# Generalized linear mixed models for small area estimation

Tomáš Košlab

Supervisor: doc. Ing. Tomáš Hobza, Ph.D.

FNSPE CTU

June 24, 2019

# Table of Contents

# Small Area Estimation (SAE)

- statistical discipline that deals with the problem of obtaining estimates of a target characteristic from a population divided into (geographic, socio-economic or other) subdomains
- *small area* - not enough data for a reliable direct estimate of the characteristic of interest


- regression models with fixed and random parameters respectively which represent the small areas - models "borrow strength" between related areas as well as from external sources

# Task of SAE

- $D$ domains, $N$ individuals, $N_d$ individuals in the $d$-th area
- $Y_{dj} \sim Be(p_{dj})$, $d = 1, \ldots, D$, $j = 1, \ldots, N_d$,
- area means are to be predicted

$$\overline{y}_d = \frac{1}{N_d} \sum_{j=1}^{N_d} y_{dj}, \quad d = 1, \ldots, D \tag{1}$$

- using uniform random sampling without replacement $n_d$ individuals are chosen into the sample from the $d$-th area
- comparison with the direct estimate

$$\hat{\overline{y}}_d^{dir} = \frac{1}{n_d} \sum_{j=1}^{n_d} y_{dj}, \quad d = 1, \ldots, D \tag{2}$$

# Proposed model

- some areas are modelled using fixed, other using random effects
- using the idea presented in Herrador et al. (2013)
- areas with more data (e.g. cities) are modelled differently to the other domains
- logistic regression model ($Y_{dj} \sim Be(p_{dj})$)
-

$$\begin{aligned}
\text{logit}(p_{dj}) &= \exp(\boldsymbol{x}_{dj}^T \boldsymbol{\beta} + m_d), \qquad d = 1, \dots, D_F \\
\text{logit}(p_{dj}) &= \exp(\boldsymbol{x}_{dj}^T \boldsymbol{\beta} + u_d), \qquad d = D_F + 1, \dots, D,
\end{aligned} \tag{3}$$

  where $m_d$ is a fixed parameter and $u_d \sim N(0, \sigma^2)$ is a random parameter of the $d$-th area

# Parameter estimation and predictions of $\overline{y}_d$

- parameters $\boldsymbol{\beta}$, $\boldsymbol{m}$, $\sigma^2$ are estimated using the PQL method (adapted for our model)
- log-likelihood function contains the integral

$$\int_{\mathbb{R}} \exp\left\{\sum_{j=1}^{n_d}\left[y_{dj}u_d - \log(1+\exp(\boldsymbol{x}_{dj}^T\boldsymbol{\beta}+u_d))\right] - \frac{u_d^2}{2\sigma^2}\right\} du_d \quad (4)$$

- predictions of area means in the $d$-th area can be expressed as

$$\hat{\overline{y}}_d = \frac{1}{N_d}\left(\sum_{j\in s_d} y_{dj} + \sum_{j\in r_d} \hat{p}_{dj}\right), \quad (5)$$

and predictors differ in the form of $\hat{p}_{dj}$

# Plug-in predictor and the empirical best predictor (EBP)

- for the plug-in predictor it holds

$$\hat{p}_{dj}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{m}}, \hat{\sigma}^2, \hat{\boldsymbol{u}}) = \begin{cases} \frac{\exp(\boldsymbol{x}_{dj}^T \hat{\boldsymbol{\beta}} + \hat{m}_d)}{1+\exp(\boldsymbol{x}_{dj}^T \hat{\boldsymbol{\beta}} + \hat{m}_d)} & d = 1, \dots, D_F \\ \frac{\exp(\boldsymbol{x}_{dj}^T \hat{\boldsymbol{\beta}} + \hat{u}_d)}{1+\exp(\boldsymbol{x}_{dj}^T \hat{\boldsymbol{\beta}} + \hat{u}_d)} & d = D_F + 1, \dots, D \end{cases} \tag{6}$$

- the empirical best predictor (EBP) can be expressed as

$$\hat{p}_{dj}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{m}}, \hat{\sigma}^2) = \mathbf{E}(p_{dj}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{m}}, \hat{\sigma}^2) | \boldsymbol{y}_s) \tag{7}$$

- integrals similar to (4) are encountered and their values are obtained using Monte Carlo simulations

