On Models for Small Area Compositions

Angela Luna
A.Luna-Hernandez@soton.ac.uk

Li-Chun Zhang
L.Zhang@soton.ac.uk

Social Statistics & Demography
University of Southampton

Acknowledgements: This work has been funded by the Office for National Statistics - ONS and the Economic and Social Research Council - ESRC.

Area	1	...	J	Total
1				$Y_{1\cdot}$
2				$Y_{2\cdot}$
3				$Y_{3\cdot}$
		Y_{aj}		
...				...
A				$Y_{A\cdot}$
Total	$Y_{\cdot 1}$	...	$Y_{\cdot J}$	$Y_{\cdot\cdot}$

Area	1	...	J	Total
1				$Y_{1\cdot}$
2				$Y_{2\cdot}$
3				$Y_{3\cdot}$
		Y_{aj}		
...				...
A				$Y_{A\cdot}$
Total	$Y_{\cdot 1}$	...	$Y_{\cdot J}$	$Y_{\cdot\cdot}$

Target: Estimate the within area cell counts Y_{aj}, using proxy information and fixed row/column margins.

Two different (fixed-effects) approaches to this problem are considered:

- ▷ Structure Preserving Models: Long tradition in SAE. Assumptions about the relationship between the interactions of two compositions in the log-linear scale. (Proxy information is required).

Two different (fixed-effects) approaches to this problem are considered:

- ▷ Structure Preserving Models: Long tradition in SAE. Assumptions about the relationship between the interactions of two compositions in the log-linear scale. (Proxy information is required).
- ▷ Regression (Generalized Linear) Models: Multinomial-Logistic: Assumptions about the relationship between the log-odds with respect to a reference category and a set of covariates. (Proxy information can be used as covariate).

In this work, we:

1. Introduce a generalization of the Structure Preserving approach, covering the SPREE and GSPREE models and also the logit-multinomial (using proxy information) as particular cases.

In this work, we:

1. Introduce a generalization of the Structure Preserving approach, covering the SPREE and GSPREE models and also the logit-multinomial (using proxy information) as particular cases.
2. Use data from 2001 and 2011 Population Censuses in England to compare the different models in terms of their Prediction Error.

In this work, we:

1. Introduce a generalization of the Structure Preserving approach, covering the SPREE and GSPREE models and also the logit-multinomial (using proxy information) as particular cases.
2. Use data from 2001 and 2011 Population Censuses in England to compare the different models in terms of their Prediction Error.
3. Show some ongoing work on a model using a mapping matrix between the proxy and desired compositions, which allows to incorporate auxiliary information at the aggregate level.

Outline

Denote by $\theta^X_{aj} = X_{aj}/X_{a\cdot}$ an auxiliary composition of exactly the same dimension as $\theta^Y_{aj} = Y_{aj}/Y_{a\cdot}$, its log-linear representation given by:

$$\gamma^X_{aj} = \alpha^X_0 + \alpha^X_a + \alpha^X_j + \alpha^X_{aj}$$

where $\gamma^X_{aj} = \log \theta^X_{aj}$, $\alpha^X_0 = \bar{\gamma}^X_{\cdot\cdot}$, $\alpha^X_a = \bar{\gamma}^X_{a\cdot} - \bar{\gamma}^X_{\cdot\cdot}$, $\alpha^X_j = \bar{\gamma}^X_{\cdot j} - \bar{\gamma}^X_{\cdot\cdot}$ and $\alpha^X_{aj} = \gamma^X_{aj} - \bar{\gamma}^X_{a\cdot} - \bar{\gamma}^X_{\cdot j} - \bar{\gamma}^X_{\cdot\cdot}$.

Denote by $\theta^X_{aj} = X_{aj}/X_{a\cdot}$ an auxiliary composition of exactly the same dimension as $\theta^Y_{aj} = Y_{aj}/Y_{a\cdot}$, its log-linear representation given by:

$$\gamma^X_{aj} = \alpha^X_0 + \alpha^X_a + \alpha^X_j + \alpha^X_{aj}$$

where $\gamma^X_{aj} = \log \theta^X_{aj}$, $\alpha^X_0 = \bar{\gamma}^X_{\cdot\cdot}$, $\alpha^X_a = \bar{\gamma}^X_{a\cdot} - \bar{\gamma}^X_{\cdot\cdot}$, $\alpha^X_j = \bar{\gamma}^X_{\cdot j} - \bar{\gamma}^X_{\cdot\cdot}$ and $\alpha^X_{aj} = \gamma^X_{aj} - \bar{\gamma}^X_{a\cdot} - \bar{\gamma}^X_{\cdot j} - \bar{\gamma}^X_{\cdot\cdot}$.

