Challenges of regional - small area statistics in higher education

Erkki Pahkinen
University of Jyväskylä, Department of Mathematics and Statistics
Mattilanniemi, Building MaD
Jyväskylä, Finland
E-mail: pahkinen@maths.jyu.fi

Abstract :

   Cross-discipline knowledge is a prerequisite for producing and interpreting regional - small area statistics. Production requires skills in statistics, computer science and applied mathematics. In addition, some empirical research experience from an appropriate substance field is necessary. Political decision-makers and mass media should master literacy on this area for the proper use of results achieved by researches of regional statistics. To fulfil these learning demands in higher education, we suggest an educational programme in five steps. An informative theme lecture helps users of small area statistics to attain at least a general point of view over this field. In turn, embedding regional -small area topic(s) in running methodological courses may give knowledge in depth of some production problems for undergraduate students. A focused, two-year academic master's programme with emphasis on regional statistics prepares graduating students for a professional career in this field. PhD students could inscribe in seminars or scientific meetings organised by research projects of this field. Distance learning serves as an example of continuing education. Links with homepages of some existing programmes are provided.

Keywords: Linked course materials, Focused master's programme

1. Introduction

A good example of the oddities of small area research is the particular terminology in use. Firstly, we can recognise that the word "subpopulation" has several such synonymous expressions as subgroup, category, class, geographical area, domain or region. According to Rao (2003), the concept "small area" refers to those subpopulations owing few or nil observations. Secondly, in this case we are interested in parameters of subpopulations but not those of the whole population. From this, it follows that also the quality measures for produced statistics shall be defined on the subpopulation level. It often happens that calculated small area estimates are biased so that the bare standard error gives an erroneous view of variation. Then a better way is to use the root mean squared error to measure the uncertainty of inference. Equally, the relative standard error or the coefficient of variation has a great importance if compared with variability among subpopulations. In addition, tailored estimators seem to be more like a rule than coincidences. For the evaluation of the properties of these estimators, many simulation experiments are put in force. To summarise, one can find out that statistical concepts and tools in use here have special properties.

   Important point of view is the political relevance. For example, political authorities aiming to allocate funds to geographical regions according to some development programme need regional – small area statistics. The redistribution of funds depends on the funding formulas and calculated figures. If the local share of funds does not meet the expectations of the local population, a political conflict may arise. In order that a redistributive action might succeed, producing small area statistics needs professional knowledge and the functioning officials and politicians should master literacy on small area statistics and its quality measures. Originating from the need of interdisciplinary knowledge, the producers of small area statistics generally assemble a research team for this task thus ensuring that the difference competencies are available. For example, the EURAREA project (2001-04) was a research programme funded by Eurostat to investigate methods for small area estimation. The research team consisted of researchers from six European countries, spread across twelve National Statistics Institutes and Universities. The Census Bureau (USA) organised a program

– Small Area Income and Poverty Estimates (SAIPE). The panel members were 17 distinguished researchers from different fields.

2. Basic concepts and analysis strategies for regional – small area statistics

The main problem arises from the fact that data sources in use are regularly unplanned for the production of small area statistics in hand. Therefore, researchers should first investigate existing data sources as 1) population census or regular register, 2) such administrative records as for tax files or vehicle register, 3) large-scale surveys as Labour Force Survey or Health Survey, and 4) local studies as satellite imaging, aerial photography. To get these data sources in use, special knowledge in computer science helps a) to link records, b) to map elements and variable values and c) to perform statistical disclosure control.

In Table 1 the data source and the number of observations are cross-tabulated, and the analysis strategy derived from these facts is proposed following some ideas of Brackstone (2002). The mass of data for a subpopulation is divided roughly into four categories a) large, b) medium, c) small and d) empty.

Table 1. Analysis strategy for a subpopulation by data source and the number of observations.

Number of observations for a subpopulation	Source of observations in subpopulation		
	Census or register; micro data	Large-scale survey; sample data	Observational data or local studies
Large	Parameter value	Design-based estimate	Parameter value
Medium	Parameter value	Model-assisted estimate	Model-based estimate
Small	Disclosure control	Model-based estimate	Model-based estimate
Empty	Model-based estimate	Model-based estimate	Model-based estimate

Direct calculation of parameter values is possible if we have observations from each element of the subpopulation, as may be the situation in micro data and some local studies. However, a special case arouses if the number of elements in a subpopulation is so small that rules of privacy protection prevent publications of figures. The need of a special skill is obvious to perform statistical disclosure control. In extremity no observations are included in the micro data. Then some model-based estimation technique applies in producing the subpopulation figures. Thus, the need of different practical and scientific knowledge is obvious if one wants to put these production strategies into practice. The need of special skills in computer science (managing databases and disclosure problems), applied mathematics (simulation studies), statistics (estimation and modelling) and survey sampling (design-based inference) is then useful. The content of each strategy could better open out if one can familiarize with following textbooks. Statistical disclosure control is well presented in Willenborg and de Waal (2001). Elliot (1997) explains different mapping systems and geographical information systems (GIS) in epidemiological context. Simulation studies are superfluous in these environments. Therefore, good knowledge in this part of applied mathematics can be helpful see Robert and Casella (1999). Rao (2003) underlines statistical modelling, and Lehtonen and Pahkinen (2004) present design-based and design-based model-assisted estimation of small area figures. Applied subjects should be selected in accordance with the topic of Master's thesis depending its subject-matter field that can be for example in economics, epidemiology, social science or demography.

3. Organising higher education on regional - small area statistics

To meet these educational needs we suggest a five-step learning programme. Because the methodology of production of regional – small area statistics is multidisciplinary and the research results are applied in different fields, each step of proposed programme is furnished with short description of aim, potential audience and teacher. Homepages of some existing programmes and courses serve as reference.

