Administrative Data
Research Network

An ESRC Data
Investment

Data quality issues in administrative data

ADRN Publication

Author: David McLennan
Series editors: Elaine Mackey & Mark Elliot

Better Knowledge Better Society

Data quality issues in administrative data

Written by David McLennan

Edited by Elaine Mackey and Mark Elliot

ADRN Publication - February 2018

This guide, ‘Data quality issues with administrative data’ was created by David McLennan, edited by Elaine Mackey and Mark Elliot and published by the Administrative Data Research Network.

David McLennan is Director and Senior Research Fellow at deprivation.org, a not-for-profit research organisation undertaking a wide range of studies in poverty, deprivation and inequality to help provide the evidence base for policies to address these three prominent socio-economic challenges. The organisation works across the world, spanning low-income, middle-income and high-income countries (www.deprivation.org). In Southern Africa, they work through the sister organisation SASPRI (Southern African Social Policy Research Insights) (www.saspri.org).

Contents

1. Introduction

1.1. Description of ADRN

The Administrative Data Research Network (ADRN) is a UK-wide partnership between academia, government departments and agencies, national statistical authorities, funders and the wider research community to facilitate new economic and social research based on routinely collected government administrative data.

The Network is establishing a new legal, secure and efficient pathway for the research community to access de-identified linked administrative datasets.

This will potentially benefit our society by providing a greater evidence base to inform policy.

The Network consists of:

- four Administrative Data Research Centres (ADRCs):
    - ADRC England: led by the University of Southampton
    - ADRC Northern Ireland: led by Queen's University Belfast
    - ADRC Scotland: led by the University of Edinburgh
    - ADRC Wales: led by Swansea University
- an overarching Administrative Data Service, which is the co-ordinating body of the Network
- administrative data owners
- the Economic and Social Research Council (the funding body)
- the UK Statistics Authority (chairing the ADRN Board)

The ADRN has commissioned this guide on data quality to support the development of knowledge and skills in the subject area.

1.2. Who this guide is aimed at

This guide is relevant for academic researchers accessing the ADRN and those in the wider social and economic sciences community. It may also be of interest to those who work in government, survey agencies, official statistics, charities or the private sector and are interested in learning about data quality issues affecting administrative datasets.

2. An introduction to administrative data

2.1 About administrative data

Administrative data are those that are routinely collected for operational purposes such as administering public services, rather than for a specific research objective. The majority of administrative datasets are collected at source as microdata, which are data about individual entities. These individual entities may be individual people (e.g. school data on pupil examination scores), individual households (e.g. Council Tax records), individual properties (e.g. Land Registry data on property sale price and date) or individual events (e.g. police data on crimes or incidents of disorder). Most administrative datasets are either collected directly by government (central, devolved or local) or collected on behalf of government. Although administrative data are collected for operational purposes, they are also becoming increasingly important as a basis for social and economic statistics and research.

One of the major attractions of the secondary analysis of administrative data for statistics and research is the financial benefit of using datasets that are already routinely collected, when compared to the high costs of primary data collection through social surveys. The process of routine data collection also facilitates certain analytical approaches. For instance, administrative data can be used to track detailed temporal trends (e.g. police recorded crime data have been used to identify peaks in offending at certain times of the day/week (e.g. Bromley and Nelson, 2002), which isn't possible through social surveys. A further major analytical attraction of administrative data is that they are often suitable for constructing statistics at small area level – unlike social surveys which are subject to survey sampling error and which are typically not representative down to small area level. The various indices of multiple deprivation that have been produced in the UK since 2000 are prime examples of how a range of administrative datasets have been used in combination to measure social and economic deprivation at small area level with a clear policy impact (e.g. Noble et al., 2000a; Noble et al., 2000b; Noble et al., 2001; Noble et al., 2003; Noble et al., 2004; Noble et al., 2005; Noble et al., 2006; Noble et al., 2007; Bradshaw et al., 2009; McLennan et al., 2010; Smith et al., 2015).

While administrative data clearly hold great potential for social and economic research and statistics, they are not explicitly collected for research and this means users should exercise a degree of caution when dealing with such data. This guide discusses some of the potential data quality issues relating to administrative datasets that users should be aware of and should investigate before and during their analysis. Awareness of these issues should also inform users' interpretation of the results.

