Small area solutions for the analysis of pollution data

Daniela Cocchi, Enrico Fabrizi, Carlo Trivisano

Università di Bologna

TIES 2002

Genova, June 22nd 2002

Summary

The small area problem in the environmental context
Theoretical framework
The linear hierarchical model
Special case: ANOVA model
Approximating criterion: least squares
Prior conjectures
Statistic chosen for conditioning
The normal case and the departure from normality
Comments on least squares approximations
A simulated experiment
Consequences of misspecifications in prior evaluations
High applicability to environmental problems
Extensions

The small area problem in the environmental context

Simultaneous estimation of parameters related to different subpopulations (domains) of a more general population

Desired property: additivity

Example 1

Erosion from agricultural land in a watershed (Opsomer, Botts, Kim, 2001)

Population of areas (160 acres plot)

Auxiliary variables available

Mixed effect linear model

Example 2

Emissions inventories

(CORINAIR project in EU)

national estimates of emission volumes (by pollutant)

many domain estimates

- spatial
- economic sectors
- time evolution

Building inventories

- Census of major pollution sources
- Indirect estimation of small and very diffuse sources
  - Activity indicators
  - Emission factors

Hierarchical classification of emission activities
SNAP system (Selected Nomenclature for Air Pollution)

$$E_{p,a,t} = A_{a,t} F_{p,a}$$

$A_{a,t}$: volume of activity a in period t subject to sampling random variation; $F_{p,a}$: emission factor for pollutant p in activity a; seldom estimated on a sample basis.

$$E_{p,t} = \sum_{a=1}^{N} E_{a,p,t}$$

Approach: bottom up.

Suggestion: ANOVA model is a realistic proposal for $A_{a,t}$ since it does not use auxiliary information which is very heterogeneous and suffers from problems of spatio-temporal definition.

Theoretical framework

Typical of the finite population context

It can be solved within

The design based approach

parameters: unknown characteristics of the subpopulations

The model based approach

where hyperparameters are considered

We discuss:

Model based approach

Hierarchical modeling

stressing Bayesian solutions

Two main motivations

A natural way of building and solving models

Managing different aggregation level of the information

The linear hierarchical model

It is a structural model for the population

$$\eta = Z_2\beta + \varepsilon \qquad \eta, \varepsilon : N \times 1 \qquad Z_2 : N \times p \qquad \beta : p \times 1$$
$$\beta = Z_1\beta_0 + \delta \qquad Z_1 : p \times q \qquad \beta_0 : q \times 1 \qquad \delta : q \times 1$$

where

$$Z_2 = diag\left[i_{Nk}\right]$$

is the small area indicator matrix and the incidental parameter β is the vector of small area means object of inference.

Assumptions on the model (only on the first two moments of the distributions):

$$C\left(\delta', \varepsilon \middle| Z_1, Z_2\right) = 0$$

$$E\left(\varepsilon \middle| Z_2, \sigma_\varepsilon^2, \beta\right) = E\left(\varepsilon \middle| \sigma_\varepsilon^2\right) = 0$$
$$V\left(\varepsilon \middle| Z_2, \sigma_\varepsilon^2, \beta\right) = V\left(\varepsilon \middle| \sigma_\varepsilon^2\right) = \sigma_\varepsilon^2 V$$

$$E\left(\delta \middle| Z_1, \sigma_\delta^2, \beta_0\right) = E\left(\delta \middle| \sigma_\delta^2\right) = 0$$
$$V\left(\delta \middle| Z_1, \sigma_\delta^2, \beta_0\right) = V\left(\delta \middle| \sigma_\delta^2\right) = \sigma_\delta^2 B_0$$

A set of assumptions of conditional independence is able to simplify the model.

Many small area models are special cases of this general model.

Role of covariates

a) At the individual level

The slightly different Battese et. al. (1988) model

$$\eta = X\beta + \nu$$

$$\nu = Z_2\delta + \varepsilon$$

$$V = I_N \quad B_0 = i_p i_p' \quad Z_1 = I_p$$

with a β value on the whole population and small area random effects

b) At the area level

$$N = p$$

$$\eta, \varepsilon : p \times 1 \qquad Z_2 = I_p \qquad \beta : p \times 1$$

Fay and Herriott (1979) model adds distributional assumptions

$$e_k \sim N\left(0, D_k\right), \qquad k = 1, \ldots, p$$

$$\delta_k \sim N\left(0, A\right), \qquad k = 1, \ldots, p$$

Frequentist solutions

Empirical Best Linear Unbiased Prediction.

