013) provide an elaborate example, the Austrian register-based census, involving seven base registers and several comparison registers which are provided with data from 35 data holders. They characterise data quality in terms of several "hyperdimensions" described by Berka *et al* (2012) (see below). Antoni (2013) links survey and administrative employment data.

Other sources which have relevant materials include:

- the UNECE Data Collection Workshops (see, for example, UNECE (2012), on new frontiers in data collection);
- the ESS Vision 2020, which includes discussion of administrative data sources and challenges (ESS, 2020). Their Administrative Data Sources project is exploring how administrative data may be used to increase data availability and reduce costs (ESS Admin, 2015);
- ESSNet has current and previous projects on administrative data topics (ESSNet, 2017). See, for example, the ESSNet Admin Data Workshop (2013);
- the US *Review of Administrative Data Sources* (Ruggles, 2015);
- the Statistics New Zealand *Guide to Reporting on Administrative Data Quality* (Statistics New Zealand, 2016);
- the use of administrative data at Statistics Canada (Statistics Canada, 2015)
- the administrative data quality assurance documents produced by the UK Statistics Authority (UKSA 2014, 2015).
- the checklist of quality of statistical outputs in van Nederpelt (2009).
- *Pros and Cons for Using Administrative Records in Statistical Bureaus*, from the Israel Central Bureau of Statistics (Israel, 2007)
- the OECD compilation *Short-Term Economic Statistics (STES) Administrative Data: Two Frameworks of Papers* (OECD, 2016) is a particularly valuable source of examples of the use and challenges of administrative data, albeit focused mainly on economic uses.

The structure of this paper is as follows. Section 2 describes a fundamental problem that is relevant to all data analysis, no matter what the source of the data, namely data quality. But the challenges - and even the recognition that there are challenges - presented by administrative data differ from those presented by other sources. We look at some of these challenges and how they differ from those of other types of data.

Section 3 addresses the notion that one might have "all" of the data. This is typically regarded as one of the particular merits of administrative data but, as we show, it is all too often an unjustified assumption.

Section 4 explores the fact that administrative data are collected for operational purposes, and not with specific research questions in mind. The consequence is that the data may be far from ideal for addressing those questions.

Sections 5, 6, and 7 look at deeper issues where the nature of administrative data impacts other aspects of analysis, including efforts to identify causation, merging data from multiple sources, and the thorny issues of confidentiality, privacy, and anonymisation of records.

Section 8 draws some conclusions.

2. Data quality

The value of administrative data for producing official statistics has attracted increasing attention recently. In large part this is in the hope that they can replace more conventional survey data, motivated on the one hand by a worldwide decrease in survey response rates, and on the other by a perceived lower cost in using administrative data, since it has already been collected. However, as the UK Statistics Authority put it, "we have been surprised by the general assumption made by many statistical producers that administrative data can be relied upon with little challenge, and, unlike survey-based data, are not subject to any uncertainties" (UKSA, 2014). Because of this, the UK Statistics Authority has produced a report on quality issues in administrative data, summarising the lessons learnt from a review of users of administrative data for statistical purposes and describing a toolkit to monitor data quality in this context (UKSA, 2014, 2015).

Other explorations of the quality aspect of administrative data include the model Daas *et al* (2008) have developed for Statistics Netherlands. Noting that a key issue with administrative data is that the source of the data is typically some other body, Daas *et al* point out that the collection and maintenance are not within the control of the analyst: when data are collected by bodies other than those undertaking the analysis, issues of data provenance and curation are critical. In their review of earlier work on administrative data quality, Daas *et al* observe that different authors have identified "a remarkable difference between the number and types of quality groups or dimensions identified for the statistical quality aspects of administrative data". They attribute this partly to the complexity of the problem and partly to the fact that different researchers had different perspectives on the topic. Their paper is then an attempt to integrate the various views into a single framework. Their conclusions include the observation that administrative data quality is a multidimensional issue, with a hierarchy of dimensions (Karr *et al*, 2006).

Other work (e.g. Eurostat, 2003) has explored the potential uses of administrative data. This is an important point when attempting to evaluate data quality, as data may be "good" for one purpose but "bad" for another: quality is not a property of the data set itself, but of the interaction between the data set and the use to which it is put. And yet other, more general, work on data quality, especially in the context of official statistics, inevitably touches on administrative sources (e.g. van Nederpelt, 2009; Memobust Handbook, 2014; Statistics Netherlands, 2014; etc.). Once again, we stress that work on the quality of administrative data has appeared in diverse publications, from a wide range of sources.

The common misconception that quality issues are less important for administrative data than they are for survey data seems to be based chiefly on the belief that data initially acquired for operational purposes must necessarily be both complete and error free, while survey data will be based on a mere sample from the population being studied, so that the results will vary between possible samples. The fact is, however, that administrative data may be neither complete nor error free. As far as "complete" is concerned, incompleteness can manifest either in the form of partial records – records in which some of the fields are missing – or in the form of entire records missing, so that the data set does not in fact cover the entire population. And as far as "error-free" is concerned, errors can arise in an unlimited number of ways. To paraphrase Leo Tolstoy: "A perfect data set is perfect in only one way; each imperfect data set is imperfect in its own way." This means that one can never be sure that all the errors have been detected. The problem is analogous to that of testing random number generators: one can look for particular kinds of departures from randomness, but there will always be kinds one has not thought of. Unfortunately, one of the lessons we have learnt from data mining practice over the past twenty years is that most of the unusual structures in large data sets

arise from data errors, rather than anything of intrinsic interest. One should be suspicious of any data set (large or small) which appears perfect. A standard check I carry out is to ask those providing the data what they have done about missing values. Often this has resulted in surprising responses which the researchers would not have thought to mention had the question not been explicitly asked. For example, it is not uncommon for researchers to have removed any incomplete records from the data set, introducing unknown selection bias. Caruana *et al* (2015) describe the development of a machine learning diagnostic system based on hospital administrative data which classified high risk asthma patients as low risk because such cases had been excluded from the training data.

While the technology of data editing and imputation has been substantially developed, with entire books being written about it (e.g. de Waal *et al*, 2011), it is not the case that detected errors can necessarily usefully be corrected. This means that commercial tools for detecting and correcting data errors are unlikely to be one hundred percent effective, whatever they may claim.

Statisticians know very well that it is common for the major part of their time on a project to be spent cleaning data prior to actual statistical analysis. This is all very well when the data set is of moderate size, but it becomes more of a problem when the data set is massive – as is increasingly the case in the "big data" world, and is particularly the case with administrative data and data which are captured automatically. Especially in such contexts, the computer is a necessary intermediary between the analyst and the data, with consequent risks of missing important shortcomings of the data – and indeed, even creating extra errors during an automatic data cleaning process. For example, rule-based correction mechanisms can distort perfectly good, though unusual, data values, and an unfortunately all-too-common strategy for coping with missing values is to substitute the mean of the observed values (so leading to an underestimate of variance).

