First EASR Conference

18-22 July 2005

BARCELONA

Variance estimation with an emphasis on small area estimation (SID:50)

Wednesday, July 20; 11:30 to 13:30

Small Area Estimation Under Auxiliary Sample Information

Alex Costa, Albert Satorra, Eva Ventura

Collaborating researchers:

Maribel Garcia

Xavier López

Survey description

- EPA:Spanish Labor Force Survey, quarterly, n ≈ 147,500 .
- CIS: survey by the Spanish Center for Survey Research, light survey of wide spectrum. Monthly, n ≈ 7400.
  - One question on self-perceived work status
- Need of estimates at the small area level (50 provinces, 17 autonomous communities…)

Estimation of an small area parameter

- To estimate small area parameters more accurately, one obvious option is to increase sample size for the area.
- In the case of the EPA survey, this is a highly costly option.

Borrowing strength

- From neighboring areas. This is the classical small area approach (shrinkage to the overall mean).
- From neighboring areas plus auxiliary sample data (CIS). This is what we address in this paper.

What follows…

- Details of the population context
- Synthetic estimators
- Composite estimators
- Monte Carlo Simulations
- Results
- Conclusion

Small area definition and size

Small area	*EPA survey*		*CIS survey*	
Almería-Granada	4,775	3,24%	229	3,07%
Málaga	3,098	2,10%	258	3,46%
Cádiz-Huelva	5,261	3,57%	242	3,24%
Córdoba-Jaén	6,589	4,47%	286	3,83%
Sevilla	6,549	4,44%	276	3,70%
Aragón	6,726	4,56%	231	3,10%
Asturias	4,507	3,06%	218	2,92%
Baleares	3,306	2,24%	138	1,85%
Canarias	7,401	5,02%	296	3,97%
Cantabria	3,342	2,27%	101	1,35%
Albacete-C.Real	4,868	3,30%	185	2,48%
Cuenca-Guadalajara-Toledo	6,235	4,23%	137	1,84%
Castilla-León	15,289	10,37%	490	6,57%
Barcelona	7,237	4,91%	934	12,52%
Gerona-Lérida-Tarragona	8,121	5,51%	251	3,36%
Alicante-Castellón	6,422	4,35%	314	4,21%
Valencia	5,837	3,96%	420	5,63%
Extremadura	6,100	4,14%	201	2,69%
La Coruña	3,444	2,34%	216	2,89%
Lugo-Orense-Pontevedra	6,954	4,72%	313	4,19%
Madrid	8,184	5,55%	971	13,01%
Murcia	4,262	2,89%	197	2,64%
Navarra-Rioja	5,14	3,49%	148	1,98%
Álava-Guipúzcoa	4,578	3,10%	189	2,53%
Vizcaya	3,241	2,20%	221	2,96%
TOTAL	147,466	100	7,462	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
- Objective: combining the information from the sources

Actual values of unemployment rates: EPA vs. CIS, year 2001.

	Unemployment rate					
	EPA			CIS		
Small area	TOTAL	MEN	WOMEN	TOTAL	MEN	WOMEN
Almería-Granada	15.70	10.98	23.42	13.94	5.99	34.98
Málaga	16.09	13.89	19.97	22.50	13.04	36.57
Cádiz-Huelva	23.47	17.37	34.05	25.25	12.69	51.01
Córdoba-Jaén	13.29	9.76	19.08	18.72	13.78	27.05
Sevilla	23.27	18.07	31.64	37.10	23.79	54.97
Aragón	4.71	3.12	7.34	11.71	7.52	18.21
Asturias	8.99	5.70	14.45	16.18	11.48	21.35
Baleares	8.96	5.97	13.72	14.76	9.85	21.03
Canarias	9.88	7.49	13.66	21.35	16.54	28.22
Cantabria	9.84	6.10	15.50	12.19	6.06	20.83
Albacete-C.Real	10.78	6.83	18.24	21.19	21.45	20.45
Cuenca-Guadalajara-Toledo	9.72	5.60	17.54	18.21	13.06	29.57
Castilla-León	10.33	6.40	16.96	14.09	8.80	23.61
Barcelona	8.33	6.60	10.74	13.30	9.26	18.97
Gerona-Lérida-Tarragona	7.83	4.75	12.34	7.54	6.74	8.64
Alicante-Castellón	8.15	6.07	11.52	12.43	8.62	19.86
Valencia	9.46	6.88	13.31	16.79	12.13	24.14
Extremadura	15.45	11.03	23.49	13.81	10.42	21.59
La Coruña	13.16	10.15	17.43	20.09	8.97	35.00
Lugo-Orense-Pontevedra	11.40	7.52	16.66	12.03	10.70	14.64
Madrid	8.04	5.50	11.67	10.33	7.67	14.21
Murcia	11.06	7.57	17.11	8.71	3.04	17.78
Navarra-Rioja	5.33	3.76	7.90	9.14	12.96	2.78
Álava-Guipúzcoa	8.24	5.26	12.68	13.47	16.01	9.87
Vizcaya	12.91	8.49	19.06	15.30	13.95	17.49

An approximate linear relationship

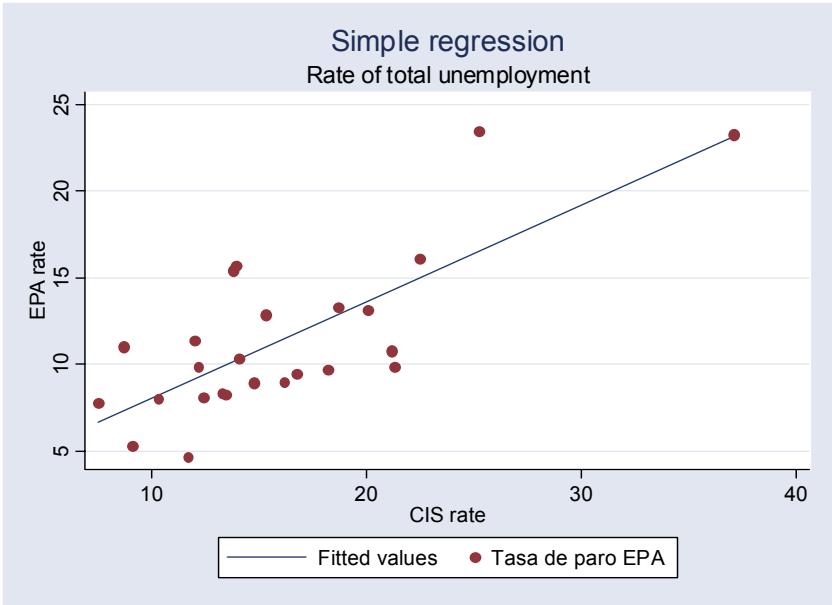
CIS: we use quarterly average of monthly rates

A naïve synthetic estimate: simple regression at one point time data

Simple OLS regression.

Total rate of unemployment, EPA regressed on CIS.

Year 2001, 1st quarter




```
Number of obs       =      25

          R-squared =  0.5755


Txdes_t   Coef.     Std. Err.      t    P>t       [95% Conf.Interval]

txdes_t   .5567852  .0997191     5.58   0.000     .3505005  .7630698
_cons     2.464651  1.710594     1.44   0.163     -1.073983 6.003284
```

Regression estimator (synthetic estimator)

- Regression estimator based on one time point data
- Regression estimator based on historical data, EPA from 3 years.

Random effects model

We integrate information from all the years

Random effects:

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

```
Random-effects GLS 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

Random effects u_i ~ Gaussian                   Wald chi2(1)       =     16.86
corr(u_i, X)       = 0 (assumed)                Prob > chi2        =    0.0000

------------------------------------------------------------------------------
       ^ t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
           +----------------------------------------------------------------
           .0798742   .0194523     4.11   0.000     .0417483        .118
            ` 06323   .6580147    15.29   0.000     8.773546    11.35292
               ------------------------------------------------------------


                          ^tion of variance due to u_i)
                             ---------------------------------------
```

u_i : area effect

Fixed effects model

Fixed effects:

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

```
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

------------------------------------------------------------------------------
      ′des_t |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
             |   .0596874   .0185123     3.22   0.001     .0232429    .0961319
             |   10.41935   .3352615    31.08   0.000     9.759338    11.07937
             ------------------------------------------------------------------
                  ?8425
                     ′7
                                (fraction of variance due to u_i)
                                 ------------------------------------------
                                     274) =    81.99             Prob > F = 0.0000
```

