Ten Years of OpenStreetMap Project: Have We Addressed Data Quality Appropriately? – Review Paper

Dr. Maythm al-Bakri
Lecturer
Department of Surveying
College of Engineering
University of Baghdad
Email: m.m.m.s.albakri@gmail.com

ABSTRACT

It has increasingly been recognised that the future developments in geospatial data handling will centre on geospatial data on the web: Volunteered Geographic Information (VGI). The evaluation of VGI data quality, including positional and shape similarity, has become a recurrent subject in the scientific literature in the last ten years. The OpenStreetMap (OSM) project is the most popular one of the leading platforms of VGI datasets. It is an online geospatial database to produce and supply free editable geospatial datasets for a worldwide. The goal of this paper is to present a comprehensive overview of the quality assurance of OSM data. In addition, the credibility of open source geospatial data is discussed, highlighting the difficulties and challenges of VGI data quality assessment. The conclusion is that for OSM dataset, it is quite difficult to control its quality. It therefore makes sense to use OSM data for applications do not need high quality spatial datasets.

Key words: OpenStreetMap; VGI; spatial data quality; geometrical similarity; positional accuracy

عشر سنوات من بداية مشروع OpenStreetMap : هل تمت السيطرة على جودة بياناته

م.د. ميثم مطشر شرقي
قسم هندسة المساحة
كلية الهندسة / جامعة بغداد

الخلاصة

أثبتت الدراسات الحديثة المتعلقة بتحليل البيانات المكانية ان الاتجاهات البحثية المستقبلية في معالجة هذه البيانات ستتركز على دراسة طبيعة وجودة البيانات المكانية المجانية المنشورة على شبكة الانترنيت Volunteered Geographic Information (VGI). يُنتج هذا النوع من الخرائط عن طريق جمع البيانات الجغرافية بالمسح الأرضي باستخدام مستقبلات نظام التموضع العالمي المحمولة Global Positioning System (GPS) المتوفرة حالياً في اغلب اجهزة الهاتف المحمول ، و كذلك بالاستعانة بمصادر حرة أخرى كالصور الفضائية المجانية المنشورة على شبكة الانترنيت. يمكن للمستخدمين تحرير المسارات والطرق وتحديثها من خلال وسائل التحرير المتاحة. شهدت العشر سنوات الاخيرة دراسة وتحليل دقة وتجانس مواقع واشكال العوارض على الخرائط المنتجة على شبكة الانترنيت وبالاخص مشروع خرائط الشارع المفتوح OpenStreetmap (OSM). خريطة الشارع المفتوح تقدم خريطة حرة للعالم بأكمله قابلة للتحرير من قبل أي شخص، شبيهة بطريقة عمل موسوعة ويكيبيديا. يهدف هذا البحث إلى تقديم نظرة شاملة عن ضمان جودة بيانات OSM ، بالإضافة الى مناقشة مصداقية

المصدر المفتوح للبيانات الجغرافية المكانية، وتسليط الضوء على الصعوبات والتحديات التي تواجه مستخدمي VGI في تقييم جودة بياناتها. إستنتج هذا البحث انه من الصعب جدا السيطرة على جودة بيانات OSM لذا فمن المنطقي استخدام هذه البيانات لتطبيقات لا تحتاج الى دقة عالية ، وللاغراض الاستطلاعية وعمل الدراسات الاحصائية.

1. INTRODUCTION

The past 30 years has seen the start and the development of the Internet technologies, which were initially used to obtain information only, **Harris, 2008**. However, in recent years, the advancement of web technologies has favoured the design of new patterns and practices models on the web extending beyond passive receiving of data. These cumulative developments are grouped under a common concept known as Web 2.0. The first official introduction of the term Web 2.0 was in the first conference of Web 2.0 (O'Reilly Media) by Tim O'Reilly in October 2004. Although the term Web 2.0 indicated a major change in the approaches of software developers on the web, it did not mean a new version of the World Wide Web. Web 2.0 has been described essentially as being a platform that can collect together different sites and software and make them easily available and useable to users, **O'Reilly, 2005**.

The emerging of Web 2.0 technologies has led to significant changes in the methods of producing, processing, sharing and spreading information through the Internet, **Rinner et al., 2008**. One consequence is enabling users to collaborate and interact among each other more easily and effectively. This is represented by the availability of a variety of social networking and communicating sites such as Facebook, Twitter, Flickr and You Tube. By using these facilities, it is possible to upload pictures or videos for public sharing; at the same time, it is also possible to make comments on the postings of others. Both of these practices facilitate the sharing of huge amounts of information on the web. Hence, nowadays not only professional users, but also non-experts, can generate and publish information on the Internet. This free sharing system was known as User Generated Content (UGC), **Krumm et al., 2008**.

Since the term UGC refers to a variety of activities and applications, it is difficult to offer a standard definition for it. For instance, in their review of UGC, **Vickery and Wunsch-Vincent, 2007** identified the meaning of UGC as the ability of creating publicly, data on the Internet which can be achieved by amateurs or professionals with limited creative efforts. However, other authors did not accept this description as a common UGC definition; for example, **Ochoa and Duval, 2008** reported that the UGC concept may be considered local rather than universal, as uploaded data may be available only for a specific group and not for common usage, or it may be simply rearranging such information and not making new contributions. Nonetheless, generally speaking, UGC can refer to information or media that may appear on the web which was contributed by volunteers without anticipation of any type of income, **Krumm et al., 2008**.

A classical example of this kind of provision of information is Wikipedia (the free encyclopaedia). Wikipedia was originally established and founded by Larry Sanger and Jimmy Wales in 2001, **Miliard, 2008**. As Wikipedia adopts an open model for uploading and editing the contributions of others, which are in most cases articles, the numbers of the registered users and the articles in Wikipedia have increased significantly according to its statistics. However, the question of accuracy will arise when comparing this freely available data with professional productions. Inaccurate structure, bad quality and wrongly edited articles might be expected. Nevertheless, this is not the case in some situations when the free data is created by a group of people rather than single person, a factor that has been emphasised by, **Goodchild and Glennon,**

2010. They reported that the information obtained from a few contributors will be less accurate than that obtained from many people. Furthermore, the Wikipedia community has developed its regulations and rules through a specific section of wiki space. This offers an opportunity to members to contact each other and decide upon standards for documented data in Wikipedia. This is why the concept of UGC expanded to several fields rapidly. For instance, the trend towards user-generated content which can be provided or shared online has had profound impacts on the geo-data scene.

