This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Fusing Heterogeneous Data: A Case for Remote Sensing and Social Media

Han Wang, Erik Skau, Hamid Krim, and Guido Cervone

***Abstract*— Data heterogeneity can pose a great challenge to process and systematically fuse low-level data from different modalities with no recourse to heuristics and manual adjustments and refinements. In this paper, a new methodology is introduced for the fusion of measured data for detecting and predicting weather-driven natural hazards. The proposed research introduces a robust theoretical and algorithmic framework for the fusion of heterogeneous data in near real time. We establish a flexible information-based fusion framework with a target optimality criterion of choice, which for illustration, is specialized to a maximum entropy principle and a least effort principle for semisupervised learning with noisy labels. We develop a methodology to account for multimodality data and a solution for addressing inherent sensor limitations. In our case study of interest, namely, that of flood density estimation, we further show that by fusing remote sensing and social media data, we can develop well founded and actionable flood maps. This capability is valuable in situations where environmental hazards, such as hurricanes or severe weather, affect very large areas. Relative to the state of the art working with such data, our proposed information-theoretic solution is principled and systematic, while offering a joint exploitation of any set of heterogeneous sensor modalities with minimally assuming priors. This flexibility is coupled with the ability to quantitatively and clearly state the fusion principles with very reasonable computational costs. The proposed method is tested and substantiated with the multimodality data of a 2013 Boulder Colorado flood event.**

***Index Terms*— Least effort principle, maximum entropy models, optimal transport, remote sensing, social media, volunteering labels.**

I. Introduction

FUSION of information from different sensors has long been of interest in Data Science, even if a comprehensive and sound formalism has eluded researchers on account of the complexity of meshing data from different and often incompatible sensor modalities. The readily available sensors and their ubiquitous deployment have further enriched the list of so-called Big Data problems well known to be of central research interest.

As noted earlier, information fusion is of broad interest and, as we further elaborate in the following, cuts across a diverse body of disciplines. In short, interest in fusion often arises in inference problems with a limited number of features obtained from a single sensor, and hence, unable to fully characterize the physical phenomenon at hand. Many types of fusion have been considered proceeding from lower to higher level data processing, e.g., sensor/data level (see [1]), feature level (see [2]), and knowledge/decision level (see [3]).

As a taxonomy of fusion methods, Dasarathy [4] distinguishes five different levels/classes of fusion, including Data In–Data Out, Data In–Feature Out, Feature In–Feature Out, Feature In–Decision Out (FEI-DEO), and Decision In–Decision Out. We note that our proposed work addresses the fusion of heterogeneous data with a goal/outcome of higher level decision-making. While our work appears to follow FEI–DEO, it is closer to Data In–Decision Out (DAI–DEO)-level fusion, and is of more practical interest, given that the features are not readily gleaned from the senor measurements.

Applications in the fusion of remote sensing data are many, and include those of environmental nature such as pollution detection/assessment and weather-related prediction which is of interest here. In density estimation of certain hazard events based on the spatially distributed data, it is common to use a diverse set of sensors, as information from a single sensor is largely insufficient and/or nonrepresentative of the phenomenon of interest. While the current technology may provide a satellite imagery of high spatial and temporal resolution, its quality may fall short on account of a variety of factors such as satellite trajectories (sampling) and atmospheric interference, or simply limited amounts of data. Adverse weather conditions, for instance, on account of their possible duration, could additionally lead to poor quality of the acquired satellite imagery. All such limitations are often mitigated by incorporating additional information sources which, when combined with limited resources on the ground, yield an effective fusion strategy with aerial and remote sensing data.

As a result, social media can become an inference empowerment sensor by volunteering geotagged labels (contextual information) of events of interest. Twitter, for instance, has increasingly become an abundant and valuable source of information—an average of around 6000 tweets tweeted every second according to the twitter statistics, among which any tweeted contents relevant to the subject of interest could be considered for fusion. Due to the notoriously noisy nature

Manuscript received April 23, 2017; revised October 9, 2017, March 5, 2018, and April 23, 2018; accepted May 9, 2018. This work was supported by the Department of Energy National Nuclear Security Administration's Office of Defense Nuclear Nonproliferation R&D through the Consortium for Nonproliferation Enabling Capabilities at North Carolina State University, under Grant DE-NA0002576. *(Corresponding author: Han Wang.)*

H. Wang is with the Department of Computer Science, Institute for Cyber Security, The University of Texas, San Antonio, TX 78249 USA (e-mail: uniwanghan@gmail.com).

E. Skau is with Los Alamos National Laboratories, Los Alamos, NM 87545 USA.

H. Krim is with the Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, NC 27695 USA.

G. Cervone is with the Geoinformatics Department, Pennsylvania State University, University Park, PA 16802 USA.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TGRS.2018.2846199

0196-2892 © 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

of social media, however—particularly twitter—caution should be exercised in handling such information.

Given the prevalence of incomplete, redundant, or incompatible data, fusing heterogeneous sensing modalities is a good motivation for a principled multimodality information fusion formalism, and hence, forms our principal objective in this paper. Our work herein, to the best of our knowledge, represents the first attempt to propose a reasonable and principled formalism for such a goal.

Our proposed approach to formulate the fusion of heterogeneous data is fundamentally based on optimizing a functional with constraints around data from many auxiliary sensing modalities. The objective is a guiding principle or task-motivated fusion criterion, with constraints tying the result to the observed data. The challenge to a successful fusion hinges on securing the integration of heterogeneous and even incompatible data and their homogenization, followed by the optimization to achieve the desired objective. Understanding the essence of heterogeneity underlying the various data modalities is crucial to a successful fusion. We also note that other factors, such as domains for different data with their associated statistics, if unaccounted for, may yield a biased estimation, and hence, reduced fusion performance, as further discussed later in this paper.

Our contribution to investigate a heterogeneous weather-related fusion problem is cast as spatial density estimation using imaging sensors and social media. To that end, and to pursue a fusion-based robust density estimation, we establish a framework based upon two simplified principles: the maximum entropy principle and the least effort principle. While used in some applications [5], [6], the maximum entropy principle has largely been considered in density estimation. As demonstrated later in the sequel, this principle turns out to be a very effective model to work with when a limited number of positive labels and a reasonably large number of features are available. We also show that the performance may be limited in the presence of various noises given the generic (also flexible) statement of the problem and the numerous modalities (e.g., Twitter data and its inherent bias-inducing nature). With such noises and limited data, we consider the least effort principle, where we apply an optimal transport technique [7] to the social media data, which is "homogenized" as labels with an empirical distribution over the fusion space. Specifically, we apply a transport strategy to establish a relationship between our noisy data to priorly known data (e.g., in our case, 100-year history flood zones) to alleviate the bias problem, as further elaborated in the following. The flowchart of the our proposed fusion method specialized to density estimation is shown for quick reference in Fig. 1, and the technical details are deferred to Section III.

Our proposed approach to the satellite imagery and social media-based flood estimation was in part inspired by Schnebele *et al.*'s work [8], and by the associated data heterogeneity. In relation to other works having also exploited social media in [9], our work here provides a more systematic and robust fusion framework.

In comparison to the state of the art in heterogeneous sensor fusion, our contributions in this paper include:

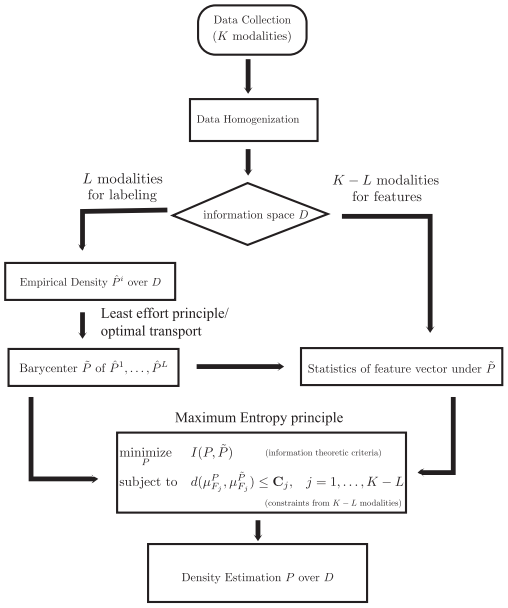


Fig. 1. Diagram illustrates the processing of data through homogenization for a density estimation. Some data modalities are considered for empirical distributions, leading to certain modified empirical distribution with a least effort principle, while other data modalities are considered as constraints in a maximum entropy model.

1) a mathematical framework for fusing heterogeneous data modalities;
2) a successful showcase of our information-based optimization fusion with remote sensing imagery and social media data-based constraints;
3) a least effort principled strategy of relocating geolocational volunteering labels from social media to overcome noisy and spatially biased label issues in the maximum entropy model.

We should also note that the convexity of our formulation makes our solution highly efficient, and lends significant flexibility and effectiveness to many other applications, beyond the flood estimation problem at hand.

The balance of the paper proceeds as follows. Section II discusses much related work. Section III describes our data set for fusion, and introduces our fusion framework, followed by the specific fusion model applied to our data set. In Section IV, we discuss our algorithmic issues, while in Section V, we discuss the experimental results of various optimization instances. Finally we conclude in Section VI with the challenges and work ahead.

