ITACOSM 2019
Florence, June 5 2019

Using entropic distance for large area definition in small area estimation methods

Livio Fenga
fenga@istat.it

Fabrizio Solari
solari@istat.it

Italian National Statistical Institute

Presentation's Outline

- Model group definition in small area estimation problems
- Complexity-Invariant Distance for time series
- Experimental study on Italian LFS
- Concluding remarks

Choosing the macro-area

- The macro-area is defined as the set of small areas for which a common model is specified (and fitted). It is connected to the model group concept
- Often NSIs identify macro-areas using preexistent territorial delimitations (e.g., regions, states, etc,), or macro regions which are geographically meaningful (e.g. north, center, south)
- This could be a practical but not always an optimal solution
- In this presentation a solution for the definition of an optimal macro-area for each small area is proposed

Choosing the macro-area

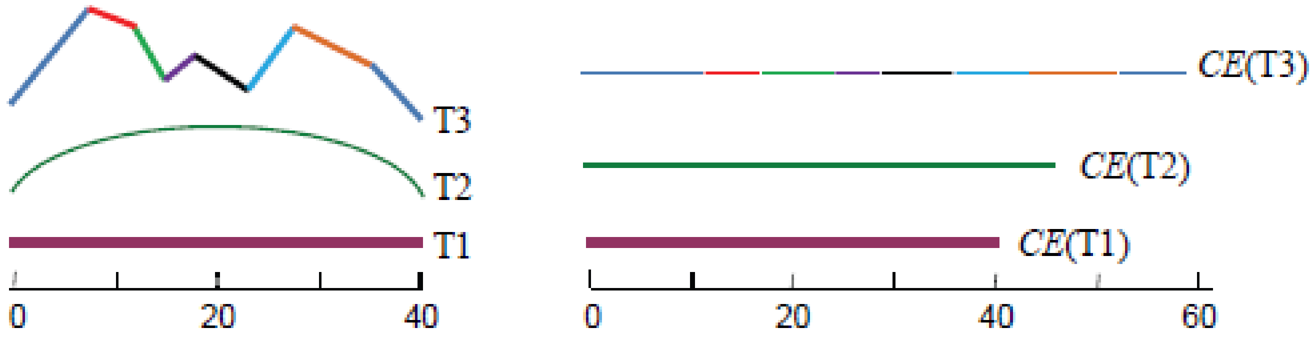
- The goal is to find sets of small areas with similar "behaviour" with respect to the model
- This could be done analyzing either predicted or residual values under the model and including in the same macro-area all the small areas with similar residuals (or predicted) values
- Problems:
  - It is not advisable to use the same data for 1) compute model residuals (or predicted values) to define the macro-areas and 2) fitting the model using the macro-areas defined in 1)
  - likely different macro-areas are expected for different times causing consistent changes in the estimates from time to time

Choosing the macro-area

- When time series data are available, it is more appropriate considering the similarity between residual (or predicted) time series data instead of considering only one point in time data
- Similarities between time series data can be evaluated using the Complexity-Invariant Distance (Batista, Keogh, Tataw & de Souza, 2014)

Complexity-Invariant Distance

- Ideally a time series is "stretched" until it becomes a straight line. As a result of that, a complex time series should result in a longer line than a simple time series



Complexity-Invariant Distance

In the case of the Euclidean distance between two time series Q and C

ED(Q,C)

complexity-invariance is achieved by introducing a correction factor:

CID(Q,C) = ED(QC) X CF(Q,C)

- CF = complexity correction factor
- CF(Q,C) = max(CE(Q), CE(C)) / min max(CE(Q), CE(C))
- CE(·) complexity estimate of time series C

Complexity-Invariant Distance

- CF accounts for differences in the complexities of the time series in order to set apart time series with different complexities
- Under same complexity time series, CID degenerates to the Euclidean distance

Experimental study

- LFS quarterly data from 2004 to 2014
- Direct estimates and sampling variances for employment and unemployment rate at Local Labour Market Area level
- Smoothing of sampling variances
- 221 out of 611 LLMAs are sampled for all the 44 quarters
- Residual and predicted values from a standard area level LM are computed (auxiliary variables: 12 cross-classification of age classes and sex)

Experimental study

- For both predicted and residuals macro-areas are defined using the Complexity-Invariant Distance:
  - for the generic small *d* area an *ad hoc* macro-area is defined including all the areas whose distance from *d* is less than a given treshold
  - a minimum number of 30 small areas is included in each *ad hoc* macro-area

