Session 1 — Methodologies and designs for multi-source processes with non-probability data

Discussion

M. Giovanna Ranalli[1]

[1]Dip. Scienze Politiche, Università degli Studi di Perugia

Second Workshop on Methodologies for Official Statistics
ISTAT — Rome, 6-7 December 2023

Outline

1. Multi-source data: new approaches for non-standard employment statistics. The Dutch and Italian experience
   - Danila Filipponi, Silvia Loriga, Mauricio Garnier Villarreal, Dimitris Pavlopoulos, Roberta Varriale
2. Setting up a statistical register of individuals and dwellings, updated with administrative sources: approach and first steps
   - Aurélien Lavergne
3. Producing U.S.Population statistics using multiple administrative sources
   - J. David Brown and Marta Murray-Close

Outline

1. Multi-source data: new approaches for non-standard employment statistics. The Dutch and Italian experience
   - Danila Filipponi, Silvia Loriga, Mauricio Garnier Villarreal, Dimitris Pavlopoulos, Roberta Varriale
2. Setting up a statistical register of individuals and dwellings, updated with administrative sources: approach and first steps
   - Aurélien Lavergne
3. Producing U.S.Population statistics using multiple administrative sources
   - J. David Brown and Marta Murray-Close

Contribution of the paper

Stream of literature on Hidden Markov Models to integrate LFS and admin data on employment.

Steps forward in this paper

- Focus on mobility trends over time → transition from flexible to permanent employment
- Cross-country comparison → harmonization
- Multiple-group Hidden Markov Models

Overall comments

- The paper is well written, but hard to decrypt: Latent Gold jargon
  - → Parametrization/specification of the models with respect to formulas
- Descriptive vs Analytic Inference
  - → Prediction problem, with a lot of data, information criteria may not be the only useful model selection tools
  - → (Cross-)Validation approach on the non-sampled part of the sample. How do estimates of the true size of temporary/permanent/other employment change according to different models in general and for subpopulations of particular interest? And what about the other transitions (to/from temporary, to/from Other)?
  - → Which is the error of the final rates' estimate?

Overall comments

- The paper is well written, but hard to decrypt: Latent Gold jargon
  - → Parametrization/specification of the models with respect to formulas
- Descriptive vs Analytic Inference
  - → Prediction problem, with a lot of data, information criteria may not be the only useful model selection tools
  - → (Cross-)Validation approach on the non-sampled part of the sample. How do estimates of the true size of temporary/permanent/other employment change according to different models in general and for subpopulations of particular interest? And what about the other transitions (to/from temporary, to/from Other)?
  - → Which is the error of the final rates' estimate?

Overall comments

- The paper is well written, but hard to decrypt: Latent Gold jargon
  - → Parametrization/specification of the models with respect to formulas
- Descriptive vs Analytic Inference
  - → Prediction problem, with a lot of data, information criteria may not be the only useful model selection tools
  - → (Cross-)Validation approach on the non-sampled part of the sample. How do estimates of the true size of temporary/permanent/other employment change according to different models in general and for subpopulations of particular interest? And what about the other transitions (to/from temporary, to/from Other)?
  - → Which is the error of the final rates' estimate?

Overall comments

- The paper is well written, but hard to decrypt: Latent Gold jargon
  - → Parametrization/specification of the models with respect to formulas
- Descriptive vs Analytic Inference
  - → Prediction problem, with a lot of data, information criteria may not be the only useful model selection tools
  - → (Cross-)Validation approach on the non-sampled part of the sample. How do estimates of the true size of temporary/permanent/other employment change according to different models in general and for subpopulations of particular interest? And what about the other transitions (to/from temporary, to/from Other)?
  - → Which is the error of the final rates' estimate?