The log-linear representation satisfies the constraints:
$\sum_a \alpha^X_a = 0$, $\sum_j \alpha^X_j = 0$, $\sum_a \alpha^X_{aj} = \sum_j \alpha^X_{aj} = 0$. Analogous for θ^Y_{aj}.

Denote by $\theta^X_{aj} = X_{aj}/X_{a\cdot}$ an auxiliary composition of exactly the same dimension as $\theta^Y_{aj} = Y_{aj}/Y_{a\cdot}$, its log-linear representation given by:

$$\gamma^X_{aj} = \alpha^X_0 + \alpha^X_a + \alpha^X_j + \alpha^X_{aj}$$

where $\gamma^X_{aj} = \log \theta^X_{aj}$, $\alpha^X_0 = \bar{\gamma}^X_{\cdot\cdot}$, $\alpha^X_a = \bar{\gamma}^X_{a\cdot} - \bar{\gamma}^X_{\cdot\cdot}$, $\alpha^X_j = \bar{\gamma}^X_{\cdot j} - \bar{\gamma}^X_{\cdot\cdot}$ and $\alpha^X_{aj} = \gamma^X_{aj} - \bar{\gamma}^X_{a\cdot} - \bar{\gamma}^X_{\cdot j} - \bar{\gamma}^X_{\cdot\cdot}$.

The log-linear representation satisfies the constraints:
$\sum_a \alpha^X_a = 0$, $\sum_j \alpha^X_j = 0$, $\sum_a \alpha^X_{aj} = \sum_j \alpha^X_{aj} = 0$. Analogous for θ^Y_{aj}.

The modelling process is focused on the relationship between α^Y_{aj} and α^X_{aj}. Marginal constraints such as $\sum_a \hat{Y}_{aj} = Y_{\cdot j}$ for $j = 1, \ldots, J$ and $\sum_j \hat{Y}_{aj} = Y_{a\cdot}$ for $a = 1, \ldots, A$ can be considered using IPF without modifying the parameter estimates. Proxy information (not just covariates) is required.

In the context of SAE, the following Structure Preserving models have been used:

In the context of SAE, the following Structure Preserving models have been used:

1. Given $\left\{\theta_{aj}^{X}\right\}, \{Y_{1\cdot}, \ldots, Y_{A\cdot}\}$:

In the context of SAE, the following Structure Preserving models have been used:

1. Given $\left\{\theta_{aj}^{X}\right\}, \{Y_{1\cdot}, \ldots, Y_{A\cdot}\}$:

   Synthetic Estimator: Adapted from Gonzalez & Hoza (1978),

   $$\hat{Y}_{aj} = \theta_{aj}^{X} Y_{a\cdot}$$

   The underlining model is $\alpha_{j}^{Y} = \alpha_{j}^{X}$, $\alpha_{aj}^{Y} = \alpha_{aj}^{X}$.

   The estimated composition is a rescaled version of the auxiliary composition.

2. Given $\left\{\theta_{aj}^{X}\right\}, \{Y_{1\cdot}, \ldots, Y_{A\cdot}\}, \{Y_{\cdot 1}, \ldots, Y_{\cdot J}\}$:

2. Given $\left\{\theta^X_{aj}\right\}, \{Y_{1\cdot}, \ldots, Y_{A\cdot}\}, \{Y_{\cdot 1}, \ldots, Y_{\cdot J}\}$:

   SPREE: Purcell & Kish (1980) use IPF to fit the two margins,

$$\hat{Y}^{(1)}_{aj} = \theta^X_{aj} Y_{a\cdot}\,, \quad \hat{Y}^{(2)}_{aj} = \frac{\hat{Y}^{(1)}_{aj}}{\hat{Y}^{(1)}_{\cdot j}} Y_{\cdot j}\,, \quad \ldots$$

   until convergency is achieved. This estimator minimizes the distance between the compositions X and $\hat{Y}$ satisfying the marginal constraints. The underlining model is $\alpha^Y_{aj} = \alpha^X_{aj}$.

