ERDC/CHL CHETN-XII-1
August 2019

US Army Corps of Engineers®

Reducing Uncertainty and Improving Precision in Coincident Geospatial Datasets Using Weight-of-Evidence: Part 1

By Jeffrey Cegan, Matthew Wood, Igor Linkov, and Drew Allan Loney

PURPOSE: This U.S. Army Engineer Research and Development Center (ERDC) Technical Note (TN) is the first of multiple TNs focused on improving environmental datasets in limited-knowledge conditions by merging multiple datasets, each with high uncertainty and low precision, together with institutionalized subject matter expert knowledge to increase accuracy and precision. This TN provides a brief overview of geospatial data fusion and uncertainty quantification for environmental datasets. Additionally, this TN details the progress and current results following an investigation of the working hypothesis that a weight-of-evidence (WOE) framework that joins qualitative and quantitative datasets can significantly improve the accuracy and precision as related to individual datasets and current data fusion algorithms.

INTRODUCTION: The accuracy and precision of geospatial data, and the results of models that utilize that data, directly impact the information available for decision makers to select courses of action in many applications. Often, multiple coincident datasets are available for a given environmental variable, and data fusion can be used to reduce the uncertainty in a combined dataset (Cao et al. 2014; Carrara et al. 2008; Sorber et al. 2015). Coincident datasets are frequently derived from disparate sampling methods – such as satellite imagery, ground measurements, and individual observation (Jongjin et al. 2016) – and no consensus method exists for combining coincident datasets (Pan et al. 2015).

To combine coincident datasets in a way that minimizes the resulting uncertainty, it is important to first quantify the error in each of the individual datasets (Mostafavi et al. 2004). Many methods exist for quantifying uncertainty in geospatial datasets, most focusing primarily on spatial and thematic (measurement) error. However, there are many other characteristics that contribute to data quality, including temporal error, data consistency, and data completeness (Veregin 1999).

This TN describes the problem with respect to fusing coincident datasets, identifies how uncertainty is quantified for each point in each dataset, and shows how fuzzy logic with WOE can be used to fuse datasets that have both quantitative and qualitative contributors to uncertainty.

PROBLEM FRAMING: The goal of this effort is to develop a combined dataset $\boldsymbol{D}$ by merging constituent datasets $\boldsymbol{D_c}$ in such a way as to reduce multiple types of uncertainty associated with $\boldsymbol{D}$. All $\boldsymbol{D_c}$ are comprised of n_c points, which each have a value v_{ci}, and uncertainty related to that value is defined by an arbitrary set of m_c parameters, $\overrightarrow{a_{ci}} = \{a_{ci1}, a_{ci2}, \dots, a_{cim}\}$. For convenience, let c correspond to the dataset index, i correspond to the index of the point, and j correspond to the index of the parameter. Definitions for all variables used throughout this TN can be found in Table 1.

Table 1. Variables and definitions.

Parameter	Definition
$\boldsymbol{D_c}$	Individual datasets
$\boldsymbol{D}$	Fused dataset
c	Index of datasets to be fused
v_{ci}	Value of point in a dataset
i	Index of points in a dataset
a_{cij}	Uncertainty parameter j acting on point i
j	Index of uncertainty parameters
n	Number of points in a dataset
m	Number of uncertainty parameters
w_j	Weight of influence for uncertainty parameter j
σ_{cij}	Standard deviation of point i due to uncertainty parameter j

To create the combined dataset $\boldsymbol{D}$, an optimization will be conducted that minimizes all a_{cij} proportional to end-user preference for minimizing those uncertainties based on the extent to which they are restrictive to the application context for $\boldsymbol{D}$. This is done under assumptions that

1. Each type of uncertainty can be represented as a property of the points within a given $\boldsymbol{D_c}$. This assumption means that the uncertainty can vary with the internal variation of a dataset. In addition, while the uncertainty can be informed in part by dataset averaged properties, it is not necessarily defined completely by such averaged properties.
2. Uncertainties a_{cij} can be assessed across datasets using normalized relative magnitudes. This assumption implies that knowledge of the true magnitude of uncertainty is not required in all circumstances. Rather, knowledge of the relative magnitude between uncertainty parameters at a point and across points within $\boldsymbol{D_c}$ is necessary.
3. Qualitative and quantitative uncertainty information for each a_{cij} can be compared, provided that qualitative assessments approximately correspond to magnitudes and can be recast as quantitative assessments using a categorical ratio scale. This follows from assumption 2 so long as the relative relationship between quantitative and qualitative parameters can be established.

UNCERTAINTY: The uncertainty of each v_{ci} can be dependent on $\overrightarrow{a_{ci}}$ in various manners. The type of dependence can be classified into three broad categories:

1. The contribution of each a_{cij} to the uncertainty of every v_{ci} in $\boldsymbol{D_c}$ is independent and uninfluenced by the state of $\overrightarrow{a_{ci}}$. Let this be known as the fully independent case.
2. The contribution of some subset of $\overrightarrow{a_{ci}}$ to the uncertainty of some subset of the points in $\boldsymbol{D_c}$ is not independent. Let this be known as the partially independent case.
3. The contribution of each a_{cij} to the uncertainty of v_{ci} in $\boldsymbol{D_c}$ depends on the state of $\overrightarrow{a_{ci}}$ in some arbitrary fashion. Let this be known as the fully dependent case.

The fully independent case is the simplest to consider as the influence of each a_{cij} on v_{ci} can be independently quantified. The fully dependent case is much more difficult as the relationship among the parameters must be determined. The complexity of the partially independent lies between these two extremes but is complicated by the need to identify and group the dependent parameters.

The distinction among the categories is relevant to capturing the multi-parametric distribution, which represents the uncertainty. The fully independent case assumes that the sources of error each contribute independently and cumulatively to the error. The shape of the marginal distribution can then be said to hold across the entirety of the remaining parametric space. The independent case is expected to be the most common case as the mechanisms that contribute to uncertainty are largely independent themselves. If true across all a_{cij}, this allows the user to directly estimate the marginal distribution for each parameter and as a result, evaluate each point relative to the distribution of all a_{cij} in $\overrightarrow{a_{ci}}$ for each v_i, providing a mechanism to develop a point-scale distribution of uncertainty for each v_i.

