A Longitudinal Model for Repeated Cross Sectional Data with Clustering

Albert Satorra* Peter M. Bentler †

Abstract

Cross-sectional data based on independent repeated samples do not obviously lend themselves to longitudinal modeling. We show that when data is clustered (e.g., repeated measures within individuals, patients nested in clinics, students nested in schools, respondents nested in areas) with invariant second-level units, longitudinal modeling at the second level is possible and meaningful. Specifically, we consider a factor analysis model with autoregressive factors whose measurement structure varies by first-level sample size m, being a 1-factor model when m is very large but a 2nd-order factor model otherwise. We consider issues of consistency, bias, alternative estimators, chi-square model tests, and factor score estimation, as well as model misspecification and effects of intraclass correlation. Among other things, we note that Bartletts factor score estimates require modification for 2nd-order factor models. The models are implemented using standard software, evaluated via simulations, and illustrated with repeated survey data on information and communication technology.

Key Words: Longitudinal data, structural equation models, factor analysis, factor score estimation, consistent estimation, small area estimation

1. Introduction

Sampling over time enables researchers to model the evolution (the change) of economic, social or behavioral processes in populations (e.g., Duncan and Kalton, 1987). When (part of the) individuals surveyed overlap during several periods, as in rotating panels, panel data methods are typically used. A vast literature on panel data methods is available (see Binder (1998); Kalton and Citro (1993); Kasprzyk, Duncan, Kalton and Singh(1989)). In contrast to panel surveys, we address issues to the case where the individuals surveyed are different in each time occasion of the survey [1], so we can not use the classical methods of panel data analysis. To be able to do longitudinal analysis, we assume hierarchical two -level grouped data where first-level units are nested in second-level units. This is the case for example, in a multi-stage survey, where primary sample units (PSU) are nested in secondary sample units (SSU), also in small areas surveys, where the individuals are nested in small geographical areas (see Ghosh and Rao (1994) and Pfeffermann (2013) for a review of small-area analysis). In contrast with other work in repeated surveys (e.g., Binder, and Dick (1989), Blight and Scott (1973), and Fuller (1990), Holt and Skinner, (1983), Jones (1980)) we consider the general case where the key variables of the longitudinal model are latent, observable just through multiple indicators. We believe this a general set-up with potential multiple applications. We give an example of application to a longitudinal analysis of household information and communication technology (ICT) for repeated surveys in a region of Barcelona (Spain). The methods presented will address both the estimation of structural parameters of a longitudinal model, as well the estimation of the levels of latent variables in the secondary sample units. The second aspect makes the work overlap with small area estimation methods (see the recent survey of Pfeffermann, 2013) which we

*Ramon Trias Fargas, 25-27, Universitat Pompeu Fabra, 08005 Barcelona, Spain

†Box 951563, UCLA, Los Angeles, CA 90095-1563

[1] The methods developed can easily be modified to cope with partial overlap of the cases in repeated surveys

frame in a general context where we have multiple indicator models and repeated surveys. The basic models to be proposed for the analysis turn out to be special models for structural equation modeling (SEM) and thus can be analyzed using standard software widely available to researchers.

Section 2 describes the data design and models to be used for the analysis; Section 3 investigates the performance of the models proposed in estimating structural parameters of the longitudinal model; Section 4 does the same for small area estimation of latent characteristics of interested (factor score estimation); Section 5 discusses an application with empirical data; Section 6 concludes.

2. Data structure and models

Consider a set of k variables $\{Y_j,\, j = 1, \ldots, k\}$ and repeated surveys $\{\mathcal{S}_t, t = 1, \ldots, T\}$ in a stratified/clustered $\mathcal{C} = \{\mathcal{C}_g,\, g = 1, \ldots, G\}$ population $\mathcal{P}$, where $\mathcal{P} = \cup_{g=1}^{G} \mathcal{C}_g$ and $\mathcal{C}_g \cap \mathcal{C}_{g'} = \emptyset$ when $g \neq g'$. Thus, we have data $\{Y_{ijgt}\}$ with Y_{ijgt} being the value of variable Y_j on individual i of cluster g for the tth survey. Note that notation Y_{ijgt} is loose in the sense that individual i is different in the repeated (varying t) surveys. One example of this set-up is repeated surveys on several variables of individuals of a population that is stratified in small areas (the gs). In section below we consider the example where we have repeated surveys on household usage of information and communication technology (ICT), where each household responds to ICT usage variables $Y_j, j = 1, \ldots, k$, and the aim is to assess the variation across time of the ICT level of the households at a small area level (in the illustration, the population is clustered in 41 small areas). Assume the survey is repeated with independent cross-sections along several years. In this paper we address the issue of longitudinal analysis for the areas as well as the estimation of the levels of ICT usage for the different areas. We will use the ICT example to make the discussion of the model more concrete, but other examples could have been used, e.g. the area (second level units) could be a hospital (or school, or teacher, etc.) and the first level units could be patients (or students).

We assume the following model for first-level unit data

$$
\begin{aligned}
y_{gti} &= \Lambda F_{gti} + \epsilon_{gti} && (1)\\
F_{gti} &= f_{gt} + w_{gti} && (2)\\
f_{g(t+1)} &= \rho f_{gt} + v_t, \quad t = 1, \ldots, T-1 && (3)
\end{aligned}
$$

where f_{tg} and F_{tgi} are first- and second-order scalar factors (latent variables) respectively, ϵ_{gti} $(k \times 1)$, w_{gti} and v_t are centered (vector/scalar) variables. In the ICT example, f_{gt} corresponds to the true (unobservable) ICT level of the area g at time t, while F_{gti} correspond to the true ICT level for individual i interviewed at time t in area g. Assume f_{gt}, ϵ_{igt}, w_{gti} and v_t are i.i.d. (in their subindexes) mutually independent random variables. No distribution is specified, except for existence of fourth-order moments of all the variables involved. We use the terminology of areas (or small areas) but the approach does not preclude that the areas could be teachers, schools or other secondary units of analysis.

