An Italian experience to evaluate the quality of integrated administrative data

Simonetta Cozzi
Marco Di Zio
Danila Filipponi
Istat

25th Meeting of the Wiesbaden Group on Business Registers
- International Roundtable on Business Survey Frames
Tokyo, 8 – 11November 2016

Istat

Administrative data

Use of administrative data in the production of official statistics has exceedingly increased in the recent years.

In the *past*, administrative data were mostly used :

- for support survey frame
- for estimation, edits, imputation as auxiliary data
- for data analysis and validation

Now, more and more often, administrative data are used for ***direct collection***

Effective use of AD requires the development of an appropriate methodological framework for assessing the quality of administrative data for statistical purpose.

Arcolaio Project

ARCOLAIO has developed ***methodologies*** for monitoring the ***quality of administrative data*** in each phase of a statistical production based on administrative data.

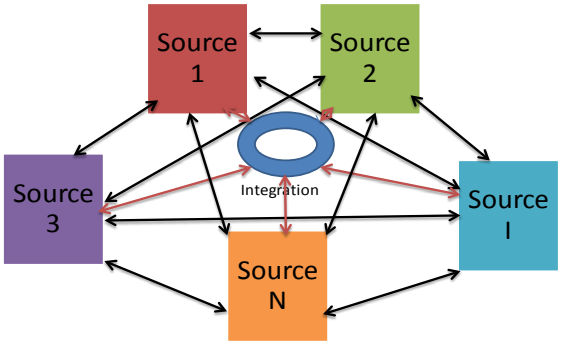
ARCOLAIO has proposed methodological solutions :

- To evaluate the quality of administrative data in the acquisition phase through an approach based on quality indicators
- To analyze the longitudinal stability of administrative data, in terms of metadata and data, and their impact on production process
- To evaluate the quality of statistics based on an integrated use of administrative data

Integrated System of administrative Microdata (SIM)

The framework within which the project has been developed is the Integrated System of administrative Microdata (**SIM)**.

SIM identifies and integrates the information present in the administrative sources to define an *infrastructure* usable for the production of *social and economic statistics.*



SIM includes:

- ***units***: individuals, economic units and places;
- The ***variables*** on these units;
- The ***relationships*** between units, between units and variables, and between variables;

Integrated System of administrative Microdata (SIM)

SIM includes social and economic data.

Subsystems of the *units*

- SIM individuals
- SIM economic units;

