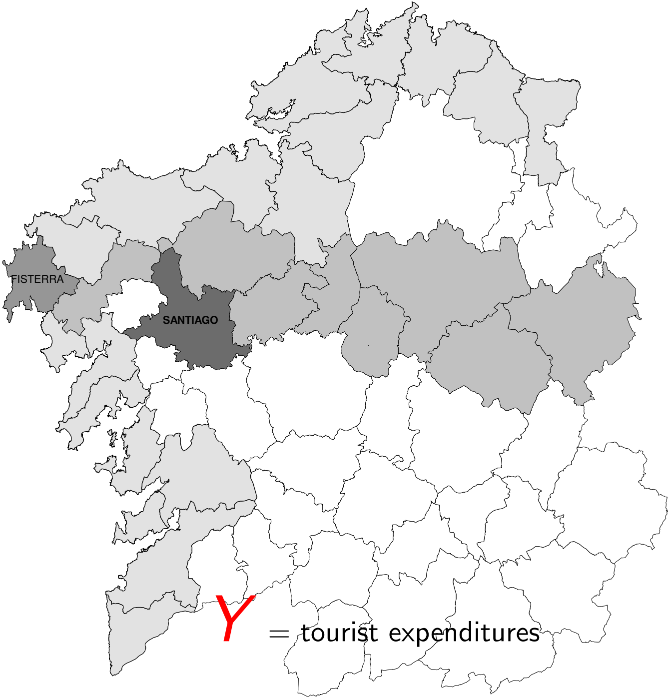
Evoking ideas from small area estimation for the timely indicator prediction on a high level of disaggregation

Stefan Sperlich

Research Center for Statistics

and

Institute for Economics and Econometrics



Y = tourist expenditures

The SAE Problem – 1

Assume, can calculate your area parameter (e.g. indicator) from y_i $i = 1, \ldots, N$ (in sample n), for areas $d = 1, \ldots, D$, defining N_d, n_d

But

- too few y_i observed in most areas for direct estimation
- such that even design-based and synthetic estimators don't *work*
- you may even have no y_i in some areas ($n_d = 0$ but $N_d > 0$)

Idea

- use auxiliary information (x_i and/or v_d) – which is useful?
- and use a prediction model relating it to Y_{id}

Note: The latter may be unspecified, then think of *machine learning*

A popular SAE model

A most general way is to think of the DGP

$$y_{id} = \varphi_d(x_i, v_d,\ \epsilon_i, u_d)$$

Model not identified that generally; so be more restrictive, i.e.

$$y_{id} = m(x_i, v_d) + \eta_d(z_i) + \epsilon_{id}\ , \qquad z_i \subseteq x_i$$

with $\eta_d(\cdot)$ tackled as *random (effects)* function, as n_d too small

In practice often set $\eta_d(Z_i) := u_d' Z_i$ with u_d random coefficients

If $m(X_i, V_d) := \beta' X_i + \alpha' V_d$ then simply linear mixed model (LMM)

Notes: 1. all random terms are centered to zero

2. $m(\cdot)$ borrows strength from entire data set

3. haven't used geography or similar, so 'area' is any *given cluster*

Special Cases

Popular examples and straightforward extensions:

- Nested error models (i.e. Z only contains constant)
- Random regression coefficients model (i.e. $m(\cdot) \equiv 0$)
- Generalized MML with known link G (e.g. logit for binary Y)
- Area-level models (i.e. only V_d, U_d, maybe area-averages of X)
- Longitudinal (or panel) random coefficients models
  (then d for subject, i for time or repeated measurement)

Note: most extensions only fully parametric

Typical Conditions – Typical Problems

D.1 Y_{id} $(d = 1, \ldots, D; i = 1, \ldots, n_d)$ are conditionally independent given $(\boldsymbol{u}_d, \boldsymbol{V}_d, \boldsymbol{X}_{di})$, (SAE often takes $\boldsymbol{X}_{id}$ as fixed)

D.2 parameter of η_d, say u_d are i.i.d. across areas, mean zero and *known* covariance structure (matrix) $\boldsymbol{\Sigma}_u \in \boldsymbol{R}^{\rho \times \rho}$,

D.3 Independence conditions of u_d w.r.t. the rest

F.1 conditions on $m(\cdot)$ and $G(\cdot)$, and maybe also conditions for existence of MLE

Extra All further conditions depend on particular model specification and data

Note: Clearly, most problematic are the independences

Steps & Main Problems in SAE

Assume, data are already collected (given)