Small area effects: fixed vs. random

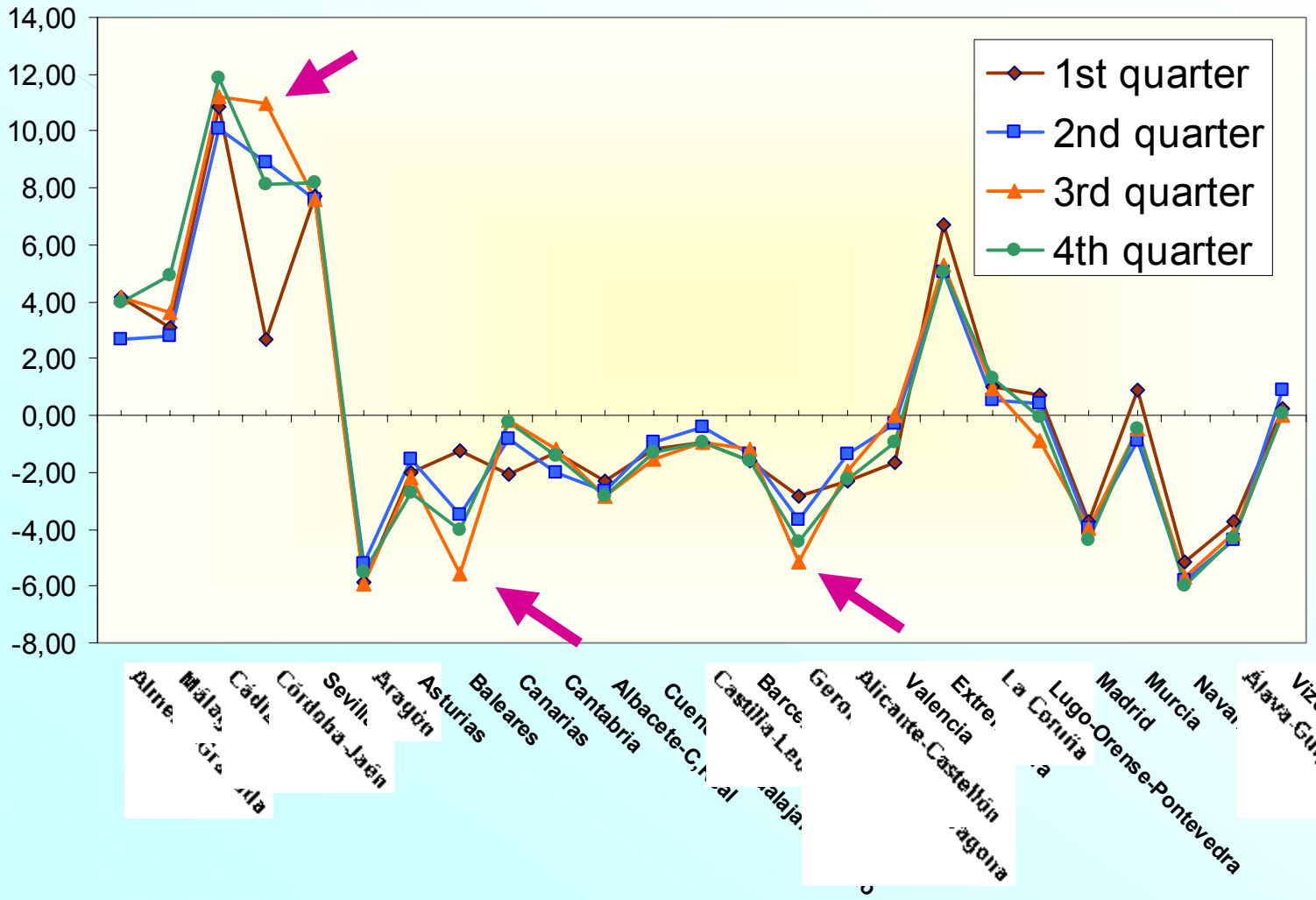
Small areas	Fixed	Random
Almería-Granada	3.67638	3.485407
Málaga	3.69148	3.50665
Cádiz-Huelva	11.07711	10.64098
Córdoba-Jaén	7.772888	7.459998
Sevilla	7.977803	7.665291
Aragón	-5.625362	-5.410994
Asturias	-2.138158	-2.18439
Baleares	-3.675362	-3.506119
Canarias	-.7510928	-.8526955
Cantabria	-1.488708	-1.395779
Albacete-C.Real	-2.679073	-2.652816
Cuenca-Guadalajara-Toledo	-1.233451	-1.137147
Castilla-León	-.7860277	-.7026417
Barcelona	-1.435128	-1.339818
ona-Lérida-Tarragona	-4.134264	-3.910424
Alicante-Castellón	-1.954283	-1.837342
Valencia	-.7174698	-.7365981
Extremadura	5.45155	5.310777
La Coruña	.9942183	.9397409
Pontevedra	.032394	.0375081
Madrid	-4.049702	-3.888889
Murcia	-.3540572	-.2392511
ioja	-5.735651	-5.513937
oa	-4.19625	-4.067385
	.2802148	.3298753

