Les défis d'utilisation de données administratives et des enquêtes par sondage dans le Système Statistique Européen*

Dario BUONO (Eurostat)
Unité B1 Qualité, Méthodologie et Recherche
Enrico INFANTE (Eurostat)
Unité B1 Qualité, Méthodologie et Recherche
Fabienne MONTAIGNE (Eurostat)
Unit F1: Unité Statistiques Sociales: Modernisation et Coordination
Jean-Marc MUSEUX (Eurostat)
Unité B1 Qualité, Méthodologie et Recherche

*** Les points de vue et les opinions exprimés dans ce document n'engagent que leurs auteurs et ne necessairement refelct ceux des institutions pour lesquelles ils travaillent**

Sommarie

- Cadre du ESS
- Utilisation de données administratives
- Fusion des données
- L'industrialisation des processus
- Principales conclusions

Introduction

- L'utilisation de données administratives est promue pour des raisons de coût et d'efficacité
- Dans le cadre du Système Statistique Européen (ESS), un aperçu des techniques qui sont nécessaires pour traiter les données administratives est donnée

L'utilisation de données administratives (1)

- L'utilisation des données administratives est l'un des principaux éléments de la "Vision" pour les Statistiques Européennes adoptées par la Commission Européenne en Août 2009, et le document "Vision" approuvé par le Comité du Système Statistique Européen (ESSC)
- Eurostat soutient activement cette initiative avec la proposition de révision du règlement 223/2009 relatif aux Statistiques Européennes
- Eurostat a entrepris le projet ESS VIP "ADMIN" afin de répondre au besoin d'informations méthodologiques sur l'utilisation des données administratives

L'utilisation de données administratives (2)

- Presque tous les Pays Membres ont augmenté l'utilisation des sources administratives, car il est nécessaire de:
  - **réduire le coût de la collecte des données**
  - **réduire la charge des répondants**
  - **recueillir des données une seule fois et l'utiliser à des fins différentes**
- Une utilisation accrue des sources administratives implique également le risque d'impact sur la qualité:
  - **La pertinence des statistiques pourrait diminuer**
  - **L'exactitude des statistiques sur de petites zones/populations pourraient augmenter**
  - **Habituellement, l'effet sur délais est négatif**
  - **La comparabilité des statistiques risque d'être réduite**

L'utilisation de données administratives (3)

Réduction des coûts

Plus output

Réduire la charge des répondants

Gestion de "field force"

statistiques sur de petites zones/populations

Pertinence

Exactitude

Délais

Comparabilité

Unique administratif source

Fusion des données (1)

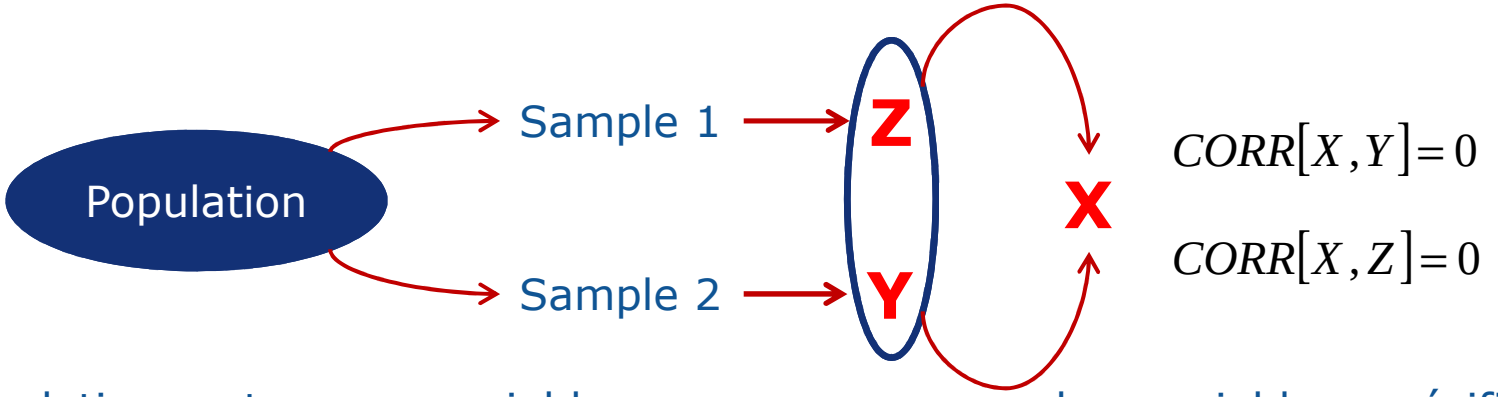
- **Fusion des données** est basée sur un approche model-based pour fournir des informations sur les variables et les indicateurs recueillis auprès de sources multiples
- Les données provenant de différentes sources qui correspondent à des unités similaires sont identifiés et liés
- L'avantage de cette approche est la possibilité d'améliorer l'utilisation complémentaire et le potentiel analytique des sources de données existantes (il augmenter l'efficacité)

