# Small Area Estimation for mapping local indicators

Monica Pratesi

Caterina Giusti

Department of Economics and Management, University of Pisa

Research Centre 'Camilo Dagum' on Advanced Statistics for the Equitable and Sustainable Development

# Need for local information

Need for a system to produce local, meaningful, 'flash' data and indicators on poverty and vulnerabilities, which are understandable and useful to policy making.

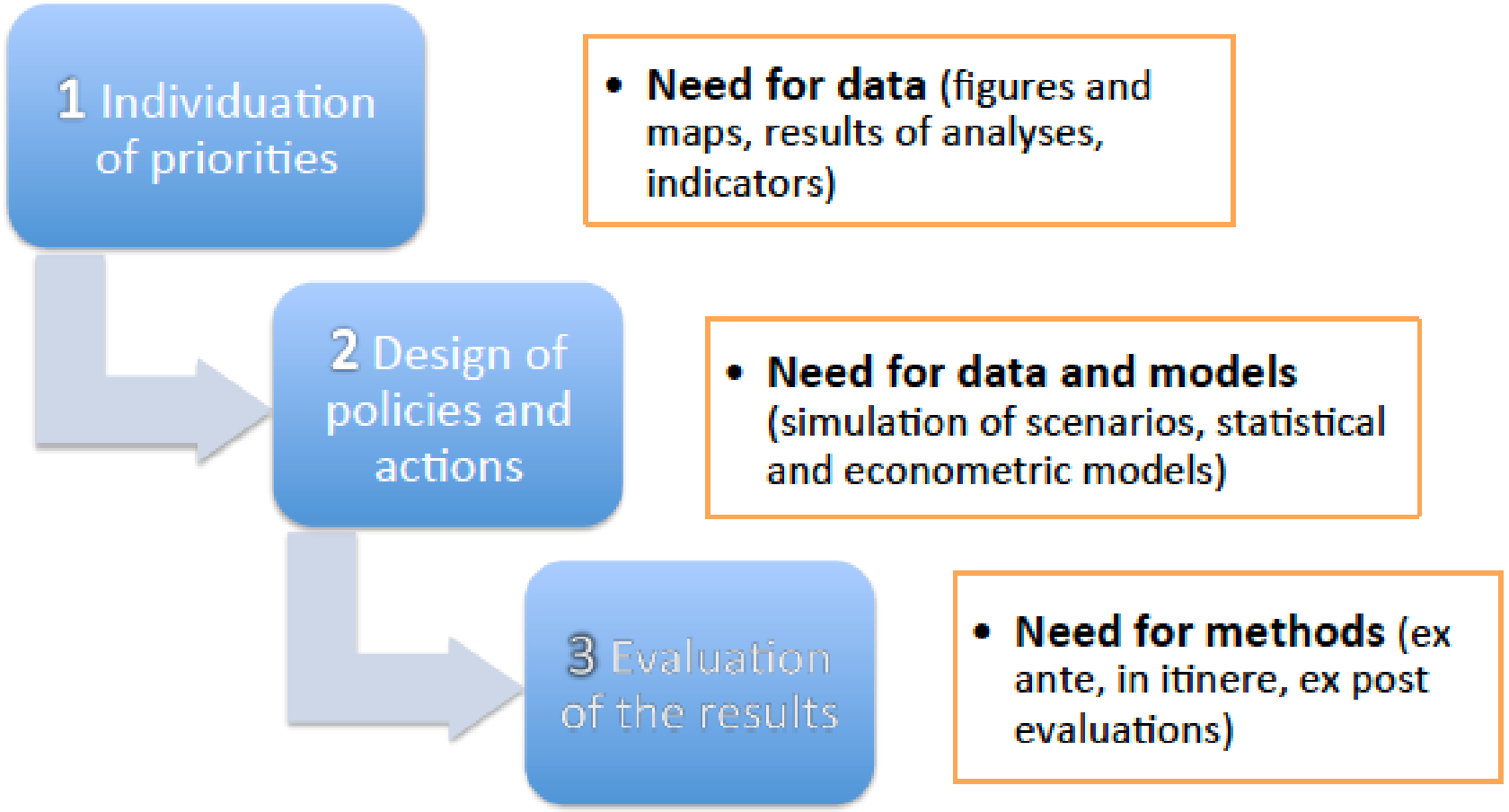
The decision on "what" data and "how" collect is not neutral, but it is a map of reality to be defined given the goal of the policy maker….

…sometimes it is important to have a signal and not an "error-free" estimate!

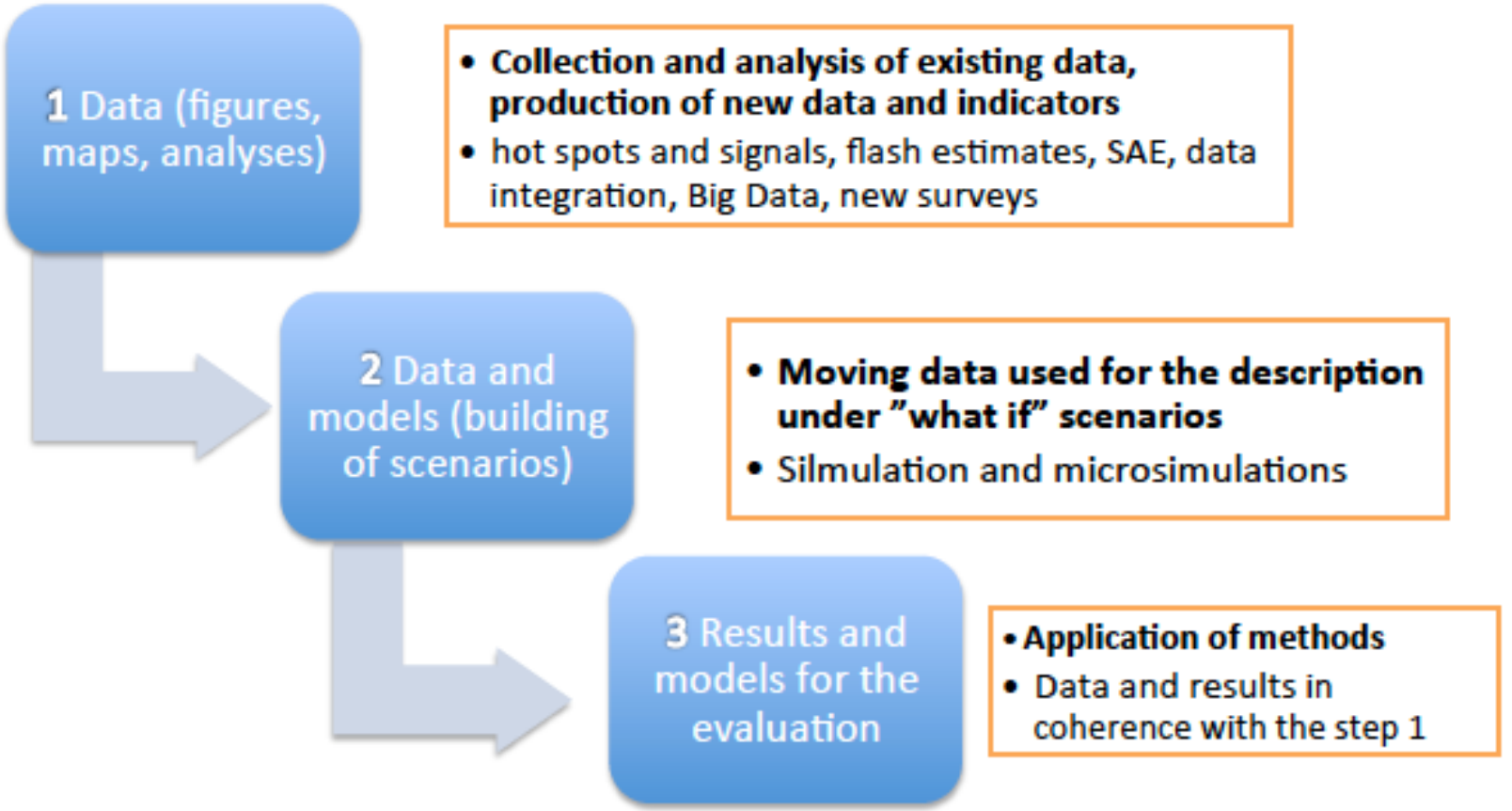
see https://www.makswell.eu/

see http://www.inclusivegrowth.eu/

# Decision Making



# Data production



# SDGs monitoring: data on poverty by Eurostat

Eurostat collects data from a harmonized set of current surveys:

- European Survey on Income and living conditions
- Household Budget survey
- Labour Force survey

Sample surveys are conducted yearly (LFS every trimester) in Member States

# Official Local data in Europe

What does **local** mean?

We need to measure poverty where it matters, in the places where people live.

"Local" is an attribute not defined once forever!

The Classification of Territorial Units for Statistics (NUTS)

French: Nomenclature des unités territoriales statistiques is a geocode standard for referencing.

The subdivisions of countries for statistical purposes.

# Official Local data in Europe

There are three levels of NUTS defined, with two level of LAUs (Local Administrative Units) below.

Note that not all countries have every level of division, depending on their size.

One of the most extreme cases is Luxembourg, which has only LAUs; the three NUTS divisions each correspond to the entire country itself.

