Contents lists available at ScienceDirect

Preventive Veterinary Medicine

journal homepage: www.elsevier.com/locate/prevetmed

Iso-population partition: An innovative epidemiological approach to mapping and analyzing spatially aggregated data

Anne Bronner [a], Eric Morignat [a], Emilie Gay [a], Timothée Vergne [b], Guillaume Fournié [b], Dirk U. Pfeiffer [b], Didier Calavas [a,*]

[a] *ANSES-Lyon, Lyon, France*
[b] *Royal Veterinary College, Hatfield, Hertfordshire, UK*

ABSTRACT

In epidemiology, data are often aggregated using administrative boundaries or regular spatial lattices. Iso-population partitioning methods allow the aggregation of small units for which population data are available into larger units that are contiguous, as compact as possible, and have a similar population size. The objective of this paper was to study the influence of three spatial data aggregation approaches on data visualization and data analysis: iso-populated units (IPUs), administrative units, and iso-geometric units. This study was conducted using results and simulations from the brucellosis clinical surveillance system for dairy cattle in France. Our findings indicate that using spatial partitioning methods for generating IPUs enhances the ability to interpret the spatial distribution of epidemiological indicators under study. In addition, it provides information on population density and improves the consistency of the power of statistical tests across units. By defining the target population size per spatial unit, IPUs can be used to control the statistical power of a study. Finally, by adding criteria based on environmental factors to generate spatial units, they can be used to control the variation of exposure to these factors within the units.

© 2015 Elsevier B.V. All rights reserved.

1. Introduction

Many spatial epidemiological analyses are based on aggregated data. In fact, the exact location of cases may not be directly available, or variables of interest (*e.g.*, social or environmental covariates) that might influence disease incidence may be missing at individual-level. This lack of high resolution locational information can be due to the spatial resolution at which information was recorded or due to data protection (Bell et al., 2006). In addition, even if high resolution locational data are available, aggregated data analysis can be used to increase the stability of modeling estimates, by preventing emergence of spurious spatial patterns caused by random variation in small numbers of events (Miyawaki and Chen, 1981). For example, aggregated data are used for disease mapping (EFSA, ECDC, 2014), analyzing the geographical distribution of disease in relation to explanatory covariates (Pfeiffer et al., 2007; Fang et al., 2013), or for syndromic surveillance (Kleinman et al., 2005).

Data are often aggregated at administrative levels that correspond to the level at which they are collected (Justin and Williams, 1995; Bronner et al., 2013; EFSA, ECDC, 2014). Data may also be aggregated using regular spatial lattices such as a hexagonal grid (Abrial et al., 2005). In either case, the population size generally varies greatly among units. As a result, the uncertainty in risk or risk ratio estimates can vary greatly (Beale et al., 2008). Bayesian methods can be used to map relative risks and their uncertainty (Jarup et al., 2002; Beale et al., 2008). Likewise, Bayesian hierarchical models, based on an adaptive smoothing approach, can be used to produce more stable estimates of the pattern of underlying disease risk (Best et al., 2005). It is acknowledged that raw risks can result in 'noisy' maps that are difficult to interpret, but over-smoothed maps may mask underlying variability in the true risk distribution (Beale et al., 2008).

As an alternative, iso-population partitioning methods allow the aggregation of small units (for which population data are available) into larger spatial units, with each of which having a similar population size, referred to hereafter as iso-populated units (IPUs). Historically, these methods were developed to define the electoral districts in the United States for ensuring that each elected member represents the same number of individuals (Justin and Williams, 1995). To our knowledge, they have neither been used in human nor in animal epidemiological studies.

The objective of this paper was to study the influence of three spatial data aggregation approaches on data visualization and data analysis: administrative units, iso-geometric units and IPUs. The study uses data from the brucellosis clinical surveillance system for dairy cattle in France.

2. Materials and methods

2.1. *Data* sources and study population

In France, the bovine brucellosis clinical surveillance system requires farmers to consult their veterinarian in case of abortion, who report the abortion to veterinary services and sample the aborted cow for a serological analysis for *Brucella* (European Commission, 1964; Anon., 2003). The study focused on abortions recorded during the 2nd semester of 2009 in dairy cattle farms located in mainland France.

For all dairy cattle premises, information about the farm (identification number, ZIP code of the municipality in which the farm is located) was extracted from the French National Cattle Register. Abortion notification data reported by farmers (*i.e.,* notifying farmers) during the 2nd semester of 2009 were extracted from the French national animal health information database SIGAL (Système d'information de la Direction générale de l'Alimentation), including the farm identification number and dates of veterinary visit.

