Multivariate approach to small area estimation: towards an area-level index of TIC

Albert Satorra and Eva Ventura, UPF
Cristina Rovira, Maribel Garcia and Marc Pardal, IDESCAT

Outline

An index of TIC at small area ("comarca") level ?

- Statistical offices, worldwide, currently invest resources on surveys of ICT (from now on, "TIC")
- The aim is not to inspect TIC activity at the individual level, rather at the country and/or small area levels. It's also of interest comparison of areas on TIC intensity.
- Several questions of the TIC surveys are designed to capture the "level" of involvement in activity TIC, we say these questions share a general "common factor"
- We aim to assess variation of this "common factor" at area level
- Our work is "linking" multivariate (factor) analysis with small are estimation

Hierarchical data, and model

Two-level data (case i, area d):

$$x_{id}\,(p \times 1);\ i = 1, \ldots, n_d;\ d = 1, \ldots, D$$

Factor model:

$$x_{id} = \mu + \Lambda f_{id} + \epsilon_{id} \tag{1}$$

$f_{id}(p \times q)$ and $\epsilon_{id}\,(p \times 1)$ are (centered) vectors of common and unique factors, with respective variance matrices Φ and Ψ; $\mu\,(p \times 1)$ is a mean vector: $\Lambda\,(p \times q)$ is the loading matrix. We assume

$$f_{id} = f_{id}^{(1)} + f_d^{(2)};\quad \epsilon_{id} = \epsilon_{id}^{(1)} + \epsilon_d^{(2)} \tag{2}$$

Here $f_{id}^{(1)}$ and $f_d^{(2)}$ and $\epsilon_{id}^{(1)}$ and $\epsilon_d^{(2)}$ are mutually independent (centered) random vectors with variance matrices Φ_w , Φ_b, Ψ_w and Ψ_b, respectively; the superscripts $^{(1)}$ and $^{(2)}$ denoting variation at the first- (individual) and second- (area) levels, respectively.

Combining (1) and (2),

$$x_{id} = \Lambda f_d^{(2)} + \Lambda f_{id}^{(1)} + \epsilon_{id}; \tag{3}$$

and averaging to level two:

Factor Model (area level)

$$x_{.d} = \Lambda f_d + u_d, \tag{4}$$

$$u_d = \Lambda f_{.d}^{(1)} + \epsilon_{.d}^{(1)} + \epsilon_d^{(2)}$$

$$x_{.d} = \sum_i x_{id}/n_d; \quad f_{.d}^{(1)} = \sum_i f_{id}^{(1)}/n_d; \quad \epsilon_{.d}^{(1)} = \sum_i \epsilon_{id}^{(1)}/n_d$$

It holds:

1. $$\Phi_f = \Phi_b + \Phi_w$$

2. $$\text{var}\left(f_{.d}^{(1)} \mid d\right) = \Phi_w / n_d$$

3. $$\text{var}\left(u_d \mid d\right) = \Psi_b + n_d^{-1} \times \left(\Lambda \Phi_w \Lambda' + \Psi_w\right)$$

4. $$\text{var}\left(u_d\right) = \Psi_b + E(n_d^{-1}) \times \left(\Lambda \Phi_w \Lambda' + \Psi_w\right)$$

5. $$\text{var}\left(x_{.d}\right) = \Lambda \Phi_b \Lambda' + \text{var}\left(u_d\right)$$

BLUP of f_d ("small-area" estimation)

From formulae in Neudecker and Satorra (2003, p. 261), [1] the BLUP of f_d in (4) can be written as

BLUP for f_d

$$\hat{f}_d = \Phi_b \Lambda' [\Lambda \Phi_b \Lambda' + \Psi_b + n_d^{-1} \times (\Lambda \Phi_w \Lambda' + \Psi_w)]^{-1} (x_{.g} - \mu) \quad (5)$$

The EBLUP replaces population parameters by estimates. There is a variety of estimates one can use. [2]

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

[2] See Satorra and Bentler (2011) for further details on two-level factor score prediction.