Estimating σ_δ^2 and σ_ε^2, and introducing the estimates in the Best Linear Unbiased Predictor of β.

Special case: ANOVA model

Some basic simplifications

$$V = I_N$$

$$Z_2 Z_1 = i_N$$

A comprehensive statistical model

A sampling model can be added to the structural model

$$S = \left(e'_{s1}, \ldots, e'_{sn}\right)' : \left(n \times N\right)$$

where e_{si} is the s_i column of I_N.

Approximating criterion: least squares

The general least squares principle

Approximation to a normal distribution with the same first two moments of the exact one.

Finite populations: this approximation is conditional on (Z, S).

Normal approximation on $(\beta, t | Z, S)$ and not on the whole (β, t, Z, S).

TYPE OF INFERENCE: Bayesian

Why approximated solutions when MCMC solutions which approximated posterior distributions are easily available?

Because they are analytically approximated and need the elicitation of a relatively small number of prior guesses.

Be X the data and ψ the parameter, for any $\beta = g(\psi)$ and $t = t(X)$, we define $l(t) = \{a't\}$.

If:

$$l_{Z,S}(t) = \{a(z,s)' t\}$$

$$E_{LS}^{Z,S}(\beta|t) = \arg\min_{l} E\left(\left\|\beta - l_{Z,S}(t)\right\|^2 |Z,S\right)$$

$$= \arg\min_{l} E\left(\left\|E(\beta|t,Z,S) - l_{Z,S}(t)\right\|^2 |Z,S\right)$$

$$= E(\beta|Z,S) + C(\beta,t'|Z,S)\{V(t|Z,S)\}^{-1}\{t - E(t|Z,S)\}$$

$$V_{LS}^{Z,S}(\beta|t) = \min_{l_{Z,S}} E\left(\left\|E(\beta|t,Z,S) - E_{LS}^{Z,S}(\beta|t,Z,S)\right\|^2 |Z,S\right)$$

$$= V(\beta|Z,S) + C(\beta,t'|Z,S)\{V(t|Z,S)\}^{-1} C(t,\beta'|Z,S)$$

(Cocchi and Mouchart, 1996)

Prior conjectures

Only on a certain number of moments

Conjecturing on moments is more immediate than conjecturing on distributional assumptions

Their type and number depend on the statistic on which conditioning

Statistic chosen for conditioning

Its choice can enrich the solution and determines its complexity

a) Solution conditional on

$t = T_0 = Z_2' S' y$: vector of small area totals

Priors to be elicitated

$$E\left(\beta_0 \middle| Z_1\right) = E\left(\beta_0\right) = b_0$$
$$V\left(\beta_0 \middle| Z_1\right) = V\left(\beta_0\right) = M_0$$
$$v_\delta = E\left(\sigma_\delta^2 \middle| Z_1\right) = E\left(\sigma_\delta^2\right)$$
$$v_\varepsilon = E\left(\sigma_\varepsilon^2 \middle| Z_2\right) = E\left(\sigma_\varepsilon^2\right)$$

b) Solution conditional on

$$t = T = \left(T_0, T_1, T_2\right)$$

i.e. the vector of small area totals and
the sum of squares

T_δ: between small areas and

T_ε: within small areas

Since T contains a polynomial of order d, prior information on moments of order $2d$ must be used.

For computing the solution the following moments are needed

$$\alpha_j = E\left(\delta_k^j \middle| Z_1, \beta_0, \sigma_\delta^2\right), \qquad j=3,4$$

$$\varphi_j = E\left(\varepsilon_k^j \middle| Z_2, \beta, \sigma_\varepsilon^2\right), \qquad j=3,4$$

New priors to be elicitated besides the group of priors above:

$$V_\delta = V\left(\sigma_\delta^2 \middle| Z, S\right) \qquad V_\varepsilon = V\left(\sigma_\varepsilon^2 \middle| Z, S\right)$$

$$a_3 = E\left(\alpha_3 \middle| Z, S\right) \qquad a_4 = E\left(\alpha_4 \middle| Z, S\right)$$

$$f_3 = E\left(\varphi_3 \middle| Z, S\right) \qquad f_4 = E\left(\varphi_4 \middle| Z, S\right)$$

$$c_{0,1} = C\left(\beta_0, \sigma_\delta^2 \middle| Z, S\right)$$

It is not a normal approximation!

It does not mean approximating normal distributions with the same first two moments: the model may contain moments up to the 4th.