In the case where partial dependency between parameters exists for common sources of uncertainty, this dependency can be described pairwise for all a_{cij}, which share some theoretical or empirical relationship. A general characterization method for uncertainties, which are found to be fully dependent on each other, remains an open area of research in the present context. If a subset of parameters are dominant, the remainder will be suggested for omission from the dataset on the basis of correlated variance. This would ideally reduce the size of the uncertainty parameter subspace such that pairwise identification is feasible. A robust method for characterization of large uncertainty parameter subsets must still be determined.

Uncertainty Parameter Source. To begin defining the point-scale uncertainty for each v_{ci} in $\boldsymbol{D}_c$, one must first consider all potential uncertainty parameters that may contribute to $\overrightarrow{a_{ci}}$. Figure 1 shows a value hierarchy that can be used to describe an arbitrary number of uncertainty parameters to quantify the error/uncertainty as well as the quality of points in a dataset (Veregin 1999). Uncertainty is presumed to result from three components representing the measurement context: spatial, temporal, and thematic. The spatial component refers to uncertainty with respect to where a measurement is taken, the temporal uncertainty is associated with when a measurement takes place, and the thematic uncertainty is associated with the type of value (soil or sediment composition, moisture value, etc.), which is being assessed. For each dimension, there are several components of quality, including accuracy (what is the variance on the dimension), precision (to how many significant figures can one measure on the dimension), and consistency (are values aligned with what else one knows to be true about the data). For example, thematic accuracy for a soil moisture measurement can be described by a z-score or some other measure of normalized variance. Similarly, temporal accuracy can be described by a confidence interval in seconds of when a measurement is expected to have actually occurred compared to the time which was recorded. Note that these first three children in the tree (Accuracy, Precision, Consistency) are predominantly concerned with measures of aleatoric uncertainty such as tolerance, variance, or other descriptions of the data or its measurement, which the user either knows given their knowledge of how the data were collected or can derive by calculating standard measures of central tendency or dispersion.

A fourth component, Completeness, reflects the coverage of the dataset with respect to the features (of both the data and its underlying model) that should be measured, the attributes of those features

which should be encoded, and the values of those attributes. It is meant to capture epistemic uncertainty associated with particular selections made by the user or the developer of the dataset related to data model configuration, sampling strategy information not captured by frequency counts or instrument tolerances, and other choices that enable the user or dataset developer to answer certain questions, potentially at the expense of being able to answer others. Completeness can be measured as a percentage, either a percentage of features that are encoded in the data model compared to what one would like to encode from the real world (feature completeness) or as a ratio of values that are filled (rather than empty) in the data schema compared to the number of expected values (value completeness). The properties of uncertainty present in this value hierarchy can be measured for each D_c.

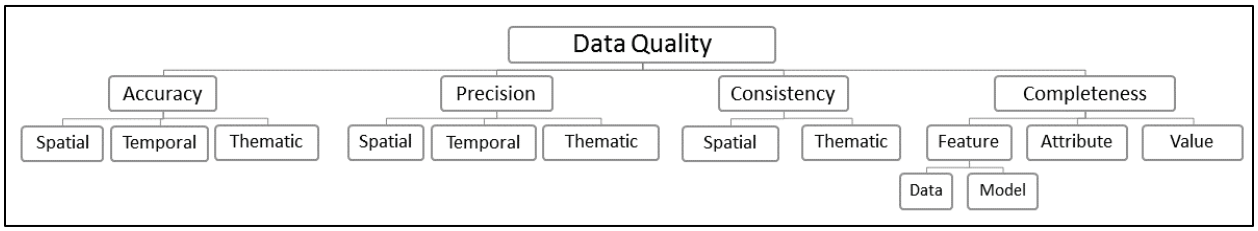

Figure 1. Categorical contributors to data quality. Lowest-level data quality categories represent uncertainty parameters, a_j, that act on individual points in a dataset, and were developed based on Veregin (1999).

Uncertainty parameters in $\overrightarrow{a_{ci}}$ can be derived from the lowest-level contributors to data quality, as seen in Figure 1. Definitions for each of these data quality components can be found in Veregin (1999). Note that not all data quality parameters are relevant for every dataset. For example, a dataset that does not vary spatially (e.g., a temperature time series for a single location) would not be subject to spatial uncertainty parameters.

Uncertainty Parameter Estimation. Estimation of $\overrightarrow{a_{ci}}$ requires a method for converting the qualitative and quantitative uncertainty parameters to distributions at the point scale in each constituent dataset. It is simplest to begin by considering the effect of each uncertainty parameter independently, assuming category 1 behavior, then aggregate the uncertainty into a single estimate. Each uncertainty parameter is assigned a quantitative uncertainty distribution that represents the contribution of the parameter to uncertainty at the point. Creating a distribution for the parameter requires that the user provide external information, either in the form of known quantitative behavior or from qualitative expert judgment. In most cases, the uncertainty parameter will be distributed normally with unknown mean and variance. An arbitrary number of parameters can be included in $\overrightarrow{a_{ci}}$ as required to fully describe the uncertainty of the point. Qualitative estimation is facilitated by the need to know the relative relationship between all parameters in $\overrightarrow{a_{ci}}$.

For a simple example, consider uncertainty related to thematic accuracy (i.e., the accuracy with which the measurement captures the true value) of value v_{ci} measured with $a_{cij} = \pm 10\%$, where j = thematic accuracy. Converting this error term to a point-scale distribution requires inferring a distribution. Possible choices include a uniform distribution with min = $v_i - 0.1v_i$ and max = $v_i + 0.1v_i$, a triangular distribution with mean = v_i, min = $v_i - 0.1v_i$ and max = $v_i + 0.1v_i$, or a normal distribution with $\mu_i = v_i$ and $\sigma_{ij} = 0.1v_i$ (Fuller 2009). The user provides additional information, either qualitative knowledge about the known shape of the error or quantitative information such as calibration metrics. This information is then stored as an uncertainty distribution on every point in $\boldsymbol{D_c}$.