By using Web 2.0 practices, an amateur can readily upload geo-data on the Internet. For example, nowadays any person can pick up the geographical information regarding their routes by using GPS in their driving or biking activities. Then, it is possible to contribute to updating and extending existing road databases on the web. It is also possible to add names or photographs to these datasets by means of the geotagging process. As these data are typically produced by volunteers, they have been labelled by **Goodchild, 2007** as volunteered geographic information (VGI). There are many alternative names and definitions for the phenomenon of geospatial information on the web. For instance, the term 'geospatial information bottom-up' was used by **Bishr and Kuhn, 2007** to refer to geo-data on the Internet, whereas according to a definition provided by **Turner, 2006**, this kind of information was coined as 'neogeography' and the same concept was also used by **Haklay et al., 2008**. On the other hand, for **Sui, 2008**, VGI means 'geography without geographers'. Whatever concepts that have been used to describe the open spatial data enabled on Web 2.0, the term VGI is the most widely adopted by many authors; see for example, **Mooney et al., 2010**, **Coleman et al., 2009** and **Elwood, 2008**. The term VGI was generally used to refer to creating, disseminating and updating geospatial data voluntarily on websites. This basically means combining the efforts of individuals or collaborative communities in such a way as to supply this new kind of geospatial data. Similar to UGC applications, VGI data can be effectively utilised by people other than the producers without any restrictions or rules.

VGI is similar to other open source productions in that producers and users of VGI data come from a variety of backgrounds because any person can become involved in this activity. In addition, there are no standard methods for uploading this kind of data. Since most VGI data are created by non-professionals, interest in integrating VGI with formal data, for instance, to develop and update formal datasets, may raise some concerns. From this point of view, **Elwood, 2008** highlighted the need to investigate the different types of VGI with specific emphasis on examining the impacts of VGI services, such as tools and procedures that were used to collect, create and share this data, on the accuracy and the validation of using VGI for multiple purposes. Therefore, these databases are easily subject to heterogeneity and errors. However, VGI does suggest a new and powerful approach that could be efficiently used to create up to date datasets. Real time geospatial data that can be obtained from VGI are necessary for several purposes such as emergency actions; for example, **Zook et al., 2010** reported that the flexibility of free web-based mapping services played a major role in aiding and rescuing people when an earthquake hit Haiti in January 2010. After the earthquake free information technologies such as aerial photographs were used in order to supply the necessary information about the most devastated areas and to produce route maps to deploy resources. Therefore, VGI can offer the most interesting and the cheapest geographic information to users and sometimes it will be the only source of information, especially for remote areas, **Goodchild, 2007**.

In view of the fact that this article focuses on OpenStreetMap (OSM) data quality investigations in general, as OSM is the leading example of VGI projects that are concerned with geospatial

data development around the world, the OSM project will be discussed in greater detail than other VGI data types.

2. CHARACTERSTICS of OPENSTREETMAP

The OpenStreetMap (OSM) is an online geospatial database launched in England (London) by Steve Coast in 2004, **Chilton, 2009**. In particular, it aims to produce and supply free editable geospatial datasets for a worldwide audience. The OSM fundamentally relies upon the collaborative volunteers' contributions for collecting and uploading geographic data to the common data base on the Internet, **Ciepluch et al., 2009**. In general, contributors can collect the OSM data by controlling handheld portable GPS devices (navigation mode) such as the Garmin series. Nowadays, it is also possible to use built-in GPS applications which are available in most mobile phones models such as iPhone. In order to map a certain area using GPS technique, for instance, the OSM community gathers volunteers through an activity called 'mapping parties'. An example of this can be found in the study carried out by **Perkins and Dodge, 2008** in which they illustrated a case study of a mapping party in Manchester, UK, in 2006. Although the GPS receivers may probably be considered as being the most important information source for the OSM project, there are also alternative data sources such as tracing data Yahoo imagery and /or Landsat images, **Ramm et al., 2011**.

The OSM database allows for every user to reproduce or edit its datasets without the necessity for any authorization although attributing the data to OSM is required. Thus, the OSM project is technically similar to the Wikipedia (free encyclopaedia) concept. The users of these systems are able to modify and add or even delete the contributions made by others. The underlying OSM map data can be uploaded by creating a user account and edited online through a wiki-like interface. There are many other services that provide mapping on the Internet freely. For example, Microsoft offers Bing Maps, and Yahoo Maps and Google Maps are readily available. However, the users of these alternative map sources have only been provided with a very limited right to use their datasets. It is not permitted for users of these services to edit or update their datasets. Compared to the OSM data, there are several restrictions and conditions for using the Google Map service, as illustrated in **GoogleMaps, 2012**. For example, the raw data of Google Maps is not available to the users at all; however, it can be used by commercial companies such as TeleAtlas and Navteq, as they pay for downloading these, while OSM data can be downloaded by any user. Consequently, the OSM project can be considered as being one of the most useful online mapping services in that it is suitable for education in schools and undergraduate studies. The survey conducted by **Bartoschek and Keßler, 2013** revealed that OSM is the most renowned online mapping service among students. The above mentioned positive aspects of OSM data, in addition to the possibility of using OSM as a base map for studies in cartography, make OSM a flexible tool in education.

The main other difference that may be noticed when comparing OSM data to other public mapping services is the level of detail as far as features are concerned. It is clear from **Fig. 1** that the site of Newcastle upon Tyne appears more complete in the OSM map than in the Google Maps version. However, the levels of detail of OSM maps vary around the world, **Ramm et al., 2011**. There are some places, such as the UK, that are mapped very well, whereas there are other parts of the world, such as Iraq for example, that have a little coverage for the centres of the big cities only. In fact, there is no detailed data for the countryside or the suburban areas, see **Fig. 2**. The detail of OSM maps is fundamentally based on the number of the volunteers that are available in each place around the globe.

The amounts of OSM data are increasing every day on the Internet. The number of registered users of this project is also growing at a remarkable rate. For example, **Haklay and Weber, 2008** reported that in 2008 the number of registered users of the OSM project was approximately 33,000, while at the time of writing this paper there are about 1,600,000 registered users, see **Fig. 3 and Table 1**, **OSM-stats, 2014**. It is clear that the number of registered people has increased by more than forty five times during the last six years. However, the number who edit is a minority of these. This view is supported by **Neis and Zipf, 2012** who concluded that the rate of the registered users who achieved at least one edit of OSM data was only 38% of the total number of members. They also found that only 5% of the registered members have contributed more than 1000 nodes. Information can also be obtained from the **OSM-stats, 2014** with regard to the total number of uploaded GPS points, nodes, ways and relations for the real time OSM database. These statistics reflect the rapid growth of OSM data on the web, as shown in **Fig. 4 and Table 2**.

