Combining sample surveys in small area estimation: a way to improve accuracy in regional statistics.

Alex Costa[1], Albert Satorra[2], Eva Ventura[2]

[1] Institut Estadístic de Catalunya (Idescat), acosta@idescat.net

[2] Universitat Pompeu Fabra, albert.satorra@upf.edu,
eva.ventura@upf.edu

Collaborating researchers:

Maribel Garcia (Idescat)

Xavier López (Idescat)

Motivation

MOTHER SURVEY (MS)	COMPLEMENTARY AREA SURVEY (CAS)
Complex questionnaire	Light questionnaire
Complex interview system PAPI: Paper and Pencil Interview	Easy interview system. CATI: Computer Assisted Telephone Interview
Large overall sample	Large sample in area of interest
Many areas	Not (always) many areas
Basic topics: general interest Expensive	Basic topics: proxy basic variable Specific topics: local interest Inexpensive
Source: INE Results: Spanish and regional basic results	Source: Regional Institute Results: regional basic results from MS plus some area information

A specific context

- MS: Active Population Survey (EPA).
  - Variable of interest: unemployment rate (men, women, and overall)
  - Quarterly data
- CAS: Sociological Research Center Survey (CIS)
  - Proxy of variable of interest: unemployment rate from self-perceived labor status.
  - Monthly data

The small areas

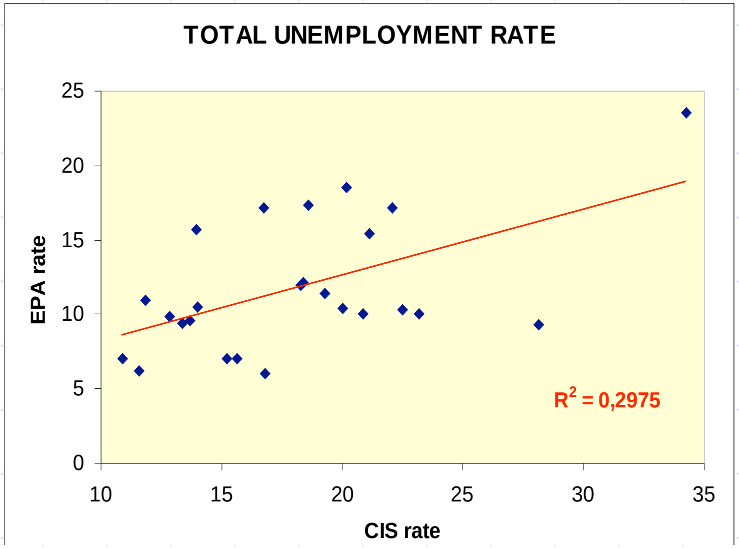
Small area	*EPA survey*		*CIS survey*	
Almería - Granada	4,864	3.29%	324	4.34%
Málaga	3,441	2.33%	235	3.15%
Cádiz - Huelva	5,287	3.58%	246	3.29%
Córdoba - Jaén	6,640	4.50%	237	3.17%
Sevilla	6,411	4.34%	248	3.32%
Aragón	6,589	4.46%	232	3.11%
Asturias	4,522	3.06%	218	2.92%
Baleares	3,539	2.40%	141	1.89%
Canarias	7,748	5.25%	297	3.98%
Cantabria	3,578	2.42%	102	1.37%
Albacete - C.Real	4,971	3.37%	150	2.01%
Cuenca - Guadalajara - Toledo	6,253	4.23%	173	2.32%
Castilla - León	15,143	10.25%	491	6.58%
Barcelona	7,448	5.04%	919	12.31%
Gerona - Lérida - Tarragona	7,721	5.23%	259	3.47%
Alicante - Castellón	6,405	4.34%	345	4.62%
Valencia	5,858	3.97%	393	5.26%
Extremadura	6,167	4.18%	201	2.69%
La Coruña	3,472	2.35%	241	3.23%
Lugo - Orense - Pontevedra	6,921	4.69%	293	3.92%
Madrid	7,765	5.26%	966	12.94%
Murcia	4,043	2.74%	198	2.65%
Navarra - Rioja	5,362	3.63%	151	2.02%
Álava - Guipúzcoa	4,489	3.04%	194	2.60%
Vizcaya	3,037	2.06%	212	2.84%
TOTAL	147,674	100	7466	100