II. Related Works

While the fusion literature is vast, it is often focused on a specific application to facilitate the discussion of the fusion problem formulation, and justify the selection of its different attributes.

As mentioned earlier, our investigation herein is motivated by the goal of developing a principled and systematic fusion of heterogeneous sensor data. With an inferential task defined, we will use social media together with other remote sensing data to primarily focus on the study of a natural flooding event in the continental U.S.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

A. Remote Sensing

Multiband and hyperspectral satellite images have actively been exploited for various purposes including environment monitoring and surveillance. For natural hazard extent estimation, remote sensing data analysis has included pixel-based segmentation [10], object-based analysis [11], supervised learning with labeling collection and classification [12], and so on. Good quality labels are, however, usually hard to acquire and often require expert knowledge. The work in [13] addressed noisy labels by proposing a generative model and training a deep neural network based on the existing maps. This unfortunately provided limited success, especially in the presence of misregistration problems due to noisy data (e.g., tweets collected during an associated event).

B. Social Media

In light of their ubiquitous emergence, social media increasingly promise to be of great value even though associated applications have thus far remained simple, and their fusion with other data has been largely *ad hoc*. Early works have been diverse and have included spatiotemporal analysis of tweets to track the progress of a forest fire in France [14], as well as the likelihood of inclusion of geographic information (of flood/wildfire event) by a statistical analysis of tweets among tweeters/retweeterst [15]. In [16], tweets were used to track city activities by police departments, while de Albuquerque *et al.* [17] exploited spatial correlations between the categories of tweets and a flooding event in Germany for tracking purposes.

C. Fusion of Two Heterogeneous Data Modalities

Twitter data were shown in [9] to have the potential of helping manage disaster events by tasking targeted aerial and satellite data collection for assessing infrastructure damage. Specifically, hot spots were localized in a space–time coordinate reference system for transportation issues by keying on road/intersection word occurrences within a 1-km^2 area and in a certain interval of time.

In this paper, we couple the geotagged information provided by tweets with satellite imagery data to perform a spatial density estimation. Our proposed methodology departs from existing work in the sense that our approach is centered around optimizing a predefined criterion functional (to perform a spatial density estimation–flood estimation/prediction) based on one or more data sensor modalities with the remaining data modalities used as constraints. Our goal is also to minimize manual adjustments for tuning the relevance of one modality over another, as is often practiced in *ad hoc* solutions to fusion.

D. Spatial Density Estimation and Optimal Transport

Spatial density estimation has been addressed largely along two approaches: a deterministic track such as Inverse Distance Weighting [18], and a statistical track such as Kriging and Kernel methods [19]. Most if not all existing methods fall short on accommodating noise due to the social media data, the principal difficulty being the inconsistency of associated geolocations inconsistent with the environmental surroundings. Heuristic weightings and bandwidth selections have for the most part been used.

In species distribution models, the spatial sampling bias issue also arose [20], and a solution to debiasing the data in sampling was proposed in [21], which primarily relied on (unrealistic) estimates based on unbiased confidence intervals.

As further discussed in the sequel, we propose a principled first attempt to accommodate a fusion of a heterogeneous set of hybrid sensors, including social media with their notorious noisy nature. We address the sampling bias of the geolocations of the labeled events from Twitter, and propose our debiasing solution using optimal transport, a technique well known in machine learning and data science for transfer learning [22]. Intuitively, this in effect transports a probability density function of a source domain to accommodate the target domain. The local nonlinear mapping property of the optimal transport is recognized for its effectively addressing transfer learning [7], together with its natural interpretation of distance of distributions in the so-called Wasserstein metric space. We demonstrate its application as a novel debiasing tool, and hence, mitigating the inherent limitation of social media. We invoke prior knowledge (in our case, historically recurrent flooded regions) to define a distribution in the target domain, and use the transported geolocations of labels in the maximum entropy model for near real time data fusion. The merits of our method include a minimal number of hyperparameter adjustments for the noisy data, and a geometrically interpretable relocation result.

III. Problem Statement and Formulation

Our primary objective is to propose a principled optimized framework to homogenize the heterogeneous large-scale data and provide a quantitative strategy to estimate/predict a scenario, which in this case, pertains to a flood over a geographical region.

Our specific hazard extent estimation objective is to estimate a flooded area as a field distribution using maximum entropy as the guiding principle. The maximum entropy principle is a well-known statistical model dating back to the 1950s [23]. The rationale of the maximum entropy is to seek the most unassuming, and hence, maximally random model.

A. Overview of the Weather-Related Data: Boulder Colorado

The proposed method was tested on data related to the 2013 Boulder Colorado Flood. This event was declared as a natural disaster starting in September 2013. From September 9 to 15, some places in Boulder County saw up to 17 in of rainfall, which was comparable to the annual average of approximately 20 in. The flood caused at least eight deaths, with several people missing or unaccounted for, and the evacuation of over 180 000 people. Financially, the flood was estimated to have caused two billion in damage.

Satellite imagery is one of the principal tools to gather data about events of interest, including floods. Landsat 8 is an American satellite, operated by National Aeronautics and Space Administration and the United States Geological

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE I

SPECTRAL INFORMATION FOR THE OLI SENSOR†

Spectral Band	Wavelength	Resolution	Solar Irrandiace
Band 1 - Coastal/Aerosol	$0.433 - 0.453\mu m$	$30m$	$2031W/(m^2\mu m)$
Band 2 - Blue	$0.450 - 0.515\mu m$	$30m$	$1925W/(m^2\mu m)$
Band 3 - Green	$0.525 - 0.600\mu m$	$30m$	$1826W/(m^2\mu m)$
Band 4 - Red	$0.630 - 0.680\mu m$	$30m$	$1574W/(m^2\mu m$
Band 5 - NI	$0.845 - 0.885\mu m$	$30m$	$955W/(m^2\mu m)$
Band 6 - SWI	$1.560 - 1.660\mu m$	$30m$	$242W/(m^2\mu m)$
Band 7 - SWI	$2.100 - 2.300\mu m$	$30m$	$82.5W/(m^2\mu m)$
Band 8 - Panchromatic	$0.500 - 0.680\mu m$	$15m$	$1739W/(m^2\mu m)$
Band 9 - Cirrus	$1.360 - 1.390\mu m$	$30m$	$361W/(m^2\mu m)$

† https://en.wikipedia.org/wiki/Landsat_8

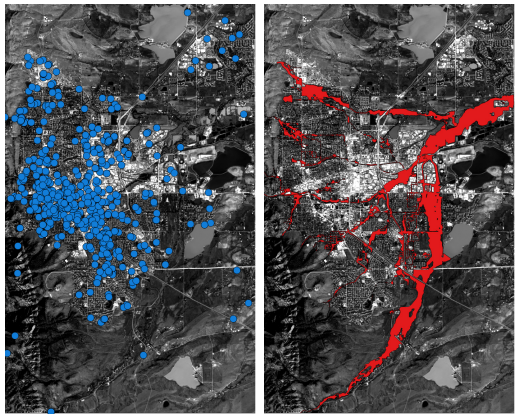


Fig. 2. Flood data set. (Left) Panchromatic image of the city of Boulder, CO, USA, taken on September 17, 2013 from Landsat 8 satellite image, overlaid with geolocated tweets. (Right) UFE as our ground truth.

Survey. The satellite carries two sensors, an Operational Land Imager (OLI) and a Thermal InfraRed Sensor. The OLI is a multispectral sensor that collects images in nine shortwave bands. Table I shows the spectral information for the OLI sensor. Shortly after the flood, the Landsat 8 satellite passed over Boulder Colorado capturing multispectral images of the flood aftermath. There is also a photograph of this region prior to the flood on August 25, 2013. The images from this satellite are publicly available. Fig. 2 shows a panchromatic image of the City of Boulder taken by Landsat on September 17, 2013.

Novel information streams, such as social media contributed videos, photographs, and text as well as other open sources, as noted earlier were also available as a means of improving situational awareness during emergencies. As contributed data with spatial and temporal information, they provide valuable volunteered geographical information, harnessing the power of "citizens as sensors" to provide a multitude of on-the-ground data, often in real time [9], [24].

Throughout the flood event, social media data were collected from Twitter. Often times, users use hashtags to somewhat contextualize their messages. During the flood, several hashtags emerged to indicate a connection to the 2013 Boulder Flood, such as "#LongmontFlood," "#boulderflood," and "#coflood." In addition, depending on the users' preferences, a small fraction $(1\% - 2\%)$ of the tweets were geotagged with the metadata of the latitude and longitude of the phone at the time of posting. We can visualize the distribution of people tweeting about the Boulder flood in a region containing Boulder city and some of its surroundings (see Fig. 2). Cervone *et al.* [9] further considered filtering the tweets content to extract geolocation information, increasing the percentage of geolocated tweets from 2% to 10%. In our experiment, we only consider the original geolocated ones for validation of our method.

Special flood hazard areas (SFHAs) are flood hazard areas identified on Flood Insurance Rate Maps (FIRMs). SFHA regions have a probability greater than 1% of being flooded each year. These areas can be expected to flood at least once every 100 years. SFHAs consist of several different types of zones, some of them are: Zones A, AE, AH, and AO. These different zones correspond to different methods of calculation, or different expected types of flooding events. For instance, Zone A regions are generally determined using approximate methodologies, while Zone AE regions are determined with detailed methodologies. AH Zones correspond to areas expected to undergo shallow pooling with depths less than 3 ft, while AO Zones correspond to the regions expected to undergo flood sheets on an incline with depths less than 3 ft. The FIRMs for Boulder Colorado as defined prior to the flood are also publicly available.