For sake of notation omit $\boldsymbol{V}_d$ here

1. Define appropriately Y and prediction model
2. Find appropriate estimator / predictor for $m(\cdot)$ and $\eta_d(\cdot)$
   will take advantage of $\boldsymbol{\Sigma} := Var[Y|\boldsymbol{X}] \quad (= \boldsymbol{Z}\boldsymbol{\Sigma}_u\boldsymbol{Z}' + \boldsymbol{\Sigma}_\epsilon$ in LMM)
3. Estimate **MSE**, construct **prediction intervals**

Note: huge literature on last issue, a lot about second, some about first

Remark: for complete SAE procedure see
From start to finish: a framework for the production of small area official statistics
JRSS-A (2018) Tzavidis, Zhang, Luna, Schmid, Rojas-Perilla

Steps & Solutions in SAE

1. Take observed, transformed, imputed values?
2. either likelihood: maximize for known C_d and weights $\boldsymbol{W}_d$

$$-\sum_{d=1}^{D}\{y_d - m(\boldsymbol{X}_d)\}'\boldsymbol{W}_d^{1/2}\boldsymbol{\Sigma}_d^{-1}\boldsymbol{W}_d^{1/2}\{y_d - m(\boldsymbol{X}_d)\} + C_d(\boldsymbol{\Sigma}_d)$$

   or (different, maybe local) 2 or 3 step least-squares method
3. For MSE, *bootstrap* methods are becoming popular
4. independence and occasionally distribution assumptions

Note: notation in step 2 for local smoother, $\boldsymbol{W}_d$ could be kernel weights; you may use sieves like splines for $m(\cdot)$ and skip $\boldsymbol{W}_d$

Fore-/Now-casting 1

To fore- or nowcast $Y_{id,t}$

obvious approaches are

- in SAE context $m(\boldsymbol{X}_{id,t-k}, \boldsymbol{V}_{d,t-k})$, choice of lag k case-dependent
- if 'sufficiently many' $y_{id,t-k}$, restrict to that sample, include them in $\boldsymbol{X}_{id,t-k}$ and see
- if not 'sufficiently many' but known to be excellent predictors, you may first impute them, and then include them
- transformations and robust methods

Remark: key for our choice is a proper *objective function* for prediction

Fore-/Now-casting 2

Less obvious,

and so far even less explored approaches are

- explore flexibility/prediction power of **ML** for $m(\cdot)$ (and $\eta_d(\cdot)$?)
- explore inclusion of more $\boldsymbol{X}_{id,t-k}$, including time trend, and select

Remarks:

apart from 1. proper *objective function*,

main challenges are 2. inclusion of covariance structure,

3. MSE estimation and further inference

Modeling Random Parts

Various proposals have been made, many based on idea

$$y_{id,t} = m(x_{id,t}, t) + u_d' Z_{id,t} + \mu_{d,t} + \epsilon_{id,t}$$

with μ_d such that $Var[\mu] = \sigma_\mu^2 \Omega$ and independent of rest, eg. for AR(1)

$$\Omega = diag\{\Omega_d\}_{d=1}^D, \quad \Omega_d = \left\{ \begin{array}{ccccc} 1 & \omega & \cdots & \omega^{T-2} & \omega^{T-1} \\ \omega & 1 & \ddots & & \omega^{T-2} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \omega^{T-2} & & \ddots & 1 & \omega \\ \omega^{T-1} & \omega^{T-2} & \cdots & \omega & 1 \end{array} \right\}$$

Recalculate $\mathbf{\Sigma} = Var[Y|\boldsymbol{X}]$ and apply similar formulas as before.

Note: this makes a big difference as $\hat{\mu}_{d,t}$ are used for prediction

Challenges for Estimation and MSE

However, until today not very popular

- for nonlinear models not efficient to work with $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{y}$, $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}$, ...
- rather think in terms of using them for efficient weighting
- automatically the case in likelihood estimation, but likelihood might be complex
- hardly software available
- for generalized models (nontrivial link, when Y is discrete)
- estimation of MSE even more complex; typically bootstrap methods proposed

Conclusions

- SAE is standard in NSOs for predicting indicators on highly disaggregated level
- Among existing methods, the GLMM assisted are the most popular ones when samples of y_is are small
- until today lively research area in official and applied statistics
- has many overlaps with biometrics, e.g. analysis of longitudinal data
- extensions for prediction in time are so far rare but obvious

References: among various

A course on SAE and mixed models (2020) Morales, Esteban, Prez, Hobza

Kernel smoothers and bootstrapping for semiparametric mixed effects models, JMVA (2013) Gonzlez-Manteiga, Lombarda, Martnez-Miranda, Sperlich

European grant proposal for training network

proposed Working Packages are

1. integrating administrative and non-official (big data) in SAE
2. Combining Machine Learning
3. and Smart Modeling (like time ans space structure)
4. Model selection, including variable selection
5. robustness issues, data transformation
6. post-selection and uniform inference