Second generation at school

Integration of different data sources in a small area perspective

Daniela Cocchi Francesco Giovinazzi

Department of Statistical Sciences
University of Bologna

Bergamo, June 10-12

ITSEW 2019 – International Total Survey Error Workshop
"Integration of surveys and alternative data sources"

Data about children and the ECDP

The European Cohort Development Project

The ECDP is a design study aimed at creating a European Research Infrastructure (**EuroCohort**) to provide comparative longitudinal survey data on children and young adults well-being.

https://www.eurocohort.eu/

Children and young adults as a **special population** in terms of both data quality and data availability.

Focus on Second Generation children

We focus on children belonging to the **Second Generation** or having a migration background.

Students can be grouped according to nationality (10) and administrative region of the attended school (21).

In 2015 the Italian Ministry of Interior and the European Fund for the Integration of third-country nationals (EFI) co-financed a survey on **Integration of the Second Generation** (ISG), which was carried out by Istat.

The problem: level of integration (in small domains)

We want to assess the level of integration of second generation students in Italy by means of a proxy variable from ISG survey.

We treat student's citizenship within each administrative region as a **study domain**, estimating the proportion of students which are more or less integrated within each domain.

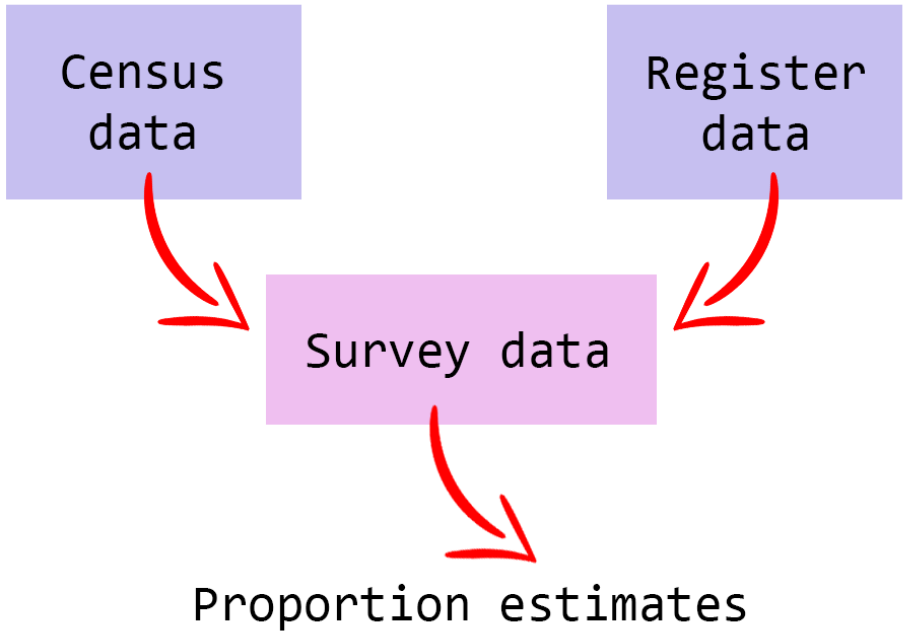
The idea: using sae methods for data integration

We are dealing with **unplanned domains** and we need to borrow strength from external data sources.

We propose to integrate auxiliary information from two different administrative data sources:

- the 15-th Italian Population and Housing Census (2011);
- the ”Anagrafe Nazionale degli Studenti” (ANS) the official register kept by the Ministry of Education, Universities and Research (MIUR).

Data integration



The ISG survey

ISG involved about 1400 lower and upper secondary schools attended by at least 5 foreign students in 20 Italian regions.

Data consist of 68127 observations on 255 items.

The questionnaire investigated many different dimensions of social inclusion, and can be divided in 6 broad sections:

A. Administrative data and migration history.
B. Use of native and local languages.
C. Relationship with schoolmates and teachers.
D. Relationship with friends, free time and social habits.
E. Composition of the family and relationship with its members.
F. Household conditions.

The ISG survey: nationalities in the sample

Foreigners
31687
(46.5%)

Italians
36440
(53.5%)

Born in Italy
9002 (28.4%)

Born abroad
22685 (71.5%)

Albania	1811 (20.1%)
Morocco	1080 (12%)
China	1015 (11.3%)
Romania	750 (8.3%)
Philippines	581 (6.4%)
Others	3765 (41.9 %)

Romania	5879 (25.9%)
Albania	2852 (12.6%)
Morocco	1555 (6.8%)
Moldova	1361 (6%)
Ukraine	1056 (4.7%)
Others	9982 (44%)

The ISG survey: a proxy for the level of integration

Among the many items in the questionnaire we selected Item A11 as best proxy of integration for a foreign student.