2.2 Purpose of the guide

The varied nature of administrative data sources and the varied analytical applications of these data mean that the options for dealing with the data quality issues may be both dataset-specific and user-specific. It is therefore not possible to give a definitive list of data quality issues or a definitive list of remedial actions in this guide because each case needs to be considered individually. So this guide aims to:

(i) raise users' awareness of the types of potential data quality issues affecting administrative data

(ii) help users choose the most appropriate form of administrative dataset for their needs

(iii) give examples of ways in which users can deal with data quality issues at various stages of the analytical process

This ADRN guide does not deal with data quality issues that are directly related to data linkage techniques – there is a separate ADRN guide Introduction to Data Linkage (Harron, 2016). Please refer to that guide for information about the different types of data linkage (deterministic versus probabilistic) and the data quality concerns that need to be explored and dealt with as part of the linkage process.

Many of the examples discussed in this guide are presented in relation to police recorded crime data, which is an administrative dataset consisting of individual criminal events. However, many of the data quality issues raised and suggested responses to deal with the issues are equally relevant for person- or household-based administrative datasets, such as social security benefits data from the Department for Work and Pensions, pupil education data from the Department for Education, patient data from the National Health Service, offender data from the Ministry of Justice, etc.

3. Types of potential data quality issues affecting administrative data

Analytical users of administrative data have long been aware of a multitude of potential data quality issues. In recent years, the United Kingdom Statistical Authority (UKSA) summarised some of the key challenges that producers of statistics (e.g. government departments and the Office for National Statistics) face when using administrative data for statistical and research purposes. Box A on the next page is reproduced from UKSA's Exposure Draft report (United Kingdom Statistics Authority, 2014, p.12).

Many of the challenges identified by UKSA in **Box A** relate to difficulties in making sure there is a consistent approach within and between different organisations contributing to a particular administrative dataset. For example, each of the regional police forces in England and Wales has responsibility for collecting administrative microdata on crimes recorded within its own force boundaries. Although there is a National Crime Recording Standard[1] which stipulates the process through which a crime should be recorded, and although there are the Home Office Counting Rules[2] which stipulate how each criminal event should be coded, there are still acknowledged differences in the recording of some crimes between different police forces[3]. Many of the challenges in Box A can be regarded as possible underlying reasons for observed data quality issues with administrative data and, as such, potential users should reflect upon these challenges when considering the strengths and weaknesses of administrative datasets.

Work has also been undertaken internationally to develop a comprehensive data quality assessment framework for administrative data (Daas et al., 2012), and this framework has been used by a number of national statistical institutions (e.g. Netherlands, Sweden, Australia). This framework distinguishes between three different views on administrative data quality, referred to as 'hyperdimensions', they are:

1. source
2. metadata
3. data

The source hyperdimension relates to issues concerning the data sharing process, while the metadata hyperdimension relates to issues concerning the quality and comprehensiveness of any accompanying metadata. Of greatest relevance to this guide is the **data hyperdimension**, which focuses on possible quality concerns with the actual data content.

[1] Reproduced as Annex A of the Home Office Counting Rules
[2] https://www.gov.uk/government/publications/counting-rules-for-recorded-crime
[3] https://www.justiceinspectorates.gov.uk/hmic/publications/crime-recording-making-the-victim-count/

Box A

Lack of standardised application of data collection:

- inconsistencies in how different suppliers interpret local guidance
- differences in the use of local systems for the intended administrative function
- the distortive effects of targets and performance management regimes
- differing local priorities, data suppliers might require higher levels of accuracy for certain variables (for example payments) but less so for other aspects that are important to the statistical producer (for example demographics)

Variability in data suppliers' procedures:

- statistical producers typically do not have direct control over the development of guidance for data entry
- local checking of the data can be variable and might not identify incorrect coding or missing values
- local changes in policy could impact on how the data are recorded or on the coverage of the statistics

Quantity of data suppliers:

- there can be a large number of data suppliers, often spread geographically
- there can be many data collectors providing their data to an intermediary organisation for supply to a statistical producer

Complexity and suitability of administrative systems:

- administrative datasets can be complex containing large numbers of variables; it takes time, and therefore resource, to extract the necessary data required by the statistical producer
- data collation can be hampered by IT changes at the data supplier level
- data might need to be manipulated by the data supplier to meet the structural requirements of the statistical producer, leading to potential for errors

Public perceptions:

- lack of knowledge about use of personal data for statistical purposes
- concern that personal data should be sufficiently anonymised and secured

The data hyperdimension is conceptualised as consisting of five component dimensions of data quality:

1. technical checks
2. accuracy
3. completeness
4. time-related
5. integrability

Each of these five dimensions is measured by a series of indicators against which a dataset is assessed. The indicators under the **technical checks dimension** measure the technical usability of the file and data in the file; for example, the readability and formatting of the data, and the degree to which the data content matches the accompanying metadata descriptions. The indicators under the **accuracy dimension** measure the extent to which data are correct, reliable and certified; for example, the frequency of implausible values or combinations of values across different variables. The indicators under the **completeness dimension** measure the degree to which a data source includes data describing the corresponding set of real-world objects and variables; for example, the extent of under/over-enumeration, duplicates, missing values or imputed values. The indicators under the **time-related dimension** relate to the timeliness of data capture and supply for example the time lag between real-world change and this change being reflected in the data as well as stability of variables and values over time (e.g. definitional consistency). The indicators under the **integrability dimension** measure the extent to which the data source is capable of undergoing integration or of being integrated in the statistical system; for example, reliability of linking variables. Whilst it is outside the scope of this Guide to discuss the framework in any great detail, users of administrative data may wish to consult this resource when considering possible data quality issues relating to a dataset of interest. Appendix 1 provides a full list of the indicators within the data hyperdimension for further reference.

4. Choosing the most appropriate form of administrative dataset

The user base for administrative data is broad and varied, and the sensitive nature of many administrative datasets (such as a child's educational attainment or an adult's receipt of social security benefits) means that many users will not be permitted to access the source microdata at individual level. They can, instead, use publicly available macrodata (aggregate statistics derived from the individual level microdata) which protect the identities of individual data subjects. In response to the growing demand from users for data on social and economic outcomes, many government departments – as well as the Office for National Statistics – are now routinely producing statistical macrodata derived from administrative data. This drive across government to expand the breadth of socio-economic statistics derived from administrative data has also stimulated a programme of action, led by the UK Statistics Authority (UKSA), to investigate and inform users of possible data quality issues. UKSA's statutory objective is to "promote and safeguard the production and publication of official statistics that serve the public good"[4]. As such, UKSA has a responsibility to make sure any official macrodata statistics based on administrative sources are of suitable quality for statistical and research purposes and are accompanied by suitable metadata. As a result, UKSA works with the relevant government departments that collect the source microdata to maximise quality assurance procedures throughout the data collection and management process. Where UKSA is satisfied that a statistical dataset is of suitably high quality, that dataset is awarded 'National Statistics' status. National Statistics are a subset of official statistics which have been certified by UKSA as compliant with its Code of Practice for Official Statistics. National Statistics accreditation indicates that a statistical dataset has been through a rigorous process of quality assurance.

In many cases, the process of producing statistical macrodata from the source microdata will involve a series of additional quality assurance checks. As noted above, this is particularly the case in respect of National Statistics. When a user has the option to access either the source microdata or the derived macrodata, an important consideration will be whether the derived macrodata is sufficient for the analytical purpose. If so, users may benefit from utilising the macrodata.

However, even when users are satisfied that the macrodata are sufficient for their analytical needs, consideration should still be given to the potential data quality issues that might relate to the underlying source microdata, and users should further consider the impacts of any data manipulations applied in the derivation of the macrodata. In light of this, the data quality issues discussed in the next section are relevant to users of both microdata and macrodata sources of administrative data.

[4] https://www.gov.uk/government/publications/how-national-and-official-statistics-are-assured/how-national-and-official-statistics-are-assured

5. Selected examples: dealing with data quality at different stages

This section presents examples of approaches that have been implemented at three different stages of the data collection/manipulation process:

1. Adjustments to the source microdata
2. Adjustments during the process of aggregating from microdata to macrodata
3. Post-aggregation adjustments implemented on the macrodata

The exact approach to dealing with data quality issues will depend on the specific dataset and the specific analytical application. The examples provided here illustrate data quality issues that might be relevant to a user's chosen administrative dataset.

1. Examples of adjustments to the source microdata

A. Data categorisation

A common challenge of working with administrative data is the potential for variable coding or categorisation to vary over time and/or between geographical areas. For example, the Home Office can legitimately change the Counting Rules which police forces use to decide which type of crime should be recorded, and this can lead to changes in the variable coding over time in the police recorded crime microdata. And, although every police force is required to provide macrodata statistics routinely on a specified set of Home Office defined notifiable offence types, many forces have their own particular coding scheme for crime types at local level, which means that the underlying microdata variable coding varies between police forces.

Where users of police-recorded crime microdata are interested in measuring levels of crime over time and/or across multiple forces, it will be necessary to construct lookup tables to convert the local crime coding schemes to a nationally consistent scheme. A new standardised crime code variable can then be added to the source microdata. The Home Office has already made efforts to tackle this issue before releasing crime microdata on the police.uk website (https://www.police.uk/). Users of the police.uk data should be mindful of the processes adopted by the Home Office in this regard, while any users of the raw microdata sourced direct from police forces (i.e. not via police.uk) should review the local crime coding schemes in detail.

B. Geocoding

Most administrative datasets contain a form of locational data. This might be in relation to the home address of an individual/household or the location at which an event took place. The geographical information in an administrative dataset may consist of full addresses including postcodes; postcodes but not full addresses; grid references; and higher level geographical identifiers (e.g. local authority district).

