Use of Record Linkage in Official Statistics and Feedbacks on Research

Marco Fortini and Tiziana Tuoto

ITACOSM 2019

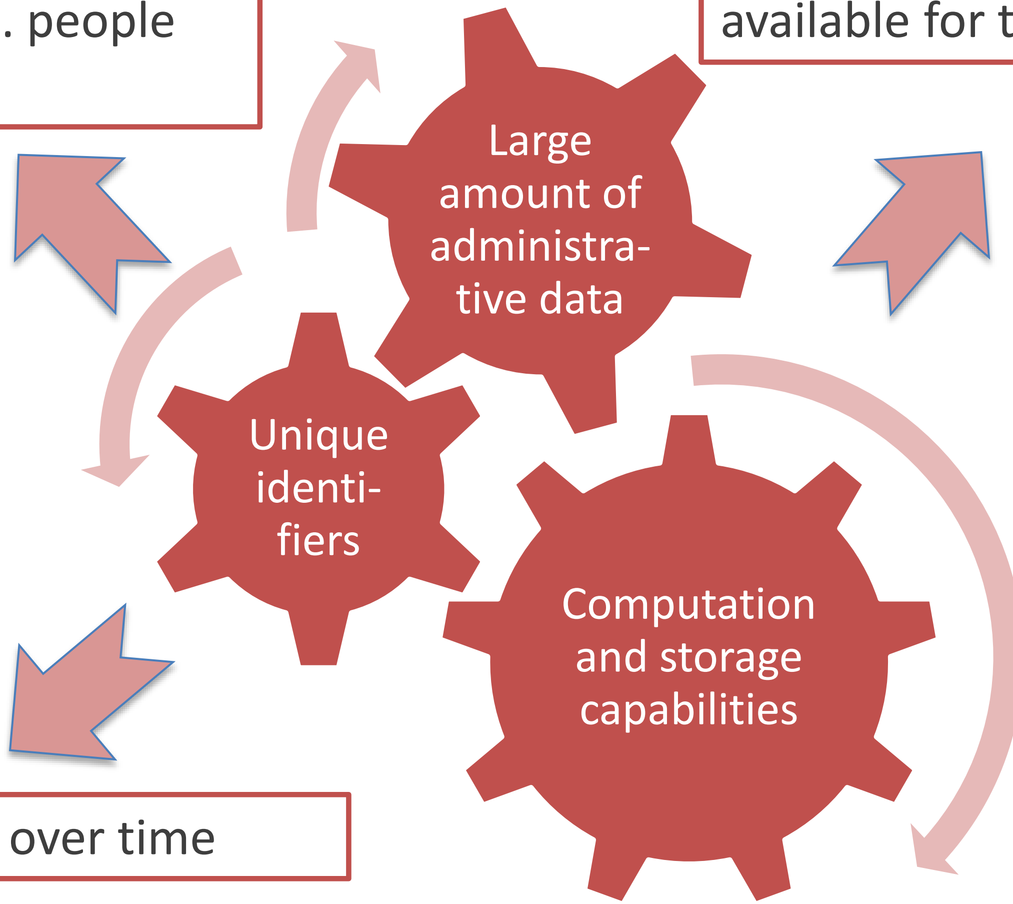
June, 7, 2019 – Florence, Italy

Istat

Toward statistical systems based on registers

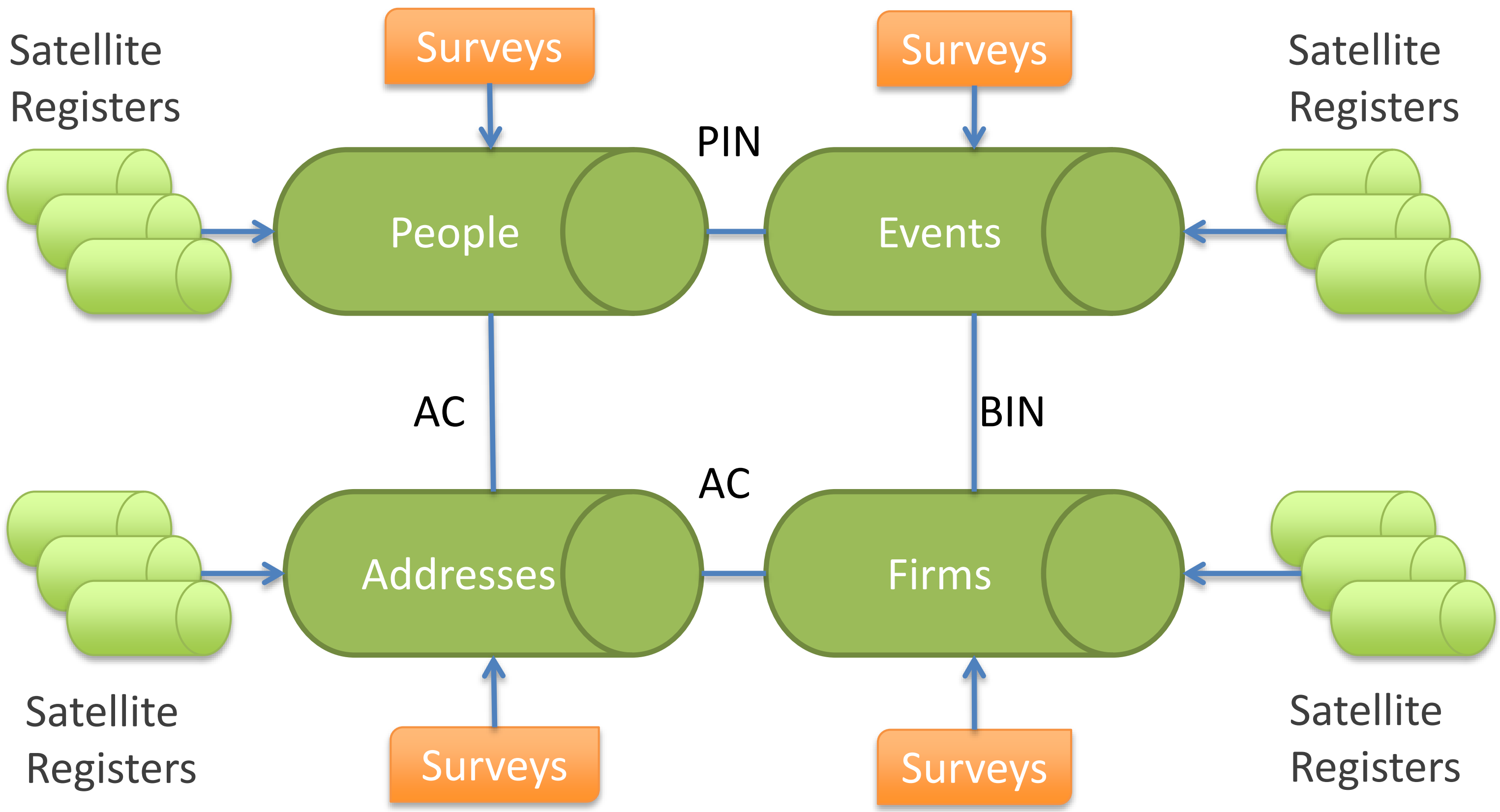
Follow relationship between different units (e.g. people and firms)

Make many variables available for the same unit



Follow changes over time

Istat Register Based Statistical System (RBSS)



AC – Address Code
PIN – Personal Identification Code
BIC - Business Identification Code

Record Linkage in Istat

- Good identification codes
- Mainly deterministic linkage
- Probabilistic linkage is however important to enhance and evaluate quality
  - Emerging phenomena (New sources of data)
  - Sub-population which ID are affected by errors (e.g. foreign people)
  - RELAIS (REcord Linkage At IStat) a specialized software
- Two research topics on possible improvements of probabilistic record linkage in official statistics will be shown

Ingredients for the Record Linkage recipe

- Goal: matching of records relating to the same unit and coming from different sources
- Files A an B of size N_A and N_B
- Pairs (a, b) → Cartesian product Ω (size $\mathrm{N}_{\Omega} = N_A \cdot N_B$)
- Partition of $\Omega = M \cup U$ with $M \cap U = \emptyset$ where
  - M set of matched pairs (same unit)
  - U set of unmatched pairs (different units)
- K common "key" variables $\mathrm{X}_{\mathrm{i},a}, \mathrm{X}_{\mathrm{i},b};\ \mathrm{i} = 1, \dots, \mathrm{K}, (a, b) \in \Omega$
- Vector $\boldsymbol{\gamma} = \{\boldsymbol{\gamma}_k, k = 1, \dots, k\}$ of agreement/disagreement between key variables (2^K possible patterns)

Istat

Linkage Probabilities (Fellegi and Sunter, 1969)

- Pairs sharing $\boldsymbol{\gamma}$ return the same evidence to be matched
- We model the 2^K frequencies $\mathrm{N}_{\boldsymbol{\gamma}}$ of pairs by patterns $\boldsymbol{\gamma}$

Matrix of observed data for 3 Key variables

γ_1	γ_2	γ_3	$N_{\boldsymbol{\gamma}}$	$p_{\boldsymbol{\gamma}}$
1	1	1	N_{111}	p_{111}
1	1	0	N_{110}	p_{110}
1	0	1	N_{101}	p_{101}
0	1	1	N_{011}	p_{011}
0	0	1	N_{001}	p_{001}
0	1	0	N_{010}	p_{010}
1	0	0	N_{100}	p_{100}
0	0	0	N_{000}	p_{000}
		Tot	N_{Ω}	1

Observed data can be seen as mixture

$$p_{\boldsymbol{\gamma}} = \mathrm{P}(M)\mathrm{P}(\boldsymbol{\gamma}|M) + \big(1 - \mathrm{P}(M)\big)\mathrm{P}(\boldsymbol{\gamma}|U)$$

or, in more compact notation

$$p_{\boldsymbol{\gamma}} = p \cdot m_{\boldsymbol{\gamma}} + (1 - p) \cdot \mathrm{u}_{\boldsymbol{\gamma}}$$

- We aim to estimate of the fraction $\pi_{\boldsymbol{\gamma}} = \mathrm{P}((a, b) \in M|\boldsymbol{\gamma}), \forall \boldsymbol{\gamma}$ of matched pairs among those showing the pattern $\boldsymbol{\gamma}$

Estimate of linkage probabilities p, m_γ and u_γ

- Estimated by frequencies N_γ with Latent class modelling and EM algorithm (Jaro, 1989)
  - At least 3 key variables
  - Conditional independence assumption

$m_\gamma = \prod_k m_k$ and $u_\gamma = \prod_k u_k$ $\rightarrow$ $\hat{u}_k, \widehat{m}_k, k = 1, \dots, K$

- Bayes rule: probability of a pair to be matched given its evidence $\boldsymbol{\gamma}$

$$\pi_\gamma = \frac{p \cdot m_\gamma}{p \cdot m_\gamma + (1-p) \cdot u_\gamma}$$

- Best patterns: $\boldsymbol{\gamma}$: $\pi_\gamma \cong 1$

Toy example 1: moderate files size, unbiased estimates

Latent distributions

Matrix of the observed data for K=4

γ_1	γ_2	γ_3	γ_4	N_γ	$N_{M,\gamma}$	$N_{U,\gamma}$
1	1	1	1	821	815	6
1	1	1	0	162	43	119
1	1	0	1	162	43	119
1	1	0	0	2256	2	2254
1	0	1	1	162	43	119
1	0	1	0	2256	2	2254
1	0	0	1	2256	2	2254
1	0	0	0	42826	0	42826
0	1	1	1	162	43	119
0	1	1	0	2256	2	2254
0	1	0	1	2256	2	2254
0	1	0	0	42826	0	42826
0	0	1	1	2256	2	2254
0	0	1	0	42826	0	42826
0	0	0	1	42826	0	42826
0	0	0	0	813692	0	813692
			Tot	1000000	1000	999000

Files

$N_A = N_B = 1000$

$k = 1, \ldots, 4$ key variables

Parameters

$m_k = 0.95; u_k = 0.05 \; \forall k$

$p = .001$

Unbiased Estimates

$$\left.\begin{array}{l}\hat{m}_k = 0.9494 \\ \hat{u}_k = 0.0500\end{array}\right\} \forall k$$

$\hat{p} = 0.001$

With large files size LCA estimates become biased

Matrix of the observed data for K=4

Latent distributions

γ_1	γ_2	γ_3	γ_4	N_γ	$N_{M,\gamma}$	$N_{U,\gamma}$
1	1	1	1	821	815	6
1	1	1	0	162	43	119
1	1	0	1	162	43	119
1	1	0	0	2256	2	2254
1	0	1	1	162	43	119
1	0	1	0	2256	2	2254
1	0	0	1	2256	2	2254
1	0	0	0	42826	0	42826
0	1	1	1	162	43	119
0	1	1	0	2256	2	2254
0	1	0	1	2256	2	2254
0	1	0	0	42826	0	42826
0	0	1	1	2256	2	2254
0	0	1	0	42826	0	42826
0	0	0	1	42826	0	42826
0	0	0	0	813692	0	813692
			Tot	1000000	1000	999000

