Data sources for statistical business registers

V. Todorov[1]

[1]United Nations Industrial Development Organization, Vienna

Regional Workshop on the Statistical Business registers for the Arab Countries
26-29 September 2016
Amman, Jordan

Outline

Outline

Introduction

Three main categories of data sources:

1. Administrative sources,
2. Statistical sources (like feedback from economic surveys, profiling and SBR improvement surveys),
3. Other sources (for example data from private data suppliers, telephone directories and the Internet),
4. New emerging sources (big data, web scrapping).

Recommendation: Create and maintain the SBR primarily using administrative sources.

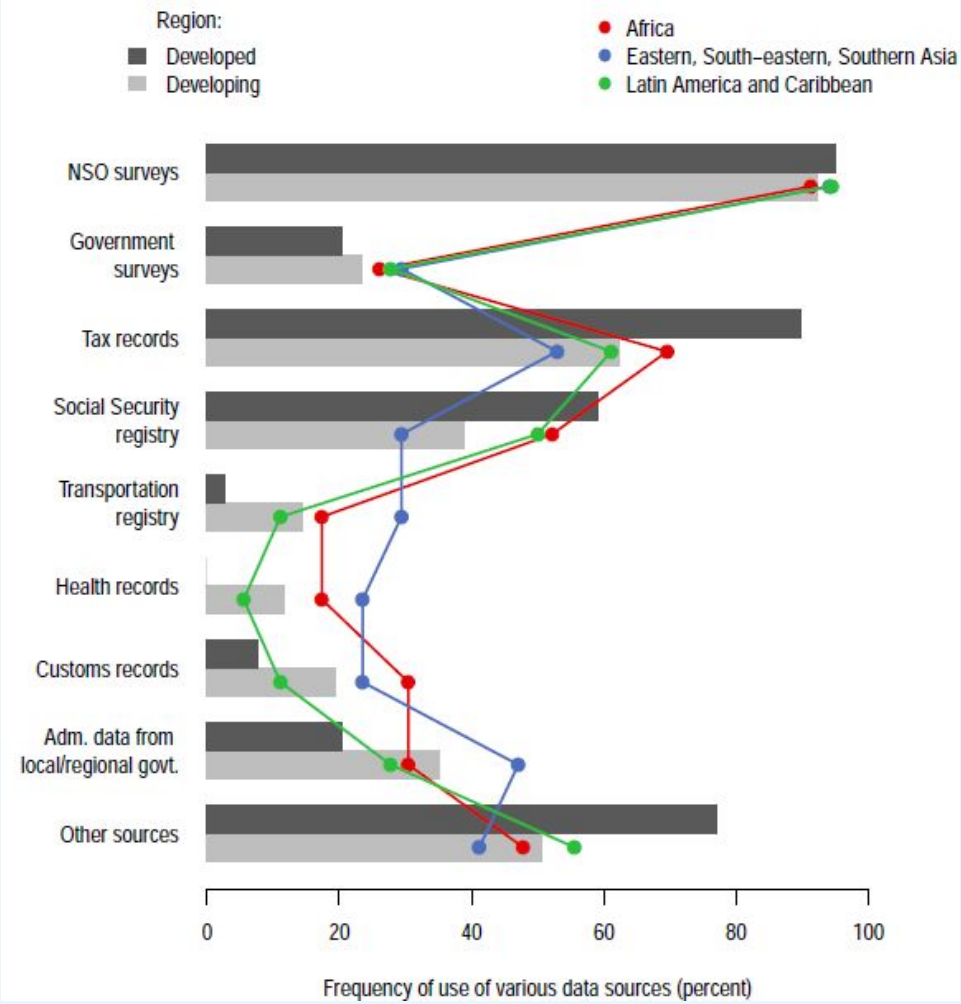
First of all review, identify potential sources and select the primary data source.

Introduction

Recommendation: Create and maintain the SBR primarily using administrative sources.

- In line with Principle 5 of the United Nations Fundamental Principles of Official Statistics:
  - "Data for statistical purposes may be drawn from all types of sources, be they statistical surveys or administrative records. Statistical agencies are to choose the source with regard to quality, timeliness, costs and the burden on respondent".
- The best choice depends on the specific situation in any given country (availability and accessability of administrative data; the scope and complexity of the national statistical system itself).

Introduction



Source: UNSD (2014)

Outline

General methods, procedures and issues

- Identifying and using external data sources
- Common identification systems
- Cooperation with data providers
- Combined administrative and statistical register
- Linking and matching
- Integration of the administrative units into the SBR
- Transforming administrative units to statistical units

Identifying and using external data sources

- First step: identify the most useful data source(s).
- Evaluate potential sources in terms of their
  - coverage,
  - content and
  - costs to the SBR of acquiring the data.
- Important to obtain all available information about each source (the definitions that are used, the coverage, the updating methods, the frequency of updates, the time lag, and how frequently it is possible to get information from the source).
- Data obtained from external sources should be stored in the SBR without change of content.
- Storing metadata about changes in data values is essential.

Examples

Examples

- Statistics Canada uses as its primary source the register of taxpayers with business income, which is maintained by the Revenue Canada.
- The Australian Bureau of Statistics (ABS) uses the Goods and Services Tax (GST) register maintained by the Australian Tax Office.
- Statistics South Africa uses the register of VAT account holders from the South African Revenue Service.

Common identification systems

Common identification systems

- A Business Number or ID number or any other identifier which is unique in the country is the ideal case.
- A unique number makes the dealings with the public sector simpler, easier, and more convenient.
- A Business Number, can be used across the country to commonly identify the business with public sector programs and services.
- Difficult to achieve but some countries have it.

Example: Canada Business Number

`http://www.bcbusinessregistry.ca/business-number.htm`

Common identification systems

Common identification systems

- Facilitates the work in combining data from administrative sources
- NSI should use any opportunity to promote it and to draw attention to the advantages of linking data in terms of more accurate and less costly statistics
- If no common ID system exists, NSI should construct an internal linkage table (containing links between units in the various sources) in order to manage the SBR and to reduce duplication or omission problems.

Cooperation with data providers

- It is vital to establish and maintain good relations with the owners of data sources, especially administrative sources.
- But can be complex and challenging.
- The purposes of cooperation are
  - to understand the concepts,
  - to ensure continuity of supply, and
  - to ensure easy linkage of data from the various sources to the SBR, preferably through a common identifier.
- $\Rightarrow$ Promote the interests of the SBR and, more generally, of the economic statistics program.
- The purposes are the same whether the SBR is dealing with an administrative, statistical or commercial data source.

Cooperation with data providers

- Cooperation includes:
  - Regular face-to-face contact and mutual visits.
  - Should involve both management level staff and operational staff who are working with the data on a daily basis
- It is good practice to set up a memorandum of understanding (MOU) or service level agreement (SLA) with an administrative source.
- MOU or SLA provides an agreed framework and gives reassurance about the services to be received.

