Visual Analytical Approaches to Evaluating Uncertainty and Bias in Crowdsourced Crisis Information

Iain Dillingham* Jason Dykes† Jo Wood, *Member, IEEE*‡

giCentre, School of Informatics
City University London

Abstract

Concerns about verification mean the humanitarian community are reluctant to use information collected during crisis events, even though such information could potentially enhance the response effort. Consequently, a program of research is presented that aims to evaluate the degree to which uncertainty and bias are found in public collections of incident reports gathered during crisis events. These datasets exemplify a class whose members have spatial and temporal attributes, are gathered from heterogeneous sources, and do not have readily available attribution information. An interactive software prototype, and existing software, are applied to a dataset related to the current armed conflict in Libya to identify 'intrinsic' characteristics against which uncertainty and bias can be evaluated. Requirements on the prototype are identified, which in time will be expanded into full research objectives.

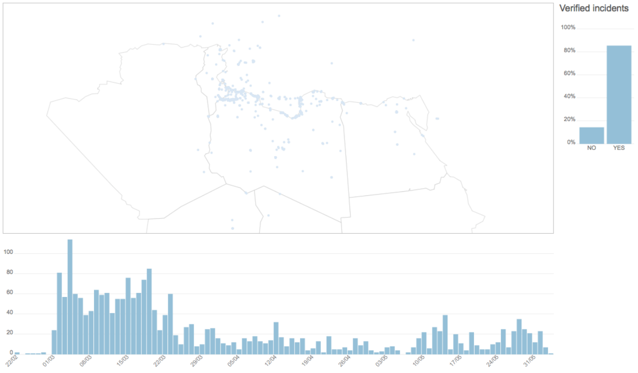

Figure 1: IncidentExplorer, Libya dataset

1 Introduction

Crowdsourcing describes the process by which tasks are completed by a heterogeneous group in response to an open call [5]. Whilst examples of crowdsourcing are generally business-focused [4, 5], recently the process has been used outside the business community to gather reports about populations directly affected by crisis events, such as the 2010 earthquake in Haiti, or the current armed conflict in Libya. However, whilst it is argued that formal responses to crisis events should accommodate crowdsourced information [8], verifying information collected during a crisis event is problematic [1]. Indeed, verification is the principal obstacle to humanitarian organisations using crowdsourced information to make decisions 'in the field' [10].

Verification, in this context, is associated with accuracy—"the inverse of error" [13, p.178]—and credibility [1]. Accuracy and credibility, alongside precision, completeness, consistency, lineage, currency, subjectivity, and interrelatedness, are components of uncertainty [7]. Many of these components have spatial, temporal, and thematic aspects [13]. Bias, by extension, can be defined as systematic error [13].

In our research, visual analytical approaches are used to evaluate the degree to which uncertainty and bias are found in public collections of incident reports gathered during crisis events. We use visual analytical approaches because they have been effective in studies with similar datasets [14], or with similar aims [16]. Our datasets relate to the 2010 earthquake in Haiti, and the current armed conflict in Libya, and were exported from the Haiti[1] and Libya[2] crisis maps; both are instances of Ushahidi,[3] an open source software platform that was built to gather information from 'tweets,' SMS messages, emails, and the web. Ushahidi allows anyone to report an incident, and incident reports are generally reviewed ('approved' and 'verified') by a restricted group before being made public. Consequently, two forms of crowdsourcing characterise Ushahidi: The first applies to reporting incidents and is consistent with the definition given above; the second applies to reviewing incident reports and is a form of moderation. However, it is important to note that not all of the information contained in each incident report is made public—the reporter's Twitter account, telephone number, and email address are not disclosed, for example—and that our research encompasses only the publicly available information.

