Estimating the Number and Characteristics of Illegal Migrants via the Center Sampling Technique: Methodology and Application to Italian Data

Gianluca Baio[1,2], Gian Carlo Blangiardo[2] and Marta Blangiardo[3]

[1] Department of Statistical Science, University College London (UK)
[2] Department of Statistics, University of Milano Bicocca (Italy)
[3] Department of Epidemiology and Biostatistics, Imperial College London (UK)
gianluca@stats.ucl.ac.uk

Joint Statistical Meeting
San Diego Convention Center

29 July 2012

Outline of presentation

1. **Description of the problem**
2. Centre sampling design
3. Some results
   - Simulations
   - Applications to Italian data
4. Conclusions

Outline of presentation

1. Description of the problem
2. **Centre sampling design**
3. Some results
   - Simulations
   - Applications to Italian data
4. Conclusions

Outline of presentation

1. Description of the problem
2. Centre sampling design
3. **Some results**
   - **Simulations**
   - **Applications to Italian data**
4. Conclusions

Outline of presentation

The problem

- Survey a group of individuals in a population, when information about the complete list of the members is missing or partially unknown
- Standard sampling methods (eg SRS) in general may not be appropriate
- Alternative methods
  - Capture-recapture schemes
  - Snowball sampling

The problem

- Survey a group of individuals in a population, when information about the complete list of the members is missing or partially unknown
- Standard sampling methods (eg SRS) in general may not be appropriate
- Alternative methods
  - Capture-recapture schemes
  - Snowball sampling

- This situation is particularly relevant in the study of migration (eg when some individuals are **unauthorised migrants** and therefore are not registered in the official records)
- The objective of our analysis is to devise a method to estimate (with reasonable precision) some selected features of the population of interest (eg the distribution of age, sex, marital and socio-economic status, political views, etc)
- Estimation of the number of unauthorised migrants is a by-product of the procedure (by integrating specific surveys with suitable external data)

Centre sampling (CS) scheme

- The basic idea is to characterise the sampling units in terms of a set of K **aggregation places** ("centres") with which they are associated
- In other words, **by necessity**, they have some form of relationship with at least one of them
  - Centres with a known list of individuals (eg healthcare facilities, schools, worship places, population registry)
  - Centres with no list of individuals (eg restaurants, bars, discos, open air places such as squares)

Centre sampling (CS) scheme

- The basic idea is to characterise the sampling units in terms of a set of K **aggregation places** ("centres") with which they are associated
- In other words, **by necessity**, they have some form of relationship with at least one of them
  - Centres with a known list of individuals (eg healthcare facilities, schools, worship places, population registry)
  - Centres with no list of individuals (eg restaurants, bars, discos, open air places such as squares)

- The centres should be selected in order to maximise the probability that a random individual from the population has some contacts with at least one of them

Centre sampling (CS) scheme

- The basic idea is to characterise the sampling units in terms of a set of K **aggregation places** ("centres") with which they are associated
- In other words, **by necessity**, they have some form of relationship with at least one of them
  - Centres with a known list of individuals (eg healthcare facilities, schools, worship places, population registry)
  - Centres with no list of individuals (eg restaurants, bars, discos, open air places such as squares)

- The centres should be selected in order to maximise the probability that a random individual from the population has some contacts with at least one of them

- The CS scheme uses a three-level procedure
  1. Sample with replacement n out of the K centres
  2. At each draw, select a subject from the N_k normally having contacts with the selected centre, k
  3. **Re-proportion** the (biased) sample obtained with this procedure, to produce suitable estimations

CS scheme (2)

- For each individual in the population, define a vector

$$\mathbf{u}(i) = [u_1(i), u_2(i), \ldots, u_K(i)]$$

where

$$u_k(i) = \begin{cases} 1 & \text{if the } i-\text{th unit has regular contacts with centre } k \\ 0 & \text{otherwise} \end{cases}$$

CS scheme (2)

