ISPRS Int. J. Geo-Inf. **2013**, *2*, 507-530; doi:10.3390/ijgi2020507

OPEN ACCESS

ISPRS International Journal of **Geo-Information**

ISSN 2220-9964

www.mdpi.com/journal/ijgi/

Article

Assessing Completeness and Spatial Error of Features in Volunteered Geographic Information

Steven P. Jackson [1],*, William Mullen [1], Peggy Agouris [1], Andrew Crooks [2], Arie Croitoru [1] and Anthony Stefanidis [1]

[1] Department of Geography and GeoInformation Science, George Mason University, 4400 University Drive, MS 6C3, Fairfax, VA 22030, USA; E-Mails: wmullen@gmu.edu (W.M.); pagouris@gmu.edu (P.A.); acroitor@gmu.edu (A.C.); astefani@gmu.edu (A.S.)

[2] Department of Computational Social Science, Krasnow Institute for Advanced Study, George Mason University, 4400 University Drive, MS 6B2, Fairfax, VA 22030, USA; E-Mail: acrooks2@gmu.edu

* Author to whom correspondence should be addressed; E-Mail: sjacks11@gmu.edu; Tel.: +1-703-362-0205.

Received: 28 March 2013; in revised form: 14 May 2013 / Accepted: 20 May 2013 / Published: 4 June 2013

Abstract: The assessment of the quality and accuracy of Volunteered Geographic Information (VGI) contributions, and by extension the ultimate utility of VGI data has fostered much debate within the geographic community. The limited research to date has been focused on VGI data of linear features and has shown that the error in the data is heterogeneously distributed. Some have argued that data produced by numerous contributors will produce a more accurate product than an individual and some research on crowd-sourced initiatives has shown that to be true, although research on VGI is more infrequent. This paper proposes a method for quantifying the completeness and accuracy of a select subset of infrastructure-associated point datasets of volunteered geographic data within a major metropolitan area using a national geospatial dataset as the reference benchmark with two datasets from volunteers used as test datasets. The results of this study illustrate the benefits of including quality control in the collection process for volunteered data.

Keywords: volunteered geographic information (VGI); OpenStreetMap; quality; error; point

1. Introduction

Improvements in communications technology and information availability are having a significant impact on the field of geography as they enable the general public to produce geospatial products for mass consumption on the Internet [1,2]. As technology continues to improve (e.g., enhancing the computing and geolocation capabilities of hand-held devices) and the Internet is accessible by more citizens, the amount of geospatial data generated by citizens without formal geographic training is expected to rapidly increase [3]. Thus, Volunteered Geographic Information (VGI) [4] is bringing the general public into the realm of map production functions traditionally reserved for official agencies. With all humans becoming potential contributors of geospatial information [4], this trend is affecting greatly the geospatial community

VGI is following the development of Web 2.0 where users are contributing in more places and more often [5]. Another significant descriptive phrase is "crowdsourcing" which describes VGI in business terms, linking resources and work assignments as suggested by Howe [6]. Crowdsourcing has many definitions in relation to the development of geospatial data. Brabham describes the approach as using on-line volunteers to solve a formerly internal production requirement of a business or agency [7]. Heipke suggested the term crowdsourcing be used to describe data acquisition by large and diverse groups of people using web technologies [8]. Stefanidis *et al.* differentiated crowdsourcing from crowd-harvesting, and VGI from Ambient Geographic Information (AGI) respectively [9]. In the former, the crowd is presented with an explicit task, and their contributions are part of this assignment, whereas in the latter broader-scope information contributed by the crowd (e.g., through social media contributions) is mined to harvest geospatially-relevant content. Harvey argued that crowdsourcing includes both data that is "volunteered" and data that is "contributed" suggesting that contributed data represents information that has been collected without the immediate knowledge and explicit decision of a person using mobile technology that records location while "volunteered" data is representative of information explicitly provided [10].

The focus of this paper is on crowdsourced VGI (rather than the AGI), and more specifically on the accuracy of such information. It has been noted that public participation in geospatial mapping on the web has allowed citizen groups to map and provide local knowledge context that significantly advances the mapping project; however, others have noted that the characteristics of the information are less rigorous than traditional scientific data collection reporting, which could impact both feature content and attribution [11,12]. Volunteered data is usually provided with little to no information on mapping standards, quality control procedures, and metadata in general [13]. Understanding and measuring the data quality of information provided by volunteers who may have unreported agendas and/or biases is a significant problem in geography today [14]. A step towards understanding the potential data quality of volunteered data would be to quantify key quality characteristics for geospatial data that can reasonably be expected to be included in contributed datasets, and then be used to compare those characteristics against reference sources of data to quantify data quality.

The assessment of the accuracy of volunteered geographical information has been the subject of earlier work, but that work has focused almost exclusively on the accuracy and completeness of linear features such as roads and walkways with a focus on analyzing nodes or points that comprise such features [15–17]. VGI errors have been shown not to be random, and occur, as Feick and Roche point out, across spatial and thematic domains yielding data that must be assessed prior to operational use [15–18].

Regarding point feature accuracy, efforts to date have focused primarily on road intersections [15], and have not evaluated the accuracy of other features that are commonly represented as points in VGI datasets. A representative example of such features is Points of Interest (POI) in OpenStreetMap (OSM) [15,19]. In order to bridge this gap and expand the current state-of-knowledge, in this study we will address point features representing schools and will assess relevant accuracy issues, namely their completeness and accuracy (spatial error). We consider for our study VGI data in the United States (US), where government-provided sources are readily available; however, the concepts presented in this study have applicability to any geographic area or dataset.

The paper is organized as follows. In Section 2, we discuss relevant VGI accuracy considerations. In Section 3, we present a methodology and case study to assess the accuracy of crowdsourced point data. The case study comprises three datasets and the results of this study, presented in Section 4, and aims to provide researchers with a better understanding of the completeness and accuracy of the different volunteered datasets. We conclude with our summary and outlook in Section 5.

2. Accuracy and Completeness Considerations for Volunteered Geographic Information

Goodchild stated that "all spatial data without exception are of limited spatial accuracy" and yet the accuracy of geospatial data remains a significant concern today, nearly 20 years after Goodchild made this comment [20]. The concern is perhaps greater now than in the past with the widespread use of new technologies for mapping across the Internet [21], presenting both challenges and opportunities to the dissemination of geographic information [22]. In addition, researchers have considered how to quantify the value of the contributions from a societal perspective as well as evaluating the quality and usability of the contributed geospatial data itself [23–26]. Feick and Roche point out that VGI data generation inherently lacks professional oversight, does not follow established quality standards, and is affected by the inherent heterogeneity of VGI across thematic, media, and spatial dimensions [18]. These issues impact the functional utility of using VGI as an alternative or complement to "authoritative" datasets, which may be available from commercial or government sources [12,27,28].

While data quality has been at the center of the research agenda since the definition of GIScience, Goodchild and Hunter presented a discussion of the method for comparing two datasets whereby the tested source of data is compared to the reference source of data [29,30]. The reference dataset is assumed to represent ground truth while the test dataset is measured against the reference dataset. Comparisons between datasets are common within the literature; however, the methods in this paper are tailored for comparing point features [30–33]. The methods outlined in Section 3 of this paper will complement the linear comparison approach of Haklay [15], extending it for application on point features. We will then illustrate how to compute completeness and accuracy of such data.

Considering the rather *ad hoc* nature of VGI approaches, completeness is as important as accuracy when it comes to assessing the quality of the contributed information. Haklay and others measured completeness of VGI data as a measure of the total length of road segments within VGI as compared to the reference data [15–17]. Brassel *et al.* noted that the concept of completeness describes the relationship between objects in a dataset and the universe of all objects (real world) and could be extended to include assessing the completeness of attribution and metadata [34]. Devillers and Jeansoulin expanded the definition of completeness to include assessment of errors of “omission” and “commission” for datasets that either under-represent or over-represent reality [32]. Therefore, knowledge of the existing (actual) number of features in a study area is a key factor when assessing the completeness of volunteered contributions of geospatial data for that area but evaluating the completeness of metadata or attribution is not included in this study.

Goodchild and Hunter examined a method of quantifying spatial accuracy using traditional statistical methods such as the Root Mean Square Error (RMSE) and the standard error to describe the spatial error of point features [30]. Al-Bakri and Fairbairn reviewed the spatial accuracy between VGI and government data and found that the RMSE is quite high for VGI [35]. These authors attributed these errors to the common methods used by VGI data collectors, which often employ low-precision instruments such as personal GPS units and commercial imagery services. Zielstra and Zipf examined the differences between VGI and commercial data sources in Germany noting that the quality of the VGI degraded considerably as the distance from the urban core increased [17].

Drummond defined positional accuracy as the “nearness” description of a real world entity in an appropriate coordinate system to that entity’s true position [36]. However, in dealing with point features representing areal features, there is an inherent ambiguity in identifying a single location to represent such a feature. Accordingly, the accuracy of such features is affected by how well a contributor can distill an areal feature to a single location and how consistently this operation can be performed across a range of different contributors. Therefore, the accuracy of such features is affected not only by mensuration (which is affected by scale, accuracy, and precision), but also by the contributor’s interpretation of the appropriate representative location of the feature [37,38]. This ambiguity or vagueness, as discussed by Worboys and Duckham, affects the traditional concept of accuracy measurements [39]. This paper suggests that representational vagueness is inherent in VGI datasets, and that assessing positional accuracy of the contributed data will provide an understanding of the reliability and variability of the reported results. Furthermore, an enhanced understanding of the positional accuracy will aid in assessing the overall quality of contributed data as a potential data source for use by mapping agencies and researchers.