Specific comments/clarifications

- Dimension of the estimation problem: dataset/sample size?
- Weighted estimates?
- Same error vs. an error?
- Why (c1)/(c2) and (d1)/(d2) have 71/70 parameters?
- Specification of the model
  - "Other" category
  - Latent state must be the same in the two countries (harmonization)
  - Covariates in the structural part (also to include a structural break such as Covid at some point)
  - Model for the transition probabilities: t and t^2 may not be enough, seasonal components
  - Covariates in the measurement part of the model (other than age and proxy: education, survey mode and metadata in LFS; time of completion in ER)

Specific comments/clarifications

- Dimension of the estimation problem: dataset/sample size?
- Weighted estimates?
- Same error vs. an error?
- Why (c1)/(c2) and (d1)/(d2) have 71/70 parameters?
- Specification of the model
  - "Other" category
  - Latent state must be the same in the two countries (harmonization)
  - Covariates in the structural part (also to include a structural break such as Covid at some point)
  - Model for the transition probabilities: t and t^2 may not be enough, seasonal components
  - Covariates in the measurement part of the model (other than age and proxy: education, survey mode and metadata in LFS; time of completion in ER)

Outline

1. Multi-source data: new approaches for non-standard employment statistics. The Dutch and Italian experience
   - Danila Filipponi, Silvia Loriga, Mauricio Garnier Villarreal, Dimitris Pavlopoulos, Roberta Varriale
2. Setting up a statistical register of individuals and dwellings, updated with administrative sources: approach and first steps
   - Aurélien Lavergne
3. Producing U.S.Population statistics using multiple administrative sources
   - J. David Brown and Marta Murray-Close

RESIL

Deep change in the production of Official Statistics in France

- Statistical registers of individuals, dwellings and households
- Reference universe based on the use of a large number of sources (more resilient than tax data only)
- Paradigm shift $\rightarrow$ *It will thus enable the transformation of statistical operations by gradually replacing survey data with administrative data*

Framework

- Very thorough overall picture of the process (plot of the movie we have seen here in the past six years!)
- I appreciate very much the approach of using other countries' experience, methods and software
- I would suggest plan the transition carefully and allow for time to test/evaluate/validate new estimation strategies

Framework

- Very thorough overall picture of the process (plot of the movie we have seen here in the past six years!)
- I appreciate very much the approach of using other countries' experience, methods and software
- I would suggest plan the transition carefully and allow for time to test/evaluate/validate new estimation strategies

Framework

- Very thorough overall picture of the process (plot of the movie we have seen here in the past six years!)
- I appreciate very much the approach of using other countries' experience, methods and software
- I would suggest plan the transition carefully and allow for time to test/evaluate/validate new estimation strategies

Suggestions

- As long as France can afford the large Rolling Census it's being conducted now ($\approx$ 5 million dwellings and 9.3 million inhabitants), keep doing it, in particular for validation of some of the estimation choices to be made in the (near) future
  - evaluation of over/undercoverage of the register $\rightarrow$ production
  - the residency index
  - model based/small area estimation methods
- Linkage service: keep track and provide as much information as possible to secondary users, such as probability of a correct match for subgroups

Suggestions

- As long as France can afford the large Rolling Census it's being conducted now ($\approx$ 5 million dwellings and 9.3 million inhabitants), keep doing it, in particular for validation of some of the estimation choices to be made in the (near) future
  - evaluation of over/undercoverage of the register $\rightarrow$ production
  - the residency index
  - model based/small area estimation methods
- Linkage service: keep track and provide as much information as possible to secondary users, such as probability of a correct match for subgroups

On the residency index

$$I(i,t) = \alpha I(i,t-1) + \beta \sum_{k=1}^{m} a_k(i,t) E_k(i,t)$$

- Many choices to be made
  - Choice of α and β
  - Choice of the weights $a_k(i,t)$
  - Choice of the threshold for $I(i,t)$
- Coverage can be very different across subpopulations/domains

On the residency index

$$I(i,t) = \alpha I(i,t-1) + \beta \sum_{k=1}^{m} a_k(i,t) E_k(i,t)$$

- Many choices to be made
  - Choice of α and β
  - Choice of the weights $a_k(i,t)$
  - Choice of the threshold for $I(i,t)$
- Coverage can be very different across subpopulations/domains