- Similar magnitude in area effects
- We adopt fixed effects

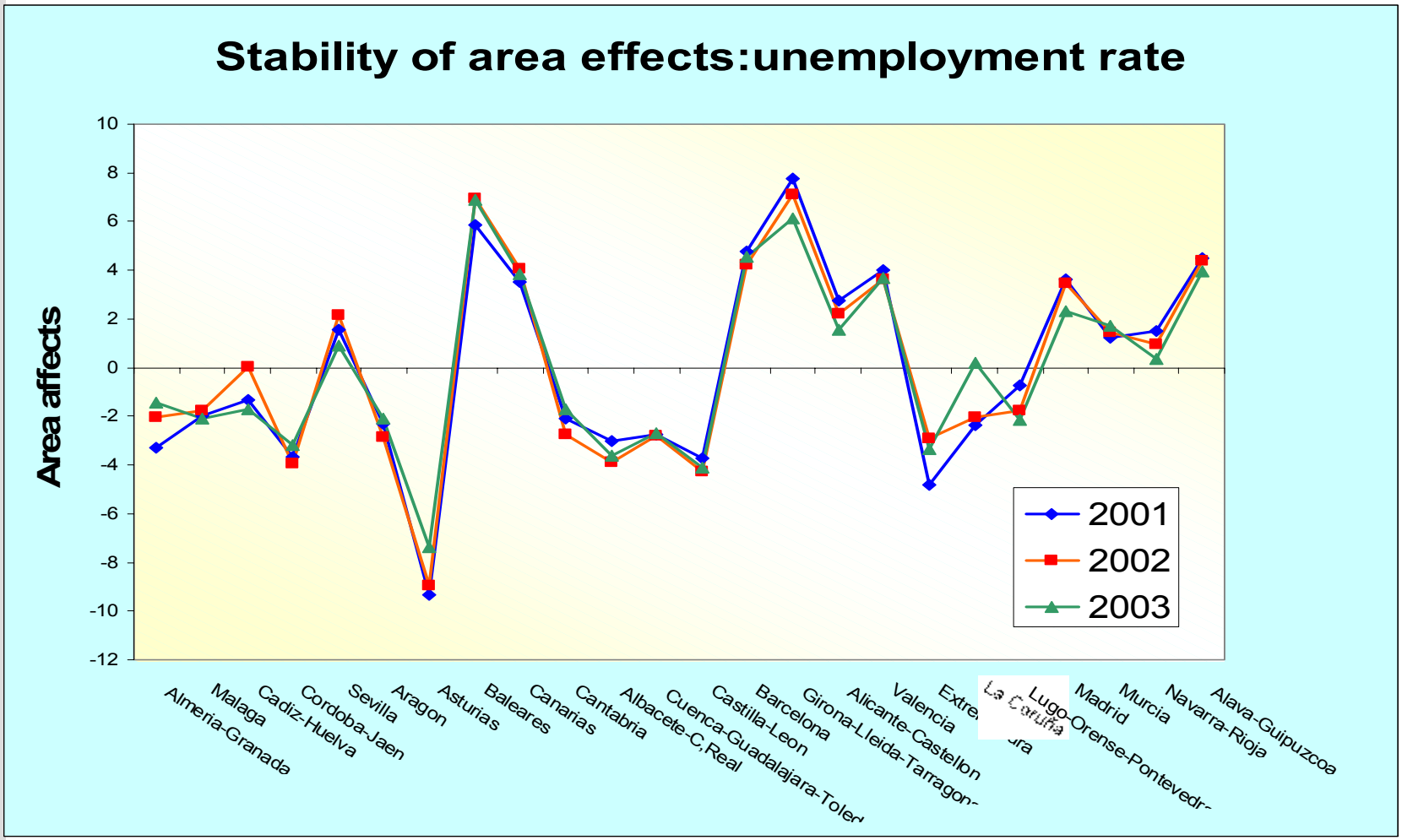
Do time of the year (quarter) matter for small area effects?