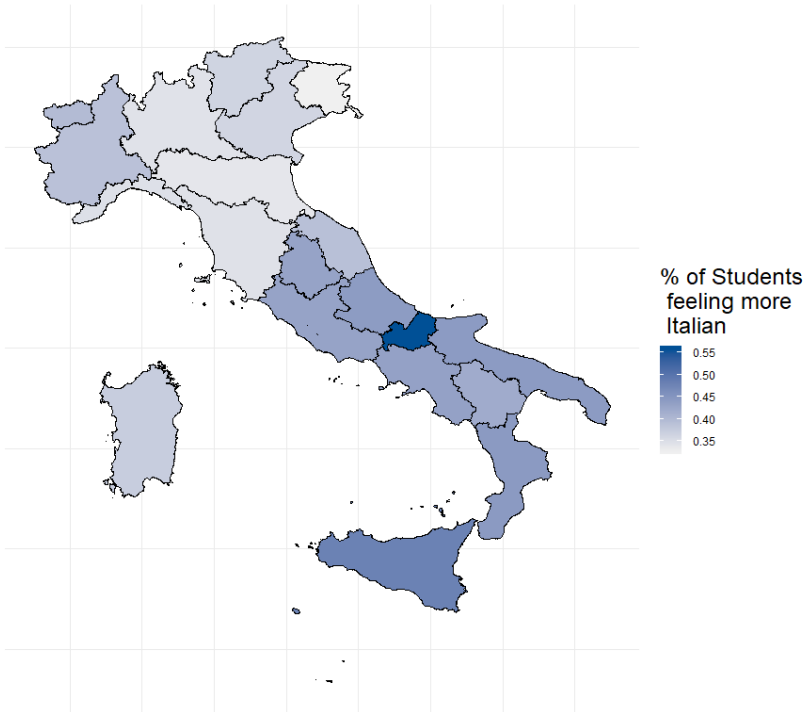
Do you feel more...

Italian | Foreigner | Don't know

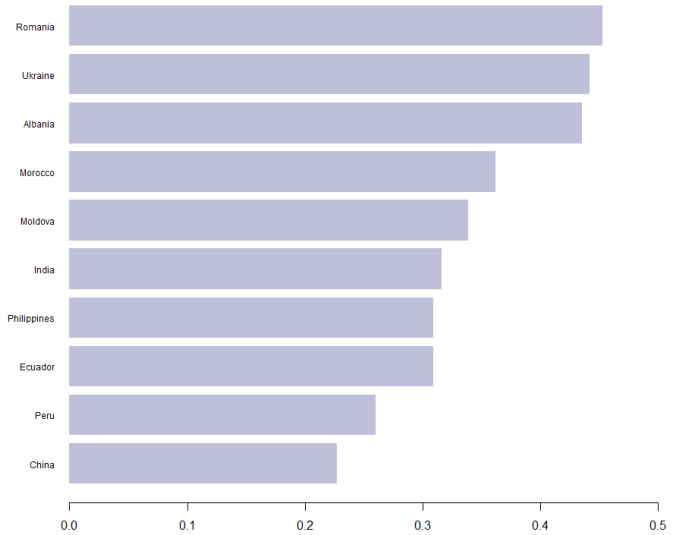
What is the proportion of students answering "Italian" in 21 Italian regions and 10 nationalities in the sample?

The ISG survey: distribution of the answer (marginals)