# Simulation experiments

- $D = 30$ areas
- $x_{dj0} = 1$, $x_{dj1} \sim Be(0.48)$, $x_{dj2} \sim Be(0.6)$

$$x_{dj3} \begin{cases} \sim Be(0.5) & \text{pre } x_{dj2} = 0 \\ = 0 & \text{pre } x_{dj2} = 1 \end{cases}$$

- regression parameters: $\beta_0 = 0.3$, $\beta_1 = 0.5$, $\beta_2 = 1$, $\beta_3 = 0.5$
- $\sigma^2 = 0.5 \qquad m_d = -0.8 - \frac{0.1d}{D_F}, \quad d = 1, \ldots, D_F$
- $K = 1000$ iterations of the algorithm
- for $k = 1, \ldots, K$ do
  1. generate $u_d^{(k)}$, $y_{dj}^{(k)}$ and calculate $\overline{y}_d^{true(k)}$
  2. choose a sample of size $n_d$ and calculate $y_d^{dir(k)}$
  3. estimate $\hat{\boldsymbol{\beta}}^{(k)}$, $\hat{\boldsymbol{m}}^{(k)}$, $\hat{\sigma}^{2(k)}$ and predict $\hat{\boldsymbol{u}}^{(k)}$
  4. predict $\hat{\overline{y}}_d^{(k)}$
  5. using the parametric bootstrap method calculate $mse_d^{(k)}$

# Simulation experiments - output

- Output:

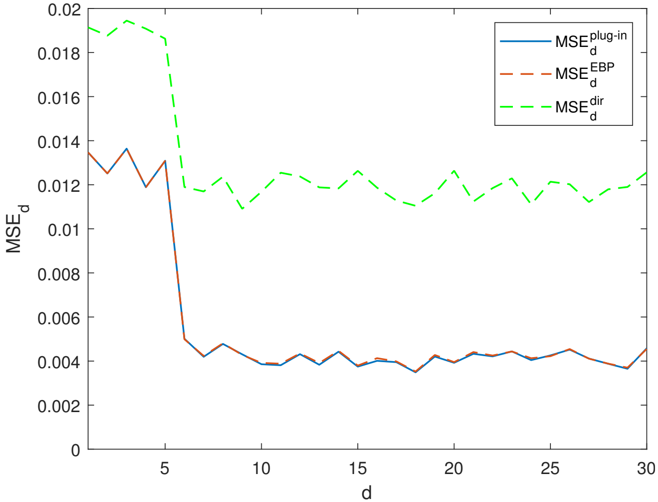
$$MSE_d = \frac{1}{K}\sum_{k=1}^{K}\left(\hat{\bar{y}}_d^{(k)} - \bar{y}_d^{true(k)}\right)^2$$

$$BIAS_d = \frac{1}{K}\sum_{k=1}^{K}\left(\hat{\bar{y}}_d^{(k)} - \bar{y}_d^{true(k)}\right) \tag{8}$$

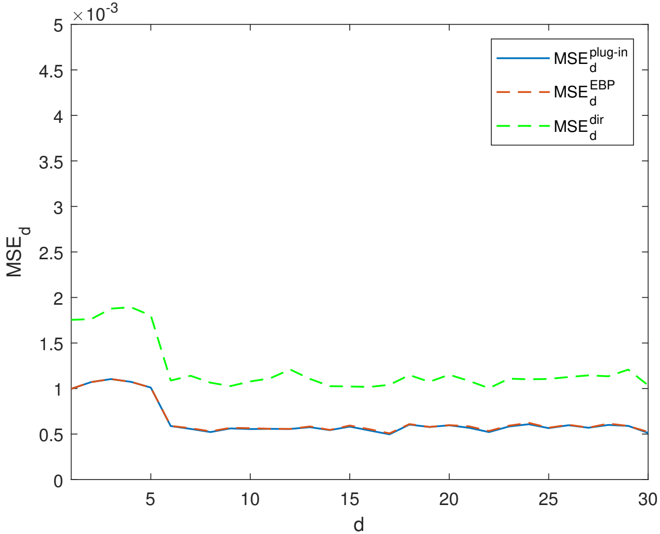
$$mse_d = \frac{1}{K}\sum_{k=1}^{K} mse_d^{(k)},$$

