Multiway sensitivity analysis of the fusion of earth observation, topography and social media data for rapid flood mapping

Didier G. Leibovici *, Julian F. Rosser

Nottingham Geospatial Institute
University of Nottingham, Nottingham, UK

*Corresponding author: didier.leibovici@nottingham.ac.uk

Abstract
Social media can be a valuable source of information for both detecting and mapping hazard events. Rapidly estimating flood inundation extent can aid decision-making during crises and help in damage assessment. A data fusion method, based on a Bayesian statistical model, which uses weights of evidence to calculate and combine the variables according to their influence in order to map the flooded areas, was previously developed for an experimental timeframe covering the flooding in Oxford, UK during January 2014. The method used three data sources: geotagged photographs, optical remote sensing and high resolution terrain mapping. In this work we aim at evaluating the sensitivity of this method. The multivariate sensitivity analysis proposed, uses an empirical non-parametric approach to generate the variations of the posterior probability of water presence from different levels of prior uncertainties of the input data sources. The sensitivity assessment is spatially depicted as a multiway correspondence analysis of the generated multi-entry table: uncertainty levels for each of the three data sources, the geographical space and the descriptive measures of the output variation*, i.e.* a 5-way array.

Keywords
crowdsourcing, social media, inundation,

I NTRODUCTION

Crowdsourcing has been used in various disaster management situations such as earthquakes (Barrington et al. 2011) and flood damage (Tomnod 2015). Typically, these crowdsourcing scenarios are task orientated. For example, the volunteers are assessing the aftermath of a disaster from satellite image interpretation. Increasingly, analysis may be undertaken on non-authoritative social media reports. For example, analysis of Tweets has been used to source data used in quasi-real time for flood extent estimation (Smith *et al.* 2015), where a hydrological flooding extent model with initialisations from the volunteered observed locations of flood. Rosser *et al.* (2016) proposed to use a data fusion method combining multiple evidence of water presence including social media data sources to estimate the inundation extent. This paper aims to perform a sensitivity analysis of that approach. The data fusion method utilises viewshed analysis (Panteras *et al.* 2015) and weights of evidence (Tehrany *et al*. 2014). The workflow adopted is shown in Figure 1 with the elements involved in the sensitivity analysis reported here highlighted in green (see section II). The three domains of data sources involved in the method are: Landsat-8 water mapping based on the Modified Normalised Difference Water Index (Xu, 2006), geotagged photographs sourced from Flickr to estimate the citizen's extent estimation based on viewshed analysis, and modelled topographic variables sourced from a high resolution Digital Elevation Model.

A classical approach to analyse the sensitivity of a model is to vary factors one-at-a-time and record the effects on the output, such as previously undertaken in multi-criteria mapping problems (Chen et al., 2013). Here, we suggest a multivariate sampling for each combination of the uncertainty levels for the three dimensions. After initial estimation of potential input uncertainties, we modulate these prior uncertainties at different levels to estimate the variation in output uncertainty from the Monte Carlo simulations. The output uncertainty of the posterior probability of water presence is summarised from the simulated distributions in each combination of the input uncertainties. The generated multi-entry table capturing the uncertainty associations (illustrating how sensitive the rapid inundation extend method is), is then analysed from multiway correspondence analysis (Leibovici 2010).