# NUTS 1

Italy: groups of Regions 5

Germany: States 16 (Bundesland)

# NUTS 2

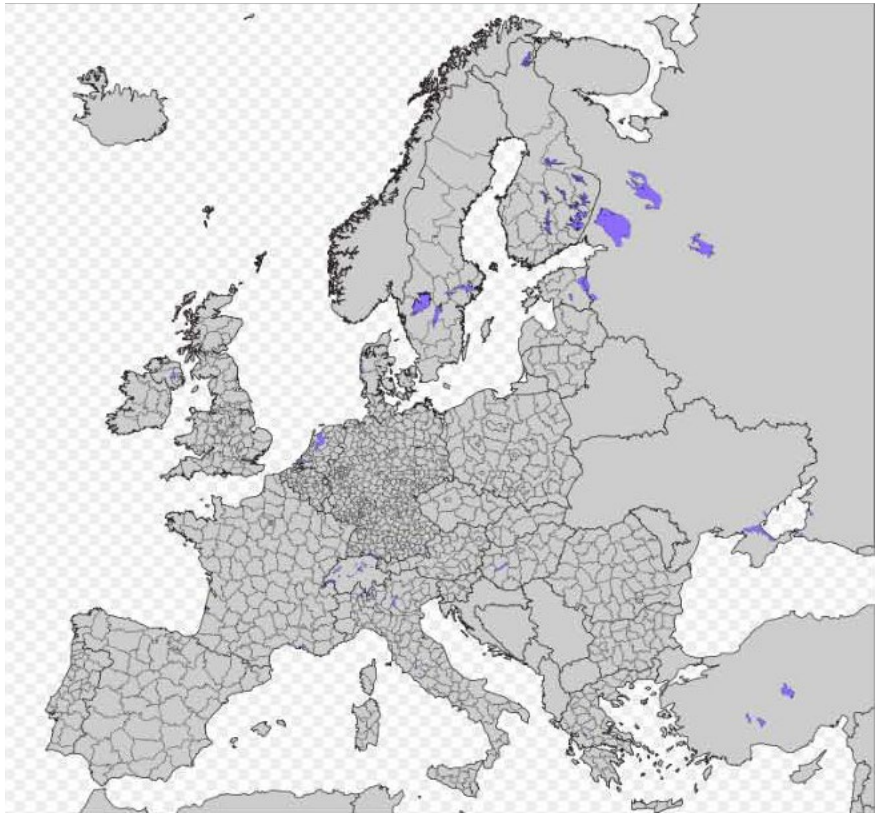
Italy: Regions 21

Germany: Government Regions 39 (Regierungbezirk)

# NUTS 3

Italy: Provinces 110

Germany: Districts 429 (Kreis)



# Local areas - NUTS3 - “small areas”

Eurostat publishes estimates at NUTS1 and NUTS2 level

These are **direct estimate**: estimate coming directly from a sample survey – design-based estimate from sample data

**Small area = domain of interest, for which the sample size is not adequate to produce reliable (accurate) direct estimates** – in EU lower than NUTS2 level

DEGURBA (EUROSTAT) creates a classification of all LAU2s into the following three categories:

- Cities (densely populated areas) (Code 1)
- Towns and suburbs (intermediate density areas) (Code 2)
- Rural areas (thinly populated areas) (Code 3)

# Accuracy

- The accuracy of statistical information is the degree to which the information correctly describes the phenomena it was designed to measure (given the measurement is valid).

- It is usually characterized in terms of error in statistical estimates and is traditionally decomposed into bias (systematic error) and variance (random error) components.

Simler K. (2016). Pinpointing Poverty in Europe: New Evidence for Policy Making. World Bank, Washington, DC

# QUESTIONS?

# Estimation: scope and purpose

Estimation is a process that approximates unknown population parameters using only that part of the population that is included in a sample.

Examples of parameters:

- simple descriptive statistics: totals, means,
- ratios and percentiles,
- complex statistics: poverty indicators
- analytical statistics: regression coefficients

# Official Local data in Europe

Proper estimation conforms to the sampling design. Sampling weights are incorporated in the estimation process (stratification, clustering, and multi-phase or multi-stage information).

Use auxiliary data whenever possible to improve the reliability of the estimates. Evaluate the use of the auxiliary data.

# Estimation: small area estimators

Statistics Canada suggestion:

> "Incorporate the requirements of small domains of interest at the sampling design and sample allocation stages (Singh, Gambino and Mantel, 1994). If this is not possible at the design stage, or if the domains are only specified at a later stage, consider special estimation methods (small area estimators) at the estimation stage. These methods *"borrow strength"* from related areas (or domains) to minimize the mean square error of the resulting estimator (Platek et al., 1987; Ghosh and Rao, 1994; Rao, 1999)."

# What's SAE?

**Small Area Estimation (SAE)** is a methodology for producing estimates for a more detailed level of geography than can be reliably obtained from direct survey estimates.

Conceptually similar to these are small domain estimates, which are disaggregated to finer classification levels (e.g. industry, income group or labor force status)

# What's SAE?

SAE combines the use of **survey data and auxiliary data sources** such as administrative data

SAE results are **new** statistics that are not otherwise available from survey or administrative data sources.

Watch out: some administrative data also can be used to produce statistics for small areas, and also Big data sources - Accuracy

# What's SAE?

SAE analytical methods may have a crucial role for producing official statistics:

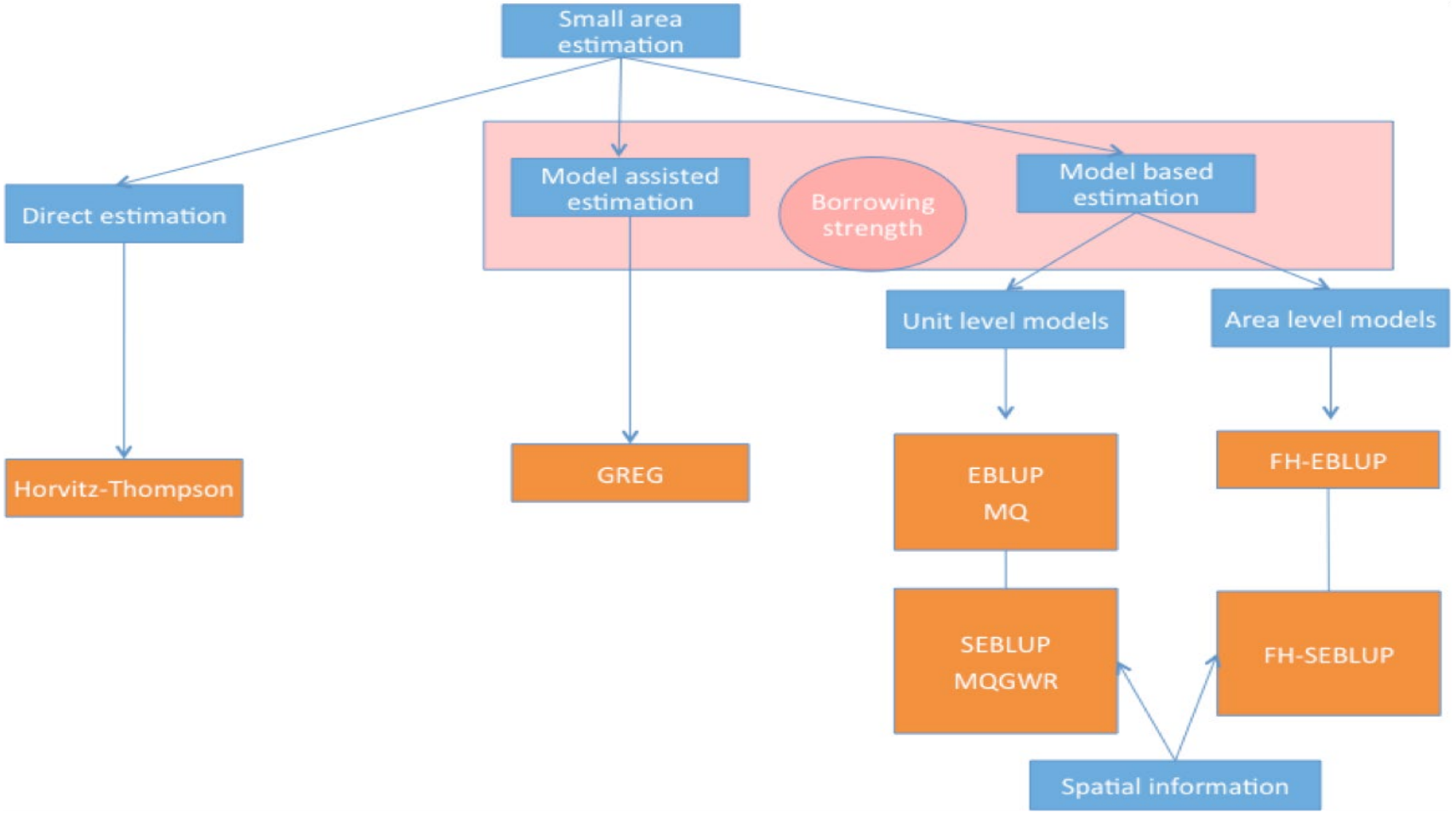
to ensure methods and assumptions are described for users

the **validity** of the modelled estimates are to be assessed

# Australian Bureau Statistics suggestion

"The choice of small area method depends on the **availability** of auxiliary data and the **relationship** between these data and the variables of interest at the small area level. In essence, we are looking to "borrow strength" from these auxiliary data to increase the accuracy of the estimates. Small area models range from the simple to the more complex, the latter requiring considerably more **time, effort, technical skill** and **available data**. A range of quantitative and qualitative **diagnostics** should be used to choose the best model for the given data."