When files grows

$$p = \frac{N_M}{N_\Omega} \to 0$$

estimates are biased

In real world

$$p < 0.001$$

Two files of 1000 units

Solution:
Filtering

Side effects:
some matches can be missed with unknown risk

Toy example 2: large files size, biased estimates

Latent distributions

Matrix of the observed data for K=4

γ_1	γ_2	γ_3	γ_4	N_γ	$N_{M,\gamma}$	$N_{U,\gamma}$
1	1	1	1	8770	8145	625
1	1	1	0	12303	429	11874
1	1	0	1	12303	429	11874
1	1	0	0	225625	23	225602
1	0	1	1	12303	429	11874
1	0	1	0	225625	23	225602
1	0	0	1	225625	23	225602
1	0	0	0	4286448	1	4286446
0	1	1	1	12303	429	11874
0	1	1	0	225625	23	225602
0	1	0	1	225625	23	225602
0	1	0	0	4286448	1	4286446
0	0	1	1	225625	23	225602
0	0	1	0	4286448	1	4286446
0	0	0	1	4286448	1	4286446
0	0	0	0	81442480	0	81442480
			Tot	100000000	10000	99990000

Files

$N_A = N_B = 10000$

$N_\Omega = 100000000$

$N_M = 10000$

$k = 1, \dots, 4$ key variables

Parameters

$m_k = .95;\ u_k = .05, \forall k$

$p = .0001$

Biased Estimates

$\widehat{m}_k = .0593;\ \hat{u}_k = .0421, \forall k$

$\hat{p} = .4641$

Idea: robust EM estimation

- Estimates $\hat{u}_k, \widehat{m}_k, k = 1, \ldots, K$ are obtained trimming expected distributions $\widehat{N}_{M,\boldsymbol{\gamma}}$ of matched and $\widehat{N}_{U,\boldsymbol{\gamma}}$ of unmatched pairs during step M of EM algorithm
- Structural zeros are included in patterns that are expected having low frequencies under true model
- Example: estimate of m_1 for K=3 (remind: $\widehat{N}_{M,\boldsymbol{\gamma}} = N_{\boldsymbol{\gamma}} \cdot \hat{\pi}_{\boldsymbol{\gamma}}$)

$$\text{Standard M step } \widehat{m}_1 = \frac{\sum_{\gamma_2\gamma_3} \widehat{N}_{M,1,\gamma_2\gamma_3}}{\sum_{\gamma_1\gamma_2\gamma_3} \widehat{N}_{M,\gamma_1,\gamma_2\gamma_3}}$$

$$\text{Robust M step } \widehat{m}_1 = \frac{\widehat{N}_{M,111}}{\widehat{N}_{M,111}+\widehat{N}_{M,011}}$$

Robust EM estimation estimates return unbiased

Latent distributions

Matrix of the observed data for K=4

γ_1	γ_2	γ_3	γ_4	N_γ	$N_{M,\gamma}$	$N_{U,\gamma}$
1	1	1	1	8770	8145	625
1	1	1	0	12303	429	11874
1	1	0	1	12303	429	11874
1	1	0	0	225625	23	225602
1	0	1	1	12303	429	11874
1	0	1	0	225625	23	225602
1	0	0	1	225625	23	225602
1	0	0	0	4286448	1	4286446
0	1	1	1	12303	429	11874
0	1	1	0	225625	23	225602
0	1	0	1	225625	23	225602
0	1	0	0	4286448	1	4286446
0	0	1	1	225625	23	225602
0	0	1	0	4286448	1	4286446
0	0	0	1	4286448	1	4286446
0	0	0	0	81442480	0	81442480
			Tot	100000000	10000	99990000

Unbiased Estimates

$$\left.\begin{aligned}\widehat{m}_k &= 0.9499\\ \hat{u}_k &= 0.0500\end{aligned}\right\}\forall k$$

$$\hat{p} = .0001$$

On simulated data with two files of 10^7 records each ($N_\Omega = 10^{14}$) estimates remain unbiased

Example – Population register vs Permits to stay

- 19,398 foreign people in population register
- 16,723 people applying for a permit to stay (new or renewal)
- N_Ω =324,392,754 pairs in Cartesian product between files
- 6 Key variables
  First and last name (single field), Gender, Code of Country citizenship, Day of birth, Month of birth, Year of birth

Probability of agreement for each key variable conditioned to match status of the pair