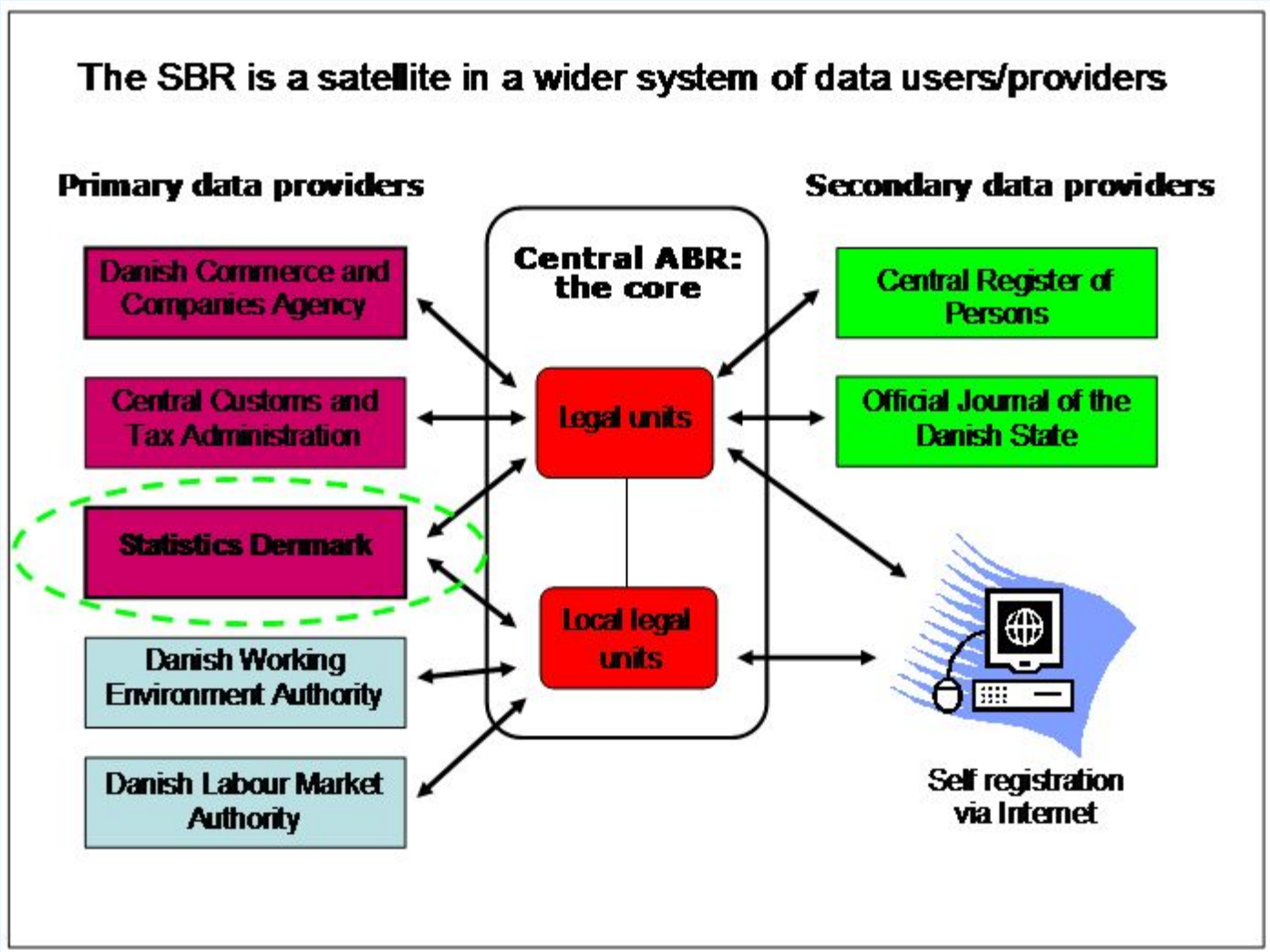
Cooperation with data providers

- Typically an SLA sets out:
  - the agreed data coverage and content to be supplied,
  - delivery timetables,
  - data security provisions,
  - checking mechanisms and quality provisions, and
  - response times within which to answer validation questions by the SBR.

Combined administrative and statistical register

- Some countries have developed a single administrative business register with multiple functions, including that of an SBR.
- This has significant advantages in that statistical data requirements are within the core of the administrative systems.
- The Danish SBR provides a good example (described in the Guidelines, Annex C1).
- Other countries have similar systems.

Combined administrative and statistical register



Source: Statistics Denmark

Linking and matching

- There are significant advantages of using data from several sources.
  - Lower risk of error, if multiple reliable data sources are used.
  - For example, VAT, provides extensive (but never complete) coverage of the economy,
  - Others may only cover one sector, e.g. financial, but more completely.

Linking and matching

- However, using two or more data sources presents two types of linking and matching challenges:
  1. Avoiding duplication without incurring omission (when the sources have a similar coverage of the economy)
  2. Dealing with the sector boundaries associated with each source, if each source covers only a limited sector of the economy, such as finance and public administration.

Linking and matching (cont)

- Probabilistic approach to linking—if there is no common identification code: based on similarities in the name and address or other characteristics, such as legal form and economic activity code.
- This process is generally referred to as matching.
- However, it can lead to units being linked in error (mismatches), as well as units not being linked (which result in duplication).
- Linking large data sets in the absence of a common identification code is difficult and requires substantial investment in software and systems.
- Matching will be discussed later in more detail.

Integration of the administrative units into the SBR

- After administrative data have been acquired, the next step is to match them with the administrative units that are already covered in the SBR.
- This is relatively easy if a common identification number is available, and there are no errors in this number in either the administrative sources or the SBR.
- The matching results should be quality checked: comparison of the values of characteristics such as economic activity code, size or legal form across the linked units.
- Maybe further clerical checks needed, particularly where larger units are concerned.

Integration of the administrative units into the SBR

- If additional data, such as turnover and imports/exports, are available from other administrative sources, these can also be used in the checking procedure.
- Useful to periodically check administrative units that have not been matched and attempt to establish further links or to determine why they do not match.
- If the non-matched units do represent active economic units: $\Rightarrow$ the failure to match may be due to timing or scope differences between the administrative source and the SBR.

Transforming administrative units to statistical units

- The final step is to build the statistical units (mainly the enterprise) on the basis of the administrative/legal units.
- In the normal case an enterprise coincides with a legal unit.
- This is the case when the legal unit is not controlled by another legal unit and thus has autonomy.
- Most enterprises belong to this category.

Transforming administrative units to statistical units

- Legal units that belong to an enterprise group may not necessarily be considered as an enterprise and may need to be combined with another legal unit of that group to form an enterprise.
- Such case are normally quite complex to handle.
- The whole task of delineating of the statistical units based on legal units is called t't'profiling" and will be further described later in this Chapter.

Outline

Administrative data sources: Definitions

Definition 1: Administrative sources

"Administrative sources are sources containing information that is not primarily collected for statistical purposes".

Definition 2: Administrative data

"data derived from an administrative source, before any processing or validation by the NSI".

Definition 3: more traditional and narrower

requires administrative data to be collected by government bodies for the purpose of administering taxes, public pension funds and other regulations.

Administrative data sources: Types

- Business registration/license register
  - can provide basic information on ID, name, address and other contact data.
  - may be run by the Tax authority, Chamber of Commerce, licensing office, or another public authority.
- Tax register
  - typically relating to VAT or employee income tax.
  - may be a source of data on economic activity, turnover, and activity status.
- Company/trade associations and chambers of commerce registers
  - can provide information on economic activity, legal form, and births and deaths.

Administrative data sources: Types

- Social security registers.
  - for businesses employing paid staff and making social contributions for employees
  - can provide identifying characteristics and stratification characteristics, such as legal form and number of employees.
- Labour and employment registers
  - can provide additional economic and social information about employees.
- Government units registers
  - maintained by government finance departments for financial management of the public sector.

Administrative data sources: Types

- Industry association registers.
  - may contain name, address, other contact information and economic activity code.
  - are likely to be up to date, but only contain members of the association, so completeness may be an issue.
- Agricultural administrative registers.
  - may cover agricultural holdings as distinct from businesses.
  - typically contain name, address, other contact information and indicators of economic activity and possibly size.
- Water supply and electric association registers.
  - maintained by public or private utility bodies.
  - typically contain name, address, other contact information and indicators of economic activity and possibly size.