Users of such data should be aware of the potential for errors in the locational data, such as incomplete or mistyped postcodes (e.g. where a numeric zero has been typed instead of a capital O, or vice versa) or grid references which indicate that an object is located in an implausible location (e.g. households appearing to be located over an expanse of open water, or crimes recorded by a particular police force appearing to have been committed many miles outside of that force's geographical boundary).

Where users are provided with postcodes and grid references in the same dataset, these should be compared to ascertain the level of agreement. Furthermore, where the dataset also contains a higher level geographical identifier, this should be used as an additional check of likely accuracy of the postcode/grid reference location. For instance, depending on the other geographical information in a dataset, it may be possible to correct mistyped postcodes in the source microdata using third party software applications.

A further potential data quality issue concerns the geocoding of records to approximate locations. For example, when the exact location of an event is not known, the record might be geocoded to an approximate location chosen by the person responsible for inputting the data (such as the centroid of the relevant ward). Users should review the geocoding in the dataset to check for unusual spatial patterns and implement a suitable correction to the source data if necessary.

2. Examples of adjustments during the process of aggregating from microdata to macrodata

A. Spatial smoothing

As noted above, although most administrative datasets are collected at source as microdata, many analytical applications use macrodata aggregated to a relevant spatial level. One of the key advantages of administrative microdata is that the presence of location information permits the aggregation of objects (people, households, institutions, events etc) to customised geographies. These geographies include not only official statistical classifications such as regions, districts and wards, but also bespoke spatial areas designed for a particular policy purpose or for a particular research purpose[5]. Users of administrative data should consider the process through which microdata is aggregated to generate macrodata. The most common technique for producing macrodata at a given spatial scale is to assign each object in the microdata to the spatial units in which it is physically located, and to count the number of objects per spatial unit[6]. However, there may be situations when the user deems a simple aggregation of this kind inappropriate by the user.

For example, in the Crime Domain of the English Indices of Deprivation, a spatial smoothing approach was adopted in order to better reflect the risk of experiencing crime between different geographical areas (see, for instance, Noble et al., 2004; Noble et al., 2007; McLennan et al., 2010; Smith et al., 2015). The motivation behind this was an acknowledgement that crimes are sometimes geocoded to locations that fall very close to the boundaries of spatial units, and in these cases there may be an unintended systematic over-count in one area and under-count in the neighbouring area. This is particularly the case when the boundaries of spatial units run along the centre line of roads. The spatial smoothing process helps to even out the spatial distribution of crimes so that areas that share a boundary running along a road also share the crimes that are recorded as happening along that road. This particular methodological approach was adopted in the Crime Domain of the Indices of Deprivation as the objective was to produce small area level counts of crime, adjusted for the geocoding issues discussed here[7].

[5] For example, the New Deal for Communities (NDC) programme in England consisted of 39 highly deprived 'neighbourhoods' which were targeted with an area-based policy intervention between 1999 and 2009. These 39 neighbourhoods did not map directly onto existing statistical boundaries but rather they were designed through local partnerships to encompass meaningful neighbourhoods. The national evaluation of the NDC programme was able to generate a wide range of statistics (on the themes such as worklessness, education, crime and health etc) for each of the 39 bespoke neighbourhood areas by aggregating administrative microdata from individual level to NDC level (see, for instance: McLennan and Whitworth, 2008; Whitworth et al., 2010; Wilkinson and McLennan, 2010; Gutiérrez Romero and Noble, 2008).

[6] For instance, calculating an unemployment rate for a given spatial area would typically consist of counting the number of unemployed people living in the area and expressing this as a percentage of the total number of working age people living in the area.

[7] The same approach was applied to the Road Traffic Accidents indicator in the Living Environment Domain of the Indices of Deprivation 2004, 2007, 2010 and 2015.

There are, however, a variety of alternative spatial smoothing techniques that can be used when the detailed geographical locations of microdata objects are known and the aim is to identify particular geographical concentrations of these objects. Chainey et al. (2008) reviewed a selection of such approaches for the purpose of predicting the locations of future crime hotspots based on the spatial patterning of recent criminal events. Of the approaches they examined, Kernel Density Estimation (KDE) was identified as the best predictor of future crime hotspots. KDE uses the point locations of objects (in this case criminal events) to calculate the density of points across any given geographical area and conveys these densities as a continuous surface that can be mapped thematically. The KDE technique allows the user to control a number of computational parameters, so this approach is flexible and can be adapted to a range of research purposes. Although the discussion here is in relation to crime hotspots, these spatial smoothing approaches can be applied to any administrative dataset where the grid references of the objects are known and the aim is to assess geographical clustering.

