Methodological proposals I: general framework, ecological sampling, and geostatistics

WP5 Mobile Phone Data

E. Esteban, S. Saldaña, **D. Salgado**, L. Sanguiao

Statistics Spain (INE)

Madrid, 7-8 June, 2017

Overview

INSTITUTO NACIONAL DE ESTADISTICA

- **Definition** of the problem
- **Framework**: general lines
- More concrete **proposals**
- Diverse **remarks**

The problem: data sets

- **Two type of data sets**:
  - Data at **mobile device** level (CDR/DDR, signaling, ...): $\mathbf{z}_k$, $k \in s$.
  - Data at an **aggregated** level (number of devices, number of tourists, ...): $\dot{Y}_{s_i}$, $i \in \mathcal{I}_s$.
- Work at **two steps**:
  - From **microdata** $\mathbf{z}_k$ to **partial aggregates** $\dot{Y}_i$, $i \in \mathcal{I}_s$.
  - From **partial aggregates** $\dot{Y}_i$ to **estimators** $\widehat{Y}_i$ of target aggregates Y_{U_i}.

The problem: target aggregates

- **Daytime** population size $\rightsquigarrow N_U$
- **Number of** inbound/domestic **tourists** $\rightsquigarrow N_{U_T}$
- **Number of commuters** $\rightsquigarrow N_{U_C}$

Methodologically reduced to a problem of ***count statistics***

Equivalently reduced to a problem of ***population size estimation***

The problem: formulation (I)

Let

- a **finite population** U of interest of **unknown size** N;
- a partition into **population domains** U_i such that $U = \bigcup_{i \in \mathcal{I}_U} U_i$, mostly **geographical** areas;
- either a **mobile device** (micro)**data** set s or a collection of **partial aggregates** $\dot{N}_{s_i}$ of **count statistics** of interest.

We want to construct:

- an **estimator** $\hat{N}$ for the population size N;
- an **accuracy measure** for $\hat{N}$;
- **estimators** $\hat{N}_{U_i}$ for the population domain sizes N_{U_i};
- **accuracy measures** for $\hat{N}_{U_i}$.

The problem: formulation (II)

Let

- a **finite population** U of interest of **unknown size** N;
- a partition into **population domains** U_i such that $U = \bigcup_{i \in \mathcal{I}_U} U_i$, mostly **geographical** areas;
- a collection of time intervals t;
- either a **mobile device** (micro)**data** set s or a time series of collections of **partial aggregates** $\dot{N}_{s_{it}}$ of **count statistics** of interest.

We want to construct:

- a time series of **estimators** $\hat{N}_t$ for the population size N_t in each time interval t;
- a time series of **accuracy measure** for $\hat{N}_t$;
- a time series **estimators** $\hat{N}_{U_{it}}$ for the population domain sizes $N_{U_{it}}$;
- a time series of **accuracy measures** for $\hat{N}_{U_{it}}$.

General setting

- Avoid **file drawer problem**
- Assembling of diverse **statistical/AI methods**
  - The **representativeness** problem $\rightsquigarrow$ ecological sampling, Heckman correction, ...
  - **Geospatial components** in our problem $\rightsquigarrow$ geostatistical modelling, misalignment problem, ...
  - Role of **official data** as auxiliary information.

Ecological sampling

Main idea

Our problem is indeed an ecological **animal abundance problem** where the **sampling protocol** is undertaken by the **mobile telecommunication network** and, in some cases (tourists, commuters, . . .), the **processing of microdata** at the mobile device level.

A naive example

Let

- N be the size of the target population U;
- Y be the number of mobile phone detected in the network;
- p the detection probability of a statistical unit $k \in U$.

We then model $Y \simeq \mathrm{Bin}(N, p)$ so that

- $\hat{N} = \frac{Y}{p}$ is an unbiased estimator for N:

$$\mathbb{E}\left[\frac{Y}{p}\right] = N.$$

- $\widehat{\mathbb{V}}\left(\hat{N}\right) = Y \cdot \frac{1-p}{p^2}$ is an unbiased estimator for $\mathbb{V}\left[\hat{N}\right]$:

$$\mathbb{E}\left[\widehat{\mathbb{V}}\left(\hat{N}\right)\right] = N \cdot \frac{1-p}{p} = \mathbb{V}\left(\hat{N}\right).$$