3. Given $\left\{\theta_{aj}^{X}\right\}, \{Y_{1\cdot}, \ldots, Y_{A\cdot}\}, \{Y_{\cdot 1}, \ldots, Y_{\cdot J}\}$ and an estimated $\left\{\theta_{aj}^{Y}\right\}$:

3. Given $\left\{\theta_{aj}^{X}\right\}, \{Y_{1\cdot}, \ldots, Y_{A\cdot}\}, \{Y_{\cdot 1}, \ldots, Y_{\cdot J}\}$ and an estimated $\left\{\theta_{aj}^{Y}\right\}$:

   Generalized Linear Structural Model (GSPREE): Zhang & Chambers (2004) propose the model

$$\alpha_{aj}^{Y} = \beta \alpha_{aj}^{X} .$$

3. Given $\left\{\theta_{aj}^{X}\right\}, \{Y_{1\cdot}, \ldots, Y_{A\cdot}\}, \{Y_{\cdot 1}, \ldots, Y_{\cdot J}\}$ and an estimated $\left\{\theta_{aj}^{Y}\right\}$:

   Generalized Linear Structural Model (GSPREE): Zhang & Chambers (2004) propose the model

$$\alpha_{aj}^{Y} = \beta \alpha_{aj}^{X} .$$

   β can be estimated using ML under the multinomial distribution, when expressing the model as:

$$\mu_{aj}^{Y} = \lambda_j + \beta \mu_{aj}^{X}$$

   for $\mu_{aj} = \log \theta_{aj} - \frac{1}{J} \sum_{l} \log \theta_{aj} = \alpha_j + \alpha_{aj}$.

3. Given $\left\{\theta^X_{aj}\right\}, \{Y_{1\cdot}, \ldots, Y_{A\cdot}\}, \{Y_{\cdot 1}, \ldots, Y_{\cdot J}\}$ and an estimated $\left\{\theta^Y_{aj}\right\}$:

   Generalized Linear Structural Model (GSPREE): Zhang & Chambers (2004) propose the model

   $$\alpha^Y_{aj} = \beta \alpha^X_{aj} .$$

   β can be estimated using ML under the multinomial distribution, when expressing the model as:

   $$\mu^Y_{aj} = \lambda_j + \beta \mu^X_{aj}$$

   for $\mu_{aj} = \log \theta_{aj} - \frac{1}{J} \sum_l \log \theta_{aj} = \alpha_j + \alpha_{aj}$.

   Given the sum-zero constraint of the α_j, the λ_j are nuisance parameters with no practical interest.

Extension of the Structure Preserving approach

All the previous models can be seen as particular cases of the more general model:

$$\begin{bmatrix} \alpha_{a1}^{Y} \\ \vdots \\ \alpha_{aJ}^{Y} \end{bmatrix} = \mathbb{B}\,\boldsymbol{\beta}\,\mathbb{B} \begin{bmatrix} \alpha_{a1}^{X} \\ \vdots \\ \alpha_{aJ}^{X} \end{bmatrix}$$

Extension of the Structure Preserving approach

All the previous models can be seen as particular cases of the more general model:

$$\begin{bmatrix} \alpha_{a1}^{Y} \\ \vdots \\ \alpha_{aJ}^{Y} \end{bmatrix} = \mathbb{B}\,\boldsymbol{\beta}\,\mathbb{B} \begin{bmatrix} \alpha_{a1}^{X} \\ \vdots \\ \alpha_{aJ}^{X} \end{bmatrix}$$

Where $\mathbb{B}_{J\times J} = \mathrm{I} - J^{-1}11'$ and $\boldsymbol{\beta}_{J\times J} = \{\beta_{jk}\}$ contains all the parameters.