Small areas	1st quarter	2nd quarter	3rd quarter	4th quarter
Almería-Granada	4,15	2,68	4,14	3,95
Málaga	3,08	2,78	3,61	4,94
Cádiz-Huelva	10,89	10,09	11,24	11,88
Córdoba-Jaén	2,68	8,90	10,95	8,15
Sevilla	7,69	7,59	7,60	8,18
Aragón	-5,88	-5,24	-5,91	-5,53
Asturias	-2,01	-1,53	-2,18	-2,72
Baleares	-1,27	-3,52	-5,57	-4,04
Canarias	-2,09	-0,81	-0,19	-0,22
Cantabria	-1,28	-2,02	-1,20	-1,40
Albacete-C,Real	-2,32	-2,67	-2,83	-2,86
Cuenca-Guadalajara-Toledo	-1,17	-0,97	-1,55	-1,31
Castilla-León	-0,95	-0,41	-0,94	-0,92
Barcelona	-1,61	-1,34	-1,17	-1,61
Gerona-Lérida-Tarragona	-2,86	-3,69	-5,13	-4,42
Alicante-Castellón	-2,32	-1,35	-1,93	-2,24
Valencia	-1,68	-0,28	0,00	-0,97
Extremadura	6,72	5,03	5,26	5,06
La Coruña	1,04	0,53	0,96	1,28
Lugo-Orense-Pontevedra	0,71	0,42	-0,89	-0,07
Madrid	-3,74	-3,94	-3,99	-4,41
Murcia	0,89	-0,90	-0,44	-0,49
Navarra-Rioja	-5,18	-5,82	-5,69	-5,97
Álava-Guipúzcoa	-3,74	-4,38	-4,15	-4,31
Vizcaya	0,25	0,87	-0,01	0,05

Stability of area effects: unemployment rate



Do years matter?



Estimators considered

[illegible]	$\hat{\bar{\upsilon}}_k$ for the small area [illegible] whole population
Classical composite	$\phi_k \hat{\theta} + (1-\phi_k)\hat{\theta}_k$
2 Regression synthetic	$\hat{\theta}_k^1(reg) = \hat{\alpha}_1 + \hat{\beta}_1 \hat{\theta}_k(C) + \hat{u}_k$ $\hat{\theta}^2(reg) = \hat{\alpha}_2 + \hat{\beta}_2 \hat{\theta}_k(C)$
2 Regression composite	$\phi_k \hat{\theta}_k(reg) + (1-\phi_k)\hat{\theta}_k$

Weights of composite estimators

- Theoretical weight (simplified, minimizes MSE)

$$\phi_k = \frac{\sigma_k^2}{\left(\theta_k - \theta_k^a\right)^2 + \sigma_k{}^2}$$

Bias term

Variance term

- Estimated weight (stability)

$$\hat{\phi}_k = \frac{\hat{\bar{\sigma}}^2}{\hat{\bar{\sigma}}^2 + b^2}$$

$$\hat{\bar{\sigma}}^2 = \frac{\sum_{k=1}^{K}(n_k - 1)\hat{\sigma}_k^2}{(n-K)}$$

$$b^2 = \frac{1}{K}\sum_{k=1}^{K}\left(\hat{\theta}_k - \hat{\theta}_k^a\right)^2$$

Synthetic estimator

Variance of direct

Direct

Monte Carlo setup

- Population: the actual EPA in 1st quarter of 2001 (147,466 units) divided into 25 small areas.
- **iid sampling:** total sample sizes of 2,500, 5,000, 10,000, 12,500 and 73,500 (proportional for each area)
- Estimators
  - Direct
  - Indirect with no complementary information
  - Classical composite with no complementary information
  - Composite regression (reg1 and reg2)
  - Direct, with sample augmented. Level of augmentation considered: 10, 25, 50 and 100% of initial sample.