First elements for the solution

- **Unbiasedness**
- **Variance estimation**
- The population size N to estimate $\rightsquigarrow$ **target process** element
- The detection probability $p \rightsquigarrow$ **observation process** element

Need for more **realistic model** (more elements)

More realistic elements

- p cannot be realistically assumed to be known:

$$p \simeq \mathrm{Model}(\theta).$$

- p can possibly vary from cell to cell in the network:

$$p_i \simeq \mathrm{Model}(\theta_i).$$

- Thus, we are estimating each local partial aggregate

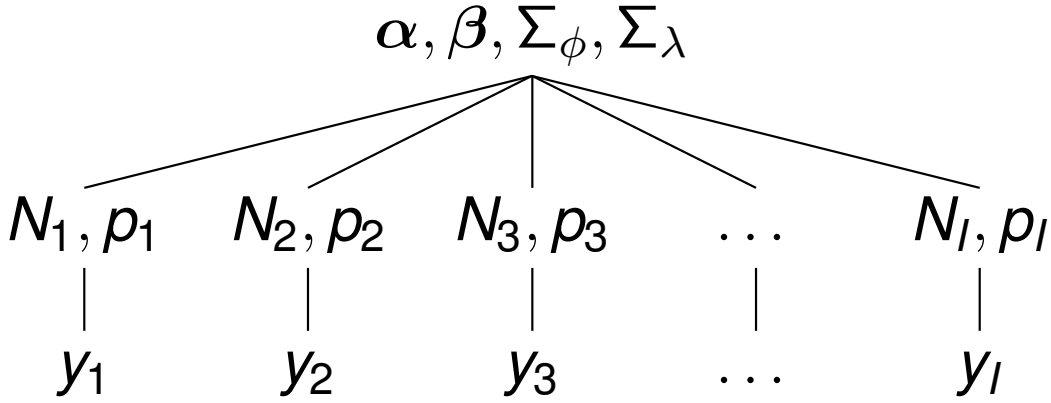
$$(Y_i, p_i) \rightsquigarrow N_{U_i}.$$

Metapopulation

Metapopulation parameters

Local parameters

Data



Hierarchical models

A (not so simple) example:

$$\begin{aligned}
[\ y_i \ | \ N_i, p_i \ ] &\simeq \mathrm{Bin}(N_i, p_i) \\
[\ N_i \ | \ \lambda_i \ ] &\simeq \mathrm{Po}(\lambda_i) \\
[\ \mathrm{logit}\,(p_i) \ | \ \beta, \Sigma_\phi \ ] &\simeq \mathrm{N}\left(\beta \cdot \mathbf{x}', \Sigma_\phi\right) \\
[\ \lambda_i \ | \ \alpha, \Sigma_\lambda \ ] &\simeq N(\alpha, \Sigma_\lambda)
\end{aligned}$$

General philosophy:

$$[\mathrm{data}|\mathrm{process}, \mathrm{parameters}][\mathrm{process}|\mathrm{parameters}][\mathrm{parameters}]$$