Extension of the Structure Preserving approach

All the previous models can be seen as particular cases of the more general model:

$$\begin{bmatrix} \alpha_{a1}^{Y} \\ \vdots \\ \alpha_{aJ}^{Y} \end{bmatrix} = \mathbb{B}\,\boldsymbol{\beta}\,\mathbb{B} \begin{bmatrix} \alpha_{a1}^{X} \\ \vdots \\ \alpha_{aJ}^{X} \end{bmatrix}$$

Where $\mathbb{B}_{J\times J} = \mathrm{I} - J^{-1}11'$ and $\boldsymbol{\beta}_{J\times J} = \{\beta_{jk}\}$ contains all the parameters.

The multiplication on left and right by $\mathbb{B}$ ensure that the sum zero constraints are satisfied by the predicted α_{aj}^{Y}, as well as the uniqueness of $\mathbb{G} = \mathbb{B}\boldsymbol{\beta}\mathbb{B}$. Denoting by $\{g_{jk}\}$ the components of $\mathbb{G}$ we can write,

$$\alpha_{aj}^{Y} = \sum_{k} g_{jk}\alpha_{ak}^{X}\,.$$

As in the GSPREE, the estimation of $\boldsymbol{\beta}$ can be done using ML under the multinomial distribution writing the model as $\boldsymbol{\eta}_a^{Y} = \boldsymbol{\lambda} + \mathbb{B}\,\boldsymbol{\beta}\,\mathbb{B}\boldsymbol{\eta}_a^{X}$.

Some particular cases:

a) SPREE: $\boldsymbol{\beta} = \mathrm{I}$

$$\alpha_{aj}^{Y} = \alpha_{aj}^{X} - \frac{1}{J} \sum_{k} \alpha_{ak}^{X}$$

Some particular cases:

a) SPREE: $\boldsymbol{\beta} = \mathrm{I}$

$$\alpha_{aj}^{Y} = \alpha_{aj}^{X} - \frac{1}{J}\sum_{k}\alpha_{ak}^{X}$$

b) GSPREE: With parameter ϕ, $\boldsymbol{\beta} = \phi\mathrm{I}$

$$\alpha_{aj}^{Y} = \phi\alpha_{aj}^{X} - \phi\frac{1}{J}\sum_{k}\alpha_{ak}^{X}$$

c) Logit-Multinomial Model: The model with J-1 parameters

$$\eta_{aj}^{Y} = \gamma_j + \phi_j \eta_{aj}^{X}$$

for $\eta_{aj} = \log\left[\theta_{aj}/\theta_{aJ}\right]$, can be written as a structural model in the form

$$\boldsymbol{\alpha}_a^{Y} = \mathbb{B}_{(J)} \boldsymbol{\beta} \mathbb{B}\, \boldsymbol{\alpha}_a^{X}$$

for $\mathbb{B}_{(J)}$ the J×(J-1) matrix resulting of dropping the column J from $\mathbb{B}$ and $\boldsymbol{\beta}$ a (J-1)×J matrix defined as

$$\boldsymbol{\beta} = \left[\operatorname{Diag}\left\{\vec{\beta}_{(J)}\right\} \middle| -\vec{\beta}_{(J)}\right]$$

for $\vec{\beta}_{(J)}$ a vector of J-1 parameters (The category J doesn't have a free parameter).

d) GSPREE with category-specific (J) parameters:

$\boldsymbol{\beta} = \mathrm{Diag}\left\{\vec{\beta}\right\}$

$$\alpha_{aj}^{Y} = \beta_j \alpha_{aj}^{X} - \frac{1}{J}\sum_{k} \beta_k \alpha_{ak}^{X}$$

The second term on the right hand, included to satisfy the sum-zero constrains without impose restrictions to the β_j, make the predictions of this model not a line anymore.

d) GSPREE with category-specific (J) parameters:

$\boldsymbol{\beta} = \mathrm{Diag}\left\{\vec{\beta}\right\}$

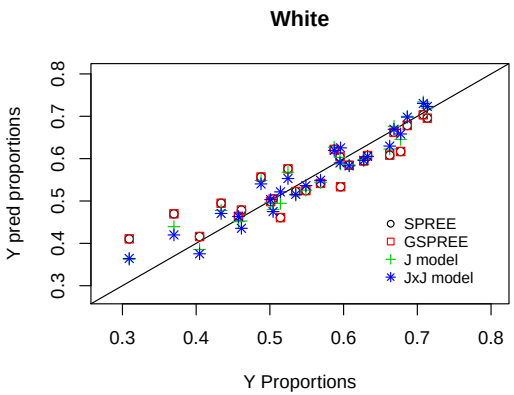
$$\alpha_{aj}^{Y} = \beta_j \alpha_{aj}^{X} - \frac{1}{J}\sum_k \beta_k \alpha_{ak}^{X}$$

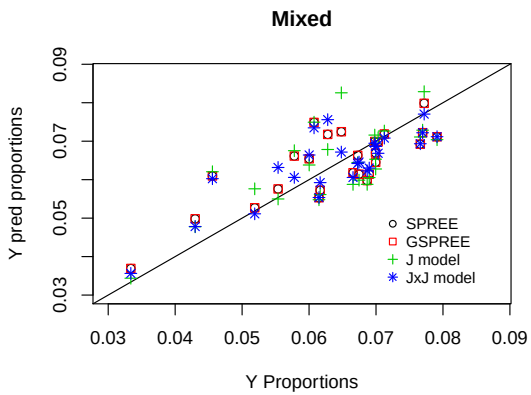
The second term on the right hand, included to satisfy the sum-zero constrains without impose restrictions to the β_j, make the predictions of this model not a line anymore.

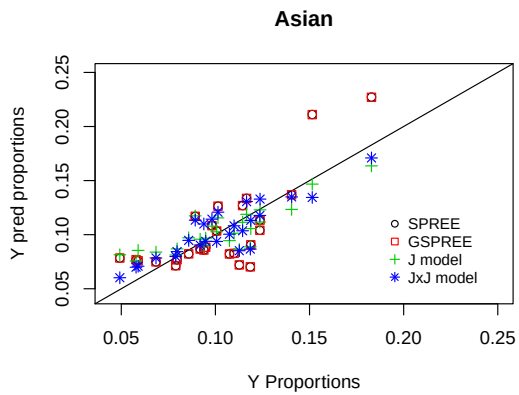
e) GSPREE JxJ model: $\boldsymbol{\beta} = \{\beta_{jk}\}$, $\mathbb{G} = \{g_{jk}\} = \mathbb{B}\boldsymbol{\beta}\mathbb{B}$

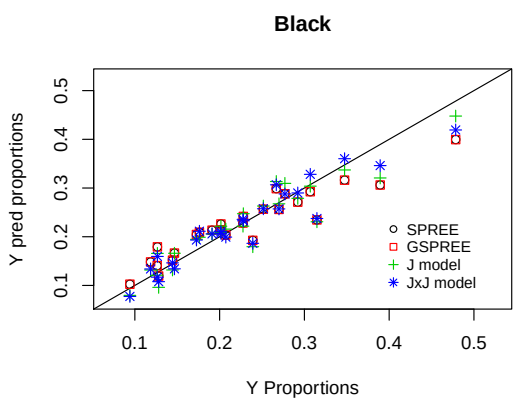
$$\alpha_{aj}^{Y} = \sum_k g_{jk} \alpha_{ak}^{X}$$

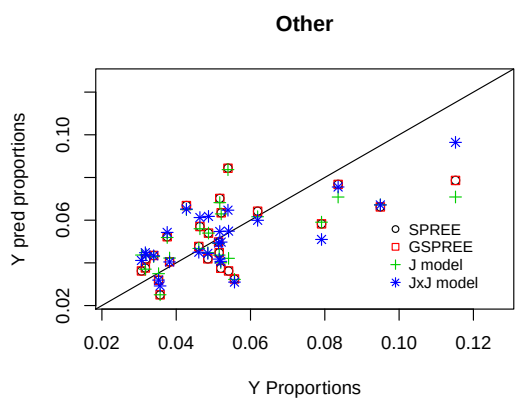
Data from 2001 and 2011 Population Census in England for the Hackney Borough. Ethnicity.