3. Examples of post-aggregation adjustments implemented on the macrodata

A. Constraining to higher-level aggregates

A key data quality concern with administrative microdata is the possibility of missing, incomplete or imprecise information. Examples of data fields where these types of data quality issues are commonly found include times and dates (of event occurrence or service registration etc), personal identifiable information (such as age, sex, ethnicity etc), unique identification codes (e.g. National Insurance Number, NHS Patient Number), and geographical location information (e.g. address, postcode or grid reference). Missing, incomplete or imprecise geographical information can be a particular problem for administrative datasets and may be due to difficulties in recording the address of a person/household (especially where the person/household moves home frequently) or the address of an event (e.g. if a crime reported as occurring 'on the high street' but without further details of exactly where on the high street).

Although the quality of detailed geographical information in administrative microdata may present some challenges to users, there may be ways to adjust the data once it has been aggregated to form macrodata statistics. For instance, certain cases in a given dataset may have information about the district in which a person lives, but detailed address information for that person may be missing. While the district identifier will allow the person in the dataset to be matched to his/her home district, it will not be possible to match them to their home ward or home neighbourhood within the district.

Users should therefore review how cases such as this are dealt with in the analysis. A common approach in this regard is to first calculate ward/neighbourhood level counts using the detailed address information for those cases for which it is known, and separately calculate district level counts using the district level identifier where it is known.

The ward/neighbourhood level counts are then rescaled across the parent district which has the effect of distributing the cases with missing address data based upon the distribution of known cases. This process of rescaling the ward/neighbourhood level counts is known as 'constraining' the data, and it produces ward/neighbourhood level counts that sum to the more complete district level count. This approach has proved particularly useful in the spatial analysis of crime patterns where variations are evident between police forces in terms of the proportion of crimes successfully geocoded. The constraining of small area level crime counts to aggregate statistics at a higher geographical level (typically Community Safety Partnership and Police Force levels) permits more comparable assessments of crime patterns and trends between small areas in different police forces (see McLennan et al., 2010, p.44, for more details).

B. Shrinkage estimation

One of the advantages of administrative data as a source for generating macrodata is that the aggregate statistics can typically be calculated for small geographical areas. However, when the macrodata statistics for a small geographical area are based on small numbers of events (for example, counts of deprived individuals or counts of road traffic accidents), the results may be viewed as potentially unreliable, due to an unacceptably high standard error[8]. Shrinkage estimation is a technique to 'borrow strength' from results for larger geographical areas to avoid creating unreliable small area statistics (Noble et al., 2006, p.177-178). The effect of shrinkage estimation is to 'move' the small area level statistic somewhat in order to produce a result that is potentially less unreliable.

In shrinkage estimation, the statistic for a small area is therefore estimated as a weighted combination of that small area's value and the mean value for the larger area within which the small area is located. Without shrinkage, some small areas might have scores which do not reliably describe the situation (e.g. the unemployment rate, or the crime rate etc) in the area due to chance fluctuations over time. This problem is most pronounced in areas where the number of events is rare. Although shrinkage estimation is not specifically related to administrative data, it may be an important approach to consider if users are intending to generate small area level macrodata from administrative microdata, or if users are intending to use pre-aggregated macrodata from administrative sources.

[8] Although not a survey sample, statistics derived from administrative data can be viewed as a sample from a superpopulation.

6. Conclusion

Great strides are being made to increase access to and support the use of administrative data for research purposes in the United Kingdom. The Administrative Data Research Networl is central to this – supporting safe and responsible access to de-identified linked administrative microdata.

As the breadth of users and research applications of administrative data increases, so too does the need to raise awareness of the importance of data quality considerations. No dataset is ever perfect. Whilst administrative datasets are limited in terms of their content by the operational purpose for which they are collected, they nevertheless hold great potential for social and economic research.

Other guides in this series highlight innovative ways in which administrative data can be used, such as linking different administrative datasets from different sources to produce new insights not previously possibly using single datasets in isolation. In this guide, we have highlighted the value of administrative data for deriving statistics at various spatial levels, including small area level. Administrative data can therefore be used for a multitude of research purposes for which social surveys are not suited.

As discussed above, the data quality concerns that users of administrative data will encounter will often be both dataset-specific and user-specific. The purpose of this guide is to highlight some of the key themes (or 'dimensions') of data quality that users should take into consideration before and during their use of administrative data, and to provide real examples of how data quality concerns have been dealt with in recent research. As the research environment evolves, and administrative data becomes increasingly prevalent, feedback from users on data quality concerns will undoubtedly help data providers to improve the source data. The Administrative Data Research Network will continue to play a central role in facilitating this vital dialogue between users and providers.