In the weeks and months following the flood, the city of Boulder created a map of inundated areas to better understand the flood as well as to potentially revise the SFHAs. With the help of hand held GPS devices, workers geotagged high water locations, and carefully produced an accurate account of the inundated areas. With additional information from community provided photograph evidence, the extent of the flooding in some regions was also obtained. This information was conglomerated to construct an urban flood extent (UFE) map. The UFE was graciously provided to us by the city of Boulder, to help us have a semblance of ground truth as a reasonable approximation of the inundated areas.

Due to imperfection of the available data, we focus on using the available data to prove our concept of fusion. We present in Section III-B our proposed mathematical formulation of the fusion problem, where tweets, Landsat 8 data, and SFHAs (as historic priors) mentioned earlier are selected exclusively for the evaluation and validation in our experiments.

In light of the significant incompatibility of these two modalities, it is clear that centering our validation around them meets our goal of addressing heterogeneous sensor information fusion.

B. Formulation of the Fusion Framework

1) Setup: We consider a probability space $(D, \Sigma, \mathcal{P}(D))$, where D is our information fusion space, Σ is the σ-algebra on D and each $s \in D$ is a variable which denotes an elemental unit in D where all information can be treated as functions $f(s)$ associated with s. $\mathcal{P}(D)$ denotes the space of probability measures over Σ. Particularly in spatial density estimation problem, D is a measurable subset of $\mathbb{R}^2$. Our goal is to find a distribution $P \in \mathcal{P}(D)$ as our density estimation optimizing some information criterion and satisfying constraints given by $f(s)$.

Moreover, suppose we have K data modalities (from K different sources) represented as maps $\{f_i\}_{i=1,\ldots,K}$ with their own domains and ranges

$$f_i : D_i \to \mathbb{V}_i$$

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

where the measurable D_i may be different from D, and $\mathbb{V}_i$ can be values taken in the field of real numbers $\mathbb{R}$, integers $\mathbb{Z}$, complex values $\mathbb{C}$, and so on. In our flood estimation, the multiband satellite images yield different maps f_i's revealing the floods spectral information.

In particular, among all different f_i's, a special map taking *categorical* values is called *labeling*, that is,

$$f_r : D_r \to \mathbb{L}$$

where D_r is the domain for the labeling f_r, and $\mathbb{L}$ is a categorical set of events. For simplicity, we restrict ourselves to single event labeling, i.e., we consider $f_r = \mathbb{1}_{D_r}$, the indicator function defined on D_r. In our flood example, all social media data can yield labeling information. Regardless of their heterogeneity, social media data may always be transformed into appropriate labels (e.g., to express flood) after certain data preprocessing steps, be it semantic analysis with tweets, or image classification with flickr photographs, and so on, leading to high-level semantic information in our fusion framework. In principle, we can assume L labeling maps among the K modalities, it is also necessary to fuse the L labeling information from different sources.

Since different data modalities are associated with different domains D_i's, we first proceed to homogenize the data into a common domain D for sensible and coherent fusion (information space), namely,

$$\phi_i : D_i \to D$$

where the measurable map ϕ_i could be nonlinear and ϕ_i^{-1} is a set-valued map (ϕ_i being many to one). This formalism allows the alignment of different domains to a common domain for fusion.

Remark 1: In our flood case study, the homogenization can be naturally addressed for Landsat 8 data and geolocated tweets. Since the multispectral remote sensing data are associated with 30×30 m^2 grid cells for most bands, one can consider those cells as elements in D, and other heterogeneous data such as geolocations of tweets can be associated with those cells in a Voronoi diagram fashion, that is, ϕ_i as a mapping from locations of the social media labeling to their nearest centers of cells based on the satellite data resolution. In general, finding ϕ_i is nontrivial and case dependent.

Next, a *fusion* operator $\mathcal{H}$ is in place for combining all f_i's through ϕ_i's to a common fusion space $\mathbb{V}$; a trivial choice for $\mathbb{V}$ could be $\prod_i V_i$, i.e., the product space of various data ranges. Following such a homogenization procedure (illustrated in Fig. 3), one would naturally expect a nice representation for the entity of interest in the fusion space through the mapping

$$\mathbf{x}(s) := \mathcal{H}(f_1(\phi_1^{-1}(s)), \ldots, f_K(\phi_K^{-1}(s))).$$

In our case study for the flood density estimation, $\mathcal{H}(f_1(\phi_1^{-1}(s)), \ldots, f_K(\phi_K^{-1}(s)))$ becomes a product of real-valued and binary vectors, namely, $\mathbf{x}(s) \in \mathbb{R}^{K-L} \times \{0, 1\}^L$, for any $s \in D$. The homogeneous map $\mathbf{x} : D \to \mathbb{V}$ is therefore our desired representation for fusion. We denote the real-valued components of $\mathbf{x}$ to be $[F_1, \ldots, F_{K-L}]$.

Finally, we focus only on labeling on finite domains, so we define the *empirical distribution* $\hat{P}_{f_r}(S_r) \in \mathcal{P}(D)$ according to a labeling f_r as a discrete measure with support on $S_r = \{s | f_r(\phi_r^{-1}(s)) = 1\} \subset D$

$$\hat{P}_{f_r}(S_r) := \{\mu | \mu = \sum_{s_i \in S_r} a_i \delta_{s_i}, a_i = |\phi_r^{-1}(s_i)|/|\phi_r^{-1}(S_r)|\} \tag{1}$$

where δ_{s_i} is the Dirac unit mass on s_i.

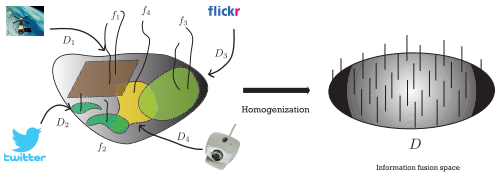


Fig. 3. With different data modalities, a data fusion scheme is to homogenize the data into an information fusion space D, where all modalities are represented in a unified way.

With the general formulation and its specialization to our case study (fusion representation vector $\mathbf{x}$) in hand, we proceed with the task of estimating a distribution according to an information criterion of choice.

2) Optimization Fusion Model: Suppose P is the underlying density of the labeled event in D of interest. Using the most unassuming probability model for P, the well-known maximal entropy model [21] (also adopted as the model of choice in this paper), implicitly defines it as the result of maximizing the *entropy* $-P \ln(P)$, with constraints given as

$$|\mathbb{E}_P F_j - \mathbb{E}_{\hat{P}_{fr}} F_j| \leq \mathbf{C}_j \tag{2}$$

for $j = 1, \ldots, K - L$. $\mathbb{E}_P$ and $\mathbb{E}_{\hat{P}_{fr}}$ in (2) denote the expectation operator (of F_j) under distribution P and $\hat{P}_{f_r}$, respectively. $\mathbf{C}_j$ is a tolerance parameter, depending on the confidence of the representative strength of F_j. Note that the intuition behind the constraints is that the first statistical moments of features under the prevailing empirical distribution should be close to those under the true distribution; higher moment constraints are also natural to consider, depending on one's desired degree of approximation of the distribution. More generally, we may view the density estimation problem as a fusion model optimizing some information criteria $I(P)$ with respect to P with the following constraints:

$$d\left(\mu_{F_j}^{P}, \mu_{F_j}^{\hat{P}}\right) \leq \mathbf{C}_j \tag{3}$$

where $\mu_{F_j}^{P}$ and $\mu_{F_j}^{\hat{P}}$ are the distributions of F_j under P and $\hat{P}_{f_r}$, respectively, and d is a metric quantifying the distance between the two distributions.

To proceed with the framework for our optimized fusion model, and to account for potentially numerous social sensors (e.g., in addition to Twitter, Flickr imagery) as well as bias correction, we assume having L labelings $f_r^1, \ldots, f_r^L$ out of K modalities ($L \leq K$) with the associated empirical distributions denoted by $\hat{P}^1, \ldots, \hat{P}^L$. Our proposed fusion model is then formulated as a two-stage optimization to first compute the *best overall prior distribution*[1] $\tilde{P}$, then

[1] $\tilde{P}$ is the *barycenter* of $\hat{P}^1, \ldots, \hat{P}^L$ as an optimizer of $\sum_{i=1}^{L} \alpha_i d_1(\tilde{P}, \hat{P}^i)$, where $\alpha_i \geq 0$ and $\sum_{i=1}^{L} \alpha_i = 1$.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

optimize some information criteria with respect to $\tilde{P}$ and the unknown P.