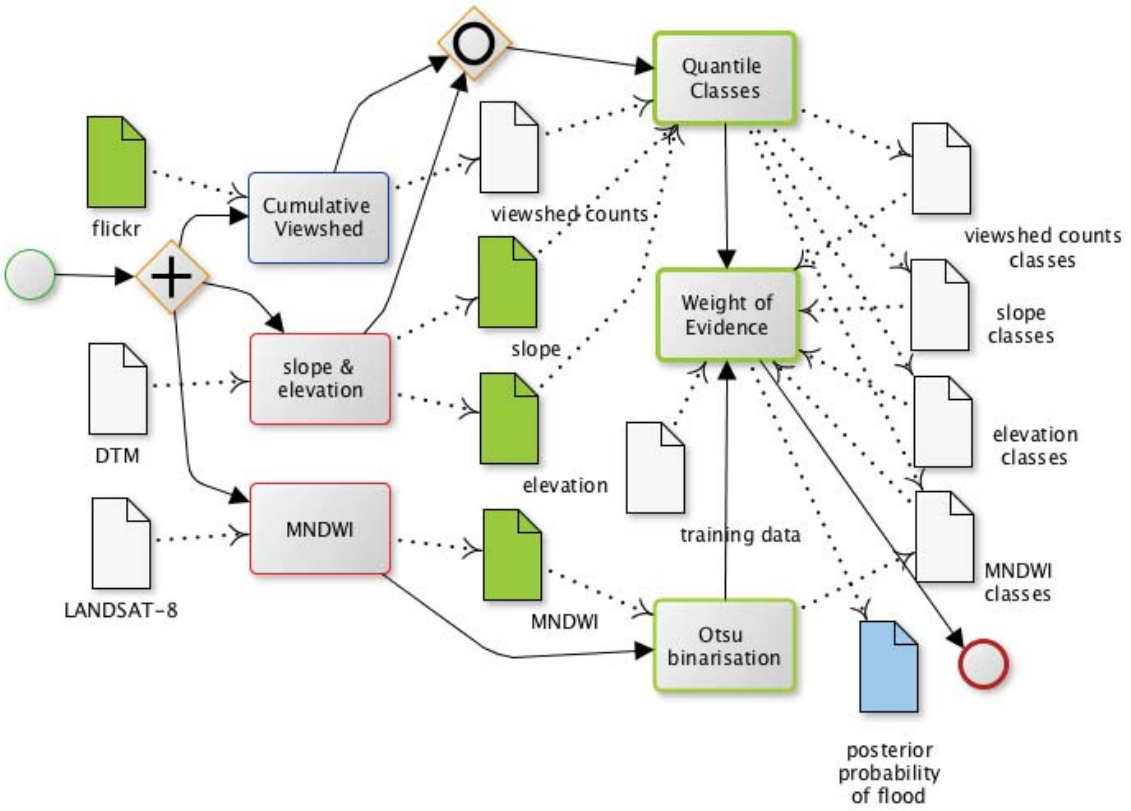

Figure 1: BPMN workflow of the rapid flood inundation data fusion method used in Rosser *et al.* (2016) – in green are the data input where uncertainty sampling takes place.

The method is quite computationally challenging as for each of the 6 174 000 pixels of the studied area, 64 000 evaluations of the workflow are performed: 4 levels of variations in each of the three dimensions with 1000 simulations. On a smaller area of 759 980 pixels (12% of the total area) 1 simulation took 108 seconds (1 core 2.7Ghz), so potentially taking 666 days for the whole study area without parallelisation. For the smaller area and parallelising the code using 20 cores and with only 100 simulations per cell, this is reduced to ~22 hours computation time and will be our first experiment.

II UNCERTAINTIES IN THE DATA FUSION WORKFLOW

Despite the use of quantile classes in the weight of evidence method (see details in Rosser *et al*. 2016) the uncertainty of the measurements or data capture could make the output posterior probability of water presence uncertain. The purpose of the sensitivity analysis is to evaluate

the effect of the uncertainties and to highlight how one data source alone, or in combination with another, can impact on the efficacy of the model.
The whole study area around the city of Oxford, UK, and its characteristics are described in Rosser *et al.* 2016; the data collection coverage for the January 2014 flood event are represented in Figure 2. Ground truth delineation of the flood extent was based on expert interpretation of colour infra-red aerial photography, provided by the Environment Agency Geomatics group (see Figure 6).