- 50 provinces, its size can be very small or even zero (CIS).
- Group into 25 areas, according to their geographical proximity and the similarity of their labor markets

EPA for a given quarter (4th quarter of 2003) and CIS agregated over three months of the same period

Association at the “population” level

	Unemployment rate					
	EPA			CIS		
Small area	TOTAL	MEN	WOMEN	TOTAL	MEN	WOMEN
Almería -Granada	15.65	10.22	23.38	13.96	12.92	16.30
Málaga	17.33	13.68	23.08	18.59	11.90	29.13
Cádiz -Huelva	23.51	18.37	31.97	34.24	32.35	38.98
Córdoba -Jaén	18.56	12.83	28.27	20.19	16.53	24.37
Sevilla	17.19	12.80	24.01	16.74	11.47	25.29
Aragón	6.20	3.71	9.94	11.58	7.60	18.26
Asturias	10.03	7.00	14.35	23.20	12.31	36.59
Baleares	9.38	8.50	10.59	13.38	6.41	24.65
Canarias	12.10	9.37	16.10	18.37	9.20	33.08
Cantabria	10.32	8.06	13.77	22.52	11.50	38.62
Albacete -C.Real	9.28	4.80	16.59	28.14	18.89	45.28
Cuenca -Guadalajara -Toledo	9.89	5.58	17.44	12.86	7.44	22.12
Castilla -León	10.91	6.09	18.37	11.85	9.50	16.80
Barcelona	9.54	7.37	12.47	13.69	11.44	16.44
Gerona -Lérida -Tarragona	7.00	5.09	9.62	10.90	5.55	18.04
Alicante -Castellón	10.38	8.16	13.67	20.04	18.07	23.28
Valencia	10.08	7.20	14.20	20.84	12.85	31.55
Extremadura	17.11	12.51	24.75	22.06	13.41	40.97
La Coruña	15.44	10.14	22.17	21.11	15.35	29.19
Lugo-Orense -Pontevedra	11.99	7.73	17.55	18.26	15.98	21.83
Madrid	7.00	5.47	9.08	15.62	10.51	21.20
Murcia	10.49	7.09	15.87	14.03	12.16	18.44
Navarra -Rioja	5.99	4.36	8.45	16.81	6.29	30.56
Álava -Guipúzcoa	7.06	5.48	9.28	15.21	6.82	26.89
Vizcaya	11.43	10.06	13.28	19.30	17.21	21.47



Positive association

Borrowing strength from CAS and small area neighbors

CAS based Composite	$\hat{\theta}_k(CBC) = \phi_k \hat{\theta}_k(CAS) + (1-\phi_k)\hat{\theta}_k$

$\hat{\theta}_k(CAS)$ is the fitted value from the OLS (current data) model or other regression alternatives.

Design of the Monte Carlo study

- Population:
  - whole EPA for a given quarter (N= 147.600 approx.).
  - Whole CIS for a three months. (N= 2550 approx. in each month)
- i.i.d. sampling within each of the 25 areas
  - EPA, area samples of size n
    - EPA, area sample of size n*(1+*r*) for computing $\hat{\theta}_k(r)$
  - CIS, area samples of size N
- Design variation:
  - Sample size of the small areas (n): average area sample size 100, 200, 400, 500, 1000.
  - Increasing factor *r* : 10%, 25%, 50% and 100%.
- Number of replications in each cell: 1000
- For each set of 1000 replications, calculate RRMSE for the different estimators

$$RRMSE_k = \frac{\sqrt{\sum_{t=1}^{1000} \left(\hat{\theta}_k - \theta_k\right)^2 / 1000}}{\theta_k}$$

A benchmark estimator: composite CAS fixed effects

- Collects maximum information from the CAS (incorporation of area effects, fixed effects regression with historical data)
- This provides a benchmark for (theoretical) improvement when using CAS.
- Another benchmark is provided by the estimator based just on current data of MS

