INSTITUTO NACIONAL DE ESTADISTICA

# Inference with Mobile Network Data

## NTTS 2019

**D. Salgado**[1], L. Sanguiao[1], B. Oancea[2]

[1]Statistics Spain (INE)

[2] Statistics Romania (INS)

ESSnet on Big Data II, WPI on Mobile Network Data

Brussels, 12-14 March, 2019

INSTITUTO NACIONAL DE ESTADISTICA

# The context: ESSnets on Big Data

- **ESSnet on Big Data I** (2016-2018)

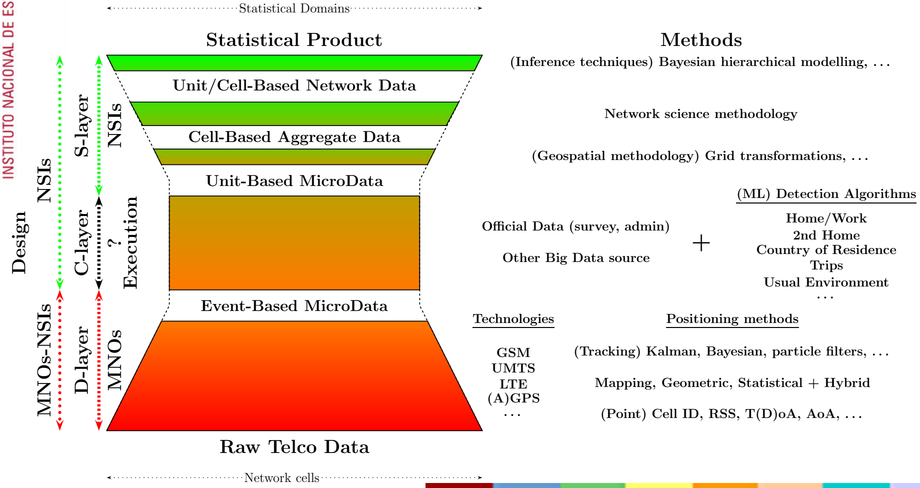
| | | |
|---|---|---|
| Online Job Vacancies | Enterprise Characteristics | Smart Meters Data |
| AIS Data | Mobile Phone Data | Early Estimates |
| Multiple Domains | Dissemination | Methodology |

- **ESSnet on Big Data II** (2018-2020)
  - Track on **Implementation**: Online Job Vacancies, Enterprise Characteristics, Smart Energy, Tracking Ships, Process and Architecture
  - Track on **New Pilots**: Financial Transaction Data, Earth Observation, Mobile Network Data, Smart Tourism, Methodology and Quality
  - Track on **Trusted Smart Statistics**

**Website: Search *ESSnet Big Data***

# Towards a production framework

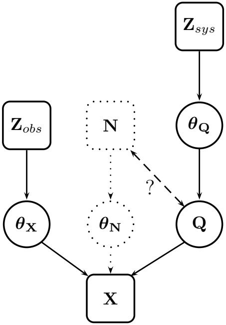
## **Reference Methodological Framework** (Ricciato et al. 2017, 2018)



# Inference Framework

- **Bayesian inference** with hierarchical models
  - **Not new** in Official Statistics: SAE, Population Statistics (Bryant and Graham, 2015).
  - **Integration of data sources**:

    $\mathbf{Z}_{obs}, \mathbf{Z}_{sys} \rightsquigarrow$ Population registers, survey/admin data, penetration rates...
  - **Quality** issues:
    - Point estimates: MAP, posterior mean and median.
    - Interval estimates: credible intervals, posterior variance.
    - Model checking: using the posterior predictive distribution




$$\mathbb{P}\left(\mathbf{Q} = \mathbf{N} \middle| \mathbf{X} = \mathbf{N}^{(D)}, \mathbf{Z}_{obs}, \mathbf{Z}_{sys}\right)$$