There are various aspects to the most important motivations for these developments, for instance, gaining advantages from the free accessibility of OSM data (licence, cost, sharing) and opening up a new paradigm of 'geo-data-people's' Spatial Data Infrastructures (SDI). Access to current OSM data without any charge is available to anybody with web connections. In addition, the wide-ranging coverage of OSM data sources (around the world) allows visitors to search a world map and download different portions from a distance for any part of the world. Although these are positive aspects, the problem of heterogeneous data quality has emerged, **Al-Bakri and Fairbairn, 2011**, as will be discussed in more details in section 3.

3. THE ASSERTION of DATA QUALITY and CREDIBILITY of OPEN SOURCE DATA

Before discussing the problems of VGI data quality, it is necessary to understand the meaning of 'data quality'. Data quality as a concept may be defined differently, depending on the context in which it applies. There are many definitions of data quality in the literature. Each varies from organisation to organisation, application to application or person to person. For instance, the term 'quality' can be defined as an indication of high degree of craftsmanship or creativity, **Veregin, 1999**. In contrast, **Jakobsson**, 2002 regards data quality as a function of the difference between a dataset and the universe of discourse, when the universe of discourse is the actual objective world view and the dataset is the identifiable collection of any related dataset. In terms of spatial data, the notion of quality has been clarified by **Korte, 2001** as being the degree of how accurately the GIS data can be represented or meet a specific accuracy standard.

In most cases, the VGI data on the web may not contain any information about their quality. From this perspective, **Flanagin and Metzger, 2008** supposed that the VGI data may improve spatial data content in general; however, the quality and accuracy of this data has still attracted the most attention to date. There are many reasons making VGI quality information extremely significant. For instance, the increasing of the decision making procedure based on the information of spatial data and the possibility of integrating different datasets which can be used for more GIS analysis and applications. The dependability of VGI data quality should be taken into consideration by people who have been collecting and disseminating this information. VGI data is usually collected by volunteers; thus its quality will vary and nobody can guess or know the value of it. This drawback has been agreed upon by authors such as, **Haklay, 2010** and **Auer and Zipf, 2009**.

There are several legitimate criticisms that make the assessment of VGI quality difficult. For example, there is an enormous variety of people who contribute VGI data and there is no unified

authority whose role is to assess the quality of spatial data. Additionally, because of the different perspectives of data developers, it is highly likely that heterogeneities will be found in resulting datasets, **Elwood, 2009**. This inspired **Exel et al., 2010** to include crowdsourced dynamics as an indicator of crowdsourced spatial data quality determination. They aimed to establish spatial data quality operational indicators for both user and feature quality. Their proposed approach fundamentally considered different crowdsourced activities such as the number of editors or edits per feature and the historical (or temporal) information of the features which includes the development of such features over time. This suggested framework may assist in measuring the density of edits to an area of crowdsourced data and ultimately assessing its data quality.

Elsewhere, **Goodchild and Li, 2012** have argued that although VGI may offer numerous advantages such as the free availability and accessibility of spatial datasets, the quality of VGI data should be considered as a vital issue as VGI data does not follow a standard structural design. Therefore, they investigated three different approaches to assess the quality of VGI data. Such quality assurance approaches are firstly, validation by crowdsourcing, secondly the social approach, relying on a hierarchy of 'trusted' individuals, and finally the geographic approach, which examines the probability of features being correctly located with reference to the surrounding context and geographical area. Subsequently, they compared these approaches with the quality assurance approach that is usually used by traditional mapping agencies. Some analysts (e.g. **Hagenauer and Helbich, 2012**) have pointed out that VGI data quality issues, especially completeness, can affect the fitness for use for such applications (e.g. urban planning). Therefore, they suggested a methodology to calculate through OSM data which urban areas in Europe are mapped or partially mapped. Their results found that the delineations of urban areas are based on the location.

The increase in the amount of data has also led to an increase in the heterogeneity between datasets. For instance, within different datasets, the features may be varying in accuracy due to the methods or skills that were employed for the purpose of collecting data. According to **Haklay, 2010**, the distribution of errors in VGI data is usually based on the carefulness of each contributor. Therefore, the concern of trust of VGI data quality is the main issue facing the GI community. These heterogeneities may be especially problematic when the integration of multi-source spatial datasets is the target, for example.

4. HISTORY of OSM GEOMETRICAL SIMILARITY MEASUREMENT

4.1 POSITIONAL DISCREPANCES

In order to understand the legitimate criticisms that the assessment of VGI quality is difficult, it is necessary to study previous researches and investigations which place emphasis on the assurance of VGI data quality. Furthermore, it is important to present research literatures related to quantitative measures for evaluating VGI quality. For instance, **Ather, 2009** carried out a research to look at the positional accuracy of OpenStreetMap data through comparison with the OS MasterMap dataset. Further map quality tests were also conducted in terms of a completeness study of road name attribution, and an analysis of number of users per area. The results of this analysis found that the positional accuracy of OpenStreetMap data to be very good in comparison to OS MasterMap, with over 80% overlap between most the road objects tested between the two datasets. The results also found there to be a positive correlation between road name attribute completeness and number of users per area.

Haklay, 2010 examined the positional quality of OSM information by comparing it with OS-Meridian 2 datasets. The Meridian 2 dataset supplies detailed data of road networks in Great

Britain such as motorways, minor and major roads. In addition to use more data sources in order to complete this investigation. These involved the 1:10,000 raster files from OS, and some data about the neighbourhood size which is based on Census from OS and national statistics office. The main focus of Haklay's work was limited to the measuring of the quality of roads or motorways of OSM datasets. The findings of Haklay's study also showed that the quality of OSM data is variable when compared to OS datasets within the average of 6m of positional accuracy.

Girres and Touya, 2010 assessed the quality of OSM datasets in France. In their investigation, many quality elements for OSM data were evaluated. The more interesting parameters that will be illustrated and described in this paper is geometric accuracy. Their analysis included the comparison of OSM data with the French National Mapping Agency geographic datasets. The results of their positional analysis of the road intersections indicated that the most frequent positional differences ranged between 2.5m to 10m, and the average value of the positional differences was nearly 6.65 m.

The differences between linear features were calculated by applying two techniques: the Hausdorf distance approach, which computes the maximum distance between the compared linear features, and the average distance approach, which takes the average distance between the compared polylines, a method suggested by McMaster (as cited in **Girres and Touya, 2010**). The principles of these methods can be seen in **Fig. 5**. The results of their study showed that the mean difference values between the compared roads were about 13.57m and about 2.19m for the Hausdorf distance and average distance methods respectively.

