Model-based Small Domain Estimation: Recent Methods & Applications in National Statistics Institutes in Europe

Nikos Tzavidis

Mexico City, October 2012

1Social Statistics & S3RI, University of Southampton
Motivation & overview of domain estimation in European NSIs

Estimation of averages (Totals) & ”complex” indicators
  - Nested error regression model
  - Design-consistent estimation
  - Alternative approaches to outlier robust estimation
  - Empirical Best Prediction

Two recent applications in European NSIs
  - Domain estimation for business surveys in the Netherlands
  - Estimating income deprivation in the UK

Current debate in UK - Beyond 2011 Census and the role of SAE methods
Motivation

Direct estimation: Use only domain-specific data

Problems with direct estimation

1. Direct estimates may suffer from low precision
2. Not applicable with zero sample sizes

Potential solution: Use of models to improve efficiency
Domain Estimation in NSIs in Europe

Examples of Domain Estimates:

- Average income
- Labour market activity (employment/unemployment)
- Income deprivation
- Business statistics

Requires use of different methodologies and auxiliary data

Increasingly relying on model-based estimation

Estimates published as experimental or National statistics
Two Popular Estimators of Domain Averages

- **Synthetic estimator**
  \[ \hat{y}_j = \bar{X}_j^T \beta_w \]

  - \( \beta_w \) is the probability weighted estimator
  - Can be biased - Homogeneity assumption **BUT**
  - Stable

- **Survey regression estimator**
  \[ \hat{y}_j = \hat{Y}_j^{HT} + (\bar{X}_j - \hat{x}_j^{HT}) \beta_w \]

  - Corrects the potential bias of the synthetic estimator **BUT**
  - Can be unstable

- Review paper (Pfeffermann (2012), Statistical Science)

Model-based domain estimation in NSIs in Europe
Model-based Methods
Nested Error Regression Model

Key Idea (Batteese et al., 1988; Rao, 2003)

- Improve efficiency of domain estimators
- Use random area-specific effects
- Capture area variation beyond that explained by covariates

\[ y_{ij} = x_{ij}^T \beta + v_j + \epsilon_{ij}, \quad i = 1, \ldots, n_j, \quad j = 1, \ldots, d \]

\[ \nu \sim N(0, \Sigma_v), \epsilon \sim N(0, \Sigma_\epsilon), \text{ OR } \epsilon \sim N(0, \Sigma_\epsilon = f(x)) \]

- Estimator of Domain Average

\[ \hat{y}_{j}^{EBLUP} = N_j^{-1} \left\{ \sum_{i \in s_j} y_{ij} + \sum_{i \in r_j} (x_{ij}^T \hat{\beta} + \hat{v}_j) \right\} \]
Extensions based on the Nested Error Regression Model I

**Nested Error Regression Model & Design Weights** *(You and Rao, 2002)*

- **Aim:** Design consistency
- Incorporates design weights in estimation
- Pseudo EBLUP

\[
\hat{y}_{jw}^{P-EBLUP} = \hat{\gamma}_{jw}\hat{y}_{jw} + (\bar{X}_j^T - \hat{\gamma}_{jw}\hat{x}_{jw}^T)\hat{\beta}_w
\]
Extensions based on the Nested Error Regression Model II

Outlier Robust estimation with the Nested Error Regression Model (Sinha and Rao, 2009)

- Effects of outliers are controlled via an influence function $\psi$
- Replace $\hat{\beta}$ by $\hat{\beta}^\psi$
- Replace $\hat{v}_j$ by $\hat{v}_j^\psi$

$$\hat{y}_{j \text{REBLUP}} = N_j^{-1} \left\{ \sum_{i \in s_j} y_{ij} + \sum_{i \in r_j} (x_{ij}^T \hat{\beta}^\psi + \hat{v}_j^\psi) \right\}$$

Model-based domain estimation in NSIs in Europe
M-quantile Model (Chambers & Tzavidis, 2006)

\[ MQ_y(q|x_{ij}) = x_{ij}^T \beta^\psi(q) \]

- \(q\) denotes a quantile. Conventionally is a-priori chosen, fixed.
- \(q_{ij}\) random variables such that \(y_{ij} = x_{ij}^T \beta^\psi(q_{ij})\)
- Estimate empirical domain effects \(\hat{\theta}_j = E(\hat{q}_{ij})\)
- \(\hat{\theta}_j\) captures between area variation
- No explicit parametric assumptions on \(\theta_j\)

\[
\hat{\bar{y}}_j^{MQ} = N_j^{-1} \left\{ \sum_{i \in s_j} y_{ij} + \sum_{i \in r_j} (x_{ij}^T \hat{\beta}^\psi(\hat{\theta}_j)) \right\}
\]
\( \hat{y}_j^{MQ} \) can be biased