In addition to thematic accuracy, spatial accuracy is an important contributor to point-scale uncertainty. Spatial accuracy is estimated as the error in v_{ci} due to uncertainty in geospatial location caused by a_{cij}, where *j* refers to spatial accuracy. For example, consider measuring soil moisture at a location (x_i, y_i) marked with a Global Positioning System, which is accurate to within a 5-meter (m) radius. On a map, that soil moisture measurement could be marked anywhere in that 5 m radius. Alternatively, any soil moisture value that exists within that 5 m radius could be the true value of v_i at position (x_i, y_i). An approach to capturing the effect of spatial error on point-scale uncertainty is to use an inverse-distance weighting approximation from the measured point to several other measured points within the 5 m radius of (x_i, y_i). A distribution is created using those simulated soil moisture measurements. The standard deviation of the simulated values around point *i*, σ_{ij}, is then used as the standard deviation for v_i due to the uncertainty caused by a_{cij}, where *j* is spatial accuracy.

The process for assigning a point-scale uncertainty, σ_{ij}, for every point *i* in $\boldsymbol{D_c}$ and every uncertainty parameter *j* is repeated until all relevant a_{cij} in $\overrightarrow{a_{ci}}$ are defined.

DATA FUSION: Once uncertainty has been characterized at the point-scale for all $\boldsymbol{D_c}$, data fusion will be done by determining which points from each $\boldsymbol{D_c}$ or portions of surfaces created using such points should be incorporated and combined into the final dataset $\boldsymbol{D}$. These data will be fused using a fuzzy-logic geographic information system – multi-criteria decision analysis approach (FLGIS-MCDA). Some subset of the uncertainty parameters shows in Figure 1 will be combined using linear weights and other rules (e.g., fuzzy logic) that represent the extent to which uncertainty from those sources is problematic for the application domain of $\boldsymbol{D}$.

Weights that express preference for data that minimizes different types of uncertainty can be established in a variety of qualitative or quantitative approaches. Qualitative approaches (e.g., logical, best professional judgment) (Weed 2005; Linkov et al. 2009) are appropriate when the problem space is simple or poorly specified. These approaches develop judgments that are dichotomous (e.g., is or is not important) or categorical (e.g., not, somewhat, very important) and convert these ratings into standardized values on a [0, 1) scale (Lipkus 2007).

Quantitative approaches, which provide greater precision in utility scores resulting from the MCDA, are appropriate as the problem space increases in complexity, and measured attributes are quantifiable, continuous, and vary enough across alternatives that they can be discriminated from each other. Here, weights are continuous decimal values with some range (e.g., [0, 1)), and the alternatives are point(s) from distinct datasets to include in a new merged dataset measured with attributes that capture how much different types of uncertainty are preferred (or not). Quantitative approaches should be preferred in the case study presented here because uncertainty scores vary sufficiently to where meaningful variations in weighting profiles leads to variations in the preference for points to include into the merged dataset.

A continuum of quantitative weighting schemes are available. Direct point allocation (Schoemaker and Waid 1982) asks users to ascribe weights to criteria directly. Simple Multi-Attribute Rating Technique (SMART) (Edwards 1977), Simple Multi-Attribute Rating Technique using Swings (SMARTS), and Simple Multi-Attribute Rating Technique Exploiting Ranks (SMARTER) (Edwards and Barron 1994) encourage users to evaluate all criteria against the one that is most important. SWING weighting (Edwards and Von Winterfeldt 1986) infers weights by asking users to rank and/or score fictional alternatives where the first is the worst on all attributes and

subsequent alternatives are the best on one attribute. A more complete review of quantitative weighting methods can be found in Alvarez-Guerra et al. (2010).

To allow a user to provide a valuation on different types of data quality and their associated uncertainty parameters, multi-criteria decision analysis will be used to quantify the extent to which each uncertainty parameter a_{cij} signals the quality of the information provided by v_{ci}. MCDA is a decision-making support process used to rank alternatives. In this case, alternatives are v_{ci} from each of several datasets that add to or transform and add to a dataset $\boldsymbol{D}$. This will be done by developing a performance function that evaluates points, based on their cumulative uncertainty, using the form

$$U(v_{ci}) = b_0 + w_1 f(a_{ci1}) + w_2 f(a_{ci2}) + \cdots w_j f(a_{cij}) \quad (1)$$

where $U(v_{ci})$ is the utility associated with submitting v_{ci} to a data fusion process, b_0 is a constant representing the baseline utility (typically set to zero), w_j is the weight associated with the users preference for data that minimize uncertainty from source a_{cij} (i.e., maximizes quality) where $\sum w_j = 1$, and $f(a_{cij})$ is a function that associates each measurement of uncertainty with the preference for that uncertainty parameter across the range of possible values of a_{cij}, each on the scale [0,1]. Weights w_j and associated uncertainty parameters a_{cij} are assumed to be independent and additive. For example, consider a river bathymetry dataset that was created using high-resolution instruments, but the dataset is 1 year old. If the river bathymetry is expected to change rapidly, then temporal accuracy may be weighted more heavily than spatial resolution, and thus a coincident dataset that was created more recently but with a lower-resolution technology may be preferred over the older dataset.

In the case where two uncertainty parameters are partially independent, an interaction term can be used to express the synergistic or antagonistic effect from a pair of uncertainties that is not captured by their independent measurement. Below is an example for two uncertainty parameters.

$$U(v_{ci}) = b_0 + w_1 f(a_{ci1}) + w_2 f(a_{ci2}) + w_{1,2} f(a_{ci1,2}) \quad (2)$$

In the case where two uncertainty parameters are completely dependent, that parameter that accounts for the most variance in overall data quality (i.e., utility) should be retained while the other parameter should be omitted from the analysis by setting its associated weight to zero. This leaves only one of dependent parameters as part of the analysis.

The extent to which some uncertainty is important to consider in the context of a data fusion process is contingent on the application for the resulting fused product. It is assumed that the user has a full understanding of relevant threats to quality for the application context and has expressed their relative preference for one form of uncertainty over others in a vector of weights w_j that sums to one. Furthermore, different expressions of $f(a_{cij})$ can be used to capture nuances of how the user prefers values of a_{cij} along its range. For instance, a step or Poisson function can be utilized if the user only desires data that are over/under some threshold value on a particular uncertainty measure but considers other values in the range of a_{cij} to be meaningless.