Administrative data sources: Types

- Sector specific sources.
  - include lists of schools from the education ministry, lists of hospitals from the health ministry, and lists of charities from regulators.
  - coverage is limited to a specific sector, however within that sector it can be very comprehensive.
- Central banks.
  - often have information for the financial sector, and on units engaged in foreign direct investments, from supervisory authorities.

Administrative data sources: Types

- Published business accounts.
  - particularly valuable as they contain information on shareholders and subsidiaries that is essential in delineating enterprise groups.
  - explore the possibilities of automatically extracting data from internal financial or management accounting systems of businesses, for example, using XBRL.
  - provide names and contact information, but also data that indicate whether or not a unit is active, its principal economic activity, its size and some other variables relevant for the SBR, such as employment and turnover.

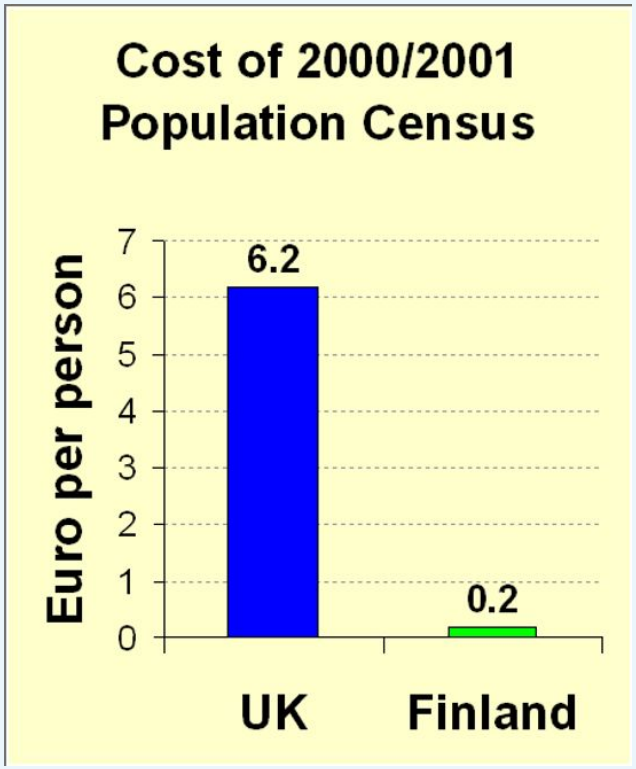
Advantages of using administrative data

- Coverage
  - Coverage is of great significance from an SBR viewpoint, given its aim of including all economically active units.
  - Sample surveys often cover only a relatively small proportion directly.
  - Administrative data give complete or almost complete coverage of the target population.
  - Eliminates sampling error.
  - Removes or significantly reduces non-response.
  - Provides more accurate and detailed estimates for various sub-populations, e.g. small geographic areas.

Advantages of using administrative data

- Cost
  - Surveys are expensive, particularly if they are conducted as censuses or involve the use of personal interviews.
  - Administrative data are often available free of charge, or for the marginal cost of extraction, particularly if they originate from the public sector.
  - However, "free of charge" does not mean "for free": set-up costs of using administrative sources can easily be as high as the set-up cost of a statistical survey.
  - But the running costs are significantly lower.

Example: Cost benefit of using administrative data



Source: Eurostat—Documentation of the 2000 round of population and Housing censuses in the EU, EFTA and Candidate Countries

Advantages of using administrative data

- Response burden
  - Using data from administrative sources involves no additional response burden.
  - The use of administrative data may, in some cases, allow statistics to be produced more frequently, with no extra cost to businesses.

Advantages of using administrative data

- Content and timeliness
  - Administrative sources may increase the quality of the SBR by providing access to more upto-date information on key characteristics, such as:
    - Name and address
    - Births and deaths of units, and the dates of these events
    - Economic activity code
    - Location
    - Size, in terms of number of employees and/or turnover
  - Positive impact of administrative data on management of Statistical registers and survey frames
  - Changes to the target population (births and deaths) are more up to date than survey information could ever be
  - However, there some cases where using administrative data can worsen the timeliness (e.g. short term indicators).

Disadvantages of using administrative data

- Administrative definitions of units and variables can deviate from statistical needs and definitions
  - Administrative data are collected according to administrative concepts and definitions
  - Administrative and statistical priorities are often different, so definitions are often different
  - Legal entities or some breakdown of these entities suitable for administrative purposes vs. statistical units (enterprises and establishments)
  - Profiling: the process of converting from administrative units to statistical units—typically is a function of the statistical business register (see EUROSTAT: Business Register Recommendation Manual, chapter 19)

Disadvantages of using administrative data

Example 1

- Statistical definition (ILO) of Unemployment
  - Out of work or
  - Available for work or
  - Actively seeking work.
- Administrative definitions of unemployment are often based on those claiming unemployment benefits

Example 2

- Turnover for VAT purposes may not include turnover related to the sales of VAT exempt goods and services, whereas the statistical system is likely to require the total turnover

Disadvantages of using administrative data

- Are known to contain inactive units
- Use of different classifications (and different use of the same classifications): conversion tables needed for different classifications
  - are not classified by economic activity or
  - use older revision of the activity classification or
  - do not provide the detail required for statistical purposes.
- Potentially less timely data
  - Data may arrive too late
  - Data may relate to a different reference period: e.g. tax year may not coincide with calendar year, necessary for industrial statistics

Disadvantages of using administrative data

- Public sources of administrative information are generally set up for the purpose of tax collection and monitoring government policies $\Rightarrow$ they are susceptible to political changes.
- Changes to administrative regulations or procedures
- Data from multiple sources: matching/linking issues, data conflicts. It is necessary to establish priority rules.

Quality of administrative data and its monitoring

- There are many aspects to quality:
  - Timeliness
  - Coverage
  - Completeness
  - Accuracy
- Administrative data will be better than survey data in some aspects but not in others.
- It is important to look at overall quality.
- Do the data meet the needs of users?

Quality of administrative data and its monitoring

- Three aspects of quality
  - Quality of incoming data
  - Quality of processing (matching, merging, ...)
  - Quality of outputs—likely to be different to survey based outputs, but are they better?

Quality of administrative data and its monitoring

- Quality measurement
  - Comparing sources
  - Quality check surveys
  - Knowledge of source (metadata)
  - Quality reports/templates

Quality of administrative data and its monitoring

- Create knowledge of the administrative sources
  - Primary purpose and the way the data are collected and processed.
  - Allows a more accurate assessment of its strengths and weaknesses.
  - To help develop and document this knowledge $\Rightarrow$ template to record information from the source on contacts, units, characteristics, quality and formats.

Quality of administrative data and its monitoring

- Quality template: Companies house data

Framework	Contract
Frequency	Quarterly updates, continuous on-line access
Timeliness	Good
Quality	Good
Delivery	CD-Rom, Internet
Key content	Legal name, Company number

Quality of administrative data and its monitoring

- Quality indicators (some examples)
  - The number and proportion of enterprises lacking a valid and complete economic activity code.
  - The number and proportion of enterprises for which the activity status (active, dormant, dead, etc.) is unknown.
  - The number and proportion of enterprises lacking a complete address.
  - Number of units with company name (address or postal code or telephone number) missing.
- Such indicators may be compiled using feedback from surveys based on the SBR and/or from SBR improvement surveys.

Quality of administrative data and its monitoring

- Dealing with conflicts
  - Administrative source and the SBR do not agree: Surveys may be used to investigate such discrepancies
  - Conflict in data from different administrative sources
    - for example conflicting industry codes.
  - Procedures and rules need to be developed to resolve these problems.
    - verify the data by contacting the enterprise, or
    - undertaking analytical work to determine which source is most reliable.
  - The goal is a set of general rules to deal with conflicts.

Legal issues

- Access guaranteed through a statistics act
  - First step in use of administrative data: ensure the NSI has access to the data.
  - There are two aspects to gaining access:
    1. legal framework and
    2. setting up and implementing the procedures for transfer of the data.
  - The preferred approach: NSI's right of access to administrative data to be enshrined in a general statistics act.