The data can be viewed as two-level data. Figure 1 gives a schematic representation of the type of data considered, where we see the two-level structure of first- and second-level units, the time dimension, as well as the multivariate dimension (of the multiple indicator model), with independent samples across time.

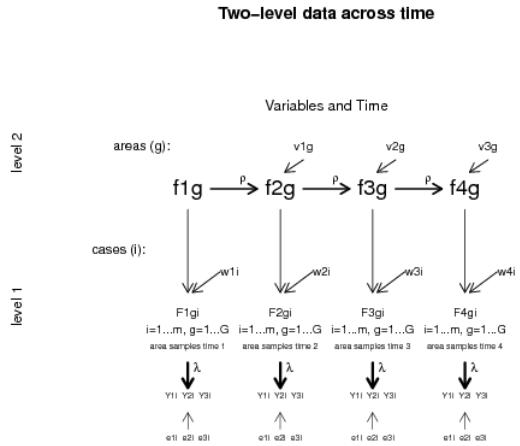

Figure 1: Two level repeated cross-section data

The between model: LFMw

We can reduce the data to a time series single-level data by aggregating (taking the mean) the variables within each area g obtaining what we can call the between (area) model. Averaging y_{gti} across i within each area and time, (1) to (3) transforms to

$$y_{gt.} = \Lambda F_{gt.} + \epsilon_{gt.} \quad (4)$$

$$F_{gt.} = f_{gt} + w_{gt.} \quad (5)$$

$$f_{j(t+1)} = \rho f_{jt} + v_t, \quad t = 1, \ldots T-1 \quad (6)$$

where $y_{gt.}$, $F_{gt.}$, $\epsilon_{gt.}$ and $w_{gt.}$ are mean variables. Note that the autoregressive equation (3) is unmodified by averaging within each area (it is the same as (6). In the structural equation model perspective, the between model is a second-order factor model where F_{gt} and f_{gt} are first- and second-order factors, respectively, and there is an autoregressive equation for the second-order factor. The model is depicted as a path diagram in Figure 2 (where for the sake of simplicity, only three indicators for each factor are considered). Since this is longitudinal factor model with first and second order factors, implied by the presence of w, the model is noted as LMw.

A simplification of this model arises when $w = 0$ at each time point (i.e. the variance of w is set to zero) The path diagram representation of this model is the same as in Figure 2 but now with the w's suppressed and thus the fs are equated to the Fs. This is now a simple first-order factor model. This model is likely to apply when the sample size within each area g, say m_g, is large, since in that case the variance of $w_{gt.}$ which involves the factor $1/m_g$ is likely to be small. That is, when sample size within each second sample unit is fairly large, then the model LFM may be a valid approximative model. A case where undoubtedly LFM is the correct model is in the case of large sample size in each area, for example in repeated surveys across countries, where the sample size is typically fairly large. Then one just uses the approximative model LFM as the model for longitudinal analysis. In the empirical example below, we will use the fit of model LFMw to verify that a valid approximative model for that data is in fact LFM.

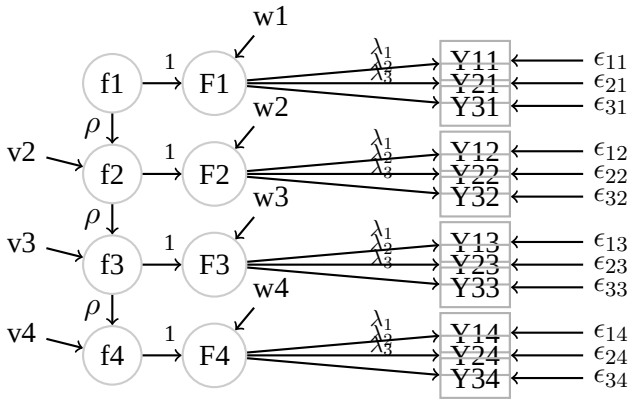

Figure 2: LFMw: second-order factor model

SEM representation of LFMw

A structural equation model (SEM) representation of the model can be obtained as follows. We stack vectors along $t = 1, \ldots, T$: $y = (y'_{g1.}, \ldots, y'_{gT.})'$, $F = (F'_{g1.}, \ldots, F'_{gT.})'$, $e = (e'_{g1.}, \ldots, e'_{gT.})'$, $w = (w'_{g1.}, \ldots, w'_{gT.})'$, $f = (f'_{g1}, \ldots, f'_{gT})'$, so that

$$\begin{aligned} y &= \Lambda F + e \\ F &= f + w \\ f &= Bf + v \end{aligned} \tag{7}$$

$v' = (f_1, v_2, \ldots, v_t)$, e and w are mutually independent with covariance matrices Φ_v, Φ_e and Φ_w. With this definition, the $(T.k) \times (T.k)$ variance matrix $\Sigma_y = \text{var}(y)$ has the moment structure

$$\Sigma_y = \Lambda[(I - B)^{-1}\Phi_v(I - B')^{-1} + \Phi_w]\Lambda' + \Phi_\epsilon.$$

That is, $\Sigma_y = \Sigma(\theta)$, where θ a vector of unconstrained parameters to be estimated that comprises the free elements of the parameter matrices involved in the moment structure. For the sake of simplicity of exposition, assume that $\Phi_w = \phi_w I_T$. This restriction can be amended when needed by proper specification of the diagonal terms in matrix Φ_w (when sample size varies highly across clusters, then we can specify linear equalities of the terms of the diagonal of Φ_w involving the factors of area sample size m_g).

