Multilevel Models and Small Area Estimation in the Context of Vietnam Living Standards Surveys

Phong Nguyen, Dominique Haughton, Irene Hudson, John Boland

General Statistics Office, Hanoi, Vietnam, Bentley University and Toulouse School of Economics, University of South Australia, University of South Australia
nphong@gso.gov.vn, dhaughton@bentley.edu, irenelena.hudson@gmail.com, john.boland@unisa.edu.au

Abstract: This talk will discuss a methodology to obtain small area estimates in the context of the Vietnam Living Standards Surveys. The presentation will proceed in three parts. First we will introduce the Viet Nam Living Standards Surveys, their historical development, topics covered, sample size issues and challenges. Second, we will briefly review main concepts in small area estimation, including the use of auxiliary data, and will contrast simple small area models with regression small area models. This will then lead to the notion of random effects in small area regression models, and to our proposed multilevel model for small area estimation at the commune level in Vietnam, to our knowledge the first such model built with Vietnam living standards data. The third part of the talk will discuss this model. Our proposed multilevel model for estimating the commune-level mean (log of) household expenditure per capita relies on independent variables available both in the 1999 Census and in the Vietnam Household Living Standards Survey of 2002. Following ideas given in work by Moura (1994, 1999), the small area estimation is performed by plugging the population means of the independent variables into the regression equation, inclusive of suitable random effects both in the intercept and in the coefficient of the dummy variable for the urban location of a household. We will discuss how the random effects in the model can also be used to examine the urban-rural gap across the country. We will also mention how to measure the accuracy of our small area estimators. Finally, we will touch upon the use of sampling weights in models such as presented in the talk.

Vietnam Living Standards Surveys

The model used in this paper relies on data from the Viet Nam Household Living Standards Survey (VHLSS) of 2002, and we refer to past work which relies on the Viet Nam Living Standards Surveys (VLSS) of 1993 and 1998. In this section we describe a few relevant features of the surveys, and the context in which the VLSS, and then the VHLSS program were established under the auspices of the General Statistics Office (GSO 1999) in Viet Nam. Further details are available from Nguyen Phong and Haughton (2006).

The VLSSs were implemented in 1993 and 1998 in Viet Nam with financial support from the United Nations Development Program (UNDP) and the Swedish International Development Agency (SIDA), and with technical support from the World Bank. The survey methodology follows the World Bank's Living Standards Measurement Study (LSMS), listed in the bibliography, covering the following areas displayed in Table 1.

The sample size was of 4,800 households in 1993 and 6,000 households in 1998, including 4,300 1993 households which were re-interviewed. The sample was divided into 10 parts and each month one tenth of the sample was covered by the VLSS, in an attempt to avoid seasonal effects.

The questionnaire wrote out the exact questions to be used by the interviewers, and data entry was performed in the field. The survey involved a very high rate of supervision, with one supervisor for every two interviewers.

Table 1. Areas covered by the VLSS surveys (1993 and 1998)

1. Income
2. Expenditure
3. Education
4. Health (including height, weight and arm circumference of all household members)
5. Employment
6. Agricultural activities
7. Non-farm business activities
8. Housing
9. Migration
10. Fertility
11. Savings and credit

The VLSS data are widely considered to be of very high quality; however some limitations include the fact that no direct estimates for provincial level were possible because of the relatively small sample size, that a long period of time elapsed between the two VLSSs, and finally that the cost of the survey was high, at $163 per household interviewed. The VHLSS program was established to try to address some of these limitations.

The VHLSS program

During 2000-2010 the plan is for the GSO to conduct a Viet Nam Household Living Standards Survey every two years: 2002, 2004, 2006, 2008 and 2010. So far, VHLSS 2002, 2004 and 2006 have been collected and VHLSS 2008 is currently in the field. Each survey year, a core module is conducted. Every four years or more, additional modules are conducted. Topics for core modules are displayed in Table 2 for core modules and in Table 3 for additional modules.

Table 2. Core module areas in the VHLSS program

1. Basic demographic information on all household members (age, sex, relationship to head)
2. Household expenditures (food, education, health, etc.)
3. Household income (wage and salary, farm production, non-farm production, remittances, etc.)
4. Employment and labor force participation
5. Education: a small number of questions (literacy, highest diploma, fee exemption)
6. Health: a small number of questions (use of health services, health insurance)
7. Housing: a small number of questions (type of housing, electricity, water source, toilet, etc.)
8. Assets and durable goods
9. Participation in poverty programs
10. A commune questionnaire with information on local infrastructure

Table 3. Additional modules in the VHLSS program

1. Detailed information on agricultural activities and non-agricultural household businesses, borrowing and lending.
2. Detailed information on health and education of household members. Questionnaires for commune health centers and local schools.
3. Infrastructure, environment, local institutions and governance.

VHLSS 2002

The data used in this paper originate from VHLSS 2002. We summarize below the main features of this survey and its questionnaire design:

- VHLSS 2002 covered only the core module.
- The 2002 VHLSS questionnaire is similar in many respects to the 1997-98 VLSS questionnaire.
- Six of the 9 sections in the questionnaire are very similar to the 97-98 VLSS: Household Roster, Education, Employment; Income; Housing; Food Expenditures and Non-Food Expenditures.
- The Health Section is similar to that in the 1997-98 VLSS, but also incorporates ideas from the 2001-2002 Viet Nam Health Survey.

The sample size in 2002 is of 75,000 households (of which 30,000 are expenditure households, which implies that questions were asked both about income and expenditures of these households).

Questionnaire design for VHLSS 2002 (and later years)

- The exact questions asked of households are printed out in the questionnaire.
- Questions were designed to ensure comparability with past surveys, especially for expenditure and income data
- The data entry was performed at the provincial level
- The field work was conducted as follows:
  - VHLSS 2002: Four rounds (four quarters)
  - VHLSS 2004 and 2006: Two rounds (May and September)
- Personnel and Training Issues:
  - Field workers are GSO staff members
  - Training for trainers was held in the North and South of the country
  - Training for interviewers was held in each province

This chapter studies whether some communes, districts and provinces had more 'effective' influence than others in promoting households' living standards, taking account of variations in the characteristics of households.
We use multilevel modeling technique as a tool to implement the study.

Brief introduction to small area estimation

Small area estimation is widely used in a number of national statistics offices over the world. References are many, but for the purposes of this paper a very useful reference is the Small Area Estimation manual by the Australian Bureau of Statistics (ABS 2005).

Small area estimation methods are often divided into two main types of methods: "simple small area methods", such as for example direct estimation (where small area estimators are obtained directly from survey data) which typically yields an unbiased estimation but with a large standard error because of small sample sizes), and methods such as broad area ratio estimators (ABS 2005). In this paper, we focus attention on small area methods which rely on a regression model. In many applications a regression model is used with independent variables available for the entire population (such as via a census), and the model is applied to obtain estimates of for example the mean of the dependent variable at the small area level. When the regression model does not include any random effects that might capture local effects, the methods is often referred to as "synthetic regression models".

In this paper, we follow up on work by Moura and colleagues (1994, 1999) who began to promote the use of random effects in regression models to obtain improved small area estimators.

Our multilevel model

The model we have constructed is a four-level model for a one-year period using the 2002 Vietnam Household Living Standards Survey (VHLSS 2002). The four levels include the household level *i* (lowest level), commune level

j, district level *k* and provincial level *l* (highest level).
The dependent variable is the logarithm of real per capita household expenditure. Independent variables include 21 variables (listed below) that reflect household characteristics (measured at the household level) from VHLSS 2002 and that are also available in the 1999 Vietnam Population and Housing Census (Census 1999). Our model in its most general form can be described as follows:

$$Y_{ijkl} = \beta_{0jkl} + \sum_{p} \beta_{pjkl} X_{pijkl} + \varepsilon_{ijkl} \quad (1)$$

$$\beta_{0jkl} = \gamma_{00} + \gamma_{01} Z_{jkl} + f_{0l} + v_{0kl} + u_{0jkl} \quad (2)$$

$$\beta_{pjkl} = \gamma_{p0} + \gamma_{p1} Z_{jkl} + f_{pl} + v_{pkl} + u_{pjkl} \quad (3)$$