Although there are compelling reasons to use the Haiti and Libya datasets specifically, they exemplify a class whose members have spatial and temporal attributes, are gathered from heterogeneous sources, and do not have readily available attribution information (i.e. information about the report, reporter, or reviewer). Visual analytical approaches are well placed to "detect the expected and discover the unexpected" in such circumstances [11, p.10]. Furthermore, exploring the relationships between the components that characterise uncertainty in different domains is a recognised research challenge in geographic information science [7]. Indeed, addressing data quality issues such as uncertainty could also benefit the wider research community [6].

In the following sections we state our aim, and describe how we have addressed our first objective using existing and new software. We describe the nature of the Haiti and Libya datasets, and conclude with possible directions for future research.

2 Exposition

The aim of our research is to evaluate the degree to which uncertainty and bias are found in public collections of incident reports gathered during crisis events. Whilst previous research used the

*e-mail:iain.dillingham.1@city.ac.uk
†e-mail:j.dykes@city.ac.uk
‡e-mail:j.d.wood@city.ac.uk

[1] http://haiti.ushahidi.com/
[2] http://libyacrisismap.net/
[3] http://www.ushahidi.com/

contribution frequency of users to evaluate bias in collections of user-generated content [9], the Haiti and Libya datasets lack attribution information. However, it should be possible to identify similar 'intrinsic' characteristics against which uncertainty and bias can be evaluated. Our first objective is to identify these characteristics.

Achieving our first objective necessitates 'getting to know' the data, a crucial component in effective data analysis [12]. Exploring the Libya dataset with existing software told us that it contains 2283 incident reports, each with two spatial (a coordinate pair and a location string), one temporal, and five 'thematic' attributes that describe and categorise each incident. The location strings are 'messy' in that they contain toponyms ("Ajdabiya Central Hospital"), coordinate pairs at different levels of precision, 'vernacular geographies' [3] ("Between Sharia as-Sayiti Street [and] Az Zawiyah Street, Tripoli, Libya"), and in some cases additional explanatory information ("Cyrinacia – older regional term meaning eastern coastal region of Libya."). Furthermore, 94% of reports in the Libya dataset are categorised as 'Geo-Located' (Ushahidi categories are similar to social media 'tags' in that they are not mutually exclusive), suggesting they are spatially accurate.

We developed an interactive software prototype called IncidentExplorer (Figure 1) to explore the Libya dataset in linked spatial (upper-left), temporal (bottom-left), and thematic (right) views. Using this tool, we see that:

- Most incidents were reported on or near to the coast, with concentrations on the north-west border with Tunisia, and in the north-east coast (Ras Lanuf to Benghazi).

- The temporal distribution of incident reports has a positive skew, with a peak on 4th March 2011 (day 10 of 102).

- Just over 80% of incident reports were 'verified.' Although on most days the proportion of 'verified' reports exceeds 'unverified' reports, the reverse is true at the 'ends' of the dataset. (All incident reports were 'approved.')

We identified several requirements on IncidentExplorer when exploring the Libya dataset. The first concerns the relationship between the coordinate pair, which locates the incident report on the spatial view, and the location string. Although roughly 78% of latitude and 75% of longitude values have six decimal places of precision, this precision does not appear to be reflected in the location strings: There are 86 "Tripoli, Libya" location strings (or similar), and 85 "Benghazi, Libya" location strings (or similar), for example. Given the desire to reach populations directly affected by crisis events, we would expect to see more location strings with greater precision (i.e. more location strings with finer spatial resolution). To explore the precision of location strings further, we wish to (1) display the location strings of the incident reports selected in the spatial and temporal views; and (2) use the location strings to classify the precision of incident reports, and represent this in IncidentExplorer. Both would allow us to assess whether the spatial precision of incident reports varies in space and time; any systematic variation would suggest bias.

Similarly, we wish to determine whether the coordinate pairs are accurate. Two published methods warrant further investigation; the point-radius method [15] and the probability distribution method [2]. The former would result in an object, and the latter a field within which the incident report is likely to be located.