Each point will be scored on each type of uncertainty subject to

$$f(a_{cij}) = \frac{a_{ij} - a_{min\,j}}{a_{max\,j} - a_{min\,j}} \text{ for } 0 \leq a_{cij} < \infty$$

Where $f(a_{cij})$is a normalized score that describes how much the amount of uncertainty a_j at point i compared to other points is preferred, a_{ij} is the particular value of the uncertainty parameter at point i, and $a_{max\,j}$ is the largest value of a_j for each uncertainty parameter.

Fuzzy logic will be developed to help the user visually represent regions of data in each $\boldsymbol{D_c}$ and the final fused dataset $\boldsymbol{D}$ that are within acceptable bounds of uncertainty across all $\vec{a}$. $U(v_{ci})$ will be calculated for each point in each dataset $\boldsymbol{D_c}$. Once established, users will be asked to assign ranges of *U(v)* for each $\boldsymbol{D_c}$ that align with an arbitrary number of categorical fuzzy linguistic quantifiers of *U(v)* but at minimum discriminate acceptable from unacceptable data for each dataset. This will be done using a visualization interface that allows the user to understand relationships between geographic coverage and data quality as determined by the MCDA analysis for each $\boldsymbol{D_c}$.

The intersection of all acceptable data in all $\boldsymbol{D_c}$ will be accepted into the final fuzzed dataset $\boldsymbol{D}$ and will be used as the starting point for subsequent operations such as value of information analysis. Underlying a_{cij} in $\vec{a}$ and associated x_{ij} will be retained for purpose of informing these methods.

CASE STUDY: This case study is intended to illustrate the application of a basic FLGIS-WOE framework to uncertainty reduction and demonstrate the feasibility of the FLGIS-WOE approach. Two sources of uncertainty, thematic and spatial, are introduced into the framework. Thematic uncertainty arises from the accuracy of the point values; spatial uncertainty occurs from the point location accuracy. Synthetically generated data are employed to better exhibit the method without confounding uncertainty sources. Two different scenarios of combining data based on assessment of uncertainty parameters via FLGIS-WOE are shown in this example: fusion with perfect knowledge of the true value of the points and fusion with imperfect knowledge that lacks the information about the true value of the points.

The input to the case study is two synthetic datasets as illustrated in Figures 2 and 3 with Thiessen polygons showing the region characterized by each point using nearest neighbor interpolation. The datasets were created by generating an 11 × 11 equally spaced grid of points on these intervals:

$$x = [0, 100]$$
$$y = [0, 100]$$

Random location noise was then added to each point by sampling an $N(0,1)$ distribution in both the x and y dimensions and summing it with the location coordinates. Introduction of normal noise to the location coordinates mimics the spatial sampling variability present in physical datasets. The value of the points was generated by sampling an $N(1,1)$ distribution to represent a uniformly valued dataset of varying accuracy. This method for introducing spatial and thematic errors yields errors that are independent and uncorrelated among the points. $\boldsymbol{D_1}$ and $\boldsymbol{D_2}$were produced in the same manner with the only variation being the errors that were introduced through the spatial and value sampling.

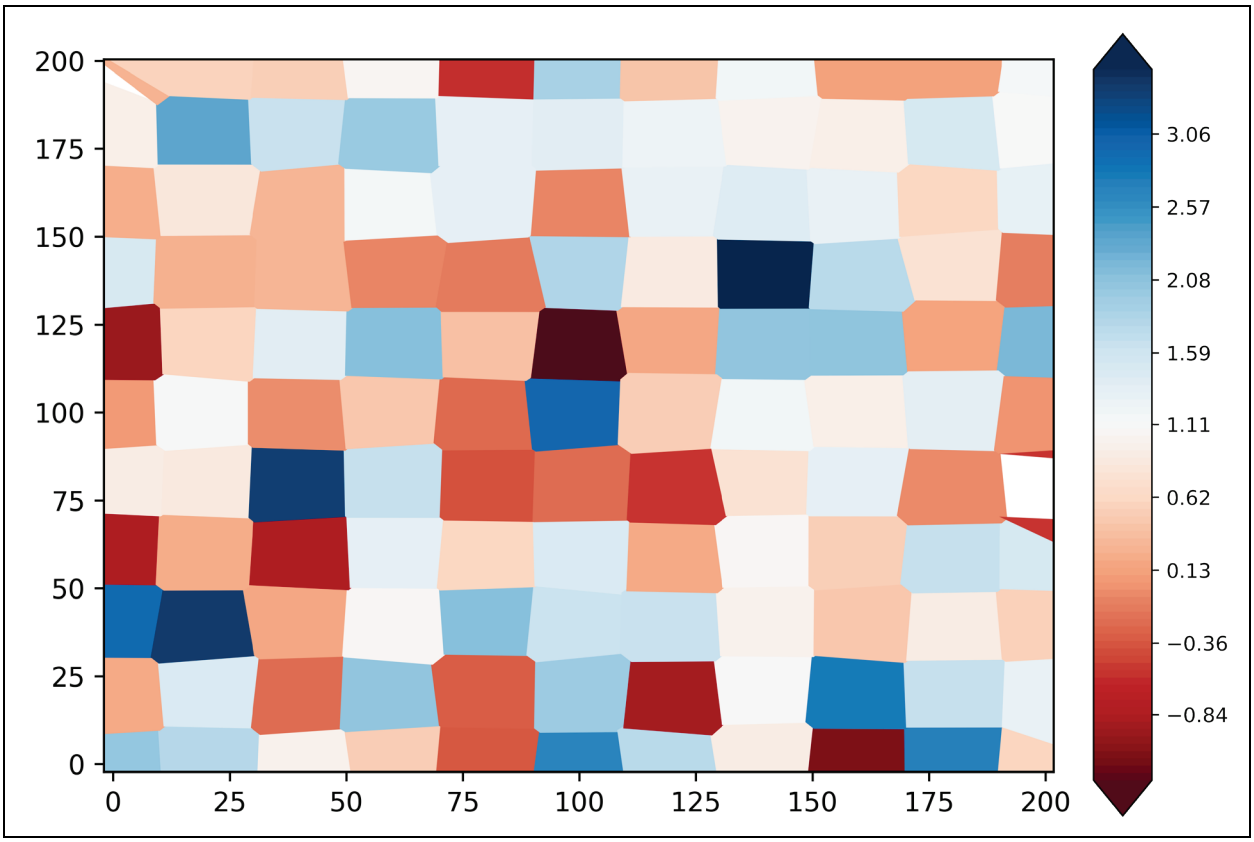

Figure 2. Point values of synthetic dataset one developed for testing WOE data fusion. The region characterized by each point is displayed by its corresponding Thiessen polygon. The colorbar indicates the deviation of each point around the dataset mean of 1.

