Perspectives in integration of data from multiple probabilistic and non-probabilistic sources: a few remarks

Pier Luigi Conti[1]

[1]Dipartimento di Scienze Statistiche
Sapienza Università di Roma

June 9, 2026

Multiple sources - 1

- Consequence of the *data deluge* phenomenon: multiple data sources are becoming more and more available for economic, social and political analyses.
- Data coming from different sources (*e.g.* Official Statistics survey data, administrative data, experimental data, observational data, data from sensor, transactions data, etc.) are collected, at different levels of granularity (aggregation), by different organizations and for different purposes.
- Collected data are frequently in the public domain, or available upon request.
- This is a part of the *Big Data Revolution*, that has opened new opportunities to access data potentially useful to investigate relationships among variables.

Multiple sources - 2

- Different sources may have not only different formats and conceptual schemes, but also different statistical characters and units.
  - A single source hardly ever contains all variables of interest for statistical analysis. Frequently, a single source just contains some of the variables of interest, but not all.
  - Different data sources could not refer to the same population.
  - Different data sources could refer to different units of the same population. For instance, this occurs in case of different samples from the same population.
  - Even if two sources refer to the same units, it could be difficult to exactly identify records pertaining to the same unit.
  - Variables from different sources could be observed with different quality levels, with different incidence of measurement errors.
  - The data collection mechanism (sampling design, in many cases) could be controlled for some sources (probability samples) and uncontrolled for some other (non-probability samples).

Multiple sources - 3

- Necessary to harmonize alternative data sources into an *integrated system* for empirical analyses, thus ensuring that data coming from different sources are *accurate*, *complete*, *logically consistent*.
- Raw and partial categorization of data sources:

  Survey data, coming well-designed sample surveys. They are gold standard, and provide high-quality information. Despite non-responses and measurement errors, survey data are relatively easy to deal with, *via* standard tools of statistical inference for survey sampling.

  - Administrative data.
  - Organic data, generated without a formal data collection processes. They often originate from everyday activities and interactions, and are typically collected passively from various sources (*e.g.* social media data, web data, transaction data, mobile data etc.).
  - Macro data and micro data.

Integrating data from different sources: paradigm change for statistical inference - 1

Old but well-established paradigm of Statistics: *High quality data come from a (frequently ad hoc) statistical sample survey, and are collected in an appropriate database.*

- Statisticians use data to make inference on population parameters, and the sources of error are essentially two.
  - Sampling error, due to the discrepancy between sample and whole population.
  - Non-sampling errors, due to measurement errors and, mainly, to non-responses.
- The main effort of statisticians is twofold.
  - Plan a good data acquisition process, a good sampling design.
  - Develop good statistical methods to analyze collected data, that require to account for both sampling and non-sampling errors.

Integrating data from different sources: paradigm change for statistical inference - 2

Among the most challenging patterns for multiple data sources, two are of particular interest (Yang and Kim, 2020).

P1. Two samples A, B are available. Sample A is a small-scale probability sample, where only a (multidimensional) variable X is observed. Sample B is a large-scale non-probability sample, where the variables X, Y are observed, Y being the outcome of interest.

P2. Two samples (either probabilistic or non-probabilistic) are observed. In sample A, the variables X, Y are observed; in sample B, the variables X, Z are observed. Here, X is the common variable to the two samples. The variable Y is specific of sample A, and the variable Z is specific of sample B. No available observations containing simultaneously all the variables X, Y, Z are available.

Integrating data from different sources: paradigm change for statistical inference - 3

- Combining different data sources implies the presence of *new potential sources of errors*, that imply a change in the traditional paradigm of statistical inference.
- Basic question: *Can the combined data be considered as observations of the population of interest?*
- Statistical data obtained by combining partial observations from different sources *do not necessarily correspond to observations of real units, because of the intrinsic uncertainty in the combination process.*
- Combined partial observations actually correspond to *virtual units*, that define, in their turn, a *virtual population.*

Estimation error under virtual population - 1

Basic framework - 1

- $\mathcal{U}$ real population.
- For each unit $i \in \mathcal{U}$, a triplet $(\mathbf{x}_i, \mathbf{y}_i)$ (values of characters $\mathbf{X}$, $\mathbf{Y}$) is defined.
- For each unit $i \in A \cup B$, the triplet $(\widetilde{\mathbf{x}}_i, \widetilde{\mathbf{y}}_i)$ is obtained through data integration.
- Due to the possible discrepancy between the two above triplets, unit i does not necessarily belong to the real population. It is a *virtual unit*.
- This underlies the concept of *virtual population* $\widetilde{\mathcal{U}}$, composed by virtual units.