# A classification of the SAE methods



# Statistical quality of SAE

- The **timeliness** of statistical information refers to the delay between the reference point (or the end of the reference period) to which the information pertains, and the date on which the information becomes available.

- SAE are timely and cost-effective flash estimates!!!

# Statistical quality of SAE

The **accessibility** of statistical information refers to the ease with which it can be obtained from the Agency.

SAE it often offered through maps (poverty mapping): questionable medium of accessibility

# Statistical quality of SAE

- The **interpretability** of statistical information reflects the availability of the supplementary information and metadata necessary to interpret and utilize it appropriately.

- SAE requires special metadata (model fitting, geography used)

# Statistical quality of SAE

- The **coherence** of statistical information reflects the degree to which it can be successfully brought together with other statistical information within a broad analytic framework and over time

- SAE requires special attention on this dimension(calibration, benchmarking, harmonization with other estimates)

# What makes for a successful small area study

- User commitment and client interaction – ability to work closely with users
- Variable(s) of interest - the variables of interest should be a reasonably common population characteristic.
- Population size of the small area - when small areas contain some sample, even if inadequate for accurate direct estimation, the modelled estimates will be more reliable (if the model fits!).
- Auxiliary data - the availability of administrative, census or other survey data with a significant relationship to the variable of interest is crucial

# Assessing the quality of small area estimates

Choice of the model for SAE

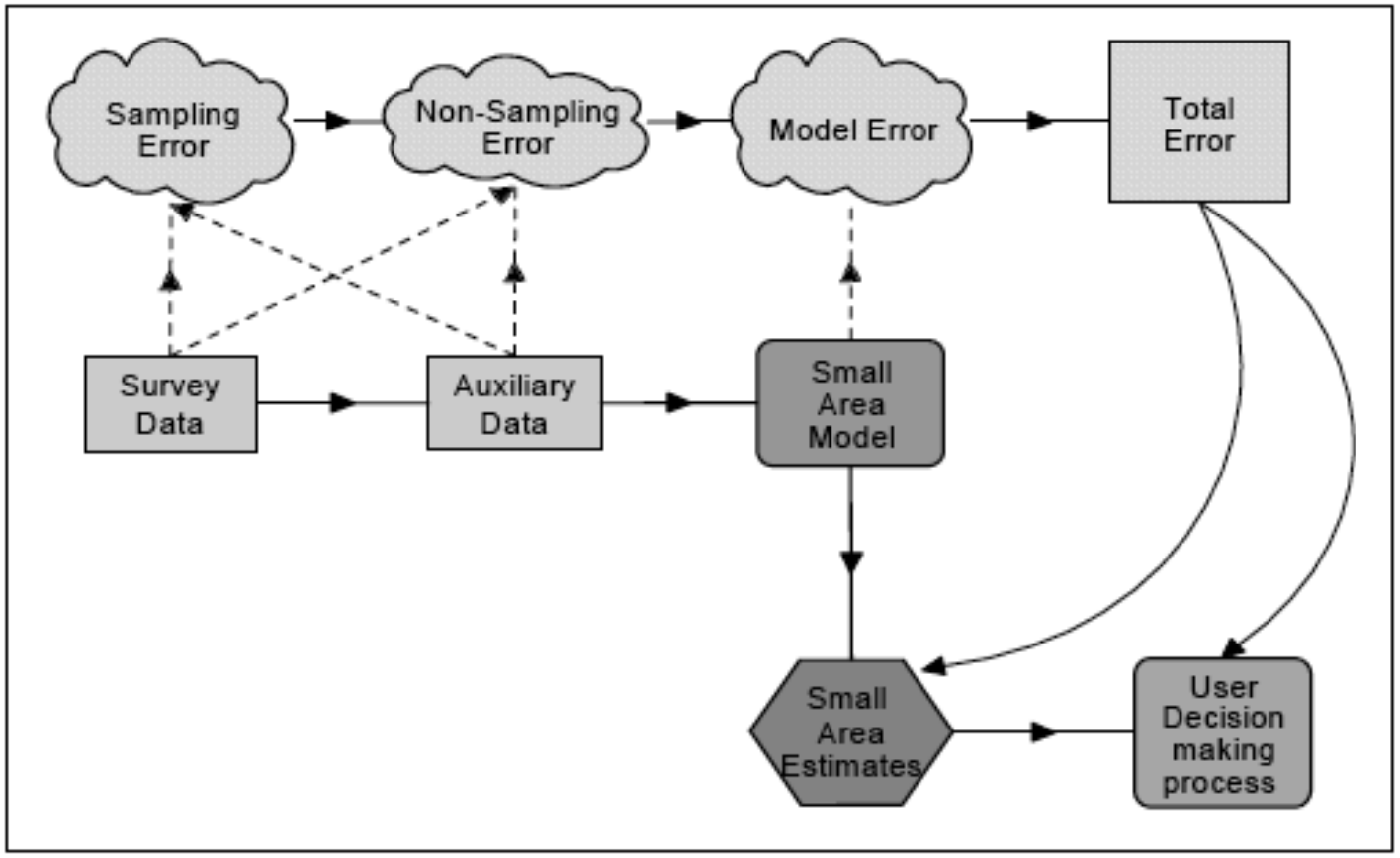
- plausibility of the model in light of previous studies or accepted wisdom;
- how well the model fits the observed data;
- accuracy of the small area estimates predicted from the model.

# Assessing the quality of small area estimates

Diagnostics:

- a bias test that compares the small area predictions with direct estimates;
- testing whether model assumptions are met and that the model is a good fit;
- checking that small area estimates add to published state or national estimates;
- local knowledge and expert advice on the spread of estimates across small areas;
- relative root mean squared errors (RMSE) – analogous to sampling errors calculated for survey estimates.

# Quantifying the quality of small area estimates



# Documenting results

- Underlying problem, scope and applicability of the estimates;
- small area estimation procedure (the specific model used, variables included, main assumptions, etc.);
- quality issues specific to different sets of small area estimates;
- guidelines on how to use the small area output;
- a summary of key issues and recommendations (e.g., aggregation of small area estimates, the need for local knowledge, etc.).

# Documenting results

- Models used plus their plausibility, validity and goodness of fit;
- how each set of small area estimates performed against specific quality diagnostics;
- other quality issues (sensitivity of the spatial model (if any), Modifiable Area Unit Problem, shrinkage effect, robustness against outliers, treatment of zero values in the study variable).

# Summary and conclusions

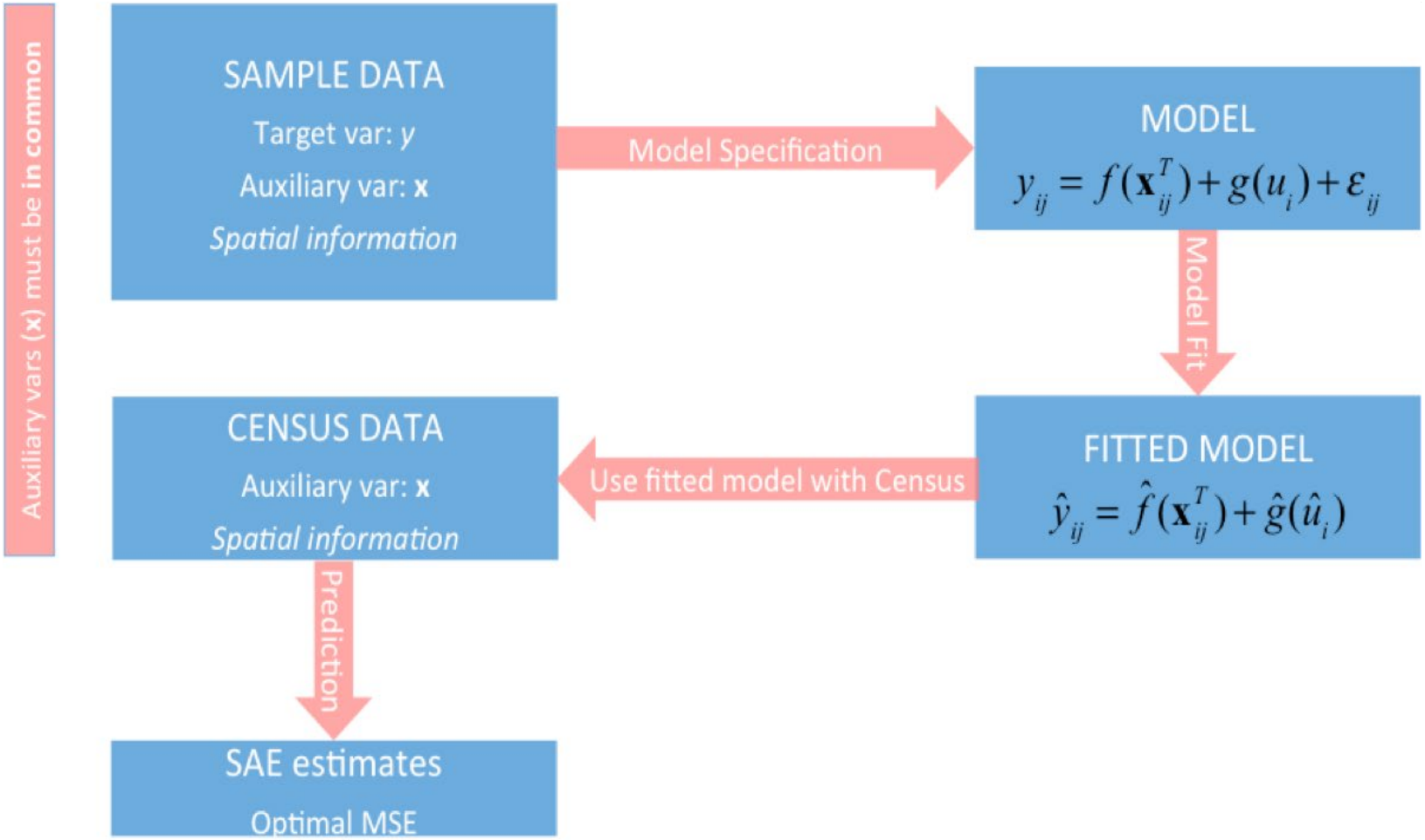
- Careful assessment of SAE
- Communication with stakeholders.
- Relevant auxiliary variables.
- Documentation on models, quality of the outputs.