The data fusion method utilized in both examples is simple point prioritization. The value of the best identified constitutive dataset point is used to represent the location in the fused dataset. This is done to test the feasibility of the WOE approach in general, under the assumption that changes to the particular data fusion methods deployed here could yield changes (ideally, improvement) in data quality independent of the WOE approach. More complex data fusion methods can be readily incorporated into the data fusion setup in future work, utilizing the uncertainty and fuzzy sets as input, with additional improvement expected over that which is demonstrated by the WOE approach alone. Similarly, fuzzy sets can be used to capture a wide range of categorical or continuous inputs at varying resolutions for description and aggregation by subsequent data fusion or WOE processes, with the number of sets and the degree of membership specificity producing substantive impacts on the quality of the resulting fused dataset.

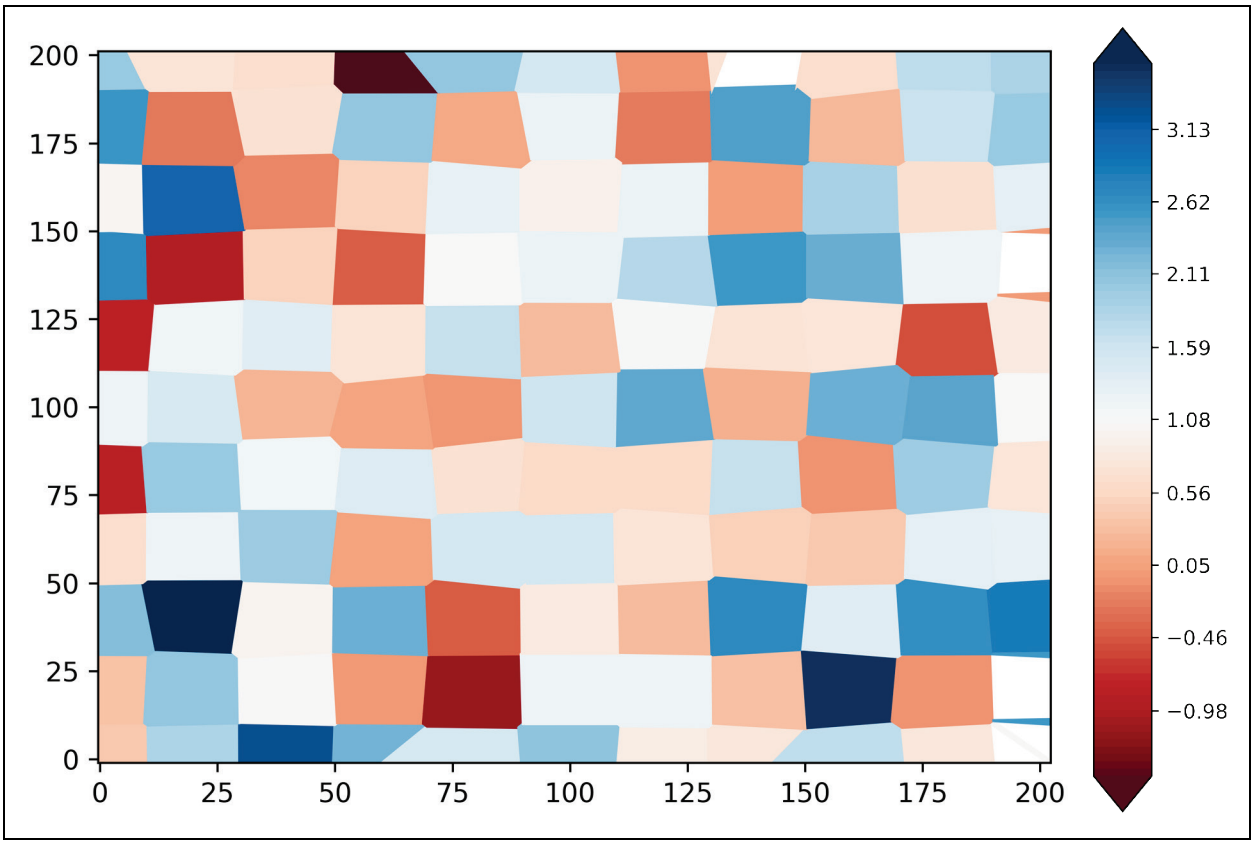

Figure 3. Point values of synthetic dataset two developed for testing WOE data fusion. The region characterized by each point is displayed by its corresponding Thiessen polygon. The colorbar indicates the deviation of each point around the dataset mean of 1.

Fusion with Perfect Knowledge. Data fusion under perfect knowledge assumes that the true value at each point is known. The true point value can then be utilized to calculate the error, and the point that minimizes the error can then be selected for inclusion in the fused dataset.

Let $(\tilde{x}_\iota, \tilde{y}_\iota)$ represent a location in the fused dataset $\boldsymbol{D}$. Additionally, assume $\tilde{v}_\iota$ to be the true value at $(\tilde{x}_\iota, \tilde{y}_\iota)$. The error for each dataset at the location is given as

$$\epsilon_{ci} = |v_{ci} - \tilde{v}_\iota|$$

where v_{ci} is the point value given by the dataset regionalization. The error for each constituent dataset $\boldsymbol{D_c}$ can be independently calculated in the above manner.

Let there also be two fuzzy sets of interest, defined for convenience as the *Minimize* and the *Maximize* sets. In this case, membership is complete and mutually exclusive; that is, a member of the Minimize must belong in whole to the set and may not also belong to the Maximize set. Membership of the constituent dataset points in each set is defined by

$$\{v_{ci} \in Minimize \mid \min\left(\epsilon_c|_{\tilde{x}_\iota,\tilde{y}_\iota}\right)\}$$

$$\{v_{ci} \in Maximize \mid \max\left(\epsilon_c|_{\tilde{x}_\iota,\tilde{y}_\iota}\right)\}$$

Membership in the Minimize set is thus defined as those points which minimize the error while membership Maximize set is defined as those points that maximize the error. Because the dataset is known to be uniform, points cannot belong simultaneously to both sets.

