Land Use Policy 75 (2018) 530–537

Contents lists available at ScienceDirect

Land Use Policy

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

Misclassification error in satellite imagery data: Implications for empirical land-use models

Austin M. Sandler[a,*], Benjamin S. Rashford[b]

[a] Department of Geographical Sciences, University of Maryland, 2181 LeFrak Hall, 7251 Preinkert Dr, College Park, MD 20740 United States
[b] Department of Agricultural and Applied Economics, University of Wyoming, Dept. 3354, 1000 E. University Ave., Laramie, WY 82071 United States

ARTICLE INFO

Keywords:
Land-use
Econometrics
Misclassification error
Satellite imagery data
Northern Great Plains

ABSTRACT

Satellite-based land-use data sets are providing new opportunities for land-use research. However, care must be used when working with these datasets due to misclassification error, which causes inconsistent parameter estimates in typical land-use models. Results from satellite imagery data from the Northern Great Plains indicate that ignoring misclassification will lead to biased results. Even seemingly insignificant levels of misclassification error (*e.g.*, 1%) result in biased parameter estimates, which alter marginal effects enough to affect policy inference. At the levels of misclassification typical in current satellite imagery datasets (*e.g.*, 35%), ignoring misclassification can lead to systematically erroneous land-use policies.

1. Introduction

Land use research has focused on developing economic models of individual landowner's decisions within a spatially explicit framework (Irwin and Geoghegan, 2001; Lynch and Geoghegan, 2011). Resulting empirical models explain the effects of land-use on environmental resources (Lewis, 2010; Rashford et al., 2010; Rashford et al., 2011; Bhattacharya and Innes, 2013), forest resources (Deininger and Minten, 2002; Munroe et al., 2002; Lewis and Plantinga, 2007; Blackman et al., 2008), agricultural resources (Lynch and Liu, 2007; Butsic et al., 2011; Skevas et al., 2016), and urban and regional planning (Irwin and Bockstael 2002; Wu and Plantinga, 2003; Wu and Cho, 2007; Irwin and Bockstael 2007; Fragkias and Geoghegan, 2010). Natural resource managers, in particular, need this spatially explicit framework to effectively evaluate the social and environmental consequences of alternative land-use scenarios (Bockstael, 1996; Untenecker et al., 2016). The spatial configuration of land-use influences many important indicators of environmental quality, including bird populations (Askins, 2002; Faaborg, 2002), amphibian populations (deMaynadier and Hunter, 2000), health of riparian systems and estuaries (Gergel et al., 2002; Hale et al., 2004; Dempsey et al., 2017), human perceptions of scenic quality (Palmer, 2004), and the extent of urban sprawl (Carrión-Flores and Irwin, 2004).

Many studies use the US Department of Agriculture's (USDA) National Resources Inventory (NRI), which provides information on land-use choices for over 800,000 sample plots across the US from 1982 to 1997 (at five-year intervals) (*e.g.*, Tanaka and Wu, 2004; Lubowski et al., 2006; Lewis and Plantinga, 2007; Langpap and Wu, 2008; Lubowski et al., 2008; Lewis et al., 2009; Rashford et al., 2010; Langpap and Wu, 2011). The NRI, however, has issues of temporal consistency and availability.

Alternatively, researchers use aggregate data because of its availability, geographic coverage, and long temporal scale (*e.g.*, Plantinga and Irwin, 2006). Most commonly, aggregate data models estimate the proportion of an area in different land-uses as a function of exogenous variables expected to influence landowner utility or profits (*e.g.*, Alig 1986; Leitch, 1989; Stavins and Jaffe, 1990; Parks and Murray, 1994; Parks and Kramer, 1995; Wu and Brorsen, 1995; Wu and Segerson, 1995; Plantinga, 1996; Hardie and Parks, 1997; Plantinga et al., 1999; Parks et al., 2000; Hardie et al., 2000; Plantinga and Ahn, 2002; Munn et al., 2002). Since aggregate data models predict aggregate land-use proportions they are only useful for understanding phenomena that respond to aggregate-level land cover characteristics. Aggregate data models cannot predict the consequences of land-use for phenomena that are sensitive to the spatial pattern of the landscape (Lewis and Plantinga, 2007).