Experimental study

- Comparison of the following model groups:
  - Italy
  - 3 large areas (North, Centre, South)
  - 5 large areas (North-East, North-West, Centre, South, Sicily + Sardinia)
  - ad hoc macro-area for each small area using the complexity-invariant distance for the residual time series
  - ad hoc macro-area for each small area using the complexity-invariant distance for the predicted values time series

Experimental study

- Standard FH model is adopted (Fay & Herriot, 1979):

$\hat{\overline{Y}}_d = \overline{Y}_d + e_d$ (sampling model)

$\overline{Y}_d = \overline{X}_d^T \boldsymbol{\beta} + v_d$ (linking model)

v_d, e_d are independent, $v_d \sim N(0, \sigma_v^2)$, $e_d \sim N(0, \sigma_d^2)$, σ_d^2 is known for all *d*

The EBLUP of $\overline{Y}_d$ is $\tilde{\overline{Y}}_d(\tilde{\sigma}_v^2) = \gamma_d \overline{Y}_d + (1-\gamma_d)\overline{x}_d^T \hat{\beta}(\tilde{\sigma}_v^2)$,

where $\gamma_d = \sigma_v^2 / (\sigma_v^2 + \sigma_d^2)$, $0 \le \gamma_d \le 1$.

Results: employment rate estimation

AARE

model group				
overall country	3 large areas	5 large areas	ad hoc large areas (residuals)	ad hoc large areas (predicted)
0.047150	0.044768	0.046352	0,050176	**0.044499**
(1.053)	(1.000)	(1.035)	(1,121)	**(0.994)**

ASE

model group				
overall country	3 large areas	5 large areas	ad hoc large areas (residuals)	ad hoc large areas (predicted)
0.000477	0.000414	0.000466	0,000559	**0.000403**
(1.152)	(1.000)	(1.125)	(1,349)	**(0.972)**

Results: unemployment rate estimation

AARE

model group				
overall country	3 large areas	5 large areas	ad hoc large areas (residuals)	ad hoc large areas (predicted)
0.276173	0.268203	0.276801	0,285767	**0.258059**
(1.030)	(1.000)	(1.032)	(1,065)	**(0.962)**

ASE

model group				
overall country	3 large areas	5 large areas	ad hoc large areas (residuals)	ad hoc large areas (predicted)
0.000473	0.000455	0.000458	0,000503	**0.000445**
(1.040)	(1.000)	(1.007)	(1,104)	**(0.977)**

Conclusions

- The macro-areas built from the complexity-invariant distance matrix of the predicted values time series outperform the standard way of defining model groups
- Not good results are produced using the complexity-invariant distance matrix of the residuals
- Likely, the residuals are not “white” residuals and some pre-whitening technique should be applied before using them as an input for the complexity-invariant distance matrix

Conclusions and future developments

- Improving model complexity:
  - Introduction of a spatial correlation structure in the model specification (Petrucci & Salvati, 2006; Pratesi & Salvati, 2008)
  - Modelling time series data (Maruenda, Molina & Morales,2013; Rao & Yu, 1994; Singh, Mantel & Thomas, 1991)
- The complexity-invariant distance can be used as an alternative distance matrix between the areas
- Define an automatic way to find an optimal value of the treshold for the complexity-invariant distance

References (for what already done)

Batista, G.E., Keogh, E.J., Tataw, O.M., and de Souza, V.M.A. (2014). CID: an efficient complexity-invariant distance for time series. Data Mining and Knowledge Discovery, 28(3), 634-669.

Fay, R.E., and Herriot, R.A. (1979). Estimation of income from small places: an application of James-Stein procedures to census data. Journal of the American Statistical Association, 74, 269-277.

References (for what to be done!!!)

Marhuenda, Y., Molina, I., and M orales, D. (2013). Small area estimation with spatio-temporal Fay-Herriot models. *Computational Statistics and Data Analysis*, 58, 308-325.

Petrucci, A., and Salvati, N. (2006). Small area estimation for spatial correlation in watershed erosion assessment. *Journal of Agricultural, Biological and Environmental Statistics*, 11, 169-182.

Pratesi, M., and Salvati, N. (2008). Small area estimation: the EBLUP estimator based on spatially correlated random area effects. *Statistical Methods & Applications*, 17, 113-141.

Rao, J.N.K., and Yu, M. (1994). Small area estimation by combining time series and cross-sectionaldata. *Canadian Journal of Statistics,* 22, 511-528.

Singh, A.C., Mantel, H.J., and Thomas, B.W. (1994), Time series EBLUPs for small areas using survey data. *Survey Methodology*, 20, 33–43.

Thanks for your attention!!!