Using historical data (fixed effects model)

Fixed effects:

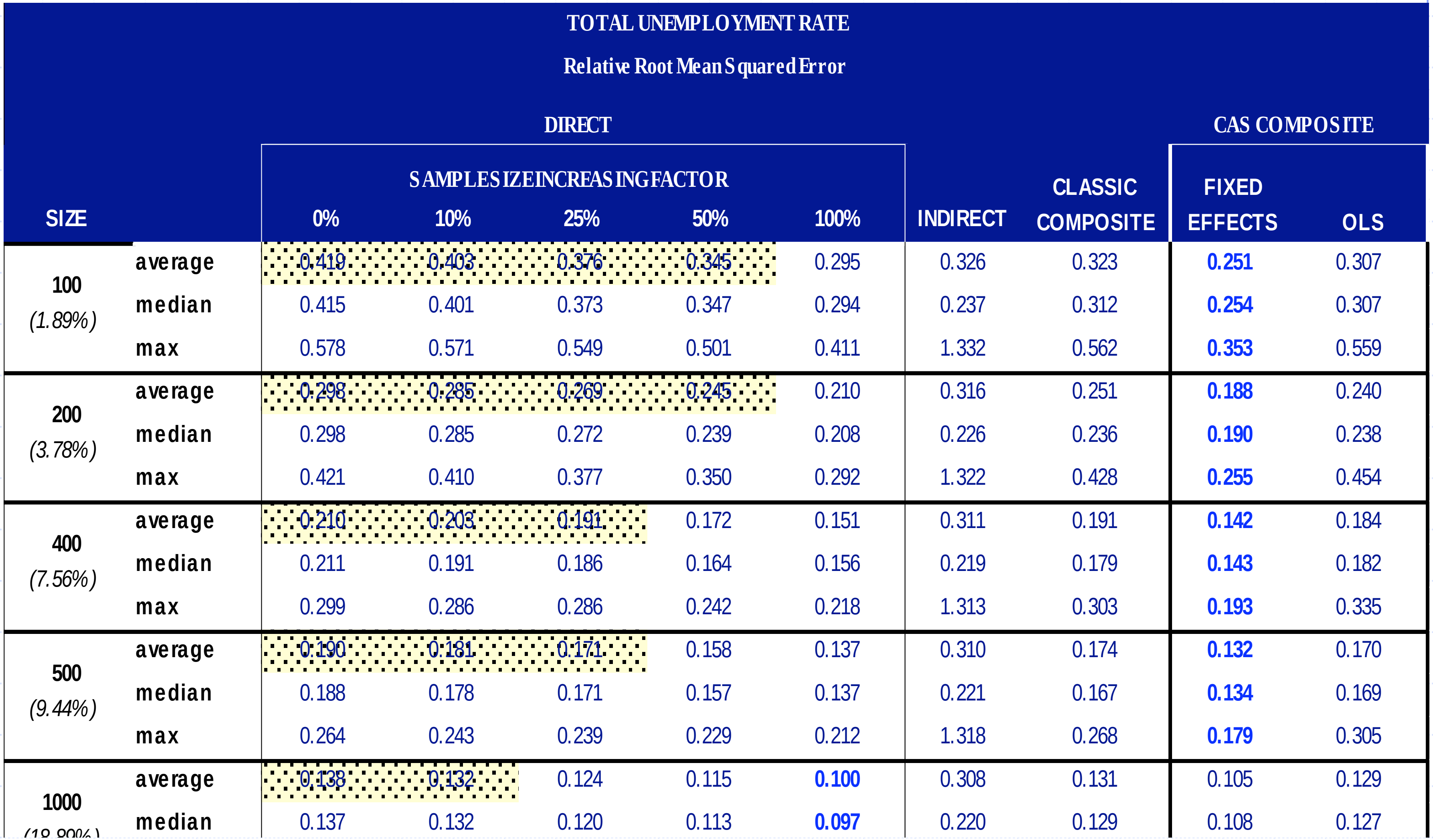
$$\text{EPA rate}_{\text{ti}} = \alpha + \beta\,\text{CIS rate}_{\text{ti}} + u_i + \varepsilon_{ti}$$

```
Fixed -effects (within) regression                  Number      of obs      =       300
Group variable (i): agrup                          Number of groups    =        25

R-sq:  within  = 0.0366                            Obs per group: min =        12
       between = 0.6869                                           avg =            12.0
       overall = 0.4149                                           max =        12

                                                   F(1,274)            =     10.40
corr(u_i, Xb)  = 0.6179                            Prob > F            =    0.0014

----------- -------------------------------------------------------------
     Txdes_t |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
------------- +------------------------------------------------------------
     txdes_t |   .0596874   .0   185123     3.22   0.001     .0232429    .0961319
       _cons |   10.41935   .3352615    31.08   0.000    9.759338    11.07937
------------- +------------------------------------------------------------
     sigma_u |  4.3638425
     sigma_e |  1.31263  17
         rho |  .91702823   (fraction of variance due to u_i)
------------------------------------------------------------------------
F test that all u_i=0:     F(24, 274) =    81.99              Prob > F = 0.0000
```