For the example of LFMw in Figure 2, the parameter matrices associated to the SEM approach are as follows (note that $T = 4$ and $k = 3$)

$$\Lambda = \begin{pmatrix} 1 & 0 & 0 & 0 \\ \lambda_1 & 0 & 0 & 0 \\ \lambda_2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & \lambda_1 & 0 \\ 0 & 0 & \lambda_2 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & \lambda_1 \\ 0 & 0 & 0 & \lambda_2 \end{pmatrix}$$

$$B = \begin{pmatrix} 0 & 0 & 0 & 0 \\ \rho & 0 & 0 & 0 \\ 0 & \rho & 0 & 0 \\ 0 & 0 & \rho & 0 \end{pmatrix}$$

Vector of free parameters:

$$\theta' = (\rho, \lambda_2, \lambda_3, \phi_{e1}, \phi_{e2}, \phi_{e3}, \phi_{e4}, \phi_w, \phi_{f_1}, \phi_{v2}, \phi_{v3}, \phi_{v4})$$

The variance of w is restricted by equality across times.

We distinguish the two models

LFM: $\phi_w = 0$ ($F_{gti} = f_{gti}$). This is a first-order factor model.

LFMw: ϕ_w free parameter ($F_{gti} \neq f_{gti}$). This is a second-order factor model.

The SEM approach allows estimation and inferences for the **structural parameters** (e.g., ρ, λ, ...) of both LFM and LFMw, with estimates being consistent, asymptotically normal, etc. Estimation and inferences will be carried using standard software for SEM analysis. In our illustration, the moment structure is to be fitted to S by EQS (Bentler, 2001-2010), with ML or LS estimation. We use normal theory and distribution free (robust) s.e. and chi-square goodness of fit test. Other SEM software, e.g., LISREL, Mplus, CALIS, `sem` of Stata, AMOS, etc. could have been used. The methods used yield inferences (asymptotically) free from distributional assumptions (Satorra and Bentler, 1990, 1994). Robust goodness of fit test of the model are also provided.

We are interested also on **factor score** estimates for latent variables at the area (second-level units) level. In the case of LFM, standard Bartlett's and Thomson's factor scores estimation are readily available in SEM software. In the case of LFMw, we need to develop specific Bartlett's and Thomson's factor scores, since we are dealing with a second-order factor model.

ML fitting function is:[2]

$$F(S, \Sigma(\theta)) = \ln \mid \Sigma(\theta) S^{-1} \mid + \text{tr}\{S\Sigma(\theta)^{-1}\} - p$$

Σ is the covariance structure for either LFM or LFMw, S is the sample covariance matrix of the second-level unit multivariate vector y. The "sample size" is G. Chi-square goodness of fit test of the model, testing equality of ρ across time, invariance of λ's, etc. are available.[3]

3. Illustration with simulated data

This section uses simulated data to assess the performance of the proposed models LFMw and LFM in analyzing repeated samples with clustering. Two - level data is generated from the following model equations:

$$\begin{aligned} Y_{jtgi} &= \lambda_j F_{tgi} + \epsilon_{jtgi}, \\ F_{tgi} &= f_{tg} + w_{tgi} \end{aligned}$$

where $\lambda_j = \lambda = (1, .6, .1)$, $\psi_\epsilon = (I_3 - \text{diag}\,(\lambda * \lambda)) + \text{diag}\,(0.2, 0.2, 0.2)$; $j = 1, \ldots, 3$; $t = 1, \ldots, 4$; $i = 1, \ldots, m$; and $g = 1, \ldots G$. We consider variation on m (the sample

[2] The WLS fitting function is $f(\theta) = (s - \sigma)' W (s - \sigma)$

[3] For the required asymptotic theory, see e.g. x Satorra, A. and P.M. Bentler (1990). Model Conditions for Asymptotic Robustness in the Analysis of Linear Relations, *Computational Statistics & Data Analysis , 10,* 235-249

size in each second-level units, considered to be constant for all the second-level units, for simplicity). Here $\lambda * \lambda$ denotes simply element-wise squaring. We further let $\rho = 0.8$, $\Psi_v = (1 - \rho^2) \times I_{T-1}$,

$$\begin{aligned} f_2 &= \rho f_1 + v_2 \\ \dots & \\ f_T &= \rho f_{T-1} + v_T \end{aligned}$$

For this type of data, we can compute the intraclass (intra-area) correlation (ICC) as the ratio of the variances within-area and total variance; that is $ICC = \phi_{f_1}/(\phi_{f_1} + \phi_w)$. The within versus between variation for the true values F_{ig} are shown in Figure 3. This is a plot of the true values of F (the ordinate scale) against the area number. The plot considers G=100 areas, and m=60 cases in each area. Big circles in the plot denote the true value f_g for each area. The graph shows variation of the F_{ig} around the values of f_g than in turn show variation across the areas. In this graph the ICC = 0.1.