where the Y_{ijkl} represent the values of the dependent variable at the first level (household level). This is in our case the logarithm of the real per capita expenditure of the i^{th} household (*i*=1, ..., n_j , level 1) in the j^{th} commune (*j*=1,..., m_k , level 2) of the k^{th} district (k=1,..., r_l , level 3) of the l^{th} province (l=1,...,61, level 4); the X_{pijkl} represent the values of the p^{th} explanatory variable measured at the first level (household level) of the i^{th} household in the j^{th} commune of the k^{th} district of the l^{th} province; here *p* =1,...,21, corresponding to the 21 variables listed below.
Note that in VHLSS 2002, the value of n_j , in principle equal by design to 25 for all communes, in fact varies: 759 communes had more than 5 households (17-25) in the sample, and 2,142 communes had 3-5 households in the sample.
The β_{0jkl} represent the regression intercepts (for each commune *j* in district *k* in province *l*), and the β_{pjkl} represent the regression coefficients (slopes) (for each commune *j* in district *k* in province *l* for each of the 21 independent variables, *p*=1,...,21).
The error terms ε_{ijkl} represent the usual residual error terms assumed to have mean 0 and variance σ^2_{ijkl} typically assumed to be constant equal to a common error variance σ^2 (a property referred to as homoskedasticity).
The Z_{jkl} denote the values of one independent variable, measured at the commune level *j* (in district *k* in province *l*); to simplify notations, we assume that we have only one such variable, but the model extends easily to more than one such variable.
The coefficients γ_{00} , γ_{01} , γ_{p0} , γ_{p1} are fixed regression coefficients, and the u_{0jkl} and u_{pjkl} are random residual error terms at the commune level, assumed to have a mean of zero and to be independent from the ε_{ijkl} . In addition the u_{0jkl} and u_{pjkl} are assumed to have a constant variance.
In a similar way, the v_{0kl} and v_{pkl} are random residual error terms at the district level, assumed to have a mean of zero and to be independent from the ε_{ijkl} , and a constant variance. Finally, the f_{0l} and f_{pl} are random residual error terms at the province level, assumed to have a mean of zero and to be independent from the ε_{ijkl} as well as to have a constant variance.
The model is made multilevel by allowing the regression linear combination for each household to shift (higher or lower) from the overall linear combination by an amount $u_{0jkl} + v_{0kl} + f_{0l} + \varepsilon_{ijkl}$.
Our multilevel model in MLwiN output format looks like below:

$$\text{lrpcexp}_{ijkl} \sim N(XB, \Omega)$$

$$\begin{aligned}\text{lrpcexp}_{ijkl} = {} & \beta_{0ijkl}\text{cons} + \text{-0.039(0.009)female}_{ijkl} + \text{-0.441(0.011)children}_{ijkl} + \text{-0.054(0.011)elderly}_{ijkl} + \\ & \text{-0.081(0.001)hhsize}_{ijkl} + \beta_{5kl}\text{urban}_{jkl} + 0.074(0.007)\text{safewater}_{ijkl} + 0.287(0.007)\text{toiletflush}_{ijkl} + \\ & 0.168(0.011)\text{toiletsuilabh}_{ijkl} + 0.190(0.006)\text{housepermnt}_{ijkl} + \text{-0.173(0.005)housetem}_{ijkl} + \\ & 0.085(0.008)\text{electricity}_{ijkl} + 0.182(0.005)\text{tv}_{ijkl} + 4.667(0.927)\text{agerescale}_{ijkl} + \text{-41.759(8.886)agerescale2}_{ijkl} + \\ & 0.073(0.009)\text{kinh}_{ijkl} + 0.018(0.001)\text{yearsedu}_{ijkl} + 0.107(0.013)\text{leader}_{ijkl} + 0.147(0.016)\text{h\_skilled}_{ijkl} + \\ & 0.058(0.013)\text{m\_skilled}_{ijkl} + \text{-0.047(0.005)noskilled}_{ijkl} + 0.006(0.001)\text{urbyearsed}_{ijkl}\end{aligned}$$

$$\beta_{0ijkl} = 7.893(0.035) + f_{0l} + v_{0kl} + u_{0jkl} + e_{0ijkl}$$

$$\beta_{5kl} = 0.076(0.018) + f_{5l} + v_{5kl}$$

$$\begin{bmatrix} f_{0l} \\ f_{5l} \end{bmatrix} \sim N(0, \Omega_f) : \Omega_f = \begin{bmatrix} 0.031(0.006) & \\ \text{-0.002(0.003)} & 0.008(0.003) \end{bmatrix}$$

$$\begin{bmatrix} v_{0kl} \\ v_{5kl} \end{bmatrix} \sim N(0, \Omega_v) : \Omega_v = \begin{bmatrix} 0.015(0.001) & \\ \text{-0.002(0.002)} & 0.006(0.003) \end{bmatrix}$$

$$\begin{bmatrix} u_{0jkl} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} 0.012(0.001) \end{bmatrix}$$

$$\begin{bmatrix} e_{0ijkl} \end{bmatrix} \sim N(0, \Omega_e) : \Omega_e = \begin{bmatrix} 0.086(0.001) \end{bmatrix}$$

*-2*loglikelihood(IGLS Deviance)* = 14782.110(29530 of 29530 cases in use)

In our model, the 21 independent variables are as follows:

Location

urban	Urban=1, Rural=0

Household size and composition

hhsize	Household size
elderly	Proportion of elderly
children	Proportion of children
female	Proportion of females

Characteristics of head of households

Ethnicity of head of household, Kinh=1, otherwise=0

agerescale	Rescaled age of household head (=age/1000)
agerescale2	Squared rescaled age of household head
yearsedu	Number of year of education,

(primary basic=5 yrs, lower secondary basic=4yrs, upper secondary basic=3yrs, secondary professional and training=3yrs, college=3yrs, university=4.5yrs, PhD=4yrs)

urbyearsed	Interaction of urban and yearsedu
leader	Leadership job, yes=1, otherwise=0, reference= Non-skilled farm worker
h_skilled	High skilled job, university and above=1, otherwise=0, reference= Non-skilled farm worker
m_skilled	Medium skilled job, secondary professional and training=1, otherwise=0, reference= Non-skilled farm worker
noskilled	Non-skilled nonfarm worker, yes=1, otherwise=0, reference= Non-skilled farm worker

Housing

housepermnt	Having permanent house=1, reference=semi-permanent
housetem	Having temporary house=1, reference=semi-permanent
electricity	Having electricity=1
safewater	Having safe water source=1
toiletflush	Having flushing toilet=1, reference=other
toiletsuilab	Having suilabh toilet=1, reference=other
tv	Having a tv set=1, otherwise=0

The regression coefficient for the intercept of *lrpcexp* is β_0; the regression coefficients for *lrpcexp* are β_1 to β_{21} corresponding to the 21 independent variables shown above.
These coefficients go together with standard errors in brackets; all are significant. For example, for the variable "children" (the proportion of children in each household) the coefficient is -0.441 with its standard error of 0.011, which is significant. Some coefficients include a random component, namely the intercept β_{0ijkl} and the coefficient β_{5kl} of the urban/rural variable.
Since all predictors except for the urban/rural dummy variable are assigned only fixed effects, the slopes of the lines are all the same except for the urban/rural variable, but the intercepts are different for each commune, since we have assigned both fixed and random effects to the intercept. For example, for the variable "children", the fitted value of the fixed coefficient is -0.441 and its standard error is 0.011 (in bracket), as mentioned above. So for all communes the slope of the variable "children" is -0.441. The estimated fixed part of the intercept is 7.893, with a fitted standard error (in bracket) of 0.035. The intercepts for the different communes incorporate the fitted level 2 residuals $u_{0\,jkl}$ which are distributed around their mean with a variance of 0.012 (standard error 0.001). The intercepts for the different districts incorporate the fitted level 3 residuals v_{0kl} which are distributed around their mean with a variance of 0.015 (standard error 0.001). The intercepts for the different provinces incorporate the fitted level 4 residuals f_{0l} which are distributed around their mean with a variance of 0.031 (standard error 0.006).
This model has random effects at the district and provincial levels which are included in the coefficient of the "urban" dummy variable. The motivation for including those random effects is to attempt to capture unexplained geographical differences in the urban/rural gap (Haughton and Nguyen 2008), known to be important in Vietnam as a source of inequality. As can be seen from the model, β_{5kl} = 0.076(0.018) + f_{5l} + v_{5kl} where the fixed effect is 0.076 with standard error 0.018, the province-level random effect (province-level residual) f_{5l} has variance 0.008 with standard error 0.003, and the district-level random effect (or district-level residual) v_{5kl} has variance 0.006 with standard error 0.003. Note that the coefficient for the "urban" dummy variable does not include a commune-level random effect, since communes are either entirely rural or entirely urban.
Our model does not include variables at a higher level than the household level, for example Z_{jkl}. However, in our second model we will use some variables from the Viet Nam 2001 Agriculture Census available for all households in rural communes to build a small area model for rural areas using the VHLSS 2002.

Calculation and interpretation of commune-level, district-level and province-level random effects

In order to see whether some communes, districts and provinces had more 'effective' influence than others in promoting living standards of a household in them, we calculated commune-level random effects ($u_{0\,jkl}$), district-level random effects (v_{0kl}), and province-level random effects (f_{0l}) using MLwiN (use Residuals in Model) and the above model (1).