Optimized fusion model

Step 1:

$$\tilde{P} = \underset{P \in \mathcal{P}(D)}{\operatorname{argmin}} \sum_{i=1}^{L} \alpha_i d_1(P, \hat{P}^i) \tag{4}$$

Step 2:

$$\begin{aligned} &\underset{P}{\text{minimize}} \quad I(P, \tilde{P}) \\ &\text{subject to} \quad d_2(\mu^P_{F_j}, \mu^{\tilde{P}}_{F_j}) \leq \mathbf{C}_j, \quad j = 1, \ldots, K - L \end{aligned} \tag{5}$$

In the above-mentioned formulation, d_1 and d_2 are metrics for distributions (and not necessarily the same). $I(P, \tilde{P})$ is the information criterion for P and $\tilde{P}$. Note that our formulation is principled and very flexible for extension. There are relevant work (see [25]) considering the fusion problem as an optimization for a barycenter of a collection of measures in a space, say Wasserstein metric space. However, our framework is more general and considers features associated with different measures; hence, the fusion could be carried out in different spaces of probability measures to incorporate multimodality data.

Remark 2: When we consider d_1 as the Wasserstein distance (we omit the definition here and refer our readers to [26] for details) in step 1 of our fusion model, $\tilde{P}$ is then called the Wasserstein barycenter. Moreover, even when $L = 1$, the optimization (4) in step 1 of our optimized fusion model can be used for bias correction for some empirical distributions from a labeling map, using optimal transport. This is closely related to our least effort principle, which we shall elaborate in the following.

3) Optimal $\tilde{P}$ for Bias Correction: As noted earlier, the labeling of social media can be very noisy and biasedly sampled. More precisely, the empirical distribution $\hat{P}_{f_r}$ generated from the labeling f_r may not be able to represent P, i.e., $d(\mu^P_{F_j}, \mu^{\hat{P}}_{F_j}) > \delta$ for some δ, contradicting the constraints (3) when $\mathbf{C}_j$ is sufficiently small. We refer to this issue as a *domain shift*, which is due to many factors. For example, in our flood case study, people could only tweet at a distance from the true flooded location. It thus makes sense to try to mitigate this inherent bias by applying a *transport* of the labeling f_r's domain D_r to $\tilde{D}_r$, where $\tilde{D}_r$ is the set of relocations which are more likely to be true flooded locations. Finding such a transport is challenging due to the non-Gaussian inherent noise in the social media data. To overcome this difficulty, we propose another principle in this paper: a *least effort principle*. The *optimal transport* in domain adaptation [7] embodies this principle, and helps us alleviate the shift issue between $\hat{P}_{f_r}$ and the desired P.

More concretely, suppose $S^{\text{prior}}(\mathbf{x})$ represents our prior criteria on feature selections, the set of relocations/likely true flooded locations $\tilde{D}_r = \{s \in D : S^{\text{prior}}(\mathbf{x}(s)) \in \mathcal{T}\}$ is characterized by a property $\mathcal{T}$ as a desirable property which

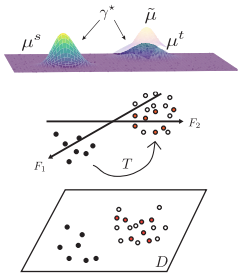


Fig. 4. Illustration of the optimal transport in the $[F_1, F_2]$ domain, leading to a transport in D.

$S^{\text{prior}}(\mathbf{x}(s))$ must satisfy. For example, in the flood case study, we may consider Modified Normalized Difference Water Index (MNDWI) as a prior of choice, and its value above some threshold could be used as a criterion for selecting the most likely flood location candidates.

Suppose we have an empirical distribution $\hat{P}(S) = \sum_{s_i \in S} a_i \delta_{s_i}$ with support on $S \subset D$, we would like to find an optimal transport $T^\star$ formulated as a minimizer of the following cost function:

$$M(T) = \sum_{s_i \in S} a_i m(s_i, T(s_i)) \tag{6}$$

where m is the cost, say Euclidean distance to move $s \in S$ to $T(s) \in \tilde{D}_r$.

Remark 3: The optimal transport $T^\star$ assigns a_i to the closest neighbor of s_i in $\tilde{D}_r$ according to the cost m; hence, it naturally results in $\tilde{P} = \sum_{s_i \in S} a_i \delta_{T^\star(s_i)}$ as a minimizer of $d_1(P, \hat{P})$ over all distributions with support in $\tilde{D}_r$, where d_1 is the Wasserstein distance. This bias correction is, hence, a special case of step 1 in our fusion model, with $L = 1$.

C. Computationally Feasible Solution

In the above-mentioned scheme for bias correction, T is, however, computationally expensive and sensitive to $\tilde{D}_r$ which depends much on an expert-based parameter tuning.

In order to minimize the parameter tuning for "unbiased" relocations, we instead propose to introduce a target distribution $\mu^{\hat{P}_t}_{S^{\text{prior}}(\mathbf{x})}$ of $S^{\text{prior}}(\mathbf{x})$ under $\hat{P}_t \in \mathcal{P}(D)$ as a proxy for prior knowledge. To transport the original distribution $\mu^{\hat{P}_s}_{S^{\text{prior}}(\mathbf{x})}$ of $S^{\text{prior}}(\mathbf{x})$ according to an empirical distribution $\hat{P}_s$ to $\mu^{\hat{P}_t}_{S^{\text{prior}}(\mathbf{x})}$, a computationally more advantageous method is to consider the *Kantorovitch* relaxed formulation [26] in the optimal transport. The idea is to find a probabilistic coupling $\gamma^\star \in \Pi(\mu^{\hat{P}_s}_{S^{\text{prior}}(\mathbf{x})} \times \mu^{\hat{P}_t}_{S^{\text{prior}}(\mathbf{x})})$ of which the marginals are $\mu^{\hat{P}_s}_{S^{\text{prior}}(\mathbf{x})}$ and $\mu^{\hat{P}_t}_{S^{\text{prior}}(\mathbf{x})}$ instead of the explicit transport T, and $\gamma^\star$ satisfies the following:

$$\gamma^\star = \underset{\gamma \in \Pi(\mu^{\hat{P}_s}_{S^{\text{prior}}(\mathbf{x})} \times \mu^{\hat{P}_t}_{S^{\text{prior}}(\mathbf{x})})}{\operatorname{argmin}} \int_{D \times D} m(s_i, s_j) d\gamma(s_i, s_j). \tag{7}$$

In general, we consider a transport in a sample space Ω. Fig. 4 illustrates the scheme in a discrete setting: to achieve an improved distribution on D and correct an intrinsic bias, a transport of the black colored dots $\{s^s_i\}_{i=1}^{n_s}$ to the white colored ones $\{s^t_j\}_{j=1}^{n_t}$ is accomplished in $\Omega \ni \omega(s) =$

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

$[F_1(s), F_2(s)]$ (F_i's are functions defined on D), where black dots are represented by $\{\omega_i^s\}_{i=1}^{n_s}$, while white ones by $\{\omega_j^t\}_{j=1}^{n_t}$. Their corresponding distributions are $\mu^s = \sum(1/n_s)\delta_{\omega_i^s}$ and $\mu^t = \sum(1/n_t)\delta_{\omega_i^t}$ in the domain Ω, and μ^s is transported to μ^t by an optimal coupling $\gamma^\star$. $\gamma^\star(\omega_i^s, \omega_j^t)$ represents the probability mass of ω_i^s being transported to ω_j^t: this yields a new distribution $\tilde{\mu}$ for the transported μ^s. When a squared (l_2) distance is considered for the cost function, the transport map T can be deduced from $\gamma^\star$, that is, a barycentric mapping $T(\omega_i^s) = \sum_j n_s \gamma^\star(i, j)\omega_j^t$ [7, eq.(15)], yielding a new distribution $\tilde{\mu}$ based on the transported black dots as red ones. One can naturally derive the transport for black dots in D, that is, $T(s_i^s) = \sum_j n_s \gamma^\star(i, j)s_j^t$, visualized as relocated red dots in D.

Remark 4: In our flood estimation problem, as we further discuss in Section IV, we derive a modified empirical distribution $\tilde{P}^{\gamma^\star}$ based on the transport of source data (locations of the original tweets) to the target data (locations of historic flood) over D.

Upon lifting the intrinsic limitations of the collected data, we proceed to formulate the following optimization problem to effectively address our flood estimation case study, the computational details are deferred to Section IV.

$$\begin{aligned} &\underset{P}{\text{minimize}} \quad P \ln(P) \\ &\text{subject to} \quad \mathbb{E}_P F_j - \mathbb{E}_{\tilde{P}^{\gamma^\star}} F_j | \leq \mathbf{C}_j. \end{aligned} \tag{8}$$

IV. Algorithmic Fusion Solution

We develop our algorithmic solution to the data-based fusion problem using the formalism developed earlier. We focus on solving the problem in (8). We use a two-step procedure: mitigate for the bias by an optimal transport followed by the solution for the maximum entropy model. Note that our fusion model now applies to a finite domain $D = \{s_1, \ldots, s_n\}$, each $s_i \in \mathbb{R}^2$. The desired distribution $P \in \mathcal{P}(D)$ and the empirical distribution $\hat{P} \in \mathcal{P}(D)$ are written as

$$P = \sum_{s_i \in D} a_i \delta_{s_i}, \quad \hat{P} = \sum_{s_i \in D} \hat{a}_i \delta_{s_i}. \tag{9}$$

Here, δ_{s_i} is the Dirac function at location s_i, and a_i's and $\hat{a}_i$'s are the coefficients in the probability simplex

$$\Sigma_n = \left\{ a \in \mathbb{R}^n | a_i \geq 0, \sum_i a_i = 1 \right\}. \tag{10}$$

In our Boulder flood case study, we are to estimate the flood density P over D. We first find $\tilde{P}$ based on $\hat{P}$ in order to remove the bias in $\hat{P}$, and derive it from geolocations of tweets.