BLUP $\hat{f}_d$: "Classical" and multivariate sae

"classical:" If $p = q = 1$ then $x_{id} = f_{id}$, $\Lambda = 1$, $\epsilon_{id} = 0$,

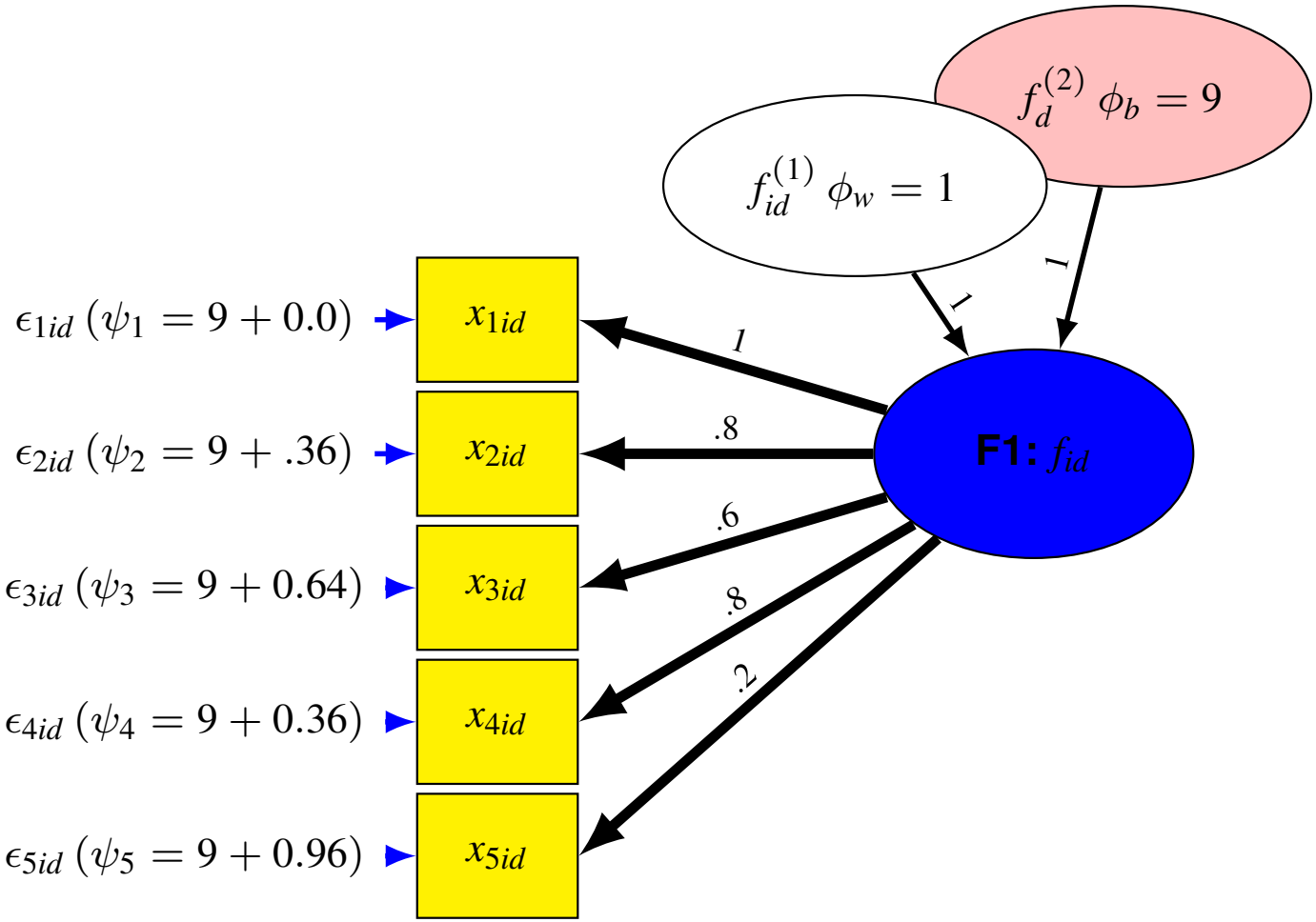
$$\hat{f}_d = \frac{\Phi_b}{\Phi_b + n_d^{-1}\Phi_w}(x_{.d} - \overline{x_{.d}}),$$

a familial expression for the "classical" small area estimator (of a centered variable); $\hat{f}_d$ "borrows strength" from neighboring areas only.

multivariate: If $p > 1$, $\hat{f}_d$ is multivariate, it "borrows strength" not only from related areas, but also from several variables (specially from the variables more correlated with the target variable f_d)