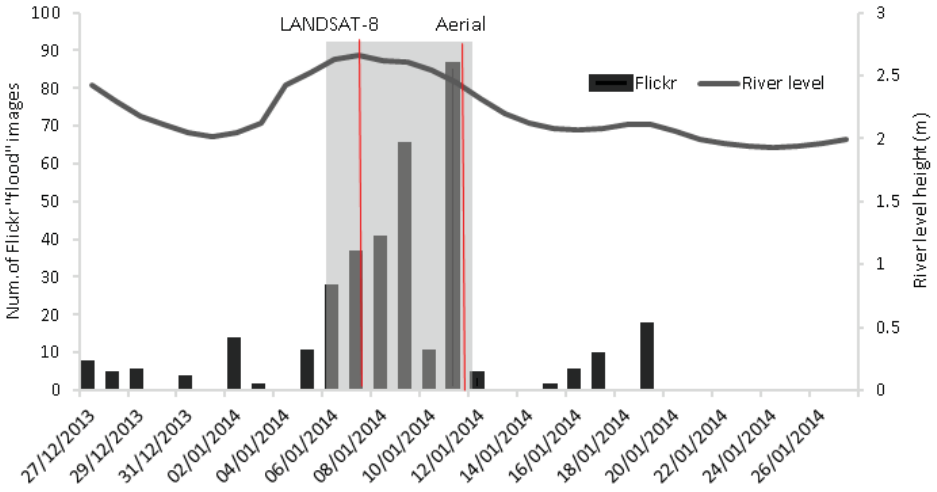

Figure 2: The number of Flickr flood images, Landsat-8 and aerial image (the latter used for ground-truth) acquisition times over the experimental time frame (shaded area). The gauged river level height is also shown.

Besides the representivity and adequation of the data sources for the study, which can be discussed as potential bias, the basis uncertainties taken into account per data source dimension are:

- Landsat-8 dimension

The satellite image matching the peak river level is used for the whole study period, so some variations could be expected in the MNDWI. In the first instance MNDWI uncertainty was estimated from the overall spatial variation of the MNDWI itself but a locally dependent (say 25 pixels' neighbourhood) variance map could be used on the index or directly on the wavelength bands. This gave a standard deviation of 0.22.

- topographic dimension

The Lidar DTM has a given horizontal and vertical uncertainty which changed due to resampling from 1m pixel size to 5m pixel size. To simplify the experiment, slope and elevation uncertainties were estimated from the observed variations globally, respectively 3 and 16 as 1 standard deviation.

- Flickr dimension

The geolocation of the social media data was resampled using as basis a 68% circular error, usually given by the mobile phone (corresponding to 1 standard deviation). As no such information was available in the Flickr data, we took as basis a value of 5m.

The weight of evidence method assumes conditional independence of the 'theme variables' (the predictors) given the outcome:

$$p(V_1 V_2 ... V_m \mid O) = p(V_1 \mid O) p(V_2 \mid O) ... p(V_m \mid O) \quad (1)$$

Assuming normal distribution of the multivariate input, we will sample the multivariate inputs as univariates with the additional assumption that the dependency to the outcome lies in the mean parameter of the distribution.

III MULTIVARIATE SENSITIVITY ANALYSIS

The multivariate sensitivity analysis applied here is based on error propagation through the data fusion workflow. Modulation of the basis uncertainties (section II) was undertaken with four levels of standard deviations: ½ *sd*, 1 *sd*, 2 *sd*, 3 *sd* where *sd* is the basis uncertainty as described in section II. Monte Carlo simulations were derived with normal distributions centred on the observed data values with standard deviations as the modulated uncertainties. A modified workflow taken from Figure 1 was used as the tasks in red ('MNDWI' and 'Slope & Elevation') did not need to run during the sensitivity analysis. Also, the other tasks ('Quantile classification', 'Otsu binarisation' and 'Weight of Evidence') were run using the initial estimated parameters or weights. Therefore the classes adopted the same cut-points and W+ and W- values for the weight of evidence associated to these classes for each simulation. Running the whole workflow with new class and weight estimations each time would encompass a sensitivity analysis, but also a reliability analysis of the workflow itself. Studying the variation of the W+ and W- values would nonetheless be an important reliability and model adequacy feature of the approach to investigate in future work.

The resulting uncertainty of the multivariate input variations has been summarised (using min, Q1, median, mean, Q3, max, var of the simulated distributions at each pixel) in the 5-way array shown in Figure 3.

