Characterising Locality Descriptions in Crowdsourced Crisis Information

Iain Dillingham, Jason Dykes and Jo Wood

giCentre, City University London

GISRUK 2012

- Who?
- Where?
- What?

- Social media are an important source of information during a crisis.
- Humanitarian organisations are reluctant to use this information in the response effort because untrustworthy and inaccurate information can cost lives.
- Organisations such as Ushahidi use crowdsourcing to address these concerns, but crowdsourcing introduces further uncertainty.
- We're interested in evaluating the uncertainty, and the potential bias, in crowdsourced crisis information.
- Crowdsourced crisis information is highly geographic; there are several characteristics of uncertainty relevant to geographic information (including trust and accuracy).
- We started with accuracy.

What were our research questions?

1. What types of locality descriptions are present in crowdsourced crisis information?
2. Are the proportions of these types different to those present in related datasets?

1. What types of locality descriptions are present in crowdsourced crisis information? (i.e. classification)
2. Are the proportions of these types different to those present in related datasets? (i.e. comparison)

How did we address our research questions?

1. Classify
2. Compare

- We classified locality descriptions in about 3,600 incident reports related to the 2010 earthquake in Haiti.
- We used a classification that was developed alongside a georeferencing method (the point-radius method). Why?
  - The method was developed to georeference locality descriptions.
  - It provides an estimate of the uncertainty associated with the georeferencing process.
  - It has been applied to a variety of related datasets:
    - Records of artefacts in natural history collections; these records are stored in MaNIS (the Mammal Networked Information System).
    - Historical records of search and rescue incidents.
- We compared the Haiti and MaNIS datasets.

- Two papers discuss the categories of locality descriptions in the MaNIS dataset. The categories they identify are slightly different. We combined them. Our combined classification has 12 categories.
- Three participants independently classified the locality descriptions in the Haiti dataset. (The classification process took about 4 hours for each participant to complete.)
- The order of the locality descriptions was randomised for each participant.
- Each participant was guided by the definitions in the table and examples from the MaNIS dataset.

How did we address our research questions?
Classification

Code	Category
U	Unsure
C	Coordinates
F	Feature
P	Path
J	Junction
FOH	Offset from a feature or path at a heading
NF	Near a feature or path
FS	Subdivision of a feature or path
FOO	Orthogonal offsets from a feature
FH	Heading from a feature, no offset
FO	Offset from a feature or path, no heading
BF	Between features or paths

Table: Combined classification of locality descriptions

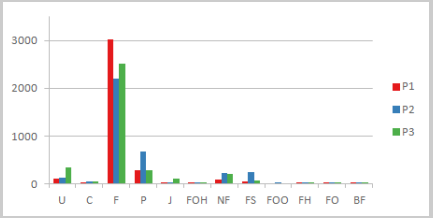

Figure: Category frequency by participant, Haiti dataset

Here we see the frequency of locality descriptions in each category, by each participant, for the Haiti dataset.

- 'Feature' is 1st for all participants.
- 'Path' is 2nd for P1 and P2, 3rd for P3.
- 'Unsure' is 2nd for P3, 3rd for P1 and 5th for P2.

Using *Fleiss' kappa*, we tested the degree to which the observed amount of agreement between the participants exceeded what would be expected if each participant were to categorise each locality description at random.

- $\kappa = 0.42$ ('moderate agreement')

(Poor; Slight; Fair; **Moderate**; Substantial; Almost perfect)

Code	Frequency
F	2570
U	419
P	295
NF	160
FS	57
C	37
J	34
BF	17
FH	13
FOH	3
FO	1
FOO	0

Table: Category frequency, Haiti dataset

Method:

- We classified partial agreement cases by simple majority vote.
- We classified disagreement cases as ‘Uncertain’.

Here we see the frequency of locality descriptions in each category, for the Haiti dataset.

- Most cases (about 71%) are ‘Feature’.
- 419 cases (about 12%) are ‘Uncertain’.

What did we find?
Comparison

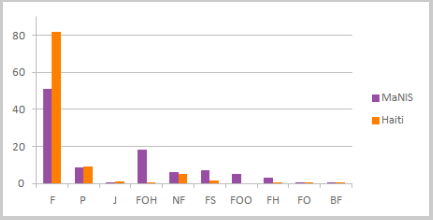

Figure: Category proportion, MaNIS and Haiti datasets

Method:

- We removed all 'Uncertain' and 'Coordinates' cases (about 13% of the Haiti dataset) to allow a like-for-like comparison with the MaNIS dataset.