Gain: this solution can consider asymmetry and kurtosis.

Solution

a) For $t=T_0$

$$E_{LS}^{Z,S}\left(\beta\middle|T_0\right)=\mathbf{a}\left[gb_0+\left(1-g\right)\bar{\bar{y}}\right]+\left(I_p-\Delta_{\mathbf{a}}\right)\mathbf{y}$$

$$=\mathbf{w}v_{\varepsilon}\left[gb_0+\left(1-g\right)\bar{\bar{y}}\right]+v_{\varepsilon}\left(v_{\varepsilon}I_p-\Delta_{\mathbf{w}}\right)\mathbf{y}:\quad p\times 1$$

$$V_{LS}^{Z,S}\left(\beta\middle|T_0\right)=v_{\delta}\Delta_{\mathbf{a}}+M_0g\mathbf{a}\mathbf{a}'$$

$$=\mathbf{w}v_{\varepsilon}\Delta_{\mathbf{w}}+M_0v_{\varepsilon}^2g\mathbf{w}\mathbf{w}':\quad p\times p$$

where

$$\mathbf{a}=\left[a_k\right]=\left[v_{\delta}^{-1}\left(v_{\delta}^{-1}+n_kv_{\varepsilon}^{-1}\right)^{-1}\right]=\left[v_{\varepsilon}\left(v_{\varepsilon}+n_kv_{\delta}\right)^{-1}\right]:\quad p\times 1$$

$$w = \left[w_k \right] = v_\varepsilon^{-1} \left[a_k \right] = \left[\left(v_\varepsilon + n_k v_\delta \right)^{-1} \right] : \quad p \times 1$$

$$\Delta_{\mathbf{a}} = diag\left[a_k \right] : p \times p \quad \Delta_{\mathbf{w}} = diag\left[w_k \right] = \left(v_\varepsilon + n_k v_\delta \right)^{-1} : p \times p$$

$$\bar{\bar{y}} = \frac{\sum_k \frac{n_k \bar{y}_k}{v_\varepsilon + n_k v_\delta}}{\sum_k \frac{n_k}{v_\varepsilon + n_k v_\delta}} = \frac{v_\varepsilon \, \mathbf{w}' \Delta_{\mathbf{n}} \mathbf{y}}{v_\varepsilon \, \mathbf{w}' \Delta_{\mathbf{n}} i_p} = \frac{\mathbf{w}' T_0}{\mathbf{w}' \mathbf{n}}$$

$$g = \left(1 + M_0 \mathbf{w}' \mathbf{n} \right)^{-1} = M_0^{-1} \left(M_0^{-1} + \mathbf{w}' \mathbf{n} \right)^{-1}$$

$$\Delta_{\mathbf{n}} = Z_2' S' S Z_2 = diag\left[n_k \right] : p \times p \qquad \mathbf{n} = \Delta_{\mathbf{n}} i_p : p \times 1 \qquad \mathbf{y} = \Delta_{\mathbf{n}}^{-1} T_0 : p \times 1$$

$$y = n^{-1} i_n' \mathbf{y} = n^{-1} i_p' T_0 = n^{-1} \mathbf{n}' \mathbf{y}$$

b) For $t=T$

heavier calculations, no elegant formula because of the difficulty of writing analytically

$$\left[V\left(T\middle|Z,S\right)\right]^{-1}=\begin{bmatrix}\mathbf{V}^{00} & \mathbf{v}^{01} & \mathbf{v}^{02}\\ \mathbf{v}^{10} & v^{11} & v^{12}\\ \mathbf{v}^{20} & v^{21} & v^{22}\end{bmatrix}:\quad (p+2)\times(p+2)$$

$$E_{LS}^{Z,S}\left(\beta\middle|T\right)=\left(I_p-C_*\right)i_p b_0+C_*\mathbf{y}+C_{**}\begin{bmatrix}T_\delta-\left[v_\varepsilon\left(p-1\right)+v_\delta\left(n-n^{-1}\mathbf{n}'\mathbf{n}\right)\right]\\ T_\varepsilon-v_\varepsilon\left(n-p\right)\end{bmatrix}$$

where

$$C_*=\left[C_0V^{00}+c_1\left(v^{01}\right)'\right]\Delta_{\mathbf{n}}:\qquad p\times p$$

$$C_{**}=\begin{bmatrix}C_0v^{01}+c_1v^{11} & C_0v^{02}+c_1v^{12}\end{bmatrix}:\qquad p\times 2$$

after defining:

$$C\left(\beta\middle|T'\right)=\begin{bmatrix}C_0 & c_1 & 0\end{bmatrix}:\qquad p\times\left(p+2\right)$$

