The SMART Approach to Comprehensive Quality Assessment of Site-Based Spatial-Temporal Data

Douglas E. Galarus
Western Transportation Institute,
Department of Computer Science
Montana State University
Bozeman, MT 59717-4250
(406) 994-5268
dgalarus@montana.edu

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

***Abstract*— There is a need for comprehensive solutions to address the challenges of spatio-temporal data quality assessment. Emphasis is often placed on the quality assessment of individual observations from sensors but not on the sensors themselves nor upon site metadata such as location and timestamps. The focus of this paper is on the development and evaluation of a representative and comprehensive, interpolation-based methodology for assessment of spatio-temporal data quality. We call our method the SMART method, short for Simple Mappings for the Approximation and Regression of Time series. When applied to a real-world, meteorological data set, we show that our method outperforms standard interpolators and we identify numerous problematic sites that otherwise would not have been flagged as bad. We further identify sites for which metadata is incorrect. We believe that there are many problems with real data sets like these and, in the absence of an approach like ours, these problems have largely gone unidentified. Our results bring into question the validity of provider-based quality control indicators. In addition to providing a comprehensive solution, our approach is novel for the simple but effective way that it accounts for spatial and temporal variation.**

Keywords*— *data quality; data stream processing; spatial-temporal data; quality control; interpolation

I. Introduction

The correct way to improve the quality of sensor-data is to start at the source – the sensors – and ensure that they yield observations as close to ground truth as possible by way of careful calibration and observation. However, it is not feasible to ensure that observations from sensors match ground truth at all times and under all conditions at numerous sites. Even if sensors are operating correctly, there are many potential points of failure: remote processing units (RPUs) that read data from the sensors, intermediate devices that poll the RPUs, aggregators that collect and redistribute the data, networks over which the data is transmitted, software that processes the data, converting it to other units or formats, etc.[1][2] And there are many data quality dimensions over which failure can occur.[3][4]

One approach that can be taken to assess data quality and identify problems in real time or near real time is comparison with neighboring observations. Assuming a correlation of observations made in spatial and temporal proximity, we can expect similarity in observation. Dissimilarity can indicate errors or differences between ground truth and observed/sensed conditions. However, when ground truth is not known, we are left to compare against estimates of ground truth. In making comparisons, we find ourselves looking at slices in time, and not being able to assess overall performance of individual sites. Thus, it is useful to compare time series from multiple, neighboring sites to assess the overall quality and fitness of an individual site relative to a given sensor.

Comparison of these time series becomes complicated when considering that sites report at different times and with varying frequencies. Errors from individual observations further complicate the process when considering the adverse impact such errors can have on least squares mappings, correlation coefficients, measures of covariance, etc. If the metadata associated with a site is incorrect, we might not even be comparing to the right reference points.

Consider the following scenarios: A site is mislocated by incorrect location metadata. In turn, its observations are compared against those from sites near the incorrect location but not with those near its actual location. Sometimes the observations are in agreement and sometimes they are not. Will the site be identified as problematic? Another site produces erroneous observations with a relatively low, but positive probability. Observations from this site are then used with an interpolation method to assess the quality of observations from a neighboring site. When the observations from this site are good, the interpolation produces a good result and the neighboring site is assessed correctly. However, when the observations are bad, the interpolation produces an erroneous result and the assessment of quality yields an incorrect result.

In the absence of ground truth data, we face the challenge of attempting to identify bad data and bad sites without a solid basis for comparison and, we may never truly know if our assessments are correct. To address this challenge, we develop a representative, artificial data set which we can treat as ground truth for development of our basic method. We then perturb a subset of sites with various types of errors. Similarly, we treat an interpolated, raster dataset as ground truth for further development of methods to identify bad sites. Subsequently we apply these methods to identify bad observations and sites in a high-profile, real world, meteorological dataset.

Our Contribution: In this paper, we present specific, practical methods to identify bad observations and sites in spatial-temporal data. We demonstrate inconsistencies in

978-1-4673-9005-7/16/$31.00 ©2016 IEEE

provider quality assessment indicators for a prominent, real-world atmospheric dataset by identifying bad sites in that data.

