SPATIAL BIAS CORRECTION OF SOCIAL MEDIA DATA BY EXPLOITING REMOTE SENSING KNOWLEDGE IN DATA-DEFICIENT REGIONS

Zhenjie Liu[1]*, Jun Li*[1]*, Javier Plaza*[2] *and Antonio Plaza*[2]

[1]Guangdong Provincial Key Laboratory of Urbanization and Geo-simulation,
School of Geography and Planning, Sun Yat-sen University, Guangzhou, 510275, China
[2]Hyperspectral Computing Laboratory, University of Extremadura, E-10003 Caceres, Spain

IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium | 978-1-7281-6374-1/20/$31.00 ©2020 IEEE | DOI: 10.1109/IGARSS39084.2020.9324479

ABSTRACT

Social media data have shown great potential for disaster response. However, the inherent limitations associated to these data (particularly, the spatial bias) restrict its precise application. In this work, we present a new spatial bias correction method based on remote sensing knowledge and spatio-temporal fusion, named locally optimal transport (LOT). Our method is first tested using a case study (2013 Boulder, Colorado flood event). Then, we apply our method to a 2016 Wuhan flood event to test its accuracy in a data deficient region. Our results show that combining remote sensing features and spatio-temporal fusion can help to address problems with a lack of prior data and limited disaster period data. According to the random ground verification points collected from news, pictures and videos, our new LOT method is able to accurately relocate spatially biased social media data to inundated areas, which are dangerous for users.

Index Terms— Social media, remote sensing, flood, cost function, relocation.

1. INTRODUCTION

With the continued warming of the climate, increasing population and rapid urbanization, the frequency and severity of disasters affecting public safety are increasing worldwide [1]. Disasters have become a major factor restricting the sustainable development of global society, economy and environment. Therefore, how to effectively collect real-time data, complete response and evaluation of disastrous events, minimize the socioeconomic losses and ensure public safety is of great significance.

Nowadays, the popularity of social media and the proliferation of its users have led to an increasingly interconnected world. Social media with global positioning (GPS) functionality have become a new source of geo-referenced data, including applications such as Twitter, Instagram, Flickr, Facebook, Weibo, etc. Due to the extensive coverage and real-time nature of social media data, they have become an important source for tracking and managing various types and stages of disasters (mitigation, preparedness, response, and recovery). Meanwhile, social media data are of great value for improving situational awareness of ongoing emergencies [2, 3]. Therefore, the effective application of accurate social media data is crucial for real-time disaster response.

Social media users can spread disaster information at a large scale within a short period when disasters occur, thus enabling anyone to obtain key and timely situational information of disasters. Related studies that leverage social media data for disaster response mainly involve two aspects, which are spatiotemporal mapping and situational awareness analysis. Social media have shown great potential in information dissemination and disaster management. However, as a non-authoritative source, social media data are ambiguous and uncertain due to the lack of metadata and the difficulty in verifying the quality and credibility of massive information [4]. This leads to challenges in the exploitation of social media data for disaster response.

As a key and reliable data source for disaster response and assessment, remote sensing is able to obtain information about disaster conditions at a large scale [5]. However, remote sensing is vulnerable to extreme weather or clouds during disaster periods, and it may not be possible to collect suitable images of areas of interest at the most urgent time due to the limitation of satellite orbit and revisit time. The combination of remote sensing and social media data can help fill the gaps in satellite images, improve the spatiotemporal resolution, and alleviate some of the inherent problems of social media data [6–8]. However, an important problem with geotagged social media is the spatial bias. For instance, the location indicated by the geotagged message is the location of the user but not always the location of the exact inundated place, which introduces severe spatial bias. Most related research has basically used geotagged social media data directly, without considering the aforementioned spatial bias problem.

The optimal transport (OT) method mainly focuses on how to find an optimal transport scheme between the original probability measure and the target probability measure. Wang et al. corrected the spatial bias of geo-referenced tweets using OT based on remote sensing images collected before and after the flood event [9]. This study obtains a high-precision flood density map using relocated geo-referenced tweets. However, the main purpose of OT is to correct geo-referenced social media data to *a priori* regions to represent the entire flood extent, instead of relocating to the actual flood areas originally indicated by the users. Besides, a prior distribution of the disaster may not be available and suitable remote sensing images may also be lacking.