The data fusion step employs the Minimize and Maximize fuzzy sets to merge the constituent datasets into a single dataset. The sets, as defined in the binary fashion, naturally separate the points into the Minimize set, which is useful for data fusion, and the Maximize set, which is not based on the error magnitude. The Maximize set can thus be discarded in its entirety. The value of the points present in the Minimize set are applied directly to the $(\tilde{x}_i, \tilde{y}_i)$ locations. This is equivalent in an MCDA analysis to fully weighting spatial uncertainty and selecting points that fall above some threshold on a value function for spatial uncertainty, $f(a_{ci\ spatial})$, which scores uncertainty inverse to its magnitude.

Figure 4 displays the fused result following the described method. Ten-thousand points were generated on a uniform grid between the minimum and maximum extents of the constituent datasets. The input constituent datasets have an area weighted error of 0.792 and 1.050, respectively. The fused dataset reduces the error to 0.464. A reduction in the fused dataset error is expected in this case because the perfect knowledge condition permits direct use of the error minimizing values. The improvement in the dataset is limited by the initial values contained in the constituent datasets.

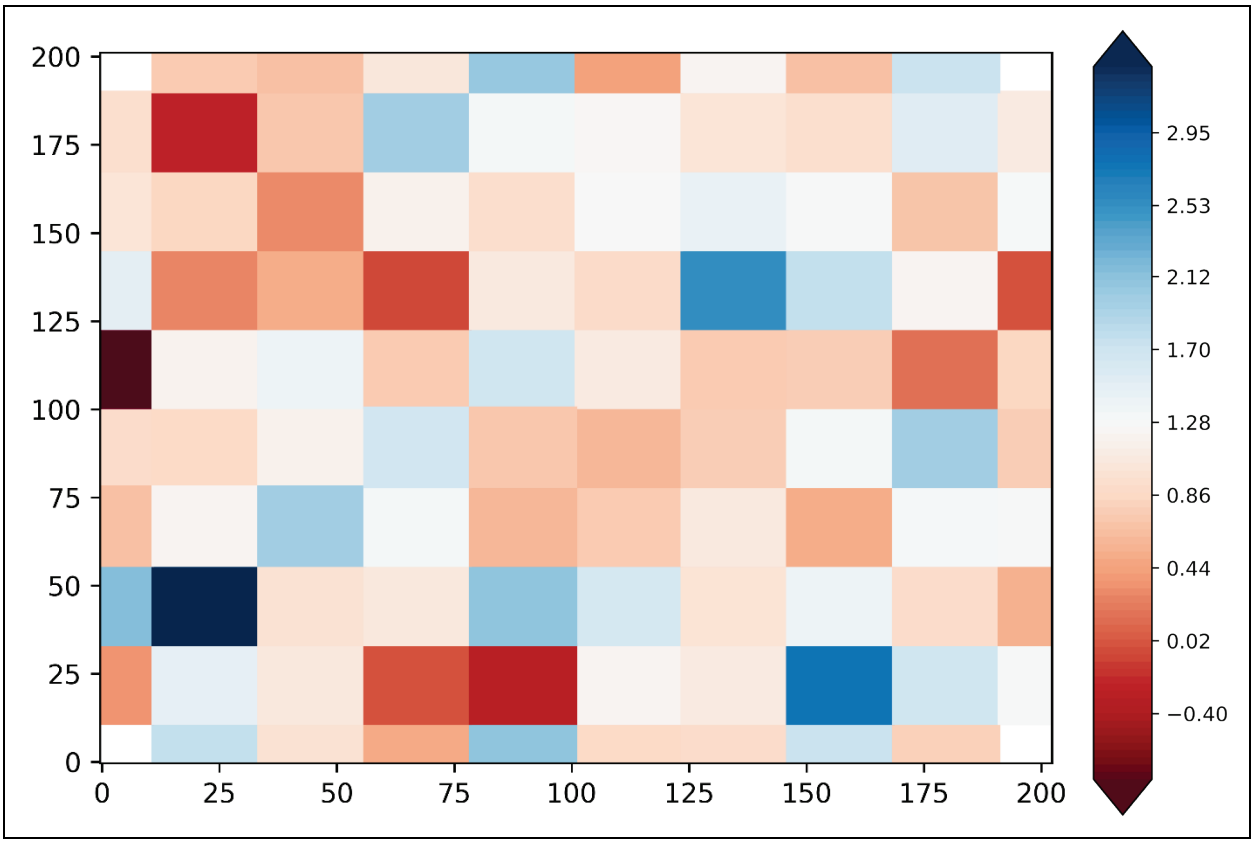

Figure 4. The fused dataset under perfect knowledge conditions. The dataset has an area averaged error of 0.465.

Fusion with Imperfect Knowledge by Combining Parameters. Data fusion under imperfect knowledge assumes that the true value at each point is unknown. This requires additional information, such as calibration information, to characterize the point for data fusion to be beneficial. Note that with a sufficient number of constituent datasets, it is possible to estimate error using methods such as collocation. Those methods are not precluded by the current method and may serve as inputs for similar analyses.

If multiple types of uncertainty are being recorded by a point, for example thematic and spatial, it facilitates the analysis to convert one type of uncertainty into another if a conversion mechanism is known. This permits judgements to be made about the relative contribution of each uncertainty type to the fuzzy set generation and subsequent data fusion. For example, thematic uncertainty can be estimated from spatial uncertainty for a point by searching for other points in nearby proximity within the characteristic spatial uncertainty length scale. The length scale can be taken as the standard error or standard deviation of the spatial resolution. An estimate for the thematic uncertainty is produced from those points that fall within that characteristic length scale. If no other points occur within the length scale, the thematic uncertainty estimate is taken as zero.

Let $(\widetilde{x_l}, \widetilde{y_l})$ represent a location in the fused dataset $\boldsymbol{D}$. Additionally, assume $\bar{\sigma}_{tc}$ to be the standard deviation of the parameter values taken as uniform across the dataset given by the calibration of the instrument. Similarly, assume that the spatial uncertainty $\bar{\sigma}_{si}$ is known from the instrument calibration as well. Converting the spatial to thematic uncertainty gives an estimate of the thematic uncertainty, $\acute{\sigma}_{tci}$ derived from the spatial uncertainty alone. Both $\bar{\sigma}_{tc}$ and $\acute{\sigma}_{tci}$ are valid estimators of thematic uncertainty resulting from different source information with the former being a uniform property across each constituent dataset and the latter being specific to each point.