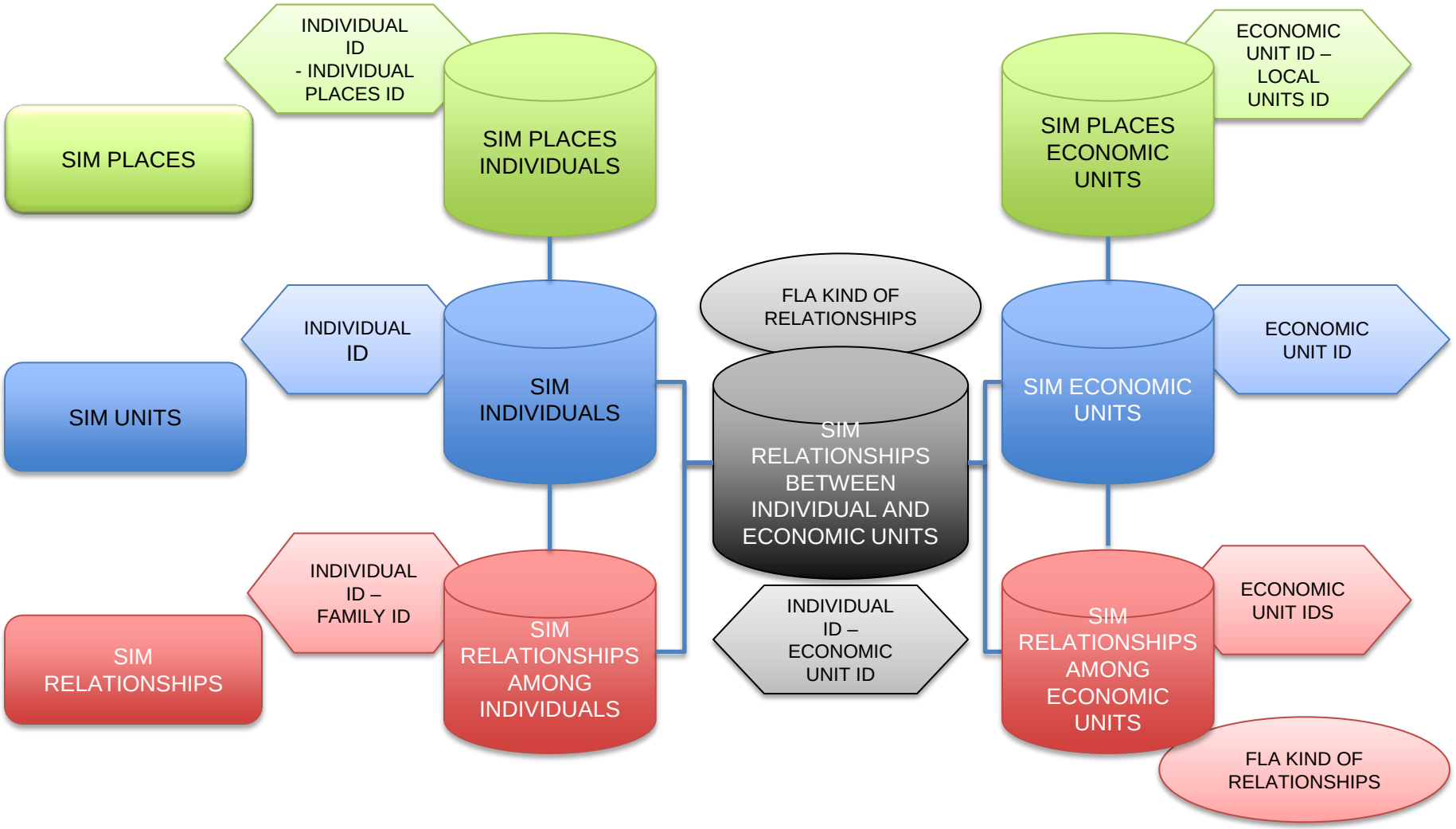
Subsystems of the *places*

- SIM economic units places
- SIM individual places;

Subsystems of the *relationships*:

- SIM relationships among economic units;
- SIM relationships among individuals;
- SIM relationships between individuals and economic units

Integrated System of administrative Microdata (SIM)



Arcolaio Project

SIM *can be thought as a* ***statistical multipurpose population*** *consisting of base elements (units, variables, relations) that appropriately combined provide population characterizing the new phenomenon of interest.*

Arcolaio defines

- methods for evaluating the quality of SIM,
- statistical methodologies to
  - derive statistical populations from an integrated system of administrative microdata
  - assess the quality of the obtained outputs.

Arcolaio Project

The project has been structured in different WP:

➢ WP1-the Administration Data Documentation System

- Define an administrative data documentation system, regarding to the structural metadata, process and quality indicators, coherent with the standard documentation systems (GSIM)

➢ WP2- Assessment of the quality of the integrated microdata

- Develop quality indicators to evaluate the quality of the available administrative sources exploiting the information gathered by the integration process

Arcolaio Project

➢ WP3-Assessment of changes in administrative data

- Develop methods for evaluating the usability of administrative source using the longitudinal information

Regulatory changes of AD may induce discontinuity producing significant impacts on the statistics production. Analysis must be carried out to verify the presence of unexpected lack of quality before AD enter into statistical production process

Arcolaio Project

- **WP4 - Statistical methods in the use of administrative data for statistical purposes and evaluation of the quality**

- Specify the inferential context in which estimates based on administrative data are produced;
- Define statistical methods for the detection and correction of representation and measurement errors of administrative sources;
- Evaluate analyses the quality of the estimates produced by administrative data based on inferential context

Arcolaio Project – WP3

WP3 developed *statistical methods* for validation of an administrative source in the input stage using temporal dimension.