Familiarity with the fact that data are often not of the highest quality has led to the development of relevant statistical methods and tools, such as detection methods based on integrity checks and on statistical properties (for example, comparing distributions with expected distributions in electoral data, or using the Benford distribution for leading digits); see, for example, Hellerstein (2008) and de Jonge and van der Loo (2013). However, this emphasis has often not been matched within the realm of machine learning, which places more emphasis on the final modelling stage of data analysis. This can be unfortunate: feed data into an algorithm and a number will emerge, whether or not it makes sense. However, even within the statistical community, most teaching implicitly assumes perfect data. This is entirely reasonable: if one is aiming to teach the basic concepts of regression, one does not want to spend time pointing out the consequences of missing data, digit heaping, or digit transposition. Nonetheless, students do need to understand the reality of data analysis. This leads to:

Challenge 1: statistics teaching should cover data quality issues.

Even if data may depart from perfect quality in an unlimited number of ways, it is important to characterise as many ways as possible, and Kim *et al* (2003) have produced a general "taxonomy of dirty data". They characterise data as dirty "if the user or application ends up with a wrong result or is not able to derive a result due to certain inherent problems with the data," and identify various possible causes of the problem, including data entry errors, data update errors, data transmission errors, and also bugs in a data processing system. Particular applications are likely to have their own characteristic types of errors, and it seems likely that an 80/20 rule will often apply, with a large proportion of errors being of just a few types, so that relatively little effort will lead to substantial initial improvement in overall quality. An illustration of this is given by Lewis and Woods (2013), who identify the main causes of error in Value Added Tax data to be just four types: scanning errors, unit

errors, incorrect quarterly data, and errors in individual responses. De Veaux and Hand (2005) give examples of data errors and their consequences, and National Statistical Institutes (NSIs) often define several dimensions of quality, including accuracy, relevance, timeliness, existence, coherence, completeness, accessibility, and security (see, for example, Eurostat (2000) and the archives of other NSIs; Meader and Tily, 2008; Biemer *et al*, 2014), though these will affect administrative data in varying degrees.

The "relevance" aspect in the NSI list is more subtle than simply finding a mistake in the data. Even perfectly accurate data may be useless for answering a particular research question if the data have not been collected with the research question in mind - as is typically the case with administrative data. Clearly one can try to ease that difficulty if one knows beforehand what questions are likely to occur, but even then difficulties can arise. For example, in a project aimed at constructing a scorecard to predict likely default on bank loans, one of the (relatively highly predictive) variables was "is the applicant a home owner or renter?" This was administrative data of the second kind mentioned above - the question was not relevant to everyday operations. But as a consequence of this the people tasked with recording the data failed to see its importance, with the result that they initially recorded it for only a small percentage of customers.

If administrative data are subject to restrictions arising from operational imperatives, they are also subject to possible constraints from the opposite direction: administrative data are often communicated, compared, and aggregated across bodies collecting the data. For example, national statistics for US states and countries within the EU will be aggregated to produce Federal statistics and EU statistics respectively. The need to do this imposes constraints on what must be collected and on its format, with particular standards requiring particular structures, formats, and protocols, as well as content.

As mentioned above, administrative data are also susceptible to changes of definition, which can adversely affect things like time series, rendering them non-comparable over time. Since much administrative data, especially that concerned with government and public policy, are subject to regulation and legislation, changes in laws can have an unfortunate impact, at least from the perspective of the statistician hoping to use the data to make inferences. Changes in what data can be stored, or the characteristics which are allowed to be used in statistical models, can mean that earlier models become unusable.

Data can be incorrectly entered, even for operational purposes. We have all heard of "fat finger" errors leading to mistaken financial transactions. Other classic examples include things like weights of 1lb being miscoded as 11lb, data being entered in incorrect columns, abbreviations leading to confusion (e.g. MS for Microsoft or Morgan Stanley), incorrect time stamps due to clocks being mis-set, mistakes in the use of measurement units, simple misspellings, and instrument failures not being detected (leading to, for example, an unnoticed stream of zero values). The list of examples is endless. Kruskal (1981) observed that "A reasonably perceptive person, with some common sense and a head for figures, can sit down with almost any structured and substantial data set or statistical compilation and find strange-looking numbers in less than an hour."

However, even data which are entered correctly and unambiguously for operational purposes can lead to errors when subjected to statistical analysis. Alternative, equally legitimate spellings or identifiers (e.g. David Hand, David J. Hand, D.J. Hand) may not be recognised as equivalent in a subsequent analysis unless they have been explicitly characterised as so. Conversely, identical entries might refer to different objects (e.g. father and son with same name). Missing values for age coded as 999 can be analysed as legitimate ages, with obvious adverse consequences. While clearly this should be flagged in the metadata, we note, again, that large data sets necessarily involve an

opacity that does not affect small data sets, in that the computer is a necessary intermediary between the data and the analyst. Mistakes and ambiguities can slip through.

The argument has been made that data errors will often affect only a very small part of the data, and so will, for example, have no significant impact on large scale conclusions. While this may be true, large scale conclusions are typically not the only ones which will be drawn from administrative data. One of the particular strengths of such data is that they are also used for small scale investigations – to explore subgroups or for small area statistics, for example. In such cases, errors in only a few records can have important consequences.

Quality issues may also arise when data sets, of adequate quality in themselves, are merged. Take, for example, time series which are out of phase, or have different frequencies of publication, or publish on different dates, or, even worse, are irregular.

These considerations lead to several challenges

Challenge 2: develop detectors for particular quality issues.

Challenge 3: construct quality metrics and quality scorecards for data sets.

Challenge 4: audit data sources for quality.

Challenge 5: be aware of time series discontinuities arising from changing definitions.

Challenge 6: evaluate the impact of data quality on statistical conclusions.

3. Data = all ?

The phrase "data = all" is sometimes encountered in the context of administrative data. This is intended to convey the notion that the data are not a merely sample from the population of objects but are its entirety: all credit card transactions, all supermarket purchases, all tax records, and so on. The implication is that having data describing the entire population means that one need not worry about sampling errors or errors arising from non-representativeness. This, however, is misleading. Administrative data tell us what happened with a particular group of people, but this group of people may or may not be the group about which we wish to make statements or from which we wish to generalise. Very often, for example, the selection process which results in them being chosen will include an aspect of self-selection.

A few examples will illustrate some of the difficulties.

Retail banks and other financial institutions construct *scorecards* to predict likely customer behaviour with financial products. For example, such models are used to predict who is likely to default on a loan, and hence who to give loans to. Administrative data are then collected on the customers awarded loans as they make their repayments. In particular, outcome data – whether they defaulted or not – are collected. Such data, the outcomes along with potential predictor variables (from application forms or behaviour on other financial products), can then be used to construct models to make loan decisions on future applicants. Unfortunately, this data set will not be representative of the population of applicants. It will only include people previously thought to be good risks. This means that models based on it could give seriously distorted predictions for people drawn from the entire population of applicants (Hand, 1993, 2001). "All" of the data are there, but they are not all of the data one needs.

An example that is currently attracting a huge amount of attention is publication bias and associated phenomena in scientific literature. We can obtain data on all papers that are published, but they

arrive at publication through a complex sociological selection process: papers reporting positive results are more likely to be submitted, editors are more likely to publish them, anomalous results may be regarded as errors so that the work does not get written up, and so on. So what we see is a distorted view of the work that is done and the results that are obtained. So much so, in fact, that John Ioannidis was able publish a paper with the title *Why most published research findings are false* (Ioannidis, 2005), stimulating a great deal of interest and subsequent work. The notion that the published scientific literature represents "all" the relevant material is simply false.