Fairbairn and Al-Bakri, 2013 also investigated the positional similarity of OSM dataset. They presented a methodology for assessing positional quality for Volunteered Geographic Information (VGI), such as OpenStreetMap (OSM) data, and authoritative large-scale data, such as Ordnance Survey (OS) UK data and General Directorate for Survey (GDS) Iraq data. The analyses are presented with a user-friendly interface which eases data input, computation and output of results, and assists in interpretation of the comparison. The results showed that a comparison of positional of OS data or GDS data, with those of OSM data, indicates that their integration for large scale mapping applications is not viable.

In another study, **Ming et al., 2013** proposed a quality analysis model for OpenStreetMap crowd sourcing geographic data. Firstly, a quality analysis framework was designed based on data characteristic analysis of OSM data. Secondly, a quality assessment model for OSM data by three different quality elements: completeness, thematic accuracy and positional accuracy was presented. Finally, take the OSM data of Wuhan for instance, the research analysed and assessed the quality of OSM data with 2011 version of navigation map for reference. The result showed that the high-level roads and urban traffic network of OSM data has a high positional accuracy and completeness so that these OSM data can be used for updating of urban road network database.

In **2014, Yang et al.** published a paper in which they developed an approach for integrating VGI POIs and professional road networks. The proposed method first generates a POI connectivity graph by mining the linear cluster patterns from POIs. Secondly, the matching nodes between the POI connectivity graph and the associated road network are fulfilled by probabilistic relaxation and refined by a vector median filtering VMF. Finally, POIs are aligned to the road network by an affine transformation according to the matching nodes. Experiments demonstrate that the

proposed method integrates both the POIs from VGI and the POIs from official mapping agencies with the associated road networks effectively and validly, providing a promising solution for enriching professional road networks by integrating VGI POIs.

4.2 SHAPE SIMILARITY MEASUREMENT

In addition to the investigation of the positional similarity issue, there are several authors have examined various issues and problems that may relate to OSM data quality assessment such as measuring the quality of the shapes. For example, **Haklay, 2010** investigated the linear quality of OSM information by comparing it with OS-Meridian 2 datasets as mentioned above. The methodology which was applied to assess the quality of motorways of OSM data was based on approaches by **Hunter, 1999** and **Goodchild and Hunter, 1997**. The method of buffer was adopted to determine the accuracy of such lengthy objects by applying a certain distance of buffer size for this test. The results of the analysis showed that the average of overlap percentages when comparing OSM with OS datasets were approximately 80%, 88% and 77% for motorways, A-roads and B-roads respectively.

The quality of VGI data has received more attention from the GI community; for example, the investigation that was carried out by **Zielstra and Zipf, 2010**. They studied the quality of the routes and roads of OSM data in Germany. The OSM information has been compared with a commercial dataset known as Tele Atlas. They based their approach on that suggested by **Goodchild and Hunter, 1997** to measure the quality of linear features. Their results found that the overlap percentages between the roads of OSM data and Tele Atlas datasets were $\geq$ 80% for most of the roads in major cities. In addition, they reported that the overlap percentage in towns of medium size was between 50% and 80%. They argued that the results of their comparisons revealed that the accuracy seems quite good and the OSM data can be used for many routing applications. However, they found that there are still shortcomings in the accuracy of the regional OSM datasets.

Girres and Touya, 2010 assessed the shape quality of OSM datasets in France. In their study, they considered the differences between polygonal objects of lakes. The surface distance method which was proposed by Vauglin (as cited in **Girres and Touya, 2010**) was adopted to quantify polygon differences. The method based on the common area of the two compared objects. The *ds* value will be zero if polygon *A* is equal to polygon *B,* while it will be one if *A* is not equal to *B*. The results of this method showed that there is a small difference between the polygons of the comparison datasets.

Koukoletsos et al., 2012 suggested an automated feature-based matching method specifically designed for VGI, based on a multi-stage approach that combines geometric and attribute constraints. It was applied to the OSM dataset using the official data from Ordnance Survey as the reference dataset. The results were then used to evaluate data completeness of OSM in several case studies in the UK. The method combined geometric and attribute constraints (road name and type) in order to deal with heterogeneous datasets, taking into account that attributes may be missing. When tested on OSM (VGI) and ITN (Reference) datasets for the selected rural and urban areas, the process lasted eight and 15 hours, respectively. Data matching proved to be efficient, with matching errors between 2.08% (urban) and 3.38% (rural areas). Data completeness of the VGI dataset is then calculated for smaller areas (tiles), giving more representative results of its heterogeneity.

Neis et al., 2012 outlined the development of Volunteered Geographic Information in Germany from 2007 to 2011, using the OpenStreetMap project as an example. Specifically, they considered the expansion of the total street network and the route network for car navigation. With a relative completeness comparison between the OSM database and TomTom‘s commercial dataset, they showed that the difference between the OSM street network for car navigation in Germany and a comparable proprietary dataset was only 9% in June 2011. The results of thier analysis regarding the entire street network revealed that OSM even exceeds the information provided by the proprietary dataset by 27%. Further analyses showed on what scale errors can be reckoned with in the topology of the street network, and the completeness of turn restrictions and street name information. In addition to the analyses conducted over the past few years, projections have additionally been made about the point in time by which the OSM dataset for Germany can be considered complete in relative comparison to a commercial dataset.

A recent study by **Fairbairn and Al-Bakri, 2013** involved creating a user-friendly interface incorporating quantitative and visual analysis of dataset comparison that could be used to assess geometrical similarity of OSM data. The interface was designed for assessing linear and shape matching, in this case comparing the formal data OS and GDS, with the rigorous field survey (FS), and the informal OSM data. The results of this analysis showed that the shape measurements of the informal OSM data do not match the formal data in any of the case study areas examined.

5. DISCUSSION and CONCLUSION

This paper has shown that there are significant incompatibilities preventing the matching of VGI with formal datasets. However, some of these can be addressed in a positive manner with regard to certain aspects such as the richness of datasets that have been offered by VGI data sources, **Ballatore and Bertolotto, 2011**. For example, the richness of feature type definition in OSM could enhance the value of integration. The growing availability of the subclasses of OSM features may enable a wealth of new opportunities to enhance and update the quality of feature classification of formal or governmental data. For instance, by comparing the children levels (i.e. fourth level) of XML schema trees of OSM datasets with those of formal datasets, as shown in **Al-Bakri and Fairbairn, 2012**, it can be seen that the schema of OSM feature classifications display more distribution and classes especially at the latest or end levels. The dynamic nature in terms of the frequency of updating and gathering detailed features of OSM data have established these aspects of feature types of OSM project as being particularly useful. This can be considered as one positive aspect of OSM or informal datasets which could enable these kinds of spatial datasets to be beneficial or advantageous.