Our research continues the trend in evaluating both completeness and accuracy, but extends the notion of completeness to the comparison of individual point features representing area features, and assesses accuracy generally. Because of the inherent vagueness of POI locations, we base our analysis on the notion of school building or school campus extent to tie the numeric results to real-world values.

3. Materials and Methods

The discussion presented here is broken down into several pieces. We start in Section 3.1 by introducing the study area and datasets that will be used in the rest of this paper. Next, in Section 3.2 we will present the rationale for the development of our method. Following that, in Section 3.3, we discuss data processing issues; this will be followed by a presentation of the automatic methods, in Section 3.4, and manual methods, in Section 3.5, which are part of our procedure. Finally, in Section 3.6, we present summary statistics, which will aid in the later analysis for completeness and accuracy.

3.1. Case Study

As has been illustrated in the previous section, geospatial data quality is an ongoing concern; however, ideas have been presented for comparing reference and test datasets with each other. Quantifying completeness and accuracy will allow users of the data to better understand the data's utility, but require that reference and test datasets be available. Fortunately, some recent work in the US has generated data that is appropriate for these analyses. This research uses three data sources including a government-provided reference source and two different VGI test sources.

The reference data is based upon information from the Department of Education's lists of public and private schools. On behalf of the Federal government, Oak Ridge National Laboratory (ORNL) was asked to geospatially improve the location accuracy of the Department of Education data using repeatable methods. The ORNL data was created by geocoding address information for the schools [40]. The resulting dataset is used extensively across the Federal government as the definitive national level database of the location and attribute information for both public and private schools in the US. In addition to this reference dataset, two test datasets are also used in this case study.

The first VGI test dataset comprises school locations from the POI layer of OSM. The POI layer represents each specific feature as a node and may include: churches, schools, town halls, distinctive buildings, Post Offices, shops, pubs, and tourist attractions as noted by the OSM wiki-site [41]. Over *et al.* indicate that the primary key in OSM for these nodes is "amenity", which is broken down into categories, including: accommodations, eating, education, enjoyment, health, money, post, public facilities and transportation, shops, and traffic [42]. From the above-referenced OSM wiki, instructions are provided to contributors to identify schools as areas when possible; however, when the boundaries of the area are unknown, the contributor is instructed to place a node in the middle of the area to represent the school compound. Because no limitations are placed on OSM contributors regarding the preferred placement of a representative location, significant variation in actual point location should be expected, whether the contribution is created from a personal GPS, smartphone, or online using heads-up digitizing with imagery of unspecified accuracy. We will use the OSM dataset to provide a direct assessment of point feature accuracy in VGI contributions.

We also use a second test dataset, which is a product of the US Geological Survey (USGS) OpenStreetMap Collaborative Project (OSMCP)—2nd Phase [43]. OSMCP represents a hybrid variant of VGI in that it introduces limited oversight to the VGI process: the data are collected through VGI processes, peer-edited by volunteers, and a government agency (USGS in this case) provides quality control feedback to the volunteers, in an effort to improve the overall accuracy of their products. The USGS provided guidelines to the volunteers and instructed them to place features at the center of the building they represent [44]; however, the OSMCP data collection method did not involve visiting any sites, but rather relied upon online research coupled with heads-up digitization using imagery that was provided to the users. The motivation behind the OSMCP effort is the desire of USGS to use such VGI data as a complement to their official datasets and incorporate the OSMCP results into *The National Map* [43].

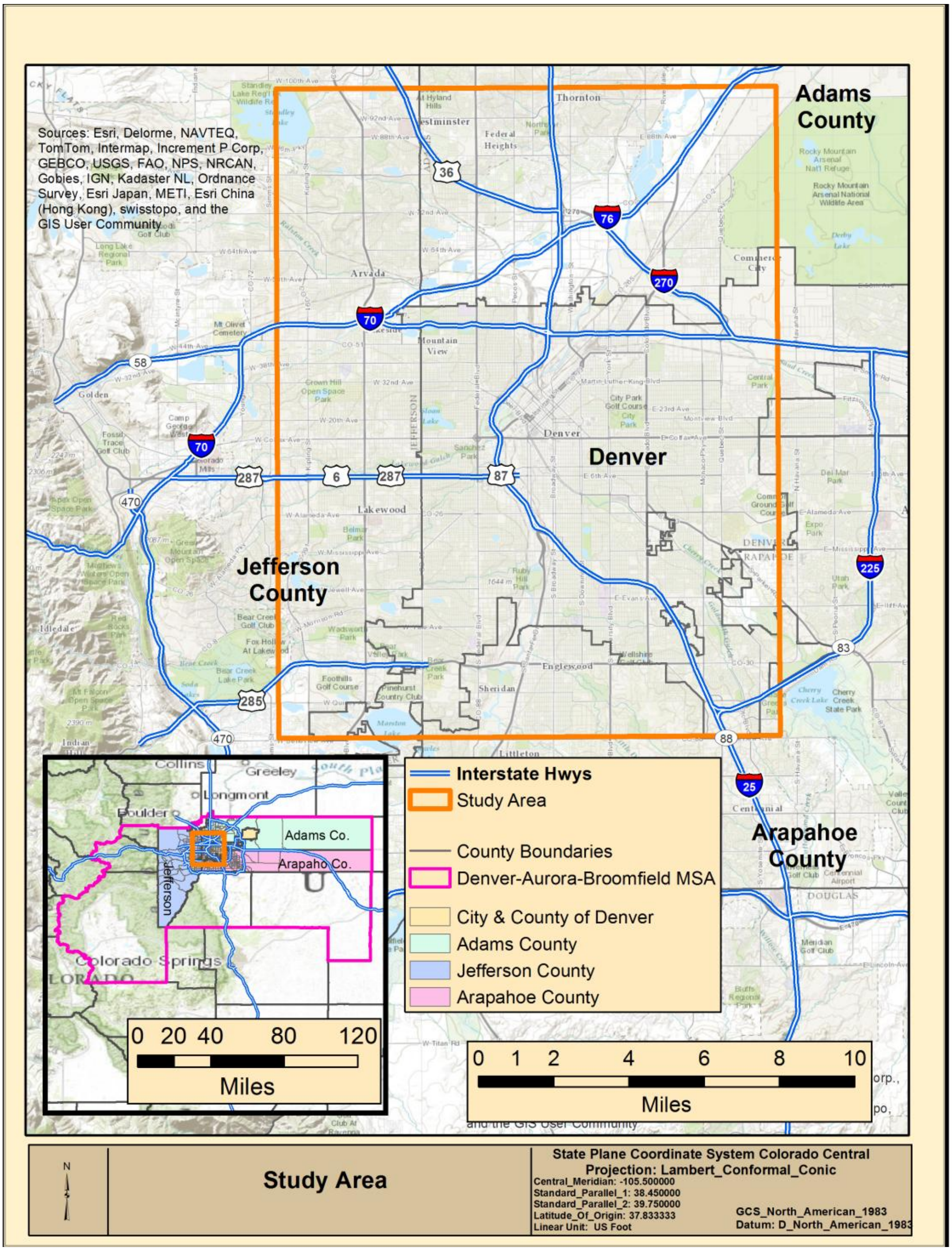
It is important to recognize that one possible source of inconsistency between the datasets is the use of different feature classification schemes between each of the data sources. The ORNL and OSMCP data represents elementary and secondary education in the US [44–46]. The OSM data includes the same definition for schools; however, the word "kindergarten" is used in OSM to represent day care facilities while in OSMCP and ORNL, it represents the first year of elementary school and as a result, some schools may be inconsistently tagged within OSM [47]. A cursory examination of the OSM data for the study area revealed no instances where "kindergarten" was used when elementary school was intended. Within the research for this paper, we noted this discrepancy; however, based upon the substantial overlap between these definitions, we do not feel that these differences strongly affect the results.

OSMCP focused on the collection of data (point features) for selected structures that are similar to the POI from OSM. In the case of OSMCP though, the data collection was performed by a select group of 85 non-expert college student volunteers, and the process also underwent an iterative but limited quality control process by volunteers and USGS [43,48]. OSMCP can therefore be viewed as a hybrid VGI effort for two reasons. First, as we mentioned above, it introduces the notion of partial oversight to VGI. Second, with a relatively small group of college student volunteers performing the data collection, it resembles focused crowdsourcing efforts like Ushahidi, rather than the participation of "large and diverse groups of people" that Heipke considered as a representative participation pattern [8,49].

The study area for the comparison presented here is dictated by the footprint of the OSMCP data since both the ORNL and OSM data include the entire US. The OSMCP study area is located completely within the Denver-Aurora-Broomfield Colorado Metropolitan Statistical Area (MSA) as defined by the Census Bureau [50]. The study area covers a large percentage of the City and County of Denver, including downtown Denver, extends into portions of the surrounding Arapahoe, Jefferson, and Adams Counties as shown in Figure 1, and encompasses approximately 228.5 square miles. The study area consists predominately of commercial, industrial and residential neighborhoods commonly found in heavily urban areas, and includes a population of just under 1,100,000 people, or approximately 43% of total MSA population of 2.54 million.