The results of the calculation can be plotted by MLwiN as below. Note that the graphs include random effects for both the intercept ('cons') and the urban/rural dummy variable ('urban'), with the exception of the commune level, where only the intercept includes a random effect. Note also that the random effects are plotted in increasing order, and with approximate 95% confidence intervals.

Figure 1. Commune-level random effects

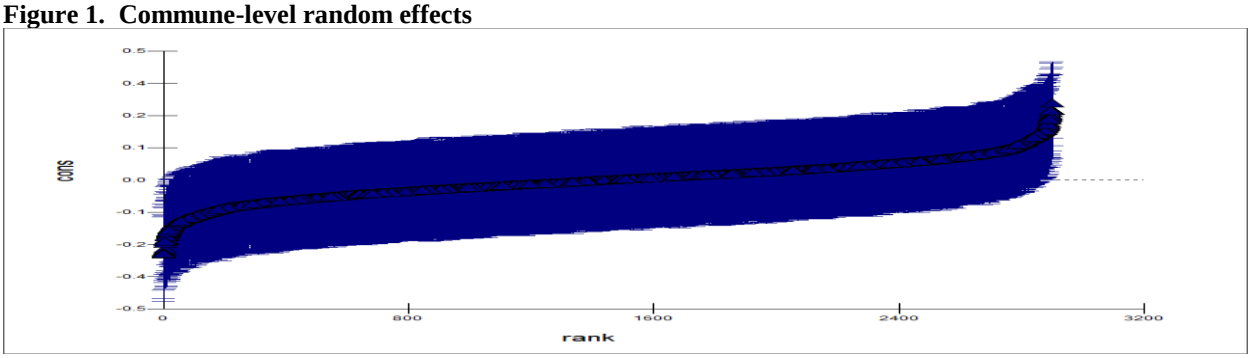


Figure 2. District-level random effects

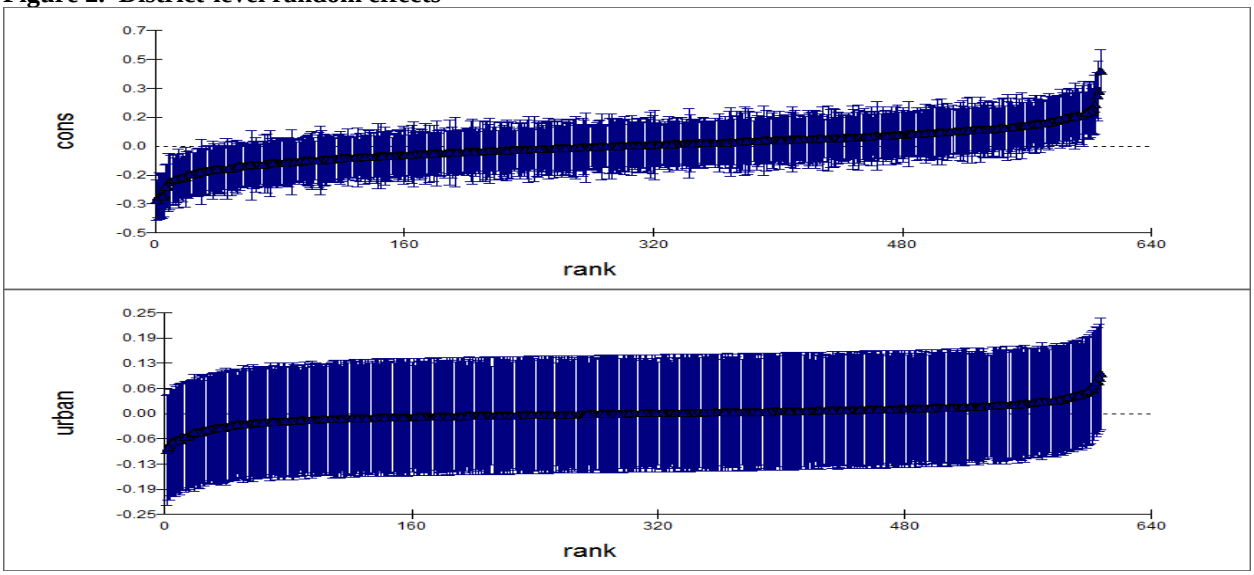
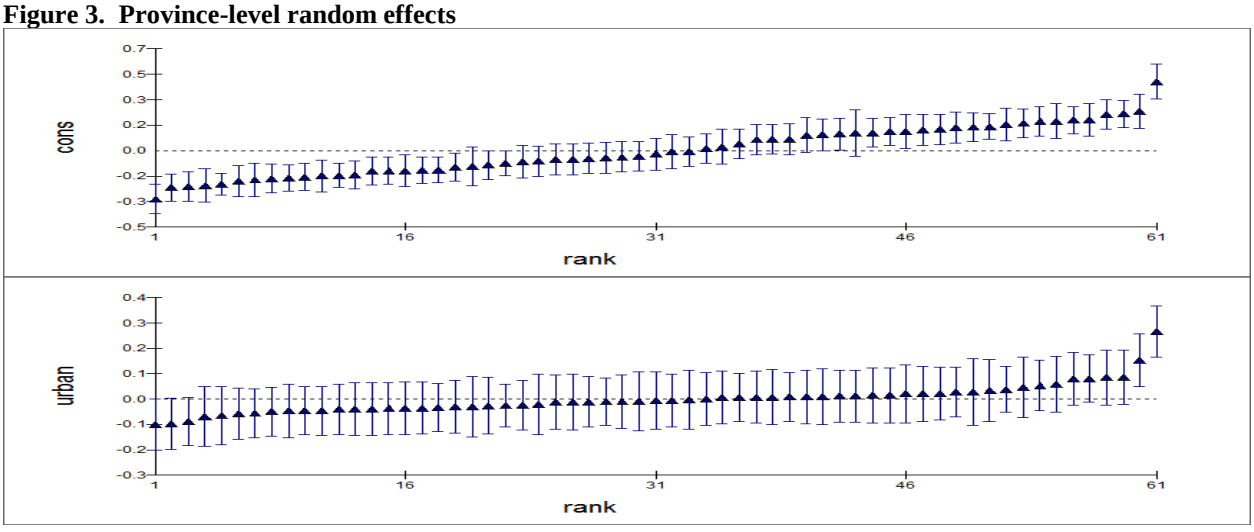


Figure 3. Province-level random effects



We have 2901 communes level 2 random effects plotted, one for each commune in the data set. These random effects represent commune departures from the overall line predicted by the fixed parameters. Looking at the confidence intervals around them, we can see that except for a small group of communes at each end of the plot where the confidence intervals for their random effects do not overlap zero, the majority of the communes do not differ significantly from the overall line at the 5% level. This means that except for a small number of communes, a

majority of the communes do not have ‘effective’ influence on household living standards. The same situation occurred for 607 district-level random effects. However, the variance for both commune and district random effects is significantly different from zero, as we have seen.
However, it is different for province random effects. We have 61 province level 4 random effects plotted, one for each province. There are 37 provinces of the plot where the confidence intervals for their random effects do not overlap zero; among them there 19 provinces have negative random effects and 18 provinces have positive random effects. This means that these provinces differ significantly from the overall line at the 5% level. In the Table xxx these provinces are marked red.
From the above results, we can say that while majority of communes and districts do not have an ‘effective’ influence on household living standards, 19 provinces exert a positive influence on household living standards, and 18 provinces exert a negative influence on household living standards.