# QUESTIONS?

# How a SAE Unit Level Model Works

- Survey data: available for the target variable y and for the auxiliary variable *x*, related to *y*
- Census/Administrative data: available for *x* but not for *y*
- Use survey data to estimate models that link y to *x*
- Combine the estimated model parameters with *x* for out of sample units, to predict the *y* values
- Use these predictions to estimate the target parameters (e.g. area totals or means)

# How a SAE Unit Level Model Works



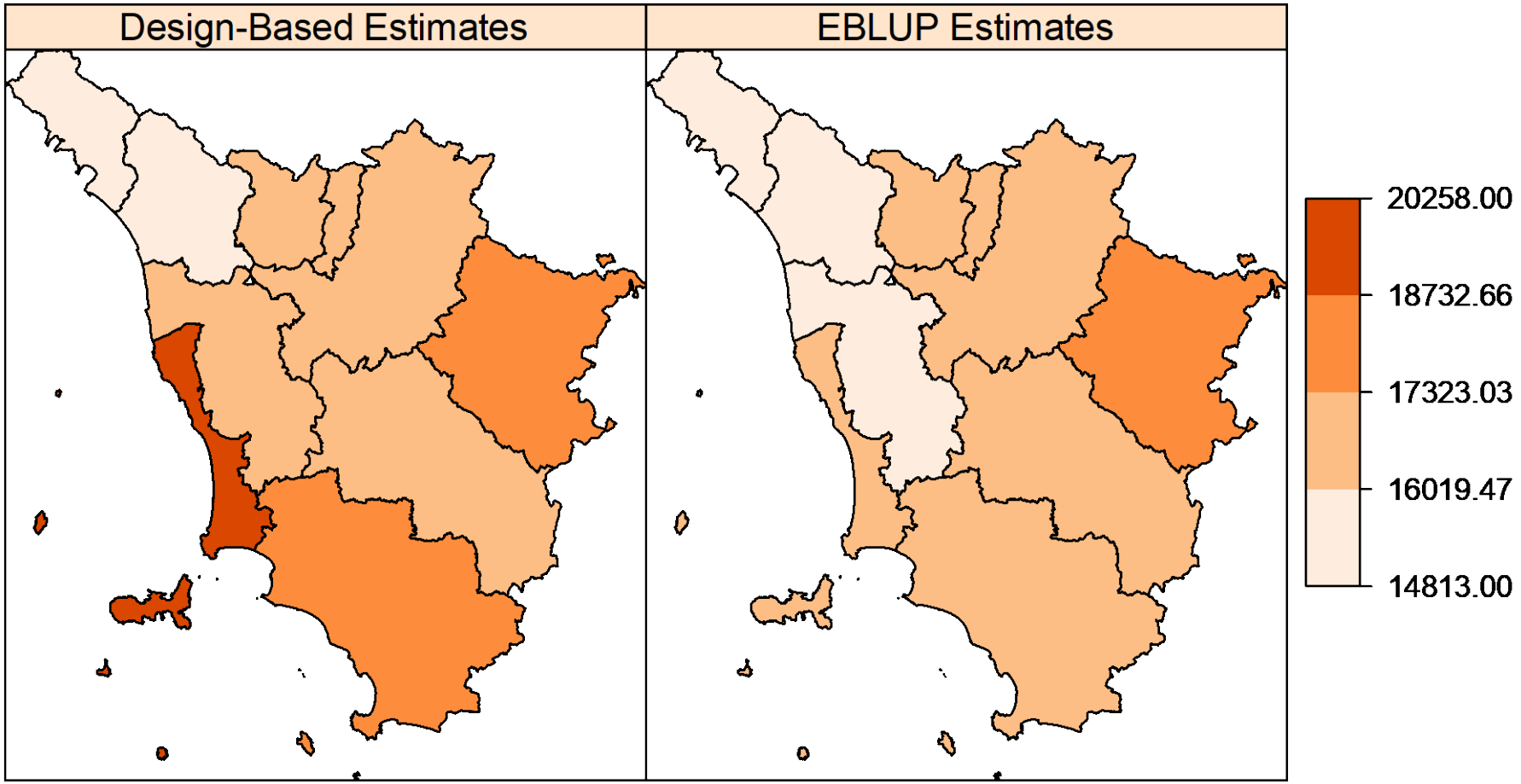
# Example: unit level EBLUP

- Data on the equivalised income in 2015 for 1525 households in the 10 Tuscany Provinces are available from the EUSILC survey 2016
- A set of explanatory variables is available for each unit in the population from the Population Census 2011
- We employ the EBLUP unit level model to estimate the mean of the household equivalised income
- The Municipality of Florence, with 125 units out of 457 in the Province, is considered as a stand-alone area

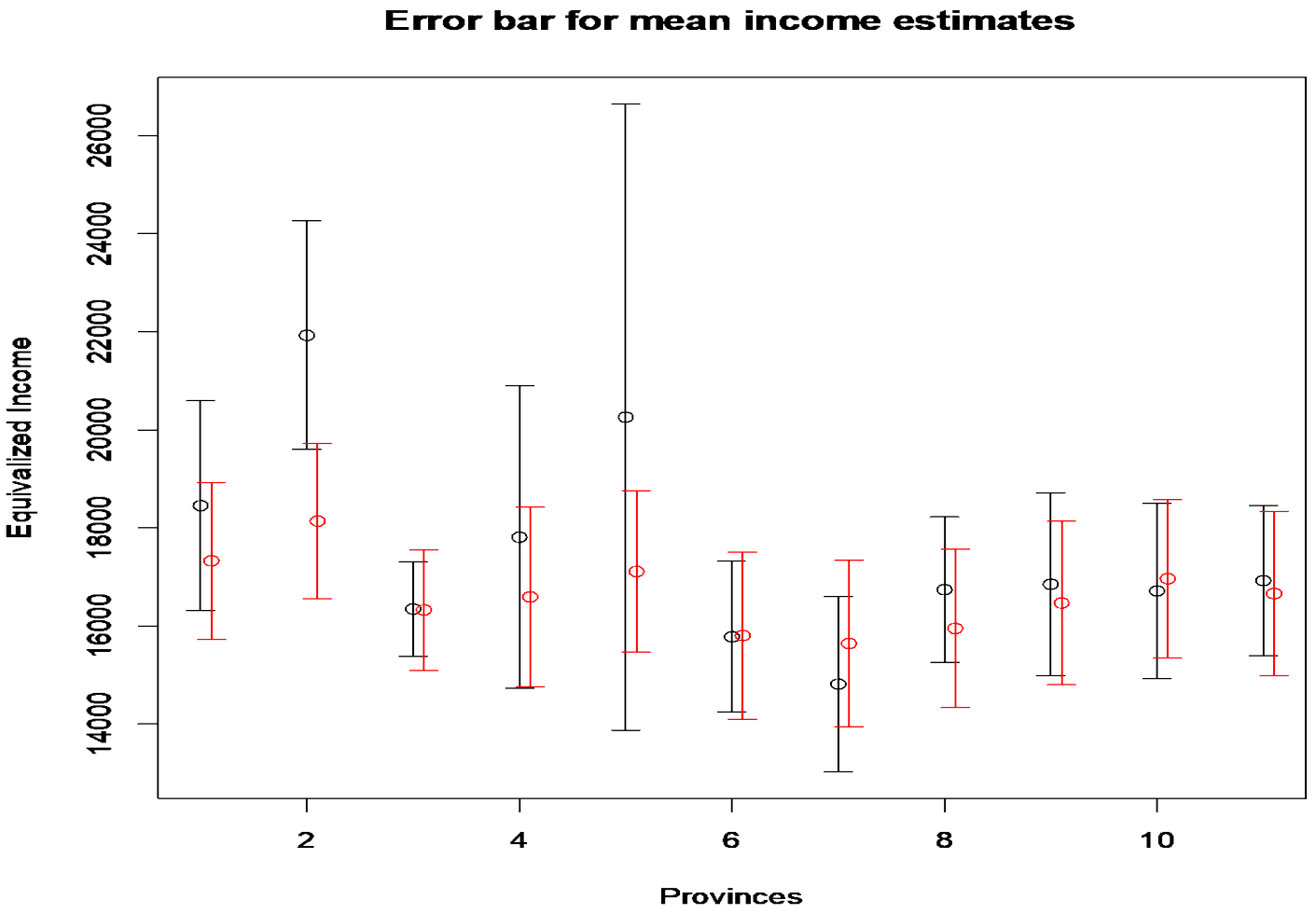
# Example: unit level EBLUP

- The selection of covariates to fit the small area model relies on prior studies on poverty assessment
  - The following covariates have been selected:
  - household size
  - ownership of dwelling (owner/tenant)
  - age of the head of the household
  - years of education of the head of the household
  - working position of the head of the household (employed/unemployed in the previous week)
- Design-based estimates of the mean income have been carried out in order to show the gain in efficiency of the EBLUP