While the ORNL data was generated using pre-prescribed methods and standards [40], it is unlikely to be perfect; however, Goodchild affirms the use of an imperfect reference source data by pointing out that many users are willing to ascribe authority to data sources which are common regardless of the methods under which they may have been developed [4]. While unlikely to be perfect, the ORNL data is likely to be more consistent than the OSM or OSMCP datasets because of the collection methods used; however, the methods for updating the ORNL data are slow and some schools, particularly newly opened or recently closed schools, are unlikely to be reflected in the ORNL data. Others have demonstrated how VGI has been shown to be more appropriate, because of their currency, than sources like ORNL during disasters [51]. Matyas *et al.* point out that data, like OSM, collected through VGI methodologies are often done so in a game-like atmosphere because the users contribute when and how they like with little or no oversight and as a result does not promote the idea of self-editing [52]. In contrast, OSMCP addressed the issue noted above by including editors from both student volunteers and USGS in a hybrid environment [53] and provided the volunteer contributors and editors with guidelines regarding the appropriate representative location for each feature. One of the stated goals of OSMCP was to gain a better understanding of the quality and quantity of data produced by volunteers by implementing the two-step quality phase after data collection [48]. The methodology used to develop the OSMCP data included an editing phase, which included a provision allowing users to add records, which were not identified in the reference dataset. As has been previously discussed, the methodology outlined in this paper uses the ORNL data as a reference dataset because it is the best available from the federal government. However, the authors acknowledge that discrepancies between the datasets will exist due to the interpretation of the contributor as discussed above. Records identified in test datasets but not present in the reference dataset are useful for completeness measures because they highlight the possible shortcomings of the reference dataset.

Figure 1. Study Area.


Sources: Esri, Delorme, NAVTEQ, TomTom, Intermap, Increment P Corp, GEBCO, USGS, FAO, NPS, NRCAN, Gobies, IGN, Kadaster NL, Ordnance Survey, Esri Japan, METI, Esri China (Hong Kong), swisstopo, and the GIS User Community
Adams County
Denver
Jefferson County
Arapahoe County
Arvada
Lakewood
Golden
Thornton
Englewood
Littleton
70
76
270
225
25
36
287
6
87
58
470
285
83
88
Boulder
Longmont
Greeley
Colorado Springs
Adams Co.
Arapaho Co.
Jefferson
0 20 40 80 120
Miles
Interstate Hwys
Study Area
County Boundaries
Denver-Aurora-Broomfield MSA
City & County of Denver
Adams County
Jefferson County
Arapahoe County
0 1 2 4 6 8 10
Miles
N
Study Area
State Plane Coordinate System Colorado Central
Projection: Lambert_Conformal_Conic
Central_Meridian: -105.500000
Standard_Parallel_1: 38.450000
Standard_Parallel_2: 39.750000
Latitude_Of_Origin: 37.833333
Linear Unit: US Foot
GCS_North_American_1983
Datum: D_North_American_1983

3.2. Rationale

Quantifying the quality aspects of the schools within OSMCP, with its relatively limited spatial extent and restricted contributor set as compared with the equivalent OSM data, offers a unique opportunity to assess the question: Do the quality controls instituted by the USGS measurably improve the completeness and accuracy of volunteered contributions? Additionally, are the spatial biases noted in OSM data consistent between linear features and the point features focused on in this study [15]? If so, are the biases present in both the OSM and OSMCP data? The findings from this study could then be applied by other researchers during implementation of other VGI projects so that they could work to mitigate any bias that may exist in data which was collected using VGI data collection methods.

Table 1 presents the total count of schools within the study area for each data source. While the total numbers of schools are reasonably close across the three sources, it is important to note ORNL and OSMCP data represent only active schools while OSM data includes approximately 12% historic schools which are likely no longer in existence and would, therefore, not match schools in either ORNL or OSMCP. A limited review of raw OSM data indicated that the majority of the records were derived from the USGS Geographic Names Information System (GNIS) which includes historic points of interest. GNIS data was likely bulk uploaded into OSM and users have not removed the historic records.

Table 1. School count by data source.

Source	School Count
Oak Ridge National Laboratory (ORNL)	402
OpenStreetMap (OSM)	406 *
OpenStreetMap Collaborative Project (OSMCP)	412

* Includes 48 historical school locations.

Following the work of Haklay, the results shown in Table 1 imply that these datasets are similar [15]. However, a deeper assessment of the schools showed that only 281 schools are common to all three datasets illustrating that simple feature count may not adequately evaluate spatial accuracy or completeness.

Figure 2 presents a portion of the study area that included 33 schools from the three datasets: 11 OSM; 12 OSMCP; and 10 ORNL, respectively. Review of Figure 2 indicates that there is very good spatial correlation across the three data sources for seven locations where each data source indicates the presence of a school. However, the review also allows identification of two ORNL reference schools that do not have either an OSMCP or OSM school located nearby (black circle). In addition, there are three areas where both OSMCP and OSM schools are indicated, but there is no ORNL school associated (black diamonds), and there is one location where ORNL and OSMCP data correlate without an associated OSM school (black triangle). Lastly, there are two locations (black squares) where only one data source indicates the presence of a school, one from OSM and one from the OSMCP data source. The above visual assessment clearly indicates that while there are similar numbers of schools within the study area, as shown in Table 1, the spatial variability suggests that a detailed evaluation of the data sources is required to understand the similarities and differences between the datasets. The above assessment only looked at the spatial "association" of features, and did not address the specific attribution associated with those locations.

Figure 2. Comparison of OSM, OSMCP, and ORNL data.

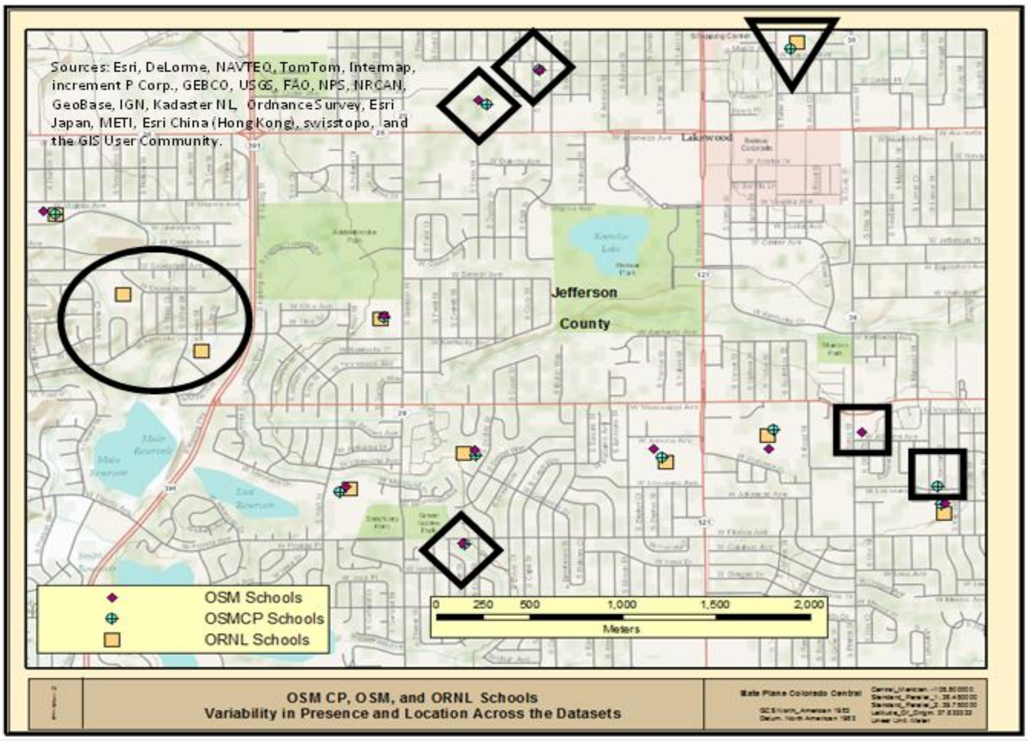
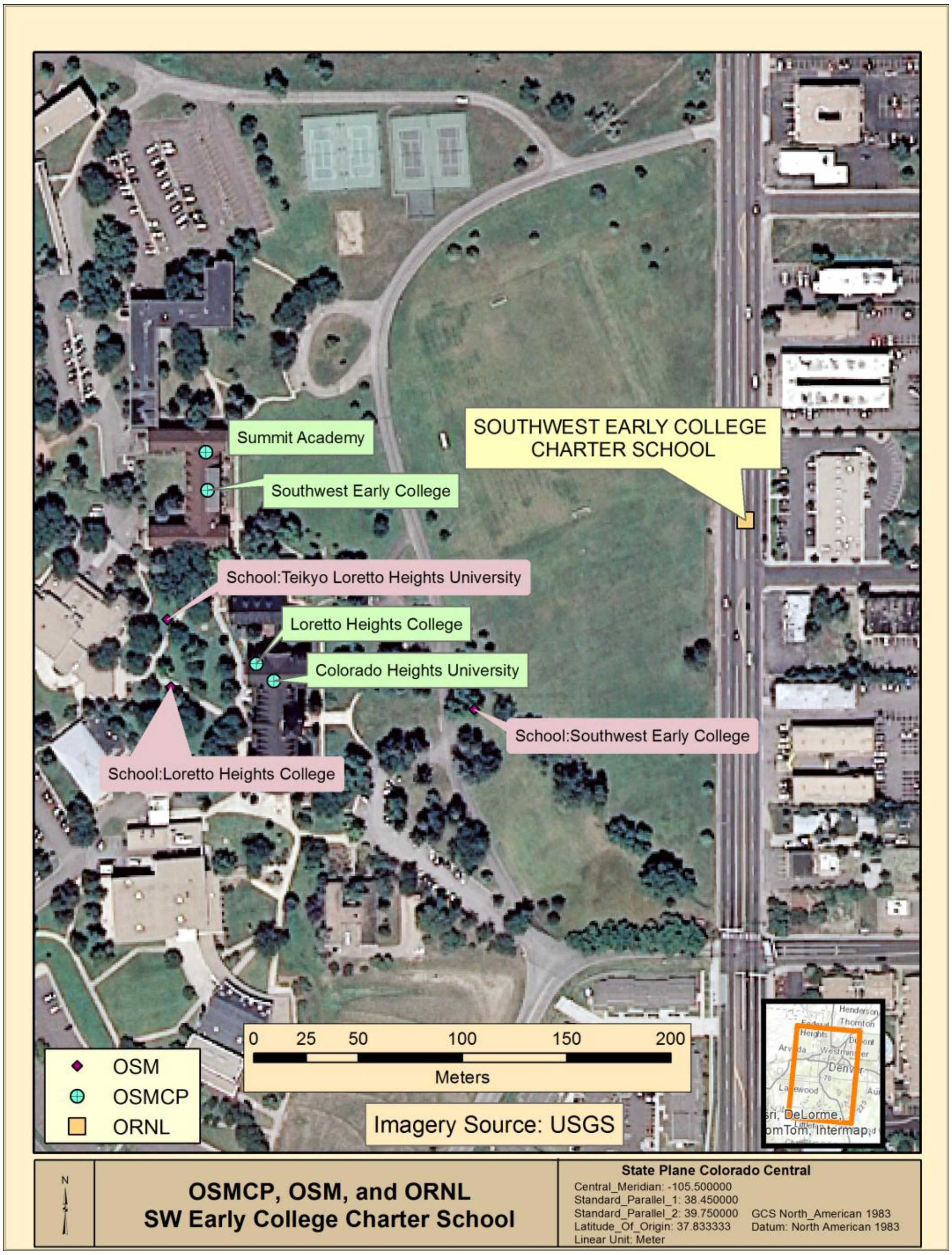