Here we see the proportion of locality descriptions in each category, for the MaNIS and Haiti datasets.

- Like the MaNIS dataset, most locality descriptions in the Haiti dataset describe features. However, the proportions are very different (MaNIS about 51%; Haiti about 82%).
- There are similar proportions of paths (about 9%) and junctions (about 1%).

Code	MaNIS (#)	Haiti (#)	MaNIS (%)	Haiti (%)
F	1	1	51.0	81.6
P	3	2	8.6	9.4
NF	5	3	6.2	5.1
FS	4	4	7.2	1.8
J	8	5	0.8	1.1
BF	10	6	0.2	0.5
FH	7	7	3.2	0.4
FOH	2	8	18.2	0.1
FO	9	9	0.4	0.0
FOO	6	10	5.2	0.0

Table: Category rank and proportion, MaNIS and Haiti datasets

Here we see the rank and proportion of locality descriptions in each category, for the MaNIS and Haiti datasets, ordered by the Haiti dataset.

- 0.5% of locality descriptions in the Haiti dataset contain offsets and headings (FH, FOH, FO, FOO—the lower four rows). The figure is 27.0% for the MaNIS dataset.
- 91.0% of locality descriptions in the Haiti dataset are either features or paths (F, P—the upper two rows). The figure is about 60% for the MaNIS dataset.

How did we interpret what we found?

Rare in the Haiti dataset:

- 'West of...'
- '10km north of...'
- '5km outside of...'
- '1km north, 3km west of ...'

- Location is seldom described using offsets and headings in the Haiti dataset. (FH, FOH, FO, FOO)
- Location is often described in terms of features (such as named places) and paths (such as roads). (F, P)

How did we interpret what we found?

Common in the Haiti dataset:

- e.g. "Lillavois 47", "Santo"
- e.g. "Rue Pierre Anselme", "Route de Tabarre"

- This finding appears to conform previous research: 'nature' and 'phase' affect the information about an emergency event. For example, during the impact and recovery phases, information from social media contains a high proportion of GI, and this GI relates to well-defined geographic objects (Vieweg et al., 2010). This finding might generalise!
- The uncertainty associated with georeferencing features and paths should be less than that associated with offsets and headings, as they have to be offsets and headings from somewhere.
- If we were to georeference the locality descriptions in the Haiti dataset, we could expect reasonably accurate results. We could compare these results to the locations produced by crowdsourcing to evaluate one aspect of the uncertainty in crowdsourced crisis information.

How did we interpret what we found?

Ambiguity The doubt associated with the classification of a phenomenon (Fisher, 1999).
Vagueness The problem of definition; the Sorites Paradox (Fisher, 1999).
Precision The amount of detail (Veregin, 1999).

- Remember *Fleiss' kappa*? There was 'moderate agreement' between the participants. (For example, about 12% of cases in the Haiti dataset are 'Uncertain'.) In other words, *ambiguity* presents a significant challenge.
- About 5% of the Haiti dataset are 'Near a feature or path'. So, there is also *vagueness*. Anecdotally, there are many *vague places* such as references to IDP (Internally Displaced Person) camps (*vernacular geography* and *naive geography*).
- We didn't investigate the *precision* (resolution, scale) of locality descriptions.

- Locality descriptions in the Haiti dataset favour more certain locations (e.g. features, paths) rather than less certain locations (e.g. that contain offsets and headings).
- Georeferencing these locality descriptions could provide a basis for comparison with the locations produced by crowdsourcing.
- But ambiguity and vagueness present significant challenges: It's complex!

We plan to:

- use alternative sources of information (e.g. OpenStreetMap) to reduce ambiguity and vagueness, and explore precision;
- explore related datasets (e.g. Libya);
- develop a geovisualization tool for exploration and analysis.

Contact

iain.dillingham.1@city.ac.uk
dillingham.me.uk

giCentre
School of Informatics
City University London
Northampton Square
London
EC1V 0HB

We would like to thank:

- Roger Beecham and Sarah Goodwin;
- those behind the Haiti Crisis Map;
- the anonymous reviewers.