# Example: unit level EBLUP



# Example: unit level EBLUP



# How a SAE Area Level Model Works

- Survey data: direct estimates of the target parameter are computed for each area
- Census/Administrative data: information is summarized at the area level to be used as auxiliary information
- Use a model to link the area direct estimated to the covariates, including area-specific random effects
- The area EBLUPs based on this model will be highly correlated to the direct estimates and will have a lower MSE

# How a SAE Area Level Model Works

The area level model includes random area-specific effects and area specific covariates $x_i$

$$\theta_i = x_i\beta + z_i u_i, i = 1, \dots, m$$

- $\theta_i$ is the parameter of interest (e.g. totals or means)
- $Z_i$ are known positive constant
- $u_i$ are independent and identically distributed random variables with mean 0 and variance $\sigma_u^2$ $\left(u_i \sim N(0, \sigma_u^2)\right)$
- $\beta$ is the regression parameters vector

# How a SAE Area Level Model Works

Assumption

$$\hat{\theta}_i = \theta_i + e_i$$

- $\hat{\theta}_i$ is a direct design-unbiased estimator
- $e_i$ are independent sampling error with mean 0 and know variance $\psi_i^2$

Fay-Harriot Model

$$\hat{\theta}_i = x_i\beta + z_i u_i + e_i, i = 1, \dots, m$$

Note: this is a special case of the general linear mixed model with diagonal covariance structure

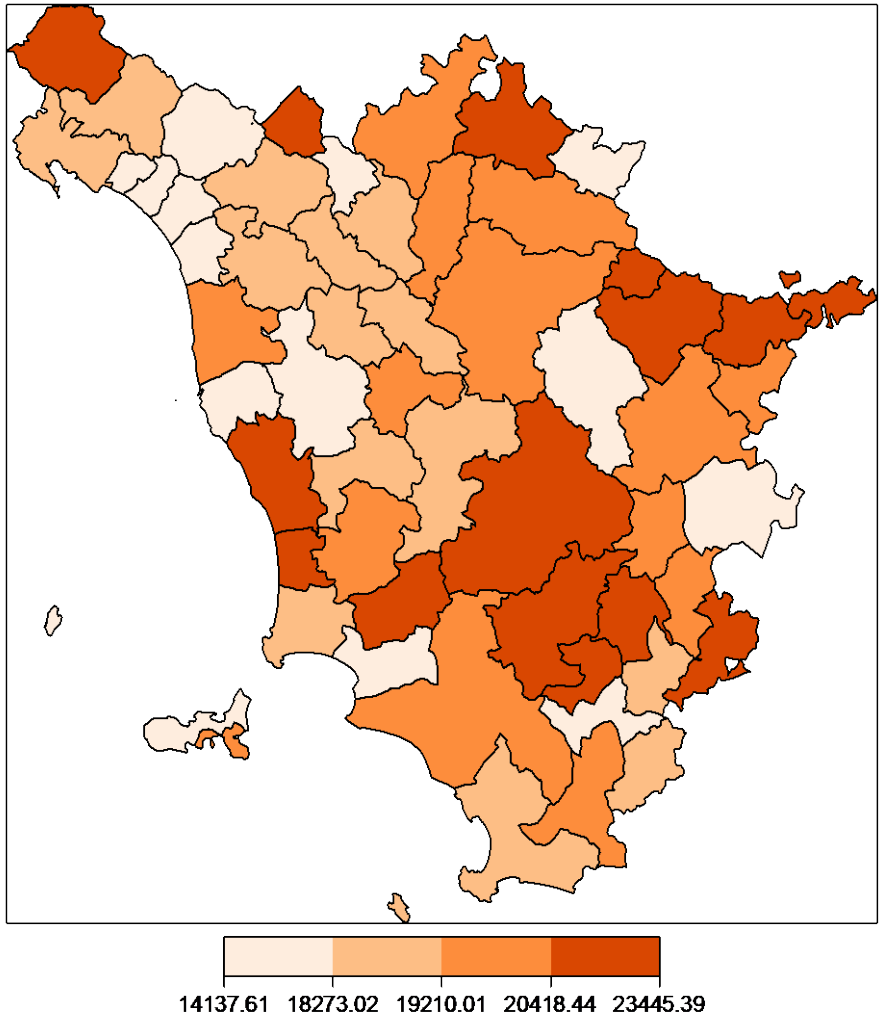
# Example: Spatial area level EBLUP

- Aim: estimate the mean of the household equivalised income for the 57 Local Labour Systems (LLSs) of the Tuscany region, Italy
- Data on household income from the 2011 wave of Italian EU-SILC survey.
- LLSs are defined as a collection of contiguous municipalities that are supposed to form a single labour market, similar to travel-to-work areas used in other countries (intermediate between LAU 1 and LAU 2 levels)
- 24 out of the 57 LLSs of Tuscany are out-of-sample areas n the EU-SILC