7. Resources

7.1. Websites

General information on the ADRN can be found at adrn.ac.uk including other guides in this series and educational material.

There is more information about the UKSA's work on quality assuring administrative data on their website: https://www.statisticsauthority.gov.uk/osr/monitoring/administrative-data-and-official-statistics/

Information on the various Indices of Deprivation in England: https://www.gov.uk/government/collections/english-indices-of-deprivation

Information on the various Indices of Deprivation in Wales: http://gov.wales/statistics-and-research/welsh-index-multiple-deprivation/?lang=en

Information on the various Indices of Deprivation in Scotland: http://www.gov.scot/Topics/Statistics/SIMD

Information on the various Indices of Deprivation in Northern Ireland: https://www.nisra.gov.uk/statistics/deprivation

The Home Office's police.uk data portal, where users can download microdata sourced from the regional police forces: https://data.police.uk/

Users can download an array of administrative macrodata resources on labour market patterns and trends on the NOMIS website: https://www.nomisweb.co.uk/

A variety of other macrodata statistics derived from administrative data are available to download at national and sub-national level from: https://www.gov.uk/government/statistics

8. References

BRADSHAW, J., NOBLE, M., BLOOR, K., HUBY, M., RHODES, D., SINCLAIR, I., GIBBS, I., MCLENNAN, D. & WILKINSON, K. 2009. Local Index of Child Well-Being - Summary Report. London: Communities and Local Government.

BROMLEY, R. D. & NELSON, A. L. 2002. Alcohol-related crime and disorder across urban space and time: evidence from a British city. Geoforum, 33, 239-254.

CHAINEY, S., TOMPSON, L. & UHLIG, S. 2008. The utility of hotspot mapping for predicting spatial patterns of crime. Security Journal, 21, 4-28.

DAAS, P. J., OSSEN, S. J., TENNEKES, M. & BURGER, J. Evaluation and visualisation of the quality of administrative sources used for statistics. Paper for the European Conference on Quality in Official Statistics, 2012.

GUTIÉRREZ ROMERO, R. & NOBLE, M. 2008. Evaluating England's 'New Deal for Communities' programme using the difference-in-difference method'. Journal of Economic Geography, July 2008 1-20.

HARRON, K. 2016. Introduction to Data Linkage. Administrative Data Research Network.

LAVRAKAS, P. J. 2008. Encyclopedia of survey research methods, Sage Publications.

MCLENNAN, D., BARNES, H., NOBLE, M., DAVIES, J., GARRATT, E. & DIBBEN, C. 2010. English Indices of Deprivation 2010. London, UK: Department for Communities and Local Government.

MCLENNAN, D. & WHITWORTH, A. 2008. Displacement of crime or diffusion of benefit: Evidence from the New Deal for Communities Programme. Oxford Social Disadvantage Research Centre, University of Oxford.

NOBLE, M., BARNES, H., SMITH, G. A. N., MCLENNAN, D., DIBBEN, C., ANTTILA, C., SIGALA, M., AVENELL, D., SMITH, T. & MOKHTAR, C. 2005. Northern Ireland Multiple Deprivation Measure 2005. Belfast: Northern Ireland Statistics and Research Agency.

NOBLE, M., MCLENNAN, D., WILKINSON, K., WHITWORTH, A., BARNES, H. & DIBBEN, C. 2007. The English Indices of Deprivation 2007. London: Communities and Local Government.

NOBLE, M., SMITH, G., PENHALE, B., WRIGHT, G., DIBBEN, C., OWEN, T. & LLOYD, M. 2000a. Measuring multiple deprivation at the small area level: The Indices of Deprivation 2000. London: Department of the Environment, Transport and the Regions.

NOBLE, M., SMITH, G. A. N., WRIGHT, G., DIBBEN, C. & LLOYD, M. 2001. Northern Ireland Multiple Deprivation Measure 2001. Belfast: Northern Ireland Statistics and Research Agency.

NOBLE, M., SMITH, G. A. N., WRIGHT, G., DIBBEN, C., LLOYD, M. & PENHALE, B. 2000b. Welsh Index of Multiple Deprivation 2000. Cardiff: National Assembly for Wales.

NOBLE, M., WRIGHT, G., DIBBEN, C., SMITH, G. A. N., MCLENNAN, D., ANTTILA, C., BARNES, H., MOKHTAR, C., NOBLE, S., AVENELL, D., GARDNER, J., COVIZZI, I. & LLOYD, M. 2004. The English Indices of Deprivation 2004. London: Neighbourhood Renewal Unit, Office of the Deputy Prime Minister.