$\mathbf{N}$ : individuals

$\mathbf{N}^{(D)}$ : devices

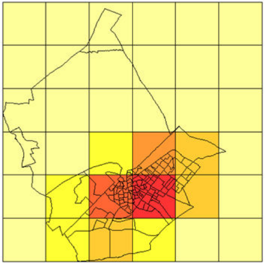
$$\mathbb{E}\left[N_r^{(\mathrm{D}),\,\mathrm{rep}} - N_r^{(\mathrm{D})} | N_r^{(\mathrm{D})}\right] \qquad \mathbb{E}\left[\frac{N_r^{(\mathrm{D}),\,\mathrm{rep}} - N_r^{(\mathrm{D})}}{N_r^{(\mathrm{D})}} | N_r^{(\mathrm{D})}\right]$$

$$\mathbb{V}\left[N_r^{(\mathrm{D}),\,\mathrm{rep}} - N_r^{(\mathrm{D})} | N_r^{(\mathrm{D})}\right] \qquad \mathbb{V}\left[\frac{N_r^{(\mathrm{D}),\,\mathrm{rep}} - N_r^{(\mathrm{D})}}{N_r^{(\mathrm{D})}} | N_r^{(\mathrm{D})}\right]$$

$$\mathbb{E}\left[\left(N_r^{(\mathrm{D}),\,\mathrm{rep}} - N_r^{(\mathrm{D})}\right)^2 | N_r^{(\mathrm{D})}\right] \qquad \mathbb{E}\left[\left(\frac{N_r^{(\mathrm{D}),\,\mathrm{rep}} - N_r^{(\mathrm{D})}}{N_r^{(\mathrm{D})}}\right)^2 | N_r^{(\mathrm{D})}\right]$$

# Inference: preliminary results. I

INSTITUTO NACIONAL DE ESTADISTICA



- **Original Data**:

  Population Register, Market & Competence Regulator

- **Simulated Data**:

  Perturbed Values $\rightsquigarrow N_i^{(\mathrm{REG})}(t_0), N_i^{(\mathrm{D})}(t_0)$

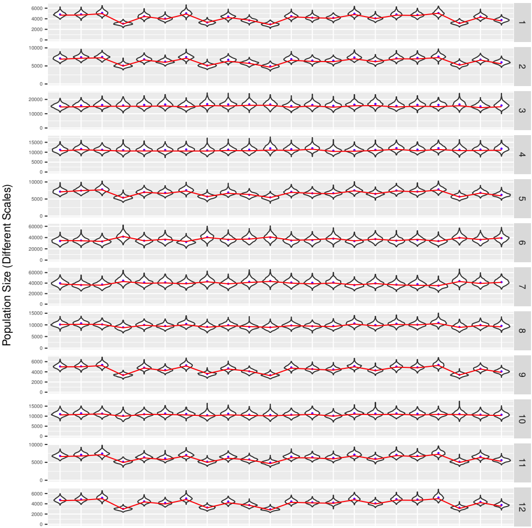
- **Simulated Dynamics**:

  $\mathbf{N}^{(0)}(t) = \mathbf{N}^{(0)T}(t_0)p(t)$, $p(t) = [p_{ij}(t) \propto d_{ij}^{-1}(t)]_{1 \leq ij \leq N_c}$

- **Toy data set**:

| ID_CELL_INI | ID_CELL_END | ID_T | N_REG | N_0 | N_D_1 |
|---|---|---|---|---|---|
| 1 | 1 | 0 | 1252 | 885 | 218 |
| 2 | 1 | 0 | 2770 | 123 | 42 |
| 3 | 1 | 0 | 16571 | 336 | 122 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 10 | 12 | 671 | 9866 | 229 | 97 |
| 11 | 12 | 671 | 2859 | 141 | 42 |
| 12 | 12 | 671 | 1130 | 789 | 183 |

# Inference: preliminary results. II




- Posterior **distributions**
- **Point estimates**: median, mean, mode...
- **Credible intervals**
- **Model checking** through the **posterior predictive distribution**
- R package: **pestim**