Estimation error under virtual population - 2

Basic framework - 2

- $F_N(\mathbf{x}, \mathbf{y})$: distribution function of $(\mathbf{X}, \mathbf{Y})$ over the *real* population.
- $\widetilde{F}_N(\mathbf{x}, \mathbf{y})$: distribution function of $(\mathbf{X}, \mathbf{Y})$ over the *virtual* population.
- $\theta(\widetilde{F}_N)$: corresponding parameter for the virtual population.
- $\widehat{F}$: estimate of F_N based on *integrated* sample data $(\widetilde{\mathbf{x}}_i, \widetilde{\mathbf{y}}_i)$, $i \in A \cup B$.

Estimation error under virtual population - 3

Problem to be considered with integrated data obtained by combining different sources: $\widehat{F}$ *is typically a consistent estimator of* $\widetilde{F}_N$*, but not necessarily of* F_N.

Estimation error for $\widehat{\theta} = \theta(\widehat{F})$

$$\underbrace{\widehat{\theta} - \theta}_{\substack{\text{Total} \\ \text{estimation error}}} = \underbrace{\theta(\widehat{F}) - \theta(\widetilde{F}_N)}_{\substack{\text{Virtual population} \\ \text{estimation error}}} + \underbrace{\theta(\widetilde{F}_N) - \theta(F_N)}_{\substack{\text{Discrepancy between} \\ \text{virtual and real populations}}} .$$

- $\theta(\widehat{F}) - \theta(\widetilde{F}_N)$: error due to observing only a *part* of the *virtual* population.
- $\theta(\widetilde{F}_N) - \theta(F_N)$: additional error due to the *discrepancy between virtual and real population*.

Estimation error under virtual population - 4

The two errors exhibit a completely different behaviour as the sample sizes n_A, n_B and the population size N increase.

$$\begin{aligned} \theta(\widehat{F}) - \theta(\widetilde{F}_N) &= O_p\left(\frac{1}{\sqrt{n_A + n_B}}\right) \text{ as } n_A,\ n_B,\ N \to \infty; \\ \theta(\widetilde{F}_N) - \theta(F_N) &= O_p(1) \text{ as } n_A,\ n_B,\ N \to \infty. \end{aligned}$$

Estimation error under virtual population - 5

- The virtual population estimation error depends on the quantitative *amount of information* provided by sample data. It is related to the *quantity* of information provided by sample data: the higher the amount of information (in terms of sample size), the smaller the error.
- The discrepancy between virtual and real populations depends on the *value of information* provided by sample data and integration process. It is a sort of *intrinsic error* due to data collection and data integration processes. Uncontrolled data collection process and/or information sources where not all variables of interest are simultaneously observed reduces the *value* of information provided by sample data. It is related to the *quality* of information provided by sample data.

Estimation error under virtual population - 6

- A large quantity of information contained in sample data (= small value of $\theta(\widehat{F}) - \theta(\widetilde{F}_N)$) cannot compensate for the poor quality of that information (= large value of $\theta(\widetilde{F}_N) - \theta(F_N)$).
- *Major source of trouble*: the quantity

$$e(\widetilde{F}_N,\, F_N) = \theta(\widetilde{F}_N) - \theta(F_N)$$

  is *not estimable*, unless special, very restrictive (and not generally testable) assumptions are made.
- The only general statement that can be made is that $e(\widetilde{F}_N,\, F_N) \in \Xi$, where Ξ plays the role of *uncertainty set* of the data integration process.
  - The wider the set Ξ, the higher the uncertainty related to the data integration process.

Estimation error under virtual population - 7

Main issues.

- Development of methodologies for data integration allowing for a safe assessment of different sources of errors. This problem arises when one has to decide how to combine data from different sources.
- This calls for *measures of uncertainty* related to the uncertainty set Ξ. Intuitively speaking, each measure of the size of Ξ could play the role of measure of uncertainty. An important requisite of a measure of uncertainty is that it should be *estimable* through sample data.
- This approach is used, for instance, in statistical matching. In that context, a bound for the maximal value of a distance between $\widetilde{F}_N(\mathbf{x}, \mathbf{y})$ and $F_N(\mathbf{x}, \mathbf{y})$ is studied, and estimated on the basis of sample data.

Estimation error under virtual population - 8

Main issues cont'd.

- A further source of complication occurs when B is non-ignorable, as it usually occurs in non-probability samples. In addition to the intrinsic uncertainty due to the data integration process, there is the uncertainty due to the unknown sample design. Steps in this direction are in Conti, Marella (2025).
- Use of secondary data, *i.e.* data already obtained by some integration process, as in secondary analysis. In this case, focus is in modeling errors and analyzing their impact on statistical inference.
- Use of secondary data, *i.e.* data already obtained by some (unspecified, unknown) integration process, as in secondary analysis.
- In this case, focus is in modeling errors and analyzing their impact on statistical inference.

Final remarks

In view of the above considerations, in my opinion Istat should adopt a clear strategy to plan and advance research in the crucial area of multi-source data. To be effective, such research should encompass both investments in experimental activities — possibly carried out in collaboration with the academic community — and efforts aimed at strengthening internal organizational processes.