**Bias correction:** Robust predictive approach (Welsh & Ronchetti, 1998; Tzavidis et al., 2010; Chambers et al., 2012)

\[
\hat{y}_j^{BC} = N_j^{-1} \left[ \sum_{i \in s_j} y_{ij} + \sum_{i \in r_j} \hat{y}_{ij}^{\psi} + \frac{N_j - n_j}{n_j} \sum_{i \in s_j} \phi \left\{ y_{ij} - \hat{y}_{ij}^{\psi} \right\} \right]
\]

- Resembles a model-based GREG
- Control bias via \( \phi \rightarrow \) Can be very unstable
- Alternative approach (Dongmo-Jiongo, Haziza, Duchesne, 2012)
Quantile Regression - A Parametric Link

A continuous random variable $y$ follows an asymmetric Laplace distribution, $y \sim ALD(\mu, \sigma, q)$ with pdf

$$p(y|\mu, \sigma, \tau) = \frac{q(1-q)}{\sigma} exp\left(-\frac{|y - \mu|}{\sigma}\right)$$

- Geraci and Bottai (2007): Quantile random effects regression
- $p(y, v|\beta, \sigma, \Gamma) = p(y|\beta, \sigma, v)p(v|\Gamma)$
- $y|v \sim ALD(x\beta + v, \sigma)$
- $p(v|\Gamma)$, Normal random effects
- $p(v|\Gamma)$, Robust random effects

$$\hat{y}_{j}^{ALD} = N_{j}^{-1}\left\{ \sum_{i \in s_j} y_{ij} + \sum_{i \in r_j} (x_{ij}^T \hat{\beta}^{ALD} + \hat{v}_{j}^{ALD}) \right\}$$
Case Study I: Domain estimation in Business Surveys

- CBS in the Netherlands interested in domain estimation
- Domain estimation of total turnover for industries in Structural Business Survey (SBS)

Task

Consider methods that account for
- the potential effect of outliers
- the role of the sampling design
**CBS Data**

- **Target:** Total tax-turnover (20 industry domains)
- Tax-turnover correlated with turnover
- **Covariates**
  - Tax-turnover in previous year (tax1)
  - Size-class (SC)
  - Number of employees (WP)
  - Population data available

- **Sampling Design**
  - Stratified: Strata combination of industries and SCs
  - SRSWOR is selected from each strata
  - Larger firms are selected with probability 1
A Starting Point - A Nested Error Model

\[ \text{TaxTurnover} = \beta_0 + \beta_1 WP + \beta_3 SC + \beta_4 \text{tax}1 + \beta_5 WP*\text{tax}1 + v_j + \epsilon_{ij} \]

- 2-level (enterprises/industries) nested error regression model
- **Diagnostics:** Residual plots and Normal probability plots
- Diagnostics produced with sample selected using SBS design
Model-based domain estimation in NSIs in Europe
The Role of Design Weights

Model-based domain estimation in NSIs in Europe
First Observations

- Potential presence of outliers
- Non-Constant level 1 variance - Function of size measure
- Use of the design weights should be beneficial if conditional variance is a function of the size measure
- **BUT** not fully protecting against outliers

Elaborating the Basic Working Model

1. 2-level model level 1 variance function of SC
   AIC-Basic=29552.24 Vs. AIC-Heteroscedastic= 19501.94

2. Basic 2-level model with design weights
Diagnostics - Robust Estimation

- Huber weights protect against outliers, negatively correlated with size

Model-based domain estimation in NSIs in Europe
Empirical Study - Business Survey Data

- Use artificial population of enterprises provided by CBS
- Select 500 samples using SBS sampling design
- Target of Estimation: Total tax-turnover for 20 domains
- Model Spec: Tax-turnover in previous year (tax1), Size-class (SC), number of employees (WP), WP*tax1

**Aim**
Empirically evaluate the properties of alternative small domain point estimators of total Tax-turnover
Small Domain Estimators

We compare the performance of estimators of industry totals

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Working Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>EBLUP</td>
<td>Nested Error</td>
</tr>
<tr>
<td>EBLUP (Var=f(SC))</td>
<td>Nested Error</td>
</tr>
<tr>
<td>P-EBLUP</td>
<td>Nested Error</td>
</tr>
<tr>
<td>REBLUP</td>
<td>Nested Error</td>
</tr>
<tr>
<td>MQ Naive (k=1.345)</td>
<td>M-quantile</td>
</tr>
<tr>
<td>MQ BC (k1=1,2,3,100)</td>
<td>M-quantile</td>
</tr>
<tr>
<td>Robust synthetic</td>
<td>Single level</td>
</tr>
<tr>
<td>Robust synthetic (Var=f(WP))</td>
<td>Single level</td>
</tr>
<tr>
<td>Direct (HT)</td>
<td></td>
</tr>
</tbody>
</table>
Summary of Results

Median values of the %Relative Bias (RB) and Root MSE (RMSE)

<table>
<thead>
<tr>
<th>Estimator</th>
<th>RB(%)</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct</td>
<td>0.07</td>
<td>663.22</td>
</tr>
<tr>
<td>EBLUP</td>
<td>-1.29</td>
<td>430.20</td>
</tr>
<tr>
<td>EBLUP ($\text{Var}=f(\mathcal{S}C)$)</td>
<td>0.53</td>
<td>279.00</td>
</tr>
<tr>
<td>P-EBLUP</td>
<td>0.35</td>
<td>300.07</td>
</tr>
</tbody>
</table>