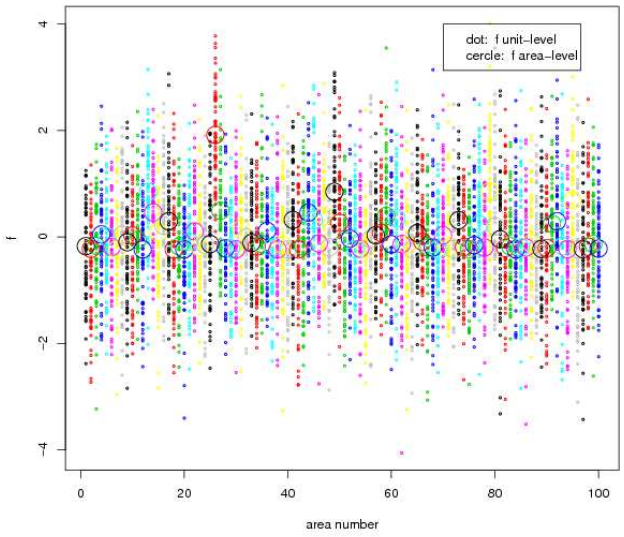

Figure 3: Two level variation when ICC = 0.1, G = 100, mg=60

SEM analysis can provide correct asymptotic inferences for normal and non-normal data. This leads us to consider simulated data that has f_g non-normally distributed. We let data arise from the scaled variable $f_g \sim (\chi^2_1 - 1)/\sqrt{2}$. The distribution of $\nu_t, t = 2, \dots, T$, as well as the distribution of the first-level variables w_{tgi} and ϵ_{jtgi}, was chosen to be normal. The histogram shown in Figure 4 corresponds to the distribution of f with $G = 1200$.

A data set of the form described is simulated and the models LFM and LFMw are fitted using EQS. Results on estimates and standard errors (both the NT and robust ones) are collected into tables 1 to 3 that vary m and ICC. The value of G is set constant equal to 1200 (a small value of G was also used but without observing difference in the results that will be reported). The tables have two parts, the first half is devoted to LFMw, the second to LFM. The first column of the table shows the parameters involved in the model; the second column shows the estimates obtained using ML estimation; the third column shows the normal theory standard errors (se); the fourth column shows the robust se (rse). A chi-square normal theory likelihood ratio test is also computed for each of the models.

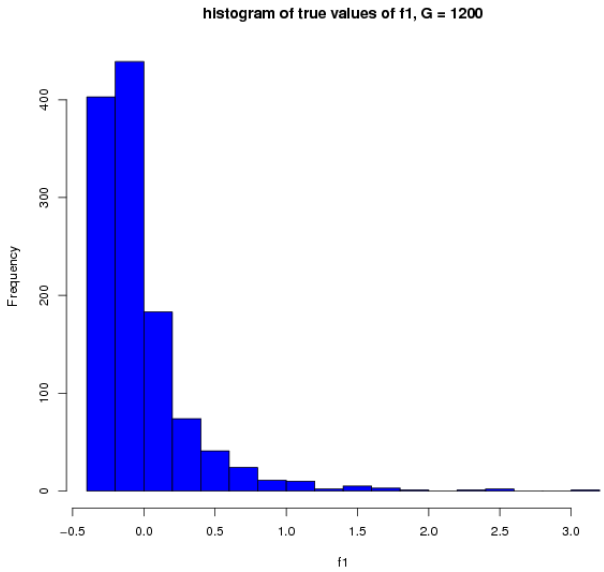

Figure 4: Histogram of true values of f (G = 1200)

Table 1: Estimates for LFM and LFMw when $m = 3$ and ICC = 0.1.

parameter	true value	$\hat{\rho}$	se	robust se
LFMw:				
ρ	.8	.807	.048	.052
λ_1	1	1	-	-
λ_2	.5	.496	.016	.016
λ_3	.1	.107	.010	.010
ϕ_f	.111	.094	.021	.023
ϕ_w	.333	.351	.033	.034
ϕ_v	.36	.356	.038	.041
Goodness of fit:	$\chi^2_{69} = 88.78$	p – value = 0.0546		
LFM:				
ρ	.8	.445	.021	.021
λ_1	1	1	-	-
λ_2	.5	.520	.016	.015
λ_3	.1	.112	.010	.010
ϕ_f	.111	.394	.026	.026
ϕ_w	-	-	-	-
ϕ_v	.36	.779	.031	.030
Goodness of fit:	$\chi^2_{70} = 127.984$	p – value = 0.00028		

Table 1 considers the case where area sample size is small, $m = 3$, and ICC is also small, ICC = 0.1. This is a case that one can expect high small area effect, in the sense that LFMw will perform differently than LFM. The autoregressive parameter ρ whose population value is 0.8 seems to be correctly estimated by model LFMw, but there is severe bias on the SEM estimates of ρ when using the approximative model LFM: the estimate is .445 when the true value is 0.8 (and the confident bounds, $\pm$ twice 0.021, do not compensate for this bias). The chi-square goodness of fit test shows a good fit for the LFMw (a p–value greater than 5%) but a very poor fit for LFM (p–value = 0.00028). Note also that the estimate of ϕ_w, the variance of the term w is large; thus, setting its value to zero, as done by LFM, incurs a severe misspecification. Note also confidence bounds of parameter estimates of the LFMw model encompass the true values of parameters. To be noted also is very minor difference between the s.e. of parameter estimates, the normal theory (se) or the robust se (rse). One curious issue to note is that in estimating $\hat{\rho}$ the model LFMw gives a much larger standard error than when using LFM. We encounter what is the payoff of using a correct model, is that bias decreases but also the se increase. Note that in this the se of a very poor estimator of ρ, the one of LFM, is half the one of the correct estimator based on LFMw.

