Global Strategy
IMPROVING AG-STATISTICS

DEVELOPING ROBUST AND STATISTICALLY BASED METHODS FOR SPATIAL DISAGGREGATION AND FOR INTEGRATION OF VARIOUS KINDS OF GEOGRAPHICAL INFORMATION AND GEO-REFERENCED SURVEY DATA

Monica Pratesi, University of Pisa

BRIEF DESCRIPTION OF THE RESEARCH TOPIC (1)

Integration, **aggregation** and **disaggregation** of multi-sourced spatial data on agri-environmental phenomena and their interoperability have received much attention in the last few years

- Maximising the use of existing data, rather than establishing new collections, avoids additional load on respondents, helps to ensure cost-effectiveness and can improve timeliness
- New official statistics based on analysis of longitudinal and small area data obtained exploiting spatial data: *spatial data integration and aggregation techniques*
- Relevance of high-precision maps in many decision-making processes – lack of needed data: *spatial disaggregation techniques*

BRIEF DESCRIPTION OF THE RESEARCH TOPIC (2)

Statistical quality of data integration, aggregation and disaggregation is a key issue

Definition and measurement of **errors** during the process of data production and analysis under an inferential approach (crucial in developing countries):

- Sources of bias in spatial interpolation and data fusion
- Robustness to outlying observations in fusion and disaggregation
- Protection against change of support problems (MAUP, ecological fallacy)
- Multivariate approach in models to aggregate and disaggregate data
- Modelling and inference on spatial relation

BRIEF DESCRIPTION OF THE RESEARCH TOPIC (3)

Many **difficulties** (other than pure statistical problems) must also be taken into consideration in order to achieve effective integration and interoperability of technical solutions. This is especially true in developing countries.

- the diversity of data providers (standards, interoperability, vertical topology, semantic, reference system, data model, metadata, format, and data quality)
- institutional, social, legal and policy requirements (collaboration, funding model,cultural issues, legislation issue, copyrights, intellectual properties, etc)

LITERATURE REVIEW (1)

Geospatial Data Integration is considered here as the process and the result of geometrically combining two or more different sources of geospatial content (GPSs, GIS, satellite imagery, remote sensing) to facilitate visualization and statistical analysis of the data.

- technical disparities including scale, resolution, compilation standards, source accuracy, registration, sensor characteristics, currency, temporality, or errors; differences in datum, projections, coordinate systems, data models, spatial and temporal resolution, precision, and accuracy; effects of different supports in different sources; **effects of these disparities on geostatistical models**

LITERATURE REVIEW (2)

Data aggregation/fusion is the process of combining information from heterogeneous sources into a single composite picture of the relevant process, such that the composite picture is generally more accurate and complete than that derived from any single source alone.

- The majority of fusion techniques are custom-designed for the problems they are supposed to solve. While it is relatively easy to define and classify types of data fusion, the same can not be said for unifying different fusion methodologies in a comprehensive framework**. Role of geostatistical models in the process of aggregation of incompatible spatial data.**

LITERATURE REVIEW (3)

Spatial disaggregation aims at interpolating spatially aggregate data (source zones) into a different spatial zoning system of higher spatial resolution (target zones). Spatial disaggregation methods are essentially based on estimation and data interpolation techniques and they can be classified according to several criteria e.g. underlying assumptions, use of ancillary data, etc.

- All these spatial disaggregation techniques generate error: assumptions about the spatial distribution of the objects; spatial relationship imposed within the spatial disaggregation process. Role and effect of the specification of **geostatistical models in small area estimation** (disaggregation)

LITERATURE REVIEW (4)

Interpolation is a method of constructing new data points within the range of a discrete set of known data points. Spatial interpolation includes any of the formal techniques which study entities using their topological, geometric, or geographic properties.

- In spatial interpolation the values of an attribute at unsampled points need to be estimated, meaning that spatial interpolation from point data to spatial continuous data is necessary. Interpolation can be done by **geostatistical models.**

SUB-TOPICS REQUIRING FURTHER RESEARCH (1)

Measurement errors in geostatistical models.

- Modelling quantities of interest over a particular geographical region is based on (usually noisy) measurements taken at a set of locations in the region. Including a measurement error component in the auxiliary variable(s) is a tool that can help inferences from models for reported areas, also with regards to systematic bias of area measurement. Many of the models developed for integration and disaggregation (say SAE models) have still to be generalized to include the possible measurement errors. Particularly the M-quantile regression models still need this extension.

SUB-TOPICS REQUIRING FURTHER RESEARCH (2)

Missing values in spatial data and in auxiliary variables

- The patterns of missingness in spatial data (as collected by GPS-based methods or remote sensing methods) and the investigation of their implications for land productivity estimates and the inverse scale-land productivity relationship constitute a very important issue. Using Multiple Imputation (MI) can constitute a useful, and still not completely explored tool, to face with the problem in agro-environmental studies.

SUB-TOPICS REQUIRING FURTHER RESEARCH (3)

Developments in small area estimation models in agro-environmental studies.

Small area estimation models can afford many of the problems in data disaggregation. Very important is the strength to be borrowed by valuable auxiliary information obtained exploiting spatial data and combining them with study variables coming from sample surveys and censuses (especially in developing countries)

- Models when the auxiliary variables are measured with error (see previous topic 1). This means trying to take into account this non-sampling error component when measuring the mean squared error of the area estimators, improving the measure of their accuracy.