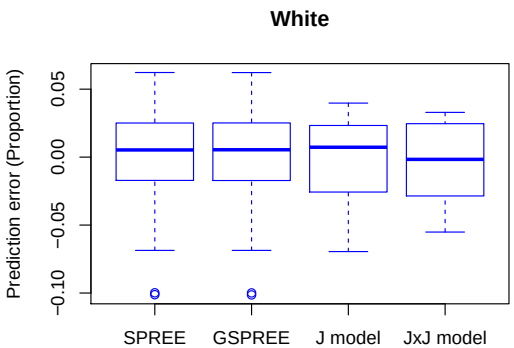


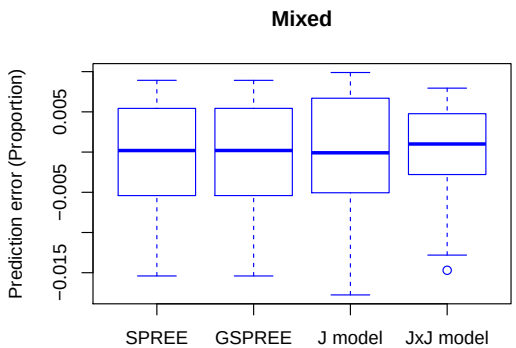


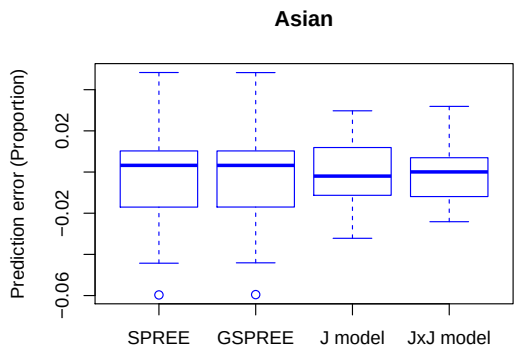


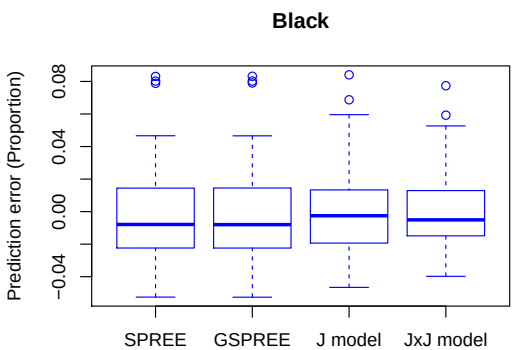


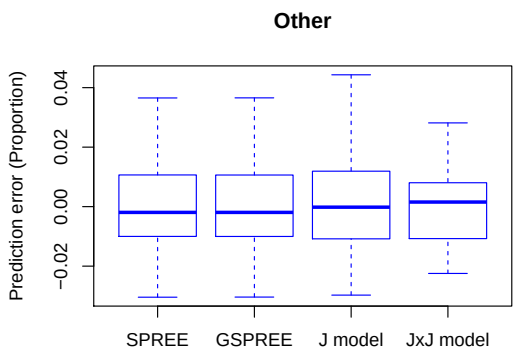
Data from 2001 and 2011 Population Census in England for the Hackney Borough. Ethnicity.