A. Correcting Distribution Shift: Label Relocation and Optimal Transport

The labelings of flood with geolocations from social media suffer from a myriad of noise sources. The constraints in the maximum entropy model (step 2 in our fusion framework) are therefore vulnerable to the labeling quality. To correct the shift between the empirical and the true density distributions, and to also induce robustness to sampling locational noise, we first make locational corrections with some expert knowledge.

It has been empirically established that *normalized difference indices* are appropriate and useable spectral signatures. Flood/wetness has been shown to be best estimated by the so-called NDWI [27], with an improved modified alternative MNDWI [28]. The calculation for the MNDWI during or after the flood relies on two Landsat bands, the green band, and the short wavelength infrared band. Another feature that is insightful for flood detection is the *Normalized Difference Vegetation Index* (refer to [29] and references therein). More precisely, the difference of NDVI before and after the flood is a meaningful signature for flood regions, and defines a feature DIFF-NDVI := $\text{NDVI}_{t_2} - \text{NDVI}_{t_1}$, with the time stamp $t_2 > t_1$. A simple way to solve (6) with m being the Euclidean distance between s and $T(s)$ is to find the nearest neighbor for s in the candidate set satisfying thresholding values. More specifically, we consider expert insight for flood features, namely,

$$S^{\text{prior}}(\mathbf{x}) = (\text{MNDWI}, \text{DIFF-NDVI}) \tag{11}$$

with a candidate set being

$$\{s | \text{MNDWI}(s) > \theta_1\} \cup \{s | \text{DIFF-NDVI}(s) < \theta_2\}. \tag{12}$$

This approach is, however, computationally heavy and tends to make our estimation unstable. In later experiments, we use this simple relocation scheme for comparison with other methods.

For proof of concept, in our case study, we consider only two empirical distributions after the *homogenization*, namely, one $\hat{P}_s$ coming from the locations of tweets, the other $\hat{P}_t$ coming from historic flood regions

$$\hat{P}_s = \sum_{s_i \in D} a_i^s \delta_{s_i}, \quad \hat{P}_t = \sum_{s_i \in D} a_i^t \delta_{s_i}. \tag{13}$$

Now, we consider Ω as the sample space for $\omega(s) = [\beta s, S^{\text{prior}}(\mathbf{x})(s)]$ (β is a constant weighting coefficient), two distributions of ω according to $\hat{P}_s$ and $\hat{P}_t$ in (13) are given as

$$\mu_\omega^{\hat{P}_s} = \sum_{i=1}^{n_s} a_i^s \delta_{\omega(s_i^s)}, \quad \mu_\omega^{\hat{P}_t} = \sum_{i=1}^{n_t} a_i^t \delta_{\omega(s_i^t)}$$

where s_i^s ranges over all locations with $a_i^s > 0$; s_i^t ranges over all locations with $a_i^t > 0$. The nonzero vector a^s belongs to the probability simplex Σ_{n_s}, and a^t belongs to the probability simplex Σ_{n_t}.

We consider the relaxed formulation of optimal transport by computing the optimal coupling $\gamma^\star$

$$\gamma^\star := \underset{\gamma \in \mathcal{B}}{\arg\min} < \gamma, M >_F \tag{14}$$

where $\mathcal{B}$ represents the set of associated discrete couplings between $\mu_\omega^{\hat{P}_s}$ and $\mu_\omega^{\hat{P}_t}$, that is,

$$\mathcal{B} = \{\gamma \in (\mathbb{R}^+)^{n_s \times n_t} | \gamma \mathbf{1}_{n_t} = a^s, \quad \gamma^T \mathbf{1}_{n_s} = a^t\}$$

and $M = (m_{i,j})$ is the cost matrix where $m_{i,j}$ denotes the cost of moving from s_i^s to s_j^t in Ω.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

In our experiment, we consider the following square l_2 cost function in Ω to take into consideration both the geotransport cost and prior feature transport cost when the transport is carried out in Ω

$$\begin{aligned} m_{i,j} := & \|\text{MNDWI}(s_i) - \text{MNDWI}(s_j)\|_2^2 \\ & + \|\text{DIFF-NDVI}(s_i) - \text{DIFF-NDVI}(s_j)\|_2^2 \\ & + \beta^2 \|s_i - s_j\|_2^2. \end{aligned} \tag{15}$$

We tune the parameter β by observing the geotransport cost in D.

The optimization in (14) is a linear programming over polygonal coupling constraints. Note that other regularization methods are also possible for a potentially more efficient computation of the optimal coupling [7], [30].

One can interpret the optimal transport $\gamma^\star(i, j)$ as how much probability mass of s_i^s is transported to s_j^t in Ω. When $p_i^s = (1/n_s)$ and $p_j^t = (1/n_t)$ (uniform distributed over the data points), the optimal relocation can be expressed as the following *barycentric mapping* based on the optimal transport $\gamma^\star(i, j)$ [7]

$$T(\omega(s_i^s)) = \sum_j n_s \gamma^\star(i, j) \omega(s_j^t) \tag{16}$$

leading to the transport in D

$$T(s_i^s) = \sum_j n_s \gamma^\star(i, j) s_j^t. \tag{17}$$

This naturally yields a modified empirical distribution $\tilde{P}^{\gamma^\star}$ on D

$$\tilde{P}^{\gamma^\star} = \sum_{i=1}^{n_s} \frac{1}{n_s} \delta_{T(s_i^s)}. \tag{18}$$

Theorem 1: The transport scheme in Ω with the given cost function for bias correction yielding $\tilde{P}^{\gamma^\star}$ is an approximation to a special case of solving (4) in step 1 of our fusion framework.

Proof: The optimal transport carried out in Ω with a squared l_2 cost is an optimal relocation in Ω which optimizes $< \gamma, M >_F$. The derived $\tilde{P}^{\gamma^\star}$ can be seen as an approximation to

$$\tilde{P} = \arg\min_P W(P, \hat{P}_t)$$

which is a special case of (4) with $L = 1$, $d_1(\cdot)$ being the Wasserstein distance $W(\cdot)$ for the cost function $m_{i,j}$ defined on D. $\hat{P}_t$ is as defined in (13) which is the empirical distribution based on historic flood regions. Furthermore, the candidate P is restricted to a range over a set of probability measures with support on at most n_s atoms, with an initial support being the same as that of $\hat{P}_s$ in (13). The solution can indeed be treated as minimizing a local quadratic approximation to the minimal Wasserstein distance $< \gamma, M >_F$ at $\{s_i^s\}_{i=1}^{n_s}$, following a Newton update projected onto D, according to [31, eq. (8)]. □

B. Solving for the Maximum Entropy Model

An equivalent formulation of the maximum entropy model can be shown to be equivalent to a maximum likelihood model with a Gibbs distribution [32]. This formulation can be written as

$$\begin{aligned} \arg\min_{P, \lambda_i} \quad & \hat{P} \ln\left(\frac{\hat{P}}{P}\right) + \sum_{i=1}^{K} \mathbf{C}_i |\lambda_i| \\ \text{subject to} \quad & p_{s_i} = e^{\sum_{i=1}^{K} \lambda_i f_i(s_i)} / Z \end{aligned} \tag{19}$$

where Z is the normalization factor over all $s_i \in D$, and λ_i is the feature mixture parameters.

This formulation aims to find the maximum likelihood Gibbs distribution that minimizes the relative entropy of the empirical distribution. From this formulation, it is evident that the optimal solution takes the form of a normalized function on D of an affine combination of features. There are several available options to solve this constrained convex problem. While there are several different methods, each with advantages/drawbacks [33], we opt for simplicity, namely, an iterative scaling algorithm [34] to solve the optimization. Consider

$$\begin{aligned} \arg\max_{x \in P} \quad & -x^T \log(x/x_0) \\ \text{subject to} \quad & Ax = Ay \end{aligned}$$

where x_0 comes from a prior distribution P_0 (uniform in our case), and y comes from the empirical distribution $\hat{P}$. Using its equivalent maximum likelihood formulation, we have

$$\begin{aligned} \arg\max_{x, \lambda} \quad & -y^T \log(x) \\ \text{subject to} \quad & x = \frac{x_0 \, e^{A\lambda}}{\left\| x_0 \odot e^{A\lambda} \right\|_1} \end{aligned}$$

which can be condensed into

$$\arg\max_{\lambda} \quad \log(x_0^T e^{A\lambda}) - y^T \log(x_0 \odot e^{A\lambda}).$$

To proceed, we take a coordinate descent approach on λ by maximizing the difference in change of the objective function. If $Q(\lambda)$ is the objective value, we then calculate

$$\begin{aligned} & Q(\lambda + \delta) - Q(\lambda) \\ & \quad = \log(x_0^T e^{A(\lambda+\delta)}) - y^T \log(x_0 \, e^{A(\lambda+\delta)}) \\ & \qquad - \log(x_0^T e^{A\lambda}) - y^T \log(x_0 \, e^{A\lambda}) \\ & \quad = -y^T A\delta + \log\left[\frac{x_0^T e^{A(\lambda+\delta)}}{x_0^T e^{A\lambda}}\right] \\ & \quad \le -y^T A\delta + \log\left[\frac{(x_0 \odot e^{A\lambda})^T (1 + (e^{\delta_j - 1}) A e_j)}{x_0^T e^{A\lambda}}\right]. \end{aligned}$$