Scope: We do not attempt to correct erroneous data or improve collection at the source. Others state correctly that correction at the source is the best way to improve data quality [5]. Our objective in this paper is to make the most of the data from providers as-is. We use mappings between sites to identify problematic observations and sites. We further investigate these sites to verify that they are problematic.

Outline: The rest of this paper is organized as follows: Section 2 provides background from a real life domain and related work, and sets the stage for our approach, which is presented in Section 3. In Section 4, we present our experimental results. In Section 5, we present conclusions and future work.

II. Background

Motivation: Since 2003, the Western Transportation Institute (WTI) at Montana State University (MSU), in partnership with Caltrans, has developed a number of web-based systems for the delivery of information from department of transportation (DOT) field devices and data from other public sources including current weather conditions and forecasts. These systems present traveler information to the traveling public and assist DOT personnel with roadway maintenance and operations. It is critical to display quality information, yet assessing the quality of the data remains a challenge.

The Weathershare system was originally developed by WTI in partnership with Caltrans to provide a single, all-encompassing source for road weather information throughout California. Caltrans operates approximately 170 Road Weather Information Systems (RWIS) along state highways, thus their coverage is limited. With a large cost for each deployment, it is unrealistic to expect pervasive coverage of the roadway from RWIS alone. Now in Phase 4, the Weathershare system is being prepared to assume greater responsibility as the repository for Caltrans RWIS data.

Weathershare aggregates Caltrans RWIS data along with weather data from other third-party aggregation sources such as NOAA's Meteorological Assimilation Data Ingest System (MADIS) [6] and the University of Utah's MesoWest [7] to present a unified view of current weather conditions from approximately 2000 stations within California. A primary benefit of the system is far greater spatial coverage of the state, particularly roadways, relative to the Caltrans RWIS network alone. In prior phases of Weathershare, automated quality control procedures were implemented for identification of "bad" data with limited success. In Phase 4 the project team revisits this challenge with the intent of assessing the quality of Caltrans RWIS data and indicating problems with RWIS sites.

Consider the temperatures shown in the Northern California area including Redding in Figure 1. By way of color-coding, two observations stand out as potentially bad: the 61°F reading shown in green and the 42°F reading shown in blue. Upon closer inspection we find that both of the sites are situated at a high elevation. The 61°F reading comes from Mt. Shasta and, while not taken from the Summit, it is reported from an elevation of 8000 ft, which certainly could account for it being colder than most other observations. The other site sits at 5700 ft. That alone doesn't account for the much lower temperature, but if we knew that the timestamp for this reading is 6 AM where-as the rest of the data shown is from approximately 3 PM, then it too could be a correct observation, albeit shown well after it occurred.

If we look at the time series of observations from individual sites, we get yet another perspective on the data and potential problems with the data. The data in Figure 2 came from the Caltrans Fredonyer Summit RWIS site on June 6, 2016. In the first plot we see that air temperature is shown as a constant 53.78°F across a full 24-hour period. The second plot shows isolated errors in pavement surface temperature on that same day with observations of 156.56°F, 2.48°F, and 169.34°F respectively. Even though these plots are small, the errors are readily apparent. Other errors may not be so obvious.

In order to assess the quality of data, mainly that of individual observations, we have taken the approach of using data from neighboring sites to validate data from a specific site. Taken in its most rudimentary form, we could simply look for anomalies on the map as we just demonstrated. Or, we could develop spatio-temporal models and compare predicted values to observed values and flag observed values as erroneous when there is a large deviation. Even if such approaches were foolproof, we have encountered little work that addresses the challenge of identifying problem sites – sites which are regularly producing erroneous data. Compounding the problem is the possibility that, like the stopped clock that is correct twice a day, sites may be producing data that is sufficiently close to that of its neighbors so-as to appear to be correct when in reality they are bad.