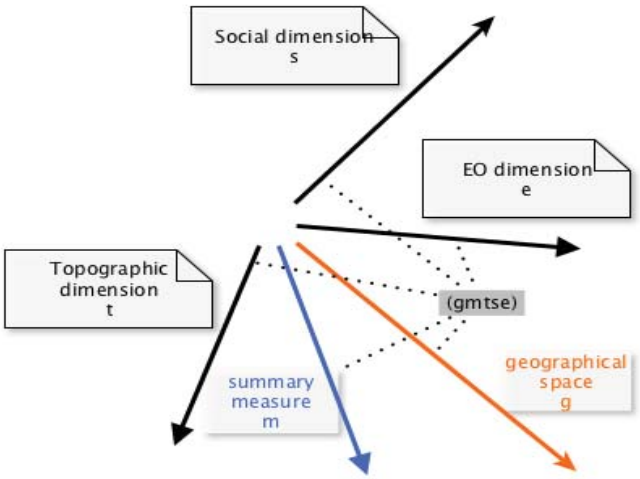

Figure 3: Multivariate sensitivity analysis output stored in a 5-way array, t, s and e indexing each the levels of input uncertainties, g the pixel location and m the distribution indicator of the output uncertainty.

To analyse this sensitivity captured in the 5-way array we can perform multiway correspondence analysis (see Leibovici 2010), allowing an association of the levels of input uncertainty to the intensity of the output uncertainty. The analysis can be performed either on the 3-way table when aggregating across the space and considering only the variance as a summary measure, or on the 4-way table keeping the space dimension. The 5-way analysis can be used to describe further potential associations within the output uncertainty distribution. Before doing so, it is worth looking at the overall variation of the output when aggregating the values over the spatial domain. Figure 4 depicts that overall the workflow is not sensitive to the earth observation

domain as the change in variation in these 4 levels is rather small. Note, also, that the maximum variance observed on the map over the 100 simulations was 4.5 E-4, and the variance over the pixels of the mean of the simulations was 1.12E-04! The patterns of variation between social media uncertainties and topographic uncertainties are very similar in each of the 4 levels of earth observation (MNDWI). Nonetheless, the earth observation uncertainty level 4 shows a larger spread over the topographic levels of uncertainty for social uncertainty level 3, therefore indicating some presence of interaction here. Globally as expected as the uncertainty increases (in social media and topographic data) so does the output variation: linear increasing trend at all earth observation levels. There is not much difference in topographic uncertainty levels except *t1* lower and with an apparent bigger increase in interaction with social dimension. The interaction topographic and social media is only convincing when looking at the difference s1-s4 for t1 versus the others.