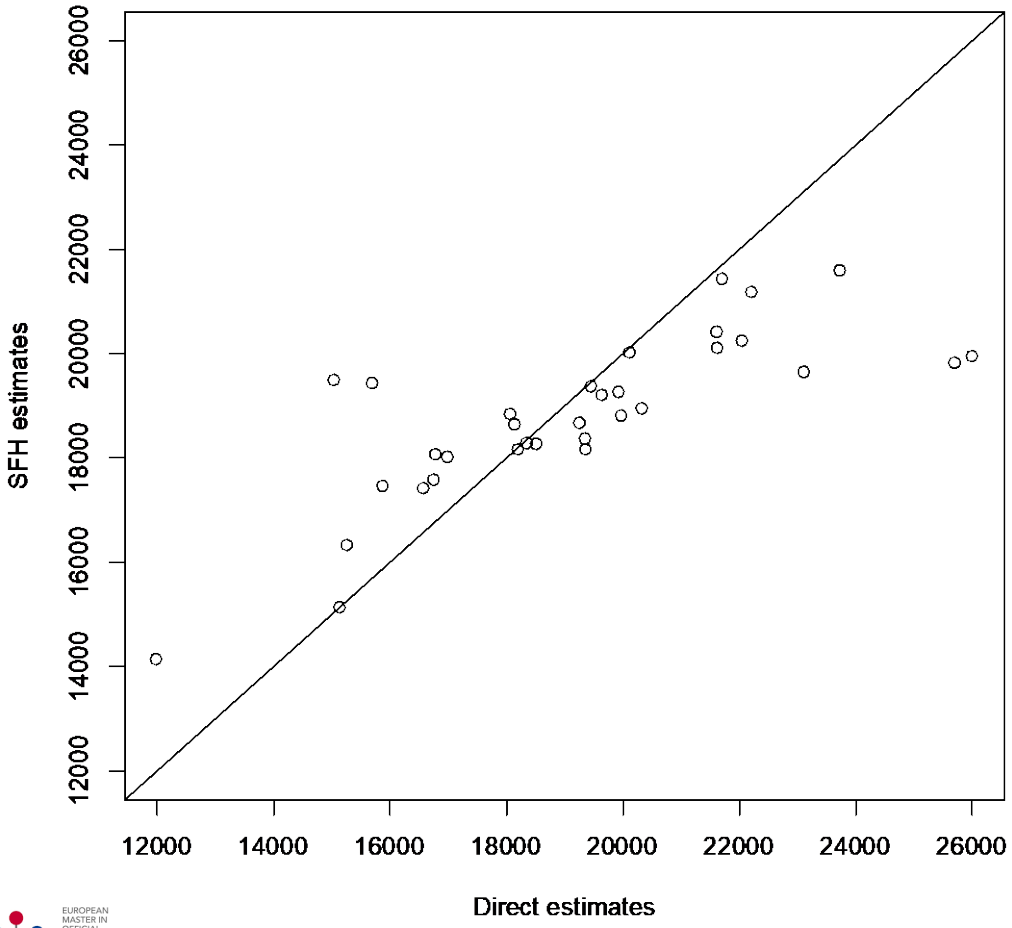
# Example: Spatial area level EBLUP

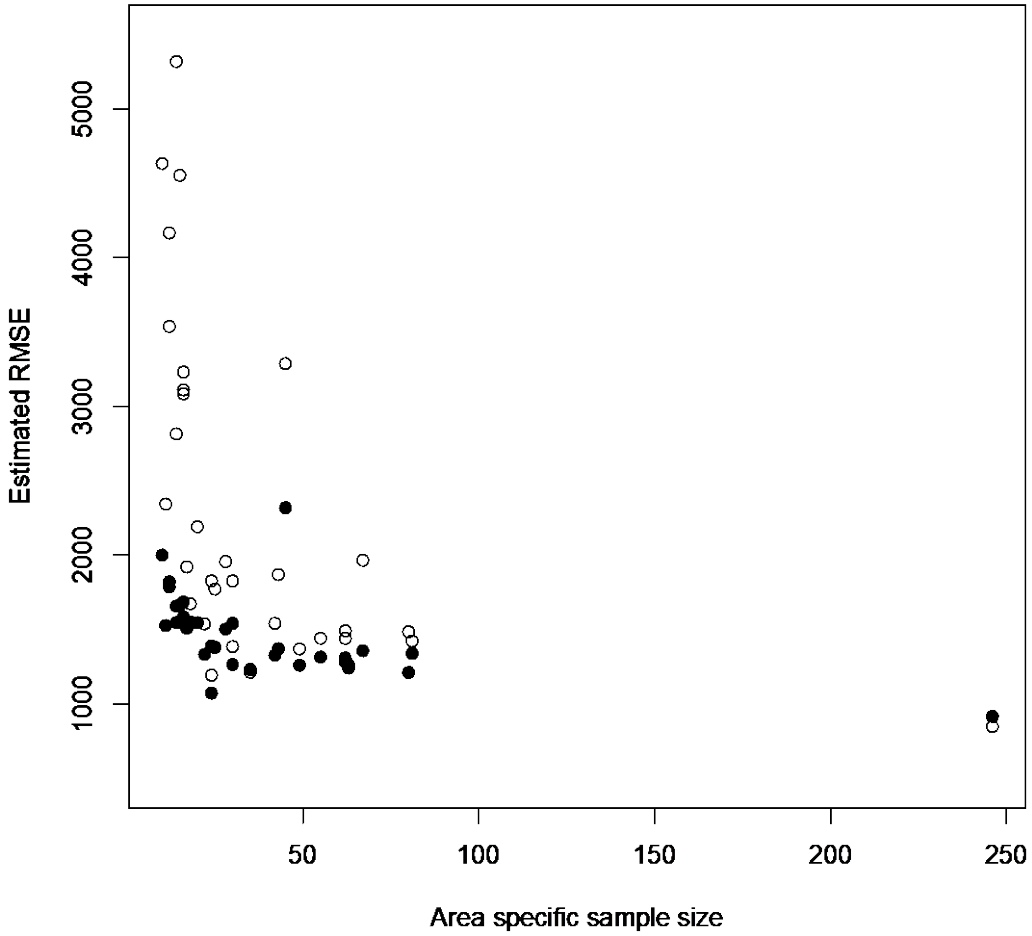
- Covariates from the Population Census 2011:
  - proportion of males aged 15-24 with low educational level
  - proportion of males aged 25-34 with low educational level
  - proportion of non-Italian males aged 25-34
  - proportion of unemployed males aged 34-65
- Standard regression model R2 equal to approximately 70%
- We used a Spatial FH model

# Example: Area level SEBLUP



# Example: Area level SEBLUP



# Local estimates of Educational Poverty

- The definition of EP dimensions and its measures are not completely developed:
  - EP has been considered as deprivation of the ability to learn, ex- periment, develop and freely ourish skills, talents, and aspirations (Watkins, 2000; Save The Children, 2018);
  - The Italian National Statistical Institute (ISTAT) (Quattrociocchi, 2018) propose a multidimensional Educational Poverty Index (EPI) that measures a mixture of problems of material, relational, cultural and environmental kind, which can limit the abilities to live in a com- plex society.
- Educational Poverty (EP) is a macro-level phenomenon, reflecting the general difficulty experienced by people in their own places.

# Local estimates of Educational Poverty

- The estimates of EPI (Quattrociocchi, 2018) produced by ISTAT are referred to the 4 macro-regions in Italy (NUTS 1 level in the European classification).
- Taking the degree of urbanisation as a braking variable in the research of the magnitudes of within-country poverty is very meaningful, as the levels, causes and solutions to poverty, and reasonably to EP, are often different in rural and urban areas (Weziak-Bialowolska, 2016).
- Interest in studying EP comparing differences in the suburbs and metropolitan areas

# Local estimates of Educational Poverty

- The degree of urbanisation
- The areas of analysis resulting from the interaction between the 20 Italian Regions (NUTS 2) and the three DEGURBA levels are 59 (no cities in Trentino Alto Adige).

**→ NUTS2 × DEGURBA → Small Areas**

# Local estimates of Educational Poverty

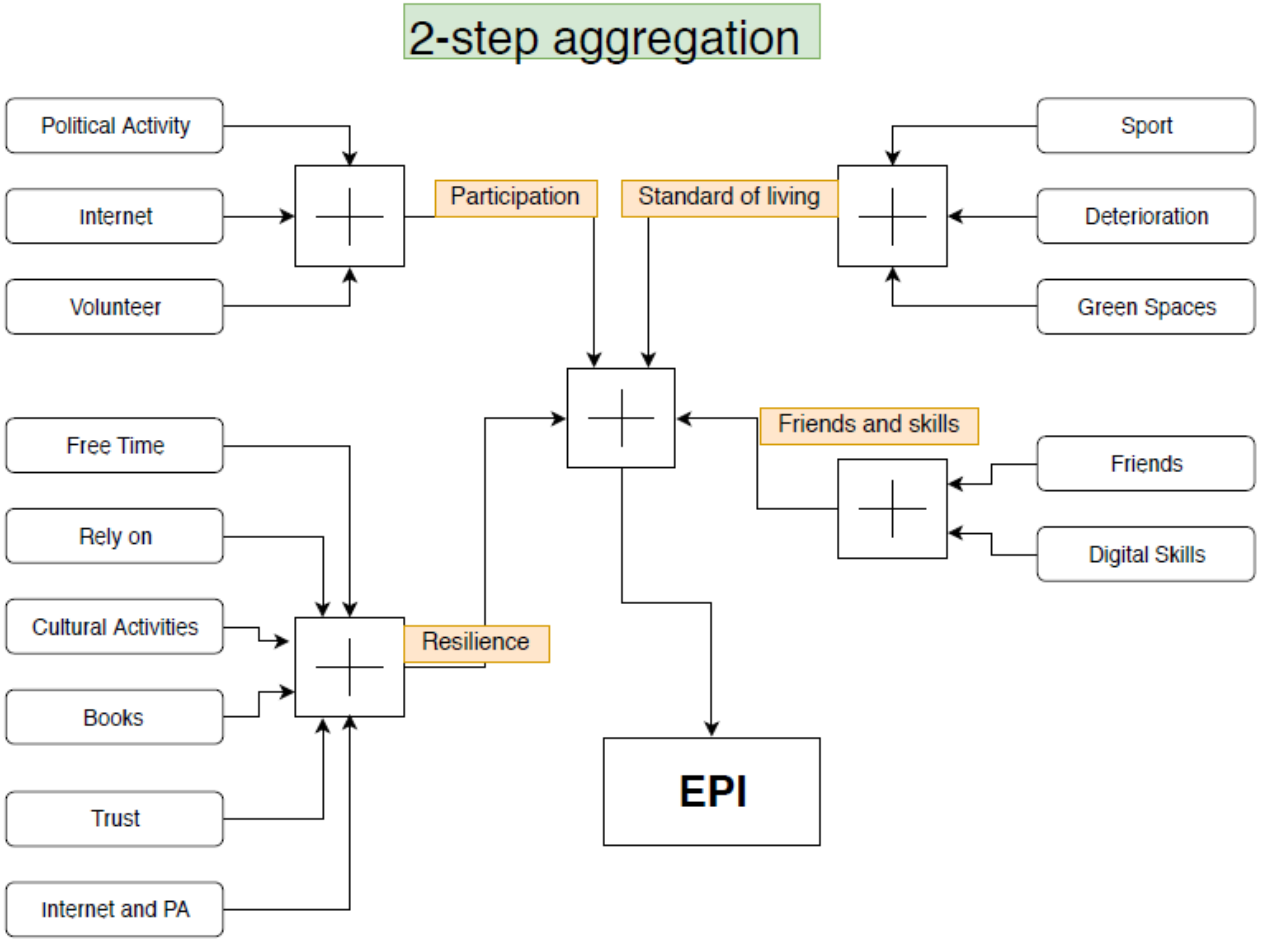
- EP appears as a latent concept with many dimensions to measure
- EPI considers 4 dimensions:
  - Participation to represent the participation of youngsters to the social life;
  - Resilience to represent the development of an attitude of trusting oneself and one's abilities;
  - Standard of living, to represent the ability to lead an inclusive, healthy and safe life having an adequate standard of living;
  - Friends and skills, to represent the ability to wave relationship with others and to achieve those skills needed to succeed in such a fast-pacing world