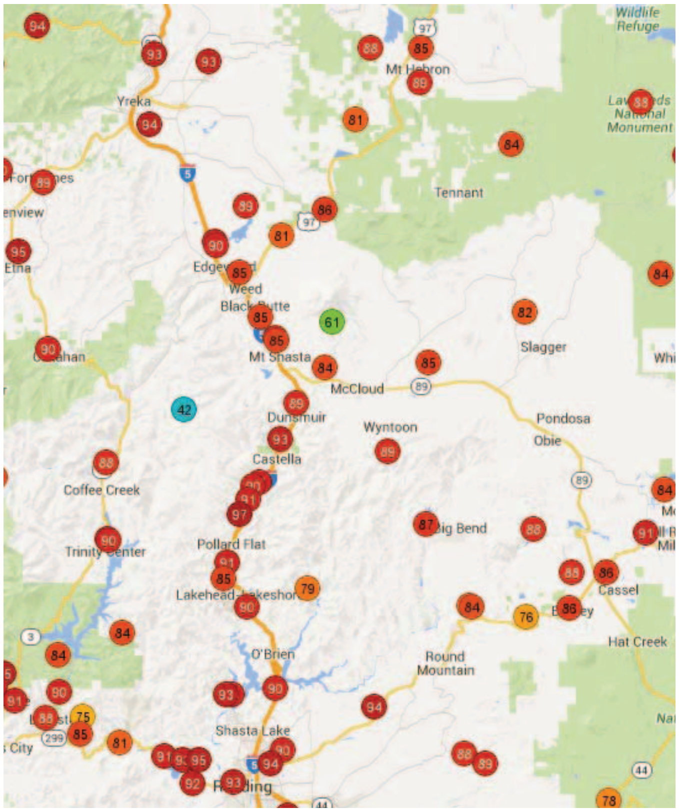

Figure 1. Air Temperature North of Redding in Early June 2015 as Shown in Weathershare

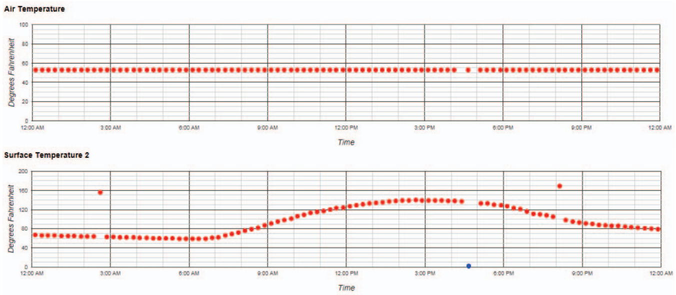

Figure 2. Errors in Air Temperature and Surface Temperature at the Caltrans Fredonyer Summit RWIS on June 6th, 2016

Literature Review: Much of the literature from the weather and road-weather communities is devoted to quality control checks for individual observations rather than for the overall site/sensor combination. For instance, these checks will help to determine if an individual temperature observation fails quality control but do not indicate if the overall temperature observation series from the site is faulty. The Barnes Spatial Test [8] is a variation of Inverse Distance Weighting (IDW) and has been used by the Oklahoma Mesonet [9] and the Federal Highway Administration's Clarus project [10]. The Barnes Spatial Test and IDW alone do not address overall data quality for a site. MesoWest [7] uses multivariate linear regression to assess data quality for air temperature. Their test does not address overall data quality for a site, although their quality control flag is assigned to the site without indication of the failing sensor [11][12].

MADIS [6] implements multi-level, rule-based quality control checks[13][14]. Their level 2 statistical spatial consistency check will flag observations as failed if 75% of the observations for the site/sensor have failed in the prior week. This check will discontinue flagging observations as bad if the failure rate for other checks drops beneath 25% in subsequent weekly statistics. While this does give an overall indication of site/sensor health, it is possible that there is a problem with a site while observations from the site still pass QC.

MADIS implements a level-3 neighbor check using Optimal Interpolation / kriging [15]. Any of the approaches mentioned (IDW, Linear Regression, kriging) could be used in this fashion to check individual observations and then flag individual observations and use cumulative statistics to flag the site as erroneous. But if the individual observation test is faulty or if site metadata is incorrect, then observations can be flagged incorrectly. Later in this paper we identify sites for which MADIS has assigned incorrect location metadata and, as a result, the neighborhood observation check for this site is faulty since it is being compared to observations from sites that are not its neighbors.

In [16], we proposed a modification of IDW that used time-series distance rather than geographic distance to assess observation quality. That work and our work in [17] and [18] focused on the use of robust methods to associate sites, but stopped with the assessment of individual observations. In this paper we extend those mappings to address the challenges of identifying bad sites. In [2] and [17] we developed quality measures that extend beyond sites, to help evaluate overall spatial and temporal coverage of a region. In that work, sites were not examined individually.