Sources: Esri, DeLorme, NAVTEQ, TomTom, Intermap, increment P Corp., GEBCO, USGS, FAO, NPS, NRCAN, GeoBase, IGN, Kadaster NL, Ordnance Survey, Esri Japan, METI, Esri China (Hong Kong), swisstopo, and the GIS User Community.
Lakewood
Jefferson
County
◆ OSM Schools
⊕ OSMCP Schools
□ ORNL Schools
0
250
500
1,000
1,500
2,000
Meters
OSM CP, OSM, and ORNL Schools
Variability in Presence and Location Across the Datasets
State Plane Colorado Central

Figure 3 depicts the ORNL, OSMCP and OSM feature locations for the vicinity of Southwest Early College within the study area and presents an example where the effects of spatial variability discussed above, as well as variation in attribution, impact dataset comparison. There is only one school identified in the ORNL reference dataset in the vicinity. The ORNL location is denoted by a tan square and is located several hundred meters east of the OSMCP and OSM locations. OSMCP data indicates four separate schools (Summit Academy; Southwest Early College; Loretto Heights College; and Colorado Heights College from north to south) on what appear to be the same school grounds while the OSM data indicate three schools (Teikyo Loretto Heights University; Loretto Heights College; and Southwest Early College, from northwest to east). The presence of multiple schools noted in the OSMCP and OSM data suggest that the reference ORNL data may have errors of omission.

Figure 3 illustrates the concept of spatial vagueness within datasets discussed above and demonstrates that the traditional concept of geographic accuracy is likely an unnecessary goal when examining crowdsourced point feature data. Using any of the sources for Southwest Early College would allow navigation to the school property, although the OSMCP point feature location visually appears more accurate in that it falls on an actual building while the OSM location is on the school grounds and the ORNL location is mapped to a road centerline, furthest from the school buildings and off the campus entirely.

Figure 3. Various identified locations of Southwest Early College.


Summit Academy
SOUTHWEST EARLY COLLEGE
CHARTER SCHOOL
Southwest Early College
School:Teikyo Loretto Heights University
Loretto Heights College
Colorado Heights University
School:Southwest Early College
School:Loretto Heights College
OSM
OSMCP
ORNL
0 25 50 100 150 200
Meters
Imagery Source: USGS
N
OSMCP, OSM, and ORNL
SW Early College Charter School
State Plane Colorado Central
Central_Meridian: -105.500000
Standard_Parallel_1: 38.450000
Standard_Parallel_2: 39.750000
Latitude_Of_Origin: 37.833333
Linear Unit: Meter
GCS North_American 1983
Datum: North American 1983
Denver
Lakewood
Westminster
Thornton

Based on the limited assessments presented above, and the discussion regarding attribution and source differences, it is clear that a systematic comparison of the datasets is required in order to evaluate the quality of the VGI test data in comparison to the reference data. The following section outlines that systematic approach.

3.3. Data Preparation

The ORNL data was provided in shapefile format projected in WGS84 but was re-projected into the Central Colorado State Plane system. The data was then clipped to the boundary of the OSMCP study area, yielding 402 school locations within the study area.

The OSM data was downloaded from the Internet [54]. The data was downloaded on 13 December 2011 and represented OSM data as of 16 November 2011. The OSM points of interest data, provided in shapefile format projected in WGS84, was extracted and then clipped to the boundary of the OSMCP study area, yielding 4,285 points. The data attribution consists of four attributes: FID, Shape, Category, and Name. Significantly, to this study, the "Name" attribute is a concatenation of function and name. For example, "Place of Worship:Applewood Baptist Church", "Restaurant:Bonefish Grill", "CaféEinstein Brothers", "Pub:Baker Street Pub and Grill" and "School:Alpine Valley School". The 4,285 OSM points of interest were queried to extract records including the phrase "School": yielding 406 schools in the study area. One concern with this method of extraction is that it relies upon the OSM data contributor to properly tag the schools; however, for this research we did not attempt to identify any improperly tagged entities to search for missing schools. The data was then projected into the Central Colorado State Plane system. Unfortunately, address was not included in the OSM extracted data and as a result, only the names can be used for comparing the data. As was discussed above, the GNIS data was used to bulk upload schools data into OSM and GNIS data does not include address information.

The OSMCP data provided by the USGS is bounded by the study area; however, it includes data beyond education, so schools were extracted from the overall dataset using an attribute called FType (Feature Type) yielding 412 schools [55]. The OSMCP data was projected into the Central Colorado State Plane system. After that, the procedures outlined in Sections 3.4 and 3.5 were repeated for each comparison of a test dataset to a reference dataset.

3.4. Automated Methods

Automated matching of the datasets was carried out using four different methods across each of the name and address fields from the dataset attributes. The automated matching process begins with verifying that the spatial reference of each dataset is the same and in units that are useful for measuring the spatial error between two features. For example, WGS84 data is unlikely to yield useful matching results because the distance measures will be in degrees whereas a projection based on feet or meters would yield more useful distance measurements. Next, a spatial join of the test dataset to the reference dataset is conducted (using the Spatial Join tool within the Analysis Toolbox of ArcGIS™) as the most-likely match for each record is the physically closest record and the spatial join identifies the closest record.

Before further analysis can begin, an attribute, called "MatchMethod" is added to the joined dataset's attribute table. This attribute is populated by a script that was developed by the authors as a part of this research utilizing the Python™ scripting language within the ArcGIS™ environment. The valid values for MatchMethod are from 0 to 11 as shown in Table 2. The script moves through the records looking for matches between the reference and test dataset and records the MatchMethod used. Initially, all records have a MatchMethod value of 0, which indicates that the records have not yet been evaluated. MatchMethod values 1 through 5 are "name" matches with the first four being automated and denoted by the 'AN' prefix. MatchMethod values 6 through 10 mirror MatchMethod values 1 through 5 in function except that the address field is used to find a match instead of the name field. Automated methods based on the address are denoted by the "AA" prefix. As was previously discussed, the OSM data did not include address information so the "AA" methods did not yield results in that specific comparison, but they are included here because the algorithm does leverage address information when it is available, such as for the OSMCP data.