Keeping all the conditions the same, Table 2 shows the results when area sample size increases to $m = 60$. Table 2 shows that the increase on sample size m reduces drastically

Table 2: Estimates for LFM and LFMw when $m = 60$ and ICC = 0.1 .

parameter	true value	$\hat{\rho}$	se	robust se
LFMw:				
ρ	.8	.754	.024	.024
λ_1	1	1	-	-
λ_2	.5	.498	.003	.003
λ_3	.1	.101	.003	.003
ϕ_f	.111	.112	.010	.012
ϕ_w	.017	.003	.010	.009
ϕ_v	.36	.381	.017	.0160
Goodness of fit:	$\chi^2_{69} = 68.423$	p – value = 0.497		
LFM:				
ρ	.8	.749	.017	.017
λ_1	1	1	-	-
λ_2	.5	.498	.003	.003
λ_3	.1	.101	.003	.003
ϕ_f	.111	.114	.005	.009
ϕ_w	-	-	-	-
ϕ_v	.36	.386	.009	.009
Goodness of fit:	$\chi^2_{70} = 68.500$	p – value = 0.528		

the inconsistency of the estimate of ρ when using LFM (even though a slight bias still persist since the 95% coverage interval $0.749 \pm 2 * 0.0167$ misses 0.8 slightly. The estimate of ρ based on LFMw seems to perform on target. Note that the se for the LFM model are still slightly smaller than for the LFMw. The chi-square goodness of fit now does not reject LFM (p-value = 0.528). Again the LFMw provides accurate estimates for all the parameter of the model (all true values under the the 95% confidence bounds of parameter estimates). The estimate of the variance ϕ_w has now been reduced drastically, as the population value

of this variance is indeed reduced since it has been multiplied by the factor $1/m$ where m is large. Slight differences are observed between the normal theory and robust se. We thus conclude that the increase of m makes LFM a valid model representation for the data. In the ICT data example to be discussed below, $m = 75$ so it is likely we are falling in that situation where LFM is the appropriate model. The exact correct model is LFMw , but setting $\phi_w = 0$ in that data context does not lead to misspecification.

For both the LFMw and LFM and the data context of the simulations, theory of asymptotic robustness (AR) as described in Satorra and Bentler (1990) ensures asymptotic correctness of the se for all the parameters except for the variance of the non-normal component. Thus, we should expect the se and rse to be very similar (since they are asymptotically equal, adhering to the AR mentioned) except possibly for the se for the estimate of ϕ_f. We see that discrepancy between robust and normal theory se is minor for all the parameters. We also see small discrepancy between the se and rse of estimate of ϕ_f in both tables. We will see that the discrepancy among se and rse will appear in the next table to be discussed. The theory of asymptotic robustness of Satorra and Bentler (1990) attests also to the asymptotic correctness of the goodness of fit tests shown in the tables, for both models and given the data generated process, despite non-normality.

We have seen the effect of increasing the area sample size m. We will now keep $m = 3$ small, but we will increase ICC to a large value, ICC = 0.833. This case is illustrated in Table 3 The results of this table show that increasing ICC has the same effect as increasing the area sample size m. LFM is now again a valid approximative model performing similarly as the true model LFMw. The only difference with Table 2 is the large difference in se and rse for the estimate of ϕ_f. This is in accordance to the AR theory of Satorra and Bentler (1990), where now we see the distortion of the normal theory se of estimator of ϕ_f now made prominent given the high ICC value. So, with regard to parameter estimation,

Table 3: Estimates for LFM and LFMw when $m = 3$ and ICC = 0.833

parameter	true value	$\hat{\rho}$	se	robust se
LFMw:				
ρ	.8	.812	.017	.016
λ_1	1	1	-	-
λ_2	.5	.486	.010	.010
λ_3	.1	.107	.009	.009
ϕ_f	.833	.852	.041	.088
ϕ_w	.056	.067	.020	.020
ϕ_v	.36	.372	.027	.025
Goodness of fit:	$\chi^2_{69} = 69.399$	p – value = 0.463		
LFM:				
ρ	.8	.775	.014	.013
λ_1	1	1	-	-
λ_2	.5	.495	.009	.009
λ_3	.1	.111	.009	.009
ϕ_f	.111	.897	.010	.087
ϕ_w	-	-	-	-
ϕ_v	.36	.445	.019	.019
Goodness of fit:	$\chi^2_{70} = 79.636$	p – value = 0.202		

we conclude that

1. LFMw produces consistent estimates of ρ (and the other parameters of the model) even for small m and small ICC;

2. Analysis based on LFM suffers from (severe) bias when estimating ρ (and other parameters) when m and ICC are small;

3. Estimates of LFMw and LFM converge to each other when m and/or ICC increase.

4. Factor score estimation

We now consider factor score estimation. We will develop estimates for the realized values of the (latent) variables f_1 and $v_1, \ldots, v_T$ for each area $g = 1, \ldots, G$.

We rewrite LFMw in factor model form

$$y = \Pi v + r \tag{8}$$

where $\Pi = \Lambda(I - B)^{-1}$, $r = \Lambda w + \epsilon$ and $v' = (f_1, v_2, \ldots, v_T)$. Let Ψ_r, Φ_v, Ψ_w and Ψ_e denote the variance matrices of r, v, w and e respectively. Clearly, $\Psi_r = \Lambda\Psi_w\Lambda' + \Psi_e$. Now applying to (8) the classical theory of factor score estimation in factor analysis, we obtain the two factor score estimators

Bartlett's:

$$\tilde{v}_B = (\Pi'\Psi_r^{-1}\Pi)^{-1}\Pi'\Psi_r^{-1}(y - \mu_y)$$

and

Thomson's:

$$\tilde{v}_T = \Phi_v\Pi'\Sigma_y^{-1}(y - \mu_y)$$

where $\Sigma_y = (\Pi\Phi_v\Pi' + \Psi_r)$ and $\mu_y = E(y)$.