Further requirements on the software prototype include extending the thematic view to include categorical and descriptive information about each incident report. The latter will require further analysis, as the Libya dataset contains 123 categories, some of which are synonymous (e.g. "Water and Sanitation" and "WATSAN").

3 Conclusion

We present a program of research on uncertainty and bias in crowd-sourced crisis information. Having developed a software prototype to address our first objective, we identify several requirements which in time will be expanded into full research objectives. Although these full objectives concern precision and accuracy, the potential exists to explore other components of uncertainty in future work.

Acknowledgements

We thank those individuals behind the Haiti and Libya crisis maps for making their data publicly available.

References

[1] D. Coyle and P. Meier. New technologies in emergencies and conflicts: The role of information and social networks. Technical report, UN Foundation–Vodafone Foundation Partnership, Washington DC and London, 2009.

[2] Q. Guo, Y. Liu, and J. Wieczorek. Georeferencing locality descriptions and computing associated uncertainty using a probabilistic approach. *International Journal of Geographical Information Science*, 22(10):1067–1090, 2008.

[3] L. Hollenstein and R. Purves. Exploring place through user-generated content: Using Flickr tags to describe city cores. *Journal of Spatial Information Science*, (1):21–48, 2010.

[4] J. Howe. The rise of crowdsourcing. *Wired*, 14(6):176–183, 2006.

[5] J. Howe. *Crowdsourcing: How the Power of the Crowd is Driving the Future of Business*. Random House, London, 2009.

[6] D. A. Keim, F. Mansmann, J. Schneidewind, and H. Ziegler. Challenges in visual data analysis. In *Information Visualization*, pages 9–16, 2006. The 10th International Conference on Information Visualization, London, 5–7 July 2006.

[7] A. M. MacEachren, A. Robinson, S. Hopper, S. Gardner, R. Murray, M. Gahegan, and E. Hetzler. Visualizing geospatial information uncertainty: What we know and what we need to know. *Cartography and Geographic Information Science*, 32(3):139–160, 2005.

[8] L. Palen and S. B. Liu. Citizen communications in crisis: Anticipating a future of ICT-supported public participation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 727–736, 2007. San Jose, CA, 30 April–3 May 2007.

[9] R. S. Purves, A. J. Edwardes, and J. Wood. Describing place through user generated content: Comparing two georeferenced image collections in the British Isles. In press.

[10] A. H. Tapia, K. Bajpai, J. Jansen, J. Yen, and L. Giles. Seeking the trustworthy tweet: Can microblogged data fit the information needs of disaster response and humanitarian relief organizations. In *Proceedings of the 8th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*, 2011. Lisbon, Portugal, 8–11 May 2011.

[11] J. J. Thomas and K. A. Cook. A visual analytics agenda. *IEEE Computer Graphics and Applications*, 26(1):10–13, 2006.

[12] A. Unwin, M. Theus, and W. Härdle. Exploratory graphics of a financial dataset. In C. Chen, W. Härdle, and A. Unwin, editors, *Handbook of Data Visualization*, Springer Handbooks of Computational Statistics, pages 831–852. Springer, 2008.

[13] H. Veregin. Data quality parameters. In P. A. Longley, M. F. Goodchild, D. J. Maguire, and D. W. Rhind, editors, *Geographical Information Systems, Principles and Technical Issues*, volume 1, pages 177–189. John Wiley and Sons, Chichester, 2nd edition, 1999.

[14] J. White and R. Roth. TwitterHitter: Geovisual analytics for harvesting insight from volunteered geographic information. In *GIScience 2010*, 2010. Zurich, Switzerland, 14–17 September 2010.

[15] J. Wieczorek, Q. Guo, and R. Hijmans. The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. *International Journal of Geographical Information Science*, 18(8):745–767, 2004.

[16] J. Wood, D. Badawood, J. Dykes, and A. Slingsby. BallotMaps: Detecting name bias in alphabetically ordered ballot papers. *IEEE Transactions on Visualization and Computer Graphics*, 17(6), 2011.