# Local estimates of Educational Poverty

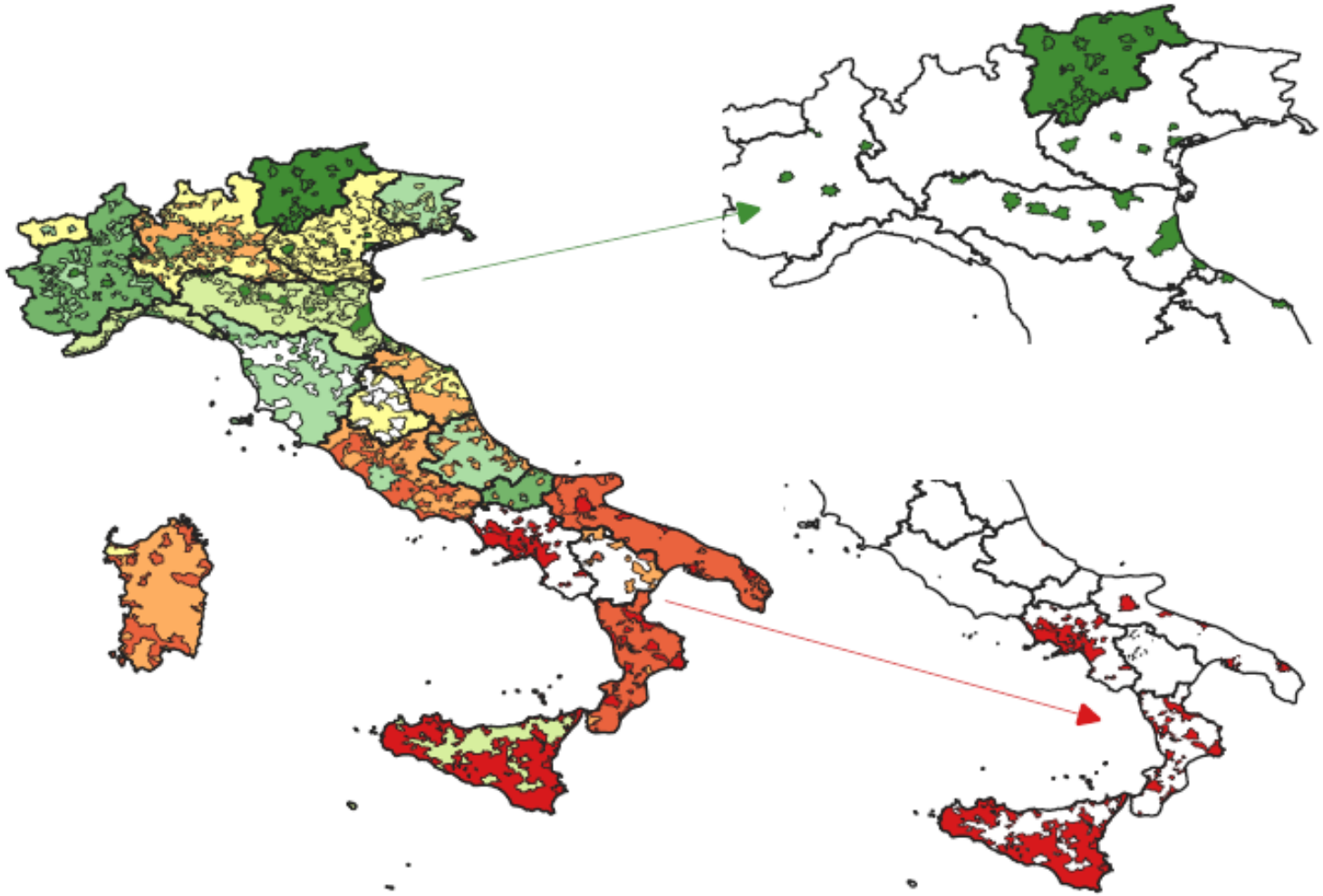
- The single dimensions were measured by indicators obtained by the sample survey on Aspects of Everyday Life (AVQ) 2016:
- ≈50,000 individuals and ≈ 20,000 households;  focus on individuals aged 15-29
- The AVQ survey is planned to obtain precise estimate at regional level

  → Small Area Estimation (SAE) is needed to obtain estimates for the desired unplanned domains.

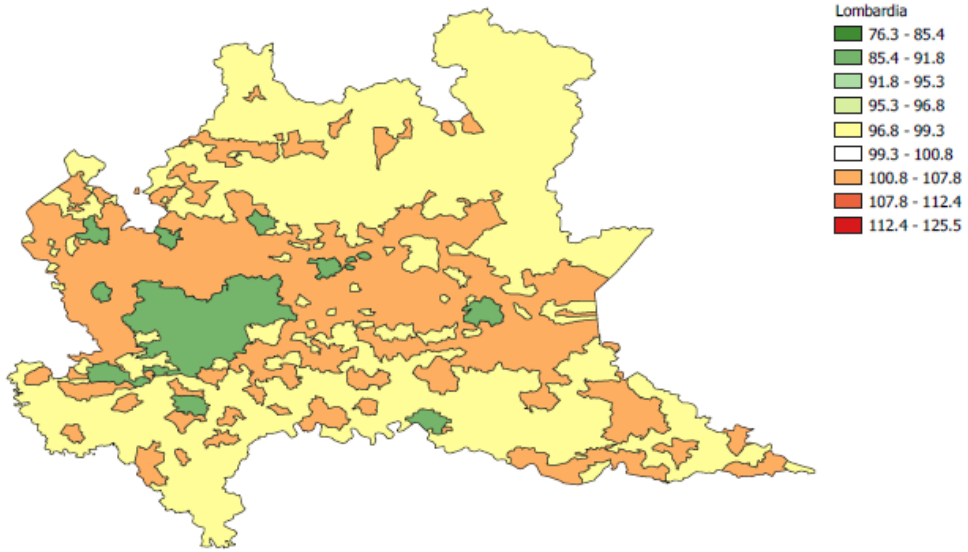
# Local estimates of Educational Poverty