- predictions of the model are examined for different sample sizes $n_d$ and the EBP and plug-in predictor are compared
- quality of predictions for models with different number of fixed effects (i.e. different $D_F$) is investigated and compared with the commonly used model with $D_F = 0$

# Results - EBP vs plug-in




(a) $n_d = 10$




(b) $n_d = 100$

Figure 1: $MSE_d$ for the respective small areas and the investigated predictors for $D_F = 5$

# Results - comparison of models

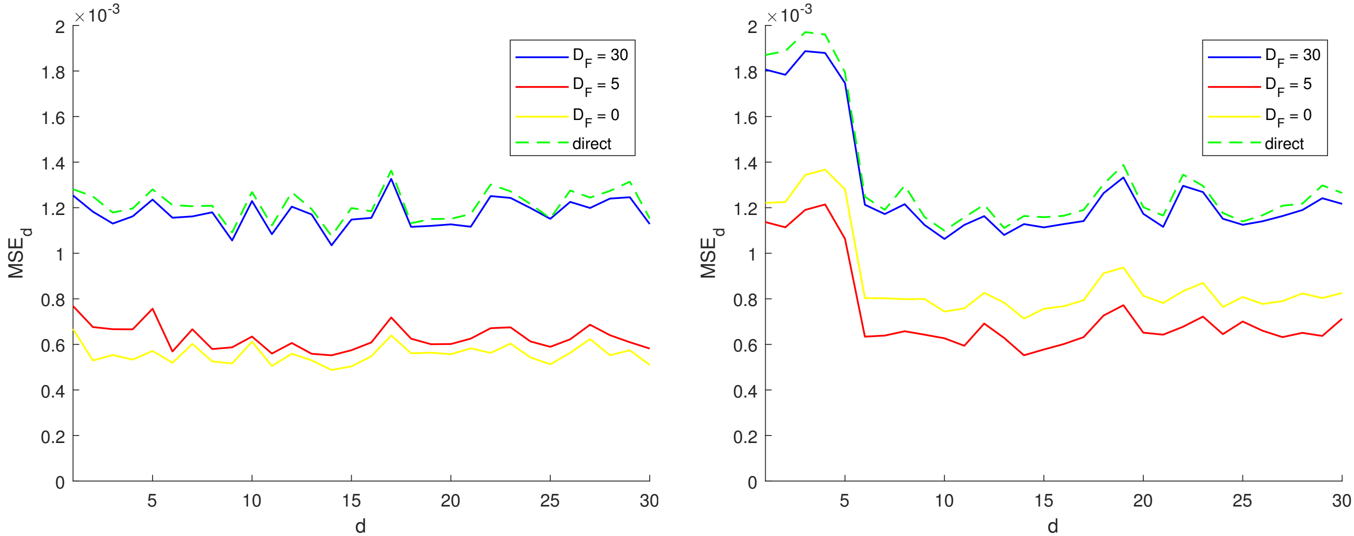



Figure 2: $MSE$ of the area means predictions for different values of $D_F$ and $n_d = 100$ using the plug-in predictor. Data are generated from the model with $D_F = 0$ (left) and $D_F = 5$ (right) respectively.