Although the current study demonstrated that a divergence exists when evaluating the geometrical elements of VGI and formal spatial data sources, **Zielstra and Hochmair 2011** showed that the integration of pedestrian routes’ accessibility to transit stations (bus and metro stations) of VGI data, such as the OSM project, into the data produced from Tele Atlas and/or NAVTEQ can be useful in US and German cities. The reason for including OSM data in this integration process as a worthy source of pedestrian routes has been explained by the authors as being a potential rich and valuable source of pedestrians’ data that can be supplied from OSM project. This merit of OSM data has been proved by the findings of the numerical analysis of, **Zielstra and Hochmair 2011**. They showed that the information about pedestrian segments of OSM data can increase the usage of the transit facilities when the commercial data is not available, as in the case studies in Germany and some US cities such as Chicago and San Francisco. It seems possible that these results are due to the effective efforts of OSM

communities in Germany and some US cities to develop comprehensive OSM pedestrian path networks.

The positive aspects of VGI data that were mentioned in Section 2, such as the easy access, free use and the rapid growth of these kinds of datasets, makes it possible to envisage further practical applications of them. For example, the combining of the richness of the OSM database into the addresses of places around the world, by following mapping mashup technologies, may produce a compatible and helpful database. This could be used for tourist or navigation purposes rather than using traditional formal and expensive tourist maps. Another useful VGI application was suggested by **Neis et al. 2010**. They explained that the integration of up-to-date OSM data into the UN Spatial Data Infrastructure for Transportation would make it possible to manage disaster relief by creating crisis maps. The obvious example of this situation was when an earthquake hit Haiti in 2010 and how the use of OSM helped in the rescue of people, as described in introduction.

The availability of other VGI data sources beyond OSM can also be useful and valuable for such applications. The other VGI data sources, especially those which do not provide pure spatial datasets, such as Flickr, may be used for such applications that do not need to consider the spatial data quality elements in the processing flowline. For instance, **Schade et al. 2011** proposed a workflow for using VGI data such as Flickr images for risk detection events. They initially described the positive characteristics of this kind of dataset and why it can be applied for this application. The massive growth of images through this website, the multiplicity of options for uploading images and the ability of users to geotag images can be considered the main positive aspects of this service. As an example of their application, Flickr images were used to detect the risk of flood in the UK for the period from 2007 to 2009. The analysis of the approach taken was basically based on the fundamental information that can be obtained from the Flickr website, such as the location of the picture, the time and date a picture was taken and the time and the data the pictures were uploaded onto the website.

Another example of a benefit gained from VGI data was illustrated by **Spinsanti and Ostermann, 2010**. They developed a methodology to test the framework of assessing the contribution of VGI data to detecting the spread of fire events, and compared this with the official information sources that can be obtained from the European forest fire information system. For their project, the authors used picture data from Flickr and text data from Tweeter. They argued that using VGI data for communicating during disasters is an effective method of crisis management which may reduce the amount of risks and losses.

The above discussion has shown that although this paper concluded that it is difficult to integrate VGI data with formal spatial datasets, there is a wide range of applications that VGI data can serve and assist with; especially those which do not need high quality spatial datasets. Updating transit maps, enhancing pedestrian navigation systems and addressing disaster scenarios are all examples of small-scale topographic data which can be more easily integrated with formal datasets. Large-scale data show more limited promise due to geometric and semantic mismatching. Therefore, the conclusion that formal data providers would not be interested in integrating their datasets with, for example, OSM data is correct. VGI users, on the other hand, are certainly keen to incorporate official datasets into their own, and the results obtained in this research would give confidence to such procedures. The issue of data quality is taken seriously throughout the OSM community, for example. The section on Quality Assurance in the OSM blog shows that a number of tools are available and in development for detection, reporting and monitoring of data quality (Wiki-OpenStreetMap, 2012).

References

Al-Bakri, M. and Fairbairn, D., 2011, *User generated content and formal data sources for integrating geospatial data,* 25th International Cartographic Conference Paris, France, pp. 1-8.

Al-Bakri, M. and Fairbairn, D., 2012, *Assessing similarity matching for possible integration of feature classifications of geospatial data from official and informal sources*, International Journal of Geographical Information Science, 26, (8), pp. 1437-1456.

Ather, A., 2009, *A quality analysis of OpenStreetMap data*. M.Sc. thesis, UCL.

Auer, M. and Zipf, A., 2009, *How do free and open geodata and open sdtandards fit together? from sceptisim versus high potential to real applications*, The First Open Source GIS UK Conference. Nottingham, UK, pp. 1-6.

Ballatore, A. and Bertolotto, M., 2011, *Semantically enriching VGI in support of implicit feedback analysis*, in Tanaka, K., Fr¨ohlich, P. and Kim, K.-S.(eds) Web and Wireless Geographical Information Systems, volume 6574 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg, pp. 78-93.

Bartoschek, T. and Keßler, C., 2013, *VGI in education–from K-12 to graduate studies*, in Sui, D., Elwood, S. and Goodchild, M.(eds) Crowdsourcing Geographic Knowledge. Volunteered Geographic Information (VGI) in Theory and Practice. Springer, pp. 341-360.

Bishr, M. and Kuhn, W., 2007, *Geospatial information bottom-up: A matter of trust and semantics,* The European Information Society - Leading the Way with Geo-information. Springer-Verlag Berlin Heidelberg, pp. 365-387.

Chilton, S., 2009, *Crowdsourcing is radically changing the geodata landscape: case study of OpenStreetMap,* 24th International Cartographic Conference Santiago, Chile, pp. 1-7.

Ciepluch, B., Mooney, P., Jacob, R. and Winstanley, A. C., 2009, *Using OpenStreetMap to deliver location-based environmental information in Ireland*, SIGSPATIAL Special, 1, (3), pp. 17-22.

Coleman, D., Georgiadou, Y. and Labonte, J., 2009, *Volunteered geographic information: the nature and motivation of producers*, International Journal of Spatial Data Infrastructures Research, 4, (1), pp. 332 - 358.

Elwood, S., 2008, *Volunteered geographic information: future research directions motivated by critical, participatory, and feminist GIS*, Geojournal, 72, (3), pp. 173–183.