Legal issues

- Formal agreements
  - Whether or not the NSI has legislated access, the NSI should try to establish some form of formal agreement with administrative data providers.
  - Legally binding contract with a private sector supplier, or a service level agreement (SLA) or a memorandum of understanding (MOU) within a public sector provider.
  - Should describe the rights and responsibilities of both parties, delivery flows, data confidentiality constraints, quality standards, frequency and format of data transfer, time frames for responding to queries and questions about the data, and procedures to follow in case of disputes.

Legal issues

- Building relationships
  - In addition to formal arrangements, good working relationships with administrative data providers should be developed.
  - These can be achieved through regular contact, preferably face-to-face.
  - It is advisable that the NSI should be coordinated when it approaches administrative sources.

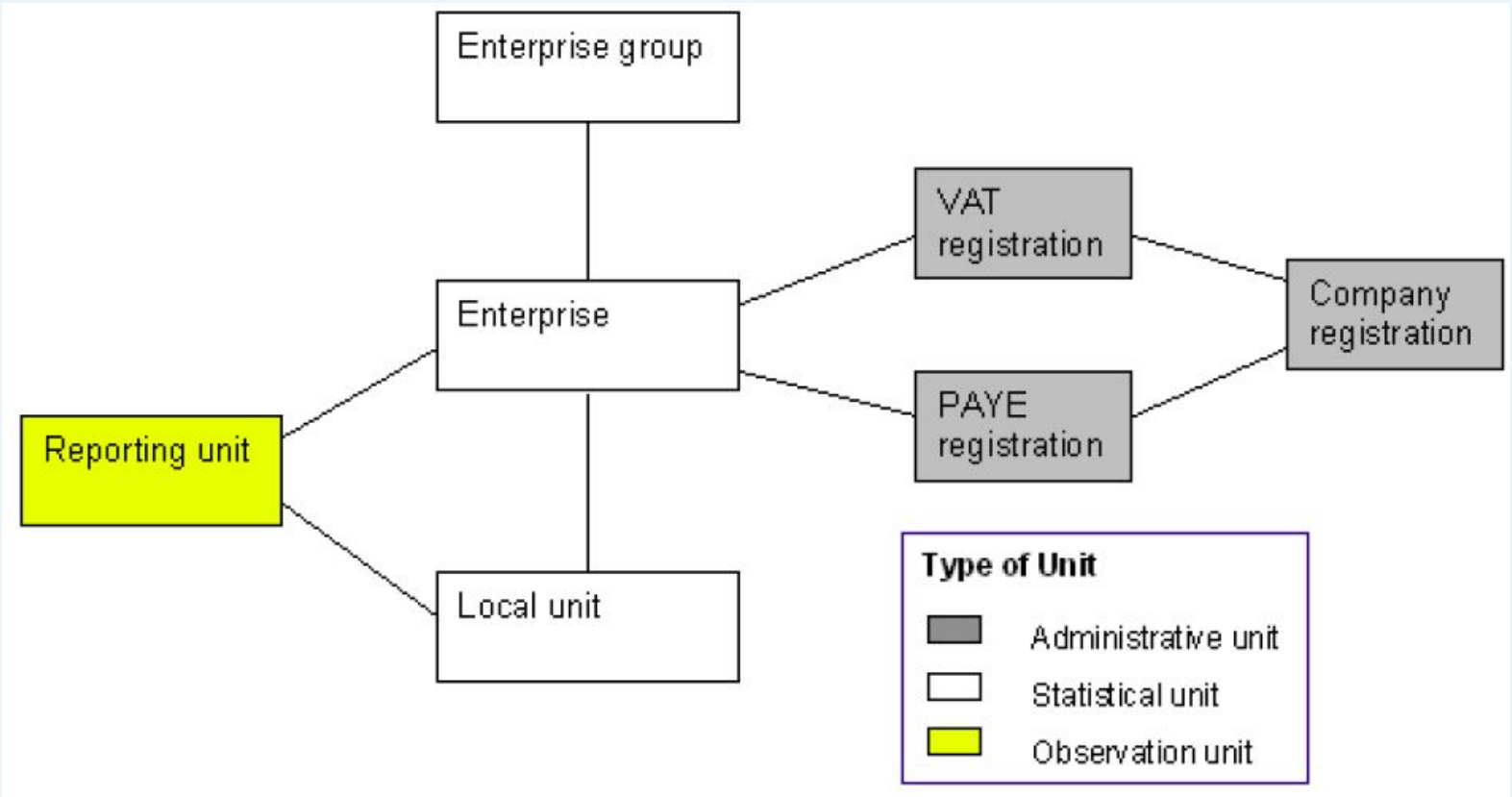
Outline

Practical guidelines for using administrative data

- Steps to be taken when creating and maintaining a statistical business register
  - when using primarily administrative data.
  - exactly what sources are used depends on data availability and accessability.
  - however, there are procedures and basic principles which apply always, whatever are the sources.
- Basic principles:
  - Keep administrative and statistical data separate
  - Establish unique identifiers
  - Use of sector specific and additional sources
  - Thresholds
  - Handling changes in administrative data

Keep administrative and statistical data separate

Statistical and Administrative units: Example from the UK's *Inter-Departmental Business Register (IDBR).*



Source: UK Business and Business Demography publications

Keep administrative and statistical data separate

- Data from administrative sources (administrative units) should be stored separately and not mixed with data for statistical units.
- Even in the case of combined administrative and statistical registers it is important to maintain clear procedures on data sources and updates to enable data received from administrative records and any statistical transformations to be auditable.

Keep administrative and statistical data separate

- Easier to discuss unusual/anomalous SBR data either with the enterprises themselves or with the administrative sources directly.
- Concrete examples to be provided if investigation is required
- Another possible reason: administrative data are being used for other statistical purposes. For example social security data may not only be used for SBR purposes, but also provide employment data for a variety of statistics.