# Results - $n_{dF} = 100$, $n_{dR} = 10$

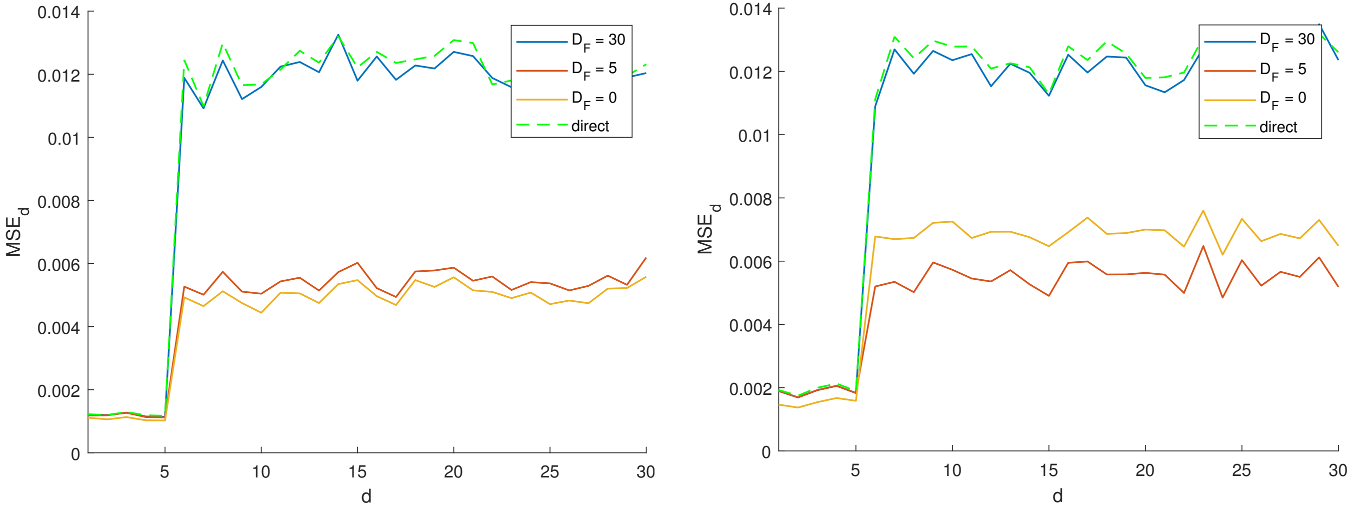



Figure 3: $MSE$ of the predictions of area means for the respective areas with $n_{dF} = 100$ and $n_{dR} = 10$ using the plug-in predictor. Data are generated from the model with $D_F = 0$ (left) and $D_F = 5$ (right) respectively.

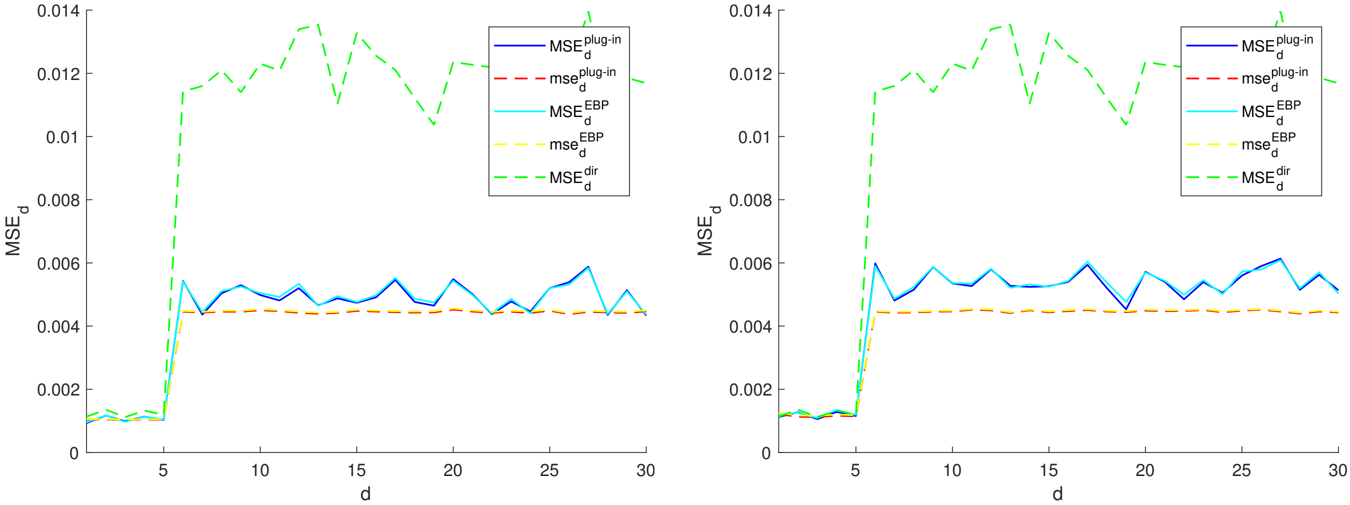



Figure 4: Comparison of $mse^{plug-in}$ and $mse^{EBP}$ for $n_{dF} = 100$ and $n_{dR} = 10$. Data are generated from the model with $D_F = 0$ (left) and $D_F = 5$ (right) respectively.