Fusion des données (2)

- Deux approches principales peuvent être définies en termes de résultats qui peuvent être obtenus grâce à la fusion:
  - **L'approche macro: est plus complet et se réfère à l'identification d'une structure qui décrit la relation entre les variables non conjointement observées**
  - **L'approche micro: est plus restrictive et se réfère à la création d'un système complet de fichier de micro-données dans lequel les données sur toutes les variables est disponible pour chaque unité**

Fusion des données (3)

Imputation des variables "target" à partir d'un "donor" à un "recipient survey"



La relation entre ces variables communes avec les variables spécifiques observés dans un seul ensemble de données (**donor**) seront explorées et utilisées pour imputer les unités de l'autre ensemble de données (**recipient**) les variables pas directement observées

Un dataset synthétiques est généré avec des informations complètes sur **X**, **Y** and **Z**

Fusion des données (4)

- Il y a quatre niveaux de validité pour une procédure de fusion:
  1) Les distributions marginales et conjointes des variables dans l'échantillon du "Donor" sont conservés dans le fichier fusionné (normalement atteint)
  2) La structure de corrélation et moments d'ordre supérieur des variables sont conservées après la fusion des données (à vérifier)
  3) La vraie distribution conjointe de toutes les variables se reflète dans le fichier fusionné (à vérifier)
  4) Les vraies valeurs, mais inconnue de la variable Z des unités "recipient" sont reproduits (normalement pasatteint)

Fusion des données (5)

Les principales conclusions sont les suivantes:

- Un premier défi dans tout exercice de fusion appliquée est l'harmonisation des différentes sources de données
- Évaluation de la qualité dans le cadre de la fusion doit tenir compte de plusieurs facteurs essentiels: la qualité et la cohérence des sources, le pouvoir explicatif des variables communes, les méthodes d'imputation appliquées et les méthodes utilisées pour calculer les estimations basées sur les ensembles de données fusionnées
- Un facteur essentiel est la possibilité de remédier aux limitations inhérentes à la fusion statistique, et de fournir une mesure de la qualité des estimations basées sur les ensembles de données appariées
- L'existence d'une information auxiliaire est un point essentiel pour toute fusion
- Fusion appliquée dans un système ex post intégré doit procéder à plusieurs étapes initiales de la réconciliation des sources avant l'application effective de l'adéquation des techniques et des besoins de contrôle de l'incertitude des estimations en raison d'hypothèses implicites

L'industrialisation des processus (1)

Le concept de l'industrialisation, qui a fonctionné dans le contexte des usines, présente des similitudes suffisantes pour suggérer cela devrait fonctionner aussi bien dans un contexte de production statistique

Dans le contexte des statistiques sociales au niveau Européen, ce serait traduite par une modernisation des enquêtes sociales, avec les objectifs d'une augmentation de leur efficacité, du potentiel analytique des données et de la réactivité des besoins des utilisateurs. Pratiquement, il s'agit d'une utilisation élargie sur des données administratives, l'alignement de la production des données et des processus d'assurance de la qualité et de la normalisation des outils et des processus

L'industrialisation des processus (2)

- Méthodes avancées de collecte de données doit être assisté par ordinateur

CAPI ⟶ **Face-to-face**

CATI ⟶ **By telephone**

CAWI ⟶ **Web based**

La collecte de données basée sur Internet a deux principaux avantages: il est moins cher que les méthodes traditionnelles, et il pourrait être utile pour augmenter les taux de réponse

Comme inconvénients, le layout et les options ont une forte influence sur les résultats

Le potentiel de la CAWI pour le codage et la vérification est très élevé

L'industrialisation des processus (3)

- La combinaison de plusieurs modes de collecte dans la même enquête est également recommandé, car il permet de choisir le mode de collecte le plus approprié pour chaque type de répondant et pour chaque type de données à recueillir
- Il y a au moins trois raisons principales pour choisir un design mode multiple enquête:
  - **Augmenter la rentabilité**
  - **Améliorer les taux de réponse**
  - **Réduire les erreurs de mesure**

L'industrialisation des processus (4)

