

**Project Title: Methodologies for an Integrated Use of
Administrative Data in the Statistical Process**

Activity A: Mapping and Overview

Output A.1 :

Usage of Administrative Data Sources for Statistical Purposes

Introduction

The informative context for many National Statistics Institutes (NSIs) has dramatically changed over the past years as administrative data sources (ADS) are becoming increasingly available. Advances in data linking methods and related computing capabilities are developing apace. This environment provides opportunity to make greater use of ADS in the production of official statistics, both replacing previous means of data collection and in the creation of new statistical products.

The MIAD project goal is to develop coherent and well founded strategies which would allow fully exploiting the use of ADSs in the production of official statistics.

The UNECE definition of an ADS is a “data holding containing information which is not primarily collected for statistical purposes”. The potential range of administrative sources that could be used for statistical purposes is large and growing consisting of both government and private sector sources. The scope in initial phases of the project will be restricted to administrative data coming from regulatory sources.

Mapping and Overview

This report constitutes the first output for the first phase of the project; ‘mapping and overview’. This phase of the project is considered a foundation for establishing a common setting for applications of ADS. The outcome would be applied in subsequent project phases, including the development of a framework to assess the quality of an ADS and its statistical utility (phase 2 of the project).

The report aims to provide an inventory of the ways in which ADS can be used in the production of official statistics, and to provide a mapping to the processes and sub-processes of the current version of the General Statistical Business Process Model (GSBPM).

The next task of this phase of the project is to outline a description of the possible informative contexts of ADS usage. This will be useful on one hand to guide possible methods to establish their quality, and on the other to help set up a schema of possible statistical methods for integrating of administrative in the statistical production.

Context

ADS are used extensively throughout NSIs in the production of official statistics in a wide array of applications. A scan of various NSIs shows a clear relationship between the usage of ADS and the different legislative environments in which NSIs operate.

A number of NSIs are supported by legislation that readily enables access to a variety of ADS which enables key statistical outputs being produced by direct tabulation (see below). Such NSIs, e.g. those in Nordic countries, are usually described as having a ‘register based’ statistical system.

ADS are often sourced from a registration process and consist of a complete set of records containing a number of attributes for each record with a regular updating mechanism. A statistical register is constructed and maintained for statistical purposes according to statistical concepts and definitions under the control of the NSI¹. A fully register based system typically involves a suite of registers built from ADS and a unique identification number that enables direct linking of records between registers. The conduct of surveys in register based systems is often to supplement or assist the ADS.

At the other end of the spectrum are NSIs that primarily adopt direct collection via a survey-based statistical system (e.g. Australia). Due to the legislative environment and other factors ADS are less readily available, often not at the statistical unit level of interest, and of variable quality. ADS in these situations can be used for direct tabulation for some statistics where quality is deemed sufficient but often without ability to form links between different sources. Traditionally ADS are used to supplement and/or assist surveys through creation of sampling frames and/or as auxiliary variables used in estimation, with ADS usage built around carefully designed surveys.

Various NSIs exist between these ends of the register based/collection based spectrum, and often within a NSI there may be various models as to how statistical registers and surveys are organised for different statistical themes.

¹Using Administrative and Secondary Sources for Official Statistics: A handbook of Principles and Practices”, UNECE, 2007

However, even within survey based organisations there is an increasing trend in attempts to make increased use of ADS, often through the creation of unit level databases for multi-purpose analyses, combining both ADS and other sources, which ostensibly are similar to a 'register based' model.

Direct and Indirect Usage

Whatever the prevailing statistical system, for this exercise we propose that ADS usage can effectively be distilled into two broad classifications: direct and indirect.

Direct usage will be defined as situations where there is an immediate link between the ADS and statistical output. The ADS may undergo various transformations, such as converting administrative units to statistical units (e.g. profiling businesses obtained from tax registers), or deriving statistical output variables from the unit's attributes, but in essence output is primarily sourced from the ADS itself.

Indirect usage of ADS describes situations where the ADS plays a supporting role in the creation of statistical output sourced primarily from either a survey or another ADS. Examples include the use of ADS to create a survey frame, or as population benchmarks used in weighting sample data.