The increasing availability of land-use data derived from satellite imagery offers researchers a greater ability to model micro-level land-use (see Holloway et al., 2007). Several papers use early versions of satellite products to model land-use (many in developing countries where other data products are not available). These models examine a range of land-use issues, including agricultural dynamics (Thompson and Prokopy, 2009; Hendricks et al., 2014), ecosystem services (Polasky et al., 2008; Lawler et al., 2014), deforestation (Chomitz and

---

* Corresponding author.
E-mail addresses: sandlera@terpmail.umd.edu (A.M. Sandler), brashfor@uwyo.edu (B.S. Rashford).

https://doi.org/10.1016/j.landusepol.2018.04.008
Received 3 November 2017; Received in revised form 20 February 2018; Accepted 3 April 2018

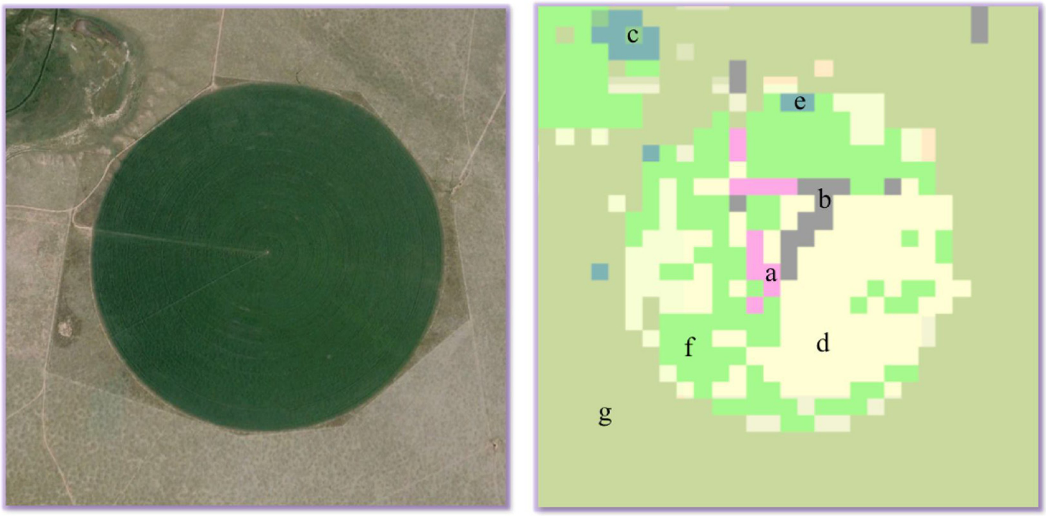

Fig. 1. National Agricultural Image Program aerial photo and corresponding Cropland Data Layer satellite rendered image showing the extent of misclassification error in an irrigation circle (approximately 41°14′21.0″N 105°39′07.7″W) (USDA, 2009a; USDA, 2009b). Each letter corresponds to a specific land cover classification; a: Alfalfa; b: Developed/Open Space; c: Evergreen Forest; d: Grass/Pasture; e: Herbaceous Wetlands; f: Non Alfalfa/Other Hay; g: Shrubland.

Gray, 1996; Nelson and Hellerstein, 1997; Mertens and Lambin, 2000; Cropper et al., 2001), wildlife habitat loss (Polasky et al., 2005; Shi et al., 2006), and climate change (Sohl et al., 2012). Recent availability of satellite-based land-use data sets (*i.e.*, raster datasets), with high resolution of contiguous spatial coverage over broad spatial extents, relatively long temporal coverage, and specific land cover classifications (*e.g.*, rye or winter wheat), are providing new opportunities for future research (Lark et al., 2017). The National Land Cover Database (NLCD) and Cropland Data Layer (CDL) in the US, for example, provide nationwide plot-level observation at high resolution (30 m × 30 m plots) classifying land into refined classifications from altered to natural land covers (*e.g.*, urban, corn, and herbaceous grassland).