- For each individual in the population, define a vector

$$\mathbf{u}(i) = [u_1(i), u_2(i), \ldots, u_K(i)]$$

where

$$u_k(i) = \begin{cases} 1 & \text{if the } i\text{−th unit has regular contacts with centre } k \\ 0 & \text{otherwise} \end{cases}$$

- This characterises the **profile** of the i−th individual in terms of the K centres
- Together with the variables of interest in the survey (eg sex, age, socio-economic status, etc.), each individual $i = 1, \ldots, n$ provides information on the centres that they normally visit, in the form of a vector $\mathbf{u}(i)$

CS scheme (3)

- Define:
  - The number of individuals in the overall population who have a given profile $\mathbf{u} = (u_1, u_2, \ldots, u_K)$, $N(\mathbf{u})$
  - The correspondent population proportion

$$\pi(\mathbf{u}) = \frac{N(\mathbf{u})}{N}$$

CS scheme (3)

- Define:
  - The number of individuals in the overall population who have a given profile $\mathbf{u} = (u_1, u_2, \ldots, u_K)$, $N(\mathbf{u})$
  - The correspondent population proportion

$$\pi(\mathbf{u}) = \frac{N(\mathbf{u})}{N}$$

- The idea behind the CS is that, **when suitably weighted**, the n sampling units should have a sampling distribution consistent with the population distribution of $\pi(\mathbf{u})$

CS scheme (3)

- Define:
  - The number of individuals in the overall population who have a given profile $\mathbf{u} = (u_1, u_2, \ldots, u_K)$, $N(\mathbf{u})$
  - The correspondent population proportion

$$\pi(\mathbf{u}) = \frac{N(\mathbf{u})}{N}$$

- The idea behind the CS is that, **when suitably weighted**, the n sampling units should have a sampling distribution consistent with the population distribution of $\pi(\mathbf{u})$
- This condition holds if the weights are defined as

$$w(\mathbf{u}) := \frac{\pi(\mathbf{u})}{\hat{\pi}(\mathbf{u})} = \frac{N(\mathbf{u})/N}{n(\mathbf{u})/n}$$

where $n(\mathbf{u})$ is the number of sample units who possess profile $\mathbf{u}$ (and similarly $\hat{\pi}(\mathbf{u})$ is the sample proportion of such individuals)

CS scheme (4) — digression

- The expected number of units with profile $\mathbf{u}$ in the sample is

$$\mathsf{E}[n(\mathbf{u})] = \sum_{k=1}^{K} n_k p_k(\mathbf{u}) \qquad \text{with} \qquad p_k(\mathbf{u}) = \begin{cases} N(\mathbf{u})/N_k & \text{if } u_k = 1 \\ 0 & \text{otherwise} \end{cases}$$

CS scheme (4) — digression

- The expected number of units with profile $\mathbf{u}$ in the sample is

$$\mathsf{E}[n(\mathbf{u})] = \sum_{k=1}^{K} n_k p_k(\mathbf{u}) \qquad \text{with} \qquad p_k(\mathbf{u}) = \begin{cases} N(\mathbf{u})/N_k & \text{if } u_k = 1 \\ 0 & \text{otherwise} \end{cases}$$

- Consequently, the corresponding expected sample proportion is

$$\mathsf{E}[\hat{\pi}(\mathbf{u})] = \mathsf{E}\left[\frac{n(\mathbf{u})}{n}\right] = \sum_{k=1}^{K} \frac{n_k}{n} \frac{N(\mathbf{u})}{N_k} u_k$$

with

$$\mathsf{Var}[\hat{\pi}(\mathbf{u})] = \frac{1}{n^2} \sum_{k=1}^{K} n_k \frac{N(\mathbf{u})}{N_k} \left[1 - \frac{N(\mathbf{u})}{N_k}\right] u_k$$

which goes to 0 (fairly quickly), for n sufficiently large

CS scheme (5)

- So, **if the sample is large enough**, the observed value of $\hat{\pi}(\mathbf{u})$ can be used a reasonable (and convenient) estimation of its unknown expected value

CS scheme (5)