Functional Outlier Detection offers some potential overlap with our problem. In [19], the authors indicate that "for functional data the study of outlier detection has started only recently." They present several useful types of functional outliers in a taxonomy including isolated outliers, which exhibit outlying behavior over short time periods; persistent outliers which produce outliers over all or nearly all of the time period investigated; shift outliers that have the same shape as other data but is shifted in value; and amplitude outliers, which have the same shape as other data but for which the scale/amplitude differs. Unfortunately, as we demonstrate later, the data that we are dealing with may include sites that produce data that is not a function. Disparate reporting frequencies, sporadic and limited reporting, and the potential for bad timestamps introduce further challenges. Minus these challenges, approaches for functional outlier detection such as functional outlier maps [20] and functional adjusted outlyingness [21] could be applicable.

While not directly related, work such as [22] investigates the ability of a modified interpolator to predict the locations of extreme values. Turning such an approach around and comparing the predicted locations of extremes to actual extremes could be a useful approach for flagging sites with bad location metadata.

Unfortunately, none of these approaches directly addresses quality control for spatial-temporal data in a way that is immediately applicable to our situation. And, while it may seem that "simple" approaches would be applicable, counter-examples can be readily produced to show that such approaches are neither robust nor comprehensive.

III. Our Approach

Our intent here is to identify "bad" observations and sites/sensors so that they can be excluded from display and computation. Our intent is not to diagnose problems, although there certainly does seem to be opportunity in this area.

The greatest challenge we face in assessing spatio-temporal data quality is lack of ground-truth data. Comparison of observations versus ground truth is the ultimate determination of error. In order to develop and evaluate a method to identify bad sites, it is desirable to have a representative data set for which ground-truth is known. We developed such a representative dataset. In doing so, it was not our intent to model a complex system such as weather but instead to develop a weather-like data set with which we could conduct research and development.

In turn, we developed a method called SMART, Simple Mappings for Approximation and Regression of Time-series, and show how this method can be used to assess site-based, spatio-temporal data quality.

A. Interpolation

Interpolation is applied to data quality assessment by estimating ground truth and comparing observed conditions to the estimate. The idea is to use the collective set of "current" observations from multiple sites to estimate ground truth for a given site at the current time. In turn, if the deviation between an observation and estimated ground truth is “large”, then the observation will be considered erroneous or at least questionable. The current observation for the given site is held out and not used by the estimator. Formally: Let S be the set of all sites. Let $s \in S$ be a site for which we are evaluating values/observations. Let $\langle s_1, \dots, s_n | s_i \in S, s_i \neq s \rangle$ be the set of sites other than site s. Then we wish 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). We could go even further and choose the best value from a set of recent values for a given site relative to the site and observation in question. This wouldn’t necessarily yield the most recent observation from each site.

B. Our SMART Interpolator

Let an individual observation be represented as $obs = \{(t, v): t = time, v = value\}$. Let obs_i be the set of observations from site i and obs_j be the set of observations from site j. Then for a given time radius r we pair the observations from sites i and j as:

$$obs_{pairs_{i,j}} = \{(x, y): (t_1, x) \in obs_i, (t_2, y) \in obs_j, |t_2 - t_1| \leq r\}$$

Selection of the time radius is not an arbitrary choice. We want to ensure that each pair of sites will have at least three groupings of observation pairs corresponding to different time offsets.

We now define a site-to-site mapping as a linear mapping of paired observations from site i and site j:

$$l_{i,j}(x) = a + bx$$

We then define a quadratic estimate of the squared error of the linear mapping relative to the time offset between the paired observations:

$$sq\_err\_pairs_{i,j} = \left\{\left(\Delta t, \left(y - \left(l_{i,j}(x)\right)\right)^2\right): (t_1, x) \in obs_i, (t_2, y) \in obs_j, \Delta t = |t_2 - t_1| \leq r\right\}$$

$$q_{i,j}(\Delta t) = a + b(\Delta t) + c(\Delta t)^2$$

The significance of the squared error estimate relative to time offset is that we generally expect an increased squared error for an increased difference in time. This model will help to estimate the squared error and it will account for reporting time offsets between observations. Note that our method does not require a complex, data-specific covariance model.

These simple mappings are the core elements of our method and we must overcome the potential impact of the erroneous data in determining the mappings. Least squares regression suffers from sensitivity to outliers. Thus, we use Least Trimmed Squares and use the approach developed in [23] to perform Least Trimmed Squares Regression. Least Trimmed Squares determines the least squares fit to a subset of the original data by removing data furthest from the fit. Given an initial fit, an iterative process is used to successively improve the fit by removing data furthest from the fit and re-computing the fit to the remaining data.

Before applying Least Trimmed Squares to determine a linear fit we must select a trim percentage to indicate the percentage of data that will be trimmed before computing the fit. The choice is not arbitrary although it is subjective. In general, this will represent our tolerance or willingness to accept sites with bad data in our models.

Subsequently, we will refer to the coefficients of the linear fit as $l.a$ and $l.b$. And we will refer to $l.mse$ as the mean-squared-error of the linear fit to the un-trimmed data from the final fit. We will refer to the coefficients of the quadratic error function as $q.a$, $q.b$, and $q.c$; and $q.mse$ is the mean-squared error of the fit. Several additional values are derived: $q.axis$ and $q.extreme$, which represent the axis of symmetry and the extreme value of the quadratic error expression respectively.

Our SMART interpolator is a generalization of IDW, and it uses our quadratic error estimate as the distance given the time lag between pairs of sites yielding observations:

$$SMART\_estimate_s(t_s) = \frac{\sum_{i=1}^{n} \left(\frac{1}{q_{s,s_i}(t_s - t_i)}\right)^g l_{s,s_i}(v_i)}{\sum_{i=1}^{n} \left(\frac{1}{q_{s,s_i}(t_s - t_i)}\right)^g}$$

We determine the exponent by minimizing error relative to ground truth. Prior to computing the weighted estimate, weights are examined and, if necessary, “re-balanced” in order to reduce the potential influence of single sites on the outcome. For instance, we find it useful to restrict the maximum relative weight a site can be given so-as to reduce the risk that a bad value from one site can over influence the resulting average.

Rather than take a simple weighted average, we use a trimmed mean to reduce the influence of any outliers on the result. We again employ Least Trimmed Squares as a metaheuristic and compute the mean relative to the weights while minimizing the (weighted) mean-squared-error relative to the non-trimmed data.

Our algorithm is as follows: Let S be the set of all sites. Let $s \in S$ be a site for which we are evaluating values/observations. Let $\langle s_1, \dots, s_n | s_i \in S, s_i \neq s \rangle$ be the set of sites other than site s. Let $maxweight \in (0, 1]$ We use $maxweight$ =0.25. We allow no single site to contribute more than one-fourth of the weight to our estimate. Let $trimpct \in (0, 1]$ We use $trimpct$ =0.1 and average of the middle 90% of the weighted estimates. Let $minvalidqfit = 0.0001 \in (0, 1]$ and $maxvalidqfit \in (0, 1]$ = 0.5, which are chosen specific to our artificial dataset. Let $iterations_{max} = 100 \in \mathbb{N}$. This limits the number of iterations in our weighted mean algorithm.

```
$SMART\_ESTIMATE(s, S, t)$
$sumweights = 0$
$weightedsum = 0$
$\mathbf{for}\ i = 1\ \mathbf{to}\ n$
    $\mathbf{if}\ VALID\_MAPPING(s_i, s)\ \mathbf{then}$
        $\mathbf{let}\ (t_{s_i}, v_{s_i}) = MOST\_RECENT\_OBS(s_i, t)$
        $x_i = l_{s_i,s}(v_{s_i})$
        $\Delta t = t - t_{s_i}$
        $qfitval = q_{s_i,s}(\Delta t)$
        $\boldsymbol{if}\ (qfitval > minvalidqfit)\ \boldsymbol{and}\ (qfitval < maxvalidqfit)\ \boldsymbol{then}$
            $weight = \frac{1}{qfitval}$
            $w_i = weight$
        $\boldsymbol{else}$
            $w_i = 0$
    $\boldsymbol{else}$
        $w_i = 0$
        $x_i = 0$
$NORMALIZE\_WEIGHTS(w)$
$BALANCE\_WEIGHTS(w, maxweight)$
$predicted = W\_TRIMMED\_MEAN(x, w, trimpct, iterations_{max})$
$\boldsymbol{return}\ predicted$
```

$NORMALIZE\_WEIGHTS(w)$ normalizes weights to sum to 1. The weights must be non-negative.

$BALANCE\_WEIGHTS(w, maxweight)$ reduces any weights that exceed the maximum specified and redistributes the excess weight proportionally to remaining elements. Iteration may be necessary in the event that a redistributed weight exceeds the maximum specified.

$W\_TRIMMED\_MEAN(x, w, trimpct, iterations_{max})$ uses the iterative Least Trimmed Squares metaheuristic to determine the optimal mean in terms of mean squared error relative to the untrimmed data.

IV. Results

A. Assessment of Individual Observations

We developed a weather-like phenomenon representing temperature using approximate fractal surfaces generated using the method of Successive Random Addition [24][25][26]. 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 introduce motion/flow.) We generated one surface and eight weather patterns, allowing us to train on one weather pattern and test on the remaining.

Observation time series were then generated by combining the surface data with the weather data along with a periodic effect and a north-south effect to simulate a weather-like phenomenon similar to the diurnal effect and general north-south variation in the Northern Hemisphere. This yields a time series of ground truth values of length n=513 for each (x, y) on the 513x513 surface.

We then selected 250 "sites" using random uniform x-y coordinates. For each site, we assigned a reporting pattern defined by m = RandInt(1..10) and s=RandInt(0..m-1) where s is the start time and m is the frequency generating a series of times: $\langle s, s+m, s+2m, \dots \rangle$. Errors were added to 25 sites in the form or random noise added to ground truth, rounding of ground truth, replacement of ground truth with a constant value, random bad values with varying probabilities, or negation of ground truth. The remaining 225 sites were left error-free.

Using the data from our artificial data set's first time period / weather pattern, we created mappings between all sites and then applied them to make estimates for other time periods. We compared results to those for IDW, Least Squares Regression and Universal Kriging. For Least Squares Regression we investigated three cases with varying radii: a maximum of 50 units, a maximum of 100 units and no maximum/all data. For Universal Kriging we used a radius of 150 units, a value for which computation time was still reasonable while not restricting too much data.

The mean-squared-errors over all predictions versus ground truth shows that our SMART method dramatically out-performs the other methods. See Table 1.

Table 1: MSE from Ground Truth for Interpolators

SMART	IDW	LSQR50	LSQR100	LSQRALL	UK150
0.088	1.110	11.8175	6.228	1.413	2.980

Next, we extracted temperature observations from the MADIS Mesonet subset from December 2015 and bounded between 38.5°N and 42.5°N latitude and -124.5°W and -119.5°W longitude, covering California north of Sacramento and overlapping a portion of Oregon, Nevada and the Pacific Ocean. We used the first week in December for a training set and the second week for testing. All total, 890 sites had observations in this subset. Note that weather was variable during this month. California finally received bad weather during the bad weather season in 2015 after several years of little or no winter.

We compared our SMART method to Inverse Distance Weighting (IDW), Least Squares Regression with a 50 mile radius (LSQ50), LSQ100 (100 mile radius), and LSQALL (include all data). For all methods, the most recent observations from individual sites were used. We were not able to get a consistent covariance model and could not get Universal Kriging to work consistently with the MADIS data. Issues were experienced with non-invertible matrices because of coincident sites, etc. Too much preprocessing would have been necessary, and so we did not make predictions using Universal Kriging on the MADIS data set.

Since we don't have ground truth for this data set, we computed mean-squared error relative to observed value and grouped by the quality control descriptors (QCD) that MADIS provides for observations. While there are some differences, the results were similar across the QCD classes, i.e., using provider quality control for evaluation. The resulting mean-squared-error by method and QCD is shown in Table 2.