The normal case and the departure from normality

In case of symmetry: $\left[V\left(T\middle|Z,S\right)\right]^{-1}=\begin{bmatrix}\left[V\left(T_0\middle|Z,S\right)\right]^{-1} & 0\\ 0 & \left[V\begin{pmatrix}T_\delta\\ T_\varepsilon\end{pmatrix}\middle|Z,S\right]^{-1}\end{bmatrix}$

when normality holds: $C\left[T_\delta,T_\varepsilon\right]=V_\varepsilon\left(p-1\right)\left(n-p\right)$

When $c_{0,1}=C\left(\beta,T_\delta\middle|Z,S\right)=0$

Then $C_{**}=0$ and $C\left(\beta\middle|T'\middle|Z,S\right)\left[V\left(T\middle|Z,S\right)\right]^{-1}=C\left(\beta\middle|T_0'\middle|Z,S\right)\left[V\left(T_0\middle|Z,S\right)\right]^{-1}$

As a consequence

$$E_{LS}^{Z,S}\left(\beta\middle|T\right)=E_{LS}^{Z,S}\left(\beta\middle|T_0\right)$$

Comments on least squares approximations

They borrow strength from the other subgroups and from the prior conjectures

The idea recalls two important points :

posterior linearity (for estimating the general mean starting from the least squares approximations of the small area parameters)

empirical Bayes solutions.

The solution is robust

since it depends only on moments.

The solution seems not to have practical drawbacks linked to the number of small areas or to the fact of not having evidence for some subgroups.

A simulated experiment

Aim: Checking the improvement of conditioning on T instead of T_0 in the approximated posterior expectations.

Data generation process: random generation of β and η from

normal - normal

lognormal - lognormal.

Lognormal parameters are set in order to have the same first two moments of the normal distribution.

Exact Bayesian least squares solutions

i.e. right prior conjectures.

Design of the experiment

1. generation of 1000 populations of N= 1000 elements with mean β_0=1.

under 4 different hypotheses on structural variability:

$\sigma_\delta = 1, \sigma_\varepsilon = 0.5$ $\quad$ $\sigma_\delta = 0.5, \sigma_\varepsilon = 0.5$

$\sigma_\delta = 2, \sigma_\varepsilon = 1$ $\quad$ $\sigma_\delta = 1, \sigma_\varepsilon = 1$

and 2 different hypotheses on the domains: p=10; p=40;

2. generation of 100 samples from each population with

n_k =10, k=1,…, p when p=10 $\quad$ n_k =5, k=1,…, p when p=40;

3. evaluation of the performances of the posterior expectations by means of mean square errors averaged over the domains.

Lognormal – Lognormal

		p=10 n_k=20	p=40 n_k=5
σ_δ=1	σ_ε=0.5	1.043	1.088
σ_δ=0.5	σ_ε=0.5	1.012	1.023
σ_δ=2	σ_ε=1	1.085	1.132
σ_δ=2	σ_ε=2	1.031	1.054

Normal –Normal ($c_{0,1} = 1$)

		p=10 n_k=20	p=40 n_k=5
σ_δ=1	σ_ε=0.5	1.012	1.021
σ_δ=0.5	σ_ε=0.5	1.001	1.009
σ_δ=2	σ_ε=1	1.023	1.044
σ_δ=2	σ_ε=2	1.011	1.014

Consequences of misspecifications in prior evaluations

a) Not important: errors in conjectures on variances

$$V(\beta_0) \quad V(\sigma_\delta^2) \quad V(\sigma_\varepsilon^2)$$

b) As expected: errors in conjectures on $E(\beta_0)$

c) The most dangerous: errors in conjectures on expectations of variances

$$E(\sigma_\delta^2) \quad E(\sigma_\varepsilon^2)$$

d) Exchanging the right lognormal and normal conjectures

For normal-normal data generation: the most dangerous are the errors on the 3rd moment.

For lognormal - lognormal data generation: the most dangerous are the errors on the 4^{th} moments.

High possibility of application to environmental problems

Suitable solution for estimating emission inventories since

a) the compositeness and heterogeneity of local emissions is so high that a random effect model is in many cases the only practicable solution

b) all methods for building emissions inventories involve expert evaluations, which can be managed as priors.

Extensions

introduction of covariates

relationships with the design