- La modularité est la construction d'un processus complexe de petits sous-systèmes qui peuvent être conçus indépendamment mais encore fonctionner ensemble comme un tout. Il donne une structure à un système compliqué
- Un module est un ensemble de variables homogènes en termes de thèmes de mesure de contenu de manière globale
- Avec la modularité, il est possible de personnaliser par une meilleure mise en place du recouvrement de données pour les besoins de sortie: inclure ou exclure un module facilement dans une période de collecte de données, les modules peuvent avoir des temps différents, mettre en oeuvre des modules en tant que suivi de sondages avec un mode de collecte de données spécifique

Principales conclusions

- Eurostat a entrepris le projet ESS VIP "ADMIN" afin de répondre au besoin d'informations méthodologiques sur l'utilisation des données administratives
- Un premier défi dans tout exercice de mise en correspondance appliquée est l'harmonisation des différentes sources de données
- Dans le cadre d'une modernisation des enquêtes sociales, avec les objectifs d'une augmentation de leur efficacité, l'utilisation de données administratives devrait être élargi
- Les principales caractéristiques à analyser pour un processus d'industrialisation d'un système moderne des enquêtes sociales au niveau européen sont les modes de collecte de données, la conception des collections de données et l'approche modulaire

Questions?

THE CHALLENGE OF USING ADMINISTRATIVE DATA VS. SAMPLE SURVEYS IN THE EUROPEAN STATISTICAL SYSTEM

Dario BUONO (Eurostat)

Unit B1: Quality, Methodology and Research

Enrico INFANTE (Eurostat)

Unit C1: National Accounts Methodology. Sector Accounts. Financial Indicators

Fabienne MONTAIGNE (Eurostat)

Unit F1: Social Statistics: Modernization and Coordination

Jean-Marc MUSEUX (Eurostat)

Unit B1: Quality, Methodology and Research

**** The views and the opinions expressed in this paper are solely of the authors and do not necessarely refelct those of the institutions for which they work***

Summary

- ESS framework
- Use of Administrative Data
- Data Matching
- Industrialization process
- Main Findings

Introduction

- The use of administrative data is promoted for cost and efficiency reasons
- In the context of the European Statistical System (ESS), an overview of the techniques that are necessary for dealing with administrative data is here given

eurostat

The use of Administrative Data (1)

- The increased use of administrative data is one of the main elements of the Vision for European Statistics adopted by the European Commission in August 2009, and the Vision implementation document endorsed by the European Statistical System Committee (ESSC)
- Eurostat actively supports this move with the proposed revision of Regulation 223/2009 on European Statistics
- Eurostat is undertaking the ESS VIP "ADMIN" in order to address the need for methodological information on the use of the administrative data

The use of Administrative Data (2)

- Almost all Member States have increased the use of Administrative Sources, as there is a need to:
  - reduce the cost of data collection
  - reduce the burden on respondents
  - collect data only once and use it for different purposes
- An increased use of Administrative Sources also implies the risk of an impact of some quality dimensions:
  - The relevance of statistics might decrease
  - The accuracy of statistics on small areas/populations might increase
  - Commonly the effect on timeliness is negative
  - Comparability of statistics risks to be reduced

The use of Administrative Data (3)

Reduction of costs

More output

Reduction of burden respondent

Management of field force

Small geo areas and/or sub-populations

Relevance

Accuracy

Timeliness

Comparability

Single Administrative Source

Data Matching (1)

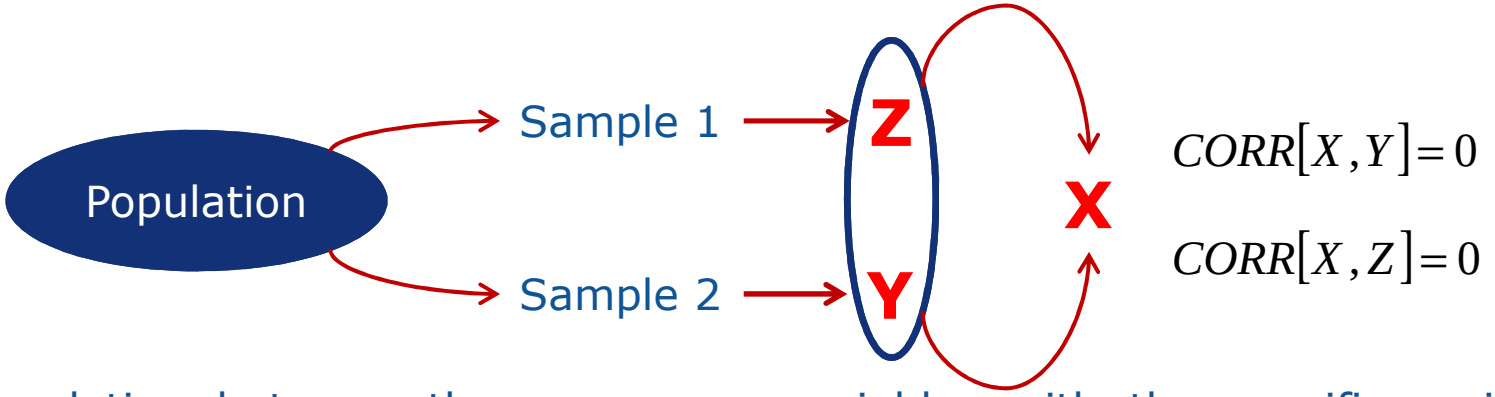
- **Statistical Matching** is model-based approach for providing joint information on variables and indicators collected through multiple sources
- Records from different sources that correspond to *similar* units are identified and linked
- The benefit of this approach is the possibility to enhance the complementary use and analytical potential of existing data sources (it increase the efficiency)