The Crimemaps system provides another example. Originally developed in Chicago, based on police recorded crimes, this gives (approximate) locations of crimes, displayed on maps so that people can see which areas are dangerous. However, research from the Direct Line insurance company in the UK suggests large numbers of people are not reporting crimes because of the potentially adverse impact it will have on house prices and hence their ability to sell or rent their house (Direct Line, 2011). The data are purely administrative – from the police databases – but can become progressively more distorted. This may well mean not only that the data are of limited value for determining which areas are risky, but also that they become less and less valuable for their original purposes. This is a straightforward illustration of Campbell's Law: "The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor." This law applies just as much to administrative data as survey or other data.

Even something as apparently automatic and complete as data from financial markets exhibit numerous errors and omissions. These can arise from the use of different timestamps on transactions, ambiguity over time resolution of transactions, extent and method of aggregation of data, whether or not certain types of transactions are present (e.g. so-called "dark pool trades"), changes to symbols identifying corporations perhaps leading to mismatches or failed matches when data are linked, confusion arising from stock splits or mergers, and so on. The financial data sources described in the Caltech Quantitative Finance Group guide to market data at http://quant.caltech.edu/historical-stock-data.html illustrate the problems.

The bulk of traditional survey work is based on known probabilities of including each person in the study, with sampling theory permitting solid inferential conclusions, but sometimes non-probability samples are chosen; for example convenience sampling, matched sampling, network sampling, and situations in which people are allowed to opt in or opt out. These kinds of non-probability samples have stimulated research on drawing valid population conclusions (see, e.g., Baker *et al*, 2013; Bethlehem, 2010) although it may not be straightforward to apply the methods to administrative data. In general, the data distortion will have different consequences in different contexts and with different problems. For example, for one particular credit scoring dataset, Crook and Banasik (2004) drew the conclusion that "even where a very large proportion of applicants are rejected, the scope for improving on a model parameterised only on those accepted appears modest. Where the rejection rate is not so large, that scope appears to be very small indeed." So, for their example at least, the problem seems not to be too bad, but it would be unwise to assume this is generally the case.

On top of all this, there are more complicated questions of what is meant by "complete". For example, many systems are dynamic and constantly changing. A database of all the people in a company today will provide at best only a snapshot – it is likely that some employees will have moved on or been recruited by next year. Indeed, it is *certain* that the individual employees will have changed by next year - if only because they will have aged, let alone possible name changes due to marriage, address changes due to moving house, and so on. This *population drift* poses interesting

statistical challenges, and it again points out the weakness of the assertion that administrative data represent “all” the data one needs.

Statistical methods have been developed for correcting for sample distortion (e.g. Heckman, 1976; Copas and Li, 1997), but they do depend on making assumptions about the form of the distortion. Statisticians can do amazing thing, but they cannot perform miracles and if the data have been chosen in an arbitrary and unspecified way there is little that can be done. Were this not the case we could always draw accurate conclusions from the most limited of data. And this is precisely why survey sampling and experimental design have grown into such elaborate disciplines: they specify and constrain how the data must be collected so that valid conclusions can be drawn from a statistical analysis. Administrative data, on the other hand, without this underlying statistical imperative, may well not be so useful for drawing statistical conclusions. They may be selected in precisely those “arbitrary and unspecified ways”.

In short, the fact that administrative data arise through an administrative process does not mean that they represent the entire population of interest. Some major successes in the world of “big data” have been achieved by simply analysing the data as they present - but some major failures - such as the initial Google Flu Trends projections (Hodson, 2014) - have also arisen from taking the data at face value.

These points lead to:

Challenge 7: explore potential sources of non-representativeness in the data.

Challenge 8: develop and adopt tools for adjusting conclusions in the light of the data selection processes.

4. Answering the right question

As the previous sections have illustrated, there can be difficulties in using administrative data to answer specific research questions. This might be because the data were not collected with those questions in mind, because of quality issues irrelevant to operations but highly relevant to subsequent statistical analysis, because of changes in definitions of the recorded data items, or for other reasons. This brings us back to a point made earlier: it can be useful, if it is possible, to have statisticians involved in the data collection process. They might be able to think ahead and expand the range of data collected so that it will be more able to answer future questions.

Statistical analysis methods are often divided into *descriptive* and *inferential*. Descriptive methods are used to summarise a body of data so that the important messages within it can be readily grasped. We might summarise a distribution of values by their mean and standard deviation, or the results of a census using a series of counts organised as cross-tabulations. It goes without saying that the summary statistics that are appropriate will depend on the subject matter and on the questions to which answers are sought. Administrative data are often used for purely descriptive purposes – perhaps especially so in official statistics contexts, where we might want to establish the characteristics of some population.

In contrast, inferential methods are used to make a statement about unobserved values or underlying mechanisms. We might be trying to infer the disease of a new patient, based on analysis of patients with similar symptoms diagnosed in the past. We might be trying to forecast whether inflation will go up or down next month. We might be trying to elucidate an underlying mechanism, so that we can understand how the data were generated, and perhaps influence things in the future.

Much of the statistical theory of inference is based on the notion of random sampling from a (possibly infinite) population of values. Because the sampling is random, solid mathematics (such as the law of large numbers and the central limit theorem) means that sound statements can be made about the characteristics of the population from summary statistics obtained from the sample. Moreover, error bounds can be put on the conclusions. We can say things such as "on average, 99 out of 100 of our intervals will cover the true population mean," so we can be confident of our results. (Always subject to data quality issues, of course.)

But administrative data are not collected by such a random sampling process. We can certainly calculate descriptive statistics, summarising the data before us and, if we are prepared to assume that the data are perfect, with no missing or distorted values (a brave assumption, as the above has illustrated) then this will accurately summarise the population which led to our data. We can make a statement such as "*this* is the true population mean".

But, if our aim is not really to summarise the data at hand but to make an inference to another (often future) population, then additional, unknown and quite possibly unquantified, sources of uncertainty may be relevant - possible incompleteness of the data set, discussed above, is but one example. This has consequences for inferential statements.

Of course, these unknown and possibly unquantified sources of uncertainty beyond sampling variation also affect the sampling approach. Indeed, it is likely that they have been underappreciated in many contexts, where the elegance of the sampling theory mathematics has distracted attention from the fact that there are other sources of uncertainty. This could be one of the drivers behind the phenomena to which John Ioannidis has drawn attention, mentioned above. Hand (2006) gives examples and implications of this oversight in a different context.

Administrative data are often highly complex. For example, a single credit card transaction leads to some 70-80 items of data being recorded, while web search and social media data have an elaborate graph structure. And this leads to the observation that data capture technology changes rapidly. In the context of the first of these examples, a recent change is a shift towards mobile banking, leading to new and additional transaction characteristics being available. In the context of the second, changes in social media platforms mean that there is a very real risk that any particular kind of model based on data recorded from web transactions may be impossible to build just a couple of years in the future, as social interaction media change and evolve. And similar problems apply in other areas - in medicine, for example, with different kinds of bioinformatic measurement methods, in the financial sector with short time series because of changing regulations, and so on. At a substantive level, this has clear implications for studying how society is developing. At a statistical analysis level, it drives home the point made above, that administrative data are collected for operational reasons, and may have serious weaknesses for subsequent analysis.