- So, **if the sample is large enough**, the observed value of $\hat{\pi}(\mathbf{u})$ can be used a reasonable (and convenient) estimation of its unknown expected value
- Therefore, setting $f_k = N_k/N$, we obtain

$$\begin{aligned}
\hat{\pi}(\mathbf{u}) \approx \mathsf{E}\left[\hat{\pi}(\mathbf{u})\right] &= \sum_{k=1}^{K} \frac{n_k}{n} \frac{N(\mathbf{u})}{N_k} u_k \\
&= \frac{N(\mathbf{u})}{N} \sum_{k=1}^{K} \frac{n_k}{n} \frac{N}{N_k} u_k \\
&= \pi(\mathbf{u}) \sum_{k=1}^{K} \frac{n_k/n}{f_k} u_k
\end{aligned}$$

CS scheme (5)

- So, **if the sample is large enough**, the observed value of $\hat{\pi}(\mathbf{u})$ can be used a reasonable (and convenient) estimation of its unknown expected value
- Therefore, setting $f_k = N_k/N$, we obtain

$$\begin{aligned}\hat{\pi}(\mathbf{u}) \approx \mathsf{E}\left[\hat{\pi}(\mathbf{u})\right] &= \sum_{k=1}^{K} \frac{n_k}{n} \frac{N(\mathbf{u})}{N_k} u_k \\ &= \frac{N(\mathbf{u})}{N} \sum_{k=1}^{K} \frac{n_k}{n} \frac{N}{N_k} u_k \\ &= \pi(\mathbf{u}) \sum_{k=1}^{K} \frac{n_k/n}{f_k} u_k\end{aligned}$$

and thus the weights can be taken as

$$w(\mathbf{u}) = \frac{\pi(\mathbf{u})}{\hat{\pi}(\mathbf{u})} = \left(\sum_{k=1}^{K} \frac{n_k/n}{f_k} u_k \right)^{-1}$$

Comments

- The basic assumption in the CS design is that we have some information on the popularity of the centres
- Using this framework (and with a large enough sample size), it is relatively straightforward to compute the weights $w(\mathbf{u})$ and the weighted sample produces consistent estimations of the relevant variables

Comments

- The basic assumption in the CS design is that we have some information on the popularity of the centres
- Using this framework (and with a large enough sample size), it is relatively straightforward to compute the weights $w(\mathbf{u})$ and the weighted sample produces consistent estimations of the relevant variables

- Obviously, this setting might be too restrictive, and two possible ways of estimating the rates f_k are the following:
  1. Selecting a popularity score of $1/K$ for each centre
  2. Estimating the value of each f_k using (possibly subjective) knowledge

Comments

- The basic assumption in the CS design is that we have some information on the popularity of the centres
- Using this framework (and with a large enough sample size), it is relatively straightforward to compute the weights $w(\mathbf{u})$ and the weighted sample produces consistent estimations of the relevant variables

- Obviously, this setting might be too restrictive, and two possible ways of estimating the rates f_k are the following:
  1. Selecting a popularity score of $1/K$ for each centre
  2. Estimating the value of each f_k using (possibly subjective) knowledge

- In either case, after weighting the sample, we obtain unbiased estimates for the variables of interest as well as for the "true" f_k, but precision is increased using the extra information on the centres' popularity

Simulation study

- Assume that the "true" population is made by $N = 20531$ individuals
- For each individual, simulate
  - A profile in terms of contacts with $K = 7$ centres, based on "true" popularity rates $\mathbf{f} = (0.074, 0.111, 0.133, 0.022, 0.237, 0.133, 0.288)$
  - Values for sex, age, marital status (married/single) and education level (no, lower, higher education and PhD)

Simulation study

- Assume that the "true" population is made by $N = 20531$ individuals
- For each individual, simulate
  - A profile in terms of contacts with $K = 7$ centres, based on "true" popularity rates $\mathbf{f} = (0.074, 0.111, 0.133, 0.022, 0.237, 0.133, 0.288)$
  - Values for sex, age, marital status (married/single) and education level (no, lower, higher education and PhD)