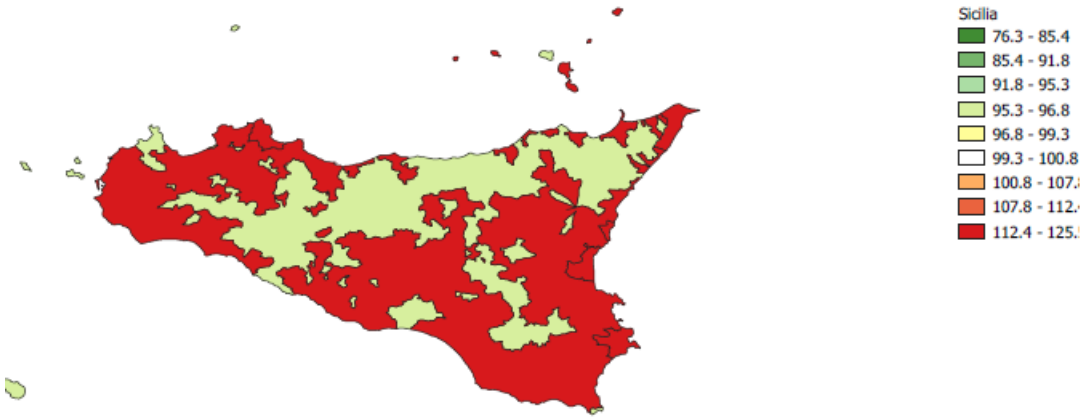


# Local estimates of Educational Poverty



# Local estimates of Educational Poverty



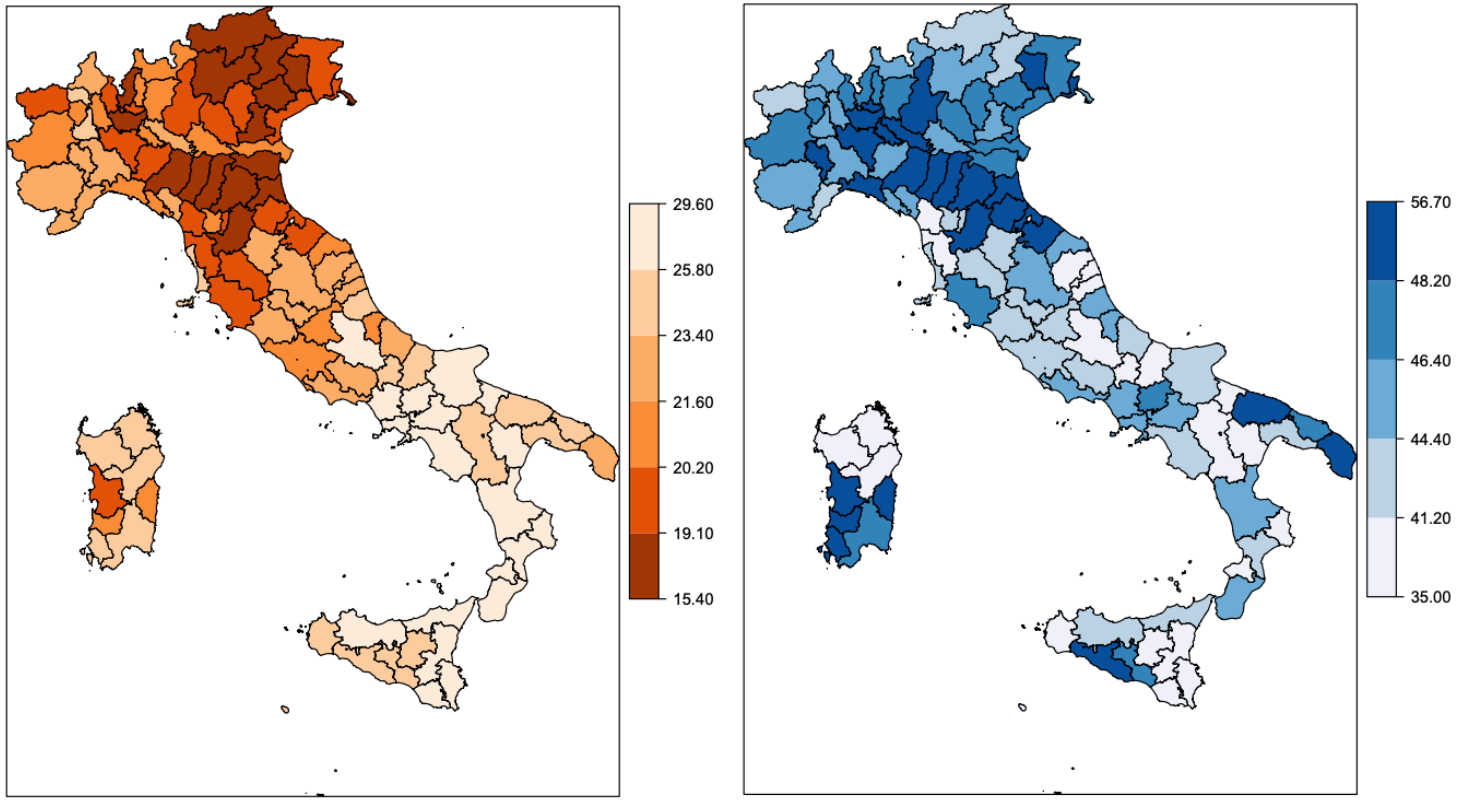



# Local estimates of Educational Poverty

| | Direct | | | FH | | |
|---|---|---|---|---|---|---|
| Indicator | $\leq 16.6$ | $16.6 - 33.3$ | $\geq 33.3$ | $\leq 16.6$ | $16.6 - 33.3$ | $\geq 33.3$ |
| Political | 21 | 26 | 12 | 59 | 0 | 0 |
| Internet | 3 | 31 | 25 | 47 | 12 | 0 |
| Volunteer | 19 | 28 | 12 | 59 | 0 | 0 |
| Free Time | 6 | 33 | 20 | 56 | 3 | 0 |
| Rely on | 0 | 20 | 39 | 0 | 36 | 23 |
| Cultural | 17 | 29 | 13 | 59 | 0 | 0 |
| Books | 19 | 30 | 10 | 59 | 0 | 0 |
| Trust | 19 | 27 | 13 | 59 | 0 | 0 |
| PA | 19 | 28 | 12 | 59 | 0 | 0 |
| Sport | 13 | 32 | 14 | 59 | 0 | 0 |
| Deterioration | 16 | 31 | 12 | 58 | 1 | 0 |
| Green | 4 | 23 | 32 | 13 | 27 | 19 |
| Friends | 0 | 17 | 42 | 2 | 41 | 16 |
| Digital | 6 | 33 | 20 | 58 | 1 | 0 |

Number of Areas with CVs of direct and FH estimates ≤ 16.6, between 16.6 and 33.3 and ≥ 33.3

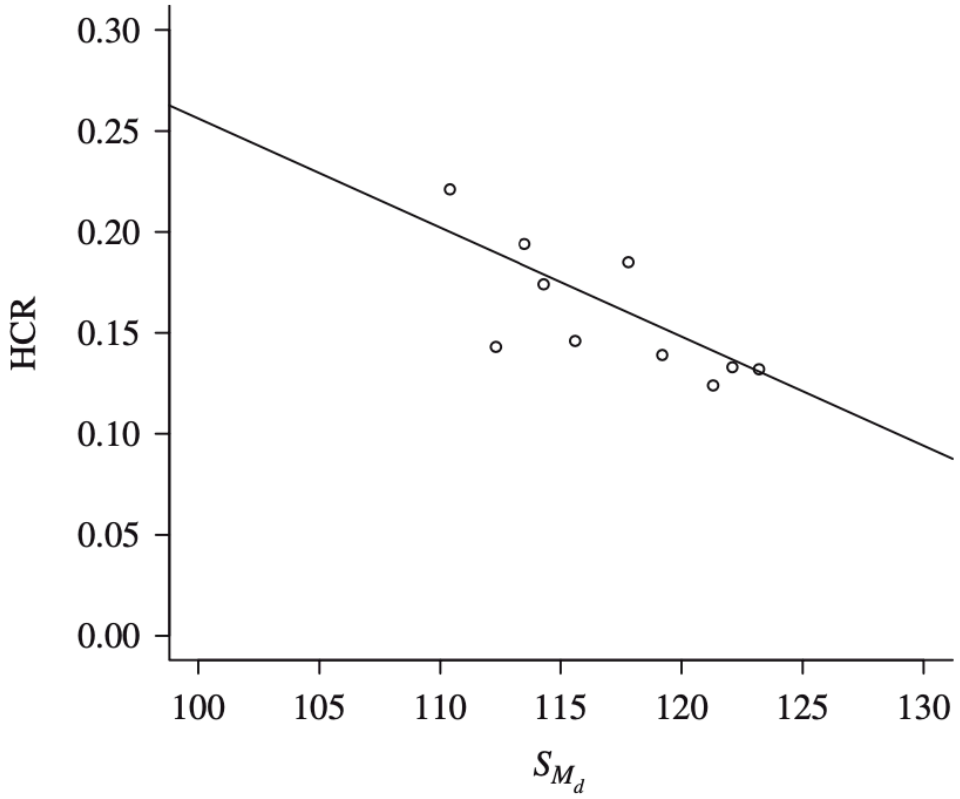
# Use of Big Data in SAE




(a) SFCE estimates (b) iHappy

Map of the FH estimates of the SFCE (a) and map of the iHappy index (b) for 110 provinces in Italy. In both the maps a darker color corresponds to a better situation.

# Use of Big Data in SAE




Scatterplot of the standard deviation of a mobility index based on GPS data vs. estimates of the HCR at province level in Tuscany, Italy.

# Projects

**MAKing Sustainable development and WELL-being frameworks work for policy analysis Project**

https://www.makswell.eu/

**InGRID-2 Integrating Research Infrastructure for European expertise on Inclusive Growth from data to policy**

http://www.inclusivegrowth.eu

**Jean Monnet Chair Small Area methods for Multidimensional Poverty and living conditions Indicators in EU – SAMPIEU**

http://sampieuchair.ec.unipi.it/

# References


Australian Bureau of Statistics (2005). A Guide to Small Area Estimation

Beręsewicz M., Lehtonen R., Reis F., Di Consiglio L., Karlberg M. (2018). An overview of methodsfor treating selectivityin big data sources. Eurostat Statistical Working Papers

Bertarelli G., D'Agostino A., Giusti C., Pratesi M. (forthcoming) Measuring Educational Poverty in Italy: a Multidimensional and Fuzzy Approach. Routledge.

Betti G. and Lemmi A. (2013). Poverty and Social Exclusion. New Methods and Analysis. Routledge.

de Jonge E. (2020). Communicating the uncertainty in official data. EMOS Webinar 2020 (emos2020events.ec.unipi.it/communicating-the-uncertainty-in-official-data/)

Eurostat (2014). ESS handbook for quality reports. Eurostat Manuals and Guidelines

FAO (2015). Spatial disaggregation and small-area estimation methods for agricultural surveys: Solutions and perspectives, Tech. Rep. Technical Report Series GO-07-2015, Global Strategy—Improving Agricultural and Rural Statistics

Giusti C., Masserini L., Pratesi M. (2015). Local Comparisons of Small Area Estimates of Poverty: An Application Within the Tuscany Region in Italy. Social Indicators Research. doi: 10.1007/s11205-015-1193-1

Loonis V. and M.P. de Bellefon (eds.) (2018). Handbook of Spatial Analysis, Theory and Application with R. INSEE Méthodes No 131, Insee-Eurostat

# References


Marchetti S., Giusti C., Pratesi M., Salvati N., Giannotti F., Pedreschi D., Rinzivillo S., Pappalardo L., Gabrielli L. (2015). Small area model-based estimators using big data sources. Journal of Official Statistics, vol. 31, p. 263-281

Marchetti S., Giusti C., Pratesi M. (2016). The use of Twitter data to improve small area estimates of households' share of food consumption expenditure in Italy. AStA Wirtschafts- und Sozialstatistisches Archiv, 10(2-3), pp. 79-93

Pratesi M., Quattrociocchi L., Bertarelli G., Gemignani A., Giusti C. (2020). Spatial Distribution of Multidimensional Educational Poverty in Italy using Small Area Estimation. Social Indicators Research

Pratesi M. (2016). Analysis of Poverty Data by Small Area Estimation. Wiley

Simler K. (2016). Pinpointing Poverty in Europe: New Evidence for Policy Making. World Bank, Washington, DC

Statistics Canada (2003). Statistics Canada Quality Guidelines.

Tzavidis N., Zhang L. C., Luna A., Schmid T., Rojas-Perilla N. (2018). From start to finish: a framework for the production of small area official statistics. Journal of the Royal Statistical Society: Series A (Statistics in Society), 181(4), 927-979

UN-GGIM: Europe (2019). The territorial dimension in SDG indicators: geospatial data analysis and its integration with statistical data. Instituto Nacional de Estatística, Lisboa.

# Thank you for your participation!

# Questions?

MONICA PRATESI
monica.pratesi@unipi.it

CATERINA GIUSTI
caterina.giusti@unipi.it