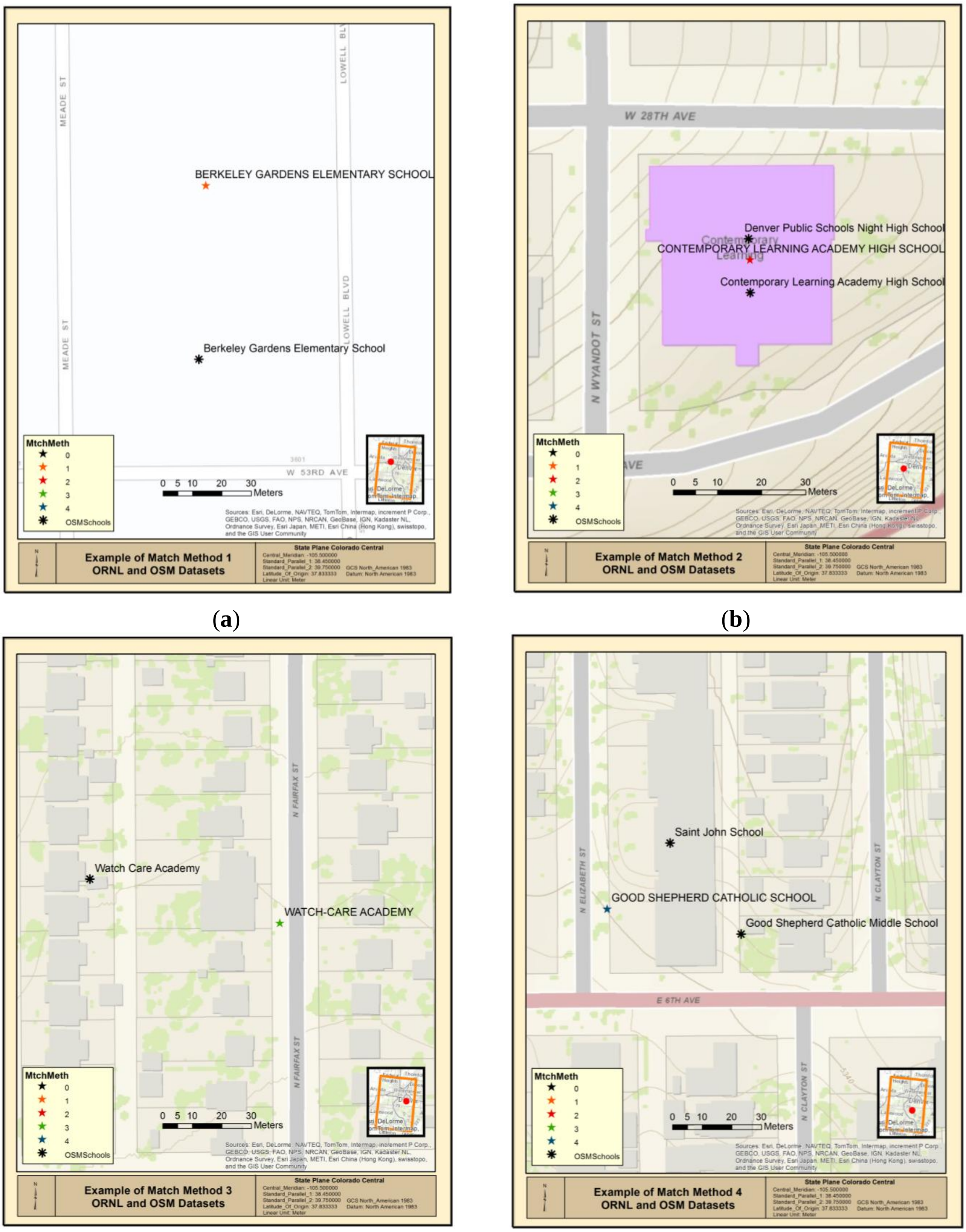
Table 2. Values used to track record matching.

Match Method Value	Description
0	Record not yet analyzed
1 (AN-C)	Fully automated name match derived from closest record
2 (AN-O)	Fully automated name match using any other record
3 (AN-DC)	Fully automated name match using difflib string comparison to closest record
4 (AN-DO)	Fully automated name match using difflib string comparison to any other record
5 (MN)	Manual match using name
6 (AA-C)	Fully automated address match derived from closest record
7 (AA-O)	Fully automated address match using any other record
8 (AA-DC)	Fully automated address match using difflib string comparison to closest record
9 (AA-DO)	Fully automated address match using difflib string comparison to any other record
10 (MA)	Manual match using map/address
11	No match identified

Initially, the algorithm attempts to identify a perfect match between the test and reference dataset. The test record nearest to each reference record was identified using a spatial join as described above. Perfect matches of the nearest records are denoted by the suffix "C" in Table 2. If the nearest record is not found to be an exact match for the reference dataset, then the algorithm examines the test records to identify another exact match within the test dataset. If one is found, then the match is denoted by the suffix "O" in Table 2. Examples of MatchMethod values AN-C and AN-O are shown in Figure 4(a,b).

The MatchMethod values ending with "DC" and "DO" use the Python™ difflib library to identify similarities in the attribute values. The difflib method is based on a pattern matching algorithm developed in the late 1980s by Ratcliff and Obershelp [56]. The two methods listed in Table 2 that used the "DC" suffix examine the nearest test record to each reference record using the SequenceMatcher class and the ratio method to find a match. The ratio method returns a value between 0 (no match) and 1 (perfect match) when comparing two values. The pattern matching methods implemented in Python™ are used in this analysis because they are fast and effective for our purposes and they provide a result, the ratio, which can be quickly interpreted. In essence, the method counts the number of matching characters between the two strings and divides that number by the total number of characters in the two strings and returns that result as a value (ratio) which can then be compared to a minimum value to determine whether an appropriate match was identified. Through trial and error, appropriate minimum ratios for matching the name and address attributes were identified. These ratio values were selected with a focus on minimizing false positives. Similarly, the methods from Table 2 that have the suffix "DO" leverage the get_close_matches method to find the closest match between the reference record and all test records. The get_close_matches method returns an ordered list of close matches. The script developed for this paper selects the highest match value and compares that value to the minimum acceptable ratio to determine whether the record is a match or not. Examples of MatchMethod values AN-DC and AN-DO are shown in Figure 4(c,d).

Figure 4. (**a**) MatchMethod value AN-C between ORNL and OSM Data; (**b**) MatchMethod value AN-O between ORNL and OSM Data; (**c**) MatchMethod value AN-DC between ORNL and OSM Data; (**d**) MatchMethod value AN-DO between ORNL and OSM Data.


MEADE ST
LOWELL BLVD
BERKELEY GARDENS ELEMENTARY SCHOOL
Berkeley Gardens Elementary School
W 53RD AVE
MtchMeth
0
1
2
3
4
OSMSchools
0 5 10 20 30 Meters
Example of Match Method 1
ORNL and OSM Datasets
State Plane Colorado Central
W 28TH AVE
N WYANDOT ST
Denver Public Schools Night High School
CONTEMPORARY LEARNING ACADEMY HIGH SCHOOL
Contemporary Learning Academy High School
Example of Match Method 2
ORNL and OSM Datasets
N FAIRFAX ST
Watch Care Academy
WATCH-CARE ACADEMY
Example of Match Method 3
ORNL and OSM Datasets
N ELIZABETH ST
N CLAYTON ST
Saint John School
GOOD SHEPHERD CATHOLIC SCHOOL
Good Shepherd Catholic Middle School
E 6TH AVE
Example of Match Method 4
ORNL and OSM Datasets

(**a**) (**b**)

(**c**) (**d**)

After completing the automated methods, one school, Adams City High School, was observed to have over a 3,000 m difference between the reference and test datasets. Review of the school information from online sources indicated that the original school buildings (represented in the ORNL data) had been closed after the development of the ORNL data and a new school, bearing the same name was built in a different location (represented in both the OSMCP and OSM data) [57]. Therefore, this school was omitted from further analysis of accuracy.

3.5. Manual Methods

MatchMethod values 5 and 10 are manual matching methods. Our automated methods are incapable of dealing with these cases; therefore, the user must manually examine the unmatched records which remain after the automated processes to determine if any potential matches were missed. In concept, the user will display both the reference and test datasets along with their labels within a Geographic Information System (GIS). The user then examines all records that have yet to be matched (*i.e.*, they retain a MatchMethod value of 0), in order to determine if a match might exist. If a record with a similar name is identified as a match, then the user would update the MatchMethod value to 5, measure the distance between the two features, and then update the distance, name, and address fields to match those of the matched record. In rare cases, the automated methods fail to match two records because their names and/or addresses are too different for the pattern matching algorithm to recognize; however, during the manual process, the user can identify these two records as a match. One example of manual match is “Escuela Tlatelolco Centro de Estudios” and “Escuela Tlatelolco” where the automated matching methods failed to identify these two records due to the difference in the length of the corresponding records. In order for the pattern-matching algorithm to identify these records as a match, the minimum ratio would have to be set so low that many false matches would be generated for other features and as was previously stated, the minimum values were selected to minimize false matches. The ratio describing the similarity of these two strings is 0.65, which is well below the minimum ratio of 0.83 that was developed through trial and error. Fortunately, very few cases of missed records, like the one illustrated above, were discovered during the manual matching process which confirmed the decision to set higher ratios to avoid false matches.