NOBLE, M., WRIGHT, G., LLOYD, M., DIBBEN, C., SMITH, G., RATCLIFFE, A., MCLENNAN, D., SIGALA, M. & ANTTILA, C. 2003. Scottish Index of Deprivation 2003: Summary Report Edinburgh: Scottish Executive.

NOBLE, M., WRIGHT, G., SMITH, G. A. N. & DIBBEN, C. 2006. Measuring multiple deprivation at the small area level. Environment and Planning A, 38, 169-185.

SMITH, T., NOBLE, M., NOBLE, S., WRIGHT, G., MCLENNAN, D. & PLUNKETT, E. 2015. The English Indices of Deprivation 2015. London: Department for Communities and Local Government.

UNITED KINGDOM STATISTICS AUTHORITY 2014. Exposure draft of a report from the UK Statistics Authority: Quality Assurance and Audit Arrangements for Administrative Data. UKSA.

WHITWORTH, A., MCLENNAN, D. & NOBLE, M. 2010. Crimes occurring and prevented in New Deal for Communities areas. Communities and Local Government, UK.

WILKINSON, K. & MCLENNAN, D. 2010. Narrowing the gap? Analysing the impact of the New Deal for Communities Programme on educational attainment. Communities and Local Government, UK.

9. Glossary

Home Office Counting Rules: the rules specifying how police forces should classify crimes into detailed crime type categories.

Hyperdimensions of data quality: overarching dimensions of data quality as proposed by Daas et al. (2012).

Indices of Deprivation: multidimensional measures of deprivation constructed and presented at small area level and often used as the basis for allocating resources to deprived areas.

Kernel Density Estimation: a methodological approach to identifying geographical concentrations of objects or events (and often used in crime hotspot analysis)

Macrodata: aggregate level statistics (often derived by summing or averaging individual level microdata)

Microdata: information about individual entities, such as individual people, properties, transactions or events.

Standard error: a measure of statistical accuracy of an estimate

Shrinkage estimation: a statistical approach sometimes applied to small area level statistics to 'borrow strength' from a higher level aggregation.

Small area level: small geographical areas, such as Census output geographies or wards.

Superpopulation: In the Encyclopedia of Survey Research Methods a superpopulation is defined as follows: "When data for a variable are gathered from a finite population and that variable is regarded to be a random variable, then the finite population is referred to as being "a realization from a superpopulation." A superpopulation is the infinite population that elementary statistical textbooks often describe as part of the enumeration of a finite population. It is because sampling theory is based on making inference for a well-defined finite population that the concept of superpopulation is needed to differentiate between a finite population and an infinite superpopulation." (Lavrakas, 2008)