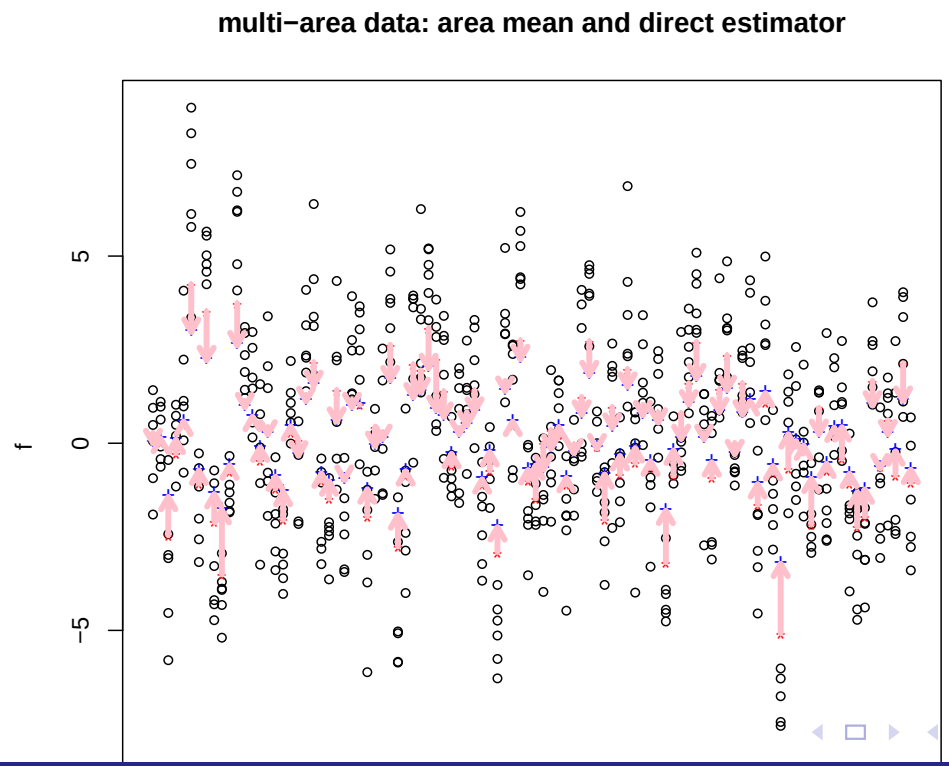
- A small Monte Carlo study will be used to evaluate the proposed multivariate small area estimator.
- Two level data, and ML estimation of two-level confirmatory factor model, and ML estimation of a single (area level) confirmatory factor model.
- The approach can be solved using software for structural equation models (e.g., `sem` in `R`; EQS (Bentler, 2009) ; Mplus (Muthén and Muthén, 2010); LISREL (Joreskog and Sorbom, 2007); the new module `sem` of `Stata`, etc.) .

Two level data generated by a one-factor model, with a two-level factor:



The two-level f_{id} (arrows from area parameter to direct estimator)

Data with a factor with two-level variation



Intraclass correlations of the variables X1 to X5

SUMMARY OF DATA

Number of clusters 100

Average cluster size 6.000

Estimated Intraclass Correlations for the Y Variables

Variable	Intraclass Correlation
X1	0.162
X4	0.129
X2	0.095
X3	0.067
X5	0.004

e.g. `sem` estimation, with `R`

```
model   <- matrix(c(
                     'f1  -> v1', NA,  1,
                          'f1  -> v2', 'lamb2', NA,
                               'f1  -> v3', 'lamb3',  NA,
                                'f1   -> v4', 'lamb4',  NA,
                                 'f1   -> v5', 'lamb5',  NA,
                   'v1  <-> v1', 'theta1',  NA,
                        'v2  <-> v2', 'theta2',  NA,
                        'v3  <-> v3', 'theta3',  NA,
                        'v4  <-> v4', 'theta4',  NA,
                        'v5  <-> v5', 'theta5',   NA,
                    'f1 <-> f1', 'phi', 10),
                     ncol=3, byrow=TRUE)
 Model Chisquare =  2.050578   Df =  5
 Parameter Estimates
       Estimate  Std Error z value  Pr(>|z|)
lamb2   0.855709 0.18621    4.59546 4.3180e-06 v2 <--- f1
lamb3   0.569625 0.17795    3.20108 1.3691e-03 v3 <--- f1
lamb4   0.964611 0.20782    4.64160 3.4572e-06 v4 <--- f1
lamb5  -0.065828 0.13196   -0.49887 6.1787e-01 v5 <--- f1
theta1  1.675759 0.42439    3.94863 7.8601e-05 v1 <--> v1
theta2  2.091036 0.40413    5.17410 2.2901e-07 v2 <--> v2
theta3  2.873968 0.44726    6.42569 1.3127e-10 v3 <--> v3
theta4  2.379975 0.48109    4.94703 7.5354e-07 v4 <--> v4
theta5  2.365462 0.33671    7.02522 2.1374e-12 v5 <--> v5
phi     1.893884 0.56945    3.32582 8.8160e-04 f1 <--> f1
```

MSE of alternative estimators

Table: Monte Carlo estimates and standard errors of the MSE of alternative small area estimators (Monte Carlo based on 200 replications).