Table 4. Province-level intercept random effects in increasing order

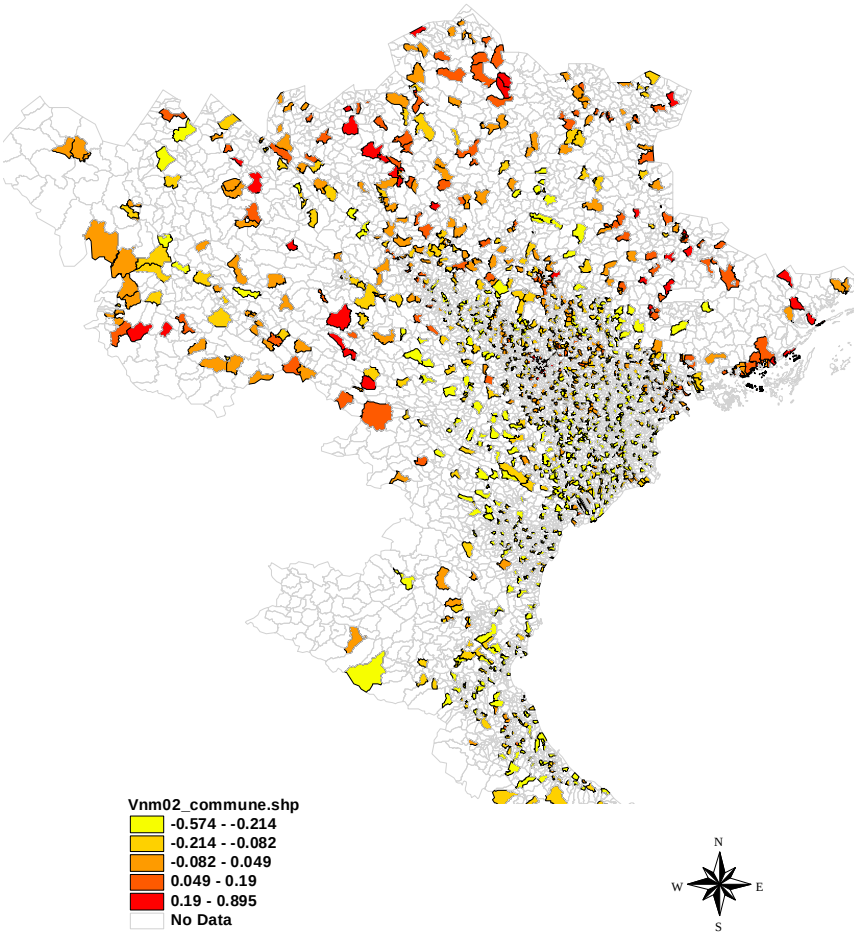
Order number	Province code	Province	Province-level random effects
1	115	Thái Bình	-0.25503
2	207	Bắc Kạn	-0.24824
3	305	Hòa Bình	-0.21035
4	401	Thanh Hóa	-0.20369
5	405	Hà Tĩnh	-0.18776
6	407	Quảng Bình	-0.18692
7	111	Hà Nam	-0.18651
8	403	Nghệ An	-0.18158
9	113	Nam Định	-0.1766
10	221	Bắc Giang	-0.15659
11	603	Gia Lai	-0.15363
12	505	Quảng Ngãi	-0.15263
13	117	Ninh Bình	-0.15111
14	107	Hải Dương	-0.12551
15	217	Phú Thọ	-0.12551
16	503	Quảng Nam	-0.11352
17	301	Lai Châu	-0.10848
18	104	Vĩnh Phúc	-0.10659
19	507	Bình Định	-0.09201
20	105	Hà Tây	-0.08662
21	109	Hưng Yên	-0.07764
22	605	Đắk Lắk	-0.07206
23	705	Ninh Thuận	-0.06928
24	607	Lâm Đồng	-0.06656
25	303	Sơn La	-0.04053
26	509	Phú Yên	-0.0379
27	411	Thừa Thiên - Huế	-0.03751
28	213	Yên Bái	-0.03668
29	205	Lào Cai	-0.03496
30	203	Cao Bằng	-0.0344

31	409	Quảng Trị	-0.03356
32	215	Thái Nguyên	-0.02441
33	211	Tuyên Quang	-0.02121
34	201	Hà Giang	-0.01602
35	106	Bắc Ninh	-0.01028
36	103	Hải Phòng	0.030395
37	511	Khánh Hòa	0.038571
38	209	Lạng Sơn	0.048548
39	809	Vĩnh Long	0.07146
40	501	Đà Nẵng	0.080296
41	817	Trà Vinh	0.090388
42	819	Sóc Trăng	0.098
43	225	Quảng Ninh	0.098579
44	803	Đồng Tháp	0.11016
45	807	Tiền Giang	0.11043
46	823	Cà Mau	0.11426
47	801	Long An	0.11433
48	709	Tây Ninh	0.12
49	601	Kon Tum	0.12138
50	715	Bình Thuận	0.15242
51	815	Cần Thơ	0.15415
52	707	Bình Phước	0.15759
53	811	Bến Tre	0.16233
54	711	Bình Dương	0.187
55	821	Bạc Liêu	0.19183
56	813	Kiên Giang	0.19261
57	101	Hà Nội	0.21304
58	717	Bà Rịa - Vũng Tàu	0.236
59	713	Đồng Nai	0.245
60	805	An Giang	0.26729
61	701	Hồ Chí Minh	0.41626

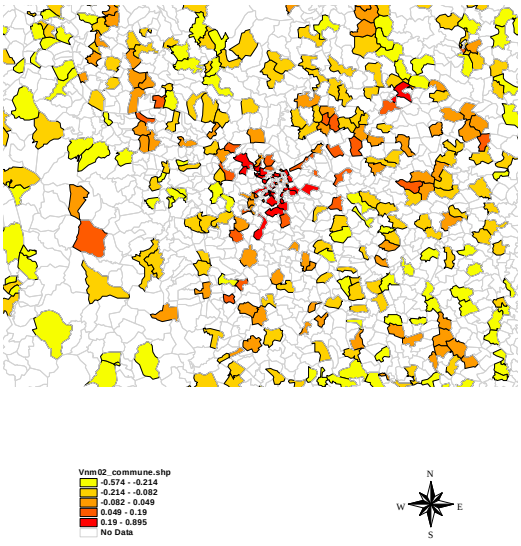
We will use visual tools to display random effects. An example of a display of total intercept random effects (province plus district plus commune-level effects) in the form of a map is given below in Figure xxx. Communes coloured red are communes whose location is associated with higher living standards, even once variables such as age, education and job status of the head of household are controlled for. On the other hand, communes coloured bright yellow are communes whose location is associated with lower living standards, controlling for those same variables. Such a display can be very useful to help identify communes in Vietnam that suffer challenges by their very location; these challenges could be due to isolation or particularly difficult climate conditions etc.

Figure 4. Total (province plus district plus commune) intercept random effects in our model for expenditure per capita