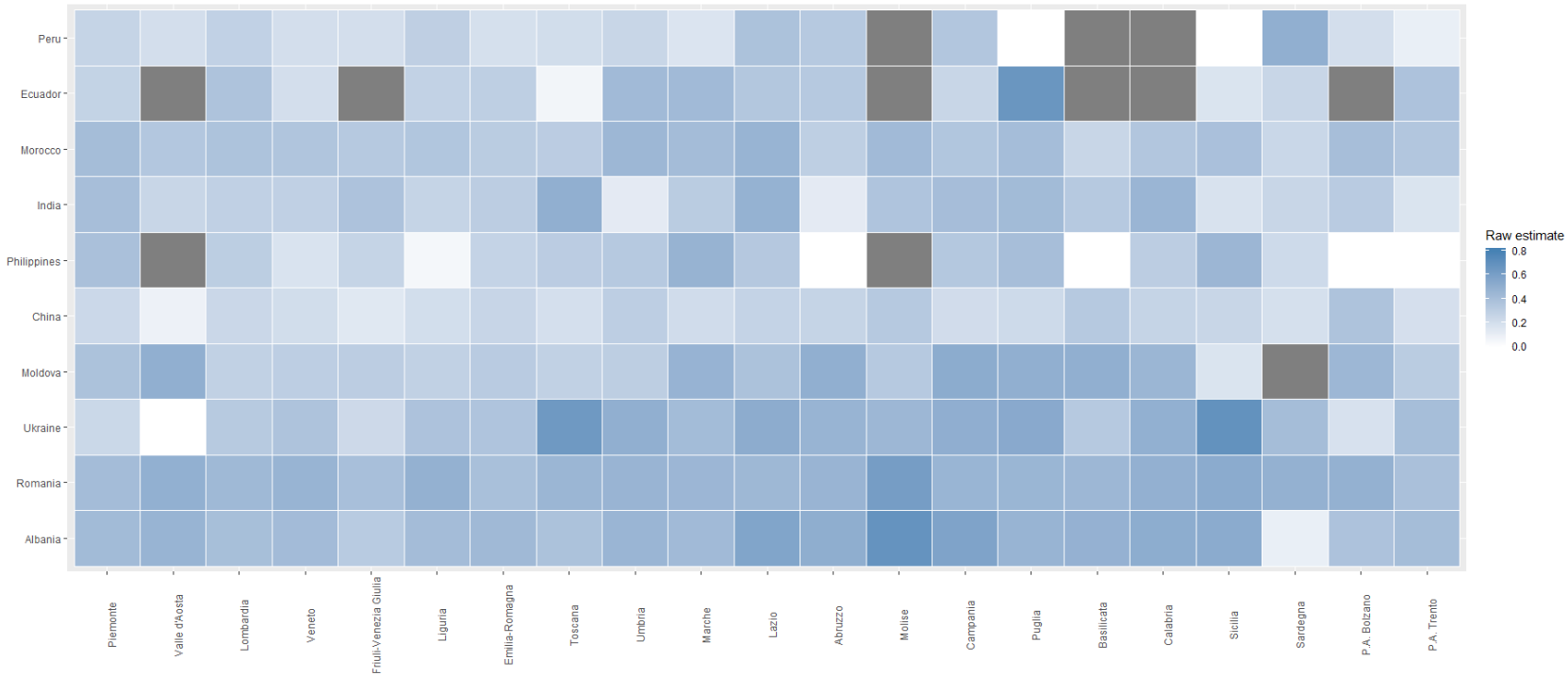
Regions



Nationalities



The ISG survey: distribution of the answer (joint)



Census data

The **15th Population and housing census** 2011 was launched by Istat in October 2011. Census data and results are available open source on Istat website.

Issues in using census data as auxiliary information:

- difference in the target population (residents vs. students)
- difference in the year of data collection (2011 vs. 2015)

ANS register data

Anagrafe Nazionale Studenti is kept by MIUR. The register collects data from schools all over the country and it is updated at the beginning of each school year in September.

Issues in using ANS data as auxiliary information:

- difference in the target population (all schools vs. schools with more than 5 for students)
- updated only on voluntary base by families
- not open source

Data organization for modeling

Given K regions $k = 1, \ldots, K$ and L nationalities $i = 1, \ldots, L$, we identify KL groups, each with size n_{ik} so that $\sum_{i,k} n_{ik} = n$.

$$\underbrace{\mathbf{y} = \begin{bmatrix} y_{111} \\ y_{112} \\ \vdots \\ y_{ikj} \\ \vdots \\ y_{LKN} \end{bmatrix}}_{\text{Survey}} \quad \underbrace{\mathbf{X} = \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_k' \\ \vdots \\ \mathbf{x}_K' \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1M} \\ x_{21} & x_{22} & \ldots & x_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{K1} & x_{K2} & \ldots & x_{KM} \end{bmatrix}}_{\text{Register}} \quad \underbrace{\mathbf{G} = \begin{bmatrix} \mathbf{g}_1' \\ \mathbf{g}_2' \\ \vdots \\ \mathbf{g}_i' \\ \vdots \\ \mathbf{g}_L' \end{bmatrix} = \begin{bmatrix} g_{11} & g_{12} & \ldots & g_{1H} \\ g_{21} & g_{22} & \ldots & g_{2H} \\ \vdots & \vdots & \ddots & \vdots \\ g_{L1} & g_{K2} & \ldots & g_{LH} \end{bmatrix}}_{\text{Census}}$$

We refer to $\hat{p}_{ik} = \dfrac{\sum_{j=1}^{n_{ik}} y_{ikj}}{n_{ik}}$ as dependent variable.

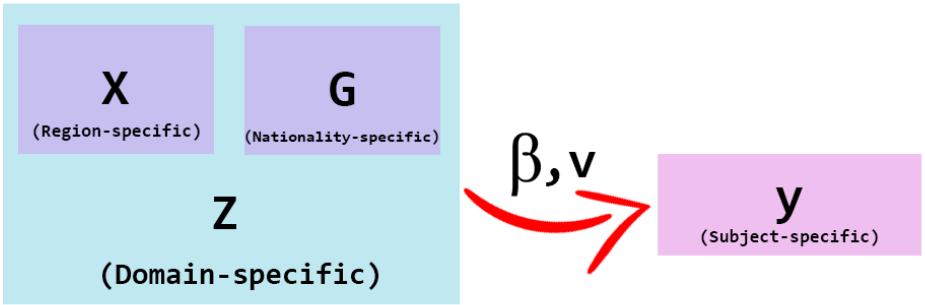
Suggestion: area level model

The Fay-Herriot model

We can combine **X** and **G** into a block matrix **Z** and fit a classic Fay-Herriot area level model

$$\log\left(\frac{p_{ik}}{1-p_{ik}}\right) = \text{logit}(p_{ik}) = \mathbf{z}'_{ik}\beta + v_{ik}$$