On the residency index

$$I(i,t) = \alpha I(i,t-1) + \beta \sum_{k=1}^{m} a_k(i,t) E_k(i,t)$$

- Residency may be seen as a latent variable hidden behind Signs Of Life
  - Continuous latent construct (IRT models $\rightarrow$ model based weights linked to the discrimination parameters)
  - Categorical variable (latent classes $\rightarrow$ hidden Markov Models) also to cluster profiles

On the residency index

$$I(i,t) = \alpha I(i,t-1) + \beta \sum_{k=1}^{m} a_k(i,t) E_k(i,t)$$

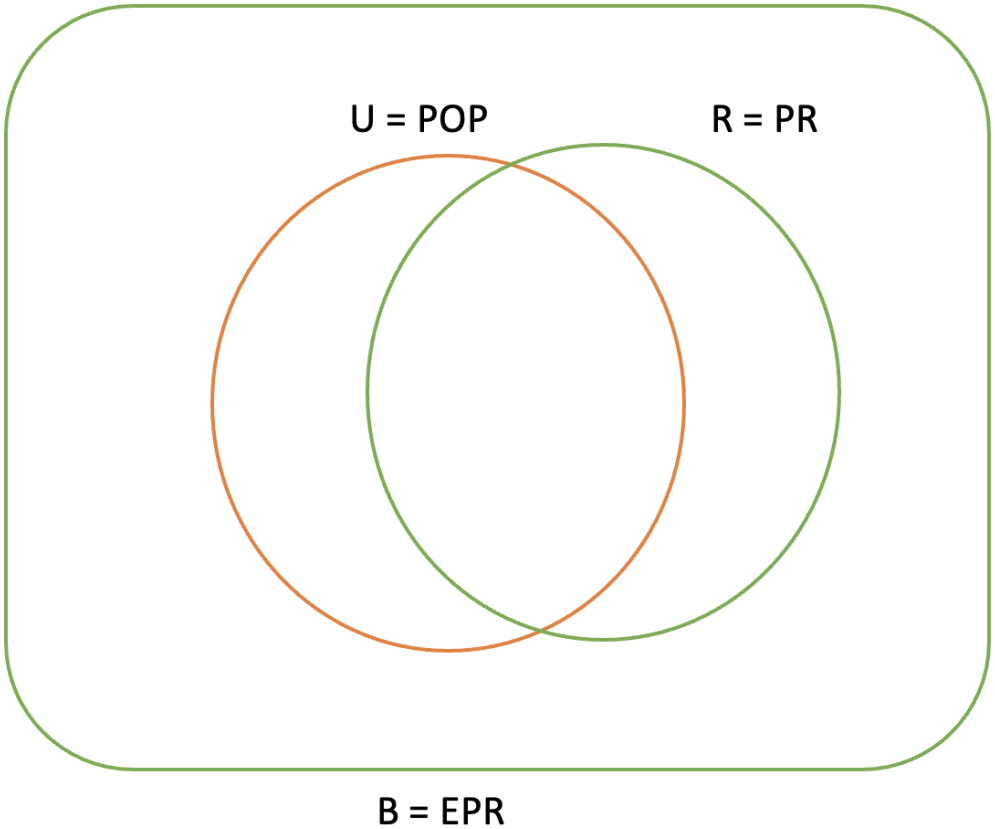
- Residency may be seen as a latent variable hidden behind Signs Of Life
  - Continuous latent construct (IRT models $\rightarrow$ model based weights linked to the discrimination parameters)
  - Categorical variable (latent classes $\rightarrow$ hidden Markov Models) also to cluster profiles

Outline

1. Multi-source data: new approaches for non-standard employment statistics. The Dutch and Italian experience
   - Danila Filipponi, Silvia Loriga, Mauricio Garnier Villarreal, Dimitris Pavlopoulos, Roberta Varriale
2. Setting up a statistical register of individuals and dwellings, updated with administrative sources: approach and first steps
   - Aurélien Lavergne
3. Producing U.S.Population statistics using multiple administrative sources
   - J. David Brown and Marta Murray-Close

Multiple AR-based population statistics

2020 AR census using the principle of redundancy → 31 sources are combined (Extended Pop Register includes SoL)



Challenges

- Locational accuracy
- Person coverage completeness and its consistency across time,
- Coverage of children
- Distinguishing international migrants from continuous U.S. residents
- Choice of demographic characteristics when multiple ones are reported or when they are missing altogether

✓ Demographic Characteristic Accuracy → the issue of race and ethnicity discrepancy between AR and Census can be handled using Hidden Markov Models as in the first paper.