A very similar process would be followed looking for an adjacent record with a similar address except that the user would update the MatchMethod value to 10 and then update the distance, name, and address. If no match can be found through manual means, then the user would update the MatchMethod value to 11 indicating that no match was found. The user would repeat the process until no record with a MatchMethod value of 0 remains, completing the matching process.

3.6. Computation of Summary Statistics

Once the matches had been identified, as outlined above, an additional processing step is required before analysis can take place. The final process in the analysis includes computation of the intersection, union, and complement. These values can then be used in the computation of the accuracy and completeness [58].

The Intersection (Reference ∩ Test) dataset includes all records that are in both of the datasets. The records are identified by selecting records that have a MatchMethod value of 1–10 in the joined dataset.

The Reference Complement Test (Reference/Test) dataset includes those records that are in the Reference dataset, but not in the Test dataset. To identify these records, the records with a MatchMethod value of 11 are selected from the joined dataset.

The Test Complement Reference (Test/Reference) dataset includes those records that are in the Test dataset, but not in the Reference dataset. The method for identifying these records is a bit less straightforward than the previous method. The reason for the added complexity is that the joined dataset does not include all of the records from the Test dataset but instead only those records that had matches. In order to determine which Test records are not in the joined dataset, the two datasets were compared using a table join on attributes using a unique identifier in the Test dataset and then all unjoined records were extracted since they were not identified in the joined dataset.

The Union (Reference ∪ Test) represents all records that are in the Reference or Test dataset. The previously described intersection and complement datasets were merged to derive the Union.

Computation of these four parameters will facilitate the completeness and accuracy calculations, which will be presented later in this paper.

4. Results

From previous discussion, Table 1 presents the total number of schools in the study area for each dataset and illustrates the similarities between the two datasets based purely on record counts. While Haklay used record counts of road segments as a surrogate for completeness, the previous discussion goes on to illustrate how pure count comparisons are insufficient to describe the differences between two point datasets [15]. The result is the method in this paper, which quantifies these differences in a more robust way.

Table 3 provides a breakdown of the types of each match that were identified using the automatic and manual matching techniques when comparing the ORNL to OSMCP datasets and ORNL to OSM datasets in an effort to determine whether or not the methods outlined in this paper meet the goal of being repeatable and reliable. For the comparison of ORNL to OSMCP, roughly 82% of the records were matched using the automated algorithm while fewer than 7% were matched using manual methods with approximately 11% unmatched. For the comparison of ORNL to OSM, the matching percentages were somewhat lower as a result of the fact that the OSM data used in the analysis did not contain address information. Roughly 56% of the records were matched using the automated algorithm while approximately 15% were matched using manual methods with over a quarter (28%) unmatched.

Table 3. Matching results for ORNL and OSMCP.

Match Method		ORNL-OSMCP		ORNL-OSM	
		Record Count	**Percent**	**Record Count**	**Percent**
1–4, 6–9	Automated Matches	329	82%	225	56%
5, 10	Manual Matches	28	7%	62	16%
11	No Match	44	11%	114	28%
	Total	401	100%	401	100%

These analyses of the match method counts indicate that the automated methods were successful in matching a majority of the records automatically even though the datasets were contributed using different, or non-existent, standards. The percentages of records that were identified using manual methods were relatively low even when the unmatched record counts were high. These successes in automated matching will ensure that the method used is repeatable and provide some validation for the method. The next step is to use these results to begin to evaluate the differences between the datasets.

Following the previous discussion, examining the results of the automatic and manual matching methods, Table 4 provides a summary of the counts that were obtained from each of the Intersection, Complement, and Union calculations. The Union and Complement counts for the OSM comparison are higher, as would be expected, because of the lower matching rates caused by the absence of the address information; however, as is the case for comparing record counts, simply comparing the counts between these comparisons is insufficient when trying to understand the meaning of the results.

Table 4. Summary of record counts for data matching.

Method	ORNL-OSMCP	ORNL-OSM
Intersection	357	287
Reference Complement Test	44	114
Test Complement Reference	63	99
Union	464	500
Reference Count	401	401
Test Count	412	406

4.1. Completeness

Assessing completeness of the contributed data provides an understanding of the reliability of the reported results and allows assessment of the usefulness of contributed data as a potential data source for use by mapping agencies and researchers. In assessing completeness, the present study considers the issues of omission and commission within the database. Brassel *et al.* focused on evaluating whether the entity objects within the database represent all the entity instances in the real world [34]. Completeness describes the difference between the real world and the database as a percentage of the total physical structures in the study area.

Using Brassel *et al.*'s approach to determine degree of completeness, this study compared the test (OSMCP and OSM) data to the reference (ORNL) data [34]. As shown in Table 3, the comparison of ORNL and OSMCP data showed that a total of 89% of the records were matched. Of that, 82% were matched automatically while the remaining 7% were matched manually indicating that the automated matching algorithm is successful.

Table 3 also summarizes the comparison of ORNL and OSM data. While the match rate for this comparison was considerably lower with only 71% of the total records matched, the manual match rate was over twice as high at 15% with the automated match rate falling to 56%. Due to the relatively poor performance for completeness, the utility of OSM data as an alternative mapping source is questionable in the study area; however, the OSMCP data, which included approximately 9 out of every 10 schools, represented a significant improvement over the unconstrained OSM results of just over 7 out of every 10 schools.

Analyses revealed that the OSM and OSMCP efforts captured schools that were not in the ORNL data. 28% of the OSM schools remained unmatched at the end of the analyses while 11% of the OSMCP records remained unmatched, as shown in Table 3. The analyses within this paper did not investigate these issues further to determine whether these unmatched records identify schools which are absent from the reference dataset or whether the test datasets captured records which are not schools according to the common definition described above in Section 3.1. Further analysis of the unmatched records would be required to study their nature to assess the quality of the reference dataset.

4.2. Accuracy

While the completeness measure is assessed by comparing the matched and unmatched records, the accuracy examines only those records which had matches as identified previously in the Intersection.

It is expected that the overall accuracy of the OSMCP data will be high considering the quality control procedures that were part of the project; however, it is important to quantify the accuracy of the data since this is a fundamental element of geospatial analysis. Accuracy comparisons across datasets from differing sources require careful understanding of the standards used for feature placement in each source. As was mentioned above, the ORNL data focused on address matching to street centerline. The OSMCP collection parameters instruct the contributors to locate the school features at the building centroids. OSM instructs contributors to create area features for school complexes, but if the complex boundary is poorly specified the contributor is instructed to estimate a location in the middle of the complex. It is important to recognize that associating a single point for the features studied in this work may result in different degrees of vagueness depending on the characteristics of a specific feature and the interpretation by the contributor. In the absence of a common method, some positional differences between any two datasets can occur. This issue of how contributors reduce areal features into point locations is a complex one, and as such it is beyond the scope of this paper [38].

The spatial error is evaluated for each match and the results are located in Table 5. While the minimum error was two meters in both comparisons, the maximum error for the OSM data was approximately four times that of the maximum error for the OSMCP. Both the mean and the standard deviation were higher for the OSM data with the latter indicating that the error within the OSM data varies more than in OSMCP. In addition, the median of the OSM data is lower than the mean indicating that the data is skewed. These results support the notion of error heterogeneity which has previously been described for OSM road networks [15–17].

Table 5. Spatial error for matched schools.

			Spatial Error (m)			
Datasets	**Count**	**Minimum**	**Maximum**	**Mean**	**St. Deviation**	**Median**
ORNL-OSMCP	357	2	487	47	50	33
ORNL-OSM	287	2	1,848	190	314	43

Because the error was so different between the two datasets, an additional effort was undertaken to examine the nature of the error distribution. As was discussed previously, VGI has no formal mechanism for enforcement of data collection standards and therefore some users may decide to place the school feature somewhere on the building while others may decide to place the school feature in the street adjacent to the building. As a result of these uncertainties of the feature placement, some error is expected. While on the low end of the spectrum, as seen in Table 5, a two meter error could be considered noise, at over 1,800 m, the high end would be considered a true error. Somewhere in between these two values, error makes the jump from noise to true error. In an effort to evaluate the place where this jump exists, two thresholds were identified. In recognition of the inherent vagueness of the definition of the location for the schools, we selected thresholds that are based upon the physical nature of the elements being represented as opposed to basing the thresholds upon the results themselves.

Thresholds are identified at 30 and 150 m respectively. The 30 meter (30 m) value was selected in response to school building size. If a school were a square with a 40,000 sq ft foot print, which would be typical for a 500 pupil middle school with two stories, then one half of one side would be approximately 30 m and as a result, any two points within 30 m of each other could be thought of as being in the middle of and on the edge of the building [59]. Similarly, school sites are ideally around 25 acres for middle schools with populations of 500 pupils [60]. One half of one side of a square 25-acre lot is a little over 150 m and therefore two points which are within 150 m of each other could be thought of as still being on the school property. Therefore 30 m is used to represent "building" accuracy and 150 m is used to represent "campus" accuracy. Using these two thresholds, Table 6 was generated to show the distribution of the error in the test datasets within these thresholds.

Table 6. Percent of schools with spatial error in each threshold.