Elwood, S., 2009, *Geographic information science: new geovisualization technologies – emerging questions and linkages with GIScience research*, Progress in Human Geography, 33, (2), pp. 256–263.

Exel, M. v., Dias, E. and Fruijtier, S., 2010, *The impact of crowdsourcing on spatial data quality indicators,* GIScience 2010. Zurich, Switzerland pp. 1-4.

Fairbairn, D.; Al-Bakri, M., 2013, *Using geometric properties to evaluate possible integration of authoritative and volunteered geographic information*, ISPRS Int. J. Geo-Inf., 2, pp. 349–370.

Flanagin, A. J. and Metzger, M. J., 2008, *The credibility of volunteered geographic information*, Geojournal, 72, pp. 137–148.

Girres, J.-F. and Touya, G., 2010, *Quality assessment of the French OpenStreetMap dataset*, Transactions in GIS, 14, (4), pp. 435-459.

Goodchild, M. F. and Hunter, G. J., 1997, *A simple positional accuracy measure for linear features*, International Journal of Geographical Information Science, 11, (3), pp. 299 - 306.

Goodchild, M. F., 2007, *Citizens as sensors: the world of volunteered geography*, Geojournal, 69, (4), pp. 211–221.

Goodchild, M. and Glennon, A., 2010, *Crowdsourcing geographic information for disaster response: a research frontier*, International Journal of Digital Earth, 3, (3), pp. 231-241.

Goodchild, M. F. and Li, L., 2012, *Assuring the quality of volunteered geographic information*, Spatial Statistics, 1, pp. 110-120.

GoogleMaps, 2012, *Google maps/earth additional terms of service* Available at: http://maps.google.com/help/terms_maps.html (Accessed: 27-05-2014).

Hagenauer, J. and Helbich, M., 2012*Mining urban land-use patterns from volunteered geographic information by means of genetic algorithms and artificial neural networks*, International Journal of Geographical Information Science, 26, (6), pp. 963-982.

Haklay, M., 2010, *How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets*, Environment and Planning B: Planning and Design, 37, (4), pp. 682 -703.

Haklay, M., Singleton, A. and Parker, C., 2008, *Web mapping 2.0: The neogeography of the GeoWeb*, Geography Compass, 2, (6), pp. 2011-2039.

Haklay, M. M. and Weber, P., 2008, *OpenStreetMap: user-generated street maps*, IEEE Pervasive computing, 7, (4), pp. 12-18.

Harris, D., 2008, *Web 2.0 evolution into the intelligent Web 3.0.* London, UK: Emereo Pty Ltd.

Hunter, G. J., 1999, *New tools for handling spatial data quality: moving from academic concepts to practical reality*, URISA Journal, Vol. 11, (No. 2), pp. 25-34.

Jakobsson, A., 2002, *Data quality and quality management - examples of quality evaluation procedures and quality management in European national mapping agencies*, in Shi, W., Fisher, P. F. and Goodchild, M. F.(eds) Spatial Data Quality. London: Taylor & Francis, pp. 216–229.

Korte, G. B., 2001, *The GIS Book, How to Implement, Manage, and Assess the Value of Geographic Information Systems.*5th ed Albany, New York: Onword Press.

Koukoletsos, T., Haklay, M. & Ellul, C., 2012, *Assessing data completeness of VGI through an automated matching procedure for linear data*, Transaction in GIS, 16, pp. 477-498.

Krumm, J., Davies, N. and Narayanaswami, C., 2008, *User-generated content*, Pervasive Computing, IEEE, 7, (4), pp. 10-11.

Mooney, P., Corcoran, P. and Winstanley, A. C., 2010, *Towards quality metrics for OpenStreetMap,* GIS: Proceedings of the ACM International Symposium on Advances in Geographic Information Systems. San Jose, CA, pp. 514-517.

Miliard, M., 2008, *Wikipediots: who are these devoted, even obsessive contributors to Wikipedia*, City Weekly. Available at: http://www.cityweekly.net/utah/article-5129-feature%20wikipediots-%09who%20are-these-devoted-even-obsessive-contributors-to-%09wikipedia.html.

Ming, W., Qingquan, L., Qingwu, H. & Meng, Z., 2013, *Quality analysis of OpenStreetMap data*, International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XL-2/W1, 8th International Symposium on Spatial Data Quality , 30 May - 1 June 2013, Hong Kong, pp. 155-158.

Neis, P., Singler, P. and Zipf, A., 2010, *Collaborative mapping and emergency routing for disaster logistics -Case studies from the Haiti earthquake and the UN portal for Afrika*, Geoinformatik 2010. Kiel, Germany, pp. 1-6.

Neis, P. and Zipf, A., 2012, *Analyzing the contributor activity of a volunteered geographic information project -The case of OpenStreetMap*, ISPRS International Journal of Geo-Information, 1, (2), pp. 146-165.

Neis, P., Zielstra, D. & Zipf, A., 2012, *The street network evolution of crowdsourced maps: OpenStreetMap in Germany 2007-2011*, Future Internet, 4, pp. 1-21.

Ochoa, X. and Duval, E., 2008, *Quantitative analysis of user-generated content on the web*, Proceedings of the First International Workshop on Understanding Web Evolution Beijing, China, pp. 19-26.

OSM-stats, 2014, *OpenStreetMap stats*. Available at: http://www.openstreetmap.org/stats/data_stats.html (Accessed: 21-05-2014).

O'Reilly, T., 2005, *What is Web 2.0: design patterns and business models for the next generation of software*. Available at: http://oreilly.com/web2/archive/what-is-web-20.html?page=1 (Accessed: 11-03-2014).

Perkins, C. and Dodge, M., 2008, *The potential of user-generated cartography:a case study of the OpenStreetMap project and Mapchester mapping party*, North West Geography, 8, pp. 19-32.

Ramm, F., Topf, J. and Chilton, S., 2011, *OpenStreetMap - using and enhancing the free map of the world.* Cambridge, England: UIT Cambridge Ltd.

Rinner, C., Kebler, C. and Andrulis, S., 2008, *The use of Web 2.0 concepts to support deliberation in spatial decision-making*, Computers, Environment and Urban Systems, 32, (5), pp. 386-395.

Schade, S., Díaz, L., Ostermann, F., Spinsanti, L., Luraschi, G., Cox, S., Nuñez, M. and Longueville, B. D., 2011, *Citizen-based sensing of crisis events: sensor web enablement for volunteered geographic information*, Applied Geomatics, Online First™, 19 July 2011. DOI 10.1007/s12518-011-0056-y, pp. 1-16.