Evaluation of estimators: relative root mean squared error

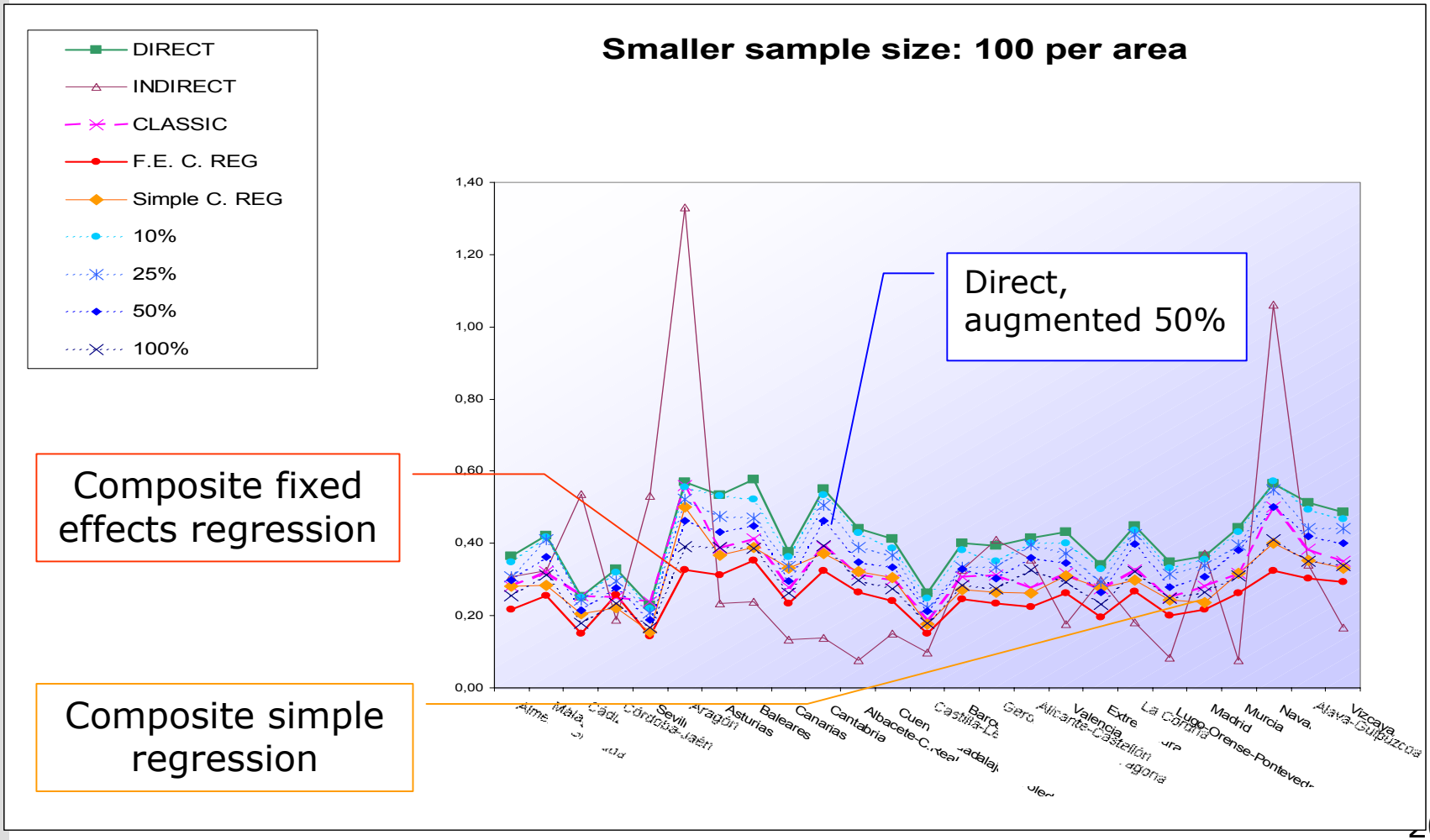
- RRMSE for Monte Carlo simulation (1000 replications)