Economic and social measures such as GDP, CPI, and national wellbeing are what are called *pragmatic* measures (see, e.g. Hand, 2004): the definition of the concept and the way it is measured are two sides of the same coin. Change the measurement procedure and you change the thing being measured, with different measures being suitable for different purposes. This is why, in the UK, we have CPI, CPIH, RPI, RPIJ, and so on. It is not a question of any of these being more "right" than the others, but simply that they measure slightly different things. This means that they have different properties and are suited to answering different questions. Increasingly, interest is turning to the possibility of using administrative data for measuring productivity and price inflation. Instead of conducting surveys of businesses to obtain data, the data can be automatically transmitted from the transaction to the database. Scanner data, such as retail purchase data obtained directly from the point of sale machine, provides an example, yielding data ideal for use in price index calculation.

Moreover, such data also gives information on volume of different goods purchased, so that weights can be chosen. But issues of selection bias still apply: not all purchases are made through such routes, and one cannot assume that those purchases which are made in this way represent a random or representative sample of all purchases.

A variant of this uses web-scraped price collection, being explored by a number of national statistical institutes. For example, the Billion Prices Project (Cavallo and Rigobon, 2016) seeks to collect massive amounts of price data from websites. Apart from its vast coverage ("big data") this means much more timely estimates can be obtained much more cheaply than traditional methods. Note, however, that Cavallo and Rigobon do not describe this as an alternative to traditional methods, but as a complement.

This approach has had some notable successes - for example, Cavallo (2013), using such online data collection, estimated Argentina's annual inflation rate between 2007 and 2011 to be over 20%, a striking contrast to the 8% claimed by the Argentinian government. Moreover the online estimates were reported daily.

The apparent simplicity of this approach risks concealing various complications. It is still necessary to decide which websites to collect data from - and this might be biased towards larger retailers. In fact, Cavallo and Rigobon say "we ... focus almost exclusively on large multichannel retailers and tend to ignore online-only retailers (such as Amazon.com)". The basket of goods to be included still has to be chosen. Cavallo and Rigobon note that online prices cover a much smaller set of retailers and product categories than is covered by the traditional approach. Also, online prices are one thing, but they say nothing about the quantity sold.

A key question, perhaps obvious in view of the earlier sections, is how representative the online prices are of prices in general. What about prices of goods or services not bought online? Moreover, in certain contexts, such as airline tickets, dynamic pricing systems operate, which introduces not only changes over time, but also the impact of gaming strategies.

One of the problems with web based tools is the rate of change of that technology. Companies appear, grow to a massive size, and vanish at a dramatic pace. Bebo, for example, launched in January 2005, and sold to AOL just three years later for $850m. But then, in May 2013 it voluntarily filed for Chapter 11 bankruptcy protection. Worse still, the algorithms used by the companies can change arbitrarily. Google's search algorithm is constantly being redeveloped. As we have noted above, this means that administrative data may have a short shelf life, in the sense that comparative data and time series may not merely have discontinuities as definitions change, but may also experience changes in ill-defined, even ill-understood ways.

Since survey data will be collected with a view to answering specific questions, the variables will be relevant by definition. Variables from administrative data may be less relevant. This means that *derived* variables will be more important for the analysis of administrative data. These are variables created by combining other variables. For example, whereas a survey question might ask about disposable income directly, to obtain the corresponding value from administrative data might require adding earned income, interest from bank and other deposits as well as other sources of income, subtracting tax paid, and so on.

As a final example of definitional difficulties, the media were recently exercised by an apparent discrepancy between the number of long term migrants to the UK, estimated by the International Passenger Survey, and the number of National Insurance Number registrations (administrative data). Close examination (ONS, 2016b) revealed that the discrepancy was due to differences in definitions.

They commented: “it is not possible to provide an accounting type reconciliation that simply ‘adds’ and ‘subtracts’ different elements of the NINo registrations to match the LTIM definitions.”

All of this leads to:

> *Challenge 9: explore how suitable the administrative data are for answering the questions. Identify their limitations, and be wary of changes of definitions and data capture methods over time.*

Administrative data, typically being observational data, permit hypothesis generating exercises. Whereas survey data have the advantage that they will be tuned to answer the survey questions, and administrative data may not be well-suited to answer those questions, the converse also applies: administrative data, often being much richer than survey data, can be used to explore other questions and to generate hypotheses based on relationships observed in the data.

Here is one example illustrating both the complexity of human behaviour and the use of administrative data in detecting unsuspected patterns in that behaviour.

Hand and Blunt (2001) sought to model the distribution of sizes of credit card transactions at petrol stations in the UK, based on administrative data recorded at petrol stations. Superficially, the distribution was as expected – roughly normal, but with some right skewness since it could take only positive values. However, closer investigation revealed a number of anomalous spikes. The size of the data set meant that these spikes could not be attributed to random variation arising from the particular period being studied, but must have represented an underlying reality. (Note, the data were *all* of the transactions, but were being analysed as a sample to make inferences about underlying mechanisms, and hence what might be expected to happen at other times). Closer investigation led to the observation that there are two different types of behaviour patterns: some people simply fill the petrol tank at each purchase, while others seek to hit a convenient whole number of pounds cost, such as £20 or £30. Noting this, and digging deeper, led us to recognise further patterns of behaviour: there was more overshoot than undershoot; people preferred to hit whole numbers of pounds of *any* magnitude than numbers ending in a non-zero number of pennies (though especially those which were a multiple of £10); subject to that, they particularly favoured numbers ending in 50p, and then 25p, and so on. Things were further complicated by the fact that a significant proportion of spend in such situations is in the forecourt shop, where goods have particular prices, often with special values of their own (e.g. ending in 99p). And as if all that was not enough, there were further features in the data which arose as a consequence of marketing initiatives run by the forecourt operations. In the end, we constructed an elaborate mixture model which tried to take all these phenomena into account. This model was purely descriptive, though inference was needed to decide whether effects were large enough (in the context of the particular data set being supposedly drawn from a superpopulation of possible such data sets) to be included. However, the aim was not merely to describe what customers had done in the past, but to use the model to inform future pricing strategies.

A particular merit of administrative data, and especially of transaction data, is that it is recorded as time progresses. Unlike data recorded at a particular time, or discrete sequence of times as in repeated surveys, it is essentially continuous. This means that administrative data can be very useful for early detection of changes in populations. Indeed, often one of the operational reasons for collecting the data in the first place will be for monitoring processes. But an assertion that a time series (of GDP or unemployment, say) has changed will typically not be intended merely as a face value assertion that the raw numbers differ, but rather as an assertion that some underlying reality has changed. And this should not be based on a simple comparison of the numbers, but rather on a

comparison of the difference between the numbers with the inaccuracy of measurement. It should answer the question: is this difference larger than we would expect, given the intrinsic uncertainty in how the measured numbers represent the reality, or is it well within the scope of what we might expect with no change in the underlying reality? And the crucial point is that this intrinsic uncertainty should include all sources of uncertainty, not merely sampling variation (if that is indeed relevant).