Estimation method	p (N_Ω)	m, u	First last Names	Gender	Country ID	Day of birth	Month of birth	Year of birth
		k	(1)	(2)	(3)	(4)	(5)	(6)
Standard	0.212	m_k	0.998	0.514	0.758	0.989	0.917	0.967
	$(324 \cdot 10^6)$	u_k	1.000	0.544	1.000	0.988	0.918	0.967
Standard blocking	0.006	m_k	0.997	0.935	0.863	0.991	0.989	0.987
	(861.000)	u_k	0.026	0.479	0.201	0.012	0.082	0.032
Robust	0.0000186	m_k	0.984	0.887	0.790	0.957	0.964	0.963
	$(324 \cdot 10^6)$	u_k	0.000	0.461	0.049	0.012	0.082	0.032

Linking with less than three variables

- Conditions on comparison variables
  1. Binary functions
  2. Conditional independence between each other
  3. At least three comparison variables
- Overcome points 1 and 3 through mixtures other than multinomial models
- Example: Geocoding of address location
  - Less than three variables
  - Real value distance between strings in [0,1] interval

RL of addresses : The model

- ONLY 2 key variables

1. Street type (ST)
   e.g. *via*, *strada*, *viale*, etc (street, avenue, square,…)
   – 0-1 variable γ_{ST}
      - 1 if Levenshtein distance$\leq$2
      - 0 otherwise
2. Street name (SN):
   – Continuous variable in [0, 1] δ_{SN}
   – Comparison via Jaccard distance

Ex: “V.le G.B. Morgagni” *.vs.* “Viale Giovanni Battista Morgagni”

$$\boldsymbol{\gamma} = (\gamma_{ST}, \delta_{SN}) = (2, 0.7)$$

RL of addresses : The model

- We propose a mixture of beta and Bernoulli distributions
- Conditional independence between beta and Bernoulli is assumed
- Street type (ST) – Bernoulli distr. on random variable γ_{ST}
  - Be(θ_M) given the pairs are in M
  - Be(θ_U) given the pairs are in U
- Street name (SN) – Beta distr. on random variable δ_{SN}
  - Beta(α_M, β_M) for the pairs in M
  - Beta(α_U, β_U) for the pairs in U

$$P(\boldsymbol{\gamma}) = p \cdot Beta(\alpha_M, \beta_M) \cdot \mathrm{Be}(\theta_M) + (1 - p) \cdot Beta(\alpha_U, \beta_U) \cdot \mathrm{Be}(\theta_U)$$

The case study: RL of addresses

- 4 small municipalities of region Umbria
- 2434 addresses in local registry have to be standardised
- Key variables: street type (ST), street name (SN)
  - Addresses can be written in several ways
- 900 standard format streets names from thesaurus
- 527,117 pairs to be assigned

Some results

- Starting values for EM algorithm

p= 0.01 α_M=1 α_U= 3 m_{ST} =.9

β_M= 1 β_U= 1 u_{ST} =.1

- EM algorithm converges after 67 iterations

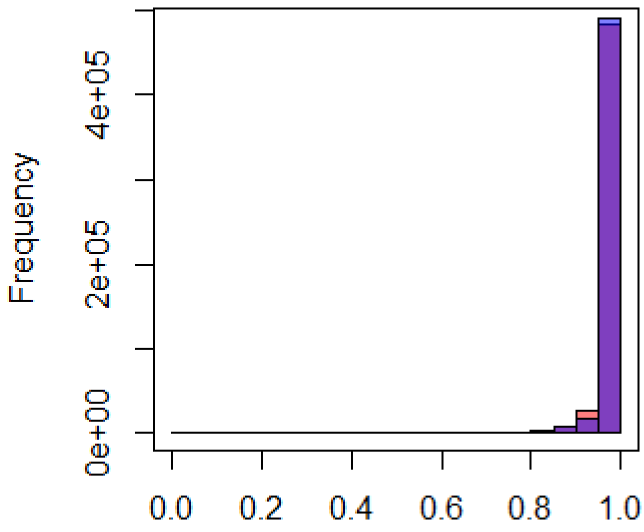
p=0.0051 α_M= 0.0888 α_U=7.5906 m_{ST} = 0.7320