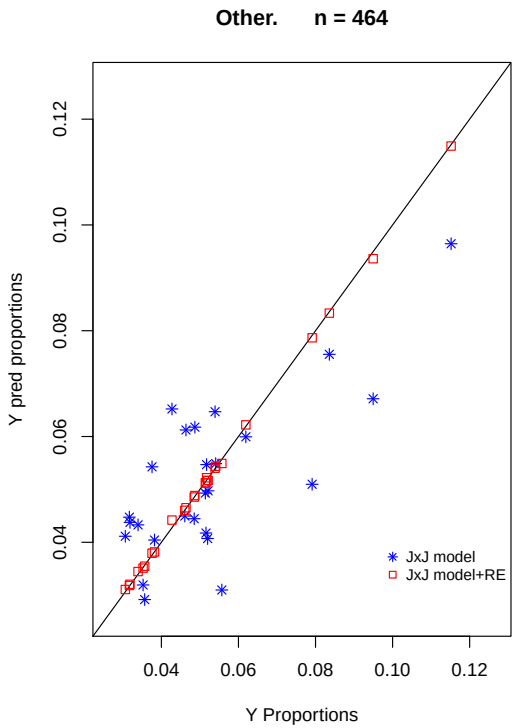


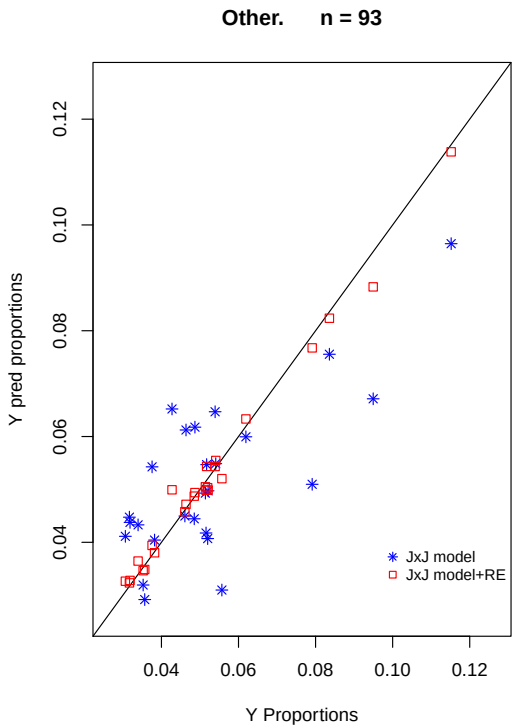


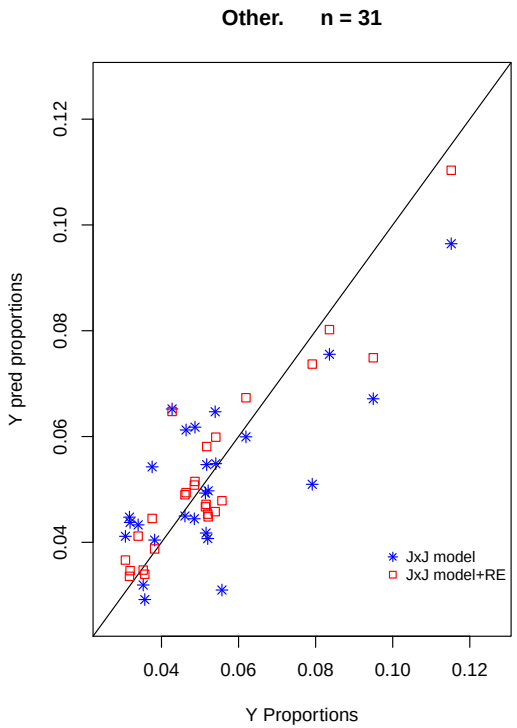




Extending the general model to include random effects







In summary...

In summary...

- Having an auxiliary composition (register/census data), an estimated (updated) composition and the current margins, we extend the GSPREE model from one to a maximum of $J \times J$ parameters.

In summary...

- Having an auxiliary composition (register/census data), an estimated (updated) composition and the current margins, we extend the GSPREE model from one to a maximum of $J \times J$ parameters.
- According to our preliminary exercises, the new models show less bias than SPREE and GSPREE models (fixed effects approach).

In summary...

- Having an auxiliary composition (register/census data), an estimated (updated) composition and the current margins, we extend the GSPREE model from one to a maximum of $J \times J$ parameters.
- According to our preliminary exercises, the new models show less bias than SPREE and GSPREE models (fixed effects approach).
- We are still working on the extension to include cell-specific random effects. As expected, for a big sample size the estimative obtained using a mixed model gets closer to the direct estimate, however, as it is borrowing strength from the auxiliary composition, it would be more stable.

In summary...

- Having an auxiliary composition (register/census data), an estimated (updated) composition and the current margins, we extend the GSPREE model from one to a maximum of $J \times J$ parameters.
- According to our preliminary exercises, the new models show less bias than SPREE and GSPREE models (fixed effects approach).
- We are still working on the extension to include cell-specific random effects. As expected, for a big sample size the estimative obtained using a mixed model gets closer to the direct estimate, however, as it is borrowing strength from the auxiliary composition, it would be more stable.
- MSE estimation is still need to be addressed.