Datasets	**Distance (m)**		
	<30	**30–150**	**>150**
ORNL-OSMCP	164 (45.8%)	178 (49.7%)	15 (4.2%)
ORNL-OSM	90 (31.4%)	186 (64.8%)	11 (3.8%)

Using the information in Table 6, the percent of matched schools within 150 m (cumulative) for both OSMCP and OSM can be shown to be 96%; however, at the 30 m level, OSMCP was able to match 46% of the records while OSM only matched 31%. These results indicate that either dataset would be equally capable of getting the user to the school property as shown by the similar percentage of records below the 150 m threshold; however, the OSMCP data has a greater potential to identify the school building as shown by its higher percentage with an error of less than 30 m. If, however, all of the ORNL data follows the example presented in Figure 3, where ORNL plots the address on the street while the OSM and OSMCP data plots a location on the property, then only the 150 m threshold would be appropriate and therefore the larger threshold is valid.

One additional assessment was undertaken in order to evaluate the accuracy of OSM *versus* OSMCP. In this final evaluation, the spatial error for matches from both datasets was compared to each other to determine which one is closer more often. Not all ORNL schools were found by both OSM and OSMCP contributors. Of the 402 ORNL schools, there were 281 schools that were matched by both OSMCP and OSM. A simple approach to determining relative accuracy is to subtract the OSM to ORNL distance from the OSMCP to ORNL distance for each of the schools that had a match. Using this method, a negative value would indicate that the OSM location is closer to the ORNL location than the OSMCP location. Table 7 presents an example of how the comparison of distance differences for matching schools between ORNL-OSM (A) and ORNL-OSMCP (B) was executed.

Table 7. Data sampling of distance differences for matched schools.

School Name	ORNL-OSM Distance (A)	ORNL-OSMCP Distance (B)	(A − B)
ALAMEDA HIGH SCHOOL	105	35	70
ALICE TERRY ELEMENTARY SCHOOL	22	7	15
ALL SOULS SCHOOL	16	42	–26
ALLENDALE ELEMENTARY SCHOOL	46	41	5
ALSUP ELEMENTARY SCHOOL	17	43	–26
ANNUNCIATION	2	6	–5
ARVADA HIGH SCHOOL	36	55	–19
ARVADA MIDDLE SCHOOL	115	98	17
…	.	.	.

The summary results of the sample analysis above for all 281 schools are located in Table 8. Of the 281 matched schools, OSMCP schools were closer for 58% of the schools; however, OSMCP also has the largest difference (224 m). Interestingly, OSMCP and OSM mean distances and standard deviations were almost identical, and the OSM data had a slightly higher median error for the matched schools.

Table 8. Accuracy comparison for matched schools.

	Count	Percent	Minimum	Maximum	Mean	St. Deviation	Median
OSMCP Closer	164	58.4%	2	224	23	32.3	26.2
OSM Closer	117	41.6%	2	159	45	32.1	28.8

5. Discussion and Conclusions

As VGI is gaining popularity, it leads to the generation of large volumes of geospatial data that can potentially complement and enhance traditional "authoritative" data sources. To enable tapping into this potential, we need a better understanding of the quality of VGI contributions, in particular their accuracy and completeness. This is even more important now, as VGI data collection is increasingly involving volunteers with little or no geographic training, who are producing geographical data. Consequently, there is a need to further study the quality characteristics of VGI.

This paper extends the current state of knowledge on this topic by focusing on completeness and accuracy of point features within VGI data. This complements prior studies, which assessed the accuracy of linear features in VGI, to improve our overall understanding of relevant quality issues. Our analysis demonstrated that simple count comparisons between two point datasets are insufficient for characterizing the differences between these two datasets, as they fail to recognize the presence of omission and commission errors. As a result, a more robust analysis is required, in order to identify and categorize discrepancies between two (or more) datasets. In particular, we introduced a semi-automatic approach to match corresponding records between two datasets, and demonstrated its effectiveness. Our analysis compared two VGI test datasets against a reference dataset and analyzed their differences. We also discussed the particularities of a hybrid variant of VGI (OSMCP) whereby a government agency is providing quality control feedback to the volunteers, and assessed its impact on the overall accuracy of the VGI product. Our observations indicate that the added rigor appears to improve both the completeness and accuracy as compared to the OSM data.

The analysis of completeness showed that the OSMCP data capture close to 90% of the records in the reference ORNL database, while the OSM data captured approximately 70% of these records. The lower completeness result observed within the OSM data can be attributed to two factors: the OSM data does not include address information, and the collection methods employed for the OSM data do not include the formal quality control processes implemented within the collection methods for OSMCP. Lastly, 70 more OSMCP schools (357) matched the reference dataset than did OSM schools (287).

Similar trends were identified with respect to positional accuracy, which reflects the spatial error between the locations of the two datasets, with OSMCP data appearing to be more accurate than OSM. Both OSM and OSMCP were within 150 m of the reference dataset 96% of the time; however, OSMCP was within 30 m more often (46% of the time) than OSM (31%). Overall, 59% of the time OSMCP schools were closer to their ORNL reference entries, compared to their OSM counterparts. In addition to providing an estimate for the accuracy and completeness, these results also suggest that OSMCP outperforms OSM.

This paper has addressed a topic, which so far has received cursory study in our community. Through this work, we have extended our understanding into these topics by explaining a method for comparing two sets of point features. In particular, we have demonstrated how this method can be used to compare reference and VGI (test) data sources. We believe that this is an important step in understanding the quality of VGI in relation to other data sources.

As VGI is evolving, both in terms of participation and scope, a better understanding of its quality, the parameters that affect it, and the practices used to produce it will help enhance the utility of its products for geospatial analysis. Based on this initial work, several areas of future work need to be explored. Unconstrained (and untrained) contributors do not always share a common understanding of the definition of "what" a feature is or "where" it should be located and the effect of the vagueness on data quality is not understood. The result of the vagueness can be degradation in the utility of VGI for decision making; however, these effects have not been studied. Lastly, there is a need to improve methods for evaluation of data that is currently labeled as "authoritative" because, as we have shown in this research, these datasets are not without error.

Acknowledgments

The authors would like to thank the US Geological Survey for allowing their data to be used and for providing support during the research. The authors would also like to thank the National Geospatial-Intelligence Agency for their continued support of education (Approved for Public Release 13-117).

Conflicts of Interest

The author declares no conflict of interest.

References

1. Rana, S.; Joliveau, T. NeoGeography: An extension of mainstream geography for everyone made by everyone? *J. Location Based Serv.* **2009**, *3*, 75–81.
2. Tulloch, D.; Shapiro, T. The intersection of data access and public participation: Impacting GIS users' success. *URISA J.* **2003**, *15*, 55–60.
3. Mooney, P.; Corcoran, P.; Winstanley, A. Towards Quality Metrics for OpenStreetMap. In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, San Jose, CA, USA, 2–5 November 2010; pp. 514–517.
4. Goodchild, M. Citizens as sensors: The world of volunteered geography. *GeoJournal* **2007**, *69*, 211–221.
5. Hudson-Smith, A.; Crooks, A.; Gibin, M.; Milton, R.; Batty, M. NeoGeography and Web 2.0: Concepts, tools and applications. *J. Location Based Serv.* **2009**, *3*, 118–145.
6. Howe, J. The Rise of crowdsourcing. *Wired Magazine* **2006**, *14*, 161–165.
7. Brabham, D.C. Crowdsourcing as a model for problem solving: An introduction and cases. *Converg. Int. J. Res. New Media Technol.* **2008**, *14*, 75–90.
8. Heipke, C. Crowdsourcing geospatial data. *ISPRS J. Photogramm.* **2010**, *65*, 550–557.
9. Stefanidis, A.; Crooks, A.; Radzikowski, J. Harvesting ambient geospatial information from social media feeds. *GeoJournal* **2013**, *78,* 319–338.
10. Harvey, F. To Volunteer or to Contribute Locational Information? Toward Truth in Labeling for Crowdsourced Geographic Information. In *Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice*; Springer Science+Business Media: Dordecht, The Netherland, 2013; pp. 31–42.
11. Hall, G.; Chipeniuk, R.; Feick, R.; Leahy, M.; Deparday, V. Community-based production of geographic information using open source software and Web 2.0. *Int. J. Geogr. Inf. Sci.* **2010**, *24*, 761–781.
12. Elwood, S.; Goodchild, M.; Sui, D. Prospects for VGI Research and the Emerging Fourth Paradigm. In *Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice*; Springer Science+Business Media: Dordecht, The Netherland, 2013; pp. 361–375.
13. Flanagin, A.; Metzger, M. The credibility of volunteered geographic information. *GeoJournal* **2008**, *72*, 137–148.
14. Neis, P.; Zipf, A. Analyzing the contributor activity of a volunteered geographic information project—The case of OpenStreetMap. *ISPRS Int. J. Geo-Inf.* **2012**, *1*, 146–165.
15. Haklay, M. How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance survey datasets. *Environ. Plan. B Plan. Design* **2010**, *37*, 682–703.
16. Girres, J.-F.; Touya, G. Quality assessment of the French OpenStreetMap dataset. *Trans. GIS* **2010**, *14*, 435–459.
17. Zielstra, D.; Zipf, A. A Comparative Study of Proprietary Geodata and Volunteered Geographic Information for Germany. In Proceedings of the 13th AGILE International Conference on Geographic Information Science, Guimar ães, Portugal, 10–14 May 2010; Volume 2010, pp. 1–15.