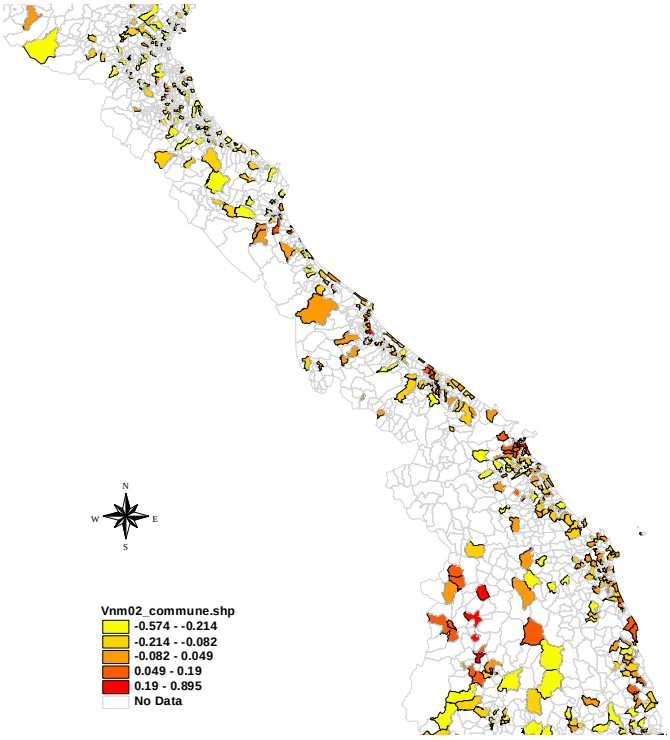
North



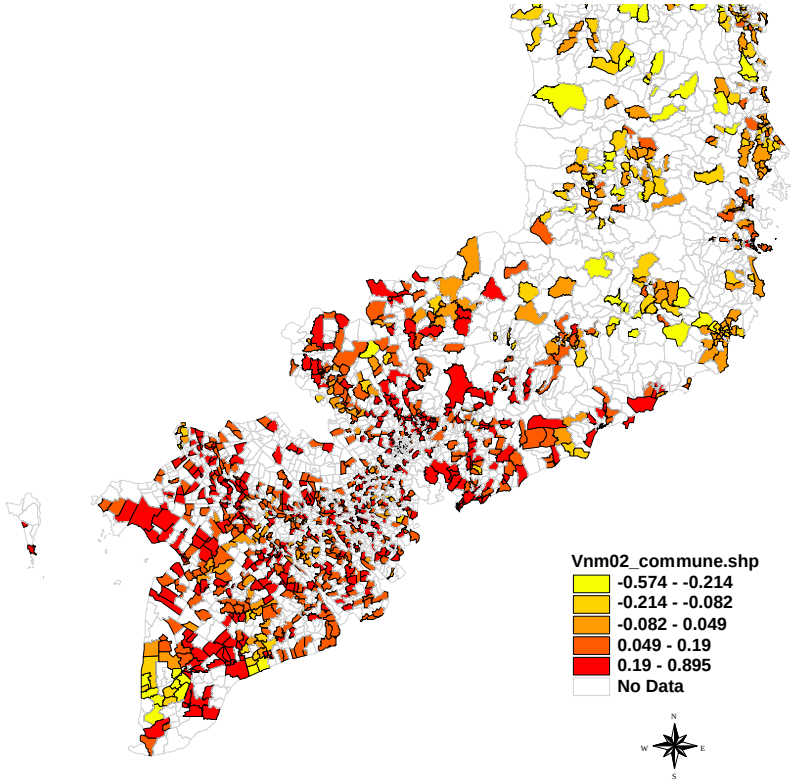
Hanoi area



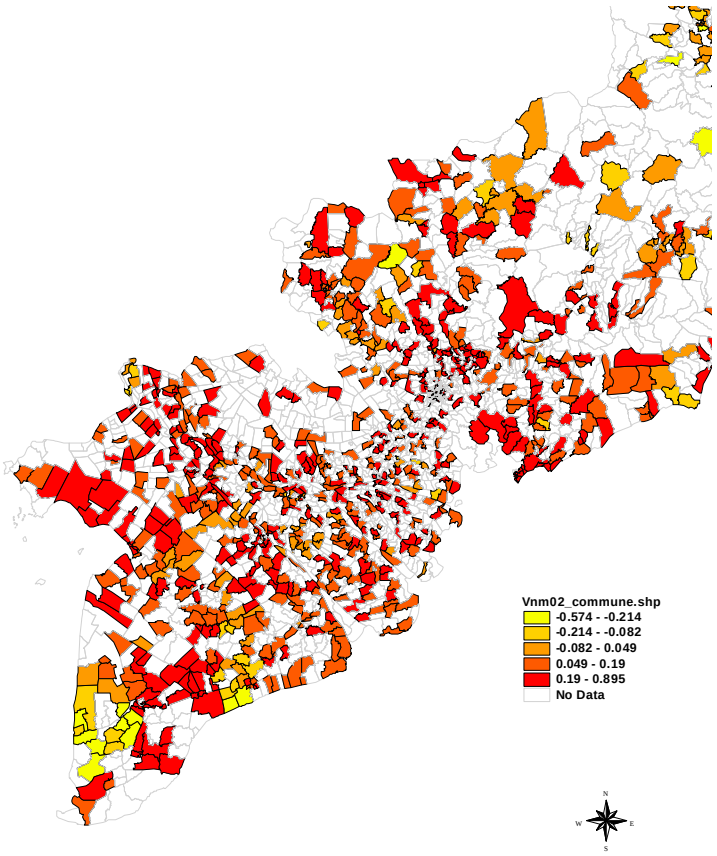
Center



South



Ho Chi Minh City area and Mekong Delta



Small area (commune) estimates

Here we use our multilevel model to obtain a predictor for the population mean of small areas, following ideas suggested by Moura (1994) and Moura and Holt (1999) briefly described below.
The generation of the random intercept model for predicting small area means is as follows:

$$Y_i = X_i B_i + \varepsilon_i \quad (8)$$

$$B_i = Z_i \gamma + \nu_i \quad (9)$$

for $i = 1, \ldots, m$, (m = number of small areas) where $Y_i = (y_{i1}, \ldots, y_{in_i})^T$ is the sample vector value of the characteristic of interest y (for example logarithm of expenditure per capita) for the i^{th} small area;

$$X_i = \begin{bmatrix} x_{0i1} \ldots\ldots x_{pil} \\ x_{0i2} \ldots\ldots x_{pi2} \\ . \\ . \\ . \\ x_{0in_i} \ldots\ldots x_{pin_i} \end{bmatrix}$$

is the n_i by $(p+1)$ matrix of explanatory variables at sample unit level for the i^{th} small area and x_{kij} is the value of the k^{th} auxiliary variable for the j^{th} sample unit in the i^{th} small area, $k = 0, \ldots, p$; also

$x_{0ir} = (1, \ldots, 1)$ for $i = 1, \ldots, m$, $r = 1 \ldots n_i$;

B_i = (B_{0i} , ... , B_{pi})^T is the (*p+1*) vector of the random coefficients for the i^{th} small area;

Z_i is the (*p+1*) by *q* (where *q* is the number of fixed parameters) design matrix of small area variables;

γ is a vector of length *q* of fixed coefficients;

ε_i = (ε_{il} , ... , ε_{in_i})^T is a random vector of length n_i for the i^{th} small area at sample unit level;

ν_i = (ν_{0i} , ... , ν_{pi})^T is a vector of length *p+1* of random effects for the i^{th} small area;

$n = \sum_{i=1}^{m} n_i$ is the total sample size.

It is assumed that ε_{ij} and ν_i are independent, $E_\varsigma \varepsilon_i = 0$, $Var_\varsigma \varepsilon_i = \sigma_\varepsilon^2 I$, where the symbol ς indicates that the expected value and the variance are calculated according to the random structure implied in equations 8 and 9. Then Moura suggested

$$\mu_i = \overline{X_i^T} Z_i \hat{\gamma} + \overline{X_i^T} \hat{\nu}_i \qquad (10)$$

as an appropriate predictor for the finite population mean $\overline{Y_i} = N_i^{-1} \sum_{j \in u_i} y_{ij}$ when N_i is large, where $\overline{X_i}$ = ($\overline{X_{0i}}$, .. , $\overline{X_{pi}}$)^T and $\overline{X_{ki}} = N_i^{-1} \sum_{j \in u_i} x_{kij}$ are the *population* means of the auxiliary variables, N_i is the population size of the i^{th} small area u_i , and $\hat{\gamma}$ and $\hat{\nu}_i$ are estimators for γ and ν_i obtained via Restricted Iterated Generalised Least Squares (RIGLS, Moura 1994, p. 39-40). The predictor in equation (10) is typically referred to as Empirical Best Linear Unbiased Predictor (EBLUP).

It follows from this framework that it is important to have at our disposal population means of the auxiliary variables in order to implement our small area estimation, as is of course common with many small area estimation methods. Moura (1994) demonstrates that this approach is optimal from the point of view of the mean squared error of the small area estimator, a measure of how much the small area estimators are expected to deviate from their true population values.

In our case from our equation (1) we have the following predictor for small area estimates

$$\mu_{ijkl} = \hat{\gamma}_{00} + \hat{\gamma}_{01} Z_{jkl} + \hat{f}_{0l} + \hat{\nu}_{0kl} + \hat{u}_{0jkl} + \sum_p (\hat{\gamma}_{p0} + \hat{\gamma}_{pl} Z_{jkl} + \hat{f}_{pl} + \hat{\nu}_{pkl} + \hat{u}_{pjkl}) \overline{X}_{pijkl}$$

where $\hat{\gamma}$ is a predictor for fixed effects; $\hat{f}$, $\hat{\nu}$, and $\hat{u}$ are predictors for province-level, district-level and commune-level random effects respectively obtained from our multilevel model (1) with data from VHLSS 2002 for the 21 variables listed. Using the ideas above we will plug the commune population means of the 21 variables from the Census conducted in 1999 into $\overline{X}_{pijkl}$ in our predictor equation (XXX) to estimate the mean logarithm of real per capita expenditure for VHLSS 2002 communes.

The results of small area estimates shows that the mean (over all communes) of small area estimates estimated with random effects (labeled lrpcexp99rnol: log of real per capita expenditure using the model and means from 1999 census *with* random effects) is a bit larger than that without random effects (labeled lrpcexp99nol: log of real per capita expenditure using the model and means from 1999 census *without* random effects). The mean of direct estimates (labeled lrpcexpvhlss02: log of real per capita expenditure from 2002 VHLSS) is larger than small area estimates *with or without* random effects.

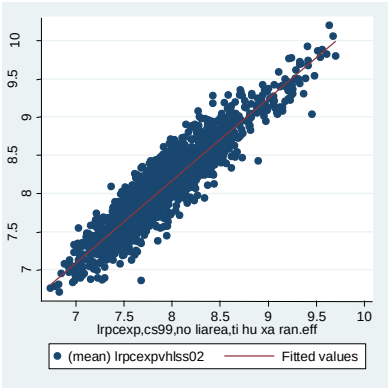
Note that we had to exclude the living area of a household as a predictor from the model because of discrepancies in the measurement of this variable between VHLSS 02 and the 1999 Census.

Variable	Obs	Mean	Std. Dev.	Min	Max
lrpcexp99rnol	2626	7.839418	.4110265	6.732353	9.699497
lrpcexp99nol	2626	7.832879	.3499648	6.784832	9.024239
lrpcexpvhlss02	2626	7.993064	.4765603	6.70638	10.19934

On the other hand, the command below indicates that for our model with random effects (and no living area), 466 communes have their direct estimate smaller than their small area estimate, while we might expect about half the communes to have their direct estimate smaller than their small area estimate. These differences could be due to changes in some independent variables between 1999 and 2002. Of course poverty mapping methods (which typically do not include random effects) encounter the same problem. However our results show that including random effects does improve the small area estimation, as we will see in see graph below.

```
.count if  lrpcexpvhlss02 <lrpcexp99rnol
  466
```

The graph proposed by Brown, Chambers, Heady and Heasman (Evaluation of small area estimation methods, Proceedings of Statistics Canada Symposium 2001) as a tool for checking model validity for a small area estimation shows that the estimates with random effects are closer to the least squares fit line, which is itself close to the 45-degree line:

Direct estimates vs. small area estimates, *with* province, district and commune random effects	Direct estimates vs. small area estimates, *without* province, district and commune random effects
	

Note: Direct estimates are on Y axis, small area estimates are on the X axis.