# Real data application - SLCS 2012

- region of Valencia, Spain - proportions of people in the risk of poverty are estimated (annual income below €6840)
- $D = 26$ domains, total population $N = 4908194$, sample $n = 2678$ people
- the model, which uses labour status (employed, unemployed, inactive, child) and domain as explanatory variables, has the form

$$\text{logit}(p_{dj}) = \beta_0 + \beta_1 x_{dj1} + \beta_1 x_{dj2} + \beta_3 x_{dj3} + \mu_d, \tag{9}$$

  where $x_{dj1}$, $x_{dj2}$, $x_{dj3}$ express the labour status of the $j$-th individual in the $d$-th area and $\mu_d$ is the (fixed or random) effect of the area in which the individual resides
- the aims are to compare the examined predictors as well as to compare our model and the model with $D_F = 0$

# Results - SLCS 2012

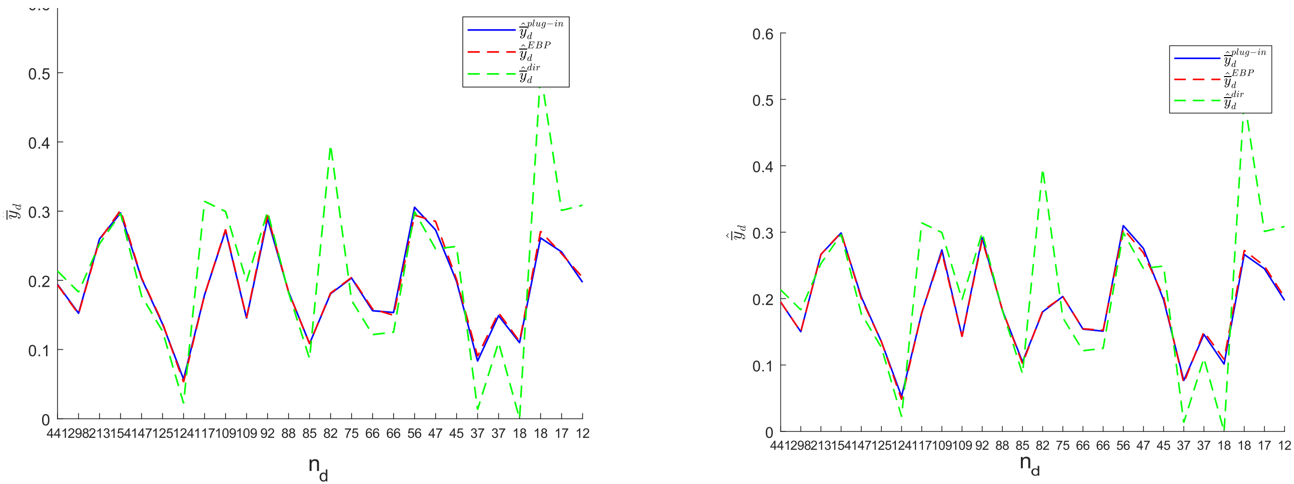



Figure 5: Mean predictions of individual areas for the model with $D_F = 0$ (left) and $D_F = 3$ (right) using the EBP and the plug-in predictor respectively. Areas are sorted in descending order according to the number of observations in the sample.

# Conclusion

- performance of EBP is very similar to plug-in and better than the direct estimate
- for larger sample sizes ($n_d = 100$) the best results are achieved by the model from which the data were generated
- proposed model is more flexible than the model with $D_F = 0$, it adadpts well on data generated from the model with $D_F = 0$ and outperforms it for data generated from model with $D_F \geq 5$
- both predictors give similar results in real data application
- reults of the proposed model on real data are comparable with the model with $D_F = 0$, error estimates could not be compared due to an implementation error

Thank you for your attention!

$$\text{logit}(p_{dj}) = \beta_0 + \beta_1 x_{dj1} + \beta_1 x_{dj2} + \beta_3 x_{dj3} + \mu_d, \tag{10}$$

| | $D_F = 0$ | $D_F = 3$ |
|---|---|---|
| $\beta_0$ | -1.2026 | -1.2334 |
| $\beta_1$ | -0.8118 | -0.8121 |
| $\beta_2$ | 0.5246 | 0.5260 |
| $\beta_3$ | -0.4032 | -0.4025 |
| $\sigma^2$ | 0.3250 | 0.3909 |

Table 1: Parameter estimates for the compared models

| Notation | Meaning |
|---|---|
| (1, 0, 0) | employed |
| (0, 1, 0) | unemployed |
| (0, 0, 1) | inactive |
| (0, 0, 0) | below 15 y. of age |

Table 2: Notation for labour status categories