1. Thomson's correspond to BLUP (best linear unbiased predictor)[4]

2. Feasible $\hat{v}_B$ and $\hat{v}_T$ are obtained by replacing population values of parameters by consistent estimates. The feasible Thomson's is the (empirical) BLUP (i.e., an EBLUP).

When $T = 1$ and $k = 1$ then

$$y_{gi} = f_g + w_{gi}$$

since $\Lambda = 1$, $\epsilon_{ig} = 0$. Let $\Phi_w = \phi_w$ $(= \sigma_w^2)$ be the variance within area, and $\Phi_{f_1} = \phi_{f_1}$ $(= \sigma_b^2)$ the between area variance. It is interesting to note that in that case ($T = 1$ and $k = 1$), Thomson's factor score estimator takes the form

$$\tilde{f}_g = \frac{\Phi_f}{\Phi_f + n_g^{-1}\Phi_w}(y_{.g} - \overline{y_{.g}}) \tag{9}$$

an expression that coincides with the classical (combined) small-area estimator for a centered variable (for an overview on the theory of small area estimation, see Ghosh and Rao, 1994). Our Thomson's factor score estimation formula generalizes thus to multivariate analysis ($k > 1$) and/or longitudinal data ($T > 1$), classical formulae for small area estimation. As naive estimator of the factor score, to be compared with the ones developed in the present paper, we consider the direct and combined estimators proposed in the literature for small area estimation (e.g., Rao, 2003).

To assess the comparative performance of the alternative estimators of the area characteristics, we consider also simulated data. The true value is known due to the simulation

[4] Neudecker, H. and A. Satorra (2003), 'On best affine prediction", *Statistical Papers*, *44*, 257-266

scheme, and the quality of the alternative estimators are evaluated by computing the mean square error

$$\text{RMSE} = \sqrt{\sum_g (\hat{f}_g - f_g)^2 / G}$$

of the corresponding estimator $\hat{f}_g$ of the realized and known true value f_g. We consider several data sets that vary on the conditions of the data with respect the size of the area sample m and the intraclass correlation coefficient ICC. The estimators considered are the following:

tLFMw, bLFMw: Thomson's and Bartlett's arising from LFMw specification

tLFM, bLFM: Thomson's and Bartlett's arising from LFM specification

D, csae: Small area estimators arising from considering the single variable $y_{ig} = 1_k' y_{ig1}/k$, where 1_k a $k \times 1$ vector of 1s (i.e. y_{ig} is the simple average of all the indicators). D is the direct estimator $y_{.g}$ and case is the composite small area estimator (9) with variances between replaced by consistent estimates. In this case, we only consider estimation of the factor score f_1 for each area g. We do not consider small area estimation of the time change variables v_t.

The table below show the results of the simulations under different data conditions. Note that the columns for D and $csae$ contains only information for the factor score f_1.

	tLFMw†	bLFMw‡	tLFM	bLFM	D	csae
			estimating f1			
f1	0.241	0.254	0.248	0.254	0.503	0.531
			estimating ν 's			
ν_2	0.275	0.309	0.291	0.307		
ν_3	0.271	0.305	0.288	0.304		
ν_4	0.272	0.300	0.286	0.300		

† t of Thomson's

‡ b of Barlett's

Table 4: The RMSE of different estimators when G= 1200, m = 40 and ICC=0.33

In the data context of Table 4, where area sample size is relatively large and also ICC is moderate, we see that the both Bartlett and Thomson estimators for f_1 have a similar performance, irregardless of whether we use the LFMw or LFM specifications. The same occurs for the estimators of the change variables v_2 to v_4. The RMSE of the small area estimators (D and csae) double approximately those of the estimators based on a model, with both D and csae performing similarly. Note that the data context considered, where both m and ICC are relatively large, amounts to a set-up where the direct estimator performs similarly as the composite small area one. We now investigate the case where we decrease the area sample size to m=3, so that it is likely that the small area formulation produces gains relatively to the standard approach.

In the data context of Table 5, where m has been reduced to 3, we see that the use of a proper small area sample model as LFMw clearly wins over the approximative LFM. We see that Bartlett's yields a higher RMSE than Thomson's. For the Thomson's estimator, LFMw is associated to a clear RMSE gain with respect the LFM: compare the RMSEs 0.586 versus 0.784 when comparing the Thomson's estimators for the LFMw and LFM specifications. Similar gains are attained for Thomson's estimators of the change variables

	tLFMw†	bLFMw‡	tLFM	bLFM	D	csae
			estimating f1			
f1	0.586	0.895	0.784	0.892	0.741	0.753
			estimating ν 's			
ν_2	0.504	1.185	0.882	1.060		
ν_3	0.500	1.217	0.901	1.097		
ν_4	0.512	1.197	0.879	1.049		