The idea of this diagnostic graph is that if the small area estimates are a good representation of the "truth" – the population means, the direct survey observations should behave as a random sample from a distribution with mean equal to the population means.

It is interesting that the inclusion of random effects clearly improves the small area estimation according to this diagnostic tool.

As can be seen above, there are 36 provinces that have random effects statistically different from 0. It means that province-level random effects play an important role in raising or lowering commune living standards. This conclusion can be supported by the following graphs which show that small area estimates without province random effects are worse:

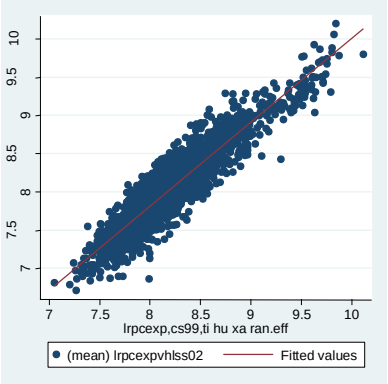
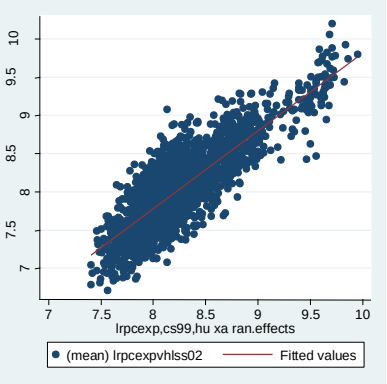
Direct estimates vs. small area estimates, *with* province, district and commune random effects	Direct estimates vs. small area estimates, *with* district and commune random effects, but NOT province random effects
	

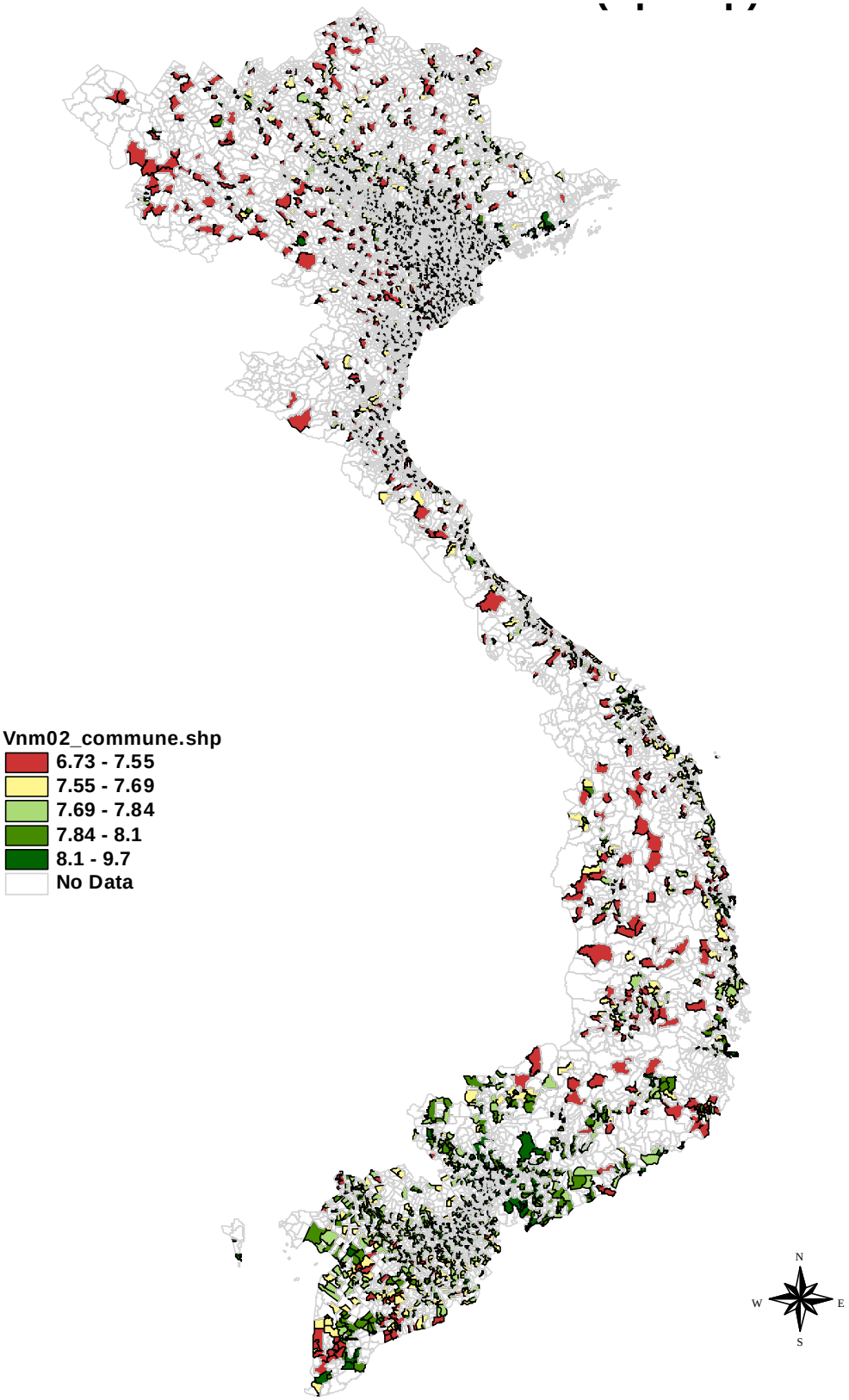
Table 5. Means for each province of direct and small area estimates with/out intercept random effects

Order number	Province code	Province name	Mean of log of per capita expenditure 2002 survey commune averages (Direct estimates)	Mean of small area estimates of log of per capita expenditure *without* random effects	Mean of small area estimates of log of per capita expenditure *with* random effects (sorted)	Difference
A	B	C	1	2	3	4=2-3
1	301	Laichau	7.444654	7.381192	7.318338	0.062854
2	603	Gialai	7.599549	7.634328	7.47031	0.164018
3	303	Sonla	7.620348	7.444576	7.470324	-0.02575
4	305	Hoabinh	7.631957	7.688017	7.504684	0.183333
5	201	Hagiang	7.633948	7.435541	7.526372	-0.09083
6	207	Backan	7.576459	7.625515	7.540832	0.084683
7	605	Daklak	7.676263	7.70176	7.603253	0.098507
8	401	Thanhhoa	7.753994	7.781063	7.610376	0.170687
9	705	Ninhthuan	7.836021	7.754676	7.613585	0.141091
10	205	Laocai	7.772395	7.650525	7.648156	0.002369
11	503	Quangnam	7.820565	7.761992	7.652136	0.109856
12	405	Hatinh	7.726047	7.774341	7.652552	0.121789
13	409	Quangtri	7.777928	7.748946	7.656882	0.092064
14	505	Quangngai	7.765183	7.711431	7.660374	0.051057
15	407	Quangbinh	7.829987	7.780575	7.663468	0.117107
16	115	Thaibinh	7.805032	7.933037	7.666295	0.266742
17	403	Nghean	7.793827	7.798377	7.685744	0.112633
18	211	Tuyenquang	7.811045	7.644742	7.700823	-0.05608
19	111	Hanam	7.877288	7.867786	7.704787	0.162999
20	221	Bacgiang	7.860664	7.785176	7.713243	0.071933
21	117	Ninhbinh	7.907673	7.858122	7.714227	0.143895
22	819	Soctrang	7.908699	7.656679	7.715765	-0.05909
23	104	Vinhphuc	7.827079	7.793365	7.723116	0.070249