2.2. Data aggregation

Dairy cattle farms and bovine abortions are recorded at the municipality level, the smallest French administrative unit. These data were aggregated using three different approaches resulting in a similar number of units: (i) administrative boundaries (at the arrondissement level,[1] (ii) a hexagonal spatial grid generated by dividing France into contiguous hexagons, having the same size and surface and (iii) partitioning methods to generate IPUs. Partitioning methods refer to the aggregation of the smallest units of a territory for which data are available (or elementary units) into larger units, fulfilling one or several criteria such as technical, ecological, social or demographic criteria (Tavares-Pereira et al., 2007). This method can be modeled using graph theory (Altman and McDonald, 2011). Each elementary unit is associated with a vertex of the graph, while a pair of contiguous elementary units defines an edge of the graph. Some numerical values, such as population size (here, the number of dairy cattle farms) or geographic locations, are associated with the edges and/or vertices. Maps (or plans) of a territory are generated using division or aggregation based methods, and exact or non-exact (heuristic) algorithms (Tavares-Pereira et al., 2007). A partition satisfying all criteria is often impossible to determine and the 'best solution' is frequently a compromise in which an optimal fit to a particular criterion results in a less optimal or even poor fit to at least one of the other criteria (Tavares-Pereira et al., 2007).

In our case, we intended to partition France into equally populated but also contiguous and compact units. This meant that (1) the population size per aggregated unit should be equal to the ratio of the total population size to the desired number of aggregated units (to ideally exactly meet the equally population criterion), (2) elementary units included in an aggregated unit should be adjacent to each other (to meet the criterion of contiguity) and (3) the ratio of the surface area of this larger unit to the area of a circle in which it is included should be close unity (to meet the compactness criterion). Using the R Bard package (Altman

and McDonald, 2011), we generated plans based on the weighted *k*-means and the random-walk based methods for generating contiguous iso-populated units (Cirincione et al., 2000). These plans were evaluated, but also used as starting points for further refinement by applying an optimization algorithm based on simulated annealing (Altman and McDonald, 2011). The random-walk-based method offered the best compromise between contiguity, compactness and iso-population criteria.

2.3. Mapping and analyzing aggregated data

To study the influence of the spatial data aggregation approach on data visualization, the distribution of the proportion of notifying farmers per spatial unit was mapped for each of the three different spatial data aggregation approaches. The spatial distribution of the proportion of notifying farmers was compared visually between partitioning approaches.

Spatially aggregated data can be used to analyze the influence of explanatory variables on the distribution of disease across units. They can also be used for syndromic surveillance that relies on the detection of a significant variation for a particular indicator, such as bovine mortality (Perrin et al., 2010) or milk yield from milk recording in dairy cattle (Madouasse et al., 2014). In either case, the probability of detecting an association if it actually exists or an abnormal variation of the indicator, *i.e.*, the statistical power of the study, is positively correlated with the population size and the magnitude of the effect in each unit. To study the influence of the spatial data aggregation approach on the statistical power of a study, we simulated in each unit an equal increase of the proportion of notifying farmers from one semester (S1) to another (S2). The number N_i of cattle farms in unit i in S1 and S2 was the same and equal to the number of dairy cattle farms recorded over the 2nd semester 2009 in this unit. The number of dairy cattle farms in which an abortion was recorded over S1 and S2 was computed as $n_{i,S1} = PN_i$ and $n_{i,S2} = \gamma PN_i$, respectively, with P being the national proportion of notifying farmers and γ the level of increase in the number of notifying farmers between S1 and S2. Several values for γ were simulated. For a given value of γ, a variation of the proportion of notifying farmers from S1 to S2 was considered as detected in unit i if the p-value of a Chi-square test was below 0.05 divided by the number of units included in the study (Bonferroni correction). For each aggregation approach and a wide range of values for γ, we calculated the proportion of units in which an increase was detected. In addition, we compared the population size of the units in which an increase was detected with the population size of the units in which no increase was detected by using a Student's *t* test.

3. Results

During the 2nd semester of 2009, at least one abortion was reported in 21% ($n = 13,986$) of the 66,382 dairy cattle farms present in mainland France. Population size per municipality varied from 0 to 79 dairy cattle farms. Using administrative boundaries resulted in data being aggregated at the arrondissement level ($n = 325$). A hexagonal spatial grid was used to define 320 units. Lastly, France was divided into 314 IPUs, with population size per unit varying from 1 to 454 (Fig. 1). The coefficient of variation[2] among IPUs (CV = 0.45) was far lower than among the administrative and hexagonal grid units (CV = 1.31 and 1.32, respectively) (Fig. 1). Fig. 2 presents the spatial distribution of the proportion of farmers who reported abortions for each of the three different spatial data aggregation approaches. The proportion of spatial units in

---

[1] An arrondissement is a French administrative unit with a mean area of 1600 km².

[2] Defined as the ratio of the standard deviation to the mean.