- Outlier analysis has been used for identifying unusual changes over time:
  - Parametric and non-parametric approach for the identification of outlier in contingency tables
  - Trend analysis

- Application with simulated and real data

Arcolaio Project – WP3

Methodological approach

Outlier identification procedures for contingency tables

Outlying observations are generally viewed as deviations from a model assumption:

- the majority of observations -*inliers*- are assumed to come from a selected model (*null model*);
- few units – *outliers*- are thought of as coming from a different model.

The *outliers identification problem* is then translated into the problem of identifying those observations that lie in an *outlier region* defined according to the selected null model (GLM).

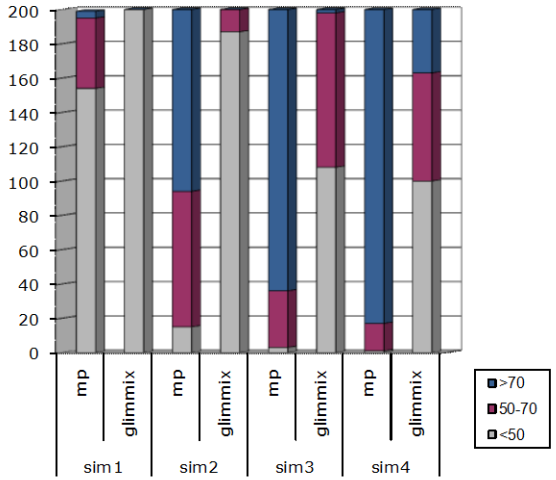
Procedures for the identification of such outliers are derived using the classical maximum likelihood estimator and an non parametric estimator based on median polish.

Kuhnt (2004) Outlier identification procedures for contingency tables using maximum likelihood and L1 estimates. Scand J. Statist. 31:431–442; Kuhnt (2010) Breakdown concepts for contingency tables. Metrika 71:281–294; Kuhnt, Rapallo, Rehage(2014) Outlier detection in contingency tables based on minimal patterns, Statistics and Computing May 2014, Volume 24, Issue 3, pp 481-491

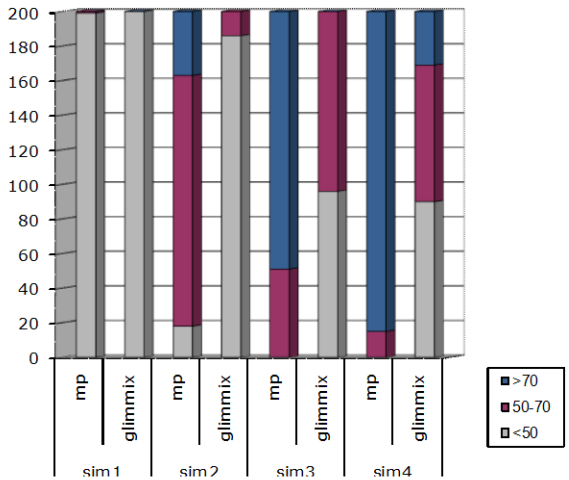
Arcolaio Project – WP3

Simulation study

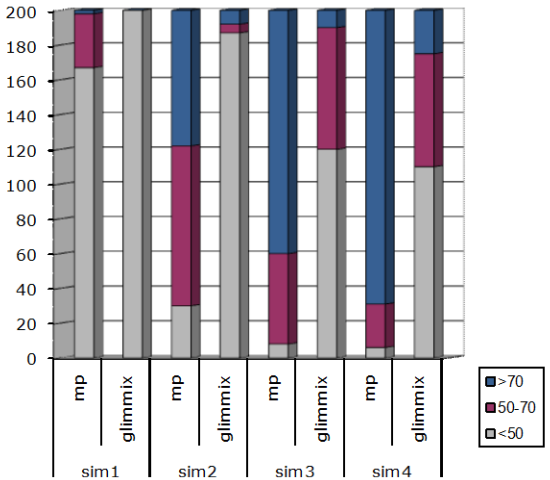
(Table dimension: 20 ×7, simulated outliers 10%)