TOTAL UNEMPLOYMENT RATE

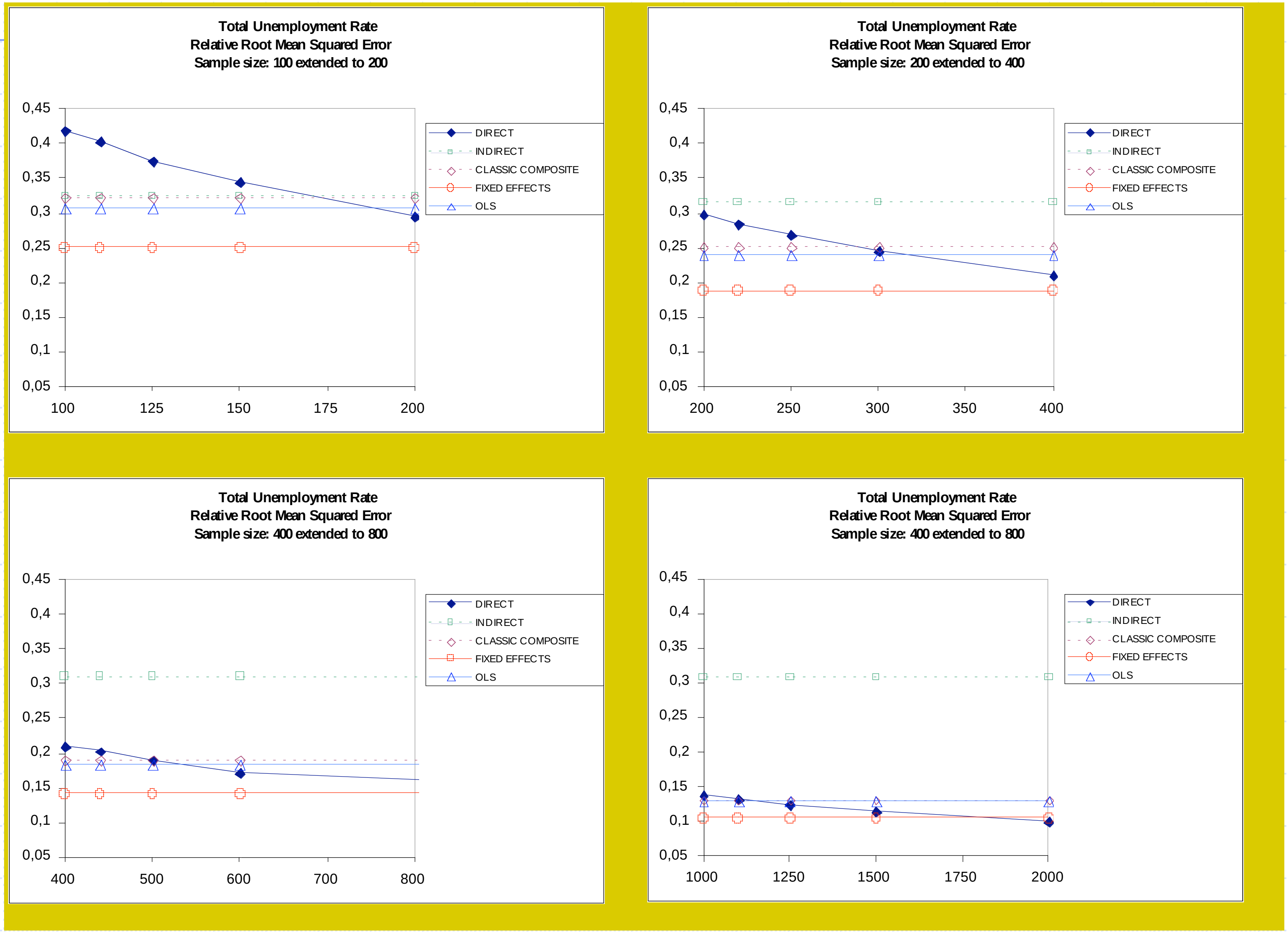
Relative Root Mean Squared Error

		DIRECT							CAS COMPOSITE	
		SAMPLE SIZE INCREASING FACTOR								
SIZE		0%	10%	25%	50%	100%	INDIRECT	CLASSIC COMPOSITE	FIXED EFFECTS	OLS
100 *(1.89%)*	average	0.419	0.403	0.376	0.345	0.295	0.326	0.323	**0.251**	0.307
	median	0.415	0.401	0.373	0.347	0.294	0.237	0.312	**0.254**	0.307
	max	0.578	0.571	0.549	0.501	0.411	1.332	0.562	**0.353**	0.559
200 *(3.78%)*	average	0.298	0.285	0.269	0.245	0.210	0.316	0.251	**0.188**	0.240
	median	0.298	0.285	0.272	0.239	0.208	0.226	0.236	**0.190**	0.238
	max	0.421	0.410	0.377	0.350	0.292	1.322	0.428	**0.255**	0.454
400 *(7.56%)*	average	0.210	0.203	0.191	0.172	0.151	0.311	0.191	**0.142**	0.184
	median	0.211	0.191	0.186	0.164	0.156	0.219	0.179	**0.143**	0.182
	max	0.299	0.286	0.286	0.242	0.218	1.313	0.303	**0.193**	0.335
500 *(9.44%)*	average	0.190	0.181	0.171	0.158	0.137	0.310	0.174	**0.132**	0.170
	median	0.188	0.178	0.171	0.157	0.137	0.221	0.167	**0.134**	0.169
	max	0.264	0.243	0.239	0.229	0.212	1.318	0.268	**0.179**	0.305
1000 *(18.89%)*	average	0.138	0.132	0.124	0.115	**0.100**	0.308	0.131	0.105	0.129
	median	0.137	0.132	0.120	0.113	**0.097**	0.220	0.129	0.108	0.127

Light blue: minimum value

Shaded: Direct estimators outperformed (in average) by both CAS composites.

