ASSESSING COMPLETENESS AND SPATIAL ERROR OF FEATURES IN VOLUNTEERED GEOGRAPHIC INFORMATION

By Steven P. Jackson, William Mullen, Peggy Agouris, Andrew Crooks, Arie Croitoru, and Anthony Stefanidis

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. As technology continues to improve and the Internet becomes more accessible, the amount of geospatial data generated by citizens without formal geographic training is expected to rapidly increase.¹ Thus, Volunteered Geographic Information (VGI) is bringing the general public into the realm of map production functions traditionally reserved for official agencies.

The focus of this article is on the accuracy of crowdsourced VGI. 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. 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 compare those characteristics against reference sources of data to quantify data quality.
Accuracy and Completeness Considerations for Volunteered Geographic Information

While data quality has been at the center of the research agenda since the definition of GIScience, geographers Michael Frank Goodchild and G.J. 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.\(^2\,^3\) 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.

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. 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.

Quantifying completeness and accuracy of VGI will allow users of the data to better understand the data’s utility. 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.

A Comparison of Data Sources

We aim to extend the current state of knowledge on the topic of VGI quality by focusing on completeness and accuracy of point features within VGI data. Our analysis compared two VGI test datasets against a reference dataset and analyzed their differences. We also examined the particularities of a hybrid variant of VGI, and assessed its impact on the overall accuracy of the VGI product.

![Figure 1](image-url) | School count by data source

<table>
<thead>
<tr>
<th>Source</th>
<th>School Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oak Ridge National Laboratory (ORNL)</td>
<td>402</td>
</tr>
<tr>
<td>OpenStreetMap (OSM)</td>
<td>406*</td>
</tr>
<tr>
<td>OpenStreetMap Collaborative Project (OSMCP)</td>
<td>412</td>
</tr>
</tbody>
</table>

* Includes 48 historical school locations.

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. 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 point of interest layer of OpenStreetMap (OSM). We also use a second test dataset, which is a product of the U.S. Geological Survey (USGS) OpenStreetMap Collaborative Project (OSMCP) Second Phase. 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 provided to the volunteers by a government agency.

Figure 1 presents the total count of schools within the study area for each data source. While the total number of schools is 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 percent historic schools which are likely no longer in existence and would, therefore, not match schools in either ORNL or OSMCP.

The results shown in Table 1 imply that these datasets are similar. 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. 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.
Automated matching of the datasets was carried out using four different methods across each of the name and address fields from the dataset attributes. Our automated methods are incapable of dealing with some cases; therefore, the user must manually examine the unmatched records which remain after the automated processes to determine if any potential matches were missed.

The final process in the analysis includes computation of the intersection, union, and complement. The intersection dataset includes all records that appear in both datasets. The reference complement test dataset includes those records that are in the reference dataset, but not in the test dataset. The test complement reference dataset includes those records that are in the test dataset, but not in the reference dataset. These values can then be used in the computation of the accuracy and completeness.

**Results of the Matching Methods**

Examining the results of the automatic and manual matching methods, Figure 2 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.

**ASSESSING 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. The comparison of ORNL and OSMCP data showed that a total of 89 percent of the records were matched. Of that, 82 percent were matched automatically, while the remaining 7 percent were matched manually, indicating that the automated matching algorithm is successful. The match rate for the ORNL and OSM data comparison was considerably lower, with only 71 percent of the total records matched. However, the manual match rate was over twice as high at 15 percent, with the automated match rate falling to 56 percent. 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.

**Figure 2 | Summary of record counts for data matching**

<table>
<thead>
<tr>
<th>Method</th>
<th>ORNL-OSMCP</th>
<th>ORNL-OSM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intersection</td>
<td>357</td>
<td>287</td>
</tr>
<tr>
<td>Reference Complement Test</td>
<td>44</td>
<td>114</td>
</tr>
<tr>
<td>Test Complement Reference</td>
<td>63</td>
<td>99</td>
</tr>
<tr>
<td>Union</td>
<td>464</td>
<td>500</td>
</tr>
<tr>
<td>Reference Count</td>
<td>401</td>
<td>401</td>
</tr>
<tr>
<td>Test Count</td>
<td>412</td>
<td>406</td>
</tr>
</tbody>
</table>

**Figure 3 | Spatial error for matched schools**

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Count</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>St. Deviation</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>ORNL-OSMCP</td>
<td>357</td>
<td>2</td>
<td>487</td>
<td>47</td>
<td>50</td>
<td>33</td>
</tr>
<tr>
<td>ORNL-OSM</td>
<td>287</td>
<td>2</td>
<td>1,848</td>
<td>190</td>
<td>314</td>
<td>43</td>
</tr>
</tbody>
</table>
Analyses revealed that the OSM and OSMCP efforts captured schools that were not in the ORNL data. 28 percent of the OSM schools remained unmatched at the end of the analyses while 11 percent of the OSMCP records remained unmatched.

**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.

The spatial error is evaluated for each match and the results are located in Figure 3. 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.

Because the error was so different between the two datasets, an additional effort was undertaken to examine the nature of the error distribution. The percent of matched schools within 150 m (cumulative) for both OSMCP and OSM can be shown to be 96 percent. These results indicate that either dataset would be equally capable of getting the user to the school property.

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. Of the 281 matched schools, OSMCP schools were closer 58 percent of the time; however, OSMCP also has the largest difference (224 m).

**Conclusions**

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 captured close to 90 percent of the records in the reference ORNL database, while the OSM data captured approximately 70 percent 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.

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.
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.

Steven P. Jackson is a Ph.D. student in Geography and Geoinformation Science at George Mason University. His specialty is geographic information systems science.

William Mullen is a Ph.D. student in Geography and Geoinformation Science at George Mason University. His expertise is in geographic information systems.

Dr. Peggy Agouris is a Professor with the Department of Geography & Geoinformation Science at George Mason University, and the director of Mason’s Center for Earth Observing and Space Research. She is also currently the interim Dean of Mason’s College of Science. She holds a Ph.D. in Digital Photogrammetry from The Ohio State University. Her areas of expertise include digital image analysis for geospatial information extraction and spatiotemporal information modeling and analysis.

Dr. Andrew Crooks is an Assistant Professor with the Department of Computational Social Science and member of the Center for Social Complexity at George Mason University. He holds a Ph.D. from the University College London’s (UCL) Center for Advanced Spatial Analysis. His area of expertise relates to working with GIS, creating and analyzing large spatial datasets, and finding patterns and insights from such data. Crooks is a recognized expert on the integration of GIS and agent-based modeling (ABM).

Dr. Arie Croitoru is an Assistant Professor with the Geography and Geoinformation Science Department at George Mason University, and a member of Mason’s Center for Geospatial Intelligence. He holds a Ph.D. in Geoinformatics from Technion University (Israel). His research focuses on three interrelated research streams: geosocial analysis and location-based mass collaboration systems (LB-MCS), spatiotemporal data mining, and digital image processing. Currently, Croitoru’s work focuses on developing novel algorithms for modeling, abstraction, and automated knowledge discovery in location-aware social media, and on the design and development of geosocial analysis platforms.

Anthony Stefanidis is a Professor with the Department of Geography and Geoinformation Science at George Mason University, and the director of Mason’s Center for Geospatial Intelligence. He holds a Ph.D. in Digital Photogrammetry from The Ohio State University. His areas of expertise include the analysis of digital imagery and video, spatiotemporal information modeling and analysis, sensor networks, and the harvesting of geospatial information from social media.

EXCERPT FROM

REFERENCES