This leads to:

> *Challenge 10: report changes and time series with appropriate measures of uncertainty, so that both the statistical and substantive significance of changes can be evaluated. The measures of uncertainty should include all sources of uncertainty which can be identified.*

5. Causality and intervention

As is well known, observational data present challenges in establishing causality. If we observe a difference in some outcome measure (e.g. income) between two groups, and we note that the groups differ in various properties (e.g. education) we cannot be sure that the observed differences in their properties explain or cause the outcome difference. To establish causality, we need to intervene to break all possible causal links except the one we wish to test (but see also Pearl *et al*, 2016). The most common way to do this is via a properly controlled experiment involving randomisation. Usually this is difficult with administrative data, not least because it requires modifying the standard operation of the organisation, although occasionally experimental designs are built into ongoing operations, enabling comparisons to be made using administrative data. In such situations the designs will typically be fairly simple, such as merely comparing two groups.

This notion of modifying an operation so that one can learn from it, as well as simply using it to carry out its normal function, can manifest in other ways. One mail order organisation we worked with enrolled a "gold sample" - a small set of people regarded as poor risks (who would normally be rejected), just so they could collect data across the entire population distribution, and hence enhance their models and improve future predictions. Scholtus *et al* (2015) have also explored the use of such a gold sample, in their case to yield estimates of intercept bias in a model.

We see from this that the needs of planned subsequent statistical analysis can sometimes influence what administrative data are to be collected. Occasionally, such intentions can lead to data being collected during the operations which are not required to run the organisation but which can be used subsequently - the second kind of administrative data mentioned in the opening section.

> *Challenge 11: be aware that administrative data are observational data, and exercise due caution about claiming causal links.*

6. Combining data from different sources

Combining data and evidence from different sources is increasingly important in statistics and elsewhere. This can be for statistical purposes, such as to yield an improved or more comprehensive estimate (e.g. Ashley *et al*, 2005; Cunningham and Jeffery, 2007), or simply because information is needed for a higher level organisation (e.g. combining statistics from several countries to give EU-wide statistics). But it can also be at the individual level, for example in detecting fraud or adverse drug reactions, or tracking terrorist activity.

Even if the data are, at least in principle, of the same type, such as combining economic statistics from different countries, they may have been collected using different methods or definitions, so that producing combinations or aggregates is not necessarily straightforward. Vâju *et al* (2015) describe "a huge number of possible sources of lack of comparability, given by combinations of (i) national legal and institutional environments, (ii) acceptable trade-off between quality dimensions at national level, (iii) appropriate trade-off between costs and benefits in terms of output data quality at national level, (iv) methodological choices to integrate the several data sources." At a lower level, problems might arise because a particular characteristic might be grouped in different ways in two data sets (e.g. age classified into ten year bands or into young versus old), or observations might be taken or recorded with different periodicities. Possible strategies for overcoming such problems include the latent variable perspective, with the observed data being regarded as a coarsened or grouped version, or state space models (Horn and Czaplewski, 2013).

The situation is further complicated because the data are often of different types - survey data, administrative data, web-scraped data, social network data, data collected from wrist health and activity monitors, and even non-numerical forms of data such as speech and image data. This is perhaps where the real opportunities, and statistical challenges, arise. Medicine, in particular, is making extensive use of such approaches, combining medical images, clinical trial reports, epidemiological data, and health registry data. Credit bureaux combine credit card transaction records from several operators to build a single database from which they can construct a generally applicable credit scorecard. An example from official statistics in the UK is the estimation of income within small geographical areas, based on linking data from the Family Resources Survey and administrative data from benefit claimant counts, council tax bandings, and tax credit claims. Vâju *et al* (2015) point out that even if the accuracy of the separate data sources can be measured, assessing the sensitivity of the accuracy of the final combined data set to the source specific errors and the integration methods can be very difficult.

As far as merging data from different sources goes, reasons include the following:

1) *Complement*. Different data sources, and different types of data, can each serve as a complement to each other by providing different types of information. This is perhaps particularly true for administrative and survey data. Some types of variables - attitudes and opinions, for example - do not normally naturally arrive in administrative data, but have to be collected by surveys (or panels, or some other purposive data collection strategy). Surveys can be designed so they shed light on tightly focused research questions, whereas with administrative data we may have to be satisfied with questions which are a little different from those we would ideally like to ask, perhaps because they are based on slightly different definitions of the concepts involved. On the other hand, administrative data sets are likely to be larger, with better population coverage (though possibly vulnerable to the other data quality issues mentioned above).

2) *Supplement*. While administrative data are often thought of as an alternative to survey data, they are at least as valuable when used in conjunction with survey data. Survey data can be used to pinpoint particular research questions, but cost necessarily limits coverage. However, relationships found from survey data can be extrapolated to yield estimates from overall populations and smaller groups using such tools as regression estimation applied to an administrative data population base. This can be useful to yield small area and regional estimates. Indeed, such statistical tools can be used to improve estimates from survey data. A further point is that surveys require sampling frames, and administrative data are central to their construction.

3) *Accuracy*. We have stressed issues of data quality above. Triangulation and imputation from multiple data sources and reconciliation between data sources are good ways to tackle these issues.

Berka *et al* (2012) give an example, exploring accuracy in the Austrian register-based census of 2011. They note the use of surveys to check register data, but point out that this is resource intensive. They evaluate the quality of data at the raw data level in terms of three "hyperdimensions", assessing documentation (e.g. plausibility, legal aspects), preprocessing (formal methods for testing for errors and inconsistencies), and comparison with an external source, respectively. The results are three measures, each scored in the interval 0 to 1. A weighted average is taken to yield an overall quality indicator for each register and attribute. The fundamental challenge here is that of combining quality indicators from different sources, and Berka *et al* explore the use of Dempster-Shafer theory to do this.

Another example is given by Romanov and Gubman (2013), who use administrative data to explore bias in answers to survey questions about income. Discrepancies pinpoint potential errors and issues to be resolved. Of course, there are complications. Errors can propagate and perhaps not all of them can be resolved. Worse, especially in the context of administrative data, this jigsaw solution is vulnerable to one of the pieces disappearing as the operational imperatives generating the administrative data change. Moreover, as we have repeatedly stressed, one must be alert to different data sources using different definitions.

A special case of merging data from different sources is matching data from different administrative databases. For example, we may have identified data on individuals, collected for different reasons and stored in two distinct databases, and we may want to combine them. But, of course, the problem is not restricted to data on individuals: Lewis and Woods (2013) describe a problem of incompatible business registers, with different identifiers in the two databases. Because of its importance the matching of corresponding records from different databases has been the focus of much research effort - see, for example, Christen (2012), D'Oriazio *et al* (2006), and Rässler (2002). It faces various data analytic challenges, including deciding when to match two records given that they do not have unique and identical identifiers, detection of duplicate records (again, because slightly different identifiers may refer to the same individual person or object), and merging of duplicate records into a single entity (or *deduplication*).

A traditional, and still widely used method, at least for small datasets, is manual matching. This has a number of obvious shortcomings, including a scalability cost (in various measures), subjectivity arising from human biases, variation between people, variation within any one person as they get tired or bored, and the difficulty of objectively improving performance. A modern variant of manual matching, for contexts where confidentiality is not important or where the data to be matched can be effectively encrypted, is crowdsourcing, enlisting the help of large numbers of people.

Computational methods can be divided into two classes: deterministic and probabilistic or statistical.

Deterministic methods simply see if two records agree on all of a specified set of identifiers. This is clearly very quick. It can be a single step procedure, or can proceed through sequential steps, beginning with stringent matching criteria and progressively relaxing them.

