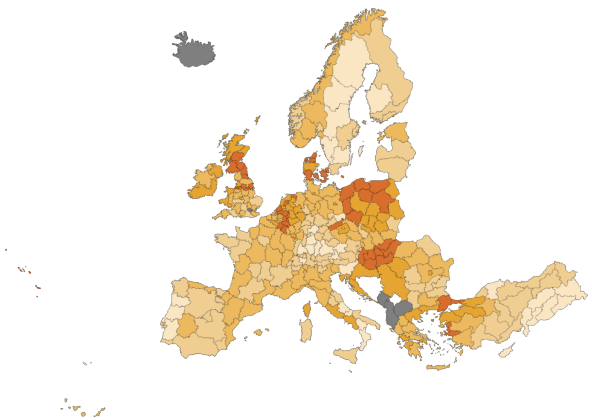
Statistical quality of small area estimates – internal and external validation of the estimates? The particular case of *disease mapping*



Lola Ugarte

Statistics and O.R. Department
Universidad Pública de Navarra
Pamplona, Spain

Conference on Small Area Methods for Monitoring Poverty and Living Conditions in EU, Pisa, 8-10 May 2018

Health in the EU

"Health is an important priority for Europeans, who expect to be protected against illness and accident and to receive appropriate healthcare services"

The third multi-annual programme of EU action in the field of health for the period 2014-2020 (Regulation (EU) No 282/2014) foresees expenditure of almost EUR 450 million over the seven-year period with a focus (among other things) on

- the increasing health inequalities between EU Member States
- the prevalence of chronic diseases

Source: Statistics Explained (*http* : *//ec.europa.eu/eurostat/statistics − explained/index.php/Health − statistics − introduced*) - 18/01/2018

Overview

- Introduction to disease mapping
- Disease mapping and EUROSTAT
- Limitations of the SMR
- Some spatial models
- Estimates validation

Introduction to disease mapping

- The representation and analysis of disease incidence or mortality has been established as a basic tool for the analysis of regional public health data
- *Disease mapping* may be defined as the estimation and representation of area summary measures of health outcomes (Wakefield, et al., 2000)
- Disease mapping is usually seen as a particular case of SAE where sampling is not involved
- The tools of this particular area of SAE are important to discover health inequalities among small areas in Europe affecting the living conditions of EU citizens and to subsequently distribute health funds

Disease mapping and EUROSTAT

- The most updated data on mortality in the EU provided by EUROSTAT relates to standardised death rates, averaged over the three-year period 2011–2013. This seems rather limited
- The data on causes of death are generally available for NUTS 2 regions (version of 2013), covering the resident population of each territory
- However, only national data are available for Slovenia, while there are no data available for the French Départements d'outre-mer (FRA), nor for London (UKI)

Disease mapping and EUROSTAT

- Appropriate disease mapping analyses rely on "good" mortality and incidence Population Data Registers. A first problem in the EU could be data availability in all targeted small areas. For example there are not incidence registers in all regions nor even mortality data for certain regions as we have seen (even at NUTS2)
- The current NUTS 2016 classification is valid from 1 January 2018 and lists 104 regions at NUTS 1, 281 regions at NUTS 2 and 1348 regions at NUTS 3 level
- SAE techniques are possibly necessary for (yearly) NUTS 2 and NUTS 3 regions
- Many cancer locations are age-dependent and so, it may be convenient to estimate rates by age and region

Limitations of classical mortality risk estimates like the SMRs

- When studying small areas or rare diseases, classical measures like the SMRs are too variable and then not reliable
- To deal with this situation it is usual to use "sophisticated" statistical models that borrow strength from neighbouring areas
- These models usually include random effects (structured and/or unstructured) for smoothing risks in low populated regions
- The more used models are hierarchical models, in particular mixed Poisson models

Spatial models in disease mapping

- Clayton and Kaldor (1987, BIOMETRICS) defined empirical Bayesian methods building from Poisson regression with random intercepts defined with spatial correlation
- This hierarchical approach provides a convenient conceptual framework wherein one induces spatial correlation across the estimated local disease rates via a conditionally autoregressive (CAR; Besag, 1974, JRSSB) random effects distribution assigned to the area-specific intercepts
- The models were extended to a fully Bayesian setting by Besag, York, and Mollié (1991, AIMS)
- In both the non-spatial and spatial settings, the amount of smoothing is determined by the data and the formulation of the model. This smoothing permits easy visualization of the underlying geographic pattern of disease

Estimates validation

- In the context of disease mapping where data are provided by Population Registers I only see INTERNAL VALIDATION
- Firstly, when considering mixed Poisson models, model identifiability is important. Appropriate constraints should be considered if the interest relies on computing spatial, temporal, and spatio-temporal components (see Goicoa et al., 2018, SERRA)
- Secondly, an appropriate model selection is also necessary. In a Bayesian framework (the most typical in disease mapping) model selection criteria like DIC and WAIC can be used. Besides, the model does not need to be the same for different causes of mortality
- Sensitivity analyses must be performed
- Methods based on the predictive distribution like the LS (based on cross-validation) or the PIT can also be used to validate the model

Estimates validation (cont.)

- Benchmarking procedures may be also considered
- For example, benchmarking the model-based estimates to direct estimates considered reliable at certain level of aggregation
- It is also possible to derive a small area estimator based on a GLMM with restrictions to guarantee the concordance between the aggregations of small area estimates and those reported by statistical agencies (EUROSTAT) for larger domains (similar ideas in Ugarte et al., 2009, TEST)
- A global measure like the MSE can be used to evaluate over-smoothing