Proceedings of Statistics Canada Symposium 2014
Beyond traditional survey taking: adapting to a changing world

Different contexts for the statistical use of administrative data

Loredana Di Consiglio - Piero Demetrio Falorsi[1]

Abstract

The project MIAD of the Statistical Network aims at developing methodologies for an integrated use of administrative data (AD) in the statistical process. MIAD main target is providing guidelines for exploiting AD for statistical purposes. In particular, a quality framework has been developed, a mapping of possible uses has been provided and a schema of alternative informative contexts is proposed. This paper focuses on this latter aspect. In particular, we distinguish between dimensions that relate to features of the source connected with accessibility and with characteristics that are connected to the AD structure and their relationships with the statistical concepts. We denote the first class of features *the framework for access* and the second class of features the *data framework*. In this paper we mainly concentrate on the second class of characteristics that are related specifically with the kind of information that can be obtained from the secondary source. In particular, these features relate to the target administrative population and measurement on this population and how it is (or may be) connected with the target population and target statistical concepts.

Key Words: Quality framework; GSBPM; Accessibility; Measurement.

1. Introduction

1.1 Description of the MIAD Project

The present work is part of the activities planned within the project Methods for use of Administrative data (MIAD; http://www1.unece.org/stat/platform/display/msis/Statistical+Network). MIAD members are the National Institute of Statistics (ISTAT) of Italy, Statistics Canada, Australian Bureau of Statistics, Statistics New Zealand and Statistics Sweden. The aim of the project is to develop coherent and well-founded strategies, which would allow fully exploiting the use of administrative data (AD) sources in the statistical process. In order to achieve this goal, first MIAD aimed at deepening the analysis of the dimensions to be considered when evaluating the usage of an external source for the statistical process. In this respect, first emphasis was given to the issue of developing a framework to assess the quality of DA and its statistical usability – prior to statistical usage, in order to determine if an AD source can be used for statistical purpose and how. In particular, the set of desired characteristics of a quality framework and the set of quality indicators before AD is introduced in the statistical process are developed. MIAD envisages also the need for guidelines to relate the results of this preliminary assessment to the effective use of AD in the statistical process, in order to suggest the thresholds for establishing how is appropriate to make use of AD, as for example, AD may be used either for direct tabulation of data (after preprocessing them), or as an auxiliary variable in the estimation process.

Another important related stream of activities that MIAD identified is mapping how AD are in fact and can be used into a statistical production process. In particular, the mapping onto the Generic Statistical Business Process Model (GSBPM) is applied as foundation. Also the GSBPM is re-analysed with the specific perspective of AD usage.

This explorative activity on possible different uses is connected with the possible actions to be performed on AD and to the overview of the most common scenarios of possible types of information frameworks the NSI faces to produce statistical information with AD.

[1]Loredana Di Consiglio, ISTAT, Via Cesare Balbo, 16 Italy, Rome, 00184 (diconsig@istat.it); Piero Demetrio Falorsi, ISTAT, Via Cesare Balbo, 16 Italy, Rome, 00184 (falorsi@istat.it)

On a successive phase of the project to be started during 2015, MIAD will focus also on the specific statistical methods for exploiting and treating AD that here are only briefly overviewed and on measures of output quality.

The present work is related to the explorative activity of the most common informative frameworks the NSI faces to produce statistical information with AD and it is mainly meant to set up a scheme of the statistical methods that are needed for integration of AD in the statistical production. These considerations would also be the basis, together with an evaluation of the input quality of the AD, to provide guidelines on effective usage of an AD source.

1.2 The main characteristics of the AD scenarios

The present work is focused on providing an overview of AD possible contexts and to start connecting the different alternative actions for statistical use to the alternative AD frameworks. In particular, here we aim to describe the most relevant background dimensions that influence the statistical use of AD and the procedures needed to integrate the AD in the statistical process; and that determine the methods to be applied to create statistical information. Some of these dimensions relate to features of the source that are connected with *accessibility*, and more in general with the environment, e.g. the relationships between the NSI and the external data providers or data owners, while other dimensions are related to the *structure* of the *objects* in the AD and its relationship with the statistical concepts.