Though satellite imagery data offers new opportunities for modeling land-use, it is not without drawbacks. Misclassification error – the phenomenon of observations of land-use being classified incorrectly (*e.g.*, Fig. 1) – can cause vagueness (*e.g.*, position of climatic zone boundaries), ambiguity (*e.g.*, a pixel with more than one class, called a mixel), positional inaccuracy (*i.e.*, data correspondence to true locations), logical inconsistency (*e.g.*, a pixel of tundra in an area of cropland), and incompleteness (*i.e.*, how well features in the data capture reality) in the data set (Bolstad, 2008).

Misclassification error does not have a single clear antecedent and is influenced by many factors. The source of error may stem from sample size, sample design, model misspecification, inference assumptions, positional uncertainty, scale misalignment, surrogate or restricted ground observations, and obfuscation of corrections, assumptions, and tolerances (Foody, 2002). Misclassification error causes mapped land-use to differ from true land-use, and although accuracy assessments attempt to disclose this disconnect often publications fail to even include this information (Olofsson et al., 2013). Our approach is distinct from remote sensing model assisted estimation techniques (*e.g.*, Foody et al., 1992; Canters, 1997; Stehman, 2009; McRoberts, 2011) due to its construction within an econometric framework and application for practitioners one step removed from the data generating process. Yet there are parallel themes that arise in deriving the conditional probabilities of land use consistently and in the fundamental understanding of the causes and effects of misclassification.

In the context of empirical land-use models, misclassification errors imply measurement error in the dependent variable of a discrete choice model. Unlike measurement error in the classic linear regression model (which only reduces the efficiency of parameter estimates), misclassification error leads to inconsistent and inefficient parameter estimates in discrete choice models (Hausman et al., 1998; Neuhaus, 1999; Hausman, 2001). Although its presence is well known and has long been considered in the epidemiology literature (Copeland et al., 1977; Magder and Hughes, 1997; Neuhaus, 1999; Lewis et al., 2012), misclassification error has largely been ignored in the land-use literature. Wright and Wimberly (2013), for example, acknowledge misclassification and use a raw data correction approach (*i.e.*, spatial smoothing) to calculate rates of land-use change using satellite imagery data. They do not, however, apply an empirical model and therefore do not consider how misclassification errors may propagate (see Kline et al., 2013).

We expand previous methods of accounting for misclassification error to make them directly applicable to empirical land-use modeling. This model specification is functionally equivalent to Hausman et al. (1998), however the interpretations of specific parameter are adapted to those commonly used in the land-use literature. We further expand this specification to account for multi-use land-use models in a more general model. This specification can allow for different misclassification levels across each land-use. We therefore expand the theory and estimation methods of Hausman et al. (1998) to account for this additional complexity. Our method is particularly applicable to studies that use satellite imagery as the basis for an econometric model of the probability of use in a cross-sectional setting. We demonstrate these developments using an application of land use in the US Northern Great Plains.

Our empirical results demonstrate that bias caused by ignoring misclassification can be substantially significant to affect policy decisions. These models are often used to derive the implications of alternative policies (*e.g.*, subsidies) or external shocks (*e.g.*, climate change). In our example we show that misclassification leads to errors in estimating the effects of alternative policies and external shocks by an order of magnitude larger than 350%. Simply put the errors constitute misconstruing areas larger than Yellowstone National Park in the United States over a total population study area only 13 times larger.

2. Theoretical modeling of land-use

Consider a landowner who faces a decision of what to do with their land. We will assume this landowner has a utility function and chooses the land-use that maximizes their utility (see Segerson et al., 2006). Specifically, the owner of plot i faces a choice among j land-use alternatives and receives varying amounts of utility (U_{ij}) from each land-use