$$v_{ik} \sim N\left(0, \sigma_v^2\right)$$

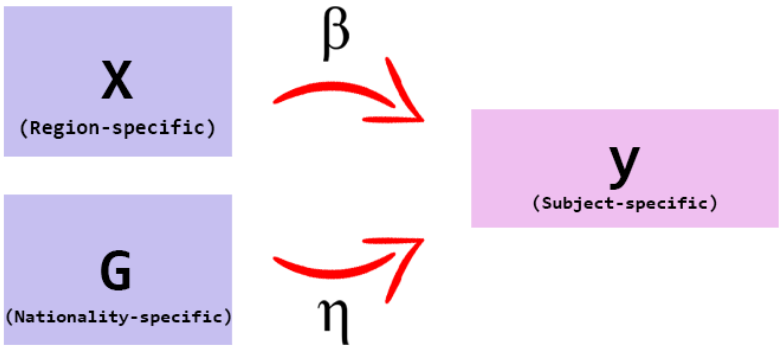


An alternative multilevel glmm

Following Malec *et al.* (1997) as cited in Rao & Molina (2015), we assume $Y_{ikj} \sim \text{Binom}(p_{ik}, n_{ik})$ with $j = 1, \ldots, n_{ik}$, and write

$$\log\left(\frac{p_{ik}}{1-p_{ik}}\right) = \text{logit}(p_{ik}) = \mathbf{x}'_k \beta_i$$

$$\beta_i | \eta, \mathbf{\Gamma} \sim N_M(\mathbf{g}_i \eta, \mathbf{\Gamma}) \qquad p(\eta, \mathbf{\Gamma}) \propto 1$$



An alternative multilevel glmm

We can write such model as:

$$\text{logit}(p_{ik}) = \mathbf{x}_k' \left[\mathbf{g}_i \boldsymbol{\eta} + \mathbf{u}_i\right] = \underbrace{\mathbf{x}_k' \mathbf{g}_i \boldsymbol{\eta}}_{\text{Fixed}} + \underbrace{\mathbf{x}_k' \mathbf{u}_i}_{\text{Random}} = \mathbf{z}_{ik}' \boldsymbol{\eta} + \mathbf{x}_k' \mathbf{u}_i$$

$$\mathbf{u}_i | \boldsymbol{\Gamma} \sim N_M(\mathbf{0}, \boldsymbol{\Gamma}) \qquad p(\boldsymbol{\eta}, \boldsymbol{\Gamma}) \propto 1$$

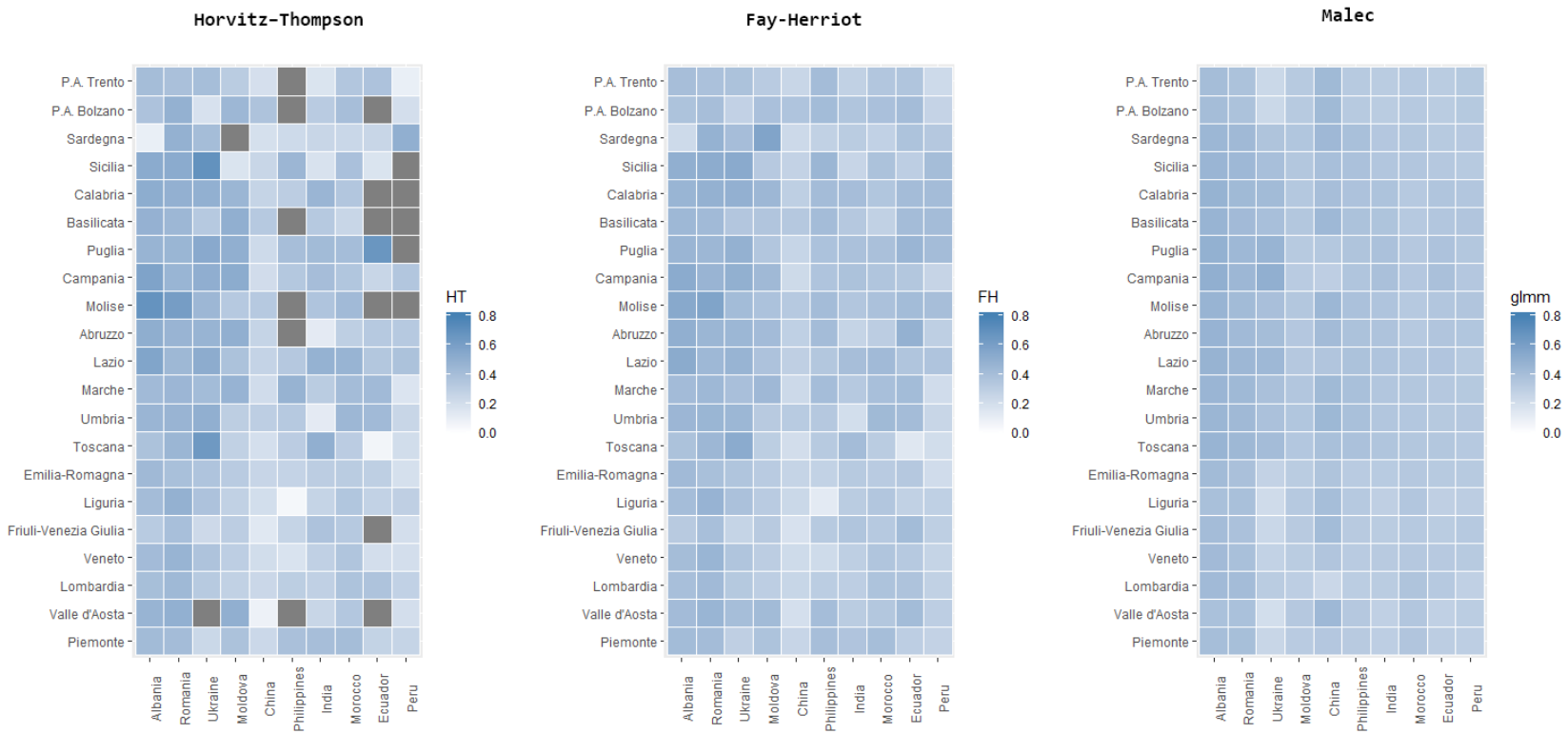
A glimpse on first results

Our interest focuses on the point estimation of 210 proportions.

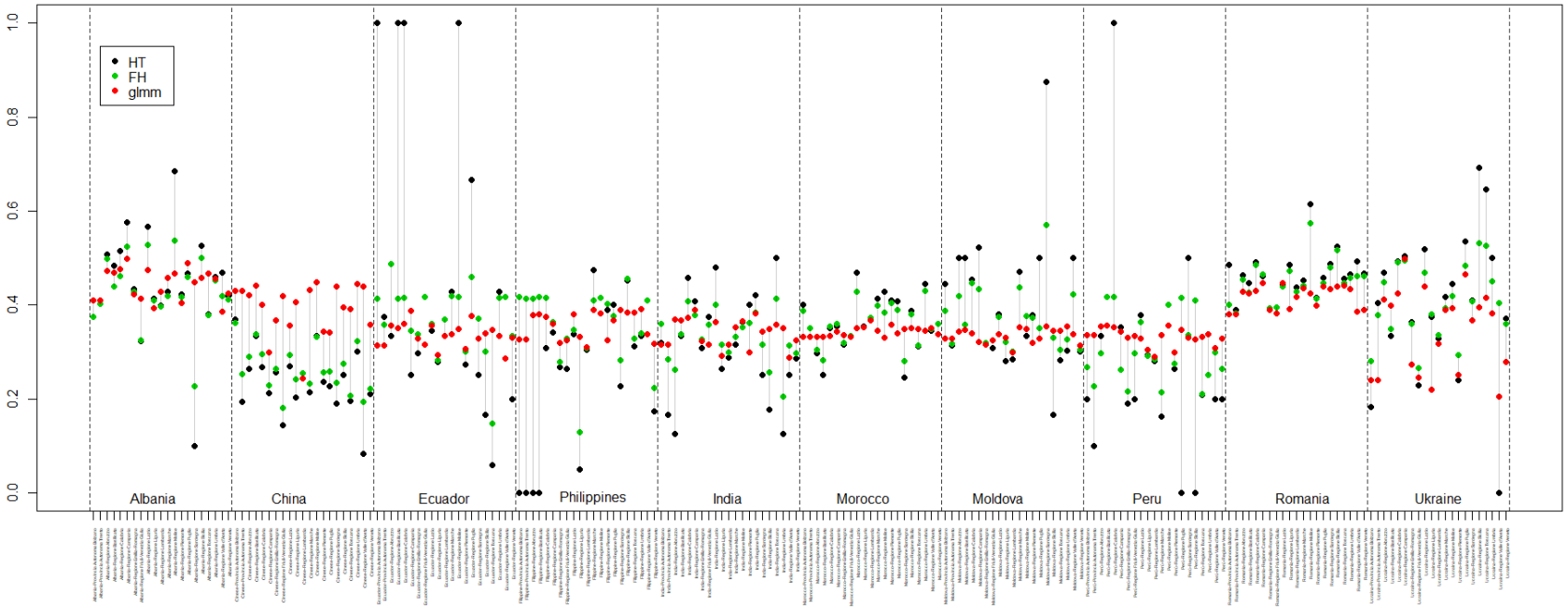
We compare:

- Horvitz–Thompson estimates [HT]
- EBLUP from a Fay-Herriot area level model [FH]
- Fitted values from the model *à la* Malec [glmm]