Establish unique identifier

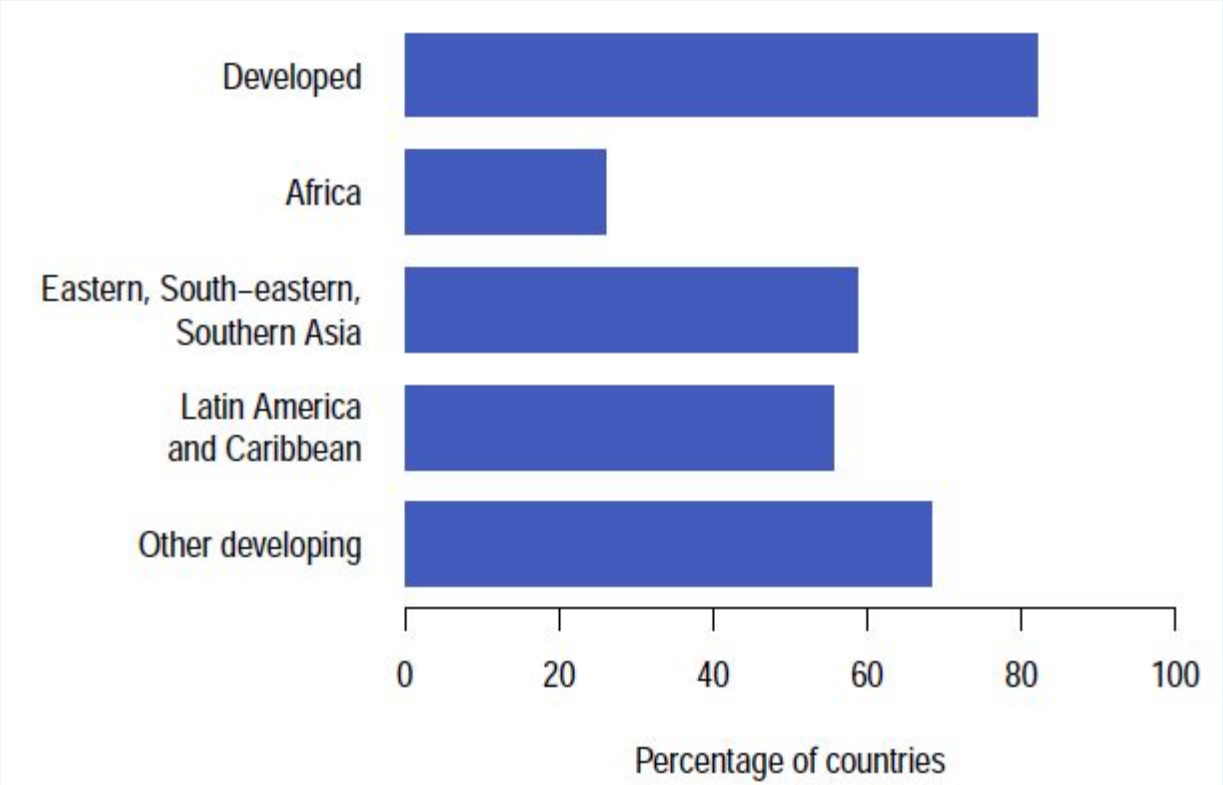
- Establishment of unique identifiers is essential for the accurate maintenance of the SBR.
- If administrative numbers are used for statistical units: added complexity in SBR systems; introduce risks such as duplication and omission of statistical units.
- Identification numbering system should be created covering each type of statistical unit.
- In case of multiple admin sources with different identification numbers linking and matching is necessary: significant resources.

Establish unique identifier

- If the NSI has influence on the development of administrative data, it should promote use of a unique business number for all relevant administrative processes in the country.
- This is difficult to achieve, but some countries have it.

Establish unique identifier

Countries that can cross-link the ID for statistical units with external ID used in other sources.



Source: UNSD (2014)

Use sector specific and additional sources

- Some trade associations, charity regulators and government ministries such as education and health may maintain data that are useful in providing coverage of certain sectors of the economy and/or additional content.
- Care should be taken to ensure that coverage of the particular sector is complete, or if not, the gaps are understood.
- Using multiple sources leads to challenges in maintenance and management.
- A source may be used: (a) to verify data from another source; or (b) a source may contribute directly to updating of statistical units.

Use sector specific and additional sources

Example 1:

The characteristic sales space may be available from an industry association for retail trade enterprises.

Example 2:

The characteristics category/number of stars and number of beds for hotels may be available from a tourism authority or bureau.

Example 3:

Import/export data may be available from the customs authority.

Thresholds

- Different thresholds in the regulations can have impact on the coverage of administrative sources
- For example an enterprise may not be required to register for VAT if its turnover is below a certain threshold.
  - If the threshold is low then the enterprises that are omitted because they are below the threshold have only a small impact on the overall estimates and their omission from the SBR is not an issue.
  - If the threshold is quite high then its impact on coverage and estimates is likely to be large.
  - $\Rightarrow$ Additional sources and methods may be needed to supplement coverage provided by the source.

Handling changes in administrative sources

- Changes in administrative sources can lead to discontinuity in the data supply:
  - changes in thresholds,
  - changes in definitions or
  - changes in computer systems.
- Work very closely with the administrative source to gain a full understanding of the change.
- Assess the impact on the SBR
- Work actively with the survey staff who are the most important SBR users to assess the impact on their outputs.
- Changes in tax thresholds are common: minor or radical.

Outline

Identifying enterprizes

- In most countries with administrative registers, a legal unit of some sort is defined for SBR purposes.
- Often the legal unit is exactly, or a close approximation to, an enterprise.
- It is necessary, at least for the bigger units, especially for those being part of an enterprise group, to be able to create an enterprise in accordance with the enterprise definition, which does not require it to be in one to one correspondence with a legal unit.
- The transformation from legal units to statistical units must follow well defined basic rules.

Identifying local units and/or establishments

- SBR should contain either local units or establishments.
- Local units/establishments enable regional economic analysis to be conducted.
- Most administrative sources (e.g. VAT) refer to a legal unit as a whole or a specific part of it, but not necessarily at local unit/establishment level.
- $\Rightarrow$ they cannot be used to populate local units/establishments in the SBR.
- Creating local units/establishment is a bigger problem than creating enterprises.

Identifying local units and/or establishments

Example 1

Employment tax systems, or social security databases, may hold information on the location of the employees—this information could be used to construct a local unit.

Identifying local units and/or establishments

- In most countries there may be no suitable administrative data source $\Rightarrow$ local units/establishments have to be established through a special SBR survey.
- Stratify such a survey to ensure that all larger businesses are included whereas smaller businesses are sampled very lightly simply to estimate the probability of smaller businesses owning more than one site.
- The resulting data can be used to prepare a typical model for small businesses.
- To reduce burden and expense it may be possible to obtain the same information through an annual employment survey.

Converting from legal units to statistical units

1. All administrative data at the level of individual enterprises (or local units) should be fed through, but not necessarily stored in, the SBR.
2. Use the administrative units as observation units and through the links in SBR transform the administrative data into statistical data at the level of each individual enterprise or establishment.

Converting from legal units to statistical units

- Check the quality and coverage of the incoming administrative data to ensure some basic conditions are satisfied:
  - The file is the expected size
  - The values of the characteristics are in valid formats and/or ranges, for example, dates are within a permissible range, text fields contain only text characters, numeric fields contain only numbers, and codes used are valid.
  - There is good coverage of the main characteristics, for example, identity numbers, addresses and economic activity codes are present for all units.

Transformation rules

- Transformation of administrative unit data into statistical unit data may involve several steps:
  - Creation of the corresponding statistical units (typically enterprise groups, enterprises, and establishments),
  - Derivation of their characteristics: algorithms and/or look-up tables.

Example: Transformation of economic activity

Economic activity codes that are peculiar to a particular administrative source are converted to the standard codes used in the SBR (ideally ISIC Rev. 4, or equivalent).

Transformation rules

- Most enterprises have a one to one relationship with the corresponding administrative units, but can be much more complex.
- The resulting transformation rules are dependent on the circumstances in each country.
- Important to determine whether it is necessary to put in place an intermediate step: creating legal units to which both administrative units and statistical units are linked.

Online delivery

- In an ideal situation the SBR is updated with information on line, unit by unit.
- In this situation there is no validation of a file as such.
- Individual units and characteristics still have to be validated.
- The same rules for creating and updating statistical units are still appropriate.

Outline

Statistical sources

- Economic census
- Feedback from establishment survey
- SBR improvement survey survey
- Profiling

Economic census

- Traditional economic census: trained field enumerators seek out each physically recognizable place of business and collect the necessary information by direct interview and observation.
- Used in many developing countries and some developed countries, including the USA.
- Very useful instrument when a country is initiating an economic statistics programme. It provides benchmark data.

Economic census

- In the past, especially in developing countries, it was a well-established method for the initial construction of an SBR. Has numerous drawbacks:
  - Very resource intensive exercise: requires large inputs of manpower and time.
  - Tend to be carried out infrequently, for example, once every five years.
  - Intercensal updating of the SBR is thus required, which is itself costly.
  - Not being able to identify and document non-recognizable places of business, or enterprises without a fixed location, for example, web-based businesses, or individual entrepreneurs.

Economic census

- In summary, economic censuses are not recommended as a means of establishing an SBR.
- On the opposite: an economic census should draw its basic frame from the SBR, possibly supplementing this by an area sample.
- However, if there is no reliable administrative source whatever, a periodic economic census is appropriate.