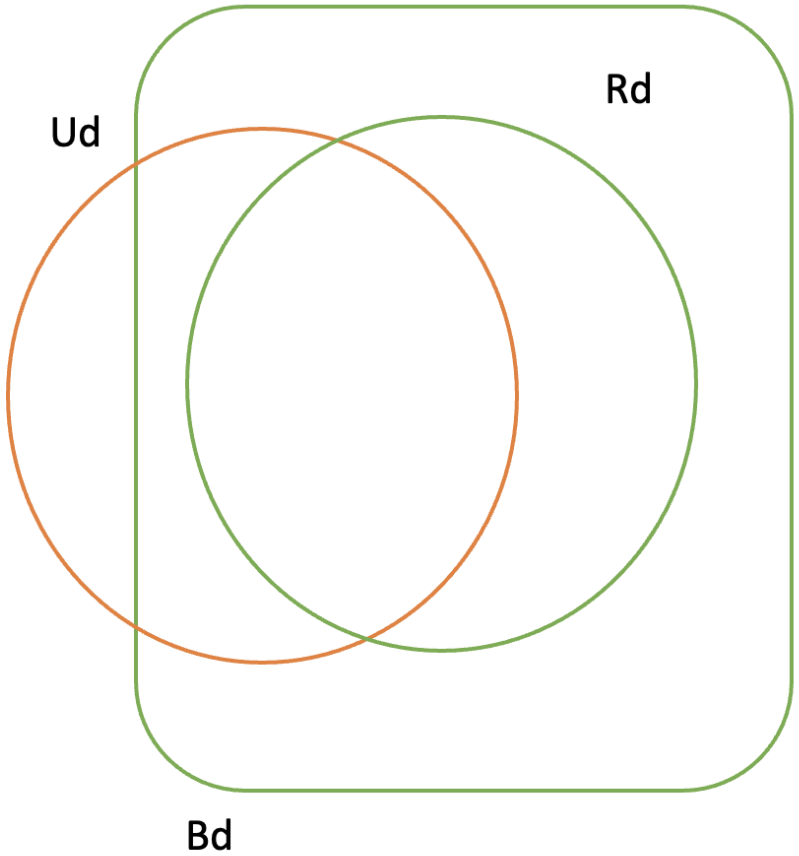
Challenges

- Locational accuracy
- Person coverage completeness and its consistency across time,
- Coverage of children
- Distinguishing international migrants from continuous U.S. residents
- Choice of demographic characteristics when multiple ones are reported or when they are missing altogether

✓ Demographic Characteristic Accuracy → the issue of race and ethnicity discrepancy between AR and Census can be handled using Hidden Markov Models as in the first paper.

Coverage varies by county

Misplacement → inter-locality over- and under-coverage of U_d



Locational accuracy

- Enhancement of AR including all addresses
- ACS to estimate probability that a given address is the person's address on the reference date
- ✓ Fractions of a person may be included in multiple locations → fuzzy assignment instead of a 0-1 assignment
- Audit surveys to validate the many choices required
- In Italy, an area sample survey is used for quality assessment of particularly hard-to-reach profiles/addresses.
- Graph sampling/indirect sampling approach of pairs of individuals and addresses from the Enhanced EPR to assess the prevalent address

Locational accuracy

- Enhancement of AR including all addresses
- ACS to estimate probability that a given address is the person's address on the reference date
- ✓ Fractions of a person may be included in multiple locations → fuzzy assignment instead of a 0-1 assignment
- Audit surveys to validate the many choices required
- In Italy, an area sample survey is used for quality assessment of particularly hard-to-reach profiles/addresses.
- Graph sampling/indirect sampling approach of pairs of individuals and addresses from the Enhanced EPR to assess the prevalent address