The fitted values

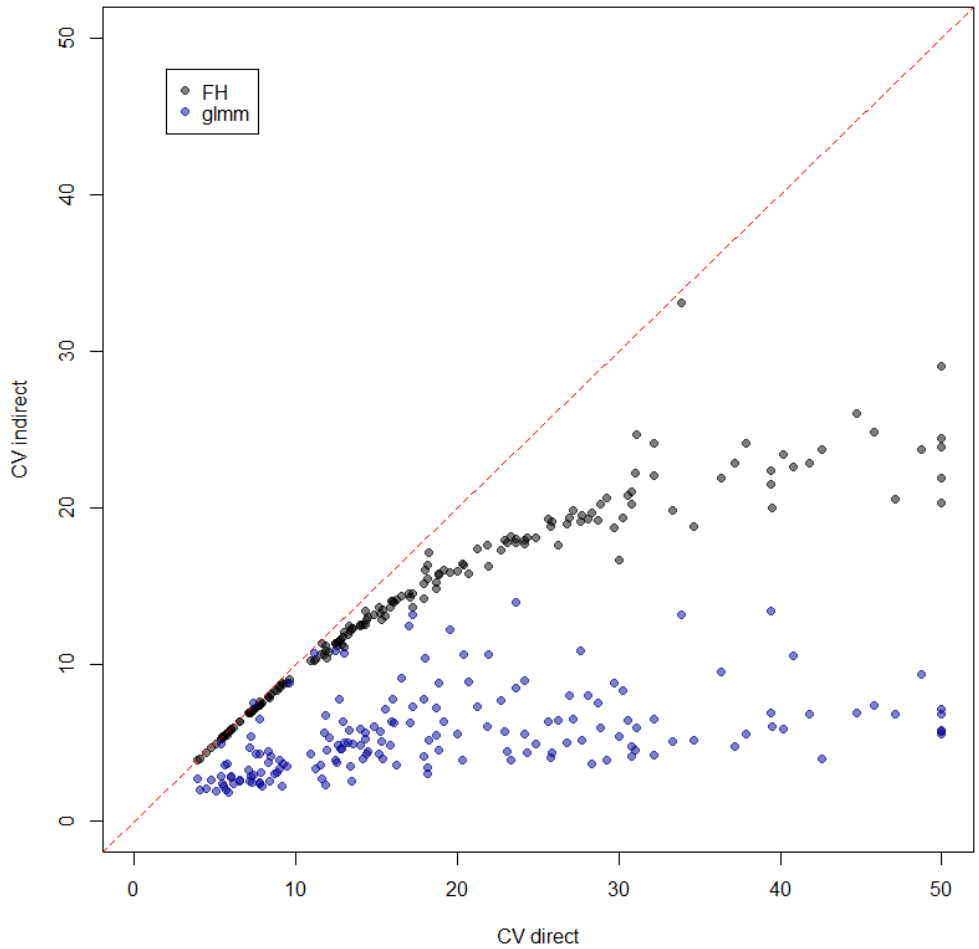


The fitted values

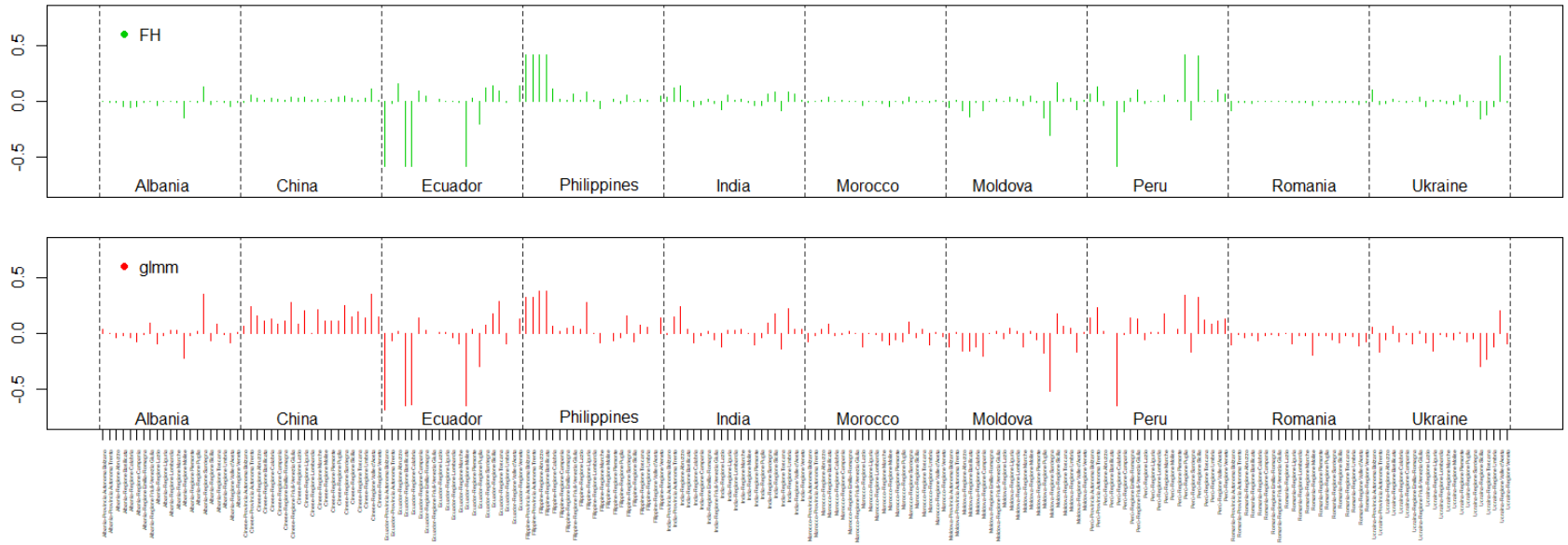


We do not compute bias and mse. Focus on coefficient of variation and comparison with Horvitz-Thompson.