Feedback from establishment survey

- A vital mechanism for updating the SBR.
- It provides information on changes in contact address, changes in the economic stratification characteristics, deaths, etc.
- Has the advantage that it is available at statistical unit level, that is, for establishments or enterprises.
- Therefore close contact between survey staff and SBR staff required.

Feedback from establishment survey

- Limitations of survey feedback:
  - It would lack new units as surveys are not designed to find births.
  - The population of enterprises would not be fully maintained as feedback would be coming only from the sampled units.
  - Even for the sampled units, use of survey feedback from sample surveys introduces the possibility of feedback bias as SBR updates are provided only for the selected enterprises.

Feedback from establishment survey

Example: Feedback bias

- Suppose that when a particular quarterly survey is first conducted, the sample is found to contain 30% dead enterprises (this is not an improbable figure).
- Based on this sample information, the dead enterprises are removed from the SBR; the next survey sample comprises the 70% live units from the previous sample plus a replacement of the 30% drawn afresh from the SBR.
- This new sample contains about 9% (30% of 30%) dead units.
- The bias will increasingly worsen with each survey repetition.

SBR improvement survey survey

- Specific SBR improvement surveys conducted by SBR staff: to obtain SBR updating information that cannot be obtained from surveys, or from the administrative sources on which the SBR is based.
- Also termed natureof-business surveys or proving surveys or SBR control surveys.
- It is usually necessary to focus improvement surveys on specific strata to measure and improve coverage and quality.

SBR improvement survey survey

Possible strategy for keeping the SBR up to date

- Conduct SBR sample surveys every year in which the biggest, the medium sized and the smallest units are sampled 100%, 50%, and 10% respectively.
  - $\Rightarrow$ Keep values of the characteristics of the units up to date in an efficient way.
  - Such surveys may also be specially designed to measure SBR accuracy:
  - $\rightarrow$ measure errors in classification by economic activity, or by size, or to estimate the proportion of falsely active units.

Profiling

Profiling

The practice of using company accounts, often accompanied by interviews with senior enterprise officials, to build and define the structure of enterprises, mainly those involved in large complex enterprise groups.

Profiling is not a primary source of data. However it does provide valuable information on the larger and more complex enterprises that individually make a significant contribution to the country's GDP. It is especially important in identifying enterprise groups.

Profiling

- The resulting profiles are used to produce a reporting structure appropriate for the surveys conducted by the NSI.
- Profiling usually involves establishing contact with the enterprise being profiled to develop a good understanding of its structure.
- It is possible, however, to complete smaller profiles simply using published accounts.

Reactive and Proactive Profiling

- Profiling is often organized so that each individual SBR profiler has an assignment of large, complex enterprise groups for which he/she is responsible for reviewing and updating.
- Reactive profiling: when profilers react to signals from various sources on enterprise that may need updating this.
- Proactive profiling: periodically performed to augment reactive profiling.

Profiling procedure

- Step 1: Determine the criteria by which to identify the enterprises to be profiled.
  - Usually focussed on large complex enterprise groups with multiple activities for which survey reporting is difficult.
  - After determining the profiling criteria, a programme of regular updates over, say, a three to four year period, should also be established.
    - A balance between the resources available and the amount of profiling that can be conducted, and this needs to be considered when determining the criteria.
    - Resources should also be planned to deal with emerging issues that occur outside the routine reviews, i.e., reactive profiling.

Profiling procedure

Example

A major merger might require a profile to be conducted ahead of the scheduled regular review.

Profiling procedure

- Step 2: Gather preparatory material for the enterprise group as currently defined.
  - Records on all legal units within the enterprise group,
  - The reporting history for surveys, and
  - SBR data such as employment, turnover and classification, etc.
  - Examine this data for consistency; try to identify reporting issues.
  - As this process is labour intensive, it is best
    - to create a standard template for the data required and
    - to automate the extraction of this information from the SBR.

Profiling procedure

- Step 2a: Gather further background information by searching enterprise websites and examining annual accounts and reports.
- In simpler cases it may be possible to conduct a profile simply on the basis of this material.
- For complex cases it is invariably necessary to meet with representatives from the enterprise.

Profiling procedure

- Step 3: After this preparation, contact is made with the controlling enterprise in the enterprise group and a visit arranged if necessary.
- Through discussion it will be possible to identify the main trading (i.e., active) enterprises within the enterprise group and to agree a mutually acceptable reporting structure covering these enterprises.
  - settle on a structure with enough detail for the statistical surveys
  - minimize the respondent burden for the enterprise group.

Profiling procedure

- Specially trained staff are required to negotiate with the enterprise as they will typically be talking to top management in the enterprise—the chief accountant and/or the company secretary.

Outline

Combining administrative and statistical sources

- A combination of administrative and statistical sources is recommended, in order to build a comprehensive SBR.
  - Administrative sources identify enterprises, but may not include all of the required characteristics.
  - Statistical sources do not identify new units but provide additional or more accurate characteristics.
- $\Rightarrow$ A strategy of using administrative and statistical sources in combination should be developed and employed.

Combining administrative and statistical sources: Examples

- Administrative sources can be used to identify legal units and transform their data to form enterprises, while local units or establishments can be identified by a survey of the enterprises, as can characteristics missing from administrative sources.
- Let us see this in action, on examples:

Combining administrative and statistical sources: Examples

Example 1

An enterprise is thought to have only one establishment but employee data from an administrative source indicate that half of the workforce lives in an area far away from the identified establishment.

This may be a signal that there is a second establishment, or that the employee data have been linked to the wrong enterprise.

Combining administrative and statistical sources : Examples

Example 2

Turnover from a VAT source does not correspond to the turnover for the same enterprise from an import/export administrative source.

This may be a signal that some part of the enterprise is missing, or the links to one or other of the administrative sources are wrong.

Combining administrative and statistical sources: Examples

Example 3

There is feedback from a survey indicating that the responding (legal) unit no longer has any economic activity.
This may be a signal that the activity has taken over by another legal unit.

Combining administrative and statistical sources: Examples

Example 4: Missing data estimation

An administrative source only contains employee numbers and industry code, it may be possible to estimate turnover using turnover/employee value for similar units.

Thus statistical sources can be used to estimate missing characteristics.

Outline

Introduction to Record linkage

- Usually, the main sources for the maintenance of the SBR are administrative data, such as data from taxation, social security or other administrative sources depending on their availability and adequacy.
- If these administrative sources have separate identification systems and if no identifier known to the SBR is provided, then record linkage methods have to be applied in order to be able to use the administrative data in the SBR.
- There are commercial and open source software products for record linkage.

Basic approach

- Record linkage is the linking of a data record in one data source A with one or more data records in another data source B.
- If two records for the same unit are brought together this is called a match.
- If no common identifier is available, identification of a match has to be based on a similarity measure that is computed using characteristics that are available both in A and in B.
- The characteristics used identify a unit uniquely (e.g. name and address, and may be legal form).