In this work, we introduce a new spatial bias correction method for social media data called locally optimal transport (LOT). Our method exploits remote sensing knowledge to accurately relocate geo-referenced tweets. We first evaluate the accuracy of our newly developed LOT in a case study with *a priori* disaster knowledge and suitable images (2013 Boulder, Colorado flood event). Then, we apply our method to a 2016 Wuhan flood event to figure out how it works in a data-deficient case study.

This work was supported by Hunan Provincial Key Research and Development Program of China under Grant 2019SK2102, National Natural Science Foundation of China under Grant 61771496, National Key Research and Development Program of China under Grant 2017YFB0502900.

2. PROPOSED METHOD

2.1. Data fusion

In our approach, we adopt the spatial and temporal adaptive reflectance fusion model (STARFM) [10] to generate images with 30m spatial resolution during the disaster period in 2016 Wuhan flood event. The Landsat and MODIS images after co-registration and atmosphere correction are imported into STARFM. This spatiotemporal fusion model assumes that changes of reflectance are consistent and comparable at coarse and fine resolutions if pixels in coarse-resolution images are pure. In this case, changes derived from coarse pixels can be directly added to pixels in fine-resolution images to get the predictions. The STARFM method works under the following expression:

$$L2 = G(L1 + M2 - M1), \tag{1}$$

where G represents a weight factor, $M1$ and $L1$ are MODIS and Landsat 8 images (collected in our case on July 23, 2016) and $M2$ is a MODIS 8-day composite surface reflectance product obtained during the flood event.

2.2. Feature extraction

In our method, the inundated areas indicated by remote sensing features are used as the prior flood distribution. The modified normalized difference water index (MNDWI) during the flood event has been proven to be an effective method for flood monitoring [7, 11]. Compared with NDWI (normalized difference water index), MNDWI can better distinguish water bodies from buildings and reduce background noise [12]. We first calculate the MNDWI values of the non-flood image and flood image as follows:

$$\mathrm{MNDWI} = \frac{\mathrm{Green} - \mathrm{MIR}}{\mathrm{Green} + \mathrm{MIR}}, \tag{2}$$

where MIR is mid-infrared band, such as band 6 of Landsat 8. The value ranges of MNDWI are [-1, 1]. In our experiment, the MNDWI threshold for water bodies or floods is set to 0.1. Then, the areas are classified as non-water bodies during non-flood period, and those areas labeled as floods during flood period are used to represent the prior distribution as follows:

$$\mathrm{PD} = \{s|(\mathrm{MNDWI}_{\mathrm{non}}(\mathrm{s}) \leq 0.1)\&(\mathrm{MNDWI}_{\mathrm{flood}}(\mathrm{s}) > 0.1)\} \tag{3}$$

where s represents the locations of pixels in the Landsat image. Since floods are contiguous and often spread to a certain extent, in order to simplify the relocation we select the regions with a continuous area larger than 5 hectares for subsequent analysis.

2.3. Domain adaptation

Before the relocation of original social media data, domain adaption of different data sources needs to be implemented. Let us assume that there are K data modalities from K different sources, represented as maps $\{f_i\}_{i=1,...,K}$:

$$f_i : D_i \to \mathbb{V}_i, \tag{4}$$

where D_i and V_i are the domains and ranges of different data modalities. Then, we homogenize the domains of different data modalities (including social media data, remote sensing images and disaster risk information) into a common information space D:

$$\phi_i : D_i \to D. \tag{5}$$

We consider the 30m Landsat 8 image as the common information space D for domain adaption. Each cell in the image is denoted as an elemental unit $s \in D$. We also use geotagged Weibo messages and ground verification points, geo-referenced to UTM-WGS84 coordinates in ArcGIS using latitude and longitude meta-data. A mapping from points to the nearest cell centers (corresponding to the resolution of Landsat 8 images) is then carried out. In this case, the domains of heterogeneous data sources are all homogenized to the common information space $D := \{s_1, ..., s_n\}$, each $s_i \in \mathbb{R}^2$.

