Spatial Accuracy 2014, East Lansing, Michigan, July 8-11.

Exploring the accuracy of crowdsourced annotations of post-disaster building damage derived from fine spatial resolution satellite sensor data.

Giles M. Foody
School of Geography, University of Nottingham, Nottingham, NG7 2RD, UK

Abstract

The accuracy of crowdsourced annotations of building damage after the Haiti 2010 earthquake is explored. Amateur annotators provided labels that were of comparable quality to those from experts and the quality of labeling varied with the nature of the satellite imagery provided. The accuracy of estimates of the amount of damaged buildings could be increased via a latent class analysis.

Keywords: Crowdsourcing, volunteered geographic information, map accuracy.

1. Introduction

Following major disasters there is often an urgent need for maps of the affected region. For example, disaster relief and associated humanitarian protection applications require maps of building damage to illustrate the extent of the damage and to aid relief planning. Often, the 'need for speed' is critical and as remote sensing has the ability to acquire imagery anywhere in the world at a range of spatial scales and at potentially high temporal frequency, it has considerable potential for the provision of information to aid post-disaster mapping activities. Similarly, distributing the mapping task among the crowd and utilizing the power of citizen science, a large task can be undertaken quickly and cheaply. A key concern though is the accuracy of the resulting maps.

Post-disaster damage mapping is typically based on visual interpretation and the potential of crowdsourcing to acquire the annotations is well-established. There are, however, many concerns, especially with the quality of the data derived. Indeed many bodies produce maps but the information they convey often differs and at times may even be contradictory, hence having the potential to hinder rather than aid post-disaster relief applications. A major mapping challenge has been identified in post-disaster mapping applications, highlighting the need for information on the accuracy of data from the various different sources contributing to the task (van Aardt *et al*., 2011; Voigt *et al*., 2011). This article aims to explore some of the key issues associated with the use of crowdsourcing in post-disaster damage mapping. In particular, it addresses the accuracy of building damage mapping, with particular regard to the accuracy of damage mapping and of estimates of the amount of damaged buildings.

2. Data and Methods

Attention is focused on 97 buildings in the Port-au-Prince region for which ground data on building damage had been acquired by field survey undertaken shortly after the

Haiti 2010 earthquake (Booth *et al*., 2012). For each building this data set yielded a label indicating the degree of damage on a standard five class classification of building damage, ranging essentially from no damage to complete destruction. An additional set of labels for each were derived from another post-disaster mapping activity, the GEO-CAN (Booth *et al*., 2012), which made use of labels provided by expert annotators.

Each building was located within a set of fine spatial resolution satellite sensor images and a small region containing each building extracted. The images were acquired by GeoEye-1 on 1 October 2009, before the earthquake, and between 13-16 January 2010 by GeoEye-1, Quickbird and WorldView-2. These satellite remote sensing systems acquired imagery with spatial resolutions of approximately 0.4-0.6 m and 1.6-2.6 m in panchromatic and multispectral modes respectively.

The image extracts were used to form a data set that was presented to a group of 8 amateur annotators willing to attempt to label the degree of building damage using the same five class classification as employed with the ground data. The data distributed to the annotators showed the buildings in four scenarios: I only a post-earthquake GeoEye-1 colour image, II a pre- and post- earthquake GeoEye-1 colour images, III the data of scenario II supplemented with panchromatic imagery from Quickbird and WorldView-2 and IV scenario II supplemented with multi-spectral images from Quickbird and WorldView-2. In this way it was hoped to assess how the nature of the available data impacted on the accuracy of annotator labeling. In total each annotator was provided with 388 cases to label.

To simplify analyses and allow comparison to other studies, the data on building damage was degraded into a binary classification, essentially reducing the damage labeling to classes of little and severe damage. For this, the three highest damage classes were grouped together to yield a high damage class while the remaining cases formed a no or little damage class. All analyses used only the data in this binary classification.