The direct/indirect dichotomy should help frame the context for the inventory of ADS usage below and assist in establishing quality dimensions required. It will also guide the next phase of this project in summarising the different 'informative contexts' for ADS use, in particular the roles ADS can play in inference when integrated with survey data.

Uses for ADS

Following is summary of ways in which administrative data sources (ADS) are currently used by NSIs in the production of official statistics. The context assumes descriptive statistical outputs (e.g. counts, totals, percentiles) are the primary objective as opposed to analytical studies, and that outputs meet NSI's quality parameters as deemed fit for relevant 'official statistics' status.

We list usages of ADS as described by Lavallée², with one dimension "survey evaluation" replaced by "Data validation/confrontation". The usages described in Lavallée's paper are survey centric, the one exception of 'direct tabulation' where an ADS has full coverage of population of interest. Cases where partial coverage ADS are utilised in combination with surveys and/or other ADS data integration exercises could be considered advances that require new classifications, however it is proposed they can still fit within the broad classifications of usage described.

Direct Usage

1. Direct Tabulation

- typically for full coverage ADS
- simple computation of totals, means, percentiles etc., though could be partial coverage with weights or imputes attached

2. Substitution and Supplementation for Direct Collection

- whole and/or partial substitution for directly collected survey variables for sub-populations and/or variables of interest
- augmentation of directly collected survey variables
- can be considered in construction of ADS databases that do not cover either the whole population or all variables of interest
- *Data 'fusion'; integrating multiple sources data representing the same object to produce synthetic data that is more informative than original (indirect usage category)

² 'Administrative Data Usage in the framework of social statistics: Current and Future picture', Lavallée P., internal Stats Canada Document, 2007

Indirect Usage

3. Creation and maintenance of survey frames
 - Sampling frames
 - Statistical units
 - Auxiliary data (e.g. size and classificatory items)
4. Construction of sampling designs
 - Measures of variability for design variables
 - size measures to improve efficiency and/or targetting sub-populations of interest (stratification, pps etc)
 - facilitation of specialised selection mechanisms e.g. screening
5. Editing and imputation
 - construction of edit rules
 - auxiliary data to construct imputation models
 - for surveys and/or other ADS
6. Indirect estimation and weighting
 - creation of population benchmarks (independent of frame)
 - Improve efficiency of estimation through a model assisted or model based framework
 - Address quality issues e.g. non-response
 - 'harmonising' – creating new figures that agree between sources
 - Prediction – using an early available ADS to predict later more reliable estimates
7. Data validation/confrontation
 - validation of survey estimates and/or other ADS
 - micro or macro level
 - assess the quality of other potential ADS

Mapping to GSBPM

The table following maps the usages of ADS described above to each phase of the GSBPM. Whilst badged as a model for “general” statistical production, the processes largely reflect the conduct of surveys. As with the classification of usage above, cases where partial coverage ADS are utilised in combination with surveys and/or other ADS data integration exercises could be considered advances that require updated business processes. The mapping below is an attempt to fit the current model as best as possible.

For each phase of the GSBPM a number of considerations specific to ADS usage are noted along with a brief description of usage for each of the classifications above where relevant. Note the intention is to describe how an ADS can be used in each phase of the GSBPM, not how each phase of the GSBPM applies to a statistical exercise using ADS, hence the lack of entries for the disseminate, archive and evaluate phases.