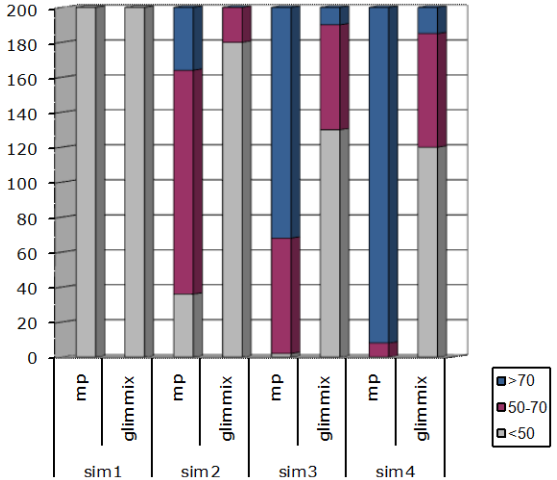
(Table dimension: 50 ×7, simulated outliers 10%)



(Table dimension: 20 ×7, simulated outliers 15%)



(Table dimension: 50 ×7, simulated outliers 15%)



Sim1=distorted data 10%
Sim2=distorted data15%
Sim3=distorted data20%
Sim4=distorted data40%

Istat

Arcolaio Project – WP3

Trend Analysis

Trend analysis examines changes in core data elements over time and includes comparisons of counts or proportions over time, as well as more sophisticated time series analysis, smoothing or curve fitting.

Methods for identifying unusual changes over time have been developed using technique* that choose the most appropriate model for smoothing the data curve/line.

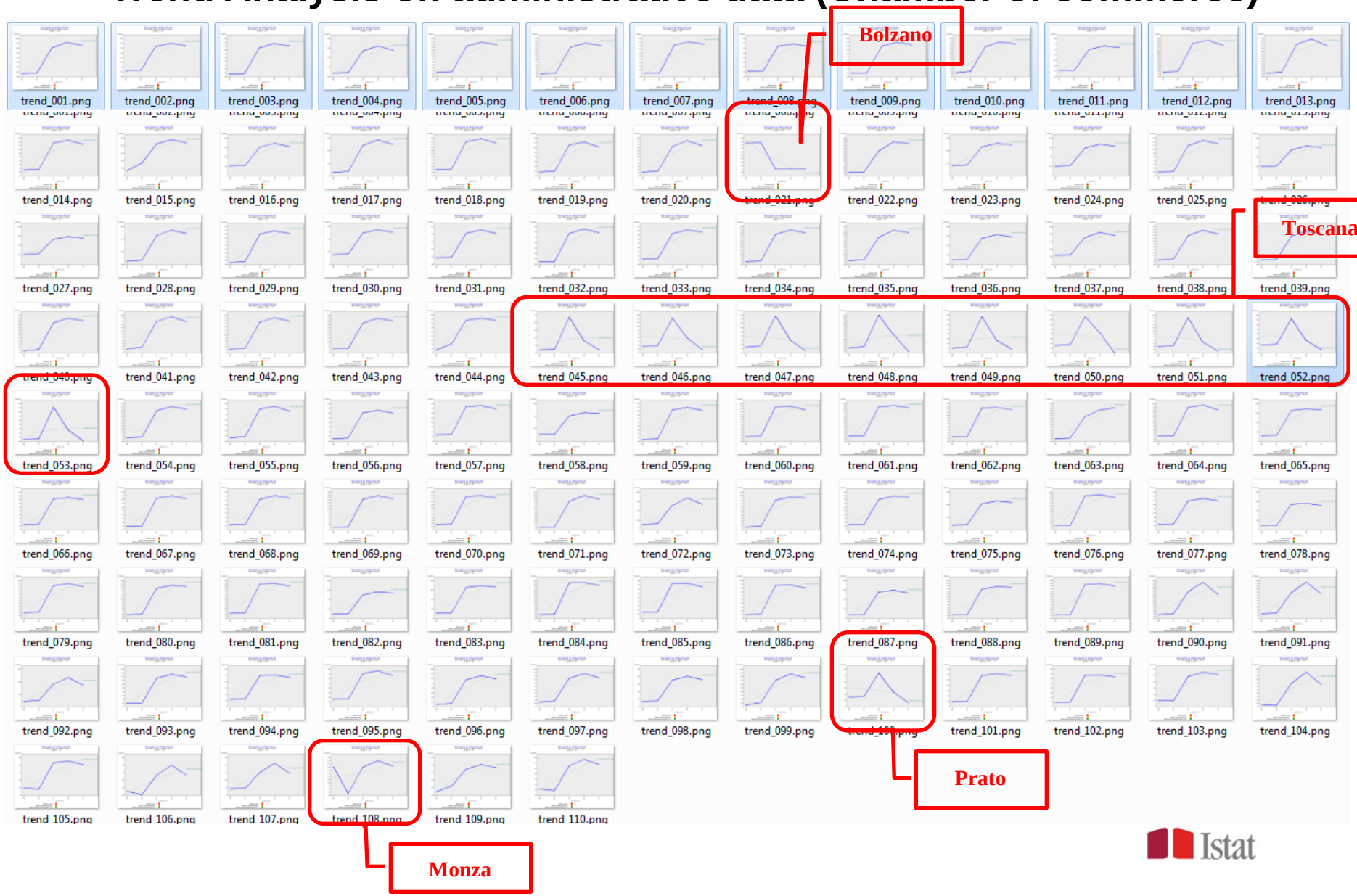
An application has been carried out on the administrative data of the Chamber of Commerce used an input file consists the number of enterprise by provinces and year t (t = 2009, ..., 2013).