2.4. Locally optimal transport (LOT)

To relocate geo-referenced social media data to the actual flood areas that users intend to indicate (rather than to represent the entire flood extent directly), we introduce a new spatial bias relocation method, called LOT. Compared with OT, LOT does not need to predefine the weight of each location in the prior distribution, which makes LOT more widely applicable. In our context, the cost function of LOT consists of different distances and remote sensing features capable of characterizing water bodies. In addition to the aforementioned MNDWI, we used the well-known normalized difference vegetation index (NDVI), which enhances the contrast between land and water. To characterize floods more accurately, a previous study calculates the difference between NDVI before and during the flood event, namely $\mathrm{DIFF\text{-}NDVI} := \mathrm{NDVI}_{t2} - \mathrm{NDVI}_{t1}$, with the time stamp $t_2 > t_1$. The value ranges of DIFF-NDVI is [-2, 2]. When DIFF-NDVI is closer to -2 or MNDWI is closer to 1, the area is more likely to be inundated.

Here, we mainly use the square l_2 cost function to consider the transport cost of distance and features between s_i and s_j. In Eq. (6), s_i tends to be transported to closer s_j, with similar DIFF-NDVI and MNDWI to s_i:

$$\begin{aligned} m(s_i, s_j) = \alpha^2 \|s_i - s_j\|_2^2 \\ + \|\mathrm{MNDWI}(s_i) - \mathrm{MNDWI}(s_j)\|_2^2 \\ + \|\mathrm{DIFF\text{-}NDVI}(s_i) - \mathrm{DIFF\text{-}NDVI}(s_j)\|_2^2, \end{aligned} \tag{6}$$

where the parameter α can be tuned by observing the distance transport cost in D. In our experiments, α is set to 100 according to [9]. In fact, LOT is a one-to-one or many-to-one transport, and there is no need to predefine the probability mass of each location in different distributions. Specifically, LOT tends to transport s_i to the most appropriate s_j based on the cost function matrix. The relocation using LOT can be expressed as follows:

$$T(s_i) = \underset{s_j \in PD}{\arg\min}\, m(s_i, s_j). \tag{7}$$

2.5. Disaster risk analysis

To estimate the density distributions of the original and relocated social media points, we use the kernel density estimation model as follows:

$$f(s) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{\omega(s - s_i)}{h}), \tag{8}$$

where K is the quartic kernel function, $\{s_1, ..., s_n\}$ are the set of independent social media points, h is the search radius, and ω describes the weight of the points. Locations with higher density usually indicate higher disaster risk. The difference between flood density calculated by relocated points and original points is given by:

$$\mathrm{DIFF\text{-}}f(s) = f_r(s) - f_o(s), \tag{9}$$

Fig. 1. Random ground verification points indicating real flood areas.

where $f_r(s)$ and $f_o(s)$ respectively indicate the flood density calculated by relocated points and original points in location s. If DIFF-$f(s) > 0$, the relocation increases the flood risk of location s. We define such locations as the areas with increased flood risk (AIFR), that is:

$$\text{AIFR} = \{s|\text{DIFF-}f(s) > 0\}. \tag{10}$$

2.6. Accuracy analysis

To evaluate the effectiveness of LOT relocation, the precision is simply given by:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{11}$$

In the 2013 Boulder flood event, TP (true positive) and FP (false positive) denote the number of relocated tweets that fall within (and not within) UFE, respectively. In the 2016 Wuhan flood event, since ground verification points represent the real flooded locations, TP and FP are the number of ground verification points that fall within (and not within) AIFR, respectively.

3. EXPERIMENTAL RESULTS

3.1. 2013 Boulder, Colorado flood event

As for the 2013 Boulder Colorado flood event, flood-related geotagged data were obtained using Web tools [6]. Other sources of data, including Landsat images, historical flood extent and ground-truth are publicly available [9]. In this event, the special flood hazard area (SFHA) can be used as the prior flood distribution, which is the area that will be inundated by the flood event having a 1-percent chance of being equaled or exceeded in any given year. The urban flood extent (UFE) is also publicly available as the ground-truth. Table 1 shows that the precision of original geo-ferenced tweets is only 11.93%. As a result, there is a serious spatial bias in the geo-referenced tweets. The direct use of biased data will no doubt affect specific disaster applications. The precision of geo-referenced tweets relocated by our LOT is increased by 43.75%, which proves the effectiveness of our relocation approach.