β_M= 0.0851 β_U= 0.0852 u_{ST} = 0.2240

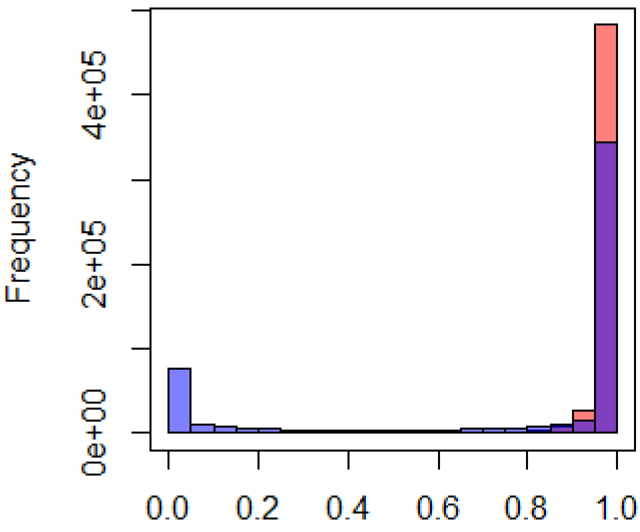
Model fitting

Histograms observed vs expected distance

Expected distance from mixture of beta distributions



Expected distance from beta distribution on marginal data

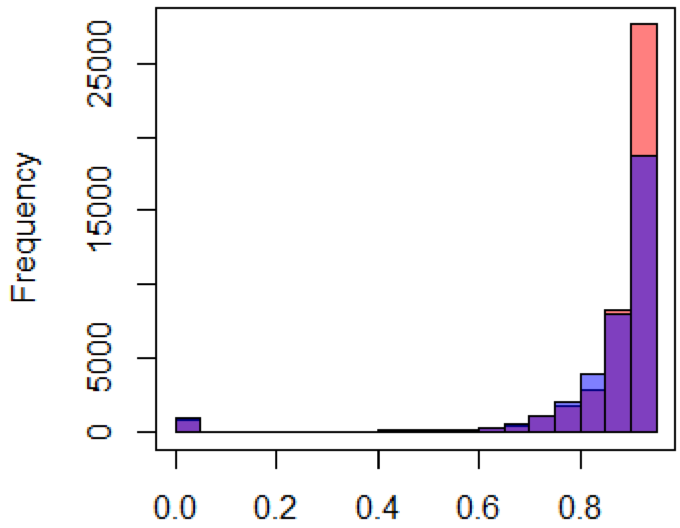


Istat

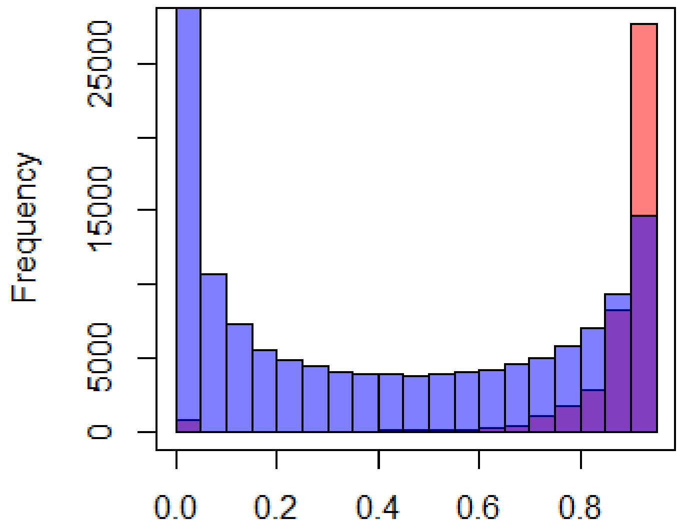
Zooming the model fitting

Histograms observed vs expected distance
Pairs which distance is less than 0.95

Expected distance from mixture of beta distributions



Expected distance from beta distribution on marginal data



Comparison with gold standard

- 2434 addresses linked to the best candidate from thesaurus
- Their match status was learned by manual checking
- False and true matches are showed according to class of posterior probability π_γ

	Classes of posterior probability π_γ				
Match	[0-0.2]	(0.2-0.4]	(0.4-0.6]	(0.6-0.8]	(0.8-1]
False	229	10	5	5	12
(%)	21.7	6.7	3.2	3.8	1.3
True	825	139	153	126	930
(%)	**78.3**	**93.3**	**96.8**	**96.2**	**98.7**

Method is **sensitive** but not very **specific**
Due to bad parsing of addresses the during pre-processing

Concluding remarks

- Linkage is a relevant procedure in official statistics
- Probabilistic record linkage helps achieving higher quality
- But it needs of improvements to better deal with real data
- We showed two relevant research topics in official statistics context
- Experimental applications to data production are at their starting point

Thank you for your attention