VGI IN AUTOMATIC INFORMATION RETRIEVAL FROM BIG RS DATA

Aneta J. Florczyk, Thomas Kemper, Mayeul Kauffmann, Pierre Soille

European Commission, Joint Research Centre (JRC), Institute for the Protection and Security of the Citizen (IPSC), Global Security and Crisis Management Unit, Via Enrico Fermi 2749, 21027 Ispra VA, Italy

ABSTRACT

This paper describes how Volunteered Geographic Information (VGI) can support automatic information retrieval (IR) from big remote sensing (RS) data on an example of the Global Human Settlement Layer (GHSL)[1]. Here, the OpenStreetMap1 (OSM) derived data are used within a satellite imagery processing framework at several stages of an information layer production. Although, we focus on built-up areas, some similarities could be identified in the methods for extracting occurrence estimation of other geographic feature produced by the RS community (e.g. forest coverage). This work introduces briefly the GHSL, the VGI derived datasets and how they are used. A discussion on levered issues related with the VGI data is provided.

1. INTRODUCTION

Information retrieval (IR) is a digital image processing technique to extract thematic information from the remote sensing imagery by means of analytics. Cloud detection, multi-spectral classification, machine learning or masking are just few examples of methods that can be involved. Many of these techniques (e.g. supervised classification) rely on ancillary data, which are also crucial for the validation of final results. First, fitness-for-use of existing products as reference data is usually evaluated (e.g. Corine Land Cover2, a European land cover product, could be used in tree coverage estimation application). Also, reference data can be produced (semi-)manually (e.g., ground truth exploration, image digitization), which is usually expensive in terms of time and money. For example, the European Land Use and Cover Area frame Survey3 (LUCAS) is an in-situ area frame survey, which means that the data is gathered through direct observations by the surveyors on the ground. Land cover data can also be obtained by photo interpreting satellite images or orthophotos.

As the RS data share all characteristics of big data (i.e. the four ‘V’), there is a need for developing computationally effective IR methods to process automatically multiple scenes with multi-resolution, multi-platform, multi-sensor and multi-temporal characteristics. In this context, also collection of reference datasets becomes a challenge, as even semi-automatic generation is not feasible, and existing products may not meet the requirements (e.g. precision, semantics, and geographic extent).

The scientific community recognizes the value of VGI, especially in preliminary steps of scientific research [2,3]. Although there is prudency in the data usage in practice mainly due to quality and licensing issues, there is growing interest in VGI data application [4,5]. There are some evidences that selected thematic subsets of OSM can be exploited for some specific application of IR from remote sensing imagery. For example, [6] and [7] show that OSM information can be better than authoritative datasets in some aspects (semantic accuracy and completeness, respectively). In general, the ongoing discussion between VGI and other communities, promises that the current obstacles can be mitigated in the future [3].

In this work we present an application of OSM derived datasets in RS imagery processing. First, we describe the usage of the reference datasets derived from OSM, including some details on their generation. Finally we provide discussion in the context of the experience.

2. VGI IN BIG RS DATA SCENARIO

An example of using VGI in big RS data processing is a map of European settlements (enhanced with green space index), derived from a large, high-resolution satellite image data collection produced in the GHSL production platform [1]. The end-to-end processing of the3143 SPOT-5 scenes (3 bands each) to produce 2.5m and 10m resolution maps of European settlements [8] took 22 days pure processing time. The image processing workflow applied uses thematic layers as references, both low- and high-resolution data. Three of them are derived from OSM data: i) the high resolution coastline, ii) the roads and iii) the building footprints. In general, we can identify four main stages when those reference datasets are required: learning, classification, post-processing and validation. The coastline data are very important in the learning stage in order to stabilize the inference engine by excluding the water pixels. The road dataset is used to differentiate between road green and other green areas (e.g. parks and gardens). These datasets are also used in post-processing stage to reduce error introduced by the IR method applied. Cars, roundabouts or even enter/exit lines of motorways can be detected as built-up, and in coastal areas some natural features (e.g. waves). In general, the masking task in post-processing stage removes noise and is important for visualization purposes. Finally, the validation is based on measurement of quality performance of a GHSL product vs. a reference dataset. The GHSL product has been validated against building footprints (part of which has been extracted

1 http://wiki.openstreetmap.org
2 http://www.eea.europa.eu/publications/COR0-landcover
from OSM for large areas (e.g. the Netherlands, Toscany) and for the entire continent with the LUCAS dataset (more than 250,000 points). In both cases, the GHSL performance is similar.

Four main steps can be identified in the production of the OSM derived datasets. First, the thematic subset extraction produces a GIS-compatible subset of information from OSM data. Next, the feature-based and dataset-based quality check and cleaning have been done by experts. Then, the fitness-for-use has been evaluated. The coast line and roads data have been identified as precise enough for the European application and selected scale. Special attention has been paid to the completeness and level of details of the building footprint dataset. Due to the data inconsistency, the pseudo-cadastral dataset has been limited to some areas of the Netherlands.

3. DISCUSSION

In general, gathering ground information for reference datasets is expensive. A common approach is a manual (or semi-manual) digitization, which in a one-scene-processing scenario has an acceptable cost. However, in case of big data scenarios (especially the global coverage and high resolution data processing), this approach fails. It is a crucial task to identify and collect existing datasets that meet quality and semantic requirements. When considering VGI as an alternative, some questions arise. Are the VGI sources, such as OSM data, appropriate for big RS data applications, and what are those applications? If yes, in which RS tasks may they be useful? What are the challenges when using VGI? Which quality metrics shall be offered, how should they be generated and represented? What kind of tools can be useful for estimation of the fitness for use? Are there other domains, where the VGI can help the RS community in their work? And how can the RS community help the VGI? The work presented partially responds to some of those questions. Although known problems of OSM data quality, it has been used to produce reference datasets due the global coverage, thematic subset extraction run took a few months) and quality (an effort of experts) aspects. The absence of a priori metrics on spatial precision, accuracy and completeness of the building footprint dataset. Due to the data inconsistency, the pseudo-cadastral dataset has been limited to some areas of the Netherlands.

4. REFERENCES