Probabilistic methods relax the requirement of an exact match, and instead calculate a dissimilarity measure for each field in the pair of records being compared. The choice of dissimilarity measure will depend on the context (e.g. approximate string matches for some text fields, matches that allow different date formats, matches that allow given name and surname to occur in the reverse order, etc). The separate field dissimilarity measures are then combined (e.g. added or used to maximise the likelihood of a match, given a probability model) to yield an overall dissimilarity measure for the record pair. In the simplest approaches, these dissimilarity measures can then be compared with a threshold to yield a match/non-match classification. More sophisticated approaches (e.g. the classic

work of Fellegi and Sunter, 1969) follow the "reject option" and define three types of decisions, match/non-match/possible-match. The third class is then subjected to a second stage of investigation, often a manual comparison. Winkler (2006) reviews linkage methods.

Clearly methods which are based on pair-by-pair comparisons run the risks of intransitivity, of several records from one database being matched to a single one in the other, and of computational intractability if all possible pairs are compared. The first two problems, at least, can be eased if a higher level view of the matching process is taken, in which constrained groups of records are compared. To take a simple example, suppose we wanted to match a collection of left shoes with a collection of right shoes, to find which shoes belonged in a pair. One strategy would be simply to calculate similarities between shoes, one from each collection, and choose the pairs which had the greatest similarity - but this would be susceptible to the first two of the problems just listed. An alternative approach would be (computation allowing) to look at all possible pairings of shoes, one from each collection, and choose the set of pairings which maximised the likelihood. Exactly this sort of approach has been used in chromosome matching.

These considerations lead to

Challenge 12: be aware of the risks associated with linked data sets and the potential impact on the accuracy and validity of any conclusions. Recognise that quality issues of individual databases may propagate and amplify in linked data. Develop better measures of overall combined data quality.

Challenge 13: continue to develop statistically principled and sound methods for record linkage and evidence assimilation, especially from non-structured data and data of different modes.

Challenge 14: Develop improved methods for data triangulation, combining different sources and types of data to yield improved estimates.

7. Confidentiality, privacy, and anonymisation

A common challenge with all data describing human beings is the need to preserve confidentiality and privacy, but this often seems to be a particularly sensitive issue with administrative data. This may be because, unlike with surveys, there may be no choice about being included (at least, if one wants access to the service or product) or perhaps because it is obvious that the identifier must be retained in the data (since it is needed for operational reasons - one cannot run a credit card operation without being able to match transactions to customers). There seems to be growing concern about the *data shadows* we all inevitably leave as we access administrative services, whether corporate or public.

Anonymisation and de-identification tools do exist – for example, based on aggregating data, perturbing data, or randomly generating data with statistical properties the same as the raw data – but they all have shortcomings. An overview of such methods is given in Duncan *et al* (2011) and see also Karr *et al* 2006), Reiter (2005), Matthews and Harel (2011), and McClure and Reiter (2016). One of the most challenging - and probably intractable - problems is that it is often possible to combine a data set with other publicly available data to identify an individual and reveal something about them. There have been several well-known public incidents of this kind, such as the identification of individual subscribers from the Netflix Prize dataset (Narayanan and Shmatikov, 2008) and the identification of the medical records of Massachusetts governor William Weld (Anderson, 2009).

From the perspective of statistical challenges, work continues to develop statistical methods of disclosure control - such as the development of differential privacy (Dwork and Roth, 2014). More generally, statistical tools are being developed to permit analysis without divulging the identity of individuals. For example, multi-party computation is a strategy to calculate aggregate statistics for a collection of individuals without requiring any individual to give away their value (Cramer *et al*, 2015).

Challenge 15: Continue to explore anonymisation and de-identification methods.

8. Conclusion

In the above, I have sought to identify and characterise what I thought were the main statistical challenges arising from administrative data. There are other challenges, including:

The communication of uncertainty: as statisticians, we are familiar with uncertainty arising from sampling variation, and with methods of communicating that uncertainty, such as confidence intervals. However, since the sources of uncertainty in administrative data are many and diverse, and may not include sampling variation, we need to find other ways to communicate (and indeed perhaps even define) such uncertainties. In some contexts this is already done. For example, the Bank of England's August 2016 Inflation Report (Bank of England, 2016), Chart 5.1 shows a fan chart with "To the left of the vertical dashed line, the distribution reflects the likelihood of revisions to the data over the past; to the right, it reflects uncertainty over the evolution of GDP growth in the future". More, however, remains to be done. Manski (2014) has a good discussion of the issues.

Statistical education: Challenge 1 above was about statistical education, although limited to the context of data quality. Administrative data are becoming so important, and so widely used (as a consequence of automatic data capture) that one can argue a case for more specialised teaching of specific methods related to administrative data.

Legal environment: The growth of awareness of modern data analysis technology has stimulated considerable legal and regulatory thought, much a consequence of the privacy and confidentiality issues discussed above. On 14th April 2016 the EU's General Data Protection Regulation (EU, 2017) was adopted by the European Parliament, and on 27th April 2017 the UK's Digital Economy Act received Royal assent (UK, 2017). These changes will certainly impact how personal data are stored, and are likely to impact statistical analyses of administrative data.

As a final comment, applied statisticians often emphasise the importance of being familiar with the data generation process. Understanding where the data come from and how they are collected can lead to the avoidance of many misunderstandings and mistakes. At first glance it might seem as if this is less critical for administrative data. This, however, is not the case. Issues of data quality, changes over time, changing regulatory and legal environments, advances in data capture and access technology, and a host of other factors are likely to impact administrative data, and their analysis. In fact, because the data will have been primarily collected for some operational purpose, these changes will almost certainly have been made without any subsequent statistical analysis in mind. They may not even be reported to the statistician who is later analysing the data. It is thus even more important - perhaps essential - that the statistician understands the data collection process. But note that this is a two way communication. If they are aware of the analyses to be undertaken later, the data producers will be able to adjust their data collection and recording processes to facilitate the subsequent analyses.

The aim of this paper is to raise awareness and stimulate discussion amongst statisticians of the need for methodological statistical work on administrative data. Such data are being used more and more widely - partly a consequence of the "big data" revolution. But drawing valid conclusions from such data encounters problems distinct from the more familiar and well-trodden paths of sampling theory inference. The problems are diverse and heterogeneous, so it is doubtful that a unifying theory as elegant as that of sampling theory can be developed. But nevertheless some principles apply. These include the need to cope with rather different kinds of data quality issues, the recognition that, despite superficial appearances, one typically does not have "all" the data, possible mismatches between the question one wants to answer and the information in the available data, challenges arising from the fact that the data are (usually) merely observational, so that elucidation of causality is difficult, the need to combine data from multiple rather different sources, and issues of confidentiality, privacy, and anonymisation which might be rather different from those of survey data.

Acknowledgements

The first draft of this paper was written as part of the Isaac Newton Institute Programme on *Data Linkage and Anonymisation*, July to December 2016. I would like to express my appreciation to the three anonymous referees and the associate editor for their detailed and helpful comments, which led to substantial improvement of the paper. The opinions expressed in this paper are the personal opinions of the author, and do not necessarily reflect those of any organisation with which the author is associated.

References

Anderson N. (2009) "Anonymized" data really isn't - and here's why not. https://arstechnica.com/tech-policy/2009/09/your-secrets-live-online-in-databases-of-ruin/ Accessed 5 March 2017.

Antoni M. (2013) Linking survey data with administrative employment data: the case of the German ALWA survey. *NTTS 2013*, 279-289.

Ashley J., Driver R., Hayes S., and Jeffery C. (2005) Dealing with data uncertainty. *Bank of England Quarterly Bulletin*, *Spring.* http://www.bankofengland.co.uk/publications/Documents/quarterlybulletin/qb050101.pdf Accessed 11 March 2017.

Baker R., Brick J.M., Bates N.A., Battaglia M., Couper M.P., Dever J.A., Gile K.J., and Tourangeau R. (2013). Summary report of the AAPOR task force on non-probability sampling. *Journal of Survey Statistics and Methodology*, **1,** 90-143.

Banasik J. and Crook J. (2004) Does reject inference really improve the performance of application scoring models? *Journal of Banking and Finance*, **28**, 857-874.

Berka C., Humer S., and Moser M. (2012) Combination of evidence from multiple administrative data sources: quality assessment of the Austrian register-based Census 2011. *Statistica Neerlandica*, **66**, 18-33.

Bethlehem J.G. (2010) Selection bias in web surveys. *International Statistical Review*, **78**, 161-188.

Biemer P., Trewin D., Bergdahl H, and Japec L. (2014) A system for managing the quality of official statistics. *Journal of Official Statistics*, **30**, 381-415.

Cavallo A. (2013) Online and official price indexes: measuring Argentina's inflation. *Journal of Monetary Economics*, **60**, 152-165.

Cavallo A. and Rigobon R. (2016) The billion prices project: using online prices for measurement and research. *Journal of Economic Perspectives*, **30**, 151-178.

Ćetković P., Humer S., Kausl A., Lenk M., Moser M., Rechta H., and Schnetzer M. (2013) Quality measurement in administrative statistics with a special focus on quality assessment of imputations. *NTTS 2013*, 247-256.

Christen P. (2012) *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Springer, Heidelberg.

Copas J.B. and Li H.G. (1997) Inference for non-random samples. *Journal of the Royal Statistical Society, Series B*, **59**, 55-95.

Cramer R., Damgård I.B., and Nielsen J.B. (2015) *Secure Multiparty Computation and Secret Sharing*. Cambridge University Press, New York.

Cunningham A. and Jeffery C. (2007) Extracting a better signal from uncertain data. *Quarterly Bulletin of the Bank of England, Q3*. http://www.bankofengland.co.uk/publications/Documents/quarterlybulletin/qb070301.pdf Accessed 11 March 2017.

Daas P.J.T., Arends-Tóth J., Schouten B., and Kuijvenhoven L. (2008) Proposal for a quality framework for the evaluation of administrative and survey data. http://www.pietdaas.nl/beta/pubs/pubs/ESSnet_Vienna_paper.pdf Accessed 26th February 2017.

De Jonge E. and van der Loo M. (2013) *An Introduction to Data Cleaning with R*. Statistics Netherlands. https://cran.r-project.org/doc/contrib/de_Jonge+van_der_Loo-Introduction_to_data_cleaning_with_R.pdf Accessed 8 March 2017.

De Veaux R.D. and Hand D.J. (2005) How to lie with bad data. *Statistical Science*, **20**, 231-238.

De Waal T., Pannekoek J., and Scholtus S. (2011) *Handbook of Statistical Data Editing and Imputation*. Wiley,

Direct Line (2011) https://www.directline.com/media/archive-2011/news-11072011 Accessed 6th May 2016.

D'Orazio M., Di Zio M., and Scanu M. (2006) *Statistical Matching: Theory and Practice*. John Wiley and Sons, Chichester.

Duncan G.T., Elliott M., and Salazar-González J-J. (2011) *Statistical Confidentiality: Principles and Practice*. Springer, New York.

Dwork C. and Roth A. (2014) The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, **9**, 211-407.

ESS (2020) European Statistical System Vision 2020. http://ec.europa.eu/eurostat/web/ess/about-us/ess-vision-2020 Accessed 18 February 2017.

ESS Admin (2015) Administrative data sources business project. http://ec.europa.eu/eurostat/documents/7330775/7339647/ADMIN+fact+sheet.pdf/cbb590b2-9d6f-439c-af2d-ca8b5e9cf1f7 Accessed 18 February 2017.

ESSNet Admin Data Workshop (2013) https://ec.europa.eu/eurostat/cros/content/essnet-admin-data-workshop-using-administrative-data-sts-evaluation-questionnaire-main_en Accessed 18 February 2017.

ESSNet (2017) https://ec.europa.eu/eurostat/cros/page/essnet_en Accessed 18 February 2017.

EU (2017) http://www.eugdpr.org/ Accessed 2 July 2017.

Eurostat (2000) *Assessment of the Quality in Statistics*. http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=FDD8499ECE9F43685CAFDD2A10EBE317?doi=10.1.1.5.8718&rep=rep1&type=pdf Accessed 8 March 2017.

Eurostat (2003) http://ec.europa.eu/eurostat/documents/64157/4374310/36-QUALITY-ASSESSMENT-ADMINISTRATIVE-DATA-STATISTICAL-PURPOSES_2003.pdf/37373e67-d69c-4215-b727-5b036393b80f Accessed 26 February 2017.

Eurostat (2011) *European Statistics Code of Practice*. http://ec.europa.eu/eurostat/documents/3859598/5921861/KS-32-11-955-EN.PDF/5fa1ebc6-90bb-43fa-888f-dde032471e15 Accessed 26 February 2017.

Fellegi I.P. and Sunter A.B. (1969) A theory for record linkage. *Journal of the American Statistical Association*, **64**, 1183-1210.

Fowler F.J. (1995) *Improving Survey Questions: Design and Evaluation*. Sage Publications, Thousand Oaks.

Hand D.J. and Henley W.E. (1993) Can reject inference ever work? *IMA Journal of Mathematics Applied in Business and Industry,* **5**, 45–55.

Hand D.J. (2001) Reject inference in credit operations. *Handbook of Credit Scoring*. Ed. Elizabeth Mays. Chicago: Glenlake Publishing. p225-240.

Hand D.J. (2004) *Measurement Theory and Practice: the World Through Quantification*. John Wiley and Sons, Chichester.

Hand D.J. (2006) Classifier technology and the illusion of progress (with discussion). *Statistical Science*, **21**, 1-34.

Hand D.J. (2008) *Statistics: A Very Short Introduction*. Oxford University Press, Oxford.

Hand D.J., Blunt G., Kelly M.G., and Adams N.M. (2000) Data mining for fun and profit. *Statistical Science*, **15**, 111-126.

Hand D.J. and Blunt G. (2001) Prospecting for gems in credit card data. *IMA Journal of Management Mathematics*, **12**, 173-200.

Heckman J.J. (1976) The common structure of statistical models of truncation, sample selection and limited dependent variables, and a simple estimator for such models. *Annals of Economics and Social Measurement*, **5**, 475-492.

Hellerstein J. (2008) Quantitative data cleaning for large databases. http://db.cs.berkeley.edu/jmh/papers/cleaning-unece.pdf Accessed 8 March 2017.

Hodson H. (2014) Google Flu Trends gets it wrong three years running. *New Scientist*, 13 March 2014. https://www.newscientist.com/article/dn25217-google-flu-trends-gets-it-wrong-three-years-running/ Accessed 5 March 2017.