3.2. 2016 Wuhan flood event

In July 2016, Wuhan was hit by severe rainstorms that caused serious floods, with 570 mm (about 22.44 in) of rainfall during July 1-7, surpassing the record of the city in 1991. According to the statistics, more than 27 people died and the economic losses reached ¥5.7 billion (about $850 million) in this catastrophic event. Central Wuhan is selected as our study area, where social media messages are widely published. We adopt Sina Weibo open platform API to collect geotagged data with latitude and longitude coordinates from June 29, 2016 to July 10, 2016, located in central Wuhan. A total of 2705 geotagged Weibo messages describing this flood event were acquired (see Fig. 2). Unlike 2013 Boulder flood event, cloudless Landsat images during disaster period, historical flood distribution, and ground-truth were lacking in 2016 Wuhan flood event, which brought additional challenges to the spatial bias correction.

Table 1. Accuracy evaluation of relocated tweets in the 2013 Boulder, Colorado flood event.

Transport	TP	FP	Precision
None	269	1985	11.93%
LOT	1255	999	55.68%

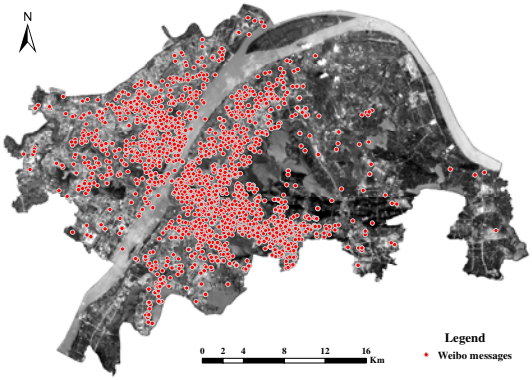

Fig. 2. Distribution of geotagged Weibo messages in the 2016 Wuhan flood event.

Since there are no cloudless 30m Landsat 8 images in central Wuhan during the flood period, we obtain replaceable image by spatiotemporal fusion. The specific images in the considered fusion model include cloudless 500m MODIS 8-day composite images during the flood period, and cloudless Landsat 8 image and MODIS images on July 23, 2016 after the floods. Cloudless Landsat 8 image on June 5, 2016 is used to provide the data before the flood event. And the inundated areas indicated by MNDWI extracted from replaceable 30m image are used as the prior distribution in LOT.

In order to test how our methodology performs in a data deficient region for disaster response, we collected flood-related news reports, pictures and videos as the ground verification points (see Fig. 1). The latitude and longitude coordinates are recorded using Amap. According to the flood features extracted from MNDWI in the flood and non-flood period, it can be seen that the floods have spread extensively in central Wuhan. Floods on both sides of the Yangtze River are particularly severe. And the overall distribution of the original Weibo points is similar to the distribution of flood features. Compared with the original Weibo points, relocated Weibo points are more concentrated, and can better show the extent of floods in local areas, as can be seen in Fig. 3.

Then, we relocated Weibo points into a continuous density map using the kernel density estimation model in Fig. 4. It is easy to find

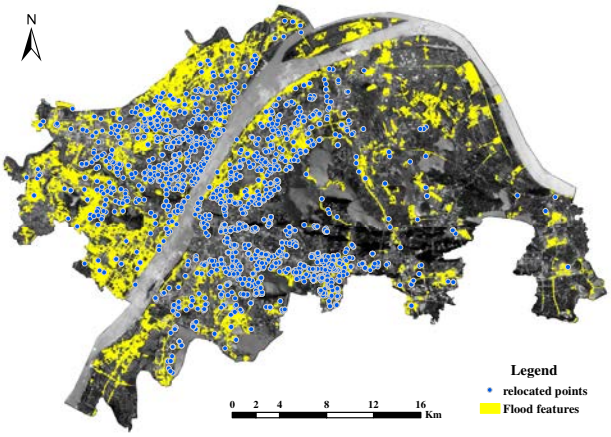

Fig. 3. Distribution of relocated Weibo messages and flood features extracted from MNDWI in the 2016 Wuhan flood event.

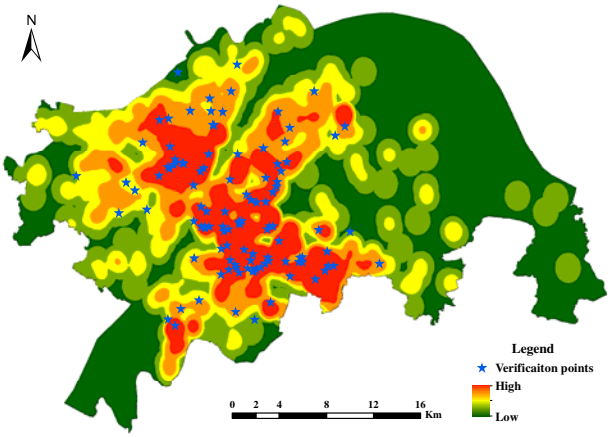

Fig. 4. Flood hazard map of the 2016 Wuhan flood event using relocated Weibo messages by LOT.

that most of the random ground verification points are located in relatively high-risk areas. This result indicates that the flood hazard map generated from relocated Weibo messages is able to reflect real flood risk to a certain degree.