MSE	dirf †	dirX1	dirX5	dirfs	newMLf	newpopf
$\widehat{MSE}$:	49.71	197.50	513.75	227.41	165.58	127.06
se$(\widehat{MSE})$:	0.37	2.45	5.92	3.10	3.32	1.41

dirf: direct estimator with known true values of f

dirX1 and dirX5: direct estimates based on X1 and X5, respectively

dirfs: direct estimator based on the factor scores of ML analysis of individual data

newMLf: EBLUP of (5); Λ and Φ_b ML- between-area data, Φ_f ML-individual data (two ML CFA analyses involved)

newpopf: BLUP of (5); parameters at their true value

† non-feasible estimates are shown in blue

TIC 2008, Catalonia

Territorial Survey on Information and Communication Technologies of the Households (TIC) of 2008, conducted by the Statistics Institute of Catalonia, IDESCAT.

- Rotating panel, with five shifts of rotation and on an annual basis.
- Stratified two-stage sample.
  - The strata are 41 Catalan administrative divisions
  - The sample is uniformly distributed among them with 75 randomly selected first stage units (dwellings) in each administrative division.
  - Second-stage unit is person aged 16 or over (also randomly selected).
- Sample size about 3000 individuals.

There are 56 survey questions, distributed among 12 blocks . We first need to create a few synthetic variables that summarize the

First stage: Constructing synthetic variables

- The second and third blocks of survey's questions contain information on equipment and access to Internet in the household. We summarize part of this information in the new variable equip
  - 0 for households without a personal computer and without access to Internet
  - 1 for households with a computer but no access to Internet
  - 2 for households with a computer and slow access to Internet
  - 3 for households with a computer and broadband access to Internet
- The two blocks also contain information on the number of operating mobile phones in the household, variable mobilcom It ranges from 0 to 5 (5 meaning more than 4 operating phones)

Constructing synthetic variables, e.g. frecInt

Variable frecInt ranges from 0 to 7. Again, the value is 0 if the household does not have access to Internet, and it is 7 if the respondent is connected to Internet over 20 hours a week.

- More than 20 hours a week: 7
- Between 5 and 20 hours a week: 6
- Between 1 and 5 hours a week: 5
- NA/DK how many hours a week, but uses it weekly: 5
- 0 to 1 hour weekly: 4
- At least once a month, but not every week: 3
- Not every week, but used in the past 3 months: 2
- Did not used PC in the past 3 months: 1
- Does not have a PC: 0

Synthetic Variables

Table: Variables In Factor Analysis

Variable	Description	Values
econ	Level of equipment in the household	0 to 3
mobile	Number of mobile phones in the household	0 to 5
sinceInt	Since when has access to Internet	0 to 4
knowPC	Number of tasks with a computer	0 to 9
knowInt	Number of tasks through Internet	0 to 9
frecPC	Intensity of use of personal computer	0 to 5
frecInt	Intensity of use of Internet	0 to 7
secbuy	Security perception: buying through Internet	0 to 3
secbank	Security perception: bank transactions through Internet	0 to 3
econ	Level of economic impact of ICT actions from home	0 to 4
social	Level of social impact of ICT actions from home	0 to 4
admin	Level of interaction with public administrations through Internet	0 to 6

The naive approach: PCA

Principal component analysis of the data aggregated by comarques gives the following results. Figure 3 shows that the solution is basically unidimensional. We will retain, however, the second component for the purpose of plotting comarcas and variables of TIC.