The last inequality results from Jenson's inequality on e^x with $x \in [0, 1]$. The update step is obtained by differentiating

$$-y^T A\delta + \log\left[\frac{(x_0 \odot e^{A\lambda})^T (1 + (e^{\delta_j - 1}) A e_j)}{x_0^T e^{A\lambda}}\right]$$

with respect to δ and solving for the critical point. When working with the bounding box regularization, there are several critical points. The critical point that maximizes the objective value corresponds to the update step.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

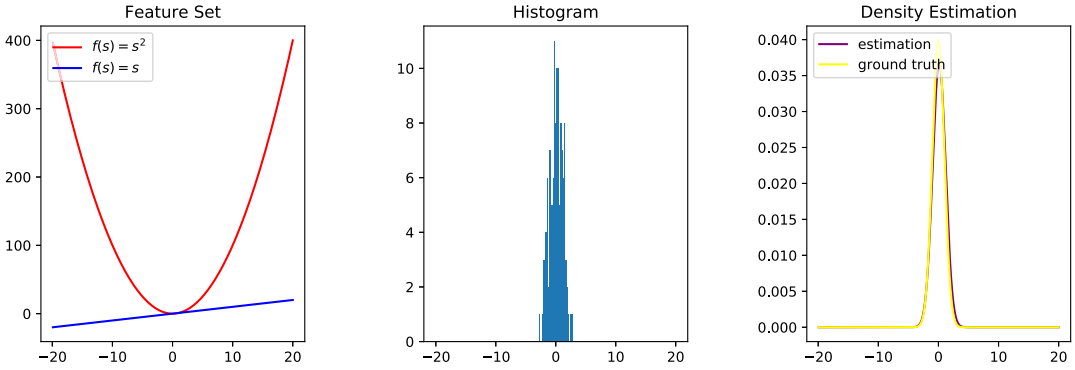


Fig. 5. (Left) Using the linear and quadratic feature functions, with 200 data points sampled from the ground truth, we can recover the distribution using the maximum entropy model.

C. Hyperparameter Tuning

In our hyperparameter tuning of our optimized fusion model, we consider in the flood density estimation the following:

1) optimal transport parameter β for the transport cost function;
2) relaxation parameter α for the constraints in the maximum entropy model.

All of the tuning requires the labeling of some true flood locations; hence, in our case study, UFE has been utilized for the tuning. In practice, the tuning can also be approximated using SFHAs for flood locations with high probability.

In principle, the feature selection is also important for an optimized fusion model. For the feature selection part in our case study, we have also considered the combination of different sets of features, including multispectral data before and after flood as well as the elevation information. Testing different feature sets using some true flood labels as our empirical flood distribution, we have decided to use all of the features. The result is shown in Section V.

The fusion algorithm is finally summarized as a pseudocode in Algorithm 1.

Algorithm 1 Optimized Fusion Algorithm

Input: Finite domain D (information fusion space);
All data modality $\tilde{f}_i : D_i \to \mathbb{V}$, $i = 1, \ldots, K$;
Optimal transport cost M;
Relaxation tolerance α for the constraints

Preprocessing data $\tilde{f}_i$ for homogenized maps $f_i : D \to \mathbb{V}_i$
for categorial maps f_i representing labeling **do**
  Compute $\hat{P}_{f_i}$
end for
$L =$ #(empirical distributions based on labelings)
if $L \geq 2$ **then**
  Assign weights $\alpha_i \in [0, 1]$ for $\hat{P}_{f_i}$, $\sum_{i=1}^{L} \alpha_i = 1$
  Compute $\tilde{P}$ as a minimizer of the sum of its Wasserstein distance to $\hat{P}_{f_i}$
end if
for f_i NOT a labeling **do**
  Normalize f_i
  Compute the moments of f_i under $\tilde{P}$
  Pose constraint $|\mathbb{E}_P f_i - \mathbb{E}_{\tilde{P}} f_i| \leq \sigma(f_i)/\alpha$
end for
Solve the Maximum Entropy density estimation model with constraints (Eq. (8)).

V. Experiments and Discussion

In the following experiments, we first demonstrate a toy example showing the intuitive idea of the algorithm performance in a simple 1-D density estimation case. We subsequently demonstrate our algorithmic solution to a real-world problem of flood density estimation by fusing heterogeneous data including satellite multispectral data and tweets.

A. Toy Simulation Example

At first, suppose the ground truth distribution is a Gaussian distribution $N(0, 1)$ over $\mathbb{R}$ with mean 0 and standard deviation $\sigma = 1$. Sampling 200 points from the distribution, we use the functions from the feature vector $\mathbf{x}(s) = [f_1(s) = s, f_2(s) = s^2]$ for additional constraints (which is equivalent to the first and second moment constraints). The maximum entropy model solution will naturally be the Gaussian distribution (see Fig. 5 for the simulation results).

When the true underlying distribution of the data is non-Gaussian, say we have a bimodal distribution following the mixture Gaussian, i.e., $P = (1/3)N(0, 1) + (2/3)N(10, 1)$, and we have 100 sample points coming from the distribution $\hat{P} = (1/3)N(1, 1) + (2/3)N(12, 1)$ (that is, the mean is a bias and shifts to the right the true mean of each mode). Moreover, suppose we have some prior distribution $P_{\text{prior}} = (1/3)N(0, 2) + (2/3)N(10, 2)$, which has the same mean in each mode as the ground truth but a larger deviation; we have 200 sample points from P_{prior}.

To show the effectiveness of our algorithm, we consider a set of feature functions more general than linear or quadratic features, since they would not suffice if the desired distribution is not Gaussian. Instead, in the simulation experiment, we randomly generated a set of bimodal feature functions $f(s) = \alpha \exp(-(x - \mu_1)^2/(2\sigma_1^2)) + (1 - \alpha)\exp(-(x - \mu_2)^2/(2\sigma_2^2))$ a mixture of Gaussians (normalized, and reflecting the true

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

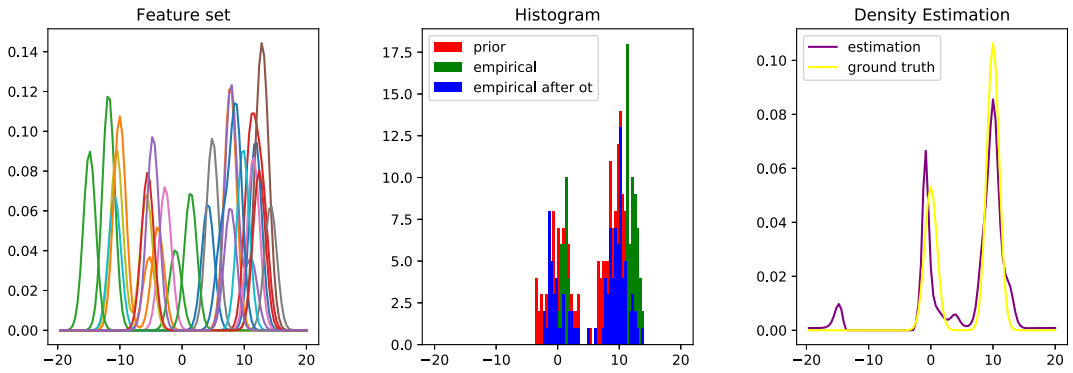


Fig. 6. (Middle) We first transport data from the empirical distribution to those from the prior distribution. (Left) Then with randomly generated feature set, we estimate the distribution (purple) and compare to the ground truth (yellow).

bimodal distribution on the line), with $\alpha, \mu_1, \mu_2, \sigma_1, \sigma_2$ sampled from uniform distributions of certain ranges. We select the "good" feature set once the maximum entropy fusion model yields an acceptable estimation given the data sampled from ground truth.[2] Using then the optimal transport with the Euclidean distance as cost function between the data points, we transport data from the empirical distribution to those from the prior distribution. This is shown in Fig. 6. In comparison, when one applies the Expectation–Maximization algorithm to the Gaussian mixture model with sample data from the empirical or the prior, the distances between the estimated ones and the ground truth (in l_2 measure) are 0.29 (empirical data points) and 0.12 (prior data points), while the distance is less than 0.12 using our algorithm (depending on the feature set we use, and we got 0.09 in some running instances).

B. Experimental Setup

As mentioned in Section III-A, we consider the Landsat 8 satellite images of date August 25, 2013 prior to flood and September 17, 2013 after the flood. We consider furthermore the first seven bands as they are most relevant to floods and are of the same resolution. The structured bands with the same resolution makes the homogenization step trivial, leading to a 14-D vector $\mathbf{f}(s)$ for the center location $s \in D \in \mathcal{R}^2$ of 30×30 m^2 cells. We also consider the elevation environmental feature elev(s) at each location s. We then normalize $\mathbf{x}(s) = (\mathbf{f}(s), \text{elev}(s))$ in each dimension over all $s \in D$, where D is the finite landscape set containing all the cell centers. For social media data, we constrain ourselves to only geolocated tweets, note that this is different from [9], in which tweets were utilized mainly for the purpose of identifying hot spots. Our primary goal in this fusion experiment is to illustrate the gain of including tweets as labeling for flood extent estimation over the whole landscape of interest. This, to a great extent, demonstrates the potential for spatial extrapolation with great flexibility when all the available data are accounted for in the fusion, thanks to the homogenization process. The *information space* derived from homogenization, together with efficient computational procedures has led to a near real-time robust natural hazard estimation with a potential for spatial extrapolation with no manual adjustments (i.e., no man in the loop). We also note that the two expert suggested features MNDWI and DIFF-NDVI in the post flood were used together with the optimal transport for bias correction in order to alleviate locational bias of tweets.