RRMSE vs. sample size expansion for different sample sizes



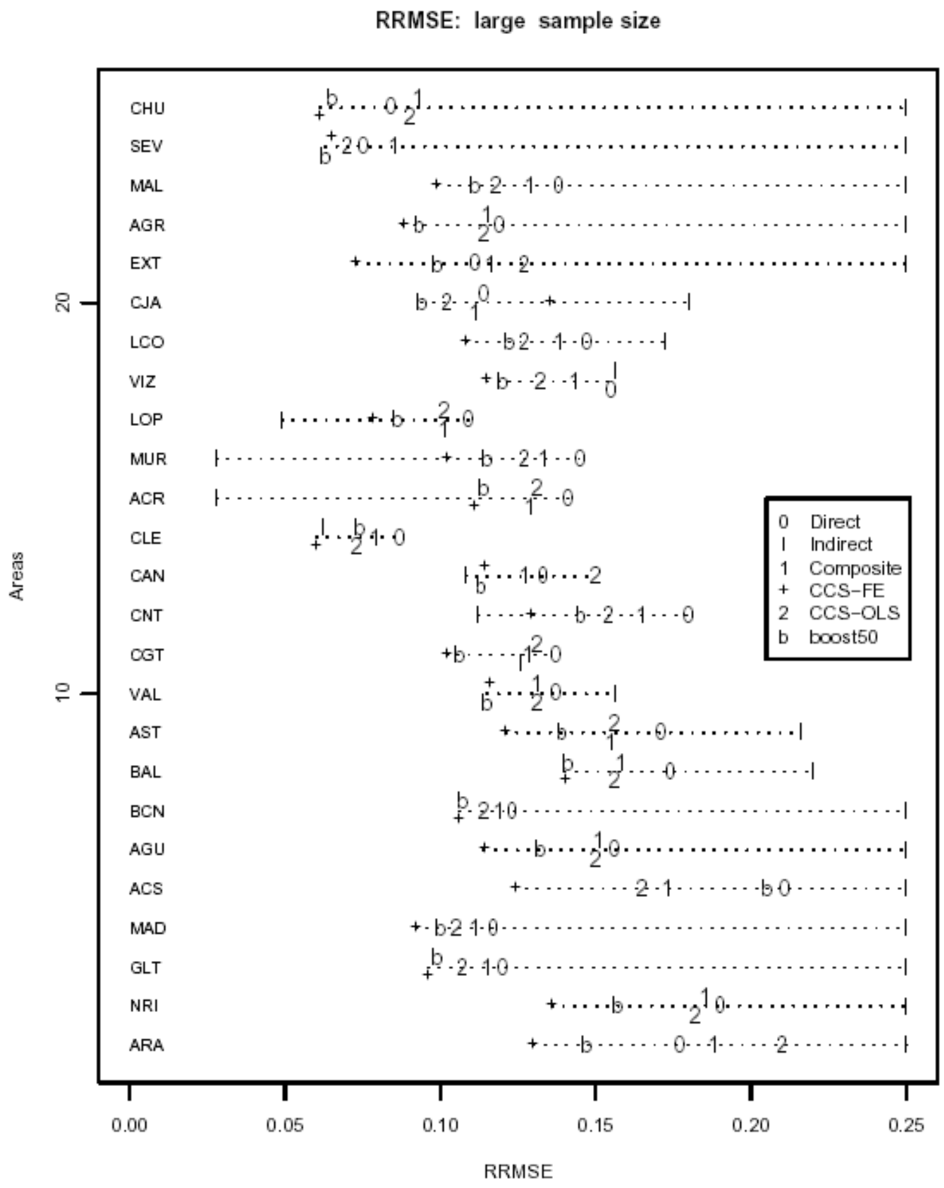


Figure 1: RRMSEs for the areas and for estimators in the case of large sample (average small area sample size is 1000). The areas have been ordered in increasing order of magnitude of their rate of unemployment. Values of RRMSE have been truncated at 0.25.