Outline

Denote by $P = \{P_{ij}\}$ the gross flow from the composition X to Y, i.e., assume that for each area:

$$\begin{bmatrix} \theta_{a1}^{Y} \\ \vdots \\ \theta_{aJ}^{Y} \end{bmatrix} = \begin{bmatrix} P_{11} & \dots & P_{1J} \\ \vdots & \ddots & \vdots \\ P_{J1} & \dots & P_{JJ} \end{bmatrix} \begin{bmatrix} \theta_{a1}^{X} \\ \vdots \\ \theta_{aJ}^{X} \end{bmatrix}$$

The column sum of P is 1.

Denote by $P = \{P_{ij}\}$ the gross flow from the composition X to Y, i.e., assume that for each area:

$$\begin{bmatrix} \theta_{a1}^{Y} \\ \vdots \\ \theta_{aJ}^{Y} \end{bmatrix} = \begin{bmatrix} P_{11} & \dots & P_{1J} \\ \vdots & \ddots & \vdots \\ P_{J1} & \dots & P_{JJ} \end{bmatrix} \begin{bmatrix} \theta_{a1}^{X} \\ \vdots \\ \theta_{aJ}^{X} \end{bmatrix}$$

The column sum of P is 1.

What is the effect of the mapping matrix P in the log-linear representation of Y?

$$\boldsymbol{\theta}_a^Y = P\boldsymbol{\theta}_a^X \qquad \stackrel{?}{\rightarrow} \qquad \log \boldsymbol{\theta}_a^Y \approx M \log \boldsymbol{\theta}_a^X$$

Using a First Order Taylor approximation over

$$\ln\left(\theta_{aj}^{Y}\right) = \ln\left(\sum_{l} \theta_{al}^{X} P_{jl}\right)$$

as function of $\ln\left(\theta_{aj}^{X}\right)$, around the distribution of X at an aggregate level, denoted by $\tilde{\theta}^{X}$, we obtain

$$\ln \theta_{aj}^{Y} - \ln \tilde{\theta}_{j}^{Y} \approx \sum_{l} \left(q_{jl} - \tau_{j} \tilde{\theta}_{l}^{X}\right) \left[\ln \theta_{al}^{X} - \ln \tilde{\theta}_{l}^{X}\right]$$

Using a First Order Taylor approximation over

$$\ln\left(\theta_{aj}^{Y}\right) = \ln\left(\sum_{l}\theta_{al}^{X}P_{jl}\right)$$

as function of $\ln\left(\theta_{aj}^{X}\right)$, around the distribution of X at an aggregate level, denoted by $\tilde{\theta}^{X}$, we obtain

$$\ln\theta_{aj}^{Y} - \ln\tilde{\theta}_{j}^{Y} \approx \sum_{l}\left(q_{jl} - \tau_{j}\tilde{\theta}_{l}^{X}\right)\left[\ln\theta_{al}^{X} - \ln\tilde{\theta}_{l}^{X}\right]$$

where $\tilde{\theta}^{Y} = P\tilde{\theta}^{X}$, $q_{jl} = \dfrac{P_{jl}\tilde{\theta}_{l}^{X}}{\tilde{\theta}_{j}^{Y}}$ is the reverse flow and $\tau_{j} = \dfrac{P_{jr_j}}{\tilde{\theta}_{j}^{Y}}$ for P_{jr_j} one cell specifically chosen for the j category.

Applying the link function

$$\mu_{aj} = \log \theta_{aj} - \frac{1}{J} \sum_{l} \log \theta_{aj} = \alpha_j + \alpha_{aj}$$

Applying the link function

$$\mu_{aj} = \log \theta_{aj} - \frac{1}{J} \sum_{l} \log \theta_{aj} = \alpha_j + \alpha_{aj}$$

we can obtain the relationship

$$\alpha^{Y}_{aj} \approx \sum_{l} (q_{jl} - \bar{q}_{.l})\, \alpha^{X}_{al} - \sum_{l} (\tau_j - \bar{\tau})\, \tilde{\theta}^{X}_{l} \alpha^{X}_{al}.$$

According to our empirical studies, the leading term in the approximation is the term associated with the reverse flow. In this sense, a model involving also auxiliary information on the reverse flow at an aggregate level could be of interest.

Thanks!