***Trend Analysis: An Automated Data Quality Approach for Large Health Administrative Databases**
Mahmoud Azimaee, Institute for Clinical Evaluative Sciences (ICES), Toronto, ON, Canada

Istat

Arcolaio Project – WP3

Trend Analysis on administrative data (Chamber of commerce)



Arcolaio Project – WP4

The main issue of the WP4 is the development of methods for using secondary data into the statistical production.

Different errors (***unit errors, measurement errors***) may occur in the use of administrative data for statistical purposes due to ***partial information, misclassification, measurements errors.***

Multi source scenario

All these kind of errors are treated into a multi source scenario.

The *estimations* are based *on multiple data sources* referring to different but overlapping populations.

Arcolaio Project – WP4

Two possible ***methodological approach*** to produce statistical output based on a ***multi-source information***:

- ***supervised***, one of the measures (typically a survey measure) is considered as error free and the administrative measures are merely used as auxiliary information.
- ***Unsupervised*** - the multiple Administrative and/or Statistical sources provide the value of a same variable of interest for the entire target population or part of it and all the measures are assumed to be imperfect, that is none of them can be considered as a benchmark.

Arcolaio Project – WP4

Supervised

- (*Y*) the error free variable is modeled as a *response* variable
- (*X*) *covariates*.

This supervised approach can be adopted in (1) a model based inference as well as (2) in a design based inference.

The choice of the methodological approach depends both on the informative content and the quality of the available data sources.

The measure of the output quality depends on the chosen inferential framework.

In a more traditional design based approach, where administrative data play the role of auxiliary variables, evaluation of accuracy is primary based on the estimate of sampling error.

Arcolaio Project – WP4

Unsupervised

A **Latent Model approach** is particularly suitable.

Y^* "true" target phenomenon (latent variables)

Y^g ($g=1,..G$) imperfect measures of the target phenomenon (variables observed from G different data source).

X^L and X^M covariates associated to latent Y^* and to the measures Y^g

$$P(Y^* \mid X^L)\,(latent\ model), \qquad (1)$$

$$P(Y^1,..,Y^G \mid Y^*, X^M) \quad (measurement\ model) \qquad (2)$$

$$P(Y^1,..,Y^G \mid X^L, X^M) = \int P(Y^1,..,Y^G \mid Y^*, X^M) P(Y^* \mid X^L) dY^* \ (marginal\ distribution) \ (3)$$

1. Estimation of model parameters using a likelihood approach, based on the data observed from the G different sources.
2. Estimation of the marginal distributions. These distributions can be used to assess the accuracy of each source and the sources can be ranked accordingly.

The estimates of the true value can be obtained in a hierarchical approach, by selecting the source with the highest level of accuracy.