Finally, we evaluate the precision of LOT relocation by recording the number of ground verification points falling inside AIFR. There are a total of 110 verification points, of which 87 are within the AIFR extent and 23 are outside the AIFR extent. Therefore, we conclude that the precision of relocation of our LOT in the 2016 Wuhan flood event reaches 79.08%.

4. CONCLUSIONS AND FUTURE LINES

In this paper, we introduce a new method for spatial correction of social media data by resorting to remote sensing knowledge in data-deficient regions. Our newly proposed method can better transport geo-referenced social media data to the actual disaster locations that the users intend to indicate. At the same time, the proposed method achieves good accuracy in regions with sufficient information (e.g., the considered 2013 Boulder, Colorado case study) and also in data-deficient regions (e.g., the considered 2016 Wuhan case study). In the future, we will apply relocated social media data to different stages of disaster response and further analyze the need to conduct more accurate relocation.

5. REFERENCES

[1] Zheye Wang and Xinyue Ye, "Social media analytics for natural disaster management," *International Journal of Geographical Information Science*, vol. 32, no. 1, pp. 49–72, 2018.

[2] Zahra Ashktorab, Christopher Brown, Manojit Nandi, and Aron Culotta, "Tweedr: Mining twitter to inform disaster response.," in *ISCRAM*, 2014.

[3] J Brian Houston, Joshua Hawthorne, Mildred F Perreault, Eun Hae Park, Marlo Goldstein Hode, Michael R Halliwell, Sarah E Turner McGowen, Rachel Davis, Shivani Vaid, Jonathan A McElderry, and Stanford A Griffith, "Social media and disasters: A functional framework for social media use in disaster planning, response, and research," *Disasters*, vol. 39, no. 1, pp. 1–22, 2015.

[4] Michael F Goodchild and Linna Li, "Assuring the quality of volunteered geographic information," *Spatial statistics*, vol. 1, pp. 110–120, 2012.

[5] Michael F Goodchild and J Alan Glennon, "Crowdsourcing geographic information for disaster response: A research frontier," *International Journal of Digital Earth*, vol. 3, no. 3, pp. 231–241, 2010.

[6] Guido Cervone, Elena Sava, Qunying Huang, Emily Schnebele, Jeff Harrison, and Nigel Waters, "Using Twitter for tasking remote-sensing data collection and damage assessment: 2013 Boulder flood case study," *International Journal of Remote Sensing*, vol. 37, no. 1, pp. 100–124, 2016.

[7] George Panteras and Guido Cervone, "Enhancing the temporal resolution of satellite-based flood extent generation using crowdsourced data for disaster monitoring," *International journal of remote sensing*, vol. 39, no. 5, pp. 1459–1474, 2018.

[8] Xiao Huang, Cuizhen Wang, and Zhenlong Li, "A near real-time flood-mapping approach by integrating social media and post-event satellite imagery," *Annals of GIS*, vol. 24, no. 2, pp. 113–123, 2018.

[9] Han Wang, Erik Skau, Hamid Krim, and Guido Cervone, "Fusing heterogeneous data: A case for remote sensing and social media," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 12, pp. 6956–6968, 2018.

[10] Feng Gao, Jeff Masek, Matt Schwaller, and Forrest Hall, "On the blending of the Landsat and MODIS surface reflectance: Predicting daily Landsat surface reflectance," *IEEE Transactions on Geoscience and Remote sensing*, vol. 44, no. 8, pp. 2207–2218, 2006.

[11] Julian F Rosser, DG Leibovici, and MJ Jackson, "Rapid flood inundation mapping using social media, remote sensing and topographic data," *Natural Hazards*, vol. 87, no. 1, pp. 103–120, 2017.

[12] H Xu, "A study on information extraction of water body with the modified normalized difference water index (MNDWI)," *Journal of Remote Sensing*, vol. 9, no. 5, pp. 589–595, 2005.