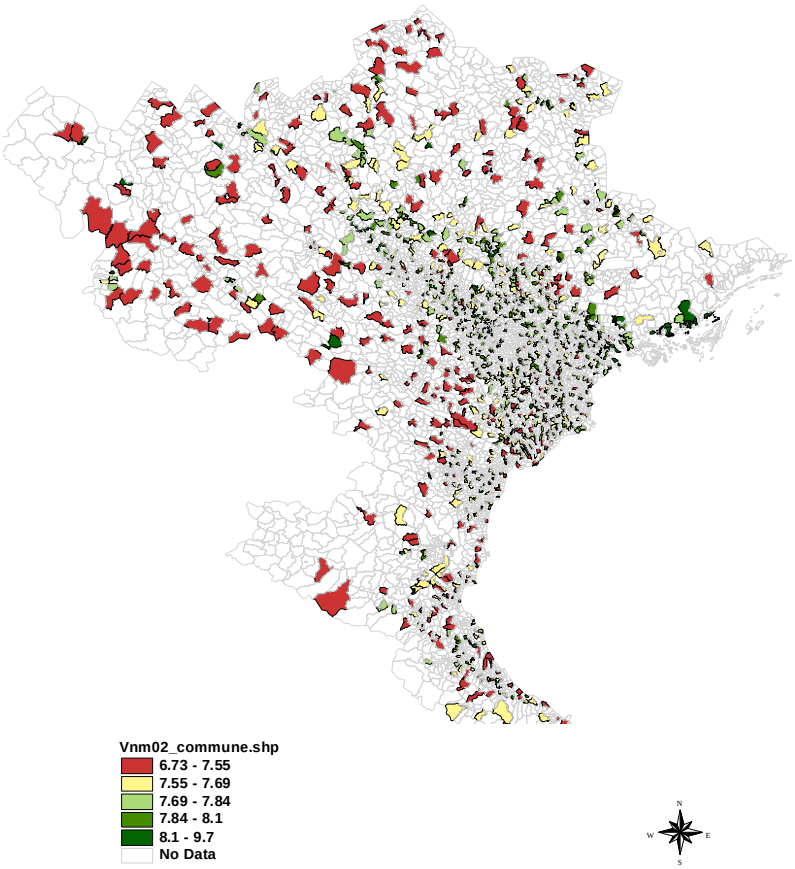
24	601	Kontum	7.798513	7.641468	7.73148	-0.09001
25	509	Phuyen	7.983709	7.790438	7.743752	0.046686
26	113	Namdinh	7.930728	7.933326	7.744675	0.188651
27	507	Binhdinh	7.947245	7.836394	7.749632	0.086762
28	817	Travinh	7.872901	7.674992	7.753805	-0.07881
29	217	Phutho	7.815461	7.818558	7.759162	0.059396
30	803	Dongthap	7.83929	7.717499	7.764334	-0.04683
31	203	Caobang	7.790983	7.668563	7.773662	-0.1051
32	213	Yenbai	7.854775	7.799317	7.775157	0.02416
33	209	Langson	7.902244	7.655753	7.785496	-0.12974
34	109	Hungyen	7.926977	7.850253	7.803892	0.046361
35	411	Hue	8.033307	7.860134	7.804657	0.055477
36	105	Hatay	7.917986	7.862491	7.809808	0.052683
37	823	Camau	8.000431	7.748756	7.818316	-0.06956
38	821	Baclieu	8.041788	7.776587	7.825221	-0.04863
39	809	Vinhlong	8.018991	7.805455	7.834547	-0.02909
40	707	Binhphuoc	7.984639	7.728679	7.844825	-0.11615
41	107	Haiduong	8.004038	7.942685	7.846064	0.096621
42	607	Lamdong	7.938713	7.926197	7.847244	0.078953
43	709	Tayninh	8.031383	7.755292	7.871876	-0.11658
44	815	Cantho	8.060887	7.799299	7.872496	-0.0732
45	813	Kiengiang	8.101149	7.748258	7.875287	-0.12703
46	215	Thainguyen	8.016177	7.876066	7.886951	-0.01089
47	801	Longan	8.116645	7.802288	7.908109	-0.10582
48	807	Tiengiang	8.108866	7.791949	7.913351	-0.1214
49	106	Bacninh	8.06555	7.869784	7.916667	-0.04688
50	811	Bentre	8.067037	7.790483	7.929554	-0.13907
51	805	Angiang	8.127779	7.731284	7.930951	-0.19967
52	715	Binhthuan	8.166921	7.846275	7.980414	-0.13414
53	511	Khanhhoa	8.290757	8.04306	8.084421	-0.04136
54	103	Haiphong	8.261419	8.152914	8.098542	0.054372
55	225	Quangninh	8.319826	8.153591	8.159599	-0.00601
56	713	Dongnai	8.303933	7.975947	8.162469	-0.18652
57	711	Binhduong	8.404768	8.002626	8.2586	-0.25597
58	717	Baria - Vungtau	8.499705	8.067645	8.279455	-0.21181
59	501	Danang	8.546833	8.296317	8.388885	-0.09257
60	101	Hanoi	8.951884	8.492667	8.782633	-0.28997
61	701	TP Ho Chi Minh	8.997186	8.348116	8.807019	-0.4589

A Geographical Information Systems (GIS) representation of our small area estimates is useful for presentation purposes and to help identifying communes with lower living standards (inclusive of contributions due to lower or higher values of predictors).