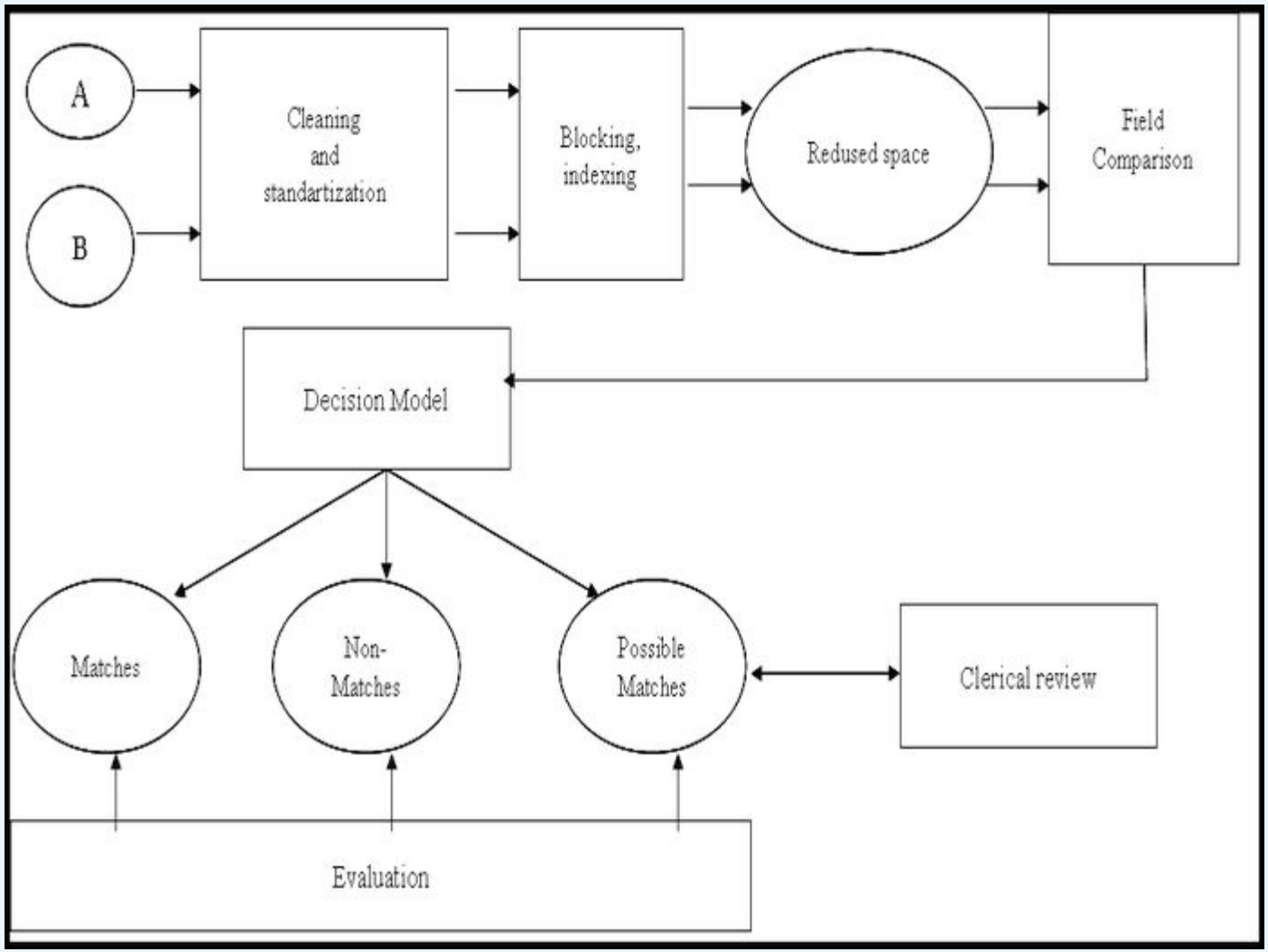
Basic approach

- However,
  - Errors are possible either in A or in B or in both.
  - It is possible that the data in B (administrative source) is not standardized, is in different format (parsing required).
  - Record linkage is computationally intensive. This can be drastically improved if blocking methods are applied.

Types of matching

- Two types of matching are possible: deterministic and probabilistic, based on the computation of numerical values expressing the similarity of a record pair.
- In case of deterministic matching, a similarity measure (or distance function) is defined.
- In probabilistic record linkage (Fellegi and Sunter) a so-called likelihood ratio is computed.
- The value obtained is compared with two thresholds: a lower threshold, below which all values are defined as non-match, and an upper threshold, above which all values are defined as match. The grey area between lower and upper thresholds defines potential matches.

Record linkage process



Preprocessing (cleaning and standardization)

- Noisy, incomplete and incorrectly formatted information
- Record linkage is highly sensitive to the quality of the data being linked—data quality assessment necessary
- Requires 75% of the total effort for implementing record linkage procedures (Gill, 2001)
- Delete null strings; convert upper/lower cases;
- Standardization: replacing various spelling of words with a single spelling (e.g. different spellings and abbreviations of "Incorporated" are replaced by "Inc.")
- Parsing of a field (e.g. name or address) into words

Matching of text strings

- The fields are considered as strings of alphanumeric characters and string comparator metrics are used to compare the two strings and to determine how much alike they are to each other.
- The values of the metrics lie in [0,1].
- There are many different possible string comparator metrics (Jaro, Winckler, Levenstein, N-grams, etc.).

Matching of text strings

Example: “Welcome” and “Welkome”

Comparing N-grams:

- The bigrams of the string “Welcome”: “_W”, “We”, “el”, “lc”, “co”, “om”, “me” and “e_”
- The bigrams of the string “Welkome”: “_W”, “We”, “el”, “lk”, “ko”, “om”, “me” and “e_”
- The 4th and 5th bigrams are different: thus the measure is 6/8=0.75.
- As an option, the first and last bigrams containing the blank can also be left out, giving a measure of 4/6=0.67.

Blocking

- Each record of A has to be compared with all records of B.
- Number of potential record pair comparisons is |A X B|.
- Reduce the number of potential record pair comparisons.
- Blocking: Use a single record field or a set of fields (blocking key) to split the database into blocks.
- Iterate the matching process with a different blocking scheme.
- Sorted neighbourhood (Hernandez and Stolfo, 1998): Sort A and B using the same key; search for possible matching records only inside a window of a fixed dimension which slides on the ordered A and B.

Computing overall similarity measure

- Combine the metrics, typically using a weighted sum, where the weights are defined according to the quality of the fields.
- Define a lower and upper threshold for this overall similarity measure.
- Determine the matching status of candidate pairs (match, non-match or possible match).
- Clerical review for the possible match: resource intensive.

Clerical review

- Case-by-case review of uncertain matches (fall between the upper and lower cutoff values).
- Add additional variables to record layout to assist the designation of the record match status.
- Apply experience, common sense and human intuition.
- To increase the reliability—preform by independent reviewers.

Software for record linkage

- Commercial Software
  - Most of them are "black box" from the users' perspective (the source code of their linkage engines is not available for inspection).
  - Specialized to a certain domain, e.g. de-duplication of customer mailing lists
  - Affordable are only smaller systems limited in their ability to process different data types and; limited functionality; can process only small data sets.

Software for record linkage

- Open Source and Free Software
  - Allow access to the source code of their linkage engines.
  - Free of charge - although this not always means “no costs”.
  - Flexible and extendible.
  - Include large number of linkage techniques.
  - Allow the practitioner to experiment with traditional as well as advanced linkage techniques; the user is able to understand up to a certain degree, many technical details.

Software for record linkage

- RELAIS
  - ISTAT
  - Implemented in Java and R (both languages are open source and can be used on different platforms)
  - Graphical User Interface (GUI) available, written in Java
  - Input and output data in relational database—mySql—also open source product
  - Available for all major platforms
  - `https://joinup.ec.europa.eu/software/relais/description`

Software for record linkage

- FEBRL
  - *Freely Extensible Biomedical Record Linkage* (FEBRL) System
  - The Australian National University, Canbera
  - Implemented in Python (free object oriented programming language)
  - Graphical User Interface (GUI) available
  - Input from text files (CSV), SQL in the future
  - Available for all major platforms
  - Contains many recently developed record linkage techniques
  - `http://sourceforge.net/projects/febrl/`

Software for record linkage

- RecordLinkage R package
  - An R package available from CRAN
  - Machine learning methods are utilized
  - Decision trees (rpart), bootstrap aggregating (bagging), ada boost (ada), neural nets (nnet) and support vector machines (svm).
  - `http://cran.r-project.org/web/packages/RecordLinkage/index.html`
  - `http://journal.r-project.org/archive/2010-2/RJournal_2010-2_Sariyar+Borg.pdf`

Outline

Other Data Sources

- Apart from the Administrative and statistical sources which we considered so far, there are other commercial sources.
- Related to the increased computing power, the technological advancement and the internet new sources are emerging which soon will be very important in official statistics.
- Telephone directories
  - Prepared by telephone companies.
  - Can be useful in adding or confirming SBR data.
  - Should not be used as sources of new enterprises.