Scope	AD-specific issues	Direct Usage		Indirect Usage				
		Direct Tabulation	Supplementation & substitution for Direct Collection	Creation and maintenance of frames	Construction of sampling designs	Data validation/ confrontation	Indirect estimation and weighting	Editing and Imputation
USER NEEDS								
Determine needs								
Consult								
Output objectives								
Identify concept								
Check data availability	Check legal framework, and if the client's needs can be met with ADS							
Prepare business case	Assessment of costs and benefits, as well as any external constraints, and the quality of the ADS. Determine the role ADS will play in production of statistics e.g. whether it will be direct or indirect usage							
DESIGN								
Output								
Variables	Study the potential alignment of ADS. Choice or definition of the statistical concepts and variables to be used	Determine if a survey is needed to augment or supplement ADS	Determine data gaps /opportunities that could be met via ADS substitution or supplementation					
Data collection	Study administrative data interfaces and possible data integration techniques. Both for direct use of ADS, or use of ADS in survey collection. This sub-process also includes the design of process-specific provider management systems. It is needed to identify rules for defining statistical units from administrative records.		Linking survey questions to ADS		Develop responsive design strategy where ADS can facilitate			
Frame and sampling				Use ADS to construct frame and/or register	Size measures to improve efficiency and/or targeting subpopulations of interest via stratification, or pps sampling			
Processing	Take into consideration of ADS in the process, including specification of routines for coding, editing, imputing, estimating, integrating, validating and finalising data sets of ADS data.							
Workflow	Take into consideration of ADS in the process, define formats and							

	timetable for acquiring ADS.							
BUILD								
Data collection	Collection instruments may also be data extraction routines used to gather data from existing statistical or administrative data sets. Direct use of ADS or ADS may be introduced in the data collection mode for either controlling survey data or assisting it when capturing survey information (additional to the current description of the GSBPM)	Use ADS in place of collecting data	Collect ADS components					Basic edit rules for quality checks
Process components	The use of ADS requires to build process components that, similarly to a census, requires management of large data sets and the tuning of complex procedures for data linkage							
Configure workflow								
Test prod system	In GSBPM this relates to software. However it is necessary to test 1) Success and quality of linkage 2) Building and identification of statistical units		Create synthetic dataset to test instrument		Use ADS as a dummy to test instrument			
Test statistical business process	In the current version this phase relates to the field activities, more in general with AD it is needed to test the channels and schedule							
Finalize								
COLLECT								
Select sample	AD source is a key element to establish and update the frame and to assist sampling as auxiliary variable.	Draw sample from ADS if necessary	Use ADS to determine what variables need to be collected, and which already exist in ADS					
Set up collection	Specifically for ADS, configure collection systems to request and receive the data; ensure the security of data to be collected							
Run collection	Contact provider for data access. Run very basic quality checks on the data				Use ADS to help target follow up			
Finalize collection		Direct tabulation of ADS	Use ADS to supplement collection					
PROCESS								
Integrate data	Derive statistical units from ADS records as a preliminary steps for integrate data (5.5)	Integrate different ADS to produce outputs	Whole and/or partial substitution for directly collected				Harmonising - creating new figures that	

	Integrate ADS sources / ADS and surveys. Several types of integration to consider - Data pooling (to increase the effective number of observations of a phenomena), data linking (extend the amount of information known about each unit), statistical matching (like data linking but assuming conditional independence), data fusion (integration followed by reduction or replacement), and micro integration (bringing together data sources at unit level) to name a few main ones"		survey variables				agree between sources	
Classify and code								Use ADS to come up with classification rules
Review validate and edit						Use ADS to validate survey outputs. Use ADS to determine consistency of data		Use ADS in the construction of edit rules
Impute								Construct imputation models (implicit + explicit) using auxiliary data
Derive new statistical unit and variables		Derive new variables directly from ADS	Derive new variables from auxiliary data					
Calculate weights							Create and use population benchmarks, auxiliary info for estimators	
Calculate aggregates		Direct tabulation of ADS						
Finalize data files		Integrate different ADS						Ensure cleanliness of data
ANALYSE								
Prepare draft outputs								
Validate outputs	For direct use of ADS, a particular					Use ADS to		

	concern is needed for assessing quality When regulations change the ADS content.					validate other ADS, or survey estimates		
Scrutinize & Explain							Address quality issues (e.g. non-response)	
Apply disclosure control								
Finalize outputs								
DISSEMINATE								
Update out system	Create output databases							
Produce dissemination products								
Manage release								
Promote								
Manage user support								
ARCHIVE								
Define rules								
Manage repository	Determine legality of keeping ADS indefinitely							
Preserve	Legal frameworks - what is the agency allowed to keep							
Dispose								
EVALUATE								
Gather								
Conduct								
Agree								