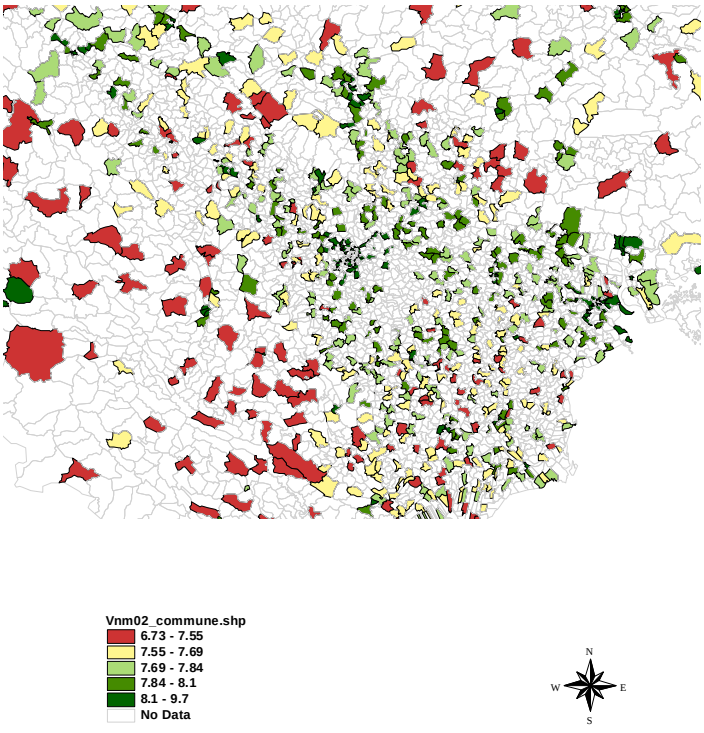
Small Area Estimates: whole country



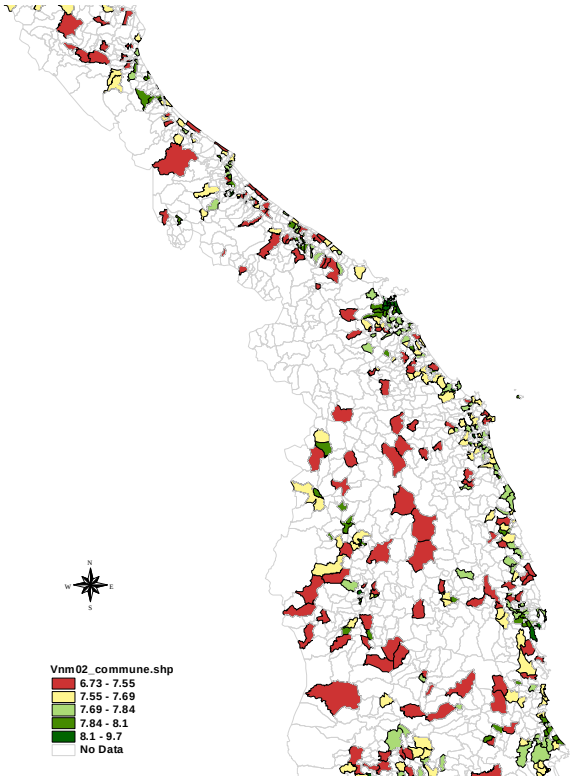
North



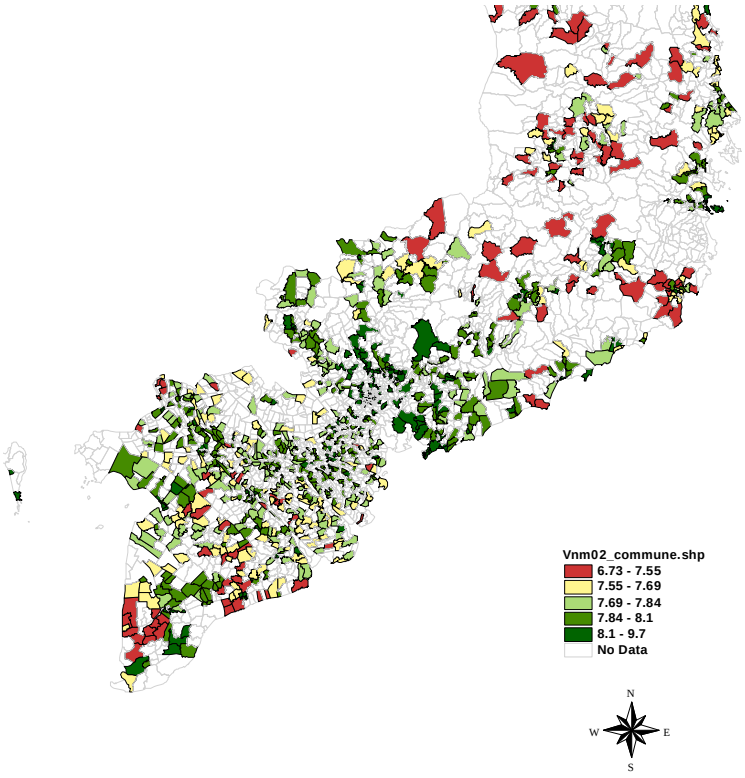
Hanoi area



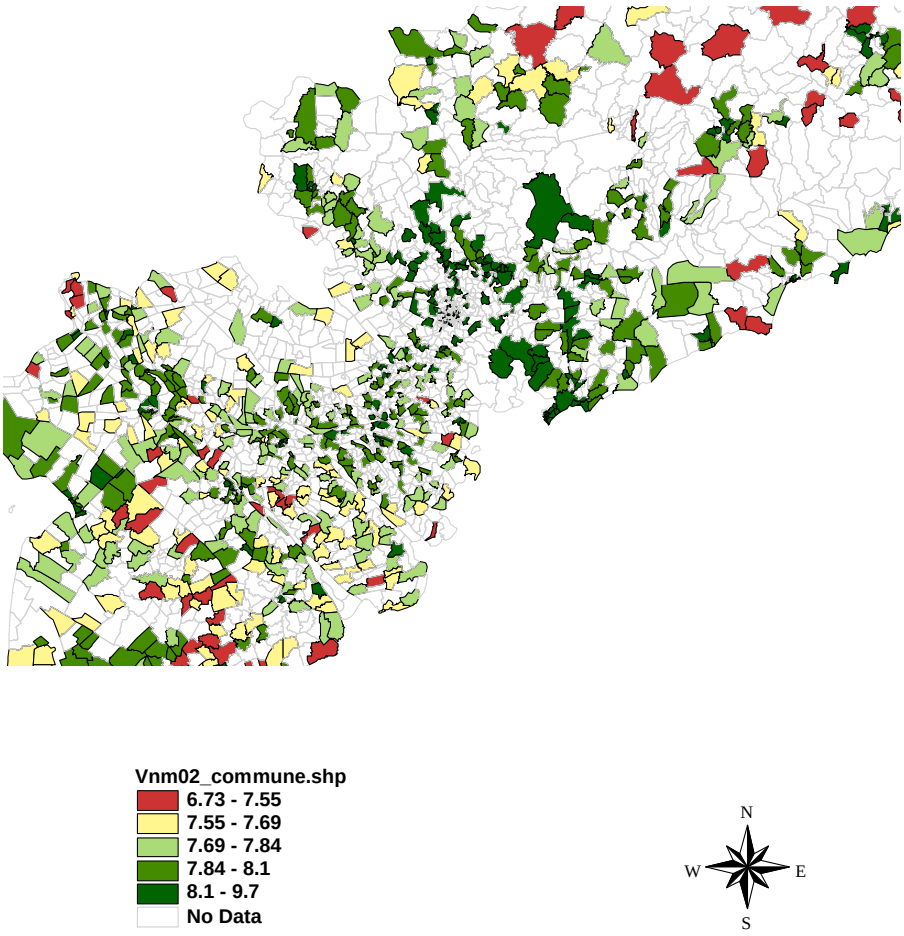
Center



South



Ho Chi Minh City area



References

Australian Bureau of Statistics (2005), *A Guide to Small Area Estimation* (http://www.nss.gov.au/nss/home.NSF/pages/Small+Areas+Estimates?OpenDocument)

Goldstein,H. (1989). Restricted unbiased iterative generalised least squares estimation. Biometrika, 76, 622-23.

Goldstein, H., 1995. *Multilevel statistical models. London* [Online]. Available at: http://www.ats.ucla.edu/stat/examples/msm_goldstein/goldstein.pdf [accessed: 24 June 2008]

Haughton, D. & Nguyen, P., 2008. Multilevel models and inequality in Vietnam, to appear, *Journal of Data Science*.

Moura, F.A.S. &, Holt, D., 1999. Small area estimation using multilevel models. *Survey Methodology*, 25(1), p.73-80.

Moura, F.A.S, 1994. Small area estimation using multilevel models. PhD thesis. University of Southampton.

Appendix: Brief description of the IGLS (Iterated Generalised Least Squares) algorithm

Here we briefly describe how the IGLS algorithm first estimates the fixed parameters and the variances and covariances of the random effects, and then, assuming those known, estimates the residuals (or random effects) in a multilevel model. To simplify the presentation we will use a model with only two levels, with random effects only in the intercept, and refer to Chapter 2 of Goldstein (1995) for more details on this particular case, and to Appendix 2.2 of Goldstein (1995) for a brief description of the estimation method of the residuals in the more general case (of more than two levels).

The basic two-level regression model can be stated as follows:

$$Y_{ij} = \beta_{0j} + \beta_1 X_{ij} + e_{0ij} \quad (1)$$

$$\beta_{0j} = \beta_0 + u_{0j} \quad (2)$$

with the u_{0j} distributed according to a normal distribution with mean 0 and variance σ^2_{u0}, and the e_{0ij} distributed independently of the u_{0j} according to a normal distribution with mean 0 and variance σ^2_{e0}. Note that the variance of the Y_{ij} is then $\sigma^2_{u0} + \sigma^2_{e0}$.

Essentially the idea of the algorithm is to first assume the components of variance σ^2_{u0} and σ^2_{e0} known, and then use generalized least squares to estimate the fixed parameters β_0 and β_1. The error covariance matrix used in the generalised least squares process is a block-diagonal matrix with blocks of size n_j, the number of level one observations in each level two set; each block has $\sigma^2_{u0} + \sigma^2_{e0}$ on the diagonal and σ^2_{u0} on the off-diagonal. This process yields estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ for β_0 and β_1.

From these estimates for the fixed parameters, we can estimate the components of variance as follows. We first construct residuals referred to as "raw residuals" by the equation $\tilde{Y}_{ij} = Y_{ij} - \hat{\beta}_0 - \hat{\beta}_1 X_{ij}$. We note that the expected value of the matrix of the cross products of the vectors $\tilde{Y}_{ij}$ and $\tilde{Y}_{i'j}$ is in fact the same as the covariance matrix of the errors in the model defined by (1) and (2). That implies that if we place all the elements of the cross-product matrix into one vector and consider this vector as the left hand side in a regression equation, the right hand side of that equation will include σ^2_{u0} and σ^2_{e0} as unknown coefficients and independent variable vectors consisting of only ones for σ^2_{u0} and ones and zeros for σ^2_{e0}. From the components of variance estimated from the previous iteration, we can obtain the covariance matrix of the vector of cross products, which in turns allows to apply generalised least squares to obtain estimates for the new components of variance. Once these components of variances are estimated, we can then return to the previous process to get new estimates for β_0 and β_1, use these estimates to get new estimates of components of variance etc, until the procedure converges.

To begin the iterative process, one can use ordinary least squares estimates as initial values for $\hat{\beta}_0$ and $\hat{\beta}_1$ and a diagonal covariance matrix as initial covariance matrix for the second step.

Assuming normality, the IGLS procedure we have described corresponds to a maximum likelihood estimation, which produces biased estimates of the components of variance since it ignores the random nature of the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ (Goldstein, 1995, chapter 2). The algorithm RIGLS (for Restricted IGLS) can produce unbiased estimates and is described in Goldstein (1989).

Once the fixed parameters and components of variance are estimated, to estimate the random effects themselves involves a maximum likelihood estimation for each *j* of the u_{0j} and of the e_{0ij} on the basis of the equation $\tilde{Y}_{ij} = u_{0j} + e_{0ij}$. Once can show that the estimates $\hat{u}_{0j}$ equal

$\hat{u}_{0j} = (n_j \sigma^2_{u0} / (n_j \sigma^2_{u0} + \sigma^2_{e0}))\tilde{Y}_j$, where $\tilde{Y}_j = \sum_i \tilde{Y}_{ij} / n_j$ and it then follows that $\hat{e}_{0ij} = \tilde{y}_{ij} - \hat{u}_{0j}$ (equations 2.14 in Golstein 1995, chapter 2).