[2]This example is to illustrate how the selection of these feature functions may similarly reflect the scenario in the flood estimation case study, where different features are more or less correlated with flood density. Our choice here is, however, by no means a criterion for feature selection, but is rather an algorithmic methodology for developing the fusion model.

TABLE II
PARAMETER TUNING FOR THE CONSTRAINT TOLERANCE $\mathbf{C} = \sigma(\mathbf{S}(\mathbf{x}))/\alpha$

α	10	20	30	40	50	60	70	80	90	100	200	400	600
AUC	0.753	0.767	0.769	0.772	0.776	0.779	0.781	0.783	0.784	0.785	0.791	0.793	0.794

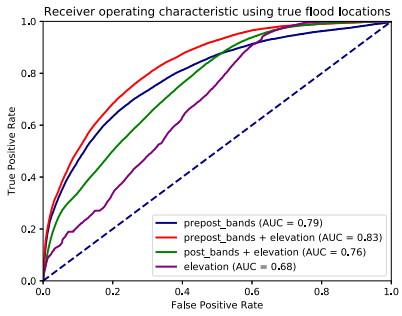
C. Evaluation of the Results and Discussions

In order to evaluate our estimation, we utilize the following assessment method. First, the receiver operator characteristic (ROC) curve is used to evaluate our binary classification of flood versus nonflood based on the density estimation. The curve is created by varying threshold and recording the true positive (TP) rate against the false positive rate. AUC denotes the area under the ROC, with values in [0, 1]. The higher the value, the better the performance is. The UFE (Fig. 2) is served as our "ground truth."

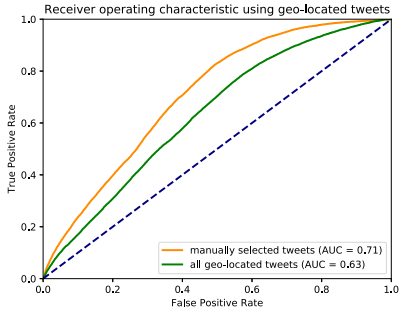
We first decide on the parameter $\mathbf{C}$ in (2), it is best to look at the performance of the fusion model with noise-free flood locations by treating UFE as true labeling directly first. Suppose $c_i = \sigma(f_i)/\alpha$, where $\sigma(f_i)$ is the standard deviation of the feature f_i under empirical distribution $\hat{P}$, and α is the tuning parameter. Working with the features from all 14 bands (seven bands information before the flood and seven after), we have the results with different α values given in Table II. The smaller the value for α, the more relaxed are the constraints in (2); therefore, a less sparse solution (more uniform) leads to less accurate estimation. Based on the results, it is harder and harder to get an increase in AUC; hence, we consider taking $\alpha = 200$ for all remaining experiments.

For achieving a reasonable feature selection $S(\mathbf{x}(s))$, and hence, avoiding additional confounding factors, we compare the following cases with a varied set of combined features for

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.



(a)



(b)

Fig. 7. (a) ROC performance with different feature selections is displayed when taking UFE as a prior/empirical distribution directly. This highlights the importance of using both presatellite and postsatellite imageries for the differential in the multiband information, together with elevation information of the region prone to flooding. (b) Using all seven band information from Landsat 8 before and after the flood, together with the elevation, the ROC performance is shown as a result of a naive use of a social media—this falls short on the potentially achievable performance [comparing to the red curve in (a)]. A manual selection of geolocated tweets only marginally improves the performance.

fusion: 1) the first seven bands from Landsat 8 both before and after the floods; 2) selection 1) with an extra feature coming from the landscape elevation elev(s); 3) only post flood bands with elevation; and 4) elevation alone. The results vary in performance, as can be seen in Fig. 7(a). This experiment unveiled that using all the information available achieves the best result. We therefore consider the fusion of all the pre/post flood remote sensing and elevation information [case (2)] in the remaining experiments.

There are several possible reasons why using ground truth falls short on yielding a "perfect" result: 1) UFE concerns only urban areas; it does not include all the flooded regions within the landscape under our study. This introduces some factors of inaccuracy in the evaluation of our estimation. The proof of concept nonetheless validates the proposed framework and 2) there is a limit in the representative features coming from satellite spectral band information, partly due to temporal mismatch of the remote sensing data. But given more timely update of satellite imagery, we expect a richer set of representative features, to yield a better timely update of the flood estimation. We plan to further explore this as additional data becomes available.

We next illustrate the behavior of the maximum entropy model when used in conjunction with biased location tweets. As noted earlier, the inherent noisy location of the labels

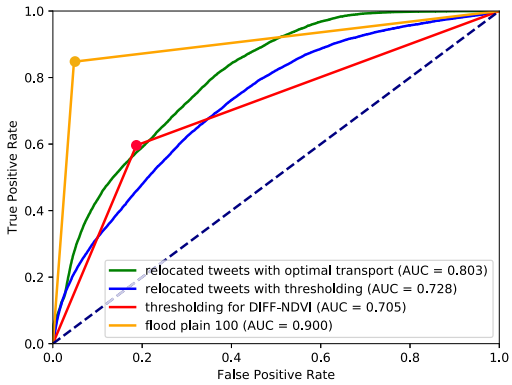


Fig. 8. Performance is enhanced by the relocation of the tweets via optimal transport. It is shown to achieve much better performance than relocation according to the thresholding of some expert spectrum features. The classification result of relying thresholding on satellite image alone is also shown as a single vertex point on the red curve; relying on history flood alone is shown as a single vertex point on the yellow curve.

introduces biases as people would normally not tweet while in the center of any flood location. The results are shown in Fig. 7(b) for a comparative assessment with the ideal performance using UFE for empirical distributions [red curve in Fig. 7(a)]. Note that with even a manual selection of the social media, the performance can be slightly improved over using all geolocated tweets.

To improve the geolocation information from the social media data, we make locational corrections with the use of expert knowledge. The manual classification of the tweets into two groups revealed that for up to 75% of observations, the tweeter may have been indirectly affected by the flooding, but are not necessarily at a flooded location. These social media observations should not be applied to the cell containing the observation, but should be attributed to nearby pixels that are likely to be flooded. Even the observations indicating flooding in the immediate vicinity may need a positional correction, as "flooding" at those precise GPS coordinates would require the observer standing in or above-mentioned water.

Instead of manual selection of tweets based on the contents, we apply the transport of the labels to the closest cells likely to be flooded using prior information. For our choice of prior, we use the MNDWI and DIFF-NDVI. The best result was achieved using relocations where DIFF-NDVI is less than -0.06. The transport is found using spatial indexing R-tree for efficient computation. As a comparison, the performance of directly thresholding DIFF-NDVI for flood region detection is inferior as can be seen from Fig. 8.

We also tested the coupling method for the optimal transport. We select $\beta = 10^2$ for the transport cost in (15). It surpasses all other methods. While one can observe in Fig. 8 that the history flood region alone achieves high TP with low negative, (meaning the history flood region is very close to our ground true), we are able to generate a well-performed actionable flood map with a probabilistic insight.

VI. Conclusion

We have addressed the largely open problem of fusing heterogeneous multimodality data by providing a formalism

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

for a principled and systematic formulation and a sound solution. This principled fusion framework provides a capacity of accounting for heterogeneous data-level input information with a decision-level output information.

Data homogenization and transport of labeling data against the domain drift problem have led to mitigate some inherent limitations of the sensing modalities in a real-world application such as the flood distribution estimation.

While the general applicability of this proposed framework is clear, many research issues remain. For example, a performance analysis along with performance bounds would further enrich this paper. Other research avenues include the development of the dynamics of the fusion problem, i.e., the spatial–temporal model for the hazard estimation. Homogenization would not only be required in space but also in time. With potentially more remote sensing data collected in time, it is possible to test our fusion model with other appropriate constraints for higher accuracy in its estimation behavior over time.

A careful and thorough evaluation of the optimal transport behavior can be achieved by taking advantage of the locational data from their environmental neighbors. Feature extraction and generation also hold much promise; approaches such as dictionary learning, for instance, can enrich the model's adaptivity to other applicable situations. It is also possible to combine with the tools introduced in [9] for hot spots detection to make our landscape selection adaptive to the labeling process. The metrics adopted in our fusion model are also considered a critical part in further developing our model.

ACKNOWLEDGMENT

The authors would like to thank the reviewers for their insightful comments and suggestions. This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

REFERENCES

[1] L. Xiao, S. Boyd, and S. Lall, "A scheme for robust distributed sensor fusion based on average consensus," in *Proc. 4th Int. Symp. Inf. Process. Sensor Netw.*, Apr. 2005, pp. 63–70.