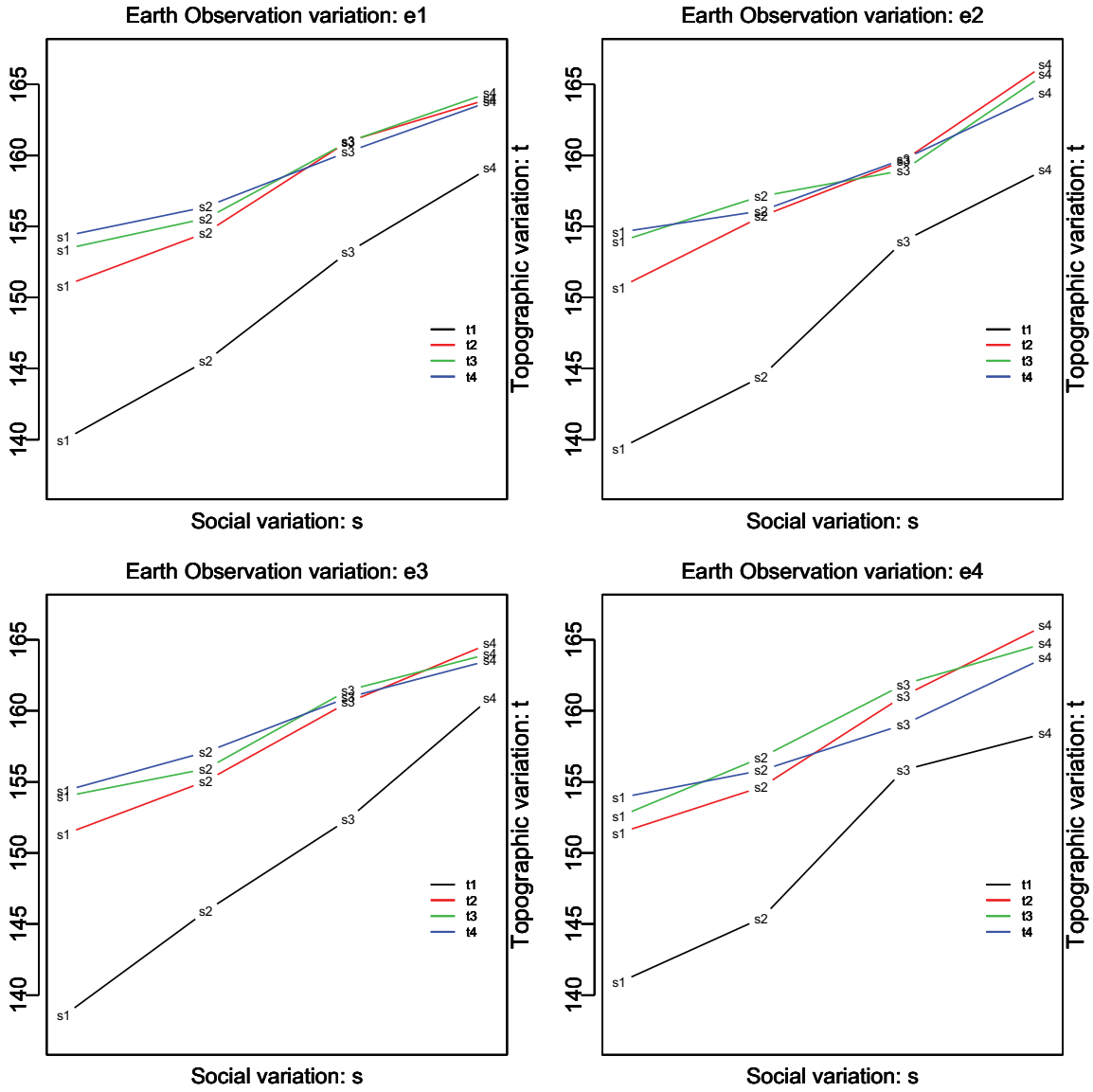

Figure 4: Spatially aggregated output variations (rescaled variance) for each uncertainty combination level (s, t, e).

When performing the multiway correspondence analysis (FCAk), the first principal tensor fitted by the algorithm is the 'independence' principal tensor (vectors of the margins which can

be also automatically set), then the following fitted tensors relate to lack of independence (Leibovici 2010). It is interesting to note that the FCAk of the spatially aggregated data (3-way table) gives 99.98% as representing the independence whilst the analysis on the 4-way table, which includes spatial differentiations, gives 91.95%. So, there is some spatial variation (in the output uncertainty), even though both are high values showing a very small lack of independence between the input uncertainty dimensions (both chi-square values are very small). For the 3-way analysis, the lack of independence has its two-way component breakdowns expressing 2% associated to the social variation margin, 2.7% associated to the marginal topological variation and 84.2% associated to earth observation variation margin, confirming no effect of the latter. Nonetheless, the pure interaction captured in the following (in the order of the decomposition) represents 7% of the lack of independence: *e3* with *t1* and *s4* opposed to *e4* with *t1* and *s1*. The 4% left (100 -(84.2+2+2.7+7)) are concentrated in 3 principal tensors representing nearly 1%, two of these as two-way decompositions associated to the 7% principal tensor. The 4-way analysis puts most of the lack of independence (51%) into a gradient in social media variation levels and spatial differentiation associated with marginal topographic and earth observation variations (Figure 5). The next principal tensor represents 7%, then a few at 1.5%, then a multitude accounting for about 30%. In Figure 5, the blue areas associated to higher social media uncertainty levels (*s3* and *s4*) coincide to the edges of the viewsheds, the brown areas correspond to areas with limited Flickr data.