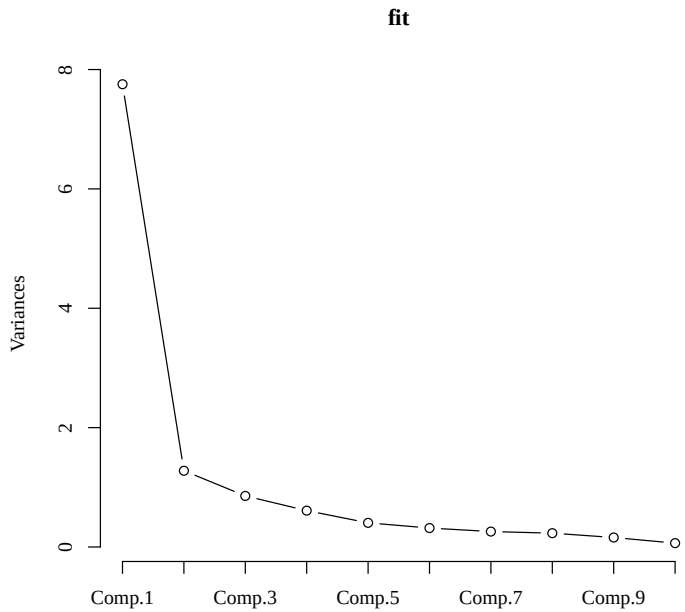


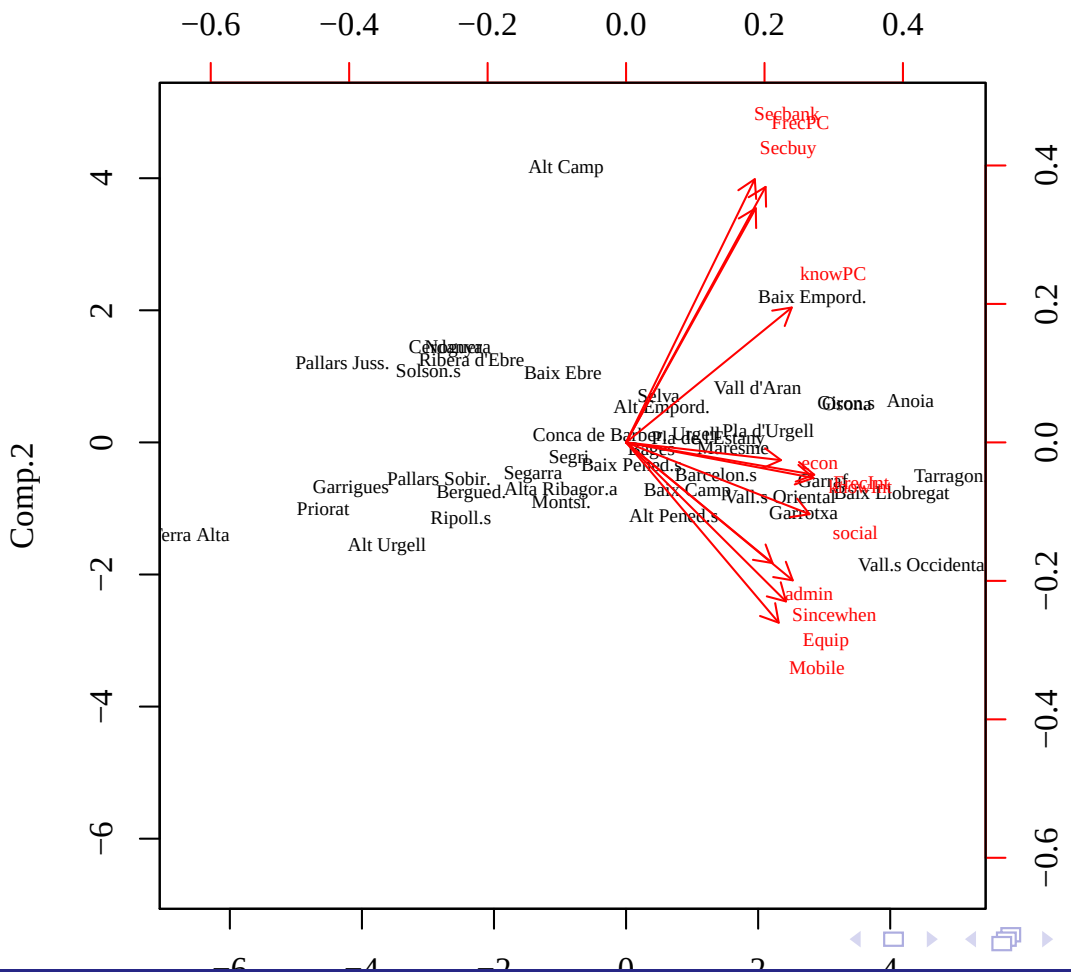
Figure: Scree plot for weighted data aggregated by commarcas