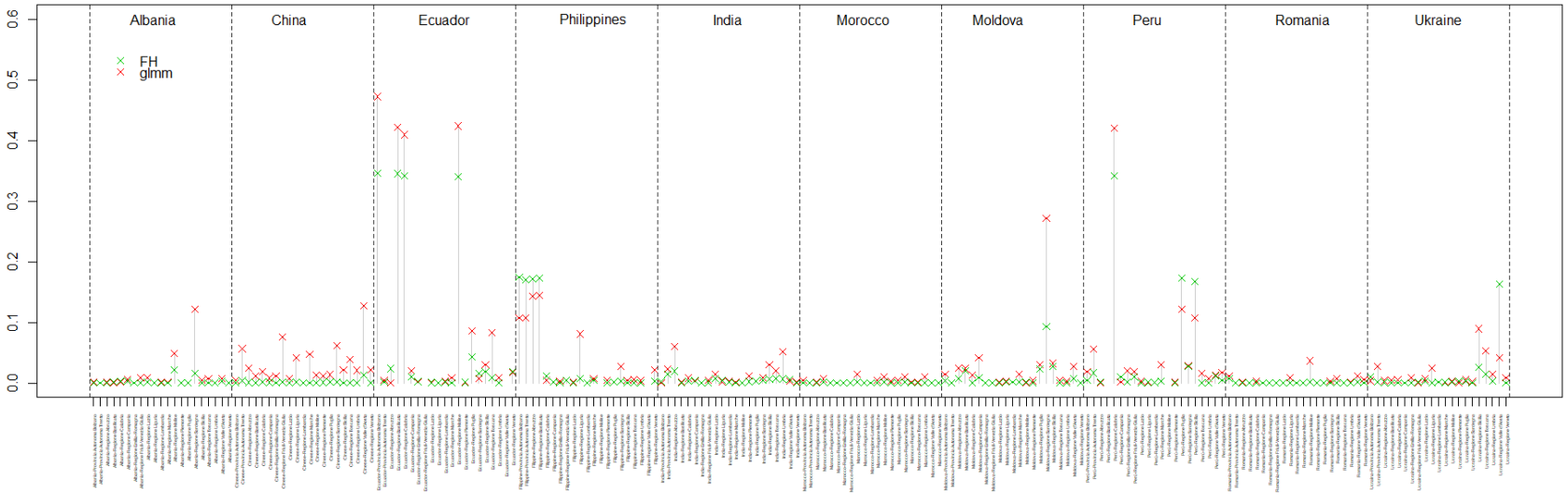
Coefficient of variation



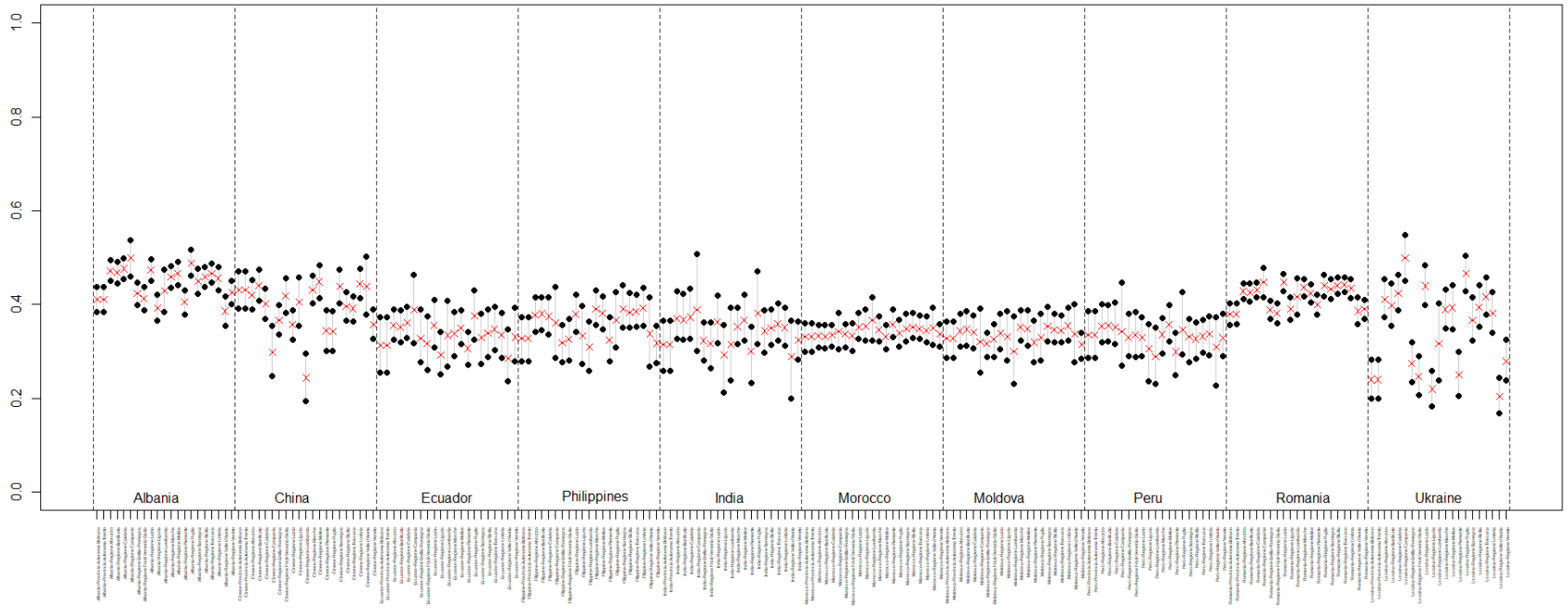
Distance from the HT estimate



Global variability with respect to the HT estimate



Credibility intervals for the glmm estimates (95%)



Final remarks

We have experimented methods to enforce the estimates of the level of integration of second generation students borrowing strength from both census data and ANS register data.

What to do next:

- develop the model with a more complex covariate structure
- define the level of integration as a latent variable