[2] J. Yang, J.-Y. Yang, D. Zhang, and J.-F. Lu, "Feature fusion: Parallel strategy vs. serial strategy," *Pattern Recognit.*, vol. 36, no. 6, pp. 1369–1381, Jun. 2003.

[3] X. L. Dong *et al.*, "From data fusion to knowledge fusion," *Proc. VLDB Endowment*, vol. 7, no. 10, pp. 881–892, 2014.

[4] B. V. Dasarathy, "Sensor fusion potential exploitation-innovative architectures and illustrative applications," *Proc. IEEE*, vol. 85, no. 1, pp. 24–38, Jan. 1997.

[5] S. J. Phillips, M. Dudík, and R. E. Schapire, "A maximum entropy approach to species distribution modeling," in *Proc. 21st Int. Conf. Mach. Learn.*, 2004, p. 83.

[6] A. L. Berger, V. J. D. Pietra, and S. A. D. Pietra, "A maximum entropy approach to natural language processing," *Comput. Linguistics*, vol. 22, no. 1, pp. 39–71, 1996.

[7] N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy, "Optimal transport for domain adaptation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 9, pp. 1853–1865, Sep. 2017.

[8] E. Schnebele and G. Cervone, "Improving remote sensing flood assessment using volunteered geographical data," *Natural Hazards Earth Syst. Sci.*, vol. 13, no. 3, pp. 669–677, 2013.

[9] G. Cervone, E. Sava, Q. Huang, E. Schnebele, J. Harrison, and N. Waters, "Using Twitter for tasking remote-sensing data collection and damage assessment: 2013 Boulder flood case study," *Int. J. Remote Sens.*, vol. 37, no. 1, pp. 100–124, 2016.

[10] U. C. Benz, P. Hofmann, G. Willhauck, I. Lingenfelder, and M. Heynen, "Multi-resolution, object-oriented fuzzy analysis of remote sensing data for GIS-ready information," *ISPRS J. Photogramm. Remote Sens.*, vol. 58, nos. 3–4, pp. 239–258, Jan. 2004.

[11] T. Blaschke, "Object based image analysis for remote sensing," *ISPRS J. Photogramm. Remote Sens.*, vol. 65, no. 1, pp. 2–16, Jan. 2010.

[12] G. Cervone and B. N. Haack, "Supervised machine learning of fused radar and optical data for land cover classification," *J. Appl. Remote Sens.*, vol. 6, no. 1, p. 063597, 2012.

[13] V. Mnih and G. E. Hinton, "Learning to label aerial images from noisy data," in *Proc. 29th Int. Conf. Mach. Learn. (ICML)*, 2012, pp. 567–574.

[14] B. De Longueville, R. S. Smith, and G. Luraschi, "'OMG, from here, I can see the flames!': A use case of mining location based social networks to acquire spatio-temporal data on forest fires," in *Proc. Int. Workshop Location Based Social Netw. (LBSN)*. New York, NY, USA: ACM, 2009, pp. 73–80. [Online]. Available: http://doi.acm.org/10.1145/1629890.1629907

[15] S. Vieweg, A. L. Hughes, K. Starbird, and L. Palen, "Microblogging during two natural hazards events: What twitter may contribute to situational awareness," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst. (CHI)*. New York, NY, USA: ACM, 2010, pp. 1079–1088. [Online]. Available: http://doi.acm.org/10.1145/1753326.1753486

[16] T. Heverin and L. Zach, "Twitter for city police department information sharing," *Proc. Amer. Soc. Inf. Sci. Technol.*, vol. 47, no. 1, pp. 1–7, 2010. [Online]. Available: http://dx.doi.org/10.1002/meet.14504701277

[17] J. P. de Albuquerque, B. Herfort, A. Brenning, and A. Zipf, "A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management," *Int. J. Geograph. Inf. Sci.*, vol. 29, no. 4, pp. 667–689, 2015.

[18] G. Y. Lu and D. W. Wong, "An adaptive inverse-distance weighting spatial interpolation technique," *Comput. Geosci.*, vol. 34, no. 9, pp. 1044–1055, 2008.

[19] M. L. Stein, *Interpolation of Spatial Data: Some Theory for Kriging*. New York, NY, USA: Springer-Verlag, 2012.

[20] S. J. Phillips *et al.*, "Sample selection bias and presence-only distribution models: Implications for background and pseudo-absence data," *Ecol. Appl.*, vol. 19, no. 1, pp. 181–197, 2009.

[21] M. Dudik, "Maximum entropy density estimation and modeling geographic distributions of species," Ph.D. dissertation, Dept. Comput. Sci., Princeton Univ., Princeton, NJ, USA, 2007.

[22] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[23] E. T. Jaynes, "Information theory and statistical mechanics," *Phys. Rev.*, vol. 106, no. 4, pp. 620–630, 1957.

[24] M. Goodchild, "Citizens as sensors: The world of volunteered geography," *GeoJournal*, vol. 69, no. 4, pp. 211–221, 2007.

[25] A. N. Bishop, "Information fusion via the wasserstein barycenter in the space of probability measures: Direct fusion of empirical measures and Gaussian fusion with unknown correlation," in *Proc. 17th Int. Conf. Inf. Fusion (FUSION)*, Jul. 2014, pp. 1–7.

[26] C. Villani, *Optimal Transport: Old and New*, vol. 338. Berlin, Germany: Springer-Verlag, 2008.

[27] B.-C. Gao, "NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space," *Remote Sens. Environ.*, vol. 58, no. 3, pp. 257–266, 1996.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

[28] H. Xu, "Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery," *Int. J. Remote Sens.*, vol. 27, no. 14, pp. 3025–3033, 2006.
[29] R. Malinowski, G. Groom, W. Schwanghart, and G. Heckrath, "Detection and delineation of localized flooding from worldview-2 multispectral data," *Remote Sens.*, vol. 7, no. 11, pp. 14853–14875, 2015.
[30] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 2292–2300.
[31] M. Cuturi and A. Doucet, "Fast computation of wasserstein barycenters," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 685–693.
[32] S. D. Pietra, V. D. Pietra, and J. Lafferty, "Inducing features of random fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 4, pp. 380–393, Apr. 1997.
[33] R. Malouf, "A comparison of algorithms for maximum entropy parameter estimation," in *Proc. 6th Conf. Natural Lang. Learn.*, vol. 20, 2002, pp. 1–7.
[34] S. J. Phillips and M. Dudik, "Modeling of species distributions with maxent: New extensions and a comprehensive evaluation," *Ecography*, vol. 31, no. 2, pp. 161–175, 2008.

Han Wang received the B.S. degree in applied mathematics from Sichuan University, Chengdu, China, in 2008, and the Ph.D. degree in mathematics from the University of Illinois at Urbana–Champaign, Champaign, IL, USA, in 2014.

From 2015 to 2018, he was a Post-Doctoral Research Scholar with the Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, NC, USA. His research interests include algebraic topology, topological data analysis, and security problems in complex systems.

Erik Skau received the B.Sc. degrees in applied mathematics and physics and the M.Sc. and Ph.D. degrees in applied mathematics from North Carolina State University, Raleigh, NC, USA.

He was a Post-Doctoral Research Associate with the Information Sciences Group (CCS3) and Applied Mathematics and Plasma Physics Group (T5), Los Alamos National Laboratory, Los Alamos, NM, USA. His research interests include signal processing, performance modeling, machine learning, and artificial intelligence.

Hamid Krim received the B.Sc. and M.Sc. degrees in electrical engineering from the University of Washington, Seattle, WA, USA, and the Ph.D. degree in electrical and computer engineering from Northeastern University, Boston, MA, USA.

He was a member of Technical Staff with the AT&T Bell Labs, where he has conducted research and development in telephony and digital communication systems/subsystems. He was an NSF Post-Doctoral Fellowship with the Foreign Centers of Excellence, LSS/University of Orsay, Paris, France. He joined the Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA, USA, as a Research Scientist, where he was performing and supervising research. He is currently a Professor in electrical engineering with the Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, NC, USA, where he is leading the Vision, Information, and Statistical Signal Theories and Applications Group. His research interests include statistical signal/image analysis and geometric machine learning with a keen emphasis on applied problems in classification and recognition using geometric and topological tools.

Guido Cervone received the B.S. degree in computer science from the Catholic University of America, Washington, DC, USA, in 1998, and the M.S. degree in computer science and the Ph.D. in computational sciences and informatics with specialization in knowledge mining from George Mason University, Fairfax, VA, USA, in 2000 and 2005, respectively.

He is currently an Associate Professor of geoinformatics, meteorology and atmospheric science with Pennsylvania State University, University Park, PA, USA, where he serves as the Associate Director of the Institute for CyberScience, the Director of the Geoinformatics and Earth Observation Laboratory, and the Faculty Affiliate of the Environmental and Energy Study Institute. He is currently an Affiliate Scientist with the Research Application Laboratories, National Center for Atmospheric Research, Boulder, CO, USA, and an Adjunct Professor with the Lamont Doherty Earth Observatory, Columbia University, Palisades, NY, USA. His expertise is in geoinformatics, machine learning, and remote sensing. The main problem domains are related to environmental hazards and renewable energy forecasting. His research interests include the development and application of computational algorithms for the analysis of remote sensing, numerical modeling, and social media spatiotemporal Big Data.