- Estimate of $\hat{\sigma}_u$ is small
- $\hat{\sigma}_u^{Rob}$ on the boundary of the parameter space
- The REBLUP is not computed
- **Alternative:** Use a robust-synthetic estimator

Model-based domain estimation in NSIs in Europe
Median values of the %Relative Bias (RB) and Root MSE (RMSE)

<table>
<thead>
<tr>
<th>Estimator</th>
<th>RB(%)</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Robust Synthetic</td>
<td>0.37</td>
<td>156.40</td>
</tr>
<tr>
<td>Robust Synthetic (Var=$f(WP)$)</td>
<td>0.20</td>
<td>146.20</td>
</tr>
<tr>
<td>MQ Naive</td>
<td>0.25</td>
<td>127.42</td>
</tr>
<tr>
<td>MQ BC (k1=1)</td>
<td>-0.05</td>
<td>120.18</td>
</tr>
<tr>
<td>MQ BC (k1=2)</td>
<td>-0.12</td>
<td>129.14</td>
</tr>
<tr>
<td>MQ BC (k1=3)</td>
<td>-0.09</td>
<td>139.72</td>
</tr>
<tr>
<td>MQ BC (k1=100)</td>
<td>0.07</td>
<td>451.26</td>
</tr>
</tbody>
</table>
Small Area Estimation of Complex Indicators

**Aim**
Previous slides focused on estimation of averages/totals. Now we present methodologies for the estimation of complex indicators. Application is in the estimation of deprivation indicators.

**Methodologies**
- The World Bank Method (Elbers et al., 2003)
- The Empirical Best Predictor (EBP) Approach (Molina & Rao, 2010)
Estimating Poverty (Income Deprivation) Indicators

- Define FGT (1984) measures $z_i(\alpha, t)$ at population level by

$$z_i(\alpha, t) = \left( \frac{t - y_i}{t} \right)^\alpha I(y_i \leq t) \quad i = 1, \ldots, N.$$  

1. $a = 0$ - Head Count Ratio / Incidence of Poverty
2. $a = 1$ - Poverty Gap
3. $a = 2$ - Poverty Severity

- In small domain $j$, $z_j(\alpha, t)$, decomposed as follows,

$$z_j(\alpha, t) = N_j^{-1} \left[ \sum_{i \in s_j} z_i(\alpha, t) + \sum_{k \in r_j} z_k(\alpha, t) \right].$$

- Target: Estimate $z_k(\alpha, t)$

Model-based domain estimation in NSIs in Europe
The EBP Approach

Uses a nested error regression model with area effects

\[ \hat{z}_j = N_j^{-1} \left[ \sum_{i \in s_j} z_i + \sum_{k \in r_j} \hat{z}_{EBP}^k \right] \]

- \( \hat{z}_{EBP}^k \) computed using the predictive density \( f(y_r|y_s) \)

- Fit the nested error regression model with area effects
- Estimate \( \beta, \sigma_v, \sigma_\epsilon \)
- Generate area effects \( u_j^* \sim N(0, \hat{\sigma}_v^2 (1 - \gamma_j)) \) & random errors \( \epsilon_{ij}^* \sim N(0, \hat{\sigma}_\epsilon^2) \)

\[ y_{ij}^* = x_{ij}^T \hat{\beta} + \hat{\nu}_j + u_j^* + \epsilon_{ij}^* \]

- Repeat the process \( L \) times each time estimating the target
Case Study II - Income Deprivation in the UK

- Office for National Statistics: Interested in producing estimates of the proportion of people below the poverty line
- Currently experimenting with the EBP approach
- Survey Data - 2001/02 Family Resources Survey
- Census Data - 2001 Census microdata
- **Target Geography:** Local Authority Districts in 2 Regions
- North-west Vs. South East (North/South Divide in the UK)
Case Study II (Cont’d)

- Nested error regression model (households nested in LADs)
- $y$: log weekly househ. equivalised income after househ. costs
- Covariates: Education, Employment, Ethnicity, Receipt of Benefits, Gender, Household characteristics
- Poverty line: £164.64 (60% of weekly median income)
Model-based domain estimation in NSIs in Europe
UK North-South Divide

Model-based domain estimation in NSIs in Europe
Thoughts on the Poverty Mapping Methods

- Powerful - End product useful for policy
- Requires access to Census/admin microdata
- Data access complicated - Use in a safe setting
- Time gap between survey and Census
- Use of administrative data sources
- How to handle PSUs? Identify PSUs in the Census
- Income is one deprivation dimension

Model-based domain estimation in NSIs in Europe
Aim is to replace the Census from 2021 onwards
New methodology to produce more frequent outputs
Consultation process:
  - What method?
  - What data sources are available and legal issues?
  - SAE potentially to play a part in the new regime
  - Investigate SPREE & Data fusion