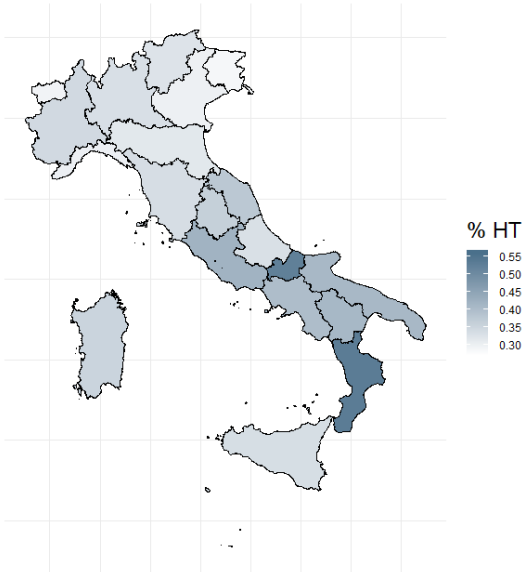
Thank you!

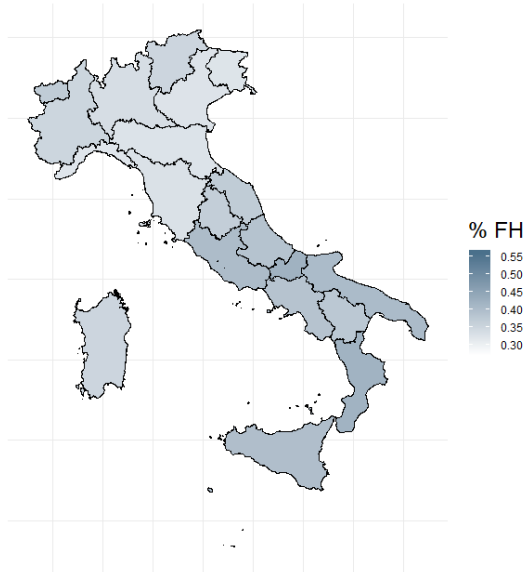
References

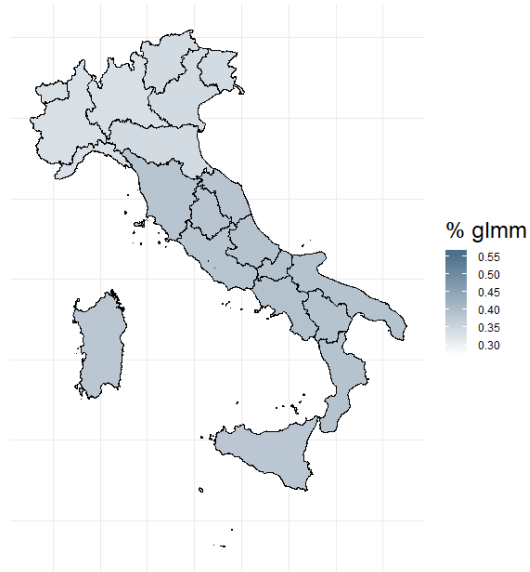
MALEC, DONALD, SEDRANSK, J, MORIARITY, CHRISTOPHER L, & LECLERE, FELICIA B. 1997. Small area inference for binary variables in the national health interview survey. *Journal of the american statistical association*, **92**(439), 815–826.

RAO, JOHN NK, & MOLINA, ISABEL. 2015. *Small area estimation*. Wiley Series in Survey Methodology.

The fitted values (regions)







The fitted values (nationalities)