Correlations of variables with principal components

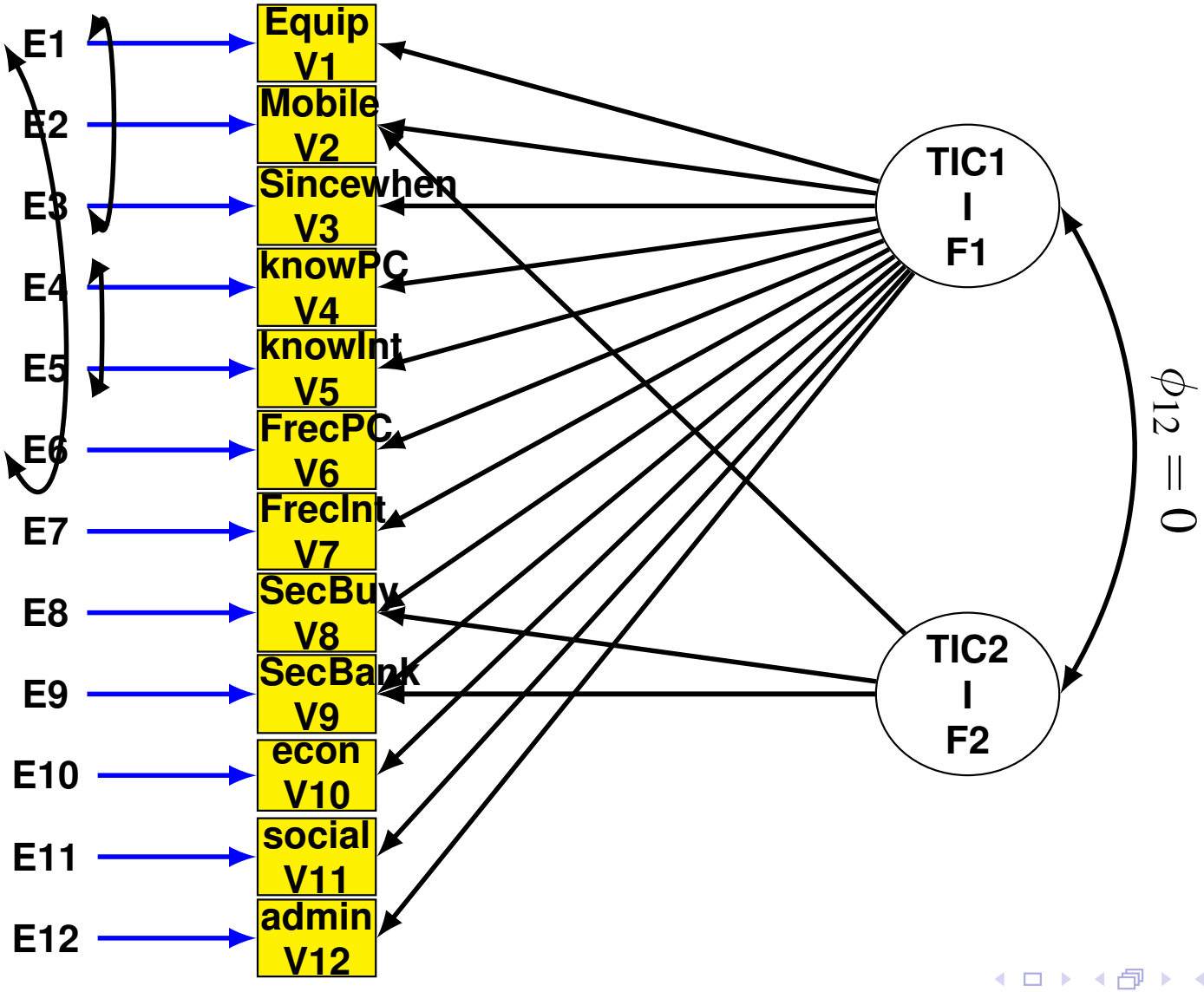
Table: Correlations Variables vs PCs

	Comp.1	Comp.2
Equip	0.80	-0.32
Mobile	0.77	-0.37
Sincewhen	0.84	-0.28
knowPC	0.83	0.28
knowInt	0.94	-0.07
FrecPC	0.70	0.52
FrecInt	0.95	-0.07
Secbuy	0.65	0.48
Secbank	0.65	0.54
econ	0.78	-0.04
social	0.92	-0.15
admin	0.74	-0.25