Spinsanti, L. and Ostermann, F. O., 2010, *Validation and relevance assessment of volunteered geographic information in the case of forest fires*, 2nd International Workshop on Validation of Geo-Information Products for Crisis Management. Ispra, Italy, pp. 1-9.

Sui, D., 2008, *The wikification of GIS and its consequences: Or Angelina Jolie's new tattoo and the future of GIS*, Computers, Environment and Urban Systems, 32, (1), pp. 1-5.

Turner, A. J., 2006, *Introduction to Neogeography.* Short Cuts Series: O'Reilly Media, Inc.

Veregin, H., 1999, *Data quality parameters*, in Longley, P., Goodchild, M., Maguire, D. and Rhind, D.(eds) *Geographical Information Systems Principles and Technical Issues.* USA: John Wiley & Sons, Inc, pp. 177-189.

Vickery, G. and Wunsch-Vincent, S., 2007, *Participative Web and User-Created Content: Web 2.0, wikis and social networking.* USA: OECD.

Wiki-OpenStreetMap, 2013, *State of the map 2012/call for venues/Tokyo*. Available at: http://wiki.openstreetmap.org/wiki/State_Of_The_Map_2012/Call_for_venues/Tokyo (Accessed: 03-06-2014).

Yang, B., Zhang, Y. & Lu, F., 2014, *Geometric-based approach for integrating VGI POIs road networks*, International Journal of Geographical Information Science, 28, (1), pp. 126-147.

Zielstra, D. and Zipf, A., 2010, *Quantitative studies on the data quality of OpenStreetMap in Germany*, Sixth International Conference on Geographic Information Science,GIScience 2010. Zurich, Switzerland, pp. 1-7.

Zielstra, D. and Hochmair, H., 2011, *Comparative study of pedestrian accessibility to transit stations using free and proprietary network data*, Transportation Research Record: Journal of the Transportation Research Board, 2217, pp. 145-152.

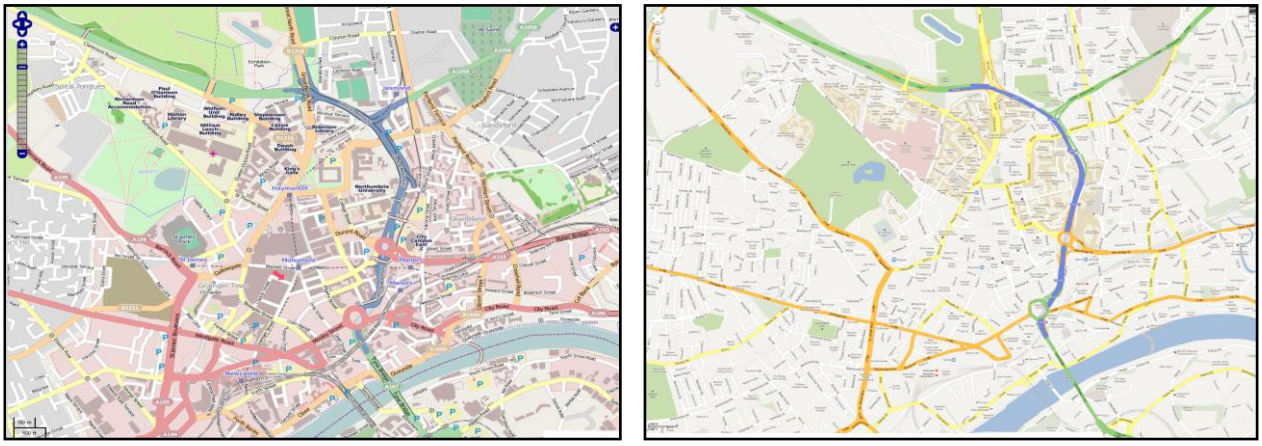
Zook, M., Graham, M., Shelton, T. and Gorman, S., 2010, *Volunteered geographic information and crowdsourcing disaster relief: a case study of the Haitian earthquake*, World Medical & Health Policy, 2, (2), pp. 7-33.

Table 1. The growth of OSM registered users (OSM-stats, 2014).

Date	Numbers of OSM Registered Users
Aug 2005	1,000
Jan 2006	1,500
Aug 2006	2,500
Jan 2007	3,000
Aug 2007	5,000
Jan 2008	10,000
Aug 2008	50,000
Jan 2009	100,000
Aug 2009	150,000
Jan 2010	200,000
Aug 2010	300,000
Jan 2011	350,000
Aug 2011	450,000
Jan 2012	530,000
Aug 2012	700,000
Jan 2013	1,000,000
Aug 2013	1,300,000
Jan 2014	1,500,000

Table 2. The growth of OSM data (OSM-stats, 2014).

Date	No. of Nodes	No. of Ways	No. of Relations
08/08/2005	8,000,000	8,000,000	8,000,000
23/01/2006	10,000,000	10,000,000	10,000,000
10/07/2006	20,000,000	15,000,000	15,000,000
25/12/2006	25,000,000	18,000,000	18,000,000
11/06/2007	30,000,000	20,000,000	20,000,000
26/11/2007	200,000,000	22,000,000	21,000,000
12/05/2008	230,000,000	24,000,000	22,000,000
27/10/2008	290,000,000	26,000,000	23,000,000
13/04/2009	320,000,000	27,000,000	24,000,000
28/09/2009	450,000,000	28,000,000	25,000,000
15/03/2010	580,000,000	31,000,000	26,000,000
30/08/2010	760,000,000	65,000,000	27,000,000
14/02/2011	1,000,000,000	80,000,000	28,000,000
01/08/2011	1,200,000,000	100,000,000	29,000,000
16/01/2012	1,300,000,000	110,000,000	30,000,000
02/07/2012	1,500,000,000	130,000,000	31,000,000
03/06/2013	1,100,000,000	200,000,000	32,000,000
05/05/2014	2,200,000,000	220,000,000	32,000,000



a-OpenStreetMap data (http://www.openstreetmap.org/). b- Google maps data (http://maps.google.co.uk/).

Figure 1. A comparison of the details of maps for the centre of Newcastle upon Tyne – UK (images sampled on 01/05/2014, both rendered at equivalent zoom levels – 16/19 for OSM, 15/18 for Google maps). This comparison facility is now available at http://tools.geofabrik.de/mc/.

a-London / UK b- Baghdad / Iraq

Figure 2. A comparison of the details of OSM data for the capital of UK (London) and the capital of Iraq (Baghdad). Images sampled on 1/05/2014, both rendered at equivalent zoom levels – 11/19 (http://www.openstreetmap.org/).