We indicate the first class of features *the framework for access* to the data and the second class of features the intrinsic *AD structure*. In fact, some settings suggest exploiting AD for indirect use only, conversely in some other cases statistical outputs can be directly obtained from the AD, used alone or within a complex system of sources. In particular, related to the framework for access, we distinguish between the legal and institutional frameworks. Regarding the former, in fact some national legal frameworks give more powers than others for access to AD for statistical purposes - set out the limits to such access, and to the uses of AD. Often there are restrictions that data can only be used for specific statistical purposes, and that the confidentiality of individual records should be maintained. The legal framework influences availability at micro or aggregate level, affects also presence of identifiers (for privacy reasons) and the kind of variables that can be shared with the NSI, then this affect the possible uses of AD and the output quality obtained though them. A legislation that allows access to AD is in fact a fundamental precondition for their use. Regarding to the latter dimension, i.e. the *Institutional framework*, we mean the organization set up to acquire AD, for the purpose of their use. It mainly relates to characteristics such as timeliness and availability of metadata, i.e. dimensions that affect quality, in particular way, input quality and they are usually included in the quality assessment frameworks.

Moreover, a strong integrated system with effective links between NSI and the data provider(s) also helps guaranteeing the possibility to know in advance of planned changes in order to plan how the statistical office can offset. These aspects are not described here in depth; we refer to UNECE (2011) and Royce (2013) for more details on existing frameworks that NSIs has set up for using AD in the statistical process.

In the next section, we focus on the second set of features that we have denoted AD structure, highlighting the aspects that affect quality assurance and methods to be implemented to obtain statistical units and concepts and to improve their quality.

2. Different AD scenarios: elements of the data structure

2.1 The two-life cycle model

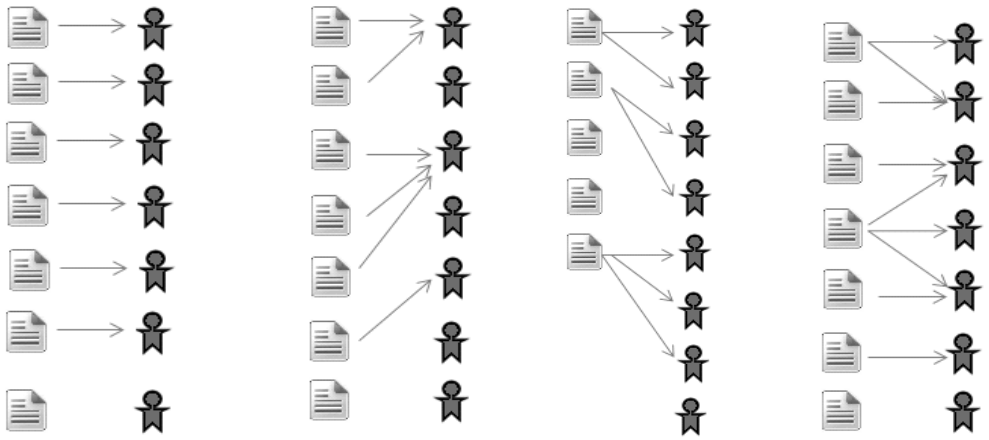
In describing the AD structure, we follow the two life cycle model by Zhang (2012). This model explicitly introduces the concept of "secondary" re-use of a source in the life cycle model. First, Bakker (2010) developed the Groves *et al*. (2004) model, that has been introduced in the survey context, to deal with the AD situation. Then, Zhang (2012) expanded Bakker model to fit in with a combined register setting. These models were explicitly introduced to consider the possible sources of errors in the process, and in the statistical re-use of a source, highlighting the process that the external source has to undergo. In particular, for our purpose, we are interested in

the second phase of the Zhang two life cycle model, where the relationships between the available AD records and the target concepts and the population NSI is interested in are described. This framework is particularly useful for describing how the different scenarios (and the related measures of input quality) impact the processes needed to obtain the desired statistical information and related risk of errors and how they affect the throughput and output quality. Here we describe the possible data structures in different situations. As we have already mentioned with data structure, we mean the types of features and measures on the administrative population and how it is (or may be) connected with the target statistical population and target statistical concepts. Following Zhang, we distinguish from representation (objects or units) and measurement – variables possible alternative setups underlying also how they are related.

2.2 Different scenarios: from administrative records to statistical units

Following the Zhang (2012) model, in the *Representation side* of the proposed framework, we can identify how well the records in the AD match with the target statistical units. Ideally, each record in the AD corresponds to one unit in the target population. In fact, due to the nature of administrative captures, often records in the AD source are events or transactions, then information for each statistical unit can be obtained only combining information of multiple records (this is of course linked to the *Measurement side* of the life cycle model) or considering the complexities of relationships between the records and units. The following figure 2.2-1 describes the possible alternative cases of relationships between the AD source content and the statistical units, representing cases of One-to-One, Many-to-One, One-to-Many or finally the Many-to-Many matching. Existence of identifiers (or key variables) is an important element of the setup of what we have called the data structure' representation side, in view of the fact that its absence would introduce higher probability of introducing errors in the linkage procedures or in general its quality would depend on quality of the linking variables. Note that the availability of identifiers, or more generally linking variables, is often constrained by confidentiality issues, i.e. by the legal framework that characterizes the accessibility scenario.