Arcolaio Project – WP4

Latent Model approach

Latent Class Models (LCM) have been used to evaluate the ***measurement errors*** of the variable 'employment status' measured from the following sources:

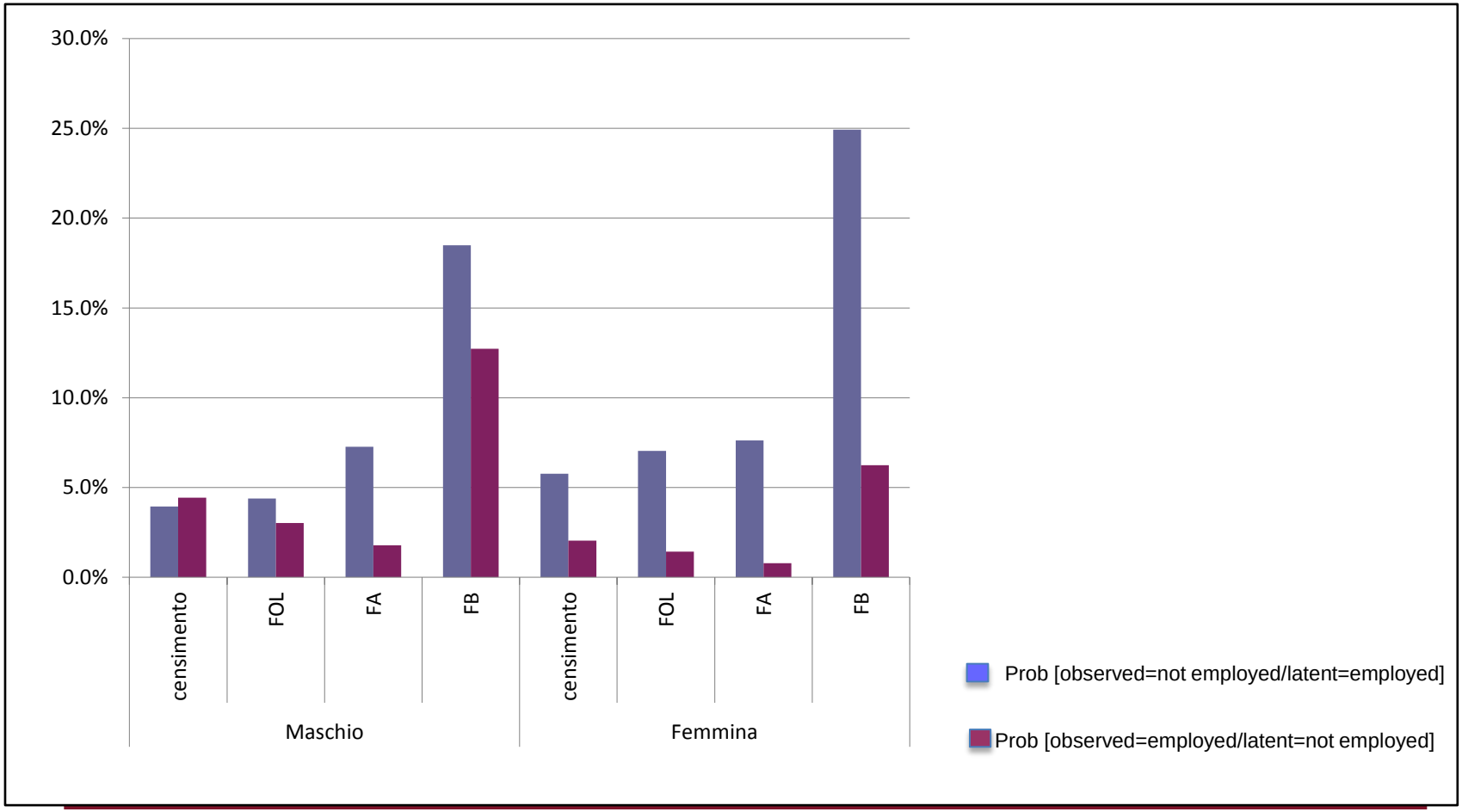
- administrative sources of contributory benefits (expected high quality - FA)
- other administrative sources, e.g., insurance against work-related injurie, Chamber of Commerce (expected low quality - FB)
- Labour Force Survey (FOL)
- Population Census

All sources have been aligned to the time reference of the Population Census

Arcolaio Project – WP4

Latent Model approach

Probability distribution of observed measures subject to the latent class ("true" value)



Istat

Thanks