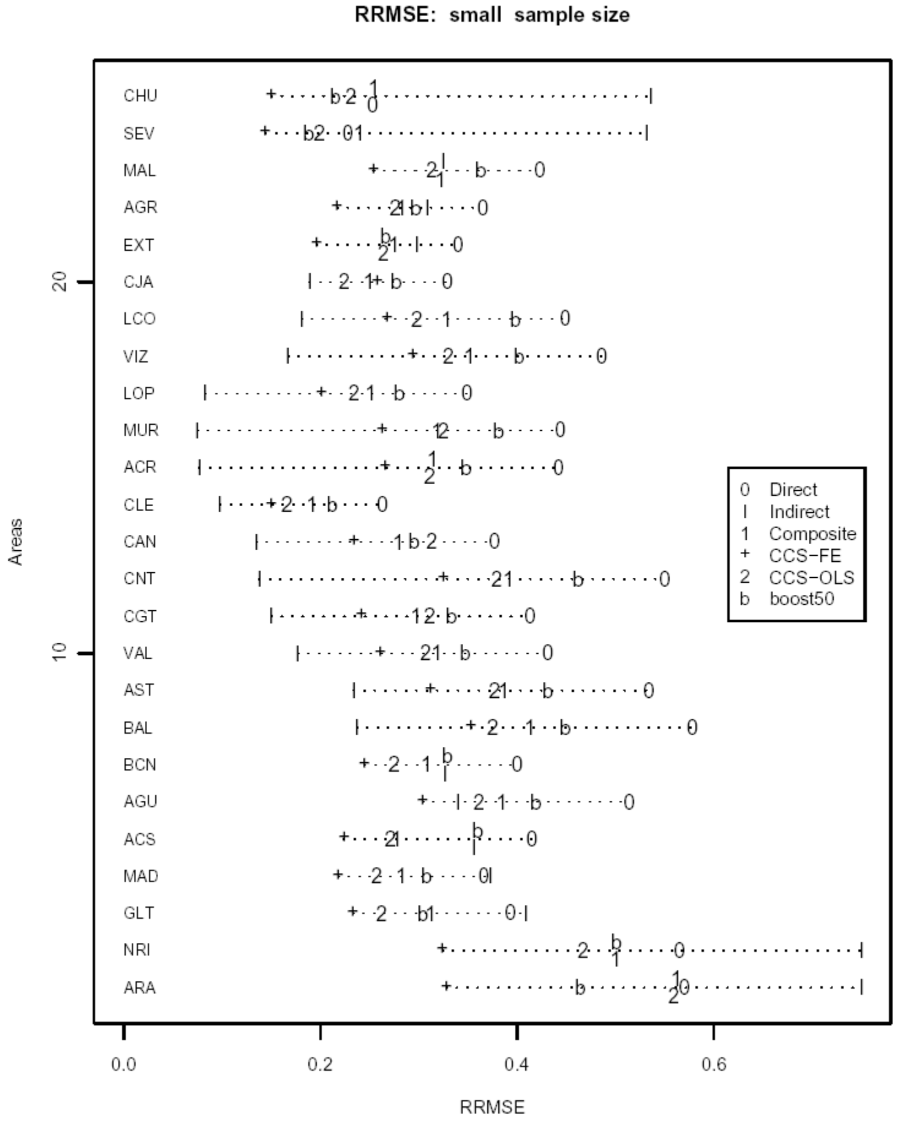


Figure 2: RRMSEs for the areas and for estimators in the case of small sample (average small area sample size is 100). The areas have been ordered in increasing order of magnitude of their rate of unemployment. Values of RRMSE have been truncated at 0.75.

Results

- Without using CAS
  - We corroborate that composite estimators outperform direct estimation in almost all cases (in terms of across areas average, median, etc.)
  - Indirect estimators are a good alternative only in case the small area sample size is very small.
- Using CAS
  - CAS based composite estimators outperform the ones that do not use auxiliary information.
    - As expected, OLS composite (based on sample data at one point time) is outperformed by the benchmark estimator based on fixed effects regression.
    - In some settings the OLS composite outperforms $\hat{\theta}_k(r)$, even for the largest value of *r.* These settings are determined by the sample size of the MS in the area of interest.
    - Only for very large samples (n = 1000) the CAS composite does not compite with $\hat{\theta}_k(r)$

...results

- An important part of the gains displayed by the benchmark estimator (using fixed effects regression) are attained by the simple OLS CAS Based Composite.
- Variation in population characteristics (male, female, total) do not lead to relevant changes in the results.
  - CAS based composite does slightly better (the values of *r* increase) in the case of the unemployment rate of women.
  - In other contexts more pronounced changes could occur.

... results

- In specifically,

  for moderate area sample sizes (say 200), the simplest CAS based composite (OLS) competes with an increase of sample size of 50%.

  This in the context of the estimation of the Spanish unemployment rates.

Costa, A., A. Satorra, & E. Ventura (2006) "Improving small area estimation by combining surveys: new perspectives in regional statistics", *WP, Department of Economics and Business, UPF.*