Chapter 3
Traditional vs. Machine-Learning Techniques for OSM Quality Assessment

Musfira Jilani
National University of Ireland, Ireland

Padraig Corcoran
Cardiff University, UK

Michela Bertolotto
University College Dublin, Ireland

Amerah Alghanim
University College Dublin, Ireland

ABSTRACT

Nowadays an ever-increasing number of applications require complete and up-to-date spatial data, in particular maps. However, mapping is an expensive process and the vastness and dynamics of our world usually render centralized and authoritative maps outdated and incomplete. In this context crowd-sourced maps have the potential to provide a complete, up-to-date, and free representation of our world. However, the proliferation of such maps largely remains limited due to concerns about their data quality. While most of the current data quality assessment mechanisms for such maps require referencing to authoritative maps, we argue that such referencing of a crowd-sourced spatial database is ineffective. Instead we focus on the use of machine learning techniques that we believe have the potential to not only allow the assessment but also to recommend the improvement of the quality of crowd-sourced maps without referencing to external databases. This chapter gives an overview of these approaches.

INTRODUCTION

While originally only *spatial professionals* (including cartographers, remote sensing scientists, photogrammetrists, etc.) were able to handle and exploit spatial data in standalone dedicated systems, nowadays a wide range of popular products and applications, used regularly by non-professionals, rely on such data (in particular maps). These include emergency response systems, location-based services, digital globes, etc. Many of these systems provide critical services and therefore require reliable and up-to-date map data. However, mapping is an expensive process which has, until recently, been carried out exclusively by official agencies and specialised companies. Because of the costs involved, authori-

DOI: 10.4018/978-1-5225-2446-5.ch003

tative maps are usually not entirely up-to-date as they are not very often edited to reflect the dynamic changes occurring, for example, in urban environments. Indeed, it has been argued that even in the most advanced countries such as the United States, the official maps produced by the US Census Bureau lack in detail and temporal accuracy primarily due to budget constraints and slow update cycles (Haklay, 2010). Furthermore, because these datasets are expensive to produce, they are often sold at high prices to users that want to incorporate them into their applications and products. This can be an impediment for many ordinary customers and companies.

It is precisely to overcome these problems that the OpenStreetMap (OSM) project started. Founded in July 2004 by Steve Coast with the aim of creating a freely accessible and editable map of the entire world (OSMwiki, 2015), OSM is considered as one of the most successful examples of Volunteered Geographic Information (VGI) (Goodchild, 2007). This new phenomenon, which refers to the crowd-sourcing of spatial (geographic) information, has been facilitated by the integration of cheap Global Positioning Systems (GPS) sensors in many popular mobile devices and the general adoption of advanced technologies such as the Web 2.0.

Similarly to other large crowd-sourced datasets collected and made available online to all, VGI repositories present both advantages and disadvantages. In particular, in addition to the lack of costs and ease of use of these datasets, benefits include the fact that the local knowledge of contributors allows them to add much finer detail (including vernacular names of places, specific features, etc.), which is typically not included by official agencies. Moreover, the global diffusion of crowd-sourcing technologies have the potential to create billions of volunteers (or human sensors) world-wide, thereby providing OSM (and other VGI projects such as wikimapia) with an unprecedented manpower that can update and refine the data in real time. Indeed, as remarked by many (Goodchild, 2007), these projects (OSM, wikimapia, etc.) are able to give rise to previously unknown applications.

However, the availability of these large amounts of data created by volunteers from all different types of backgrounds poses several different challenges. These range from the difficulty to efficiently store and manage this type of Big data, to the effective extraction of relevant information from it that matches the needs of users, to the reliability of this data for the application at hand. Indeed, while some parts of the OSM database have been mass imported from sources such as the TIGER database of the US Census Bureau (Zielstra et al., 2013), most of the map data is produced by local contributors which do not necessarily have any experience in mapping practices. As a consequence, many concerns regarding the quality and credibility of this type of geographic information arise.

Quality assessment is an important research issue because, in order to be able to rely on the use of a dataset for their applications, users want to know that it is of high quality. Even knowing that the data is not of very good quality is an important piece of information that provides an estimate of the confidence with which the data can be used. Therefore, many researchers have dedicated their efforts to tackling this issue. In particular, some authors (e.g., (Guptill and Morrison, 2013) and (Kresse & Fadaie, 2004)) have identified several dimensions of spatial data quality including geometric accuracy, attribute accuracy, temporal accuracy, lineage, logical consistency, and completeness. While the majority of studies have considered geometric accuracy in their approaches to quality assessment, other dimensions, including semantic accuracy did not receive the same level of attention.

The work presented in this chapter covers a brief overview of proposed approaches for analysing the quality of VGI, with particular emphasis on OSM which is the focus of the greatest majority of the literature. We do not claim to provide an extensive review of all the approaches, as this is covered in other chapters. Instead we focus more on approaches based on machine learning techniques, which have