https://github.com/MobilePhoneESSnetBigData/pestim

# Inference at $t_0$: no structure for the system

| Model | Posterior | MAP | Mean |
|---|---|---|---|
| **Sys:** $N_r \stackrel{\text{indep}}{\simeq} \delta_N$<br>**Obs:** $N_r^{(\mathrm{D})} \mid N_r, p_r \stackrel{\text{indep}}{\simeq} \mathrm{Bin}\left(N_r, p_r\right)$<br>$p_r \stackrel{\text{indep}}{\simeq} \delta_{p_r^*}$ | $N_r^{(\mathrm{D})} + \text{Neg-Bin}\left(N_r^{(\mathrm{D})} + 1, 1 - p_r^*\right)$ | $N_r^{(D)} + \left\lfloor \frac{N_r^{(D)}}{p_r^*} - N_r^{(D)} \right\rfloor$ | $\frac{N_r^{(D)}}{p_r^*}$ |
| **Sys:** $N_r \stackrel{\text{indep}}{\simeq} \delta_N$<br>**Obs:** $N_r^{(\mathrm{D})} \mid N_r, p_r \stackrel{\text{indep}}{\simeq} \mathrm{Bin}\left(N_r, p_r\right)$<br>$p_r \stackrel{\text{indep}}{\simeq} \mathrm{Beta}(\alpha_r, \beta_r)$<br>$(\alpha, \beta) \stackrel{\text{indep}}{\simeq} \delta_{\alpha^*} \times \delta_{\beta^*}$ | $\approx N_r^{(\mathrm{D})} + \text{Neg-Bin}\left(N_r^{(\mathrm{D})} + 1, \frac{\beta_r^*}{\alpha_r^* + \beta_r^*}\right)$ | $N_r^{(D)} + \left\lfloor \frac{N_r^{(D)}}{\frac{\alpha_r^*}{\alpha_r^* + \beta_r^*}} - N_r^{(D)} \right\rfloor$ | $\frac{N_r^{(D)}}{\frac{\alpha_r^*}{\alpha_r^* + \beta_r^*}}$ |
| **Sys:** $N_r \stackrel{\text{indep}}{\simeq} \delta_N$<br>**Obs:** $N_r^{(\mathrm{D})} \mid N_r, p_r \stackrel{\text{indep}}{\simeq} \mathrm{Bin}\left(N_r, p_r\right)$<br>$p_r \stackrel{\text{indep}}{\simeq} \mathrm{Beta}(\alpha_r, \beta_r)$<br>$\left(\mathrm{logit}\left(\frac{\alpha_r}{\alpha_r + \beta_r}\right), \alpha_r + \beta_r\right) \mid \gamma_0, \gamma_1, \sigma_\gamma^2, P_r \stackrel{\text{indep}}{\simeq}$<br>$\mathrm{N}\left(\log\left(\gamma_0 \left(\frac{P_r}{1 - P_r}\right)^{\gamma_1}\right), \sigma_\gamma^2\right) \times \mathrm{Gamma}\left(1 + \xi, \frac{e^{\mu_r}}{\xi}\right)$<br>$\left(\gamma_0, \gamma_1, \sigma_\gamma^2\right) \simeq \delta_{\gamma_0^*} \times \delta_{\gamma_1^*} \times \delta_{\sigma_\gamma^{*2}}$<br>$(\xi, \mu_r) \stackrel{\text{indep}}{\simeq} \delta_{\xi^*} \times \delta_{\mu_r^*}$ | $\approx \mathbf{N}^{(\mathrm{D})} + \prod_r \int_{\mathbb{R}} du_r\, f_r(u_r) \cdot \text{Neg-Bin}\left(N_r^{(\mathrm{D})} + 1, \frac{1}{1 + e^{u_r}}\right)$<br>$f_r(u) \equiv n\left(u; \log\left(\gamma_0^* \left(\frac{P_r}{1 - P_r}\right)^{\gamma_1^*}\right), \sigma_\gamma^{*2}\right)$ | (num) | (num) |
| **Sys:** $N_r \stackrel{\text{indep}}{\simeq} \delta_N$<br>**Obs:** $N_r^{(\mathrm{D})} \mid N_r, p_r \stackrel{\text{indep}}{\simeq} \mathrm{Bin}\left(N_r, p_r\right)$<br>$p_r \stackrel{\text{indep}}{\simeq} \mathrm{Beta}(\alpha_r, \beta_r)$<br>$\left(\mathrm{logit}\left(\frac{\alpha_r}{\alpha_r + \beta_r}\right), \alpha_r + \beta_r\right) \mid \gamma_0, \gamma_1, \sigma_\gamma^2, P_r \stackrel{\text{indep}}{\simeq}$<br>$\mathrm{N}\left(\log\left(\gamma_0 \left(\frac{P_r}{1 - P_r}\right)^{\gamma_1}\right), \sigma_\gamma^2\right) \times \mathrm{Gamma}\left(1 + \xi, \frac{e^{\mu_r}}{\xi}\right)$<br>$\mathbb{P}\left(log\left(\gamma_0\right), \gamma_1, \sigma_\gamma^2\right) \simeq \frac{1}{\sigma_\gamma^2}$ (or other weakly informative)<br>$\mathbb{P}\left(\xi\right) \simeq \frac{\zeta_0}{(\zeta_0 + \xi)^2}$ (or other weakly informative)<br>$\mu_r \mid \delta_0, \delta_1, \sigma_\delta^2, N_r^{\mathrm{REG}} \stackrel{\text{indep}}{\simeq} \mathrm{N}\left(\log\left(\delta_0 \left(N_r^{\mathrm{REG}}\right)^{\delta_1}\right), \sigma_\delta^2\right)$<br>$\mathbb{P}\left(log\left(\delta_0\right), \delta_1, \sigma_\delta^2\right) \simeq \frac{1}{\sigma_\delta^2}$ (or other weakly informative) | $\approx \mathbf{N}^{(D)} + \int_{\Omega_\gamma} d\mathbb{P}(\gamma_0, \gamma_1, \sigma_\gamma^2) \prod_r \int_{\mathbb{R}} du_r\, f_r(u_r) \cdot \text{Neg-Bin}\left(N_r^{(D)} + 1, P(u_r)\right)$<br>$f_r(u) \equiv n\left(u; \log\left(\gamma_0 \left(\frac{P_r}{1 - P_r}\right)^{\gamma_1}\right), \sigma_\gamma^2\right)$<br>$P(u) \equiv \frac{1}{1 + e^u}$ | (num) | (num) |

# Inference at $t_0$: system with structure

| Model | Posterior | MAP | Mean |
|---|---|---|---|
| **Sys:** $N_r \mid \sigma_r \overset{\text{indep}}{\simeq} \text{Poisson}\left(A_r \cdot \sigma_r\right)$<br>$\sigma_r \overset{\text{indep}}{\simeq} \delta_{\sigma_r^*}$<br>**Obs:** $N_r^{(D)} \mid N_r, p_r \overset{\text{indep}}{\simeq} \text{Bin}\left(N_r, p_r\right)$<br>$p_r \overset{\text{indep}}{\simeq} \delta_{p_r^*}$ | $N_r^{(\mathrm{D})} + \text{Poisson}\left(A_r \sigma_r^* (1 - p_r^*)\right)$ | $N_r^{(D)} + \lfloor A_r \sigma_r^* (1 - p_r^*) \rfloor$ | $N_r^{(D)} + A_r \sigma_r^* (1 - p_r^*)$ |
| **Sys:** $N_r \mid \sigma_r \overset{\text{indep}}{\simeq} \text{Poisson}\left(A_r \cdot \sigma_r\right)$<br>$\sigma_r \mid \xi, \mu_{\sigma,r} \overset{\text{indep}}{\simeq} \text{Gamma}\left(1 + \xi, \frac{e^{\mu_{\sigma,r}}}{\xi}\right)$<br>$(\xi, \mu_{\sigma,r}) \overset{\text{indep}}{\simeq} \delta_{\xi^*} \times \delta_{\mu_{\sigma,r}^*}$<br>**Obs:** $N_r^{(D)} \mid N_r, p_r \overset{\text{indep}}{\simeq} \text{Bin}\left(N_r, p_r\right)$<br>$p_r \overset{\text{indep}}{\simeq} \delta_{p_r^*}$ | $N_r^{(\mathrm{D})} + \text{Neg-Bin}\left(N_r^{(\mathrm{D})} + \xi^* + 1, \frac{A_r e^{\mu_{\sigma,r}^*}}{\xi^* + A_r e^{\mu_{\sigma,r}^*}} \cdot (1 - p_r^*)\right)$ | $N_r^{(D)} + \left\lfloor \frac{N_r^{(D)} + \xi^*}{p_r^* \cdot A_r \cdot e^{\mu_{\sigma,r}^*} + \xi^*} (1 - p_r^*)(A_r e^{\mu_{\sigma,r}^*}) \right\rfloor$ | $\left(1 - \frac{\xi^* (1 - p_r^*)}{\xi^* + p_r^* A_r e^{\mu_{\sigma,r}^*}}\right) \frac{N_r^{(D)}}{p_r^*} + \frac{\xi^* (1 - p_r^*)}{\xi^* + p_r^* A_r e^{\mu_{\sigma,r}^*}} A_r e^{\mu_{sigma,r}^*}$ |