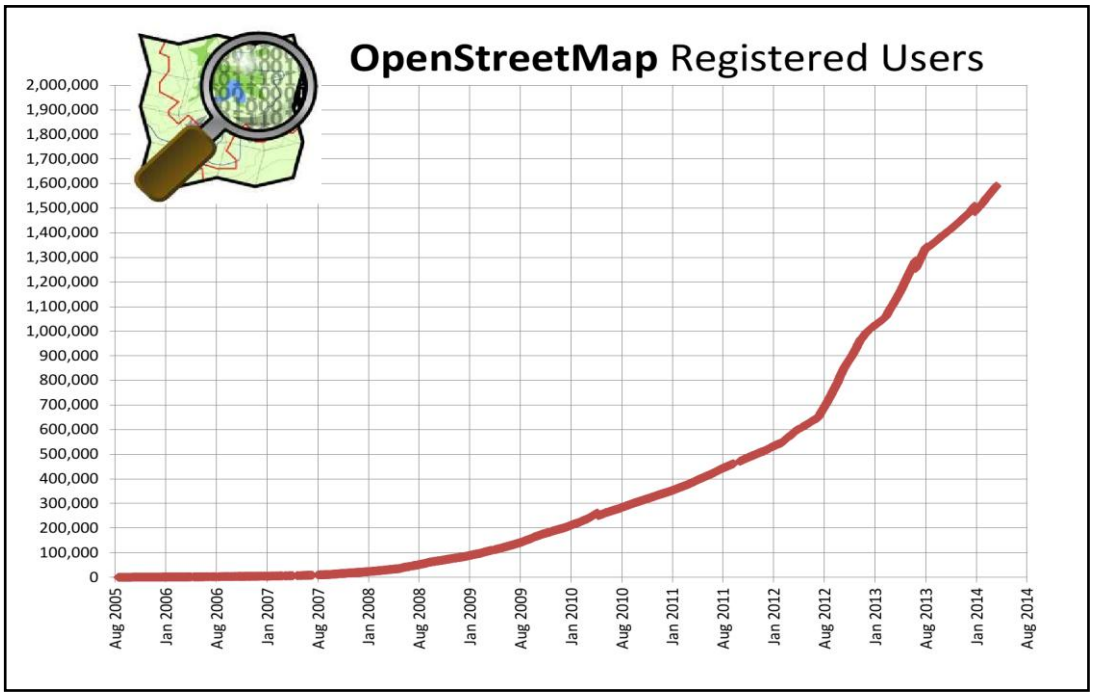


Figure 3. Statistics account graph reflecting the growth of OSM registered users (OSM-stats, 2014).

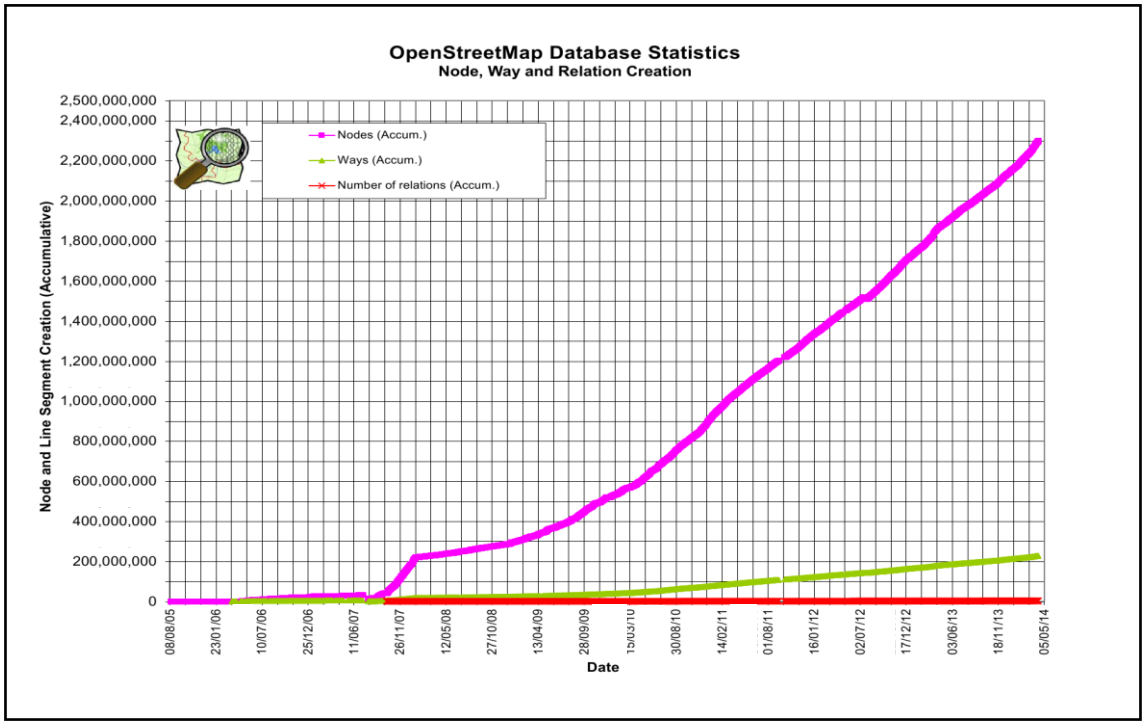
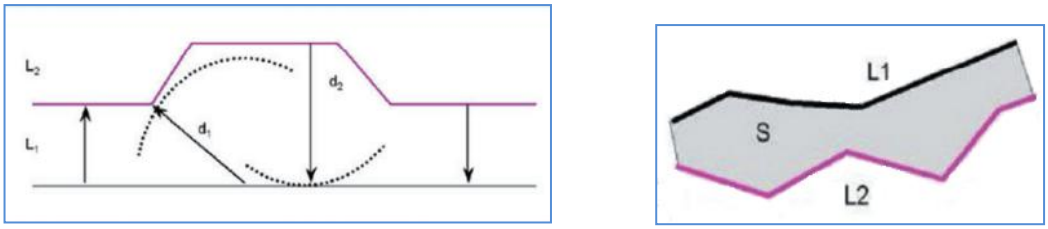


Figure 4. Statistics account graph reflecting the growth of OSM data (nodes, ways, and relations) on the web (OSM-stats, 2014).



a-Hausdorff distance method b- average distance method

Figure 5. The methods that have been adopted by Girres and Touya (2010) to determine the linear differences between road features.