Data Quality Challenges in Weather Sensor Data, Including Identification of Mis-located Sites

Douglas E. Galarus
Computer Science Department
Utah State University
Logan, UT 84322-4205, United States
douglas.galarus@usu.edu

Rafal A. Angryk
Department of Computer Science
Georgia State University
Atlanta, GA 30302, United States
angryk@cs.gsu.edu

Abstract—**There are many challenges in developing and evaluating methods including: real-world cost and infeasibility of verifying ground truth, non-isotropic covariance, near-real-time operation, challenges with time, bad data, bad metadata, and other quality factors. In this paper, we demonstrate the challenges of evaluating spatio-temporal data quality methods for weather sensor data via a method we developed and other popular, interpolation-based methods to conduct model-based outlier detection. We demonstrate that a multi-faceted approach is necessary to counteract the impact of outliers. We demonstrate the challenges of evaluation in the presence of incorrect labels of good and bad data. We also investigate, in depth, the challenge of identifying mis-located sites.**

Keywords-Data Quality; Spatial-Temporal Data; Quality Control; Outlier; Inlier; Bad Data; Ground Truth; Bad Metadata

I. INTRODUCTION

In our research, we address near-real-time determination of outliers and anomalies in spatiotemporal weather sensor data, and the implications of quality assessment on computation from the perspective of the data aggregator. This paper extends the work presented in the conference paper at GeoProcessing 2019 [1]. Data might not reflect the conditions they measure for a variety of reasons. The challenges go beyond identifying individual outlying observations. A sensor might become "stuck" and produce the same output over an extended period. A sensor's output may conform to other nearby observations and fall within an acceptable range of values, but not reflect actual conditions. A sensor may drift, reporting values further from ground truth over time. A sensor may report correct values, but the associated clock may be incorrect, resulting in bad timestamps. An incorrect location may be associated with a site. In fact, many sites may be mis-located. These and related problems cause challenges that are far more complex than simple outlier detection.

Sensor-level quality control processes often utilize domain-specific, rule-based systems or general outlier detection techniques to flag "bad" values. NOAA's Meteorological Assimilation Data Ingest System (MADIS) [2] applies the range [-60° F, 130° F] to check for air temperature observations [3] while the University of Utah's MesoWest [4] uses the range [-75° F, 135° F] [5] for validity checks. These ranges are intended to represent the possible air temperature values in real world conditions, at least within the coverage area of the provider. If an observation falls outside the range, then the provider flags that observation as having failed the range test and the observation will, for all practical purposes, be considered "bad". Range tests are not perfect. The record high United States temperature would fail MADIS's range test, although it would pass MesoWest's test. Both MADIS and MesoWest further employ a suite of tests that go beyond their simple range tests. "Buddy" tests compare an observation to neighboring observations. MADIS uses Optimal Interpolation in conjunction with cross-validation to measure the conformity of an observation to its neighbors [3]. MesoWest estimates observations using multivariate linear regression [6]. A real observation is compared to the estimate, and if the deviation is high, then the real observation is flagged as questionable.

These approaches are flawed in that they do not account for bad metadata, such as incorrect timestamps or incorrect locations. They do not account for chronically bad sites which produce bad data including data that may sometimes appear correct. Of even greater concern, they may not do a good job in assessing accuracy and may be incorrectly labeling bad data as good and good data as bad.

The consequences of ignoring data quality are great. How can we trust our applications and models if the inputs are bad? In turn, how can we better assess data for quality so that we can be confident in its use?

In this paper, we present evaluation results for our previously published method including evaluation with several data sets. These results are significant in that they demonstrate the challenges of evaluation of methods for data quality assessment of spatio-temporal weather sensor data. We also investigate in depth the problem of identifying mis-located sites. The rest of this paper is organized as follows: Section II presents relevant literature, Section III identifies general challenges, Section IV defines our approach, Section V documents evaluation results, Section VI presents new investigation of mis-located sites, and Section VII gives our conclusions.

II. LITERATURE REVIEW

The data mining process includes data preprocessing and cleaning as critical components. Outlier analysis, is addressed within these headings by Han et al. [7], and the impact of outliers is covered by Nisbet et al. [8]. Robust re-

gression techniques are employed in data mining to overcome outliers and low quality data in the process of data cleaning by Witten et al. [9]. The handling of errors and missing values is presented by Steinbach and Kumar [10], along with quality attributes, such as accuracy and precision, as well as the adverse impact that outliers can have on clustering algorithms. Such examples demonstrate the chicken-egg nature of the problem in which a method used to identify outliers is adversely impacted by outliers.

Aggarwal [11] presents a number of useful, general observations: Correlation across time series can help to identify outliers, using one or multiple series to predict another. Deviations between predicted and actual values can then be used to identify outliers. When used on temporal snapshots of data, spatial methods can fall short because they do not address the time component. Decoupling the spatial and temporal aspects can be suboptimal. Neighborhoods can be used to make predictions, yet it is a challenge to combine spatial and temporal dimensions in a meaningful way. Domain-specific methods can be used to filter noise, but such filtering can mask anomalies in the data.

Shekhar et al. [12] present a unified approach for detecting spatial outliers and a general definition for spatial outliers, but they do not address the spatio-temporal situation. Klein et al. [13]–[17] present work on transfer and management challenges related to the inclusion of quality control information in data streams and develop optimal, quality-based load-shedding for data streams in. A missing component is the spatial aspect.

The weather and road-weather communities employ detailed accuracy checks for individual observations. The Oklahoma Mesonet uses the Barnes Spatial Test [18], a variation of Inverse Distance Weighting (IDW) (see Shepard [19]). MesoWest [4] uses multivariate linear regression to assess data quality for air temperature, as described by Splitt and Horel in [20] and [21]. MADIS [2] implements multi-level, rule-based quality control checks including a level-3 neighbor check using Optimal Interpolation / kriging [3][22][23]. These approaches (IDW, Linear Regression, kriging) can be used to check individual observations for deviation from predicted and flag individual observations as erroneous or questionable if the deviation is *large*. But if interpolated values are erroneous, then the quality assessment will be bad too. If metadata, such as location or timestamps associated with a site, is erroneous, then the quality control assessment may be bad because of comparison with the wrong data from the wrong sites. None of these approaches identify incorrect location metadata. One provider, Mesowest, attempts to identify bad timestamps, yet their approach only identifies one of the most obvious timestamp-related problem – timestamps that cannot possibly be correct because they occur in the future relative to collection time. Our own experience in this domain has been that sites are often mis-located. We found in one instance that multiples sites were mis-located with different locations across four systems from which we were extracting data.

Many spatial approaches use interpolation for quality assessment, so it is useful to examine work that compares and enhances traditional interpolation methods. Zimmerman et al. [24] use artificial surfaces and sampling techniques, as well as noise level and strength of correlation, to compare Ordinary kriging (OK) and Universal kriging (kriging with a trend) (UK) and IDW. They found that the kriging methods outperformed IDW across all variations they examined. Lu and Wong [25] found instances in which kriging performed worse than their modified version of IDW, where they vary the exponent depending on the neighborhood. They indicate that kriging would be favored in situations for which a variogram accurately reflects the spatial structure. Mueller et al. [26] show similar results, saying that IDW is a better choice than OK in the absence of semi-variograms to indicate spatial structure.

In prior work, we proposed a modification of IDW that used a data-based distance rather than geographic distance to assess observation quality [27][28]. That work focused on the use of robust methods to associate sites for assessment of individual observations. In [29][30][31], we extended the mappings to better account for spatio-temporal variation and observation time differences when assessing observations. In [32] and [33], we developed quality measures that extended beyond sites, to help evaluate overall spatial and temporal coverage of a region.

IDW is widely applied, including applications which involve outlier detection and mitigation. Xie et al. [34] applied it to surface reconstruction, in which they detect outliers using distance from fitted surfaces. Others extend the method in different ways including added dimensions, particularly time. Li et al. extend IDW in [35] to include the time dimension in their application involving estimated exposure to fine particulate matter. Grieser warns of problems with arbitrarily large weights when sites are near in analyzing monthly rain gauge observations [36], and mitigates the problem in a manner that Shepard originally used by defining a neighborhood for which included points are averaged with identical weights in place of the large, inverse distance weights.

Kriging and Optimal Interpolation were developed separately and simultaneously as spatial best linear unbiased predictors (blups) that are for practical purposes equivalent. L. S. Gandin, a meteorologist, developed and published optimal interpolation in the Soviet Union in 1963. Georges Matheron, a French geologist and mathematician, developed and published kriging in 1962, named for a South African mining engineer, Danie Krige, who partially developed the technique in 1951 and later in 1962. For further information, refer to Cressie [37].

Kriging is easily impacted by multiple data quality dimensions and its applicability is hindered unless data quality issues in the inputs are addressed. Kriging will downweight observations that are clustered in direction, as indicated by Wackernagel et al. [38]. This may be beneficial. However, a near observation can also shadow far observations in the same direction, causing them to have small or

even negative weights. This is problematic in the case that the near observation is bad.

Kriging is typically used to interpolate values at locations for which measurements are unknown using observations from known locations. As such, covariance is typically estimated. This estimate usually takes the form of a function of distance alone and is determined by the data set. A principal critique of kriging is that while it does produce optimal results when the covariance structure is known, the motivation for using kriging is questionable when the covariance structure must be estimated. Handcock and Stein [39] make such an argument. Another critique is that kriging will yield a model that matches data input to the model, giving the (false) impression that the model is perfect, as stated by Hunter et al. [40].

Unfortunately, none of these approaches alone directly addresses outlier and anomaly detection for spatio-temporal data in a robust and comprehensive manner that meets our needs. None identify bad sites and metadata in a comprehensive manner. Even so, the data quality attributes presented are of some benefit and the methods used by the weather data providers appear to be state of the art for assessment of accuracy.

III. CHALLENGES

Our research involves (fixed) site-based, spatio-temporal sensor big data, acquired and evaluated for data quality with real-time potential. There are many computational challenges associated with our problem. We focus subsequent evaluation on scalability and accuracy.

Scalability. Our data sets include thousands of sites, with potential to expand to tens of thousands of sites. Sites have varying reporting frequencies ranging from every minute to hourly or longer. These sites collectively generate millions of observations daily. We desire to run our algorithms in near real-time, and scalability is key to achieving this goal.

Accuracy. The underlying data has many data quality challenges. Accurately modeling the data is challenging, because the modeled data will inherently include errors. We desire robust, accurate models that can be used to assess the quality of individual observations.

There are many indirect issues causing challenges that must be overcome. These all influence or are influenced by computation in one way or another.

Real-World Cost and Infeasibility of Verifying Ground Truth. Agencies cannot verify ground truth on a regular basis across hundreds or thousands of sites. Human-required resolution processes can be focused if problems are identified automatically. Third-party data aggregators have no control over original data quality. Assessment of quality is essential for use.

Non-Isotropic Covariance. Distance cannot be treated equally in all dimensions nor in all directions. There are differences between the time dimension and spatial dimensions. Elevation, proximity to the ocean, terrain, microclimates, prevailing weather patterns, the diurnal effect, seasonal change, etc. also cause differences in covariance.

Near-Real-Time Operation. We intend for our processes to run in near-real-time when observations are acquired. We store and use only the most recent observations for near real-time presentation and comparison. We do not intend to store third-party historical data on our production systems. This does not preclude the potential for offline preprocessing and analysis that makes use of historical data. Even if providers apply their own quality control measures, near-real-time operation may require us to use observations that have not been fully quality-checked.

Further Challenges with Time. Sites report observations at discrete times resulting in granularity and non-uniformity. Observation frequencies and reporting times vary across sites. Network latency and batch processing further disrupt timeliness.

"Bad" Data. Bad data includes but is not limited to erroneous observation data – individual observations that differ from ground truth; "bad" sites – sites that chronically produce erroneous data; and "bad" metadata including incorrect locations and/or incorrect timestamps. Bad data may include items that are not individually considered outliers.

Other Quality Factors. There are many other quality factors including reliability (site, sensor, communication network), timeliness of data, imprecision of data, and imprecision of metadata.

IV. DEFINITIONS AND APPROACH

A. General Definitions

An individual site refers to a fixed-location facility that houses one or multiple sensors that measure conditions. A measurement and associated metadata are referred to as an observation. The set of all sites, represented by S, is the set of sites for which observations are available for a time period and geographic area of interest.

An observation, obs, is represented as a 4-tuple, $obs = \langle s, t, l, v \rangle = \langle obs_s, obs_t, obs_l, obs_v \rangle$ consisting of the site/sensor s, timestamp t, location l (spatial coordinates), and an observed value v. We investigate observations from a single sensor type, so we assume that s identifies both the site and sensor. The set of all observations, represented by O, consists of observations from sites in S over a time-period of interest.

Ground-truth is the exact value of the condition that a given sensor is intended to measure at a given location and time. Ground-truth will rarely be known because of sensor error, estimation error, and high human costs, among other reasons. Human cost is a huge challenge, with agencies struggling to accurately inventory assets and technicians unable to service and maintain all equipment, including situations where they may not even be able to find the equipment.

We wish to evaluate observations to determine if they are erroneous. To do so, we compare observations to estimates of ground-truth. For our purposes, these estimates will be determined via interpolation, which is commonly used in the GIS community, as well as in the weather and road-weather communities.

B. Approach

Identification of Outlyingness and Outliers. We measure outlyingness as the absolute deviation between an observed value and ground truth. Ground truth may not be known, so we estimate outlyingness as the absolute deviation between an observation and modeled ground truth corresponding to the observed value in time and location. Given the degree of outlyingess (exact or estimated), we identify outliers using a threshold. If the degree of outlyingness for an observation meets or exceeds the threshold, then we flag the observation as an outlier. Otherwise, we flag it as an inlier. The degree of outlyingness is more informative than an outlier/inlier label.

Our approach is consistent with general model-based approaches for outlier detection found in Han et al. [7], Tan, Steinbach and Kumar [10] and Aggarwal [11], and follows the general data-mining framework of Train, Test and Evaluate.

C. Interpolation to Model Ground Truth

IDW estimates ground truth as the weighted average of observation values using (geographic) distance from the site for which an observation is to be estimated as the weight, raised to some exponent h. If ground truth is known, a suitable exponent h can be determined to minimize error. Isaaks and Srivastava [41] indicate that if h=0, then the estimate becomes a simple average of all observations, and for large values of h, the estimate tends to the nearest neighboring observation(s). This simple version of IDW does not account for time, so it is assumed that observations fall in temporal proximity.

Least Squares Regression (LSR) estimates observed values using the coordinates of the sites. We only use *x*-*y* coordinates in our experiments for LSR. There could be benefit in using elevation and other variables including time. However, doing so compounds problems related to bad metadata, such as incorrect locations, bad timestamps and inaccurate elevations.

UK estimates observed values using the covariance between sites, the coordinates of the sites, and the observed values. In our experiments, we used a Gaussian covariance function of distance and estimated the related parameters to minimize error relative to ground-truth for our training data using data from the present time window. Refer to Huijbregts and Matheron [42] for further information on UK. We implemented a fitter/solver for the estimation of the covariance function parameters using the Gnu Scientific Library (GSL) non-linear optimization code [43]. Refer to Bohling [44] for additional covariance functions.

These methods can be applied using a restricted radius or a bounding box to alleviate computational challenges and to focus on local trends. Other interpolators could be applied in a similar manner. There are obvious risks in using interpolators. Outliers and erroneous values will have an adverse impact on interpolation, causing poor estimates. Lack of data in proximity to a point to be estimated can also result in a poor estimate. For these reasons, we developed our own robust interpolator in prior work.

D. Our SMART Approach

In prior work, we developed a representative approach for data quality assessment of site-based, spatio-temporal data using what we call Simple Mappings for Approximation and Regression of Time series (SMART) [26-32]. We used the SMART mappings to identify bad (inaccurate) observations and "bad" sites/sensors, so that they can be excluded from display and computation, and to subsequently estimate (interpolate) ground truth.

Site-to-Site Mappings. Let an observation be represented as $obs = \{(t, v): t = time, v = value\}$, pairing the value with the reported time. Let obs_i be the set of observations from site i and obs_j be the set of observations from site j. For a given time radius r we pair the observations from sites i and j as $obs_{pairs_{i,j}} = \{(x, y): (t_1, x) \in obs_i, (t_2, y) \in obs_j, |t_2 - t_1| \leq r\}$. We then define a site-to-site mapping l as a linear function of the x-coordinate (the observed value from site *i*) of the paired observations $obs\_pairs_{i,j}: l_{i,j}(x) = a + bx$. We determine this function to minimize the squared error between the values of the function and the y-coordinates (the observed values from site j) for the paired observations.

We next determine a quadratic estimate q of the squared error of the linear mapping relative to the time offset between the paired observations. We expect an increased squared error for increased time differences. This model estimates the squared error and accounts for time offsets between observations. Our method does not require a complex, data-specific covariance model.

These simple mappings are the core elements of our approach, and we must overcome the potential impact of erroneous data in determining them. LSR suffers from sensitivity to outliers. We use the method from Rousseeuw and Van Driessen to perform Least Trimmed Squares Regression [45]. Least Trimmed Squares determines the least squares fit to a subset of the original data by iteratively removing data furthest from the fit. Before applying least trimmed squares to determine the linear mapping, we select the percentage of data that will be trimmed. We can interpret the trim percentage either as our willingness to accept bad data in our models or our estimate of how much data is bad. We used a trim percentage of 0.1 throughout.

For the quadratic error mappings, we experienced problems with local minima when attempting quadratic least trimmed squares. Instead we group data into intervals, determine the trimmed mean for each group, and then compute the least squares quadratic fit for the (*time difference, trimmed mean*) pairs.

We then check the coefficients and derived measures of the linear and quadratic mappings for outlying values relative to all other mappings. If we find outlying values, we flag the mapping as unusable. For instance, if the axis of symmetry of the quadratic error mapping is an outlier relative to that for another pairing, then there may be a problem with the timestamps of at least one of the two sites.

SMART Interpolator. Our SMART interpolator uses these mappings. Formally: Let S be the set of all sites. Let $s \in S$ be a site for which we are evaluating observations. Let $\langle s_1, \ldots, s_n | s_i \in S, s_i \neq s \rangle$ be the set of sites other than site s. We want to estimate $obs_s(t_s)$, the value of the observation at site s at time t_s using the most recent observations from the other sites relative to time t: (t_i, v_i).

Our SMART interpolator is like IDW, using our quadratic error estimates instead of distance given the time lag between observations and using our SMART linear mappings to yield estimated ground truth producing an estimate. Neither distance nor direction are directly used. The linear mappings and quadratic error estimates account for similarity between sites. No attempt is made to down-weight clustered sites, although there may be benefit in doing so.

We determine the exponent g by minimizing error relative to ground truth, if available, or estimated ground truth. Prior to computing the weighted estimate, we examine the weights and, if necessary, "re-balance" to reduce the potential influence of single sites on the outcome. We found it useful to restrict the maximum relative weight a site can be given to 0.25 to reduce the risk that a bad value from one site will overly influence the resulting average. Rather than take a simple weighted average, we use a trimmed mean to further reduce the influence of outliers.

E. Artificial Data Set

We developed a weather-like phenomenon representing temperature as approximate fractal surfaces produced using the method of Successive Random Addition. For further information on Successive Random Addition, refer to Voss [46], Feder [47], and Barnsley et al. [48]. Fractional Brownian processes were used by Goodchild and Gopal to generate random fields representing mean annual temperature and annual precipitation for the purpose of investigating error in [49]. We used a similar approach to model time series in [50]. A 513x513 approximate fractal surface, $surface(x, y)$, was generated with Hurst Exponent H=0.7 and $\sigma^2 = 1.0$, representing elevation. A 1025x513x513 fractal-like weather pattern, $weather(x, y, t)$, was also generated with Hurst Exponent H=0.7 and $\sigma^2 = 1.0$. The larger x-coordinate allowed us to simulate motion/flow. We generated one surface and eight weather patterns, allowing us to train on one weather pattern and test on those remaining.

We generated time series of "ground truth" data by combining the surface data with the weather data, a periodic effect and a north-south effect to simulate a weather-like phenomenon like the diurnal effect and general north-south variation in the Northern Hemisphere respectively. We added the weather data as is, with varying offsets in the x-coordinate to represent a west to east flow in the weather pattern. The surface value is subtracted so that low points are "warmer" than high points. The periodic effect represents warming during the day and cooling at night. The north-south effect yields warmer points to the south and cooler points to the "north". Our approach yields a time series of length n=513 for each (x, y) on the 513x513 surface.

We selected 250 "sites" using random uniform x-y (spatial) coordinates. For each site we assigned a reporting pattern with a random frequency and offset. We added errors to the observations from 25 sites via: random noise added to ground truth (NOISE), rounding of ground truth (ROUNDING), replacement of ground truth with a constant value (CONSTANT), replacement with random bad values with varying probabilities (RANDOMBAD), or negation of ground truth. The remaining 225 sites were left error-free.

V. EVALUATION

We evaluated the performance of the various interpolators including our SMART Method in-depth, in terms of computation and ability to identify bad data. We compared our SMART method, IDW, LSR, UK and OK. We measured performance and scalability using run-time in milliseconds. We measured accuracy using mean-squared-error (MSE) between estimated and known ground-truth. We compared means using t-tests when multiple runs were available. We used Area Under the ROC Curve (AUROC) analysis to evaluate accuracy of outlier classification given varying "threshold" values for outlier/inlier determination.

We analyzed our artificial data set, MADIS air temperature for Northern California from December 2015, MADIS air temperature for Montana from January 2017, and Average Daily USGS Streamflow for Montana from 2015, 2016, 2017.

A. Evaluation Using our Artificial Data Set

We performed an in-depth comparison of the various algorithms using our artificial data set. We enhanced the standard algorithms by randomly choosing neighboring sites using set inclusion percentages (0.1, 0.2, 0.3, …, 0.9, 1.0). For instance, a 0.9 inclusion percentage corresponds to selecting neighboring sites individually with 0.9 inclusion / 0.1 exclusion probability. We varied the radius (50, 75, 100, …, 175, 200) over which sites were included relative to the location of the site whose observation we were testing. We repeated this procedure 10 times for each parameter combination (inclusion percent and radius) and used the median of the resulting estimates as the estimate for that parameter combination. By randomly holding out sites, bad data will be held out in some of the resulting combinations. By taking the median of the results, we eliminate the extreme estimates, particularly those impacted by bad data, and ideally determine a robust estimate.

We ran the methods in aggregate over the eight time periods spanning 512 time units. For each time period, there were 37,293 observations total from the 250 sites. We iterated through the observations in order by time and estimated ground truth for each observation as if computing in real time as the observations become known. Only observations that occurred at the same time as or prior to each observation were used for prediction, simulating real-time operation of the system. We averaged the MSE and run time for each configuration (inclusion radius and inclusion percent).

We compared the results of the various runs of the methods. The run time for the SMART method was 6336.6 ms, and the MSE was 0.1026. The SMART method was comparable in run time to IDW, but the accuracy achieved was far better than for any of the other methods.

We measured the ability of each method to distinguish increasing percentages of the bad data from good data using an AUROC analysis. True outliers were defined as data that differs from ground-truth – i.e., data that was modified to be erroneous. Predicted outliers were data that differed from estimated ground truth by a given threshold. We varied thresholds for outlier/inlier cutoffs and compared results with the actual labels identifying whether the data was truly an outlier or inlier. The AUROC (area under the ROC curve) values are shown in Table I. The AUROC values show better discriminative power for the SMART method versus the other methods. No method will be perfect in identifying all errors. Some errors are small and impossible to distinguish from interpolation error. Known ground truth and known error from ground truth yields perfect labels.

TABLE I. AUROC VALUES FOR ARTIFICIAL DATASET

Method	SMART	UK	LSR	IDW
AUROC	0.827	0.740	0.739	0.708

Our SMART method's computation time is comparable to IDW and is far better than LSR and UK, but we still should account for the preprocessing computation time required for determining the linear mappings and quadratic error functions. The overall amount of preprocessing time required to determine the linear mappings and quadratic error functions was comparable to run time required for UK. This was encouraging. Generation of the mappings will be done as an offline, batch process, so the observed time required is still within reason to help facilitate the faster and more accurate, online process. Additional benefits, such as identification of bad sites and bad metadata, come from these mappings, further justifying the effort required. Optimization can reduce the overall time needed to compute the mappings. The benefits and potential to improve the run time outweigh the amount of required preprocessing time.

B. December 2015 MADIS California Data

We analyzed Northern California December 2015 ambient air temperature data from the MADIS Mesonet subset. We used a bounding box defined by $38.5° \leq latitude \leq 42.5°$ and $-124.5° \leq longitude \leq -119.5°$, yielding 888 sites. We excluded observations that failed the MADIS Level 1 Quality Control Check. This range check restricts observations in degrees Fahrenheit to the interval [-60°F,130°F]. Many values failing this check fall far outside the range and can have a dramatic impact on the interpolation methods. Our SMART method performs very well in the presence of extreme bad data, and it would have easily out-performed the other methods in the presence of the range-check failed data.

There were over 2 million observations. MADIS flagged 73.5% of these observations as "verified" / V, slightly less than 4% as "questioned" / Q, and 22.5% as "screened" / S, indicating that it had passed the MADIS Level 1 and Level 2 quality checks, but that the Level 3 quality checks had not been applied.

Training. Verified (V) observations from the first week in December 2015 were used to train all methods, including our SMART method. In the absence of range-failed data, the "enhanced" (iterated subset) versions of the other algorithms showed little improvement in accuracy while consuming excessive computation time, particularly "enhanced" UK. In some cases, it would have taken days to compute results. Because of this, we used the methods directly, without enhancement. We also tested OK (refer to Bailey and Gatrell [51] for further information). Since we do not know "ground truth" for this data, the verified data is the closest to ground truth. We trained all methods on this data to MSE of predicted versus actual. We used a 50-mile inclusion radius due to the density of sites to avoid excessive computation time for the kriging approaches.

The SMART mapping coefficients and derived values were examined for outliers, and ranges were determined for valid mappings. If any coefficient or derived value for a given SMART mapping fell outside these ranges, then the SMART mapping was considered bad, and that mapping was not used for predictions.

Our SMART method produced significantly better results than all other methods for the training data in terms of estimation of ground truth measured by MSE, as shown in Table II. A paired, one-sided t-test was used for significance testing using paired squared errors from predicted values. Only the verified (V) data was used in this comparison since it best approximates ground truth. The SMART method was compared pairwise with the other methods and results were aggregated over instances where both methods produced predictions.

TABLE II. MSE FOR MADIS CALIFORNIA TRAINING DATA

Method	**MSE**	**Method**	**MSE**
SMART	2.8322	IDW	7.6212
SMART	2.8322	LSR	17.1446
SMART	2.8046	OK	18.4989
SMART	2.8046	UK	16.5289

Testing. Testing was conducted using all data from the entire month of December 2015, minus the range-check-failed data. We computed the MSE for the verified (V) data since it best represents ground truth, but all observations were used in making estimates. The testing results indicate the robustness of methods in the presence of bad data. In comparisons across all other methods, the SMART method significantly out-performed all other methods in terms of MSE, as shown in Table III.

We conducted an AUROC analysis to compare classification ability of the methods based on the MADIS quality control flags. We considered the following flags from MADIS to be good/inlier data: V/verified, S/screened,

good. The Q/questioned, was treated as bad/outlier data. Recall that we excluded the observations having a QC flag of X, those that failed the range test, from our evaluation. Even if we accept the MADIS quality control flags as being correct, and we do not, this approach is problematic. The MADIS QC flag S corresponds to data for which not all the QC checks have been run. While this data had not failed any quality control checks that have been applied, it possibly would have failed the higher-level checks.

In terms of AUROC, IDW, LSR and SMART were comparable, with IDW finishing slightly ahead, as shown in Table IV. While these AUROC values seem reasonable, they are affected by incorrect outlier/inlier labels, and our SMART method suffers the greatest impact because the distance-based methods approximate the MADIS Level 3 quality control check. OK and UK fall short because they fail to make predictions for many observations.

TABLE III. MSE FOR MADIS CALIFORNIA TESTING DATA

Method	MSE	Method	MSE
SMART	4.4611	IDW	9.1306
SMART	4.4611	LSR	16.5223
SMART	4.3360	OK	16.0868
SMART	4.3360	UK	14.2086

TABLE IV. AUROC FOR MADIS CALIFORNIA TESTING DATA

	AUROC
IDW	0.7906
LSR	0.7578
SMART	0.7317
OK	0.6458
UK	0.6062

C. December 2017 MADIS Montana Data

We investigated ambient air temperature for Western Montana / Northern Idaho from the MADIS Mesonet and the MADIS HFMetar subset in January 2017. We added the HFMetar data set to account for aviation AWOS/ASOS sites that had previously been included in the Mesonet data set. We used a bounding box defined by $44° \leq latitude \leq 49°$ and $-116° \leq longitude \leq -110°$, resulting in observations from 497 sites. This bounding box is comparable in size to the one used for Northern California, although the density of sites is less. We excluded observations that failed the MADIS Level 1 Quality Control Check.

All total there were over 1 million observations. MADIS flagged 71.2% of these observations as "verified" / V; 10.3% of as "screened" / S, indicating that they had passed the MADIS Level 1 and Level 2 quality checks, but that the Level 3 quality checks had not been applied; and a relatively large 18.5% of the data as "questioned" / Q. This is over four times the percentage of questioned data as there was for the California data set.

Training. Verified (V) observations from the first week in January 2017 were used to train all methods, including our SMART method. We used a 100-mile inclusion radius due to a low density of the Montana/Idaho sites. The SMART mapping coefficients and derived values were examined for outliers, and bad mappings were identified as any mapping associated with such values. The quality of the mappings as measured by MSE was noticeably less than that for the Northern California data set. We found problems with many of the timestamps in this data set. Recognizing that much of the Idaho data comes from the Pacific Time Zone while the Montana data comes from the Mountain Time Zone, there appeared to be many sites for which the conversion to UTC time was not consistent. The Northern California data all falls within Pacific Time, and we did not see this problem in that data set. In terms of MSE, the SMART method produced significantly better results than each of the other methods for the training data, as shown in Table V.

TABLE V. MSE FOR MADIS MONTANA TRAINING DATA

Method	MSE	Method	MSE
SMART	8.1513	IDW	16.7217
SMART	8.1513	LSR	29.8039
SMART	10.6726	OK	47.0028
SMART	10.6726	UK	33.9863

Testing. Testing was conducted using data from the remainder of January 2017. All data was used for this test except for the observations that failed the MADIS Level 1 range test. The SMART method significantly outperformed all other methods in terms of MSE, as shown in Table VI.

TABLE VI. MSE FOR MADIS MONTANA TESTING DATA

Method	MSE	Method	MSE
SMART	21.4714	IDW	38.3208
SMART	21.4714	LSR	38.5063
SMART	23.1496	OK	50.5078
SMART	23.1496	UK	36.6762

We conducted an AUROC analysis to test classification ability based on the MADIS quality control flags in the same way as described for the Northern California data set in the previous section. As noted in that section, many of the MADIS QC flags are incorrect. In terms of Area Under the ROC curve, LSR, IDW and SMART were comparable, with LSR finishing ahead, as shown in Table VII. These AUROC values are less than those for the Northern California data set at least in part because all methods adversely affected by incorrect outlier/inlier labels.

TABLE VII. AUROC FOR MADIS MONTANA TESTING DATA

Method	LSR	IDW	SMART	OK	UK
AUROC	0.6900	0.6697	0.6393	0.5432	0.5476

This data set includes a large percentage of observations (18.5%) that are flagged as "questionable" by MADIS. These were considered "bad" / outliers for the purposes of our analysis. It also includes a large percentage (10.3%) that are flagged as "screened" by MADIS, indicating that not all QC checks have been conducted. These are considered "good" / inliers for our analysis.

There were many observations flagged as "questionable" / outliers in the HFMetar subset that should have been

flagged as "good" / inliers. This data alone accounts for most of the questionable data in the data set. Aviation weather sites are well-maintained and regularly calibrated, so it is hard to believe that these sites would produce data that is entirely bad. We checked this data against predicted values, as well as neighboring sites, and it was very close, so it is unclear why the data was labeled as questionable.

Numerous sites were flagged by our SMART method as "bad" and all observations from those sites were labeled as bad. MADIS flagged some observations from these sites as good when they were close to predicted values. In some cases, this may have been reasonable, but in others it was a random occurrence. There were some sites that produced bad data for the training period but then produced good data for at least a portion of the test period. One could argue that for such sites all associated observations should be questioned. If a site was identified as bad by the SMART method, then the V and S observations would adversely impact the SMART method in the AUROC analysis. The chance situations in which the other methods came close to the "good" values and far from the "bad" values improved their performance.

D. December 2015-2017 USGS Streamflow Data

Mean daily streamflow (ft^3/sec) was downloaded for all sites in Montana from the USGS [52] for every day from January 1st, 2015 through April 24th, 2017. There were 145 sites having data than spanned this period, and these sites were analyzed. This data set is far different from the air temperature data used for prior analysis. Since daily averages were used, there is no visible diurnal effect. There is a seasonal effect which varies with elevation and location relative to watersheds. Due to the dramatic fluctuations that occur in this data during times of peak runoff, the base-10 logarithm of the data was used for analysis.

This data set includes quality flags. Daily values are flagged as "A", approved for publication, and "P", provisional and subject to revision. Values may further be flagged as "e" for estimated. Values transition from provisional to approved after more extensive testing is conducted, so provisional values aren't necessarily bad. These flags were of limited use to us and we did not use them for analysis. We treated the data as being all good and subsequently introduced errors into some of the observations, making them known bad. There were 122,380 total observations.

Training. All data from 2015 was used to train all methods, including our SMART method. We assume this data, which was mostly "approved", to be ground truth. We trained over this data to minimize MSE of predicted versus actual. We used a 200-mile inclusion radius. The SMART mapping coefficients and derived values were examined for outliers. If any coefficient or derived value for a given SMART mapping was an outlier, then the SMART mapping was considered bad, and it wasn't used for predictions. In terms of MSE, the SMART method produced significantly better results than the other methods for the training data, as shown in Table VIII.

Testing. Testing was conducted using the 2016-2017 data. The SMART method significantly outperformed all other methods in terms of MSE, as shown in Table IX.

TABLE VIII. MSE FOR USGS TRAINING DATA

Method	MSE	Method	MSE
SMART	0.0174	IDW	0.8751
SMART	0.0174	LSR	0.9611
SMART	0.0174	OK	0.9431
SMART	0.0174	UK	0.9617

TABLE IX. MSE FOR USGS TESTING DATA (NO ERRORS)

Method	MSE	Method	MSE
SMART	0.0429	IDW	0.9031
SMART	0.0429	LSR	0.9869
SMART	0.0429	OK	0.9755
SMART	0.0429	UK	0.9874

Testing was then conducted using the 2016-2017 data, with errors introduced into 10% of the observations. A random normal value with mean zero and standard deviation one was added to each of the observations in the 10% group. The MSE was computed relative to the known, original observations which represent ground truth, and all observations (including bad observations) were used in making estimates. The testing results help to indicate the robustness of methods in the presence of bad data. The SMART method significantly outperformed all other methods in terms of MSE, as shown in Table X.

TABLE X. MSE FOR USGS TESTING DATA (WITH ERRORS)

Method	MSE	Method	MSE
SMART	0.0453	IDW	0.9103
SMART	0.0453	LSR	0.9907
SMART	0.0453	OK	0.9776
SMART	0.0453	UK	0.9914

We conducted an AUROC analysis to test the methods on classification ability based on whether observations had been altered to be erroneous by our process of randomly selecting 10% of the observations and adding a normal random variable with mean 0 and standard deviation 1 to those observations. The altered observations were labeled "bad"/outlier and the unaltered observations were labeled as "good"/inlier. Our SMART method performed far better than all the other methods, achieving an AUROC value of 0.8722, as shown in Table XI. The other methods had values between 0.6 and 0.63.

TABLE XI. AUROC VALUES FOR USGS TESTING DATA

Method	SMART	IDW	OK	UK	LSR
AUROC	0.8722	0.6241	0.6136	0.6046	0.6031

E. Evaluation Summary

For all four data sets and for every training and testing instance compared, our SMART method performed significantly better in terms of accuracy (MSE) than all other methods. Its computational performance was competitive

even though no effort was made to optimize it. For the two MADIS data sets, its performance for AUROC analysis of classification and discrimination capability showed it to be competitive with the best of the other methods. This comparison and evaluation made use of MADIS data quality labels for which we have found numerous problems. As such, all methods underperformed, and the SMART method was penalized most by mislabeling. For the other two data sets (artificial and USGS) in which ground truth is known or assumed and errors were introduced relative to ground truth, the SMART method outperformed the other methods by a wide margin. This further supports our assertions regarding the impact of bad labels on the MADIS data, and the need for better methods and benchmark data sets for data quality assessment.

OK and UK both failed to produce estimates for many observations, likely due to singular matrices. They were not competitive in terms of run time and their accuracy was no better than the other methods. UK and LSR are prone to occasional very large errors if the predicted surface slopes in an extreme manner.

Our SMART method identifies "bad sites" that chronically produce bad data and does not use data from these sites in estimating ground truth for other sites. Similarly, data from these "bad sites" is labeled as all bad. The SMART method falls short in cases where a site exhibits chronic behavior during training but recovers to produce good data during a testing period.

The USGS streamflow data exhibits correlation between sites, but the correlation corresponds to sites close to each other and in the same river/stream. Correlation will not necessarily be high for sites that are close but in different rivers. For rivers that have dams and other features that may influence streamflow in unusual ways, sensors will be correlated on each side of such features, but not as much on opposite sites, and certainly not as much with sites on rivers that do not have similar features.

The SMART method identifies like sites, yielding better correlations. IDW and LSR will not perform well in this circumstance. And, the kriging methods will not perform well either if a stationary, isotropic covariance function is used. Such an assumption is typical, and we used this assumption in determining the covariance matrices for the kriging tests.

VI. Analysis of Mislocated Sites

Bad metadata presents significant challenges regarding data quality assessment. As discussed earlier in this paper, many if not most techniques for spatial-temporal quality assessment depend on distance and time directly and/or related assumptions. If timestamps on some data are incorrect, then that data will be compared against data from other time periods, times in which conditions may be dramatically different. If location metadata is incorrect, specifically if a site is mis-located, then it may be compared against sites that are far away, and quality assessment will suffer. Conditions may vary dramatically by location, even if timestamps are correct. This causes a chicken-egg problem: to assess the quality of data, we need quality data for which the quality has been determined via quality assessment. How do we find and filter the bad data such as that with bad location metadata amid other data quality issues? In this section, we present a new analysis of this problem.

A. Gibson near Castella

For several years, we were aware of a mis-located site in the MADIS feed that was displayed in our WeatherShare system. WeatherShare users identified this site as mis-located in correspondence with us. Site GISC1 (Gibson near Castella) was mis-located at latitude 38.56556° N, longitude -121.485° W in downtown Sacramento prior to a correction that was made sometime in 2016. Note that MADIS obtained data for GISC1 from the National Weather Service's Hydrometeorological Automated Data System (HADS) system. Subsequently the location of GISC1 was corrected in the MADIS feed to latitude 41.022° N, longitude -122.399° W, 175 miles to the north near the Caltrans Gibson Maintenance yard and near the town of Castella. This relocation makes sense given the name of the site and the error reports from our WeatherShare users. There was no apparent indication why/how this site was mis-located, and we are unsure of how it was relocated. It may have been initially assigned the coordinates of another site.

We didn't have a mechanism for dealing with issues like this. We could have taken our users' word and manually relocated the site within our system by changing the latitude and longitude to what they reported. But, how could we confirm they were correct? And, if we manually changed the location, then we would also need a mechanism to detect if the location was subsequently modified in the feed, perhaps giving a better indication of the true location. We could contact the provider and ask them to correct the situation, but they too may not know the true location of the site. Even in instances where errors were known, it has taken providers years to address data quality issues we have informed them of. Instead, we chose to suppress the display of this site, and implemented a mechanism allowing us to manually select and suppress the display of any site. Of course, this out of sight (no pun intended) out of mind approach was not perfect either because again, the location of the site might be corrected in the feed at some point. Subsequently, we made a choice to again display all data with the caveat that users would have to decide for themselves what data was good and what was bad. That approach is less than ideal and could give the perception of poor quality of a system overall.

At the same time, we wondered if the mis-location of GISC1 was isolated or if there were more mis-located sites. Given our experience with other data quality issues in this data set, we suspected the latter. But the reality was that we had no way of knowing for sure in the absence of working directly with owners and operators of the equipment in the field and/or conducting site visits. That was well outside the scope of our work and would otherwise have been cost-prohibitive and infeasible. A more practical question was whether we could find such errors automatically. We have

spent significant time since then investigating this challenging and interesting problem.

In this section, we present a new analysis of MADIS ambient air temperature data from sites located in Northern California between 2014 and 2019. We look in retrospect at the 2014 data to see what more we could have done at that time to identify such problems and to assess the state of the problem at that time. We look at temperatures from the month of January, as considerable variation in temperature occurs in that month in conjunction with the bad weather season in Northern California. We selected sites that reported at least once within 90% of the 15-minute intervals during the given month. Overall there were 575 sites that met these criteria for 2014. We used all data regardless of MADIS quality control assessment.

B. Relocated Sites in the MADIS Dataset

One possible indication of erroneous location metadata is the subsequent revision of site locations in the feed. We investigated the 2014 data versus the subsequent 2015 through 2019 data for each site and identified all changes in location. We found that 5 sites were subsequently relocated by more than 20 miles from the locations provided in the 2014 data. GISC1 was relocated furthest at 176.6 miles, matching our earlier observations. Another site, KLHM, the Lincoln Regional Airport, was relocated 37.6 miles from its 2014 location. The three other sites were unfamiliar to us and trying to confirm their true locations could be challenging. See Table X.

TABLE X. 2014 SITES RELOCATED 20 MILES OR GREATER

Site	Year of Change	Distance in miles from Original Location
GISC1	2017	176.6
TS389	2018	60.0
KLHM	2018	37.6
TT109	2016	33.2
SNWC1	2016	27.7

Overall, 109 of the 575 sites were relocated at least once between 2015 and 2019 relative to their 2014 locations. While this may seem like a large proportion, and it is, most of the changes were relatively small. 81 sites were relocated less than a mile from their 2014 locations. Perhaps greater precision was used in specifying their locations: for instance, using two or more digits beyond the decimal for specification of latitude and longitude versus one digit. That could account for changes in location of several miles, and there were 93 sites overall that were relocated 5 miles or less from their 2014 locations. That leaves only 10 sites relocated by between 5 and 20 miles, plus the five sites shown in Table X that were relocated by 20 miles or greater. See Table XI. As such, and if the relocations are correct, we could say that 15 of the 575 sites were mis-located in the 2014 data feed, but we truly do not know. In fact, we found some sites that were relocated multiple times including one site, TR180, that was relocated in 2018 to a point 24.5 miles from its 2014 location and then relocated again in 2019 to a point 0.015 miles from its 2014 location. It was moved some distance away and then move back to almost the same original location. Which location, if any of the three, was right?

TABLE XI. 2014 SITE RELOCATIONS COUNTS BY DISTANCE

Range	Count
0 to 1	81
1 to 2	6
2 to 5	7
5 to 10	2
10 to 20	8
20 to 50	3
50 to 100	1
100+	1
TOTAL	**109**

C. Identifying Mis-located Sites with our SMART Method

As discussed earlier, our SMART mappings can provide a robust, consistent measure of dissimilarity between sites in the presence of bad data. We did not filter the original data other than by time and location. We did not use the MADIS quality control flags to filter at all. Bad data is certainly included. Because of this, measures of correlation or covariance would be adversely affected by bad data. We have found that the mean-squared-error (MSE) of the SMART mappings provides a robust alternative for measuring dissimilarity. We also expect to find, in general, that near sites are more closely related than far sites (Tobler's First Law of Geography). This relationship can be exploited to identify mis-located sites – at least those that are severely mis-located.

Before proceeding, we needed to be cognizant of the challenges presented earlier in this paper, as they would certainly have an impact on the results. In order to overcome some of these challenges, particularly bad data and challenges with non-uniform time reporting, we chose somewhat loose parameters for our SMART mappings: a time radius of 90 minutes for pairing observations, a 10% cutoff for trimmed least squares regression, and a 10% trim percentage for computing trimmed means. While these choices help to overcome the stated challenges, they may also blur the relationships between sites, which can cause challenges in the presence of non-isotropic covariance. In future work we plan to investigate parameter selection further.

Now we look at the relationship between distance and MSE of SMART mapping from other sites to GISC1 using the 2014 data, including the incorrect location of GISC1 in the 2014 data. Again, we expect near sites to be more closely related (low MSE) than far sites. But we find instead that sites falling over 150 miles away have the lowest MSE for the SMART mappings. See Figure 1.

Next, we look at the same plot for the 2014 GISC1 data, but with the location corrected to the 2017 location. See Figure 2. In this figure, sites nearest the corrected location have the lowest MSE values. And, there appears to be an apparent, positive trend in which MSE increases by distance. But there is also a lot of variation. This variation is

most likely attributable to non-isotropic covariance and bad data. Tightening the parameters used for the SMART mappings might help to reduce this variation, and we intend to investigate this in future research. Regardless, the plot does appear to confirm that GISC1 was relocated to the corrected location or at least near the correct location.

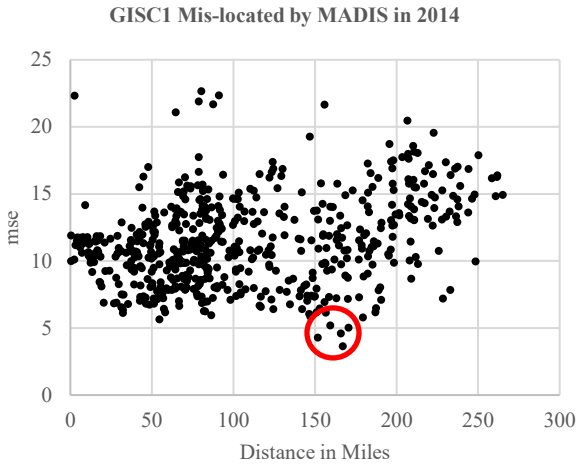

Figure 1: GISC1 (Incorrect Location) Distance versus MSE of SMART Mapping by Site

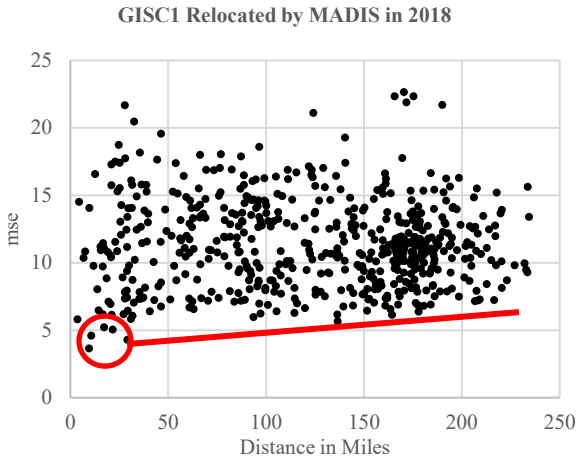

Figure 2: GISC1 (Corrected Location) Distance versus MSE of SMART Mapping by Site

Next, we look at site KLHM, the Lincoln Regional Airport. Recall that the Lincoln Regional Airport site was relocated subsequent to 2014 by 37.6 miles in the MADIS feed. In plots of distance versus MSE, we checked to see if the same relationships hold. See Figure 3 and Figure 4. Again, we see that the sites having the lowest MSE for the SMART mappings fall approximately the same distance from the mis-location as the subsequent re-location. And, we see a more prominent positive trend when the site is relocated. In this case, it is easier to confirm the correct location since the location of the Lincoln Regional Airport is known. Still, an airport occupies a lot of space, and the exact location of the weather sensor at the airport is not known to us with certainty.

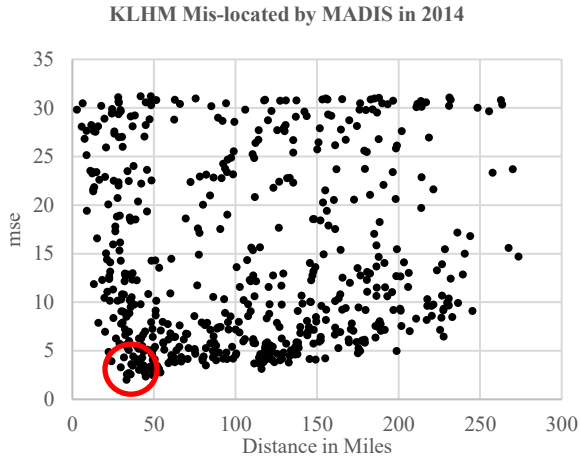

Figure 3: KLHM (Incorrect Location) Distance versus MSE of SMART Mapping by Site

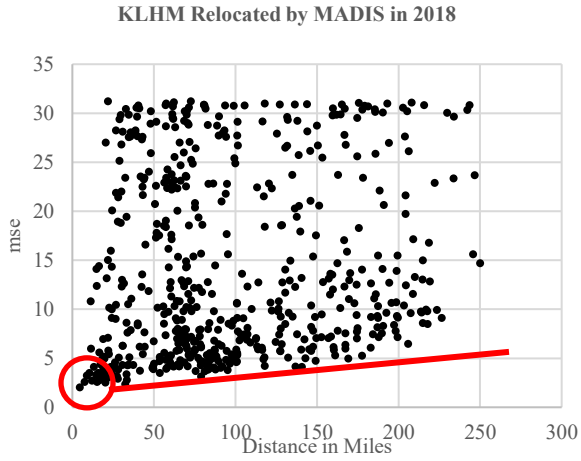

Figure 4: KLHM (Corrected Location) Distance versus MSE of SMART Mapping by Site

Before proceeding to develop a more formal method to identify situations where sites may be mis-located like GISC1 and KLHM in the 2014 MADIS data, we will artificially mis-locate several sites for which we know the correct locations, and we will see if these relationships hold. For this we will use the weather station at the Redding Airport, KRDD, and the weather station at Sacramento International Airport, KSMF. Both sites do appear to be correctly located in the 2014 MADIS data with the caveat that their precise location at each airport is unknown to us and the specification of location may lack some precision.

For the Redding Airport, KRDD, we artificially relocate the site 130 miles to the south. We see the same signature patterns in the corresponding plots. See Figure 5 and Figure

6. When correctly located, sites having the lowest MSE for the SMART mappings fall closest to the site and there is a general upward trend in the data. When the site is incorrectly located, sites have the lowest MSE for the SMART mappings fall at a distance corresponding to the distance between the correct and incorrect locations.

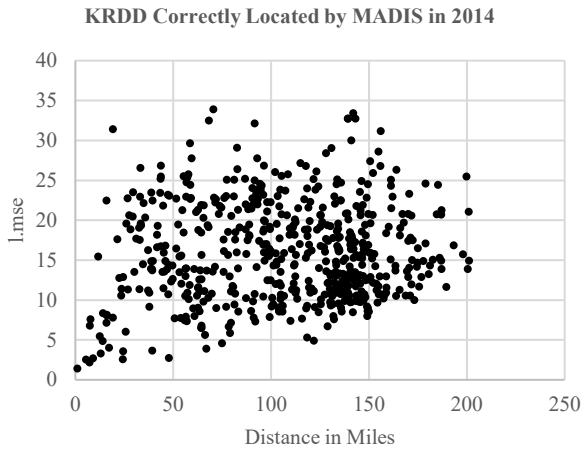

Figure 5: KRDD (Correct Location) Distance versus MSE of SMART Mapping by Site

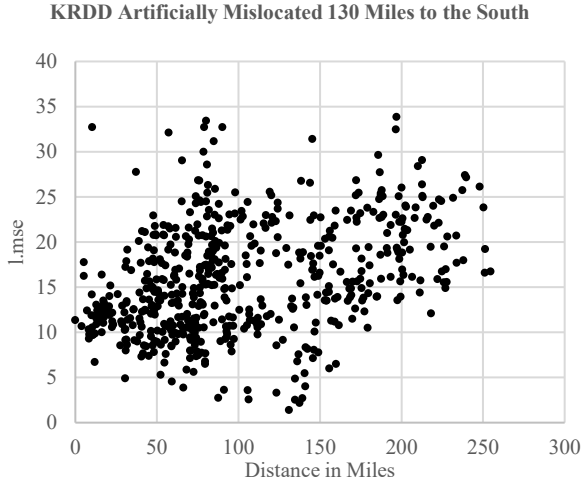

Figure 6: KRDD (Incorrect Location 130 miles south) Distance versus MSE of SMART Mapping by Site

For the Sacramento International Airport, KSMF, we artificially relocate the site 175 miles to the north. (This is similar but in the opposite direction of the mis-location of GISC1.) Again we see the signature patterns in Figure 7 and Figure 8.

Site TS389 was relocated in the 2018 MADIS feed 60 miles from its 2014 location. Given the evidence we presented above, it should be apparent from similar plots if this relocation was correct. As we just saw with the artificial relocations of KRDD and KSMF, we can also spot situations in which a site was incorrectly re-located. This may be the case with site TS389. See Figure 9 and Figure 10. It appears that the 2014 location was correct. And, it appears that the 2018 relocation of the site was incorrect. Realize though that these plots show relationships for 2014 data. It could be the case that this site truly was moved in 2018 or that the old site ceased operation and a new site was given the same name. This may seem strange, but given vague naming on some of these sites, it is possible. While not the case here, there is certainly the potential for "mobile" site data to be incorporated into the feed, in which a portable sensor suite is moved from location to location as needed. Further challenges would occur in assessing constantly moving sites such as vehicles equipped with weather sensors. Such data is being collected by department of transportations and others, and that data is being incorporated into data sets such as MADIS.

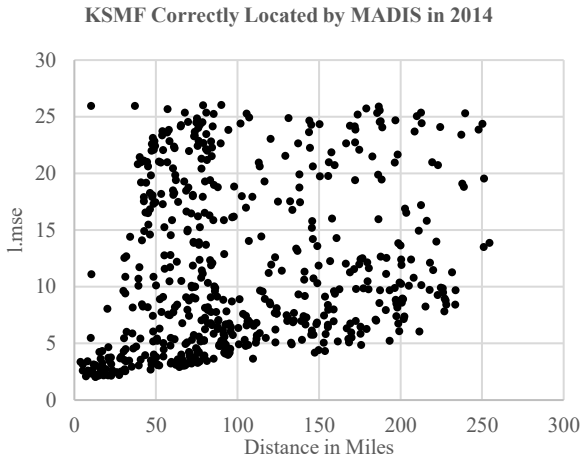

Figure 7: KSMF (Correct Location) Distance versus MSE of SMART Mapping by Site

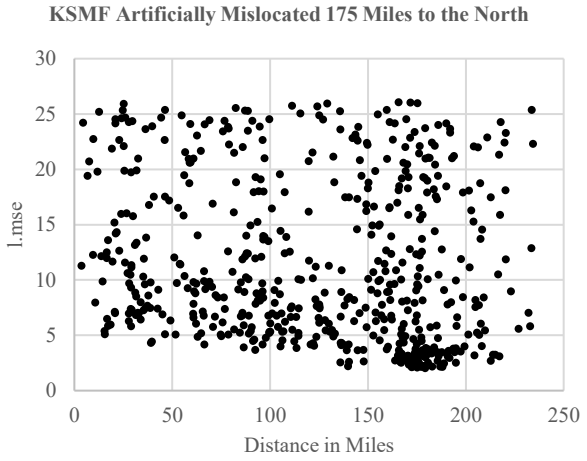

Figure 8: KRDD (Incorrect Location 175 miles north) Distance versus MSE of SMART Mapping by Site

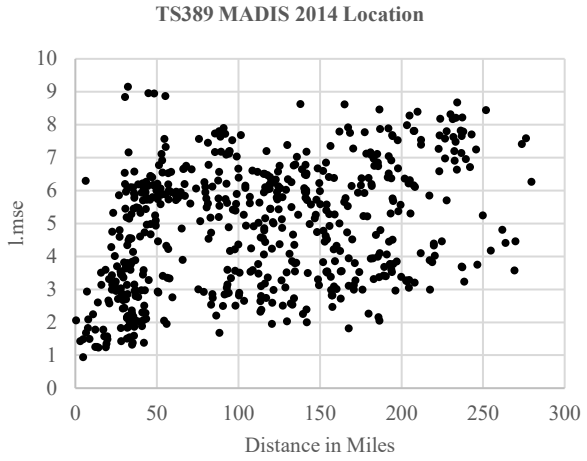

Figure 9: TS389 (Likely Correct Location) Distance versus MSE of SMART Mapping by Site

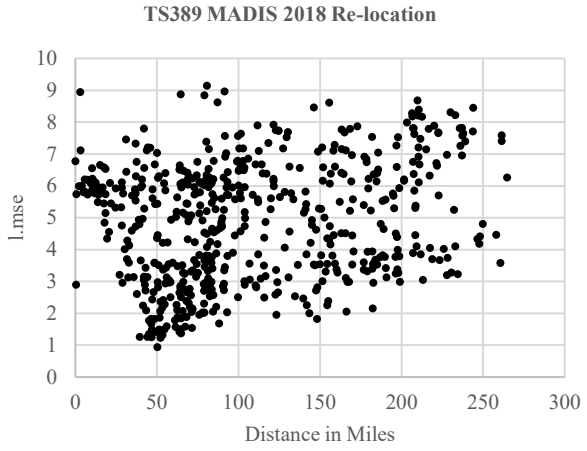

Figure 10: TS389 (Likely Incorrect 2018 Location for 2014) Distance versus MSE of SMART Mapping by Site

Site TT109 was relocated by MADIS in 2016 to a point 33.2 miles from its 2014 location. The plots of the 2014 data relative to these locations do appear to show that the relocation was correct. We omit the plots here for the sake of brevity. However, the plots for SNWC1, which was relocated by MADIS in 2016 to a point 27.7 miles from its 2014 location, are inconclusive. It may be the case that this site has other issues, including chronically bad data, that make it difficult to compare against neighbors.

While these signature patterns seem apparent visually in the plots in most cases, they can be challenging to identify in an automated fashion because of error and variation. We tried a few approaches including using the slope of the best-fit regression line, Spearman's Correlation Coefficient for the ranks of distances versus MSE and found that neither provided a reliable mechanism for identifying situations like those shown above. Instead, we developed a nearest neighbor approach that shows promise.

Let D_k be the set of nearest k neighbors to the site in question by distance. Let L_k be the set of nearest k neighbors by MSE from SMART mappings to the site in question. Then let $J_k = |D_k \cap L_k|/|D_k \cup L_k|$. This is the Jaccard Index, which measures the amount of overlap in the two sets. In our situation, it measures the amount of overlap between the nearest sites in terms of distance and the nearest sites in terms of MSE for SMART mappings. Intuitively, sites that are correctly located should have a high Jaccard Index and sites that are incorrectly located should have a low Jaccard Index. But the selection of *k*, the number of neighbors, could be tricky in the presence of errors and variation. For this reason, we compute $M_L = \max(J_k)$ for $k \in \{5, 10, 15, 20, 25, 30\}$, taking the maximum value as our measure of overlap. The rationale for doing this is that variation and errors in data, including other mis-located sites, may result in low, inaccurate values for small numbers of neighbors. High values of k will be influenced by variation in the data and eventually by the inclusion of most of the sites. Taking the max of the Jaccard index values for the given values of k helps to mitigate these issues.

For the mis-located 2014 location of GISC1, we get M_L = max(0.0, 0.0, 0.0, 0.0, 0.0, 0.0) = 0.0. There is no overlap between the sets of neighbors. For the relocated location of GISC1 we get M_L = max(0.11, 0.18, 0.15, 0.14, 0.14, 0.15) = 0.18. While this value is not close to the maximum possible value for the Jaccard Index of 1, it is an improvement over no overlap. The low value might be explained by significant variation in proximity to the GISC1 site as well as other erroneous data that is nearby. GISC1 sits in the mountainous area north of Redding along the Sacramento River. There is a lot of variation in terrain and variation in weather in that area, particularly during the bad weather season.

For the mis-located 2014 location of KLHM, we also get M_L = max(0.0, 0.0, 0.0, 0.0, 0.0, 0.0) = 0.0, indicating no overlap. For the relocated location of KLHM we get M_L = (0.25, 0.18, 0.2, 0.29, 0.32, 0.33) = 0.33. Relocating the site shows improvement from zero overlap and helps to confirm that the site was correctly relocated.

Given that we know that the GISC1 and KLHM sites were mis-located and we know that they were re-located at, or at least close to their correct locations, we can make such comparisons and see if there is improvement. But when we don't know the correct location for a site, we are left with just the original M_L value. If that value is low or especially if it is zero, then we might suspect that a site is mis-located. But there could be other problems including that a site is producing chronically bad data and there is no relationship between it and any other sites. (We address that problem in separate work.) To address this, we can gather further evidence if we can find prospective locations for which the M_L value is greater or even optimal relative to a set of candidate locations. We can do this using a grid search to identify candidate locations for sites suspected to be mis-located. This leads to the following general logic for identifying sites that we believe are mis-located:

IF the M_L value for the original location is *low* AND the M_L value for the optimal location is *high* AND the optimal location is *far* from the original location, THEN the site may be mis-located, and it should be investigated further.

We realize that this logic is vague, and for the purposes of this paper we will leave it vague. We identified such relationships by manual inspection of the ML values and distances to optimal locations. This process could certainly be automated, but we save doing so for future work.

Now we turn our attention to identifying sites that we suspect are mis-located but that MADIS did not subsequently re-locate. We use the logic above to identify these. We find sites for which there is little or no overlap between nearest sites in terms of distance and MSE of SMART mappings, we identify optimal locations for re-locating those sites, and if the distance between the locations is large, we inspect them further.

For site BUPC1 and its location specified in the MADIS feed, we get M_L = max(0.0, 0.0, 0.0, 0.0, 0.0, 0.02) = 0.02, a small value. The optimal candidate re-location point yields a value of M_L = max(0.11, 0.25, 0.43, 0.38, 0.28, 0.28) = 0.43, a relatively high value. This is a dramatic improvement and deserves further investigation, so we created plots for the original location and the possible (optimal) re-location point. See Figure 11 and Figure 12. These plots do appear to show that site BUPC1 is mis-located relative to the 2014 MADIS data.

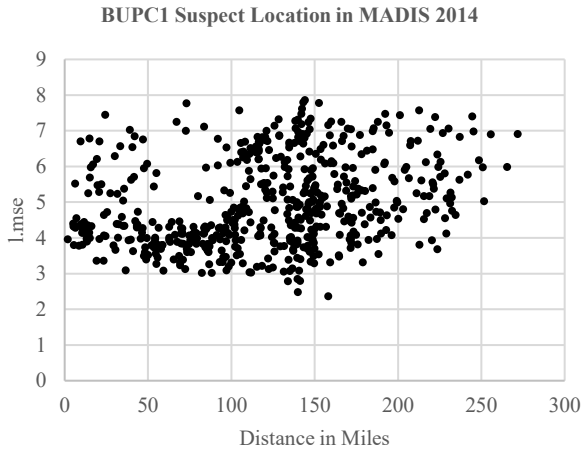

Figure 11: BUPC1 (Likely Incorrect Location) Distance versus MSE of SMART Mapping by Site

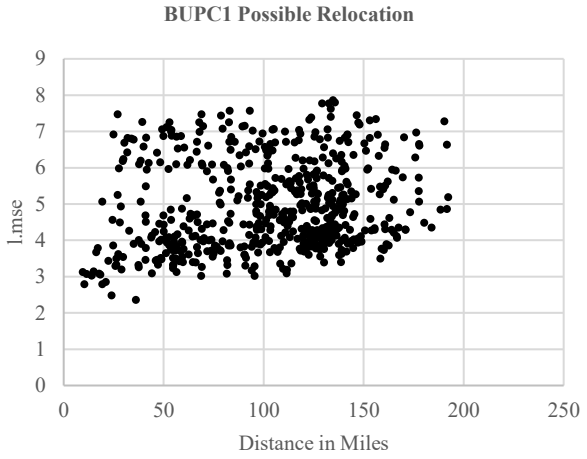

Figure 12: BUPC1 (Candidate Re-location) Distance versus MSE of SMART Mapping by Site

Another suspect site is E3738. For the location given by MADIS, we get M_L = max(0, 0, 0, 0, 0, 0) = 0. The candidate optimal location yields M_L = max(0.67, 0.33, 0.30, 0.38, 0.32, 0.36) = 0.67. See Figure 13 and Figure 14. This site also appears to be mis-located.

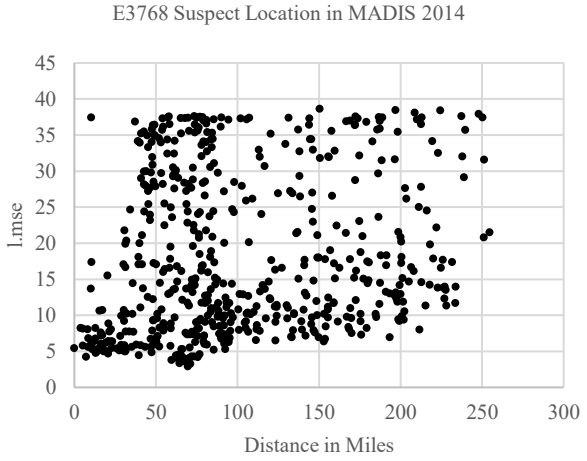

Figure 13: E3768 (Likely Incorrect Location) Distance versus MSE of SMART Mapping by Site

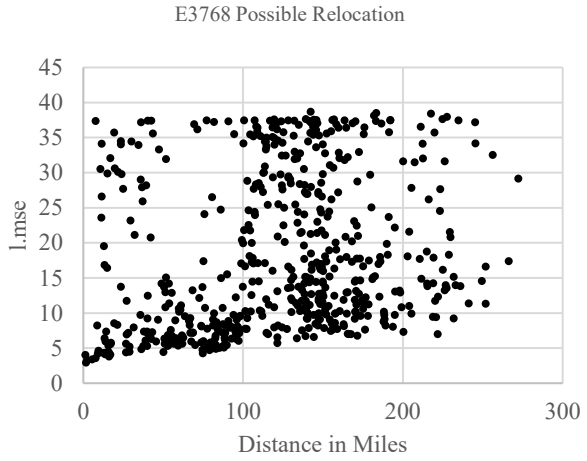

Figure 14: E3768 (Candidate Re-location) Distance versus MSE of SMART Mapping by Site

These are just several of the sites we identified in the 2014 data set that appear to mis-located and that have not subsequently been corrected by the provider. We suspect that the problem is a large one, greater than what is reflected by subsequent changes in the data, with some sites dramatically mis-located and others by lesser amounts. As demonstrated, MADIS did update locations of many sites between 2014 and 2019, so perhaps the problem is lesser now. It is hard to tell since we truly do not know with certainty which sites are mis-located.

We looked at the 2019 data and examined 625 sites meeting the same criteria as described for the 2014 data and identified further examples of what we suspect to be mis-located sites. Our analysis is not complete, but it does raise suspicion that the problem has not been alleviated. Figure 15 and Figure 16 show 2019 data for site PSWC1 and they appear to show that the site is mis-located. This is just one of multiple examples we identified using the logic presented above.

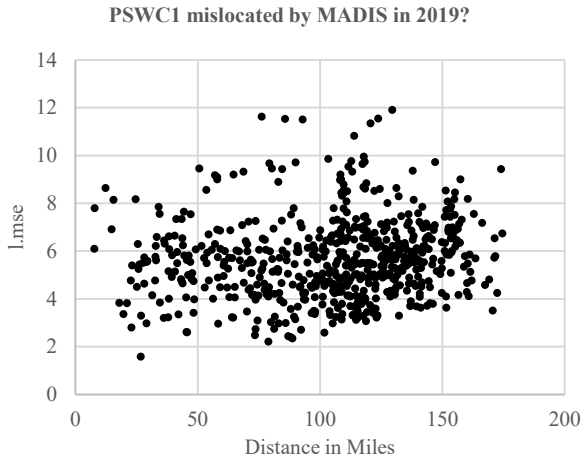

Figure 15: PSWC1 (2019 Location) Distance versus MSE of SMART Mapping by Site

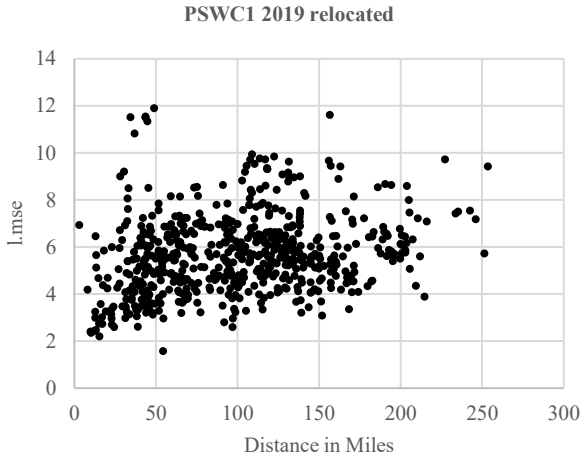

Figure 16: PSWC1 (Candidate Re-location) Distance versus MSE of SMART Mapping by Site

There is more work to do. The analysis above seems to work, but it could work better. Tightening the parameters used for the SMART mappings may help, and this could be done relative to individual sites. Iteration could be incorporated by removing the most suspect sites and recomputing. And, there would be value in removing individual data points identified as bad. All of this would help but would add to the complexity of the approach. Further investigation is merited.

VII. CONCLUSION

While our SMART method out-performed the other methods in nearly all instances in terms of accuracy of prediction of original data and classification of bad data, it was not our intent present it as the "best" method. Instead, we presented it as representative of the type of approach needed to overcome challenges of spatio-temporal data quality assessment.

It makes no assumption of isotropic covariance and does not require the determination of a specific covariance function. While it requires preprocessing time, it is suitable for near-real-time, online use. It accounts for disparate reporting times and frequency of reporting across sites. It not only helps to identify "bad data", but it also works well in the presence of bad data. It helps to identify and mitigate erroneous observations, "bad sites", and bad metadata. It uses multiple, robust methods to mitigate the impact of bad data on its estimates. Other methods, such as LSR and the various kriging approaches, could (and should) be modified in a similar manner to produce better, more robust results. Further, it is important to recognize the impact of bad data quality labels on evaluation. It is necessary to develop and use benchmark datasets with known, correct data quality labels.

A further advantage of our SMART method is that the SMART mappings provide a robust measure for comparing dissimilarity of sites. In turn, we showed how the SMART mappings could be used to identify mis-located sites. We

demonstrated that our approach works with known mis-located sites. We also demonstrated that there may be many mis-located sites for which the locations have not been corrected. Further work needs to be done in this area.

In general, we demonstrate that the quality assessment process must be an iterative process, with continual improvement and data incorporated. Figure 17 illustrates this process in general terms. A critical component in this process is evaluation.