The biplot


−0.6
−0.4
−0.2
0.0
0.2
0.4
Comp.2
4
2
0
−2
−4
−6
0.4
0.2
0.0
−0.2
−0.4
−0.6
Secbank
FrecPC
Secbuy
Alt Camp
knowPC
Baix Empord.
Pallars Juss.
Solson.s
Ribera d'Ebre
Baix Ebre
Selva
Alt Empord.
Vall d'Aran
Osona
Anoia
Pla d'Urgell
Maresme
Segri
Segarra
Baix Pened.s
Barcelon.s
Garrigues
Pallars Sobir.
Bergued.
Alta Ribagor.a
Baix Camp
Vall.s Oriental
Tarragon
Priorat
Montsi.
Garrotxa
Baix Llobregat
Ripoll.s
Alt Pened.s
social
Alt Urgell
Vall.s Occidenta
admin
Sincewhen
Equip
Mobile

CFA: TIC 2008


E1
E2
E3
E4
E5
E6
E7
E8
E9
E10
E11
E12
Equip V1
Mobile V2
Sincewhen V3
knowPC V4
knowInt V5
FrecPC V6
FrecInt V7
SecBuy V8
SecBank V9
econ V10
social V11
admin V12
TIC1 I F1
TIC2 I F2
$\phi_{12} = 0$

CFA: 2008 data. Correlations of variables with F1 and F2.

Table: Correlation of variables and factors

	F1	F2
Equip	0.76	0.01
Mobile	0.72	-0.02
Sincewhen	0.80	0.12
knowPC	0.79	0.14
knowInt	0.98	-0.01
FrecPC	0.65	0.30
FrecInt	0.99	0.02
Secbuy	0.50	0.54
Secbank	0.52	0.86
econ	0.76	0.04
social	0.96	-0.01
admin	0.68	-0.00

```
CHI-SQUARE = 55.607 BASED ON 48 DEGREES OF FREEDOM PROBABILITY VALUE FOR THE

CHI-SQUARE STATISTIC IS 0.21010
```