† t of Thomson's ; ‡ b of Barlett's

Table 5: RMSE for different models and estimators. G=1200, m = 3 and ICC=0.33

v_ts. To be noted is that Bartlett factor score estimators are unaffected for whether we use the LFMw or LFM. Table 6 gives a variation of the data set-up, where now the ICC has been decreased to ICC = 0.1. The results are similar as the ones of Table 5, except that now csae wins over D. This is a case where estimators based on small area clearly win over the ones that ignore the small area formulation, and one could say that LFMw is a small areas model with respect to LFM. When ICC is large, we see an increase of 67% (0.494/0.296=1.668919) of the RMSE of LFM relatively to LFMw. This increase was smaller in the case of ICC = 0.33, where we observed only a 34% increase on RMSE (0.784/0.586=1.337884) when comparing RMSE of LFM versus LFMw.

	tLFMw†	bLFMw‡	tLFM	bLFM	D	sae
			estimating f1			
f1	0.296	0.684	0.494	0.681	0.492	0.376
			estimating ν 's			
ν_2	0.424	0.880	0.594	0.756		
ν_3	0.466	0.900	0.648	0.799		
ν_4	0.458	0.854	0.630	0.764		

Table 6: RMSE for different models and estimators, G = 1200, m= 3 and ICC = 0.1

On factor score estimation, we thus conclude: (1) The Thomson's tLFMw (a genuine small area estimator) wins over the alternatives for all the data contexts considered; (2) The gains of tLFMw with respect to tLFM increase with the decrease of the area sample size and/or the decrease of the ICC; (3) RMSE of Thomson's estimator outperforms the ones of the Bartlett's estimator in general, but they tend to the same size when m is large and/or ICC is large. (4) Estimates of the change variables v_t can be obtained similarly as the factor estimators of f_1. Those estimators may allow specific characterization of the areas $g = 1, \ldots, G$.

5. Empirical analysis: ICT data

Statistical offices worldwide invest resources on Information and Communication Technology (ICT) surveys. For policy purposes, interest is in the level of ICT at small area levels. We are interest also in changes of level of ICT. We use ICT for small areas: four repeated surveys, 41 counties, $m = 75$, population of Catalonia, Spain.

Synthetic variables extracted from the *Territorial Survey on Information and Communication Technologies of the Households (ICT)'* of 2008-2011, conducted by the Statistics Institute of Catalonia, IDESCAT, Barcelona, Spain, and the estimated loadings for the de-

scribed LFMw model are shown in Table 7. The data consist on empirical data for 12 variables, 41 counties ("comarques", small areas), with a sample size m = 75 for each area. We consider 4 repeated surveys, years 2008–2011. [5]

Table 7: 12 synthetic ICT variables and the loading's estimates corresponding to LFMw. Numbers in brackets are the standard errors (robust) of estimates.

Variable	Description	$\hat{\lambda}$	rse
econ	Level of equipment in the household	1.043	(.094)
mobile	Number of mobile phones in the household	.430	(.073)
sinceInt	Since when has access to Internet	.998	(.149)
knowPC	Number of tasks with a computer	1.024	(.042)
knowInt	Number of tasks through Internet	1.000†	
frecPC	Intensity of use of personal computer	1.162	(.084)
frecInt	Intensity of use of Internet	1.187	(.041)
secbuy	Security perception: buying through Internet	.444	(.056)
secbank	Security perception: bank transactions through Internet	.163	(.016)
econ	Level of economic impact of ICT actions from home	.641	(.045)
social	Level of social impact of ICT actions from home	1.144	(.060)
admin	Level of interaction with public administrations through Internet	.371	(.044)

† fixed to 1 to set the scale of the factor

For this data, we fitted the LFMw model obtaining the loading estimates shown on the last columns of Table 7. The estimated autoregression was

$$\begin{array}{rcl} f_{g(t+1)} & = & .94\, f_{gt} + v_t \\ & & (.091) \end{array}$$

$\phi_f = .280(.093)$, $\phi_v = .054(.038)$, $\phi_w = .03659(.03651)$

The analysis was carried out using the SEM software EQS-6.2. The estimation method used was LS with distribution free (robust) s.e. and test statistics.[6]

The fit of LFM with the ICT data gave very close loading estimates as the ones reported above, and the following fitted autoregression equation

$$\begin{array}{rcl} f_{g(t+1)} & = & .898\, f_{gt} + v_t \\ & & (.050) \end{array}$$

$\phi_f = .306(.077)$, $\hat{\rho}_{LFM} = .898(.050)$ compares with $\hat{\rho}_{LFMw} = .94(.091)$. The same software was used for the LFM analysis. [7]

This data shows a high stability across time of ICT at area level, so there does not seem to be much variation of ICT across areas in the period of investigation. This is attested by the very high value of the estimated autoregression coefficient ($\hat{\rho} = .94$). Very similar results are obtained wether one use the "small area" model approach of LFMw or the "large

[5] Thanks are due to Eva Ventura from Universitat Pompeu Fabra, and Maribel Garcia, Marcos Pardal and Cristina Rovira, from IDESCAT, for providing access to this data.

[6] ML was not feasible, since S is a 48×48 matrix ($k = 12$ variables and $T = 4$ time points) and "sample size" for SEM analysis is $G = 41$. Mean and variance adjusted Chi2 =14.644 with 10 D.F. (P value= 0.146) (see Satorra and Bentler (1994) for the robust statistics used in conjunction to the LS estimation used.

[7] Mean and variance adjusted Chi2 = 13.132 with 10 D.F. (P value= 0.21637) for LFM.

area" approach of LFM. Note that the simulation results of Table 2 showed that for a low intraclass correlation and area sample size of $m = 75$, as the one used in this ICT data, LFM and LFMw performed similarly.