18. Feick, R.; Roche, S. Understanding the Value of VGI. In *Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice*; Springer Science+Business Media: Dordecht, The Netherland, 2013; pp. 15–30.
19. Haklay, M.; Weber, P. OpenStreetMap: User-generated street maps. *IEEE Pervasive Comput.* **2008**, *7*, 12–18.
20. Goodchild, M.F. Data Models and Data Quality: Problems and Prospects. In *Environmental Modeling with GIS*; Oxford University Press: New York, NY, USA, 1993; pp. 94–103.
21. Haklay, M.; Singleton, A.; Parker, C. Web mapping 2.0: The neogeography of the GeoWeb. *Geogr. Compass* **2008**, *2*, 2011–2039.
22. Feick, R.; Roche, S. Valuing Volunteered Geographic Information (VGI): Opportunities and Challenges Arising from a New Mode of GI Use and Production. In Proceedings of the 2nd GEOValue Workshop, Hamburg, Germany, 30 September–2 October 2010; HafenCity University: Hamburg, Germany, 2010; pp. 75–79.
23. Elwood, S. Volunteered geographic information: Future research directions motivated by critical, participatory, and feminist GIS. *GeoJournal* **2008**, *72*, 173–183.
24. Elwood, S. Geographic information science: Emerging research on the societal implications of the geospatial web. *Prog. Hum. Geogr.* **2009**, *34*, 349–357.
25. Sieber, R. Public participation geographic information systems: A literature review and framework. *Ann. Assoc. Am. Geogr.* **2006**, *96*, 491–507.
26. Cooper, A.; Coetzee, S.; Kaczmarek, I.; Kourie, D.; Iwaniak, A.; Kubik, T. Challenges for Quality in Volunteered Geographical Information. In Proceedings of the AfricaGEO 2011 Conference, Cape Town, South Africa, 31 May–2 June, 2011; AfricaGEO: Cape Town, South Africa, 2011; p. 13.
27. Haklay, M. Citizen Science and Volunteered Geographic Information: Overview and Typology of Participation. In *Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice*; Springer Science+Business Media: Dordecht, The Netherland, 2013; pp. 105–124.
28. Coleman, D. Potential Contributions and Challenges of VGI for Conventional Topographic Base-Mapping Programs. In *Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice*; Springer Science+Business Media: Dordecht, The Netherland, 2013; pp. 245–264.
29. Goodchild, M.F. Geographical information science. *Int. J. Geogr. Inf. Syst.* **1992**, *6*, 31–45.
30. Goodchild, M.F.; Hunter, G.J. A simple positional accuracy measure for linear features. *Int. J. Geogr. Inf. Sci.* **1997**, *11*, 299–306.
31. Chrisman, N. The Error Component in Spatial Data. In *Geographic Information Systems and Science*; Longley, P., Goodchild, M., Maguire, D., Rhind, D., Eds.; John Wiley & Sons: Hoboken, NJ, USA, 1991; Volume 1, pp. 165–174.
32. Devillers, R., Jeansoulin, R., Eds. *Fundamentals of Spatial Data Quality*; ISTE: London, UK, 2006.
33. Haklay, M.; Basiouka, S.; Antoniou, V.; Ather, A. How many volunteers does it take to map an area well? The validity of linus' law to volunteered geographic information. *Cartogr. J.* **2010**, *47*, 315–322.

34. Brassel, K.; Bucher, F.; Stephan, E.; Vckovski, A. Completeness. In *Elements of Spatial Data Quality*; Elsevier: Oxford, UK, 1995; pp. 81–108.
35. Al-Bakri, M.; Fairbairn, D. Assessing the Accuracy of "Crowdsourced" Data and its Integration with Official Spatial Data Sets. In Proceedings of the Ninth International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, Leicester, UK, 20–23 July 2010; pp. 317–320. Available online: http://www.spatial-accuracy.org/system/files/img-X06165606_0.pdf (accessed on 23 May 2012).
36. Drummond, J. Positional Accuracy. In *Elements of Spatial Data Quality*; Elsevier: Oxford, UK, 1995; pp. 31–58.
37. Weng, Q. Remote Sensing and GIS Integration, Theories, Methods, and Applications; McGraw Hill: New York, NY, USA, 2010.
38. Longley, P.; Goodchild, M.; Maguire, D.; Rhind, D. *Geographic Information Systems and Science*, 3rd ed.; John Wiley and Sons: Hoboken, NJ, USA, 2011.
39. Worboys, M.; Duckham, M. *GIS A Computing Perspective*, 2nd ed.; Taylor Francis: London, UK, 2004.
40. EPA. *Environmental Dataset Gateway*. Available online: http://edg.epa.gov/metadata/catalog/search/resource/details.page?uuid=%7B794F5D6E-5347-41ED-9594-DCE122FFC419%7D (accessed on 28 November 2012).
41. *Points of Interest—OpenStreetMap Wiki*. Available online: http://wiki.openstreetmap.org/wiki/Points_of_interest (accessed on 3 May 2013).
42. Over, M.; Schilling, A.; Neubauer, S.; Zipf, A. Generating web-based 3D city models from OpenStreetMap: The current situation in Germany. *Comp. Environ. Urban Syst.* **2010**, *34*, 496–507.
43. Poore, B.S.; Wolf, E.B.; Korris, E.M.; Walter, J.L.; Matthews, G.D. *Structures Data Collection for the National Map Using Volunteered Geographic Information*. 2012; p. 34. Available online: http://pubs.usgs.gov/of/2012/1209 (accessed on 12 February 2013).
44. *Overview of Structure Features*. Available online: http://navigator.er.usgs.gov/help/vgistructures_userguide.html (accessed on 3 May 2013).
45. *Private School Universe Survey (PSS)—Overview*. Available online: http://nces.ed.gov/surveys/pss/ (accessed on 4 May 2013).
46. *Common Core of Data (CCD)—What Is the CCD?* Available online: http://nces.ed.gov/ccd/aboutCCD.asp (accessed on 3 May 2013).
47. *OpenStreetMap Wik—Schools*. Available online: http://wiki.openstreetmap.org/wiki/Tag:amenity%3Dschool (accessed on 1 May 2013).
48. Wolf, E.B.; Matthews, G.D.; McNinch, K.; Poore, B.S. *OpenStreetMap Collaborative Prototype, Phase 1*; US Geological Survey Open-File Report 2011–1136; US Geological Survey: Reston, VA, USA, 2011; p. 23. Available online: http://pubs.usgs.gov/of/2011/1136/ (accessed on 25 July 2012).
49. Ushahidi: Open Source Software for Information Collection, Visualization and Interactive Mapping. Available online: http://www.ushahidi.com/ (accessed on 18 February 2013).
50. *US Census Bureau Geographic Definitions*. Available online: http://www.census.gov/geo/www/geo_defn.html#CensusTract (accessed on 27 July 2012).

51. Zook, M.; Graham, M.; Shelton, T.; Gorman, S. Volunteered geographic information and crowdsourcing disaster relief: A case study of the Haitian earthquake. *World Med. Health Policy* **2010**, *2*, 6–32.
52. Matyas, S.; Matyas, C.; Mitarai, H.; Komata, M.; Kiefer, P.; Schlieder, C. Designing Location-Based Mobile Games: The CityExplorer Case Study. In *Digital Cityscapes: Merging Digital and Urban Playspaces*; Peter Lang Publishers: New York, NY, USA, 2009; pp. 187–203.
53. Wolf, E.B.; Poore, B.S.; Caro, H.K.; Matthews, G.D. Volunteer Map Data Collection at the USGS. US Geol. Survey Fact Sheet 2011–3103; US Geological Survey: Reston, VA, USA, 2011; p. 2. Available online: http://pubs.usgs.gov/fs/2011/3103/ (accessed on 25 July 2012).
54. *CloudMade Downloads*. Available online: http://downloads.cloudmade.com/americas/northern_america/united_states/colorado#downloads_breadcrumbs (accessed on 12 February 2013).
55. Wolf, E.B. Re: Notes/Actions from Call Today. 3 November 2011.
56. *7.4. Difflib–Helpers for Computing Deltas–Python v2.7.3 Documentation*. Available online: http://docs.python.org/2/library/difflib.html (accessed on 29 November 2012).
57. Whaley, M. Old Adams City High School to be Renovated. *Denver Post* 15 December 2012. Available online: http://www.denverpost.com/news/ci_22196962/old-adams-city-high-school-be-renovated (accessed on 12 February 2013).
58. Walpole, R.; Myers, R. *Probablity and Statistics for Engineers and Scientists*, 2nd ed.; McMillan Publishing Co.: New York, NY, USA, 1978.
59. Abramson, P. *Hard Economy Hits Construction Planning*; pp. 2–8. Available online: http://www.peterli.com/spm/pdfs/SPM-2009-02-SUPPLEMENT.pdf (accessed on 12 February 2013).
60. Weihs, J. *School Site Size—How Many Acres are Necessary?*; Issuetrak, Council of Educational Facility Planners International: Scottsdale, AZ, USA, 2003; p. 7. Available online: http://media.cefpi.org/issuetraks/issuetrak0903.pdf (accessed on 12 February 2013).