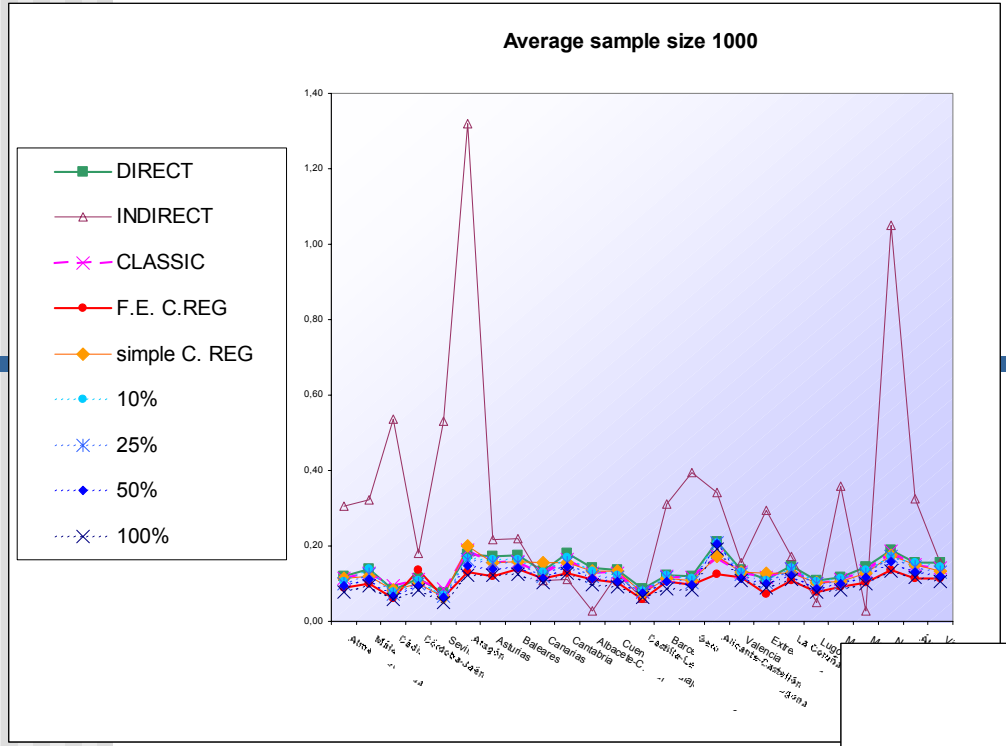
$$RRMSE_k = \frac{\sqrt{\sum_{t=1}^{1000}\left(\hat{\theta}_k - \theta_k\right)^2 \Big/ 1000}}{\theta_k}$$