Horn S. and Czaplewski R (2013) Combining survey and administrative data using state space models. *NTTS 2013*, 174-183

Ioannidis J. (2005) Why most published research findings are false. *PloS Medicine*, **2**, 696-701.

Israel (2007) *Pros and Cons for Using Administrative records in Statistical Bureaus.* http://www.oecd.org/std/41143741.pdf Accessed 8 March 2017.

Karr A.F., Kohnen C.N., Oganian A., Reiter J.P., and Sanil A.P. (2006) A framework for evaluating the utility of data altered to protect confidentiality. *The American Statistician*, **60**, 224-232.

Karr A.F., Sanil A.P., Banks D. L. (2006) Data quality: A statistical perspective, *Statistical Methodology*, **3**, 137-173.

Kim W., Choi B-Y., Hong E-K., Kim S-K., and Lee D. (2003) A taxonomy of dirty data. *Data Mining and Knowledge Discovery*, **7**, 81-99.

Kloek W. and Vâju S. (2013) The use of administrative data in integrated statistics. *NTTS 2013*, 128-138.

Lewis D. and Woods J. (2013) Issues to consider when turning to the use of administrative data: the UK experience. *NTTS 2013*, 549-557.

Manski C. (2014) Communicating uncertainty in official economic statistics. National Bureau of Economic Research Working Paper 20098. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2432840 (accessed 27 August 2016)

Matthews G. J. and Harel O. (2011) Data confidentiality: a review of methods for statistical disclosure limitation and methods for assessing privacy. *Survey Statistics*, **5**,1–29.

McClure D. and Reiter J.P. (2016) Assessing disclosure risks for synthetic data with arbitrary Intruder knowledge, *Statistical Journal of the International Association of Official Statistics*, **32**, 109 - 126.

Meader R. and Tily G. (2008) Monitoring the quality of national accounts. *Economic and Labour Market Review*, **2**, 24-33.

Memobust Handbook (2014) Quality of statistics module. https://ec.europa.eu/eurostat/cros/system/files/Quality%20Aspects-01-T-Quality%20of%20Statistics%20v1.0.pdf Accessed 26 February 2017.

Narayanan A. and Shmatikov V. (2008) Robust de-anonymization of large sparse datasets. https://www.cs.cornell.edu/~shmat/shmat_oak08netflix.pdf Accessed 5 March 2017.

Nordbotten, S. (2010). The use of administrative data in official statistics - past, present and future: with special reference to the Nordic countries. In Michael Carlson, Hans Nyquist, Mattias Villani, *Official statistics : methodology and applications in honour of Daniel Thorburn*, pp. 205-223, Published by Department of Statistics, Stockholm University, and Statistics Sweden.

NTTS (2013) *New Techniques and Technologies for Statistics: The Meeting Place for Researach in Official Statistics*. https://ec.europa.eu/eurostat/cros/system/files/NTTS2013%20Proceedings_0.pdf Accessed 17 February 2017.

NTTS (2015) *New Techniques and Technologies for Statistics: Reliable Evidence for a Society in Transition*. https://ec.europa.eu/eurostat/cros/system/files/NTTS2015%20proceedings.pdf Accessed 17 February 2017.

OECD (2016) *Short-Term Economic Statistics (STES) Administrative Data: Two Frameworks of Papers.* http://www.oecd.org/std/short-termeconomicstatisticsstesadministrativedatatwoframeworksofpapers.htm#Definition Accessed 8 March 2017.

ONS (2016) http://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/bulletins/crimeinenglandandwales/yearendingmar2016 Accessed 2 September 2016.

ONS (2016b) https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/articles/noteonthedifferencebetweennationalinsurancenumberregistrationsandtheestimateoflongterminternationalmigration/2016 Accessed 28 August 2016.

Pearl J., Glymour M., and Jewell N.P. (2016) *Causal Inference in Statistics: A Primer*. John Wiley and Sons, Chichester.

Presser S., Rothgeb J.M., Couper M.P., Lessler J.T., Martin E., Martin J., and Singer E. (eds). (2004) *Methods for Testing and Evaluating Survey Questionnaires*. New York: John Wiley & Sons, Inc.

Rässler S. (2002) *Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches*. Springer-Verlag, New York.

Reiter J.P. (2005) Estimating risks of identification disclosure for microdata. *Journal of the American Statistical Association*, **100**, 1103-1113.

Romanov D. and Gubman Y. (2013) Estimation of measurement error in categorical income survey data. *NTTS 2013*, p78-87.

Ruggles P. (2015) *Review of Administrative Data Sources*. https://www.census.gov/content/dam/Census/library/working-papers/2015/acs/2015_Ruggles_01.pdf Accessed 18 February 2018.

Scholtus S. and Bakker BFM (2013) Estimating the validity of administrative and survey variables by means of structural equation models. *NTTS 2013*, 290-299.

Scholtus S., Bakker B.F.M., and van Delden A. (2015) Modelling measurement error to estimate bias in administrative and survey variables. *NTTS 2015*, p451-455.

Statistics Canada (2015) http://www.statcan.gc.ca/pub/12-539-x/2009001/administrative-administratives-eng.htm Accessed 18 February 2018.

Statistics Netherlands (2014) *Quality Guidelines 2014: Statistics Netherlands Quality Assurance Framework at Process Level*. http://ec.europa.eu/eurostat/documents/64157/4374310/01-Quality-Guidelines-2014-Statistics-Netherlands-Quality.pdf/292b18bc-9bfd-426d-9282-785aabc43126 Accessed 26 February 2017.

Statistics New Zealand (2016) *Guide to Reporting on Administrative Data Quality*. http://www.stats.govt.nz/methods/data-integration/guide-to-reporting-on-admin-data-quality/explaining-framework.aspx# Accessed 18 February 2018.

Trépanier J, Pignal J., and Royce D. (2013) *Administrative Data Initiatives at Statistics Canada*. http://www.copafs.org/UserFiles/file/fcsm/G1_Trepanier_2013FCSM.pdf Accessed 18 February 2018.

UK (2017) http://www.legislation.gov.uk/ukpga/2017/30/contents/enacted Accessed 2 July 2017.

UKSA (2014) *Quality Assurance and Audit Arrangements for Administrative Data*. UK Statistics Authority.

UKSA (2015) *Administrative Data Quality Assurance Toolkit*. UK Statistics Authority.

UNECE (2016) http://www1.unece.org/stat/platform/display/Collection/2012+Data+Collection+Seminar%3A+Documents Accessed 18 February 2017.

Vâju S., Agafiţei M., Gras F., Kliek W., and Reis F. (2015) Measuring the quality of multisource statistics. *NTTS 2015*, p456-459

van Nederpelt P.W.M (2009) *Checklist quality of statistical output*. https://www.cbs.nl/NR/rdonlyres/4119715F-7437-4379-9A70-90A0893F949E/0/2009ChecklistQualityofStatisticalOutput.pdf Accessed 18 Frebruary 2017.

Wallgren A. and Eallgren B. (2014) *Register-based Statistics: Statistical Methods for Administrative Data, 2nd ed*. Wiley, Chichester

Winkler W.E. (2006) Overview of record linkage and current research directions. *Statistical Research Division, U.S. Census Bureau, Washington DC.*