Intensive Courses in the context of the Jean Monnet Chair:

Big data in official statistics

Block 1: Introduction

14 december 2018,

University of Pisa

Jan van den Brakel

Introduction

Official statistics:

1. Repeated sample survey
2. Problems:
   (a) expensive
   (b) not very timely
   (c) non reponse
   (d) response burden
   (e) ...

Big data:

1. Large data sets that are generated as a by-product of processes not directly related to statistical production purposes.
2. Examples of these include:
   (a) time and location of network activity available from mobile phone companies,

(b) social media messages from Twitter and Facebook

(c) internet search behaviour from Google Trends

(d) administrative data like tax registers

Use of Big data in official statistics:

1. Primary data source (see lecture slides by M. Puts)

2. Covariates in small area estimation models or models for now casting

   (a) Area level model (Fay and Herriot, 1979):

   - Uses cross-sectional correlations
   - Avoids matching unstructured big data sources with survey data on the unit level
   - Marchetti et al. (2015) uses mobility of cars tracked with GPS as a covariate for predicitng poverty in a Fay-Herriot model

   (b) Official statistics:

   - Repeated surveys

- Therefore time series models are more appropriate
- For this course we focus on structrual time series models

Outline course:

- Block 2: Introduction structural time series models
- Block 3: Bivariate state space model for nowcasting
- Block 4: Dynamic factor models

References

Fay, R. and Herriot, R. (1979). Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data. *Journal of the American Statistical Association* 74 (366), 269–277.

Marchetti, S., Giusti, C., Pratesi, M., Salvati, N., Giannotti, F., Perdreschi, D., Rinzivillo, Pappalardo, L., and Gabrielli, L. (2015). Small area model-based estimators using Big data sources. *Journal of Official Statistics* 31, 263–281.