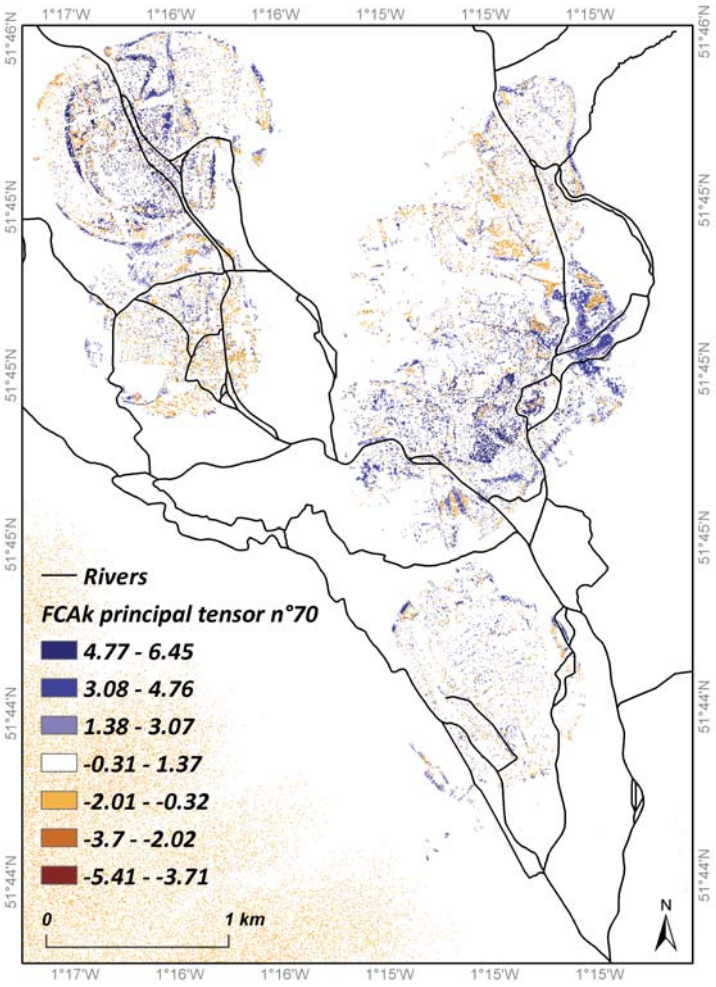

Figure 5: FCAk principal tensor n°70 associated margins to *t* and *e* representing 51% of lack of independence: gradient of *s* is *s1=-1.2*, *s2=-0.7*, *s3=0.4* and *s4=1.4 (*Values are mapped using an equal intervals).

IV DISCUSSION AND CONCLUSIONS

This paper illustrates a multivariate sensitivity analysis for a data fusion algorithm to estimate a flooding extent from a series of predictors. The approach defines each input variable as part of a sampling dimension, with a choice of prior levels of uncertainty per dimension. The multi-way correspondence analyses provide a characterisation of the potential multivariate associations sensitive to the outcome of the data fusion workflow. The example showed a small sensitivity globally and very small influence of the earth observation dimension (EO) without interaction with the other dimensions. If, overall, the increase in input uncertainty implied greater output uncertainty, the topographic dimension showed mostly a difference between its lowest uncertainty level (*t1*) against the other ones and in its interaction with the social media data dimension: bigger increase under t1 as the social dimension uncertainty than the other ones. Spatially, most of the sensitivity relates to the edges of the viewsheds when sampling with high uncertainty levels in the Flickr data, which can be expected.