Data fusion starts with classification of the constituent points into fuzzy sets. A fuzzy set represents some characteristic of the underlying data and the degree to which each point represents that characteristic. While fuzzy sets will in general span constituent datasets, as evident in the previous perfect knowledge example, equally valid are fuzzy sets that are wholly comprised of a single dataset when the constituent datasets have no shared characteristics relevant to the data fusion. Because the uncertainty estimators are derived from different sources, they must be classified into separate fuzzy sets, α and β, given by the following:

$$\{v_{ci} \in \alpha \mid (c = 1)\}$$
$$\{v_{ci} \in \beta \mid (c = 2)\}$$

where c is the constituent dataset number. Other equally valid fuzzy set configurations may exist, such as classifying both constituent datasets into the same set, so long as compensating changes are made as necessary throughout the remainder of the analysis.

The data fusion step then places weight on each fuzzy set and each type of uncertainty to facilitate the fusion process. As both constituent datasets were generated in the same manner, there is no rationale for weighting either fuzzy set α or β more highly than the other. However, the uncertainty types are a feature that merit different scores. Because the $\bar{\sigma}_{tc}$ is uniform, it is less useful for identifying poor performing points than $\acute{\sigma}_{tci}$. Without additional information, it is therefore appropriate to weight the spatially derived thematic uncertainty to be twice as important as the

calibration thematic uncertainty to the fused dataset. This weighting can change based on the type of data being considered as well as the available uncertainty information about the data.

The use of the uncertainty weights will vary by the data fusion method being implemented. Using point prioritization, it is most direct to sum the thematic uncertainty distributions with the weights within each constituent dataset to obtain a combined thematic distribution for each point prior to comparing datasets. The combined dataset for each constituent point becomes

$$\mu_{combined} = \left(0.34 * \mu_{thematic} + 0.66 * \mu_{thematic,spatial}\right)$$
$$\sigma^2_{combined} = \left(0.34 * \sigma^2_{thematic} + 0.66 * \sigma^2_{thematic,spatial}\right)$$

The combined distribution for each constituent dataset at each $(\tilde{x}_i, \tilde{y}_i)$ location is queried and the value of the point with the smallest standard deviation accepted as the value at the location.

Figure 5 displays the fused result following the described imperfect knowledge method. One hundred points were generated on a uniform grid between the minimum and maximum extents of the constituent datasets. The input constituent datasets have an area weighted error of 0.792 and 1.050, respectively. The fused dataset reduces the error to 0.474, a significant reduction and comparable to the perfect knowledge case. This improvement is result of additional uncertainty information being included as part of the data fusion process. To verify the sensitivity of the improvement to selected uncertainty weights, it is necessary to vary the selected weights. This characterization is left to future work.

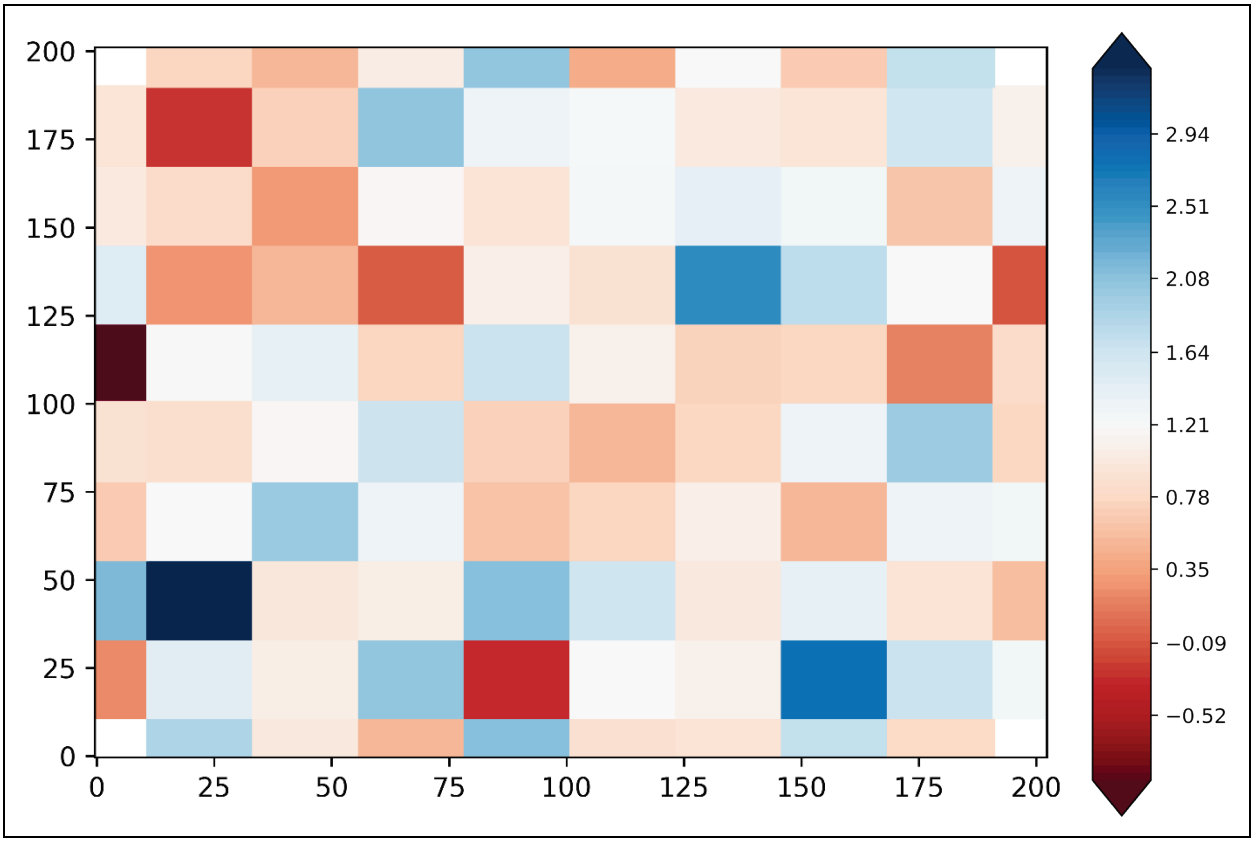

Figure 5. The fused dataset under imperfect knowledge conditions. The dataset has an area averaged error of 0.474.