Table 2. Ordered list of higher education programmes concentrated on regional statistics

Title and definitions	Description of pedagogical solution: aim, homepage of some current programme (if exists), audience and recommendation for speaker/teacher.
1. Theme lecture(s)	- aim to clarify interpretation of research results in regional statistics - audience: political decision-makers and mass media - speaker/teacher: distinguished researcher
2. Linked course materials with running stats/maths courses	- aim to focus regional statistics as an item embedded in some regular course of maths/stats. See homepages: **www.iastate.edu** (USA) and www.uow.edu.au/informatics/maths/ - audience: undergraduate students in maths/stats - speaker/teacher: university lecturers
3. Retrospective Master's programme in regional - small area statistics	- aim to help students to find a career in this area. See homepages: **www.jpsm.umd.edu/jpsm** and **www.socstats.soton.ac.uk/moffstat/** - audience: graduate students in maths/stats - speaker/teacher: supervisor(s) of MS theses, visiting experts
4. Scientific meetings and research projects	- aim to give theoretical and practical view-in-depth to current problems in regional statistics. See homepages **www.scorus2006.ae.wroc.pl** and www.statistics.gov.uk/eurarea/ - audience: graduate and/or PhD students - speaker/teacher: experts in this area, supervisors of PhD theses
5. Distance learning or continuing learning	- aim to strengthen one's methodological or pedagogical knowledge in regional statistics See homepages: http://www.rsscse.org.uk/activities/ and http://mathstat.helsinki.fi/VLISS - audience: teachers, researchers and producers of this area - speaker/teacher: administrative/planning group of learning centre

We define a Master's Programme as a retrospective one. In this case, if it is tailored as an ending phase of academic studies, especially for those students, who are interested in regional – small area statistics. This programme could include some courses as in statistics, applied mathematics and informatics, whose contents are useful in analysis of regional data. In addition, there is need for some special courses in order to give general view over this issue. It is purposeful to share course materials and teaching activities to both academic teachers (50%) and experienced researchers (50%), who are from some institution, which produces regional statistics. For students, these kinds of courses give a real touch for practical work. Naturally, the topic of MS thesis concerns some method or application in this field.

The homepages of two running Master's programmes are checked so that could we construct from their course offers an MS Programme with emphasis on regional statistics:

1. The Joint Program in Survey Methodology (JPSM) runs jointly by the University of Maryland, the University of Michigan and Westat Inc. In addition, the Census Bureau (USA) is an important co-operative unit of this program.

2. The Diploma/MSc in Official Statistics is unique in the UK in offering a programme of study tailored specifically to the needs of professional statisticians working within the field of official statistics. This MSc programme (MOFFSTAT) has been developed jointly by the University of Southampton and the Office for National Statistics (UK).

Focused teaching activities in the interdisciplinary field of regional statistics are compiled in Table 3. If it exists, the course is marked as temporary or regular.

Table 3. Selected courses from two running Master's programmes with frequency

Topics of Small Area Oriented Courses	Master's Program	
	JPSM (USA)	MOFFSTAT(UK)
Statistical disclosure control	Temporary	Permanent
Computer science (Record linkage and/or GIS)	Temporary	Temporary
Applied mathematics (Simulation studies)	Regular	Regular
Survey sampling (Design-based estimation)	Regular	Regular
Statistical modelling (Model-based estimation)	Regular	Regular
Applied subject (Economics, Demography etc.)	Optional	Optional
Regional - small area oriented master's thesis	Allowed	Allowed

4. Conclusion

Expertise in productions of regional statistics demands specific skills in several disciplines. This multidisciplinary property is challenge for higher education. To solve this problem, we presented five-step learning programme in higher education, which presumes pedagogical arrangements from the part of teaching organisations. This can be characterised by cooperation between the fellow lectures to construct linked courses and to plan a small area focused master's programme. On the other hand, students engaging in this topic should select courses from different majors to get an understanding of different skills necessary in the production of small area statistics.

REFERENCES

Brackstone ,G.,J. (2002). Strategies and Approaches for Small Area Statistics. *Survey Methodology*, 28 117-123

Elliot, P, (Editor) (1997) *Geographical and Environmental Epidemiology: Methods for Small-area Studies*. Oxford: Oxford U.P.

Lehtonen, R. and Pahkinen, E. (2004*). Practical Methods for Design and Analysis of Complex Surveys*. 2nd edition. Chichester: John Wiley &Sons.

McCulloch, C.E. and Searle, S.R. (2001). *Generalized, Linear, and Mixed Models*. New York: Wiley.

Rao, J.N.K. (2003*). Small Area Estimation*. Hoboken NJ: John Wiley &Sons.

Robert, C. P. and Casella, G. (1999).*Monte Carlo Statistical Methods*. New York: Springer.

Willenborg, L. and de Waal, T. (2001). *Elements of Statistical Disclosure Control*. Lecture Notes in Statistics, Vol. 155 New York: Springer.

Projects:

Eurarea: http://www.statistics.gov.uk/eurarea/downloads/EURAREA_PRV_1.pdf

Saipe: http://www.cencys.gov.uk/hhes/www/saipe93.html

RESUMÈ

La transdisciplinarité est un préalable à la production et l'interprétation de statistiques sur petits domaines au niveau régional. Produire de telles statistiques requiert des compétences en statistique, informatique et mathématiques appliquées. De plus, une expérience en recherche empirique dans le champ d'application est nécessaire. Les décideurs politiques et les médias de masses devraient maîtriser ce domaine afin d'utiliser convenablement les résultats obtenus par les recherches en statistique régionale. Dans le but de satisfaire des demandes d'initiation au niveau universitaire, nous proposons un programme d'apprentissage en cinq étapes. Des liens internet à des programmes existants seront fournis.