Appendix 1

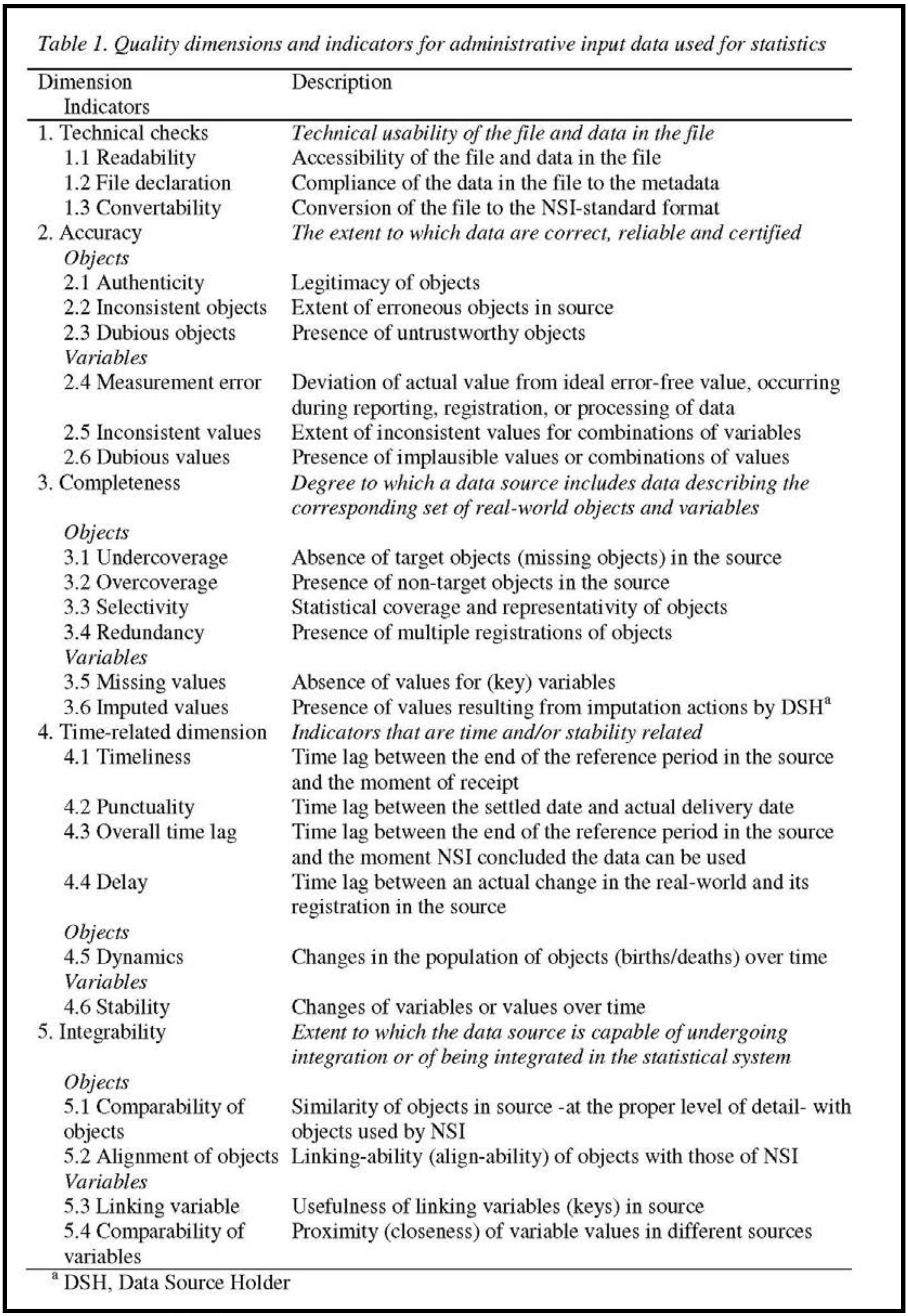

Table 1. Quality dimensions and indicators for administrative input data used for statistics

Dimension	Description
Indicators	
1. Technical checks	*Technical usability of the file and data in the file*
1.1 Readability	Accessibility of the file and data in the file
1.2 File declaration	Compliance of the data in the file to the metadata
1.3 Convertability	Conversion of the file to the NSI-standard format
2. Accuracy	*The extent to which data are correct, reliable and certified*
Objects	
2.1 Authenticity	Legitimacy of objects
2.2 Inconsistent objects	Extent of erroneous objects in source
2.3 Dubious objects	Presence of untrustworthy objects
Variables	
2.4 Measurement error	Deviation of actual value from ideal error-free value, occurring during reporting, registration, or processing of data
2.5 Inconsistent values	Extent of inconsistent values for combinations of variables
2.6 Dubious values	Presence of implausible values or combinations of values
3. Completeness	*Degree to which a data source includes data describing the corresponding set of real-world objects and variables*
Objects	
3.1 Undercoverage	Absence of target objects (missing objects) in the source
3.2 Overcoverage	Presence of non-target objects in the source
3.3 Selectivity	Statistical coverage and representativity of objects
3.4 Redundancy	Presence of multiple registrations of objects
Variables	
3.5 Missing values	Absence of values for (key) variables
3.6 Imputed values	Presence of values resulting from imputation actions by DSH[a]
4. Time-related dimension	*Indicators that are time and/or stability related*
4.1 Timeliness	Time lag between the end of the reference period in the source and the moment of receipt
4.2 Punctuality	Time lag between the settled date and actual delivery date
4.3 Overall time lag	Time lag between the end of the reference period in the source and the moment NSI concluded the data can be used
4.4 Delay	Time lag between an actual change in the real-world and its registration in the source
Objects	
4.5 Dynamics	Changes in the population of objects (births/deaths) over time
Variables	
4.6 Stability	Changes of variables or values over time
5. Integrability	*Extent to which the data source is capable of undergoing integration or of being integrated in the statistical system*
Objects	
5.1 Comparability of objects	Similarity of objects in source -at the proper level of detail- with objects used by NSI
5.2 Alignment of objects	Linking-ability (align-ability) of objects with those of NSI
Variables	
5.3 Linking variable	Usefulness of linking variables (keys) in source
5.4 Comparability of variables	Proximity (closeness) of variable values in different sources

[a] DSH, Data Source Holder

Taken from Daas et al. 2012

ADRN PUBLICATION
Produced by the Administrative Data Service
University of Essex, Wivenhoe Park, Colchester, Essex, CO4 3SQ
T. +44 (0) 1206 87 2976 E. help@adrn.ac.uk W. adrn.ac.uk

Funded by

E·S·R·C
ECONOMIC
& SOCIAL
RESEARCH
COUNCIL