Following both of the presented examples, symbology can then be associated with points in each range and be presented via any geographic information system mapping tool to show the user where current data are adequate or conversely, where one may want to collect more data or stand up a value of information analysis designed to evaluate whether more data in an identified region would improve/change the decision being made with the data. This symbology can follow a heat map representation as done above, polygons describing regions where one's efforts are best focused, or others that describe relationships between geographic area, data quality, and other factors.

SUMMARY AND FUTURE WORK: To date, this work has focused on developing theory for geospatial data uncertainty parameterization and ultimately data fusion. By considering multiple disparate sources of uncertainty, the true distribution of expected values for a given location can be more accurately quantified (Koks and Challa 2003). In additional to developing theory for uncertainty parameterization, this work has thus far simulated data uncertainty using multiple uncertainty parameters for given datasets. Future work will include testing and evaluation of the proposed data fusion process using relevant datasets.

ADDITIONAL INFORMATION: For additional information, contact Drew A. Loney, Hydrologic Systems Branch, Coastal and Hydraulics Laboratory, Vicksburg, MS, (601) 634-3490, or email: Drew.A.Loney@erdc.dren.mil. This research was funded by the USACE Military Engineering program. This CHETN should be cited as follows:

> Cegan, J., M. Wood, I. Linkov, and D. A. Loney. 2019. *Reducing Uncertainty and Improving Precision in Coincident Geospatial Datasets Using Weight-of-Evidence: Part 1.* ERDC/CHL CHETN-XII-1. Vicksburg, MS: U.S. Army Engineer Research and Development Center. http://dx.doi.org/10.21079/11681/33663.

REFERENCES

Alvarez-Guerra, M., L. Canis, N. Voulvoulis, J. R. Viguri, and I. Linkov. 2010. "Prioritization of Sediment Management Alternatives Using Stochastic Multicriteria Acceptability Analysis." *Science of the Total Environment* 408(20): 4354–4367.

Cao, G., E. Yoo, and S. Wang. 2014. "A Statistical Framework of Data Fusion for Spatial Prediction of Categorical Variables." *Stochastic Environmental Research and Risk Assessment* 28(7): 1785–1799. *https://doi.org/10.1007/s00477-013-0842-7*

Carrara, P., G. Bordogna, M. Boschetti, P. A. Brivio, A. Nelson, and D. Stroppiana. 2008. "A Flexible Multi-Source Spatial-Data Fusion System for Environmental Status Assessment at Continental Scale." *International Journal of Geographical Information Science* 22(7): 781–799. https://doi.org/10.1080/13658810701703183

Edwards, W. 1977. "How to Use Multiattribute Utility Measurement for Social Decisionmaking." *IEEE Transactions on Systems, Man, and Cybernetics* 7(5): 326–340.

Edwards, W., and F. H. Barron. 1994. "SMARTS and SMARTER: Improved Simple Methods for Multiattribute Utility Measurement." *Organizational Behavior and Human Decision Processes* 60(3): 306–325.

Edwards, W., and D. von Winterfeldt. 1986. *Decision Analysis and Behavioral Research*. Cambridge, UK: Cambridge University Press.

Fuller, W. A. 2009. *Measurement Error Models,* Vol. 305. Hoboken, NJ: John Wiley & Sons.

Hofstra, N., M. Haylock, M. New, P. Jones, and C. Frei. 2008. "Comparison of Six Methods for the Interpolation of Daily, European Climate Data." *Journal of Geophysical Research* 113(D21). https://doi.org/10.1029/2008JD010100

Jongjin, B., P. Jongmin, R. Dongryeol, and C. Minha. 2016. "Geospatial Blending to Improve Spatial Mapping of Precipitation with High Spatial Resolution by Merging Satellite-Based and Ground-Based Data: Merging Methods of Ground and Satellite-based Precipitation." *Hydrological Processes* 30(16): 2789–2803. https://doi.org/10.1002/hyp.10786

Koks, D., and S. Challa. 2003. *An Introduction to Bayesian and Dempster-Shafer Data Fusion.* Technical Report No. DSTO-TR_1436. Sydney, Australia: DTIC Document. https://apps.dtic.mil/docs/citations/ADA417895

Linkov, I., D. Loney, S. Cormier, F. K. Satterstrom, and T. Bridges. 2009. "Weight-of-Evidence Evaluation in Environmental Assessment: Review of Qualitative and Quantitative Approaches." *Science of the Total Environment* 407(19): 5199–5205.

Lipkus, I. M. 2007. "Numeric, Verbal, and Visual Formats of Conveying Health Risks: Suggested Best Practices and Future Recommendations." *Medical Decision Making* 27(5): 696–713.

Mostafavi, M.-A., G. Edwards, and R. Jeansoulin. 2004. "An Ontology-Based Method for Quality Assessment of Spatial Data Bases." In *Third International Symposium on Spatial Data Quality,* Vol. 1, 49–66. https://hal.inria.fr/inria-00000447/

Pan, M., C. K. Fisher, N. W. Chaney, W. Zhan, W. T. Crow, F. Aires, and E. F. Wood. 2015. "Triple Collocation: Beyond Three Estimates and Separation of Structural/Non-Structural Errors." *Remote Sensing of Environment* 171: 299–310. https://doi.org/10.1016/j.rse.2015.10.028

Schoemaker, P. J., and C. C. Waid. 1982. "An Experimental Comparison of Different Approaches to Determining Weights in Additive Utility Models." *Management Science* 28(2): 182–196.

Sorber, L., M. Van Barel, and L. De Lathauwer. 2015. "Structured Data Fusion." *IEEE Journal of Selected Topics in Signal Processing* 9(4): 586–600. https://doi.org/10.1109/JSTSP.2015.2400415

Veregin, H. 1999. "Data Quality Parameters." *Geographical Information Systems* 1: 177–189.

Weed, D. L. 2005. "Weight of Evidence: A Review of Concept and Methods." *Risk Analysis* 25(6): 1545–1557.