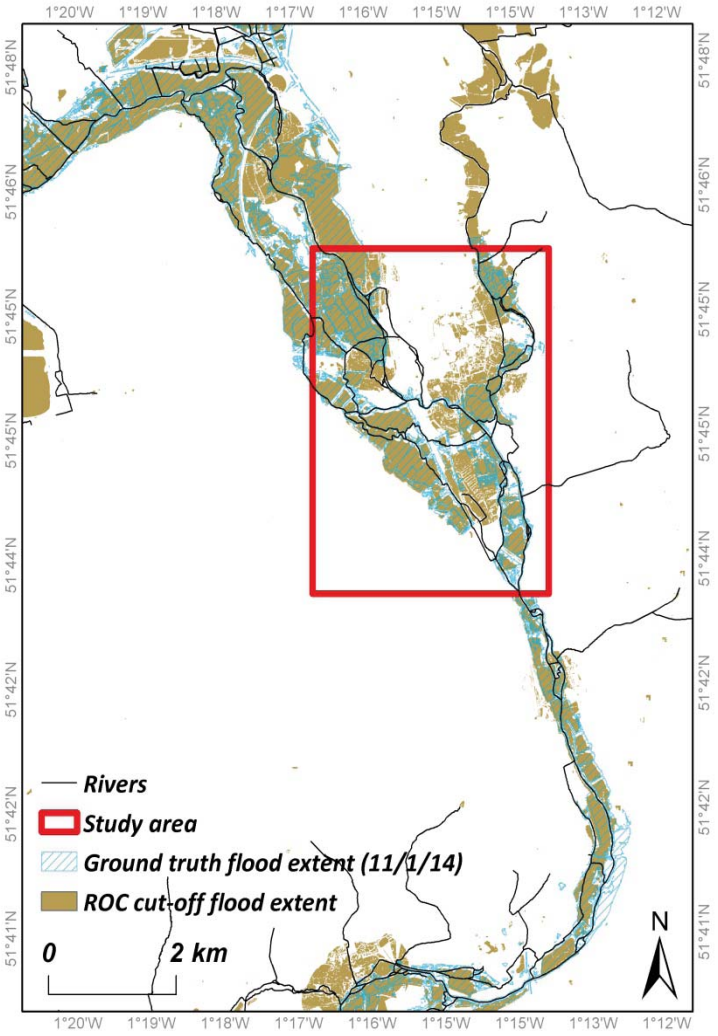

Figure 6: Sensitivity analysis study area, ground-truth and estimated flood extent (derived from Receiver Operating Characteristic).

In this sensitivity analysis, the spatial autocorrelation in the uncertainty sampling has been neglected or reduced simply to the difference in means of the distributions (normal distributions

centred on the observed values). A simple local uncertainty estimation discussed in section II could compensate some of this drawback as the variances could be locally correlated (even though this will still mean sampling at pixel level). Computationally, this would add a potentially expensive step in the initial uncertainty estimation of the inputs but would be done only once. The weight of evidence method producing the posterior probability of flood at a pixel given the predictors comes from the general form:

$$logit(p(O = 1 \mid V_1 V_2 ... V_m)) = logit(p(O = 1)) + \sum_j W^{\pm}(V_j) \quad (2)$$

using the prior and the weights of evidence (see Rosser *et al.* 2016 for more details). Our initial prior was uniform over the area and taken from Rosser *et al.* (2016) in order to allow direct comparison. Note that a Bayesian updating could be performed through deriving a *Beta* distribution from the Monte Carlo simulation (estimating the parameters from the mean and variance of the simulated sample) and using this to generate a new prior in the process. In the last 10 years Bayesian non-parametric methods (*e.g.* Gelfand et al. 2005) have developed which should allow flexibility in modelling spatial and multivariate dependency. However, for large datasets this could be still computationally prohibitive (Berrocal 2016).

References

Berrocal, V. J. (2016). Identifying Trends in the Spatial Errors of a Regional Climate Model via Clustering. *Environmetrics* 27 (2): 90–102.

Chen Y., Yu J. Khan S., (2013). The spatial framework for weight sensitivity analysis in AHP-based multi-criteria decision making, *Environmental Modeling Software*. *48* 129–140.

Gelfand, A.E., Kottas, A., MacEachern, S.N. (2005) Bayesian nonparametric spatial modeling with Dirichlet process mixing. J. Am. Stat. Assoc. 100, 1021–1035

Leibovici D.G., (2010). Spatio-temporal Multiway Decomposition using Principal Tensor Analysis on k-modes: the R package PTAk. *Journal of Statistical Software* 34(10), 1–34.

Panteras G., Wise S., Lu X., Croitoru A., Crooks A., and Stefanidis A., (2015). Triangulating Social Multimedia Content for Event Localization Using Flickr and Twitter. *Transactions in GIS* 19 (5): 694–715. doi:10.1111/tgis.12122.

Rosser J., Leibovici D., Jackson M., (2016). Rapid flood inundation mapping using social media, remote sensing and topographic data. Submitted to: *Natural Hazards.*

Smith, L., Liang, Q., James, P. and Lin, W., (2015). Assessing the utility of social media as a data source for flood risk management using a real-time modelling framework. *Journal of Flood Risk Management, (Article first published online: 27 MAR 2015 DOI: 10.1111/jfr3.12154)*

Tehrany MS, Pradhan B, Jebur MN (2014) Flood susceptibility mapping using a novel ensemble weights-of-evidence and support vector machine models in GIS. *J Hydrol 512:332–343. doi: 10.1016/j.jhydrol.2014.03.008*

Xu H., (2006). Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery. *International Journal of Remote Sensing. 27.* 3025-3033.