Data Matching (2)

- Two main approaches can be delineated in terms of outputs that can be obtained through matching:
  - The *macro approach*: is more comprehensive and refers to the identification of any structure that describes the relationship among the variables not jointly observed of the datasets
  - The *micro approach*: is more restrictive and refers to the creation of a complete micro-data file where data on all the variables is available for every unit

Data Matching (3)

Imputation of the target variables from a donor to a recipient survey



The relation between these common variables with the specific variables observed only in one dataset (**donor**) will be explored and used to impute to the units of the other dataset (**recipient**) the variables not directly observed

→ A synthetic dataset is generated with complete information on **X**, **Y** and **Z**

Data Matching (4)

- There are four levels of validity for a matching procedure:
  1) The marginal and joint distributions of variables in the donor sample are preserved in the statistical matching file (*normally reached*)
  2) The correlation structure and higher moments of the variables are preserved after statistical matching (*can be checked*)
  3) The true joint distribution of all variables is reflected in the statistical matching file (*can be checked*)
  4) The true but unknown values of Z variable of the recipient units are reproduced (*normally not attained*)

Data Matching (5)

The **main findings** are:

- A first challenge in any applied matching exercise is the harmonization of different data sources
- Quality evaluation in the framework of matching needs to account for several critical factors: the quality and the coherence of the sources, the explanatory power of common variables, the matching/imputation methods applied and methods used to compute estimates based on the matched datasets
- A critical factor is the possibility to address the limitations inherent in statistical matching, and provide a measure of quality for estimates based on matched datasets
- The existence of auxiliary information is an essential point for any matching
- matching applied in an ex post integrated system needs to undertake several initial steps of reconciliation of sources before the actual application of matching techniques and needs to control for the uncertainty of estimates due to implicit assumptions

Industrialization processes

The concept of industrialization, that worked in a manufacturing context, has sufficient similarities to suggest it should work equally well in a statistical production context

In the context of social statistics at European level, this would be translated in a modernization of social surveys, with the aims of an increase of their efficiency, of the analytical potential of data and of responsiveness of user needs. Practically, this is about expanded use of administrative data, aligning of data production and quality assurance processes and the integration of IT architecture, standardization of tools and processes

Industrialization processes

Advanced data collection methods can only be computer assisted

CAPI ⟶ **Face-to-face**

CATI ⟶ **By telephone**

CAWI ⟶ **Web based**

The web based data collection has two main advantages: it is cheaper than traditional methods, and it could be helpful in increasing response rates

As drawbacks, the lay out on the screen, the options and the buttons have a strong influence on the results

The potential of the CAWI for coding and checking is very high

Industrialization processes

The combination of several modes of collection in the same survey is also recommended, as it allows choosing the most appropriate collection mode for each respondent type and for each type of data to be collected

There are at least three main reasons for choosing a multiple mode survey design:

1) **Increase cost-effectiveness**
2) **Improve response rates**
3) **Reduce measurements errors**

Industrialization processes

- Modularity is building a complex product or process from smaller subsystems that can be designed independently yet function together as a whole. It gives structure to a complicated system
- In social statistics a module is a set of variables (with strong association within the module and very weak between them) homogeneous in terms of content measuring topic comprehensively
- With the modularity approach it is possible to customise by better fitting the data collection to output needs: include or exclude a module easily in a data collection period, modules can have different timings, implement modules as follow up surveys with a specific data collection mode
- The modularity approach can be used in the integration of different social surveys

Main findings

- Eurostat is undertaking the ESS Vision Infrastructure Project (VIP) "ADMIN" in order to address the need for methodological information on the use of administrative data
- A first challenge in any *applied matching exercise* is the harmonization of different data sources
- In the context of a modernization of social surveys, with the aims of an increase of their efficiency, the use of administrative data should be expanded
- The main features to be analysed for an industrialization process of a modern system of the social surveys at European level are the data collection modes, the data collections design and the modularity approach

Questions?