A series of analyses were undertaken with the available data. This article will focus on the evaluations of the accuracy with which each annotator labeled the cases, assessments of inter-annotator labeling accuracy, a comparison of annotator accuracy relative to a more authoritative set of labels derived from the GEO-CAN and finally an assessment of the ability to increase the accuracy of estimates from the crowdsourced data by means of ensemble and latent class analyses. In the latter it was assumed that the annotators would struggle on the same cases and so that errors were correlated.

3. Results and Discussion

Five key trends were evident in the results. First, the accuracy of the crowdsourced labels was low and varied greatly between annotators. As a guide, the percentage of cases labeled correctly by the 8 annotators was typically ~65% and the largest kappa coefficient observed was 0.25. Second, while low, the level of accuracy observed was actually broadly comparable to that derived from more expert sources; the kappa coefficient for the comparison of GEO-CAN labels against the reference data set was 0.14. Third, the nature of the data provided to the annotators was important. With some annotators it was evident that accuracy increased when they were provided with several images and preferably with multispectral rather than panchromatic imagery. However, the one universal trend in relation to the nature of the data used was that availability of pre-disaster imagery increased the accuracy of labeling; there was typically an increase in accuracy of ~7% when pre- and post-disaster images were available to inform the

labeling. Fourth, simple ensemble approaches, such as the dominant label allocated to a case determined over all annotators, could be used to try and allow for differences between annotators in performance. With this approach the degree of agreement between the modal class label derived from the 8 annotators relative to the reference data set rose with progression from scenario I to IV: kappa coefficients of 0.04, 0.14, 0.18 and 0.19 for the scenarios in order from I to IV. The modal class label was also more strongly related to the GEO-CAN labels than with the reference data, with a kappa coefficient of up to 0.43 observed for the assessment based on the modal label derived from the use of pre- and post-disaster images supplemented with panchromatic fine spatial resolution imagery (scenario III). Fifth, although the annotators provided data sets that showed little agreement and were highly imperfect their use in a latent class model enabled useful data to be derived. In particular, the latent class model provided estimates of the accuracy of labeling from each annotator and enabled derivation of an estimate of the amount of building damage that was close to reality to be derived. The latter is of particular interest given the desire to estimate the amount of building damage after a disaster. As one example, the estimate of buildings in the high building damage class derived from the use of the ensemble approach for scenario III was 13.4%. The estimate derived from the latent class analysis was 40.4%, which was much closer to that depicted in the ground reference data set: 44.3%. Thus while the annotators tended, as is common, to underestimate the amount of damage substantially the estimate derived from their combined contributions was much closer to reality.

4. Conclusions

Crowdsourcing has considerable potential as a source of data for damage mapping. Here amateur annotators derived results of comparable accuracy to experts. The absolute accuracy was low but could be enhanced by provision of appropriate pre- and post-disaster images. A latent class analysis may also be used to generate enhanced estimates of building damage.

Acknowledgements

This work benefitted from funding the EPSRC (reference EP/J0020230/1) and British Academy (reference SG112788) as well as the reviewers comments.

References

Booth, E. Saito, K and Madabhushi, G (2012) "*The Haiti Earthquake of 12 January 2010: A Field Report by EEFIT*," The Earthquake Engineering Field Investigation Team, The Institution of Structural Engineers, London.

van Aardt, J. A. N., McKeown, D., Fauiring, J., Raqueno, N., Caterline, M., Renschler, C., Eguchi, R., Messinger, D., Krzaczek, R., Cavilla, S., Antalovich, J., Philips, N., Bartlett, B., Salvaggio, C., Ontiveros, E. and Gill, S. (2011) "Geospatial disaster response during the Haiti earthquake: a case study spanning airborne deployment data collection, transfer, processing and dissemination", *Photogrammetric Engineering and Remote Sensing*, vol. 77, 943-952.

Voigt, S., Schneiderhan, T., Tweie, A., Gahler, M., Stein, E. and Mehl, H. (2011) "Rapid damage assessment and situation mapping: learning from the 2010 Haiti earthquake", *Photogrammetric Engineering and Remote Sensing*, vol. 77, pp. 923-931.