Index Comarcal de TIC (F1). CFA: 2008 data

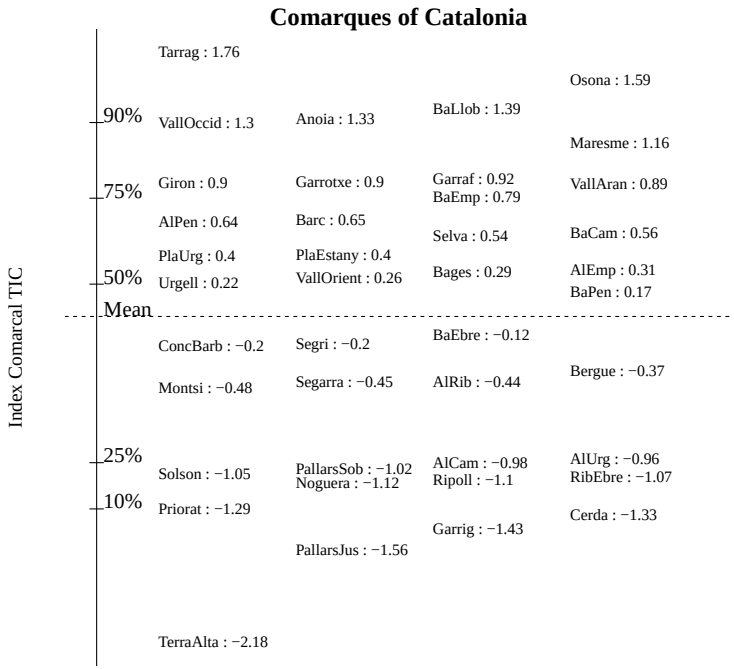


Figure: Index Comarcal de TIC (F1). CFA: 2008 data

Index Comarcal de TIC (F1). CFA: 2008 data, using Mplus

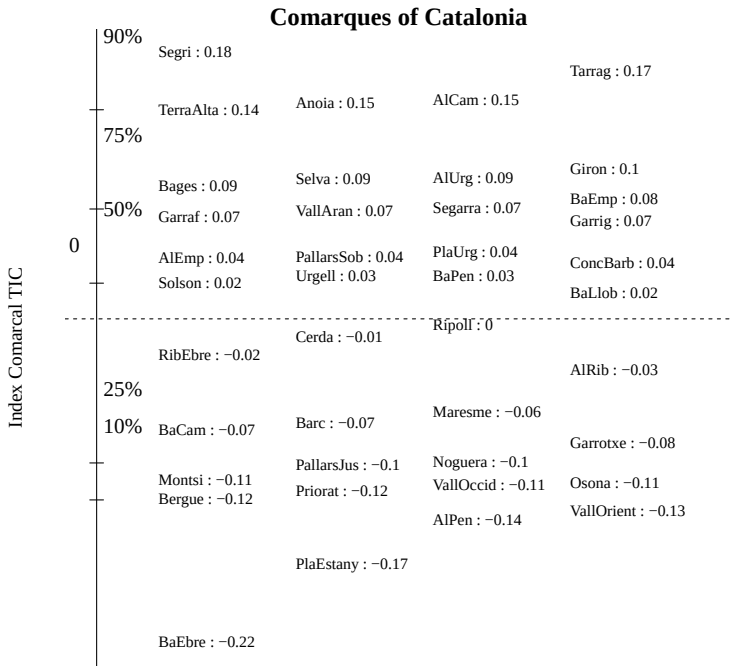


Figure: Index Comarcal de TIC (F1). CFA: 2008 data

CFA: 2010 data

Table: Correlation of variables and factors

	F1	F2
Equip	0.73	-0.07
Mobile	0.26	-0.30
knowPC	0.94	0.00
knowInt	0.99	-0.02
FrecPC	0.96	-0.08
FrecInt	0.98	-0.05
Secbuy	0.54	0.84
Secbank	0.57	0.49
econ	0.90	-0.05
social	0.98	0.04
admin	0.77	0.22

```
CHI-SQUARE = 53.897, 40 DEGREES OF FREEDOM PROBABILITY VALUE 0.06998
Significant free error correlations:  6,2
```

TIC-F1 in 2010

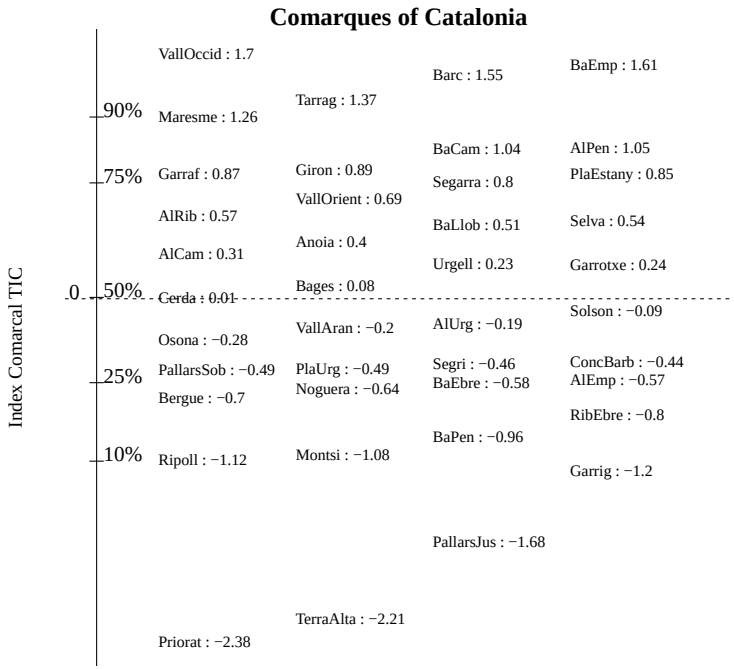


Figure: Index Comarcal de TIC. F1, 2010. The mean is at 0.

Conclusions

- TIC surveys permit summary measures of intensity of TIC in small areas
- A BLUP estimation of the area level parameter has been proposed based on classical factor score estimation of a confirmatory factor model
- The proposed BLUP contains as a particular case the classical sae for single variable analysis
- The MSE of theoretical and empirical versions of the proposed BLUP, as well as other naive predictors, for the area level factor have been compared in a Monte Carlo study
- The proposed EBLUP has been used to obtain index of level of TIC for the "comarcas" of Catalonia, for two repetitive surveys, years 2008 and 2011

conclusions (cont.)

- In contrast to the naive exploratory PCA, our method allows comparing variation across years of area levels within the frame of the same model. Invariance of model is assessed using a classical chi-square goodness of fit test.
- We have disentangle a common ground between sae and structural equation model for two level data. This common ground is worth to be exploited (since user friendly software for SEM is widely developed, e.g. EQS, LISREL, Mplus, CALIS of SAS, Stata, . . .)
- Our approach should be or relevance also to other areas of official statistics, where an area index needs to be extracted from survey data.