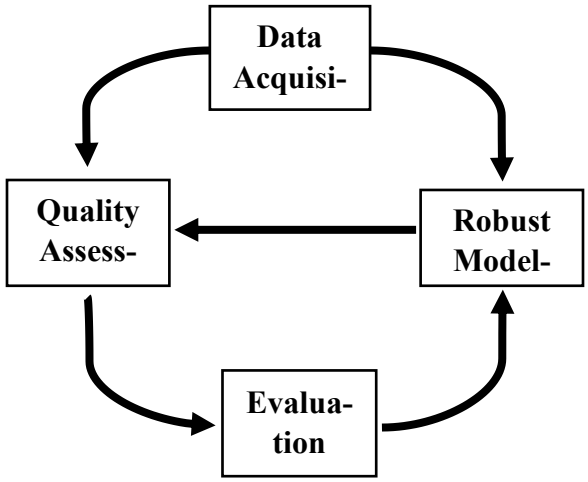

Figure 17: Iterative Data Quality Assessment Process

In this research, we investigated relatively simple situations and data sets involving ambient air temperature. We intend to expand our work to further examine other measures including wind and precipitation, as well as CCTV camera images. Departments of Transportation use CCTV camera images to verify road weather conditions reported by sensors. Yet, these images also suffer from poor data quality. Further research is needed to develop methods for detecting bad CCTV image data and for using CCTV image data to confirm sensor conditions and vice-versa. We intend to further develop benchmark datasets with known, good data quality labels.

REFERENCES

[1] D. E. Galarus and R. A. Angryk, "Challenges in Evaluating Methods for Detecting Spatio-Temporal Data Quality Issues in Weather Sensor Data," in *GEOProcessing 2019*.

[2] NOAA, "Meteorological Assimilation Data Ingest System (MADIS)." [Online]. Available: http://madis.noaa.gov/. [Accessed: 26-Dec-2015].

[3] NOAA, "MADIS Meteorological Surface Quality Control." [Online]. Available: https://madis.ncep.noaa.gov/madis_sfc_qc.shtml. [Accessed: 26-Dec-2015].

[4] U. of Utah, "MesoWest Data." [Online]. Available: http://mesowest.utah.edu/. [Accessed: 26-Dec-2015].

[5] U. of Utah, "MesoWest Data Variables." [Online]. Available: http://mesowest.utah.edu/cgi-bin/droman/variable_select.cgi. [Accessed: 26-Dec-2015].

[6] M. E. Splitt and J. D. Horel, "Use of multivariate linear regression for meteorological data analysis and quality assessment in complex terrain," in *Preprints, 10th Symp. on Meteorological Observations and Instrumentation, Phoenix, AZ, Amer. Meteor. Soc*, 1998, pp. 359–362.

[7] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.

[8] R. Nisbet, G. Miner, and J. Elder IV, *Handbook of statistical analysis and data mining applications*. Academic Press, 2009.

[9] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[10] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to data mining*. Pearson Education, Inc., 2006.

[11] C. C. Aggarwal, *Outlier Analysis*. Springer Publishing Company, Incorporated, 2013.

[12] S. Shekhar, C. T. Lu, and P. Zhang, "A unified approach to detecting spatial outliers," *Geoinformatica*, vol. 7, no. 2, pp. 139–166, 2003.

[13] A. Klein and W. Lehner, "Representing Data Quality in Sensor Data Streaming Environments," *J. Data Inf. Qual.*, vol. 1, no. 2, pp. 1–28, 2009.

[14] A. Klein and W. Lehner, "How to Optimize the Quality of Sensor Data Streams," *Proc. 2009 Fourth Int. Multi-Conference Comput. Glob. Inf. Technol. 00*, pp. 13–19, 2009.

[15] A. Klein, "Incorporating quality aspects in sensor data streams," *Proc. {ACM} first {Ph.D.} Work. {CIKM}*, pp. 77–84, 2007.

[16] A. Klein, H. H. Do, G. Hackenbroich, M. Karnstedt, and W. Lehner, "Representing data quality for streaming and static data," *Proc. - Int. Conf. Data Eng.*, pp. 3–10, 2007.

[17] A. Klein and G. Hackenbroich, "How to Screen a Data Stream." [Online]. Available: http://mitiq.mit.edu/ICIQ/Documents/IQ Conference 2009/Papers/3-A.pdf. [Accessed: 26-Dec-2015].

[18] S. L. Barnes, "A technique for maximizing details in numerical weather map analysis," *J. Appl. Meteorol.*, vol. 3, no. 4, pp. 396–409, 1964.

[19] D. Shepard, "A two-dimensional interpolation function for irregularly-spaced data," *23rd ACM Natl. Conf.*, pp. 517–524, 1968.

[20] M.E. Splitt and J. Horel, "Use of Multivariate Linear Regression for Meteorological Data Analysis and Quality Assessment in Complex Terrain." [Online]. Available: http://mesowest.utah.edu/html/help/regress.html. [Accessed: 26-Dec-2015].

[21] U. of Utah, "MesoWest Quality Control Flags Help Page." [Online]. Available: http://mesowest.utah.edu/html/help/key.html. [Accessed: 26-Dec-2015].

[22] NOAA, "MADIS Quality Control." [Online]. Available: http://madis.noaa.gov/madis_qc.html. [Accessed: 26-Dec-2015].

[23] S. L. Belousov, L. S. Gandin, and S. A. Mashkovich, "Computer Processing of Current Meteorological Data, Translated from Russian to English by Atmospheric Environment Service," *Nurklik, Meteorol. Transl.*, no. 18, p. 227, 1972.

[24] D. Zimmerman, C. Pavlik, A. Ruggles, and M. P. Armstrong, "An experimental comparison of ordinary and universal

kriging and inverse distance weighting," *Math. Geol.*, vol. 31, no. 4, pp. 375–390, 1999.

[25] 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.

[26] T. G. Mueller, N. B. Pusuluri, K. K. Mathias, P. L. Cornelius, R. I. Barnhisel, and S. a. Shearer, "Map Quality for Ordinary Kriging and Inverse Distance Weighted Interpolation," *Soil Sci. Soc. Am. J.*, vol. 68, no. 6, p. 2042, 2004.

[27] D. E. Galarus, R. A. Angryk, and J. W. Sheppard, "Automated Weather Sensor Quality Control.," *FLAIRS Conf.*, pp. 388–393, 2012.

[28] D. E. Galarus and R. A. Angryk, "Mining robust neighborhoods for quality control of sensor data," *Proc. 4th ACM SIGSPATIAL Int. Work. GeoStreaming (IWGS '13)*, pp. 86–95, Nov. 2013.

[29] D. E. Galarus and R. A. Angryk, "A SMART Approach to Quality Assessment of Site-Based Spatio-Temporal Data," in *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (GIS '16)*, 2016.

[30] D. E. Galarus and R. A. Angryk, "The SMART Approach to Comprehensive Quality Assessment of Site-Based Spatial-Temporal Data," in *2016 IEEE International Conference on Big Data (Big Data)*, 2016, pp. 2636–2645.

[31] D. E. Galarus and R. A. Angryk, "Beyond Accuracy - A SMART Approach to Site-Based Spatio-Temporal Data Quality Assessment," *(Accepted). Intell. Data Anal.*, vol. 22, no. 1, 2018.

[32] D. E. Galarus and R. A. Angryk, "Quality Control from the Perspective of the Real-Time Spatial-Temporal Data Aggregator and (re)Distributor," in *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL '14)*, 2014, pp. 389–392.

[33] D. E. Galarus and R. A. Angryk, "Spatio-temporal quality control: implications and applications for data consumers and aggregators," *Open Geospatial Data, Softw. Stand.*, vol. 1, no. 1, p. 1, 2016.

[34] H. Xie, K. T. McDonnell, and H. Qin, "Surface reconstruction of noisy and defective data sets," in *Proceedings of the conference on Visualization'04*, 2004, pp. 259–266.

[35] L. Li, X. Zhou, M. Kalo, and R. Piltner, "Spatiotemporal interpolation methods for the application of estimating population exposure to fine particulate matter in the contiguous US and a Real-Time web application," *Int. J. Environ. Res. Public Health*, vol. 13, no. 8, p. 749, 2016.

[36] J. Grieser, "Interpolation of Global Monthly Rain Gauge Observations for Climate Change Analysis," *J. Appl. Meteorol. Climatol.*, vol. 54, no. 7, pp. 1449–1464, 2015.

[37] N. Cressie, "The origins of kriging," *Math. Geol.*, vol. 22, no. 3, pp. 239–252, 1990.

[38] H. Wackernagel, *Multivariate geostatistics: an introduction with applications*. Springer Science & Business Media, 2013.

[39] M. S. Handcock and M. L. Stein, "A Bayesian analysis of kriging," *Technometrics*, vol. 35, no. 4, pp. 403–410, 1993.

[40] G. J. Hunter, A. K. Bregt, G. B. M. Heuvelink, S. De Bruin, and K. Virrantaus, "Spatial data quality: problems and prospects," in *Research trends in geographic information science*, Springer, 2009, pp. 101–121.

[41] E. H. Isaaks and R. M. Srivastava, *An introduction to applied geostatistics*. Oxford University Press, 1989.

[42] C. Huijbregts and G. Matheron, "Universal kriging (an optimal method for estimating and contouring in trend surface analysis)," in *Proceedings of Ninth International Symposium on Techniques for Decision-making in the Mineral Industry*, 1971.

[43] M. Galassi and Et-al, *GNU Scientific Library Reference Manual (3rd Ed.)*. Free Software Foundation.

[44] G. Bohling, "Introduction to Geostatistics and Variogram Analysis." [Online]. Available: http://people.ku.edu/~gbohling/cpe940/Variograms.pdf.

[45] P. J. Rousseeuw and K. Van Driessen, "Computing LTS regression for large data sets," *Data Min. Knowl. Discov.*, vol. 12, no. 1, pp. 29–45, 2006.

[46] R. F. Voss, "Random fractal forgeries," in *Fundamental algorithms for computer graphics*, Springer, 1985, pp. 805–835.

[47] J. Feder, *Fractals*. Springer Science & Business Media, 2013.

[48] M. F. Barnsley *et al.*, *The science of fractal images*. Springer Publishing Company, Incorporated, 2011.

[49] M. F. Goodchild and S. Gopal, *The accuracy of spatial databases*. CRC Press, 1989.

[50] D. E. Galarus, "Modeling stock market returns with local iterated function systems," 1995.

[51] T. C. Bailey and A. C. Gatrell, *Interactive spatial data analysis*, vol. 413. Longman Scientific & Technical Essex, 1995.

[52] USGS, "USGS Water Data for the Nation." [Online]. Available: https://waterdata.usgs.gov/nwis/. [Accessed: 28-Apr-2017].