| **Sys:** $N_r \mid \sigma_r \overset{\text{indep}}{\simeq} \text{Poisson}\left(A_r \cdot \sigma_r\right)$<br>$\sigma_r \mid \xi, \mu_{\sigma,r} \overset{\text{indep}}{\simeq} \text{Gamma}\left(1 + \xi, \frac{e^{\mu_{\sigma,r}}}{\xi}\right)$<br>$\mathbb{P}(\xi) = \frac{\zeta_0}{(\zeta_0 + \xi)^2}$ (or other weakly informative)<br>$\mu_{\sigma,r}^* \mid \beta_0, \beta_1, \sigma_\beta^2, \sigma_r^{\mathrm{REG}} \overset{\text{indep}}{\simeq} \mathrm{N}\left(\log\left(\beta_0 \left(\sigma_r^{\mathrm{REG}}\right)^{\beta_1}\right), \sigma_\beta^2\right)$<br>$\mathbb{P}\left(\log \beta_0, \beta_1, \sigma_\beta^2\right) \propto \frac{1}{\sigma_\beta^2}$ (or other weakly informative)<br>**Obs:** $N_r^{(D)} \mid N_r, p_r \overset{\text{indep}}{\simeq} \text{Bin}\left(N_r, p_r\right)$<br>$p_r \overset{\text{indep}}{\simeq} \delta_{p_r^*}$ | $\approx \mathbf{N}^{(\mathrm{D})} + \int_{\mathbb{R}} d\mathbb{P}(\xi) \int_{\Omega_\beta} d\mathbb{P}(\beta_0, \beta_1, \sigma_\beta^2) \prod_r \int_{\mathbb{R}} d\mu_{\sigma,r}\, f_r(\mu_{\sigma,r}) \cdot \text{Neg-Bin}\left(S_r(\xi), P_r(\mu_{\sigma,r}, \xi)\right)$<br>$f_r(\mu) \equiv n\left(\mu; \log\left(\beta_0 \left(\sigma_r^{\mathrm{REG}}\right)^{\beta_1}\right), \sigma_\beta^2\right)$<br>$S_r(\xi) \equiv N_r^{(\mathrm{D})} + \xi + 1$<br>$P_r(\mu_{\sigma,r}, \xi) \equiv \frac{A_r e^{\mu_{\sigma,r}}}{\xi + A_r e^{\mu_{\sigma,r}}} \cdot (1 - p_r^*)$ | (num) | (num) |

# Conclusions: work in progress

INSTITUTO NACIONAL DE ESTADISTICA

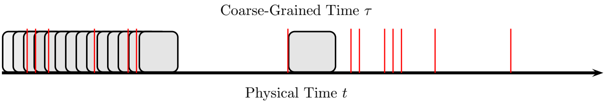
- **Models for t $>$ t$_0$**: only dependent on mobile phone data?

$$
\begin{array}{ll}
\text{(Closed Population)} & N_r(t) = \sum_s N_s(t-1)\pi_{sr}(t-1,t) \\
 & \pi_{sr}(t-1,t) \rightsquigarrow \pi_{sr}^{(D)}(t-1,t) \\
\text{(Open Population)} & ?
\end{array}
$$

- **Individual:Device 1:2**

$$
\rightsquigarrow N_r^{(D)}|N_r, p_r^{(i)}, \omega \overset{\text{indep}}{\simeq} \omega \cdot \mathrm{Bin}\left(N_r, p_r^{(1)}\right) + 2\cdot(1-\omega)\cdot \mathrm{Bin}\left(N_r, p_r^{(2)}\right)
$$

- Time **coarse-graining** procedure $t \rightsquigarrow \tau$




- Link with **geolocation of network events**:

**Event location (C-location)**

- Generation of **synthetic data**