Figure 2.2-1
Possible relationships between objects and statistical units



As an example of a complex many to many matching situation, we mention the case in agricultural statistics where AD on farms may be available (a farm register or a register of requests for subsidies), whereas the target statistical units are rural households related to these administrative units. Other typical cases are transcription of events related to population of students or information on job positions; in these cases typically we would encounter a many to one matching. Moreover, these examples make evident that many different target statistical populations can be investigated starting from the AD set of records, i.e. the farms themselves or conversely the rural households; job

positions or employees or employers and AD scenario would obviously be of different informative complexity for different statistical objectives. Zhang (2012) envisages also more composite situations, when complex statistical units should be derived from AD objects, for example “legal entities” in the business field or “households” in the social field. Zhang (2011) proposes a unit error theory based on a matrix representation of linkage between simple and complex units. An important aspect of the source setup is given by coverage issues. First, coverage issue may be due to administrative target set that may be different from the population, being for example a subset, and furthermore, under and over coverage with respect to its own target may occur also because of delays in registration of events. Coverage errors can also be caused by errors in linkage of units in different sources. Finally, as it will be illustrated in the following paragraph, different sets of the population may be covered by a diverse amount of administrative information.

2.3 Different scenarios: from measurements to statistical concepts

Let us now consider the measurements recorded in the external source. First of all, we can consider the case where the administrative measurements can be used as an auxiliary variable with a strong relationship with the target concept but they be used to replace the target variables themselves. This setting represents the traditional situation and it would imply possibly an indirect usage of the source. Conversely, some of the recorded variables in the AD source can be considered as the variables of interest. However, even in this latter case, they may not reflect the accurately the statistical concepts. First, for the reason that there may be a difference due to dissimilar definitions or diversity in the classifications that are applied; therefore, mapping or derivation of new variables is necessary.

The need for derivation of new statistical variables is also related with the kind of relationships between objects and statistical units that we have described in the previous section. In fact adjoining would be needed to determine the values of the target variable(s) on the new derived complex units (see Zhang, 2012 for more details). Moreover, administrative variables may differ from the required statistical ideal measure due to the fact that the collection method itself produces a *measurement error*. Furthermore, when the same concepts are measured on different (AD) sources, one has to choose how to combine the possibly dissimilar measurements. In an integrated system of different AD sources, figure 2.3-1 represents a possible pattern of information.

Figure 2.3-1
Possible pattern of information

	Measurements						
	AD1	**AD2**			**AD3**		
		Y1	Y2	Y3	Y1	Y2	Y4	Y5
Units								

This scheme illustrates two typical issues that are encountered in a system of sources. First, as mentioned also in the previous section, coverage in terms of the statistical population, secondly, alternative measurements of the same variable.

Summarizing, in an ideal situation, the AD contains exactly the objects and the measures needed to gain information for statistical production. In this case the output quality is the data input quality. In practice, we will never encounter this ideal setup and a process will be carried out on the original data, similarly to what happen to collected data in a statistical survey. New errors may be included during the process necessary to obtain the integrated final statistical

data. These considerations are very important when assessing the usability of the source(s). This assessment should be done taking into account of the complete system of administrative sources.

3. Different scenarios for AD usage: a first overview of methods

In Wallgren and Wallgren (2013) an extensive overview of methods and operations needed in a register based framework is illustrated. Here, we briefly mention the main statistical methods. First of all, regarding the operations needed for identification of units, it is needed to proceed with matching and linkage: both to augment information of a base register but also to check duplicates in one single source. In this case, AD scenario would typically influence the kind of linkage method one can actually apply. In fact, deterministic record linkage is commonly applied whenever an identifier is available, otherwise, probabilistic record linkage methods (Fellegi and Sunter, 1969) can overcome its absence. In latter case, estimation should take into account of this process in the subsequent analysis. See Chambers (2009) for unbiased estimation when record linkage is applied and Di Consiglio and Tuoto (2014) for a sensitivity analysis of the effect of record linkage errors on linear and logistic regression analyses.

The complex framework in a multi sources AD context can be exploited also when designing a sample survey. Falorsi and Righi (2012) proposed optimal survey strategies in a model assisted approach, dealing with a joint use of survey and AD where the latter cover only population sub-sets. Regarding the measurements that can be obtained in a multi sources framework, as we have already mentioned, when we have a complex system of different sources with different degree of coverage of the statistical target population, some specific issues arise. First of all, when integrating several data sources, the data consistency becomes an essential aspect, because the integration increases the possible conflicts into the available information. Determination of statistical variables when measurements are present in more than one source may be achieved giving a hierarchy to the different sources, on the basis of a quality assessment of the variable in the chosen source. However, more recently, literature has focused on studying measurement errors of all the involved sources without setting a priori a given source (or survey) as reference source. In the context of questionnaire design, there is a well-established tradition of using linear structural equation models (SEM) to assess the measurement quality of survey variables. This approach has been extended to administrative context in Bakker (2012) and Scholtus and Bakker (2013). Pavlopoulos and Vermunt (2013) applied a complex latent class (an hidden markov chain) model to assess the measurement of a categorical variable. In the latter case, the model is also used to get a measure, not only to evaluate the validity of the observed variables in the sources for the target concept. Meijer et al. (2013) provided alternative predictors in a similar background for a continuous target variable, assuming that the administrative register measurement may be affected by errors due to the effect of mismatch between AD source and sample survey. Figure 2.3-1 illustrates cases of incomplete information in the AD source, in this case model relationship can be used: on one side, mass imputation can be applied; otherwise, weighting or re-weighting (on other available registers) can be applied to combine registers (with surveys), see Renssen et al. (2001). When AD are used as a source of auxiliary variable, standard statistical methods can then be applied in the different settings: model based estimators have been explored in many applications (see ESSnet, 2011 for application of AD in business surveys). A model assisted approach is pursued in Kim and Rao (2012) where a projection estimator is proposed to combine different sources. See also Luzi et al. (2014) for an application of project estimation in a real case.

References

Bakker, B. (2010) "Micro-integration: State of the Art." *Paper for the Joint UNECE/Eurostat Expert Group Meeting on Register-Based Censuses*, The Hague, The Netherlands

Bakker, B. (2012) "Estimating the validity of administrative variables" *StatisticaNeerlandica* **66**, 8–17

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

Chambers R. (2009) "Regression analysis of probability-linked data", *Official Statistics Research Series*, Vol. 4.

Di Consiglio L., TuotoT. (2014) "When adjusting for bias due to linkage errors: a sensitivity analysis" *Proceedings of the European Conference on Quality in Official Statistics (Q2014).* Vienna, 3-5 June

ESSnet on use of administrative and account data in business statistics (2011) Deliverables of WP3 andWP4, http://www.cros-portal.eu/content/admindata-sga-3

Falorsi P, Righi P (2012)"Optimal Survey Strategies in the Multivariate Multidomain Context WithMultiple Sources of Administrative Information Covering Different Population Subsets." In: Electronic proceedings, of the Seminar on New Frontiers for Statistical Data Collection,http://www.unece.org/stats/documents/2012.10.coll.html. UNECE.

Groves, R. M., F. J. Fowler Jr., M. Couper, J. M. Lepkowski, E. Singer and R. Tourrangeau (2004), *Survey methodology*, New York: Wiley

Kim J. K. and Rao J. N.K. (2012) "Combining data from two independent surveys: a model assisted approach." *Biometrika*, 99, 1, pp. 85-100

Luzi O., Guarnera U., Righi P. (2014). "The new multiple-source system for Italian Structural Business Statistics based on administrative and survey data". *Proceedings of the European Conference on Quality in Official Statistics (Q2014).* Vienna, 3-5 June

Meijer, E., Rohwedder, S., & Wansbeek, T. (2012) "Measurement error in earnings data: Using a mixture model approach to combine survey and register data." *Journal of Business & Economic Statistics*, 30, 191–201.

Pavlopoulos, D., and Vermunt, J.K. (2013). "Measuring temporary employment. Do survey or register data tell the truth?" Survey Methodology *in press*

Renssen, R.H., Kroese, A.H, and Willeboordse, A.J. (2001). "Aligning estimates by repeated weighting." Research paper, BPA-no 491-01-TMO, Statistics Netherlands, Heerlen.

Royce, D. (2013.) "A Survey of International Frameworks for the Statistical Use of Administrative Data." Administrative Data Secretariat. Statistics Canada.

Scholtus S. and Bakker B. (2013) "Estimating the Validity of Administrative and SurveyVariables by Means of Structural Equation Models", *Proceedings of NTTS 2013 available on line*

UNECE (2011) "Using Administrative and Secondary Sources for Official Statistics A Handbook of Principles and Practices"

Wallgren, A. and Wallgren, B. (2014), *Register based statistics: Administrative Data for Statistical Purposes*, New York: Wiley

Zhang, L.-C. (2011). "A Unit-Error Theory for Register-Based Household Statistics", Journal of Official Statistics, 27(3), 415–432.

Zhang L-C (2012) "Topics of statistical theory for register-based statistics and data integration" *Statistica Neerlandica* Vol. 66, nr. 1, pp. 41–63