Comparing Expert and Non-expert Conceptualisations of the Land: An Analysis of Crowdsourced Land Cover Data

Alexis Comber[1], Chris Brunsdon[2], Linda See[3], Steffen Fritz[3], and Ian McCallum[3]

[1] Department of Geography, University of Leicester, Leicester, LE1 7RH, UK
[2] Department of Geography, University of Liverpool, Liverpool, L69 3BX UK
[3] International Institute of Applied Systems Analysis (IIASA), A-2361, Laxenburg, Austria
ajc36@le.ac.uk, Christopher.Brunsdon@liverpool.ac.uk,
{see,fritz,mccallum}@iiasa.ac.at

Abstract. This research compares expert and non-expert conceptualisations of land cover data collected through a Google Earth web-based interface. In so doing it seeks to determine the impacts of varying landscape conceptualisations held by different groups of VGI contributors on decisions that may be made using crowdsourced data, in this case to select the best global land cover dataset in each location. Whilst much other work has considered the quality of VGI, as yet little research has considered the impact of varying semantics and conceptualisations on the use of VGI in formal scientific analyses. This study found that conceptualisation of cropland varies between experts and non-experts. A number of areas for further research are outlined.

Keywords: Volunteered Geographical Information (VGI), Land Cover, Geo-Wiki, Geographically Weighted Kernel.

1 Introduction

The increase in web-based technologies has resulted in many new forms of data creating and sharing. Individual citizens generate, upload and share a wide range of different types of data via online databases, much of which increasingly has a spatial reference. There is the potential of what is referred to as volunteered geographic information (VGI) [1] to change the nature of scientific investigation as one of its critical advantages of VGI is the potential increase in the volumes of data describing spatially referenced phenomena. Such citizen science activities are supported by the rise of digital, location enabled technologies which offer increased opportunities for the capture of spatial data and for citizens to share information about all kinds of processes and features that they observe in their daily lives.

One of the critical issues to be overcome for crowdsourced data to be included in scientific investigations relates to the quality of the information. In much scientific research data are collected under a formal experimental design which frequently includes consideration of sampling frameworks, quality assurance checks etc. Data

T. Tenbrink et al. (Eds.): COSIT 2013, LNCS 8116, pp. 243–260, 2013.

are collected using well-established methods, by particular instruments or by people with appropriate training and expertise. Thus one of the critical issues in using crowdsourced data relates to the nature of crowd, their familiarity with the domain under investigation and consideration of any impacts of a lack of expertise on the quality of the data that are collected.

This paper generates measures of correspondence between crowdsourced data about land cover / land use and different global land cover datasets. The correspondences between locations identified as being 'cropland' by volunteers was statistically related to measures of the proportions of cropland at those locations from the global datasets. The largest correspondences were used to infer which global dataset best predicted the cropland identified by the crowd. The variation in these correspondences and inferences were examined by considering data contributed by remote sensing experts and by non-experts – ie with contrasting degrees of domain familiarity. In this way the paper explores the impacts of the differences between naïve and expert contributors of VGI and the conceptualisations of landscape features that they hold. This study uses data on land cover, collected as part of the Geo-Wiki project (www.geo-wiki.org), where the registration process captured self-reported measures of contributor expertise. The crowdsourced land cover data captured in this way has a number of potential applications and could, for example, be used to train or to validate statistical classification of remote sensed data. The results of this analysis will inform future uses of VGI by determining whether the differences in expertise about the domain under consideration (in this case remote sensing of cropland land cover) are important and should be considered in such future work.

2 Background

There are many examples of crowdsourced data being exploited that have resulted in novel scientific discoveries such as unravelling protein structures [2], discovering new galaxies [3], reporting of illegal deforestation [4]. The use of VGI is now commonplace in many areas of scientific investigation, from conservation [5] to urban planning [6], and VGI has been found to be particularly useful in endeavours to manage and understand important emerging problems such as ash dieback [7] and post-disaster damage assessments [8, 9]. The latter exemplify the critical advantage of crowdsourced data: the ability to rapidly collect and share large volumes of data describing many kinds of phenomena. These activities are facilitated by ubiquitous ability to capture and share data using many electronic devices (e.g. digital cameras, smartphones, tablets, etc) that are location-enabled, eg through in-built GPS capabilities. The result is an increasing amount of spatially referenced or geo-located data, captured through ubiquitous and low costs citizen owned sensors, that can directly and instantaneously capture and share data of the immediate environment, that are available for formal scientific analysis. This presents a number of challenges associated with use of VGI in formal analyses that relate to questions over data quality and reliability: