Spatial Statistics 23 (2018) 91–108

Contents lists available at ScienceDirect

Spatial Statistics

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

Disaggregating statistical data at the field level: An entropy approach

António Xavier [a,b], Maria de Belém Costa Freitas [c], Maria do Socorro Rosário [d], Rui Fragoso [e,b,*]

[a] Universidade do Algarve, Campus de Gambelas, 8000-117 Faro, Portugal
[b] CEFAGE-UE (Center For Advanced Studies in Management and Economics), Portugal
[c] Professora Auxiliar c/Agregação, Universidade do Algarve, Faculdade de Ciências e Tecnologia, Campus de Gambelas, 8000-117 Faro. ICAAM, Portugal
[d] Direção de Serviços de Estatística, GPP (Gabinete de Planeamento e Políticas), Portugal
[e] Professor Auxiliar c/Agregação, Universidade de Évora, Largo dos Colegiais, 7000 Évora, Portugal

ARTICLE INFO

Article history:
Received 19 June 2017
Accepted 14 November 2017
Available online 7 December 2017

Keywords:
Data disaggregation
Entropy
HJ-Biplot
Cluster analysis
Dasymetric mapping

ABSTRACT

This paper provides an alternative approach to disaggregating agricultural data concerning land-use at the detailed pixel level. The proposed approach combines several techniques, such as Hj-Biplot, cluster analysis, dasymetric mapping and cross-entropy, and it is implemented in two steps. First, prior information is estimated based on the application of a HJ-Biplot and cluster analysis and using a dasymetric mapping methodology. Then, the estimated prior information is used in a cross-entropy model to disaggregate data at the pixel level in a context of incomplete information. This approach is applied to the Algarve region in southern Portugal. The results show a significant correlation between observed and estimated land-uses and are relevant in terms of information gains.

1. Introduction

Analyses of agricultural economics and policy, as well as environmental issues, require spatial information on land-use at a local, farm, sub-regional or regional scale (Gocht and Röder, 2014; Leip et al., 2013; Butterbach-Bahl et al., 2011). Crop distribution at a detailed scale can reveal the agricultural

* Corresponding author.
E-mail addresses: amxav@sapo.pt (A. Xavier), mbfreitas@ualg.pt (M.d.B.C. Freitas), socorro@gpp.pt (M.d.S. Rosário), rfragoso@uevora.pt (R. Fragoso).

https://doi.org/10.1016/j.spasta.2017.11.005
2211-6753/

production pattern over different periods and global change on agricultural production (Foley et al., 2005; Portmann et al., 2010; Tan et al., 2014).

Generally, most studies of land-use include aggregate data at a regional or national scale (Chakir, 2009; Plantinga and Ahn, 2002). In the European Union, the Farm Accountancy Data Network (FADN) is the main source of agricultural data. Data is collected in an annual survey carried out by all Member States and provided in an aggregated format at the regional level, according to the Nomenclature of Territorial Units for Statistics II (NUTS II). The Agricultural Census provides more detailed information, but is only implemented every 10 years. No information is available at a detailed scale, such as the municipal or parish level. Cartographical land-cover information is also incomplete since its classification is different from farms' land-use presented in statistics.

This lack of information has motivated the development of some data disaggregation processes (Howitt and Reynaud, 2003; Fragoso et al., 2008; Martins et al., 2011; Xavier and Costa Freitas, 2014). Several studies aiming to analyse the impact of policies and technological innovation on agricultural sustainability have tried to overcome the lack of information, creating prior information to improve the process of recovering new information (Kempen et al., 2005; You and Wood, 2006; You et al., 2009; Chakir, 2009; Louhichi et al., 2012; You et al., 2014; Tan et al., 2014; Lamboni et al., 2016). Logistic regressions, expert knowledge, homogeneous units, Bayesian methods, entropy approaches and different sources of data have been used in these studies.

Some disaggregation approaches (Reibel and Agrawal, 2007; Chakir, 2009; Lamboni et al., 2013; Gocht and Röder, 2014; Lamboni et al., 2016), use prior information to predict land-use areas within fine scale units, based on land-use statistics, geographically-referenced or remote sensing data and ground-based and point-based observations (namely, the LUCAS survey). Generally, these studies first estimate the prior distribution of land-use at a finer scale. Then, the predicted areas are constrained, according to statistical data using Bayesian methods or entropy approaches, to guarantee consistency with the statistical data available (You and Wood, 2006; You et al., 2009).

Nevertheless, these previous studies do not include less used techniques such as HJ-Biplot methodology (Galindo, 1986) or some relevant dasymetric mapping techniques (Martins et al., 2012). In fact, the dasymetric mapping technique can be used to create prior information, using just the existing land-cover maps and statistical data (Gallego and Peedell, 2001; Gallego, 2010). Martins et al. (2012) propose an iterative procedure for redistributing the statistical variables by land-cover class, using only the land-cover cartography as covariates.

The HJ-Biplot is a multivariate technique of data reduction, which allows graphic representation in a low-dimensional space, the interrelations between individuals and variables being captured visually. Generally, the results are better than those of the classic biplot methods (Galindo, 1986; Garcia-Talegon et al., 1999; Cabrera et al., 2006), and can be used in an additional cluster analysis to create homogeneous areas (Xavier and Costa Freitas, 2014; Serafim et al., 2012; Gallego-Alvarez et al., 2014). According to previous studies, this procedure has clear advantages over other factorial methods (Galindo Villardón et al., 1996).

The increased availability of geo-referenced data and geographical information systems (GIS) provide better opportunities for spatial analyses (You et al., 2009). Flexible information recovery methods have also been used in spatial analysis, such as the information entropy theory introduced by Shannon (1948). The maximum entropy principle proposed by Jaynes in 1957 in statistical inference provided a constructive criterion for establishing probability distributions, using partial or incomplete knowledge (Golan et al., 1996). Good (1963) developed the concept of minimum cross entropy to measure the discrepancy between two probability distributions. Golan et al. (1996) developed generalised cross entropy (GCE), which considers the unknown distribution and measurement of errors. This GCE approach has been used successfully by several authors in agricultural data disaggregation (You and Wood, 2006; You et al., 2009; Chakir, 2009; Martins et al., 2011; Xavier et al., 2014; You et al., 2014; Tan et al., 2014).

Other interesting approach that uses an entropy method is the Penalised Maximum Entropy Dasymetric Model (P-MEDM) proposed by Nagle et al. (2014). This model deals with uncertainty in dasymetric mapping, integrating population and ancillary or prior information from different data sources as well as information about uncertainty of data used (population and ancillary) and its estimated relationship. Fine resolution population estimates can be obtained with quantifiable uncertainty.