Estimate

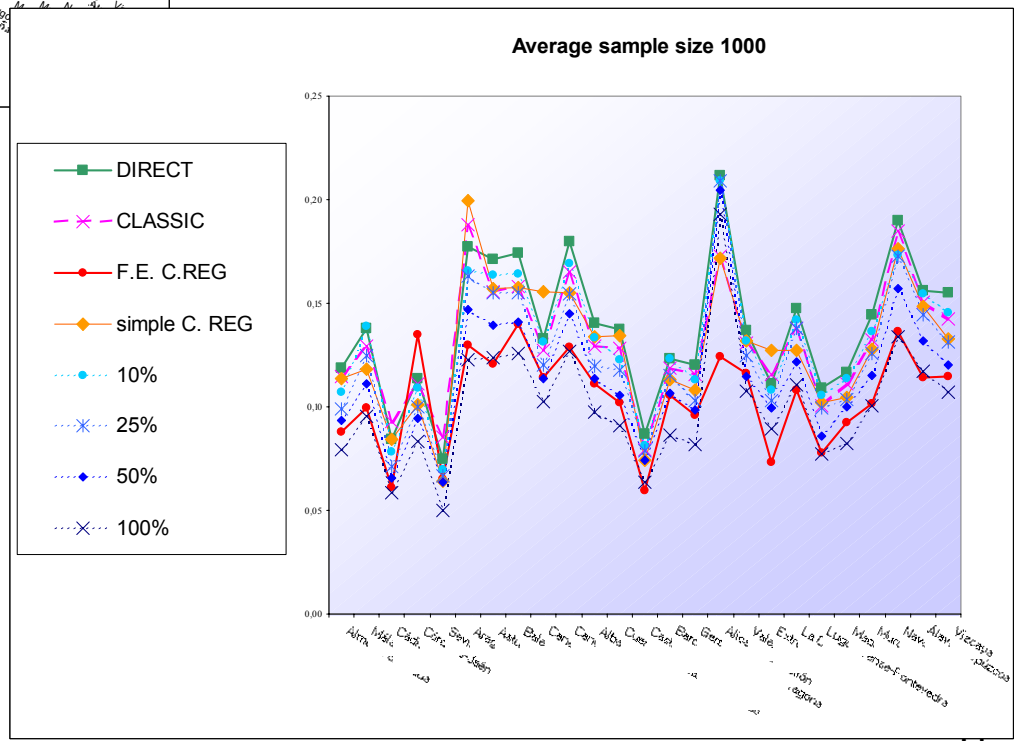
Area parameter

For small samples





For large sample



TOTAL UNEMPLOYMENT RATE
Relative Root Mean Squared Error

SAMPLE SIZE		DIRECT	INDIRECT	COMPOSITE CLASSIC	COMPOSITE REG. EFFECTS	COMPOSITE REG. SIMPLE	DIRECT WITH AUGMENTED SAMPLE 10%	DIRECT WITH AUGMENTED SAMPLE 25%	DIRECT WITH AUGMENTED SAMPLE 50%	DIRECT WITH AUGMENTED SAMPLE 100%
100	**average**	0.419	0.326	0.323	0.251	0.299	0.403	0.376	0.345	0.295
(1.89%)	**median**	0.415	0.237	0.312	0.254	0.298	0.401	0.373	0.347	0.294
	max	0.578	1.332	0.562	0.353	0.500	0.571	0.549	0.501	0.411
200	**average**	0.298	0.316	0.251	0.188	0.235	0.285	0.269	0.245	0.210
(3.78%)	**median**	0.298	0.226	0.236	0.190	0.239	0.285	0.272	0.239	0.208
	max	0.421	1.322	0.428	0.255	0.404	0.410	0.377	0.350	0.292
	average	0.210	0.311	0.191	0.142	0.182	0.203	0.191	0.172	0.151
400	**median**	0.211	0.219	0.179	0.143	0.184	0.191	0.186	0.164	0.156
(7.56%)	**max**	0.299	1.313	0.303	0.193	0.302	0.286	0.286	0.242	0.218
500	**average**	0.190	0.310	0.174	0.132	0.169	0.181	0.171	0.158	0.137
(9.44%)	**median**	0.188	0.221	0.167	0.134	0.170	0.178	0.171	0.157	0.137
	max	0.264	1.318	0.268	0.179	0.287	0.243	0.239	0.229	0.212
1000	**average**	0.138	0.308	0.131	0.105	0.129	0.132	0.124	0.115	0.100
(18.88%)	**median**	0.137	0.220	0.129	0.108	0.128	0.132	0.120	0.113	0.097
	max	0.211	1.319	0.188	0.140	0.199	0.209	0.209	0.205	0.193
2949	**average**	0.089	0.306	0.087	0.075	0.087	0.085	0.081	0.075	0.068
(50%)	**median**	0.086	0.218	0.084	0.075	0.085	0.082	0.077	0.072	0.064
	max	0.190	1.318	0.168	0.127	0.165	0.189	0.189	0.183	0.184

Results

- Composite regression (either simple or fixed effects model) does better than direct or classical composite with the initial sample. For small sample size, on average outperforms augmenting the sample (even doubling!).
- Composite regression based on fixed effects model always does very well. Even better than doubling the sample (exception: very large sample). Note that it uses historical information (3 years).
- Indirect based on whole sample with no auxiliary information performs badly on average always.
- Classical composite with no auxiliary information does better than direct but it is clearly outperformed by composite regression.
- The gain of the composite estimate with auxiliary sample decreases as sample size increases.

Concluding remarks

- A “light” survey such as CIS helps to improve accuracy in estimation of small area parameters of EPA.
- The use of auxiliary information (CIS) can be cheaper, and quicker to implement, than augmentation of the main sample (EPA).