SUB-TOPICS REQUIRING FURTHER RESEARCH (4)

- Models for space (and time) varying coefficients. That is model allowing the coefficients to vary as smooth functions of the geographic coordinates. These could increase the efficiency of the SAE estimates identifying local stationarity zones. Extensions are possible for multivariate study variables.
- Theory for “zero inflated” SAE models (some zeros in the data that alter the estimated parameters) as this is a common situation in survey data in agro-environmental field.
- Benchmarking and neutral shrinkage of SAE models. That is taking into account the survey weights (if any) in spatial SAE models to benchmark to known auxiliary totals.
- Multiple frame SAE modelling. When auxiliary data come from several areas or list frames and units appear in different frames SAE modelling could take advantage of the multiple information and in any case should take into consideration how the linkage of the information affect the accuracy of the estimates. This in comparison with the alternative of using only separate, unlinked data sources.

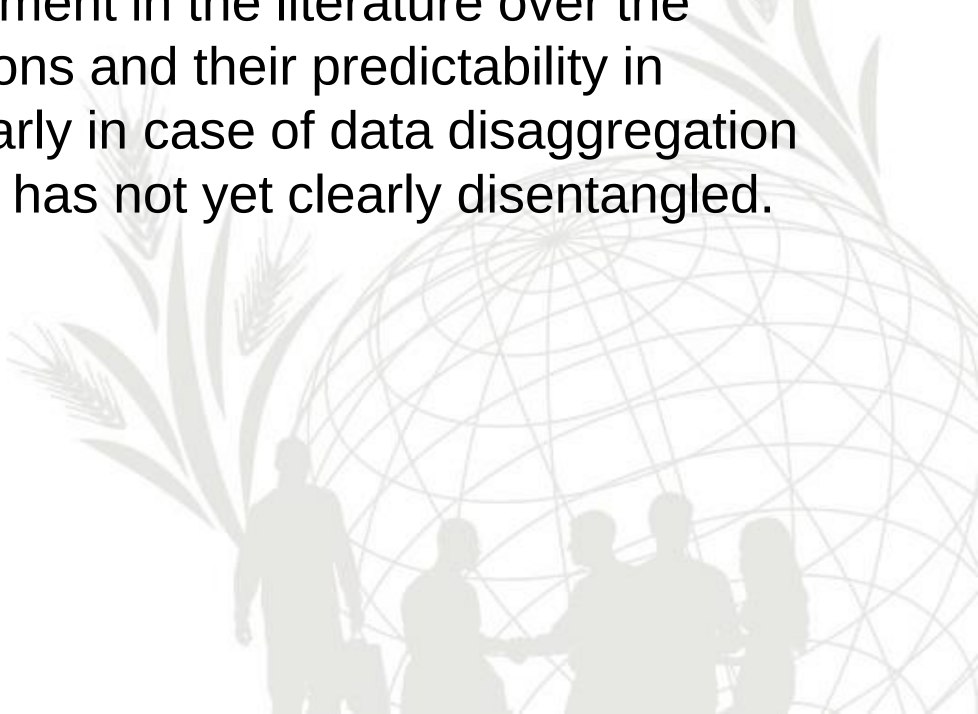
SUB-TOPICS REQUIRING FURTHER RESEARCH (5)

	Model Assisted	Model based (Linear Mixed Models)	Model based (M-Quantile models)	SAE binary/count data
Applicability in Developing Countries	Reliability of data sets	Reliability of data sets	Reliability of data sets	Reliability of data sets EBP computations
Recommendations on the method proposed in the literature	GREG design consistency	Efficiency/ Sensitive to departures from the model and to outliers	No model assumption/ robust to outliers	Model parameters by iterative procedureMPQL /REML
Further research	GREG+ spatial info	Spatial EBLUP + robustness	Spatial Temporal MQ	Spatial EBP

SUB-TOPICS REQUIRING FURTHER RESEARCH (6)

The statistical treatment of the so-called COSPs in SAE context.

- Many of the concepts interlinked with the modifiable area unit problem and other change of support problems have still to be solved and there is no agreement in the literature over the precise scope of its implications and their predictability in statistical inference. Particularly in case of data disaggregation via SAE models the problem has not yet clearly disentangled.



SUB-TOPICS REQUIRING FURTHER RESEARCH (7)

COSPs/MAUP	Optimal zoning	Model with grouping variables/area level models	Geographically Weighted Regression – MQGWR	Kriging Block kriging
Applicability in Developing Countries	Reliability of data sets/ Access to indiv. geo data	Reliability of data sets/ Access to indiv. geo data	Reliability of data sets/ Access to indiv. geo data	Reliability of data sets/ Access indiv. geo data
Recommendations on the method proposed in the literature	Conditioned to constraints/ impractical	Zoning effect/ Homogeneity inside groups?	Local parameters/ robust to outliers?	Average expected value is generated
Further research	Multivariate zoning	Homogeneity Spatial dependency	Spatial Temporal MQ/a-posteriori area effects	Co-kriging, multivariate kriginng

THANK YOU!

Monica Pratesi, m.pratesi@ec.unipi.it