Table 2: MSE by MADIS QCD for Interpolators

	SMART	IDW	LSQ50	LSQ100	LSQALL
Q	37.64	397.24	378.36	358.46	337.20
S	7.16	107.26	79.92	66.84	60.67
V	4.16	125.40	93.11	74.64	61.06
X	85525.63	95891.65	96555.19	96803.05	97299.10
ALL	97.41	254.49	225.33	209.36	199.24
V=verified/good, S=subjective/good, X=bad, Q=questionable					

The MSE results in Table 2 are very good. We now look at results for the Caltrans Dunsmuir RWIS site, CTDUN. See Figure 3 and Table 3.

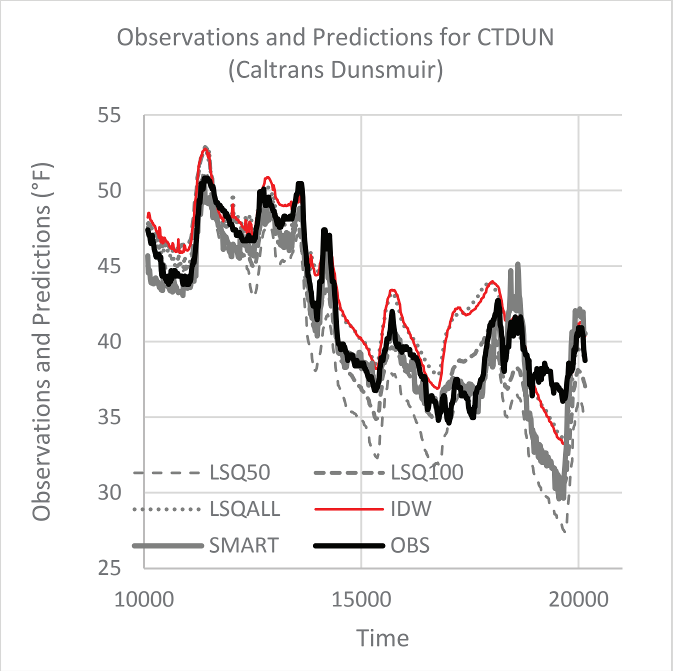

Figure 3: Predictions for CTDUN from the Interpolators

Table 3: MSE for Caltrans CTDUN

SMART	IDW	LSQ50	LSQ100	LSQALL
4.073	5.770	14.759	5.791	5.799
QCD Counts: S=301, V=369, Total=670				

According to MADIS, all of the CTDUN data is good although nearly half of the data required a subjective (manual) test to determine its quality. Even though our SMART method tracks very closely with the observed values, it deviates over the low (and prior high) at the end of the period. The other methods show differences throughout. In terms of MSE from observed, SMART is best, with IDW and LSQ100 next best. All of the observations during this period have QCD of V or S, so they are considered to be good by MADIS. The variability in the weather during this time of year and across the training and test period may have contributed to the error here as well. Otherwise, the other methods may have performed relatively well due to a lack of erroneous data from sites in proximity to CTDUN.

Now we look at Site GISC1, which MADIS located in downtown Sacramento during December 2015. Our SMART method tracks reasonably close to the observed values for this site whereas all of the other methods tend to overestimate. Notice that the observations look similar to those for CTDUN even though MADIS located them 170 miles apart. See Figure 4 and Table 4.

The MADIS QCD indicators show uncertainty regarding the quality of observations from this site, with a mix of Q, S and V values. However, when we compare the corresponding observations to the error of predictions from our SMART method, there is no apparent pattern.

For site GISC1, our SMART method performs better overall than the other interpolators, but worse than LSQ100 for the V data. Our SMART method performs better for the Q and S data than for the V data, particularly for the Q data. One way to make sense of this would be if the data was mislabeled. In fact, we are confident that it is mislabeled because we know that this site was mislocated and, as a result, the MADIS level 3 check is invalid. This site is not located in downtown Sacramento but is instead located near CTDUN, 170 miles to the north of Sacramento. The reason it fails MADIS level 3 quality control resulting in Q values for QCD is that it often doesn't agree with sites located in downtown Sacramento. There are times when it comes close, and those times correspond to S and V QCD indicators, but many times the differences are large enough to cause MADIS to flag observations from this site as failing. Next we demonstrate how our SMART method helps to identify sites having bad metadata including location metadata.