Other Data Sources

- Internet
  - The Internet can be considered as a data source (belonging to the vast category of Big Data), that may be harnessed in substitution, or in combination with, data collected by means of the traditional instruments of a statistical survey—Barcaroli (2014).
  - Cannot currently be reliably used for identifying new enterprises.
  - Can provide information on the economic activity, on the production profile, on up-to-date addresses, etc.
  - $\Rightarrow$ Internet is becoming an important source.

Other Data Sources

- Payroll, taxation and accounting service providers
  - Provide enterprises with services that involve paying an enterprise's staff and/or making returns to the taxation authorities on its behalf and/or managing its accounts.
  - Possible to build agreements with such service providers and their client enterprises that allow the service providers to provide data for the enterprises directly to the NSI.
  - $\Rightarrow$ Reduced response burden; provides faster and more efficient data flow to the NSI.
  - Identification of the relevant enterprises is required.
  - It may still be necessary to contact the enterprises directly to obtain information that the service providers cannot provide, for example economic activity classification

Commercial data providers

- Payroll, taxation and accounting service providers
  - A number of commercial enterprises provide global, regional and domestic company information.
  - These data are also valuable for maintaining an SBR, in particular by providing information on enterprise group structures.
  - Usually based on publicly accessible information; could also be obtained directly by the NSI.
  - However, additional effort and costs for processing of these free data might be necessary.

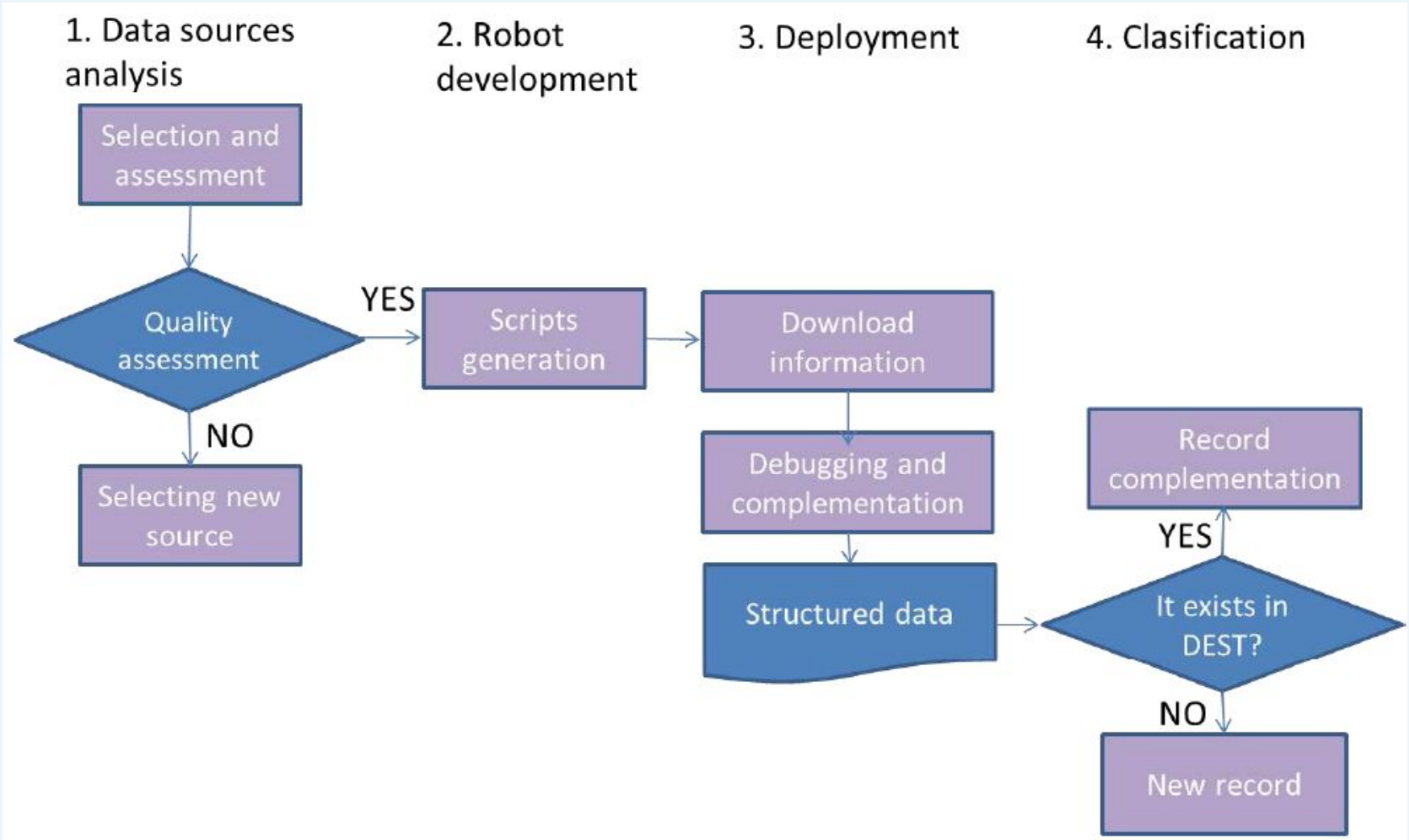
Big data

- Big data
  - Big data comes in many shapes and sizes. can be used in many different ways: real-time fraud detection, web display advertising and competitive analysis, call center optimization, social media and sentiment analysis, intelligent traffic management and smart power grids, to name just a few.
  - Large, possibly unstructured data sets that are potentially available in real time.
  - Opportunities for using big data in Official statistics are still developed.
  - Structuring and editing the data might also require substantial investment but be worthwhile if accompanied by significant benefits in terms of coverage and timeliness.

Big data

- ISTAT (Italy): Project "Potential use of web scraping techniques and text mining to complement the survey of Information and Communication Technology in business" (Barcarioli (2014).
- NSO (UK): Prototypes of web robots to capture prices from supermarket chains.
- DANE (Columbia): Project "Big Data as input for updating the statistical Business Register (DEST for its acronym in Spanish), using web-scraping techniques".

Big data



Source: `http://cepei.org/portfolio/big-data-as-input-for-updating-the-statistical-business-register-dest-using-web`

Outline

References I

Thomas N. Herzog, Fritz J. Scheuren, William E. Winkler (2007)
Data Quality and Record Linkage Techniques, Springer 2007.

Barcaroli et al.
Use of web scraping and text mining techniques in the Istat survey on "Information and Communication Technology in enterprises", 2014, European Conference on Quality in Official Statistics (Wien 2014)
`http://www.q2014.at/fileadmin/user_upload/Iad_in_ICT_survey_PAPER.pdf`.

ONS.
Further information about IDBR sources, structure and updating for publications, UK Business and Business Demography publications, 2015,
`http://webarchive.nationalarchives.gov.uk/20160105160709/http://www.ons.gov.uk/ons/guide-method/method-quality/specific/business-and-energy/business-population/index.html`.

Inclusive and Sustainable Industrial Development

Creating shared prosperity | Safeguarding the environment

UNIDO
UNITED NATIONS
INDUSTRIAL DEVELOPMENT ORGANIZATION