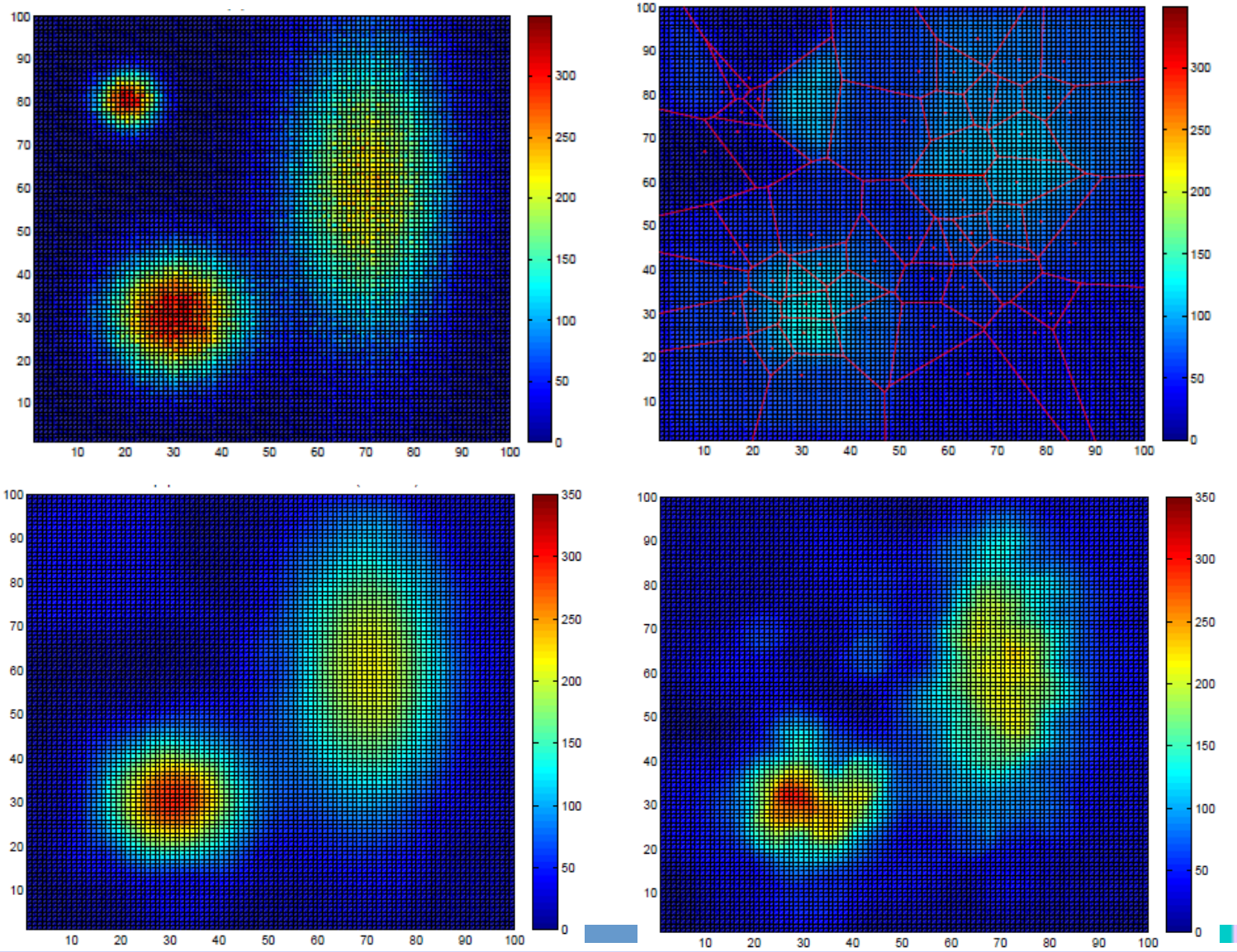
Then:

$$n_i = \mathbb{E}\left[N_i|\mathrm{data}\right]$$

Geospatial considerations

- **Spatial variations** seem apparently to play a relevant role.
- Spatial variations can be accounted for by using **geostatistical modelling techniques** $(\Sigma_\phi, \Sigma_\lambda)$.

Misalignment problem (Ricciato et al., 2015)



Role of official data

- Calibrate to national official population N^{official}

  Example (Deville et al., 2014):

  $\rho_{\text{official}} = \alpha \cdot \sigma_{\text{MD}}^{\beta}$, then $\rho = \frac{P^{\text{official}}}{\hat{P}} \cdot \alpha \cdot \sigma_{\text{MD}}^{\beta}$.

- Do not calibrate to national official population, but only use official figures as covariates **x**

  Example $\text{logit}\,(p_i) = \beta_0 + \beta_1 \cdot \text{skewness}_i^{\text{income}} + e_i$.

Mobile devices per statistical unit

- Estimates on the **number of statistical units**, **not of mobile devices**.
- Multiplicity of devices accounted by **penetration rates** τ (see Internal Doc. 3).
- **Model jointly** $\tau_i \cdot p_i$ instead of p_i.

From microdata to partial aggregates

- **Mining microdata sets** in MNOs' hands (with exceptions): restrictions due to data availability (e.g. longitudinal analysis).
- **Internal technical report** by Positium.
- Apparently, mining algorithms are **highly sensitive to domain** (human populations, tourists, mobility–commuters,. . .).

Heckman correction

INSTITUTO NACIONAL DE ESTADISTICA

See presentation by INSSE (B. Oancea).

Admin data methodology

INSTITUTO NACIONAL DE ESTADISTICA

See presentation by INSSE (C. Alexandru).

Inference philosophy

- **Frequentist** vs. **Bayesian** methods in models.
- A priori, Bayesian approach seems **more natural**.
- **Pragmatic** approach $\rightsquigarrow$ **availability of computer tools** at first stage.