The small value of the estimated variance of w obtained when fitting LFMw ($\phi_w = .037(.037)$) it also suggest LFM will be a valid approximative model for this data. Thus longitudinal analysis of the described CIT data based on a first order longitudinal factor model (LFM) is justified; with this ICT data, just proceed using LFM. The detail analysis of the LFM model for this ICT data is reported in work of Satorra and Ventura (in preparation).

The estimation of the factor scores for f and v_2 to v_4 were also computed for both models, though not reported here for conciseness; both the LFMw and LFM lead also to very similar factor score estimates, in parallel to the simulation results of Table 4 that considers a large value of m.

6. Discussion

A model has been proposed to perform longitudinal analysis with repeated cross section data that shares the same clustering across surveys. A model was proposed, the LFMw, that allowed consistent estimation of longitudinal parameters from independent cross-section data with clustering.

One example of application is small area estimation, where we have areas (clusters, second level units) and repeated surveys on that population with the same area structure. A small area context is the one that had served as an empirical illustration. A model has been specified, the LFMw, that takes care of the small area effect induced by small samples within each area, and/or small values of the intraclass (intra-area) correlation coefficient.

In contrast with classical small area estimation, where the target is just characteristics of the areas, the LFMw allowed not only to improve on classical naive small area estimators, but also allows to consistently estimate structural parameters of the longitudinal model, like for example the autoregression coefficient. In the simulations we showed that very small sample size in each area, and also very small ICC, allowed consistent estimation of key structural parameters of the LFMw longitudinal model using classical tools of SEM analysis. When sample within each area is not large, or ICC is small, consistency of parameter estimates for the structural parameters of the longitudinal model is lost when we take an approximative model LFM instead of LFMw. The use of LFM is valid when we have large samples within each area, or very large values for ICC. Both to analyze LFMw and LFM we used the widely available software EQS (Bentler, P. M. (2000-08), thus the methods proposed are readily available to researchers. Other software of structural equation modeling (e.g., LISREL, Mplus, AMOS, or `sam` of Stata) could as well have been used.

LFMw improves RMSE of factor score estimation, over naive alternatives, and improves small area estimation. For large size clusters and/or high ICC, LFMw converges to LFM. The large gains of LFMw are on low values of m and small values of ICC. On our ICT data, LFMw produced similar results as LFM (here $m = 75$, m is large).

Other applications remains to be explored where the use of the LFMw gives a clear gain over LFM, in allowing consistent estimation of parameters of a longitudinal model and more precise estimates of area level characteristics. From the theory and simulation work of this paper, we know that these applications should consist of repeated clustered surveys with modest values of intraclass correlation coefficient and small area sample size.

REFERENCES

Bentler, P. M. (2000-08), *EQS 6 structural equations program manual*. Encino, CA: Multivariate Software, Inc.

Binder, D. A. and Dick, J. P. (1989). "Modelling and estimation for repeated surveys" *Survey Methodology*, 15, 29-45

Binder, David A. (1998), "Longitudinal surveys: Why are these surveys different from all other surveys?", *Survey Methodology*, 24, 101-108

Binder, D. A. and Dick, J. P. (1989), "Modelling and estimation for repeated surveys", *Survey Methodology*, 15, 29-45

Blight, B. J. N. and Scott, A. J. (1973), "A stochastic model for repeated surveys", *Journal of the Royal Statistical Society, Series B: Methodological*, 35, 61-66

Duncan, G. J. , and Kalton, G. (1987), "Issues of design and analysis of surveys across time", *International Statistical Review*, 55, 97-117

Fay, R. E., and Herriot, R. A. (1979), Estimates of Income for Small Places: An Application of JamesStein Procedures to Census Data, *Journal of the American Statistical Association*, 74, 269 – 277

Fuller, W. A. (1990), "Analysis of repeated surveys", *Survey Methodology*, 16, 167-180

Holt, D. and Skinner, C, J, (1983), "Components of change in repeaters surveys", *International Statistical Review*, 57. 1–18

Jones, R. G. (1980), "Best linear unbiased estimators for repeated surveys", *Journal of the Royal Statistical Society, Series B: Methodological*, 42, 221–226

Kalton, G., and Citro, C. F. (1993), "Panel surveys: Adding the fourth dimension", *Survey Methodology*, 19, 205-215

Kasprzyk, D., Duncan, G., Kalton, G., and Singh, M. P. (eds) (1989), *Panel Surveys*, John Wiley & Sons, New York

Ghosh, M. and Rao, J. N. K. (1994) "Small area estimation: An appraisal"(with discussion), Statistical Science 9, 65–93

Neudecker, H. and A. Satorra (2003), "On best affine prediction", *Statistical Papers*, 44, 257–266

Pfeffermann, D., Feder, M. and Signorelli, D. (1998), "Estimation of autocorrelations of survey errors with application to trend estimation in small areas", *Journal of Business & Economic Statistics*, 16, 339-348

Pfeffermann, D. (2013),"New Important Developments in Small Area Estimation", *Statistical Science*, 28, 40–68

Rao, J. N. K. (2003). *Small Area Estimation*, Wiley Series in Survey Methodology, New York.

Satorra, A. and P.M. Bentler (1990), "Model Conditions for Asymptotic Robustness in the Analysis of Linear Relations", *Computational Statistics & Data Analysis*, 10, 235–249

Satorra, A. and P.M. Bentler (1994), "Corrections to test statistics and standard errors in covariance structure analysis," in *Latent variable Analysis in Developmental Research*, eds. A. van Eye and C.C. Clogg, California: Sage Publications, pp. 285–305.