- Use the CS design to sample $n = 800$ subjects ($\sim$ 4% of the population)
  - Assume complete response and complete observation (no missing data)
  - Consider several scenarios, upon varying the selection of centres (fewer than in the population) and the popularity rates (nearly correct vs completely wrong)

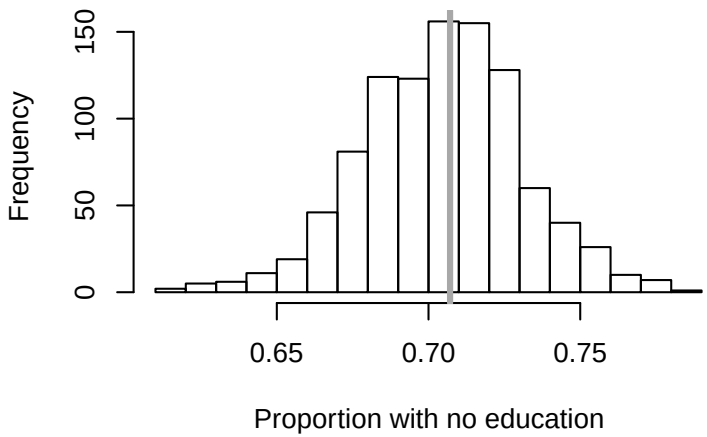
Simulation study

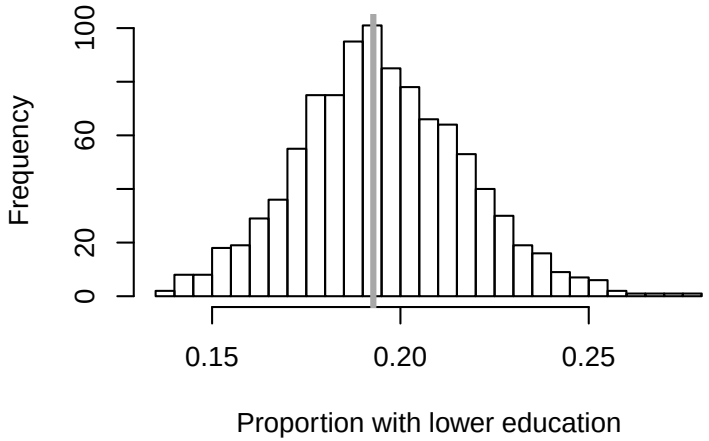
- Assume that the "true" population is made by $N = 20531$ individuals
- For each individual, simulate
  - A profile in terms of contacts with $K = 7$ centres, based on "true" popularity rates $\mathbf{f} = (0.074, 0.111, 0.133, 0.022, 0.237, 0.133, 0.288)$
  - Values for sex, age, marital status (married/single) and education level (no, lower, higher education and PhD)

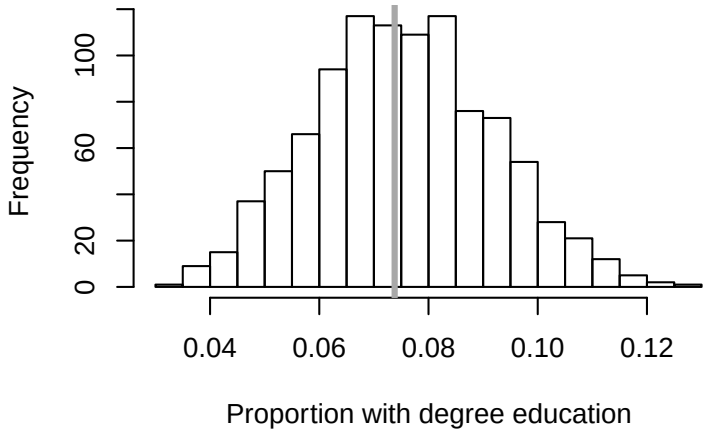
- Use the CS design to sample $n = 800$ subjects ($\sim$ 4% of the population)
  - Assume complete response and complete observation (no missing data)
  - Consider several scenarios, upon varying the selection of centres (fewer than in the population) and the popularity rates (nearly correct vs completely wrong)

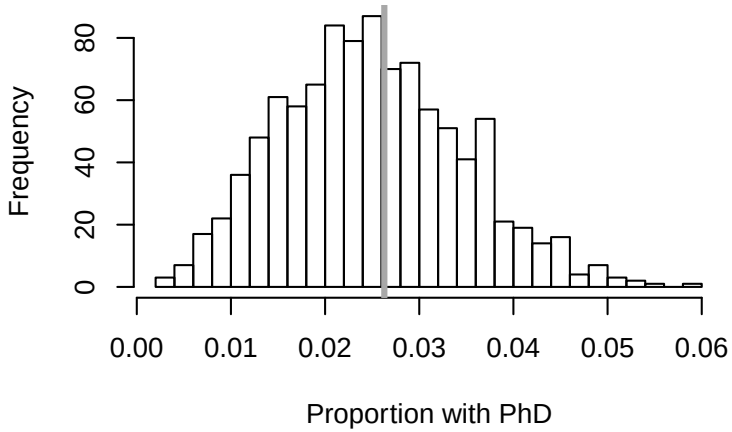
- Repeat sampling for a large number $n_{\text{sim}} = 1000$ of times to assess the performance of the method

All centres, correct popularity rates

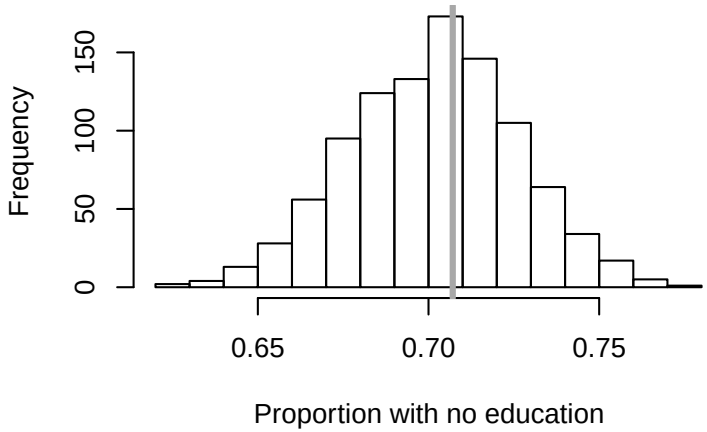


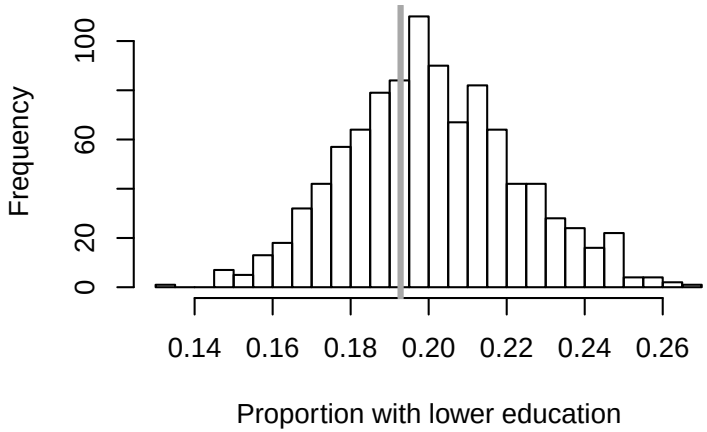


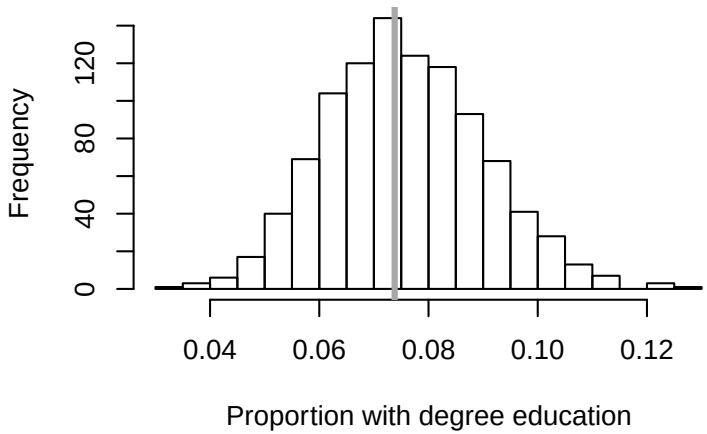


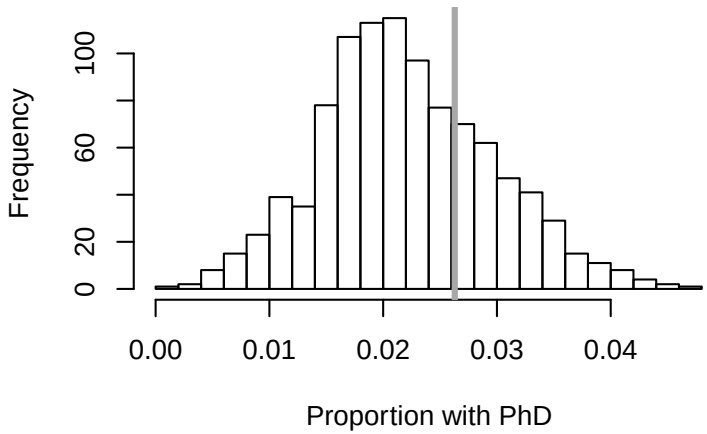


Only the 2 most popular centres

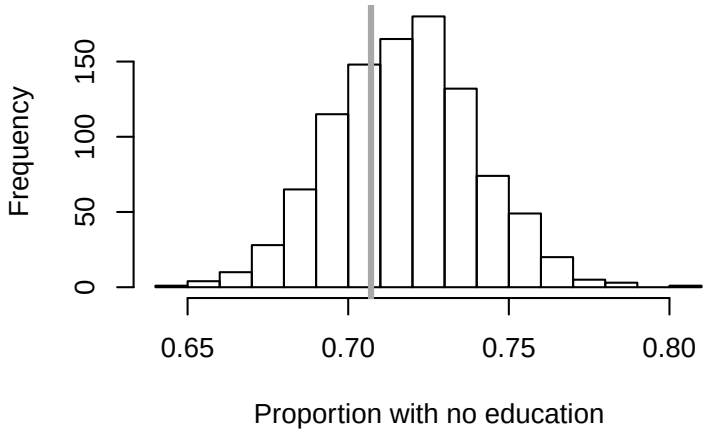


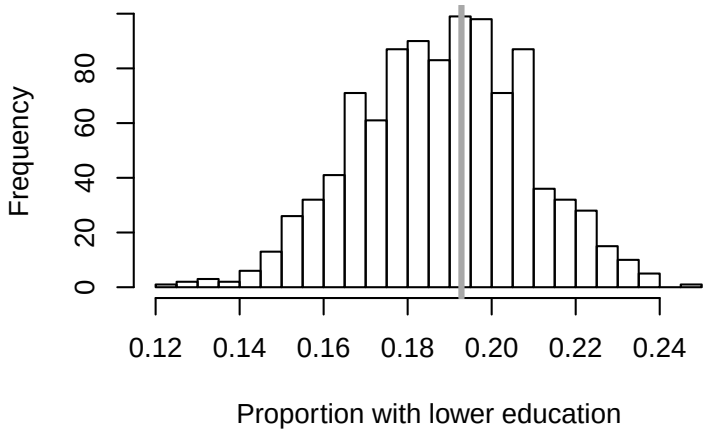


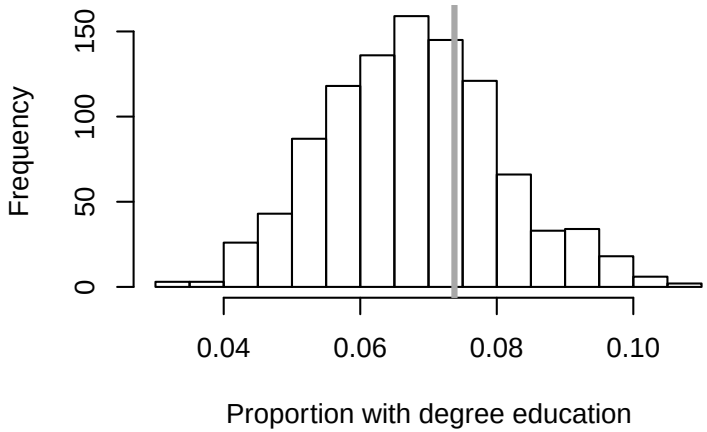


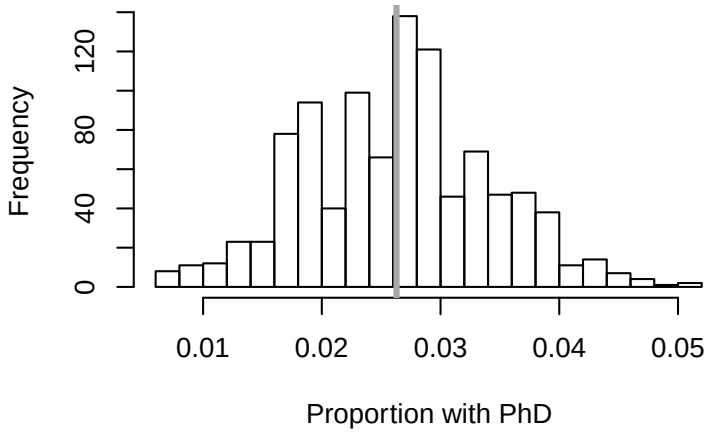


Only 2 centres, excluding the most popular









Using the CS to estimate the number of unauthorised migrants

The information obtained from the CS design can be used to post-process official data (eg from Census or population registry, PR) and provide suitable estimations (eg number of unauthorised migrants in a given area)

Using the CS to estimate the number of unauthorised migrants

The information obtained from the CS design can be used to post-process official data (eg from Census or population registry, PR) and provide suitable estimations (eg number of unauthorised migrants in a given area)

Example: Estimating the number of Ukrainians in Milan

1. Collect data from official sources and combine them with those obtained from the CS design
   - Ukrainians in the population registry of Milan (official data): 3628
   - Rate of Ukrainian in the PR (estimated by CS): 67.5%
   - Rate of Ukrainian authorised migrants (estimated by CS): 79.8%

Using the CS to estimate the number of unauthorised migrants

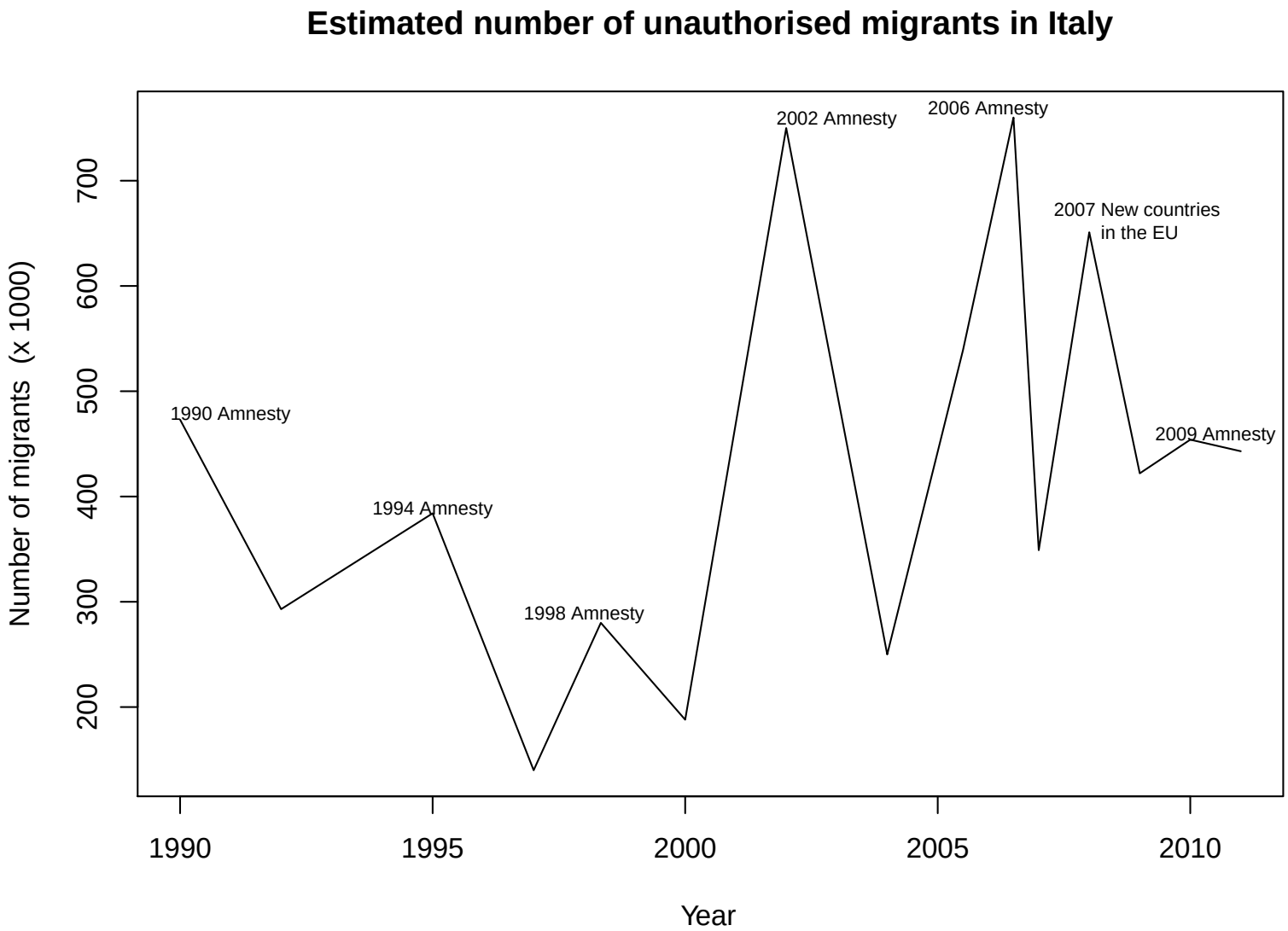
The information obtained from the CS design can be used to post-process official data (eg from Census or population registry, PR) and provide suitable estimations (eg number of unauthorised migrants in a given area)

Example: Estimating the number of Ukrainians in Milan

1. Collect data from official sources and combine them with those obtained from the CS design
   - Ukrainians in the population registry of Milan (official data): 3628
   - Rate of Ukrainian in the PR (estimated by CS): 67.5%
   - Rate of Ukrainian authorised migrants (estimated by CS): 79.8%
2. Re-proportion the official data
   - Estimated number of Ukrainian in Milan: $\frac{3628}{0.675} = 5375$
   - Estimated number of unauthorised migrants: $\frac{3628}{1-0.798} = 1086$
   - Estimated number of authorised non-resident migrants: $5375 - 3628 - 1086 = 661$

Source: Ismu Foundation, Regional Observatory for Integration and Multiethnicity

Using the CS to estimate the number of unauthorised migrants



Conclusions

- The CS design is an effective method to integrate survey data with extra information on the aggregation centres

Conclusions

- The CS design is an effective method to integrate survey data with extra information on the aggregation centres

- It has been tested and applied to several surveys (mainly in Italy) to estimate
  - The underlying characteristics of unauthorised migrants populations
  - The size of such populations

Conclusions

- The CS design is an effective method to integrate survey data with extra information on the aggregation centres
- It has been tested and applied to several surveys (mainly in Italy) to estimate
  - The underlying characteristics of unauthorised migrants populations
  - The size of such populations
- Fundamental elements
  - Selection of the centres — prior knowledge of the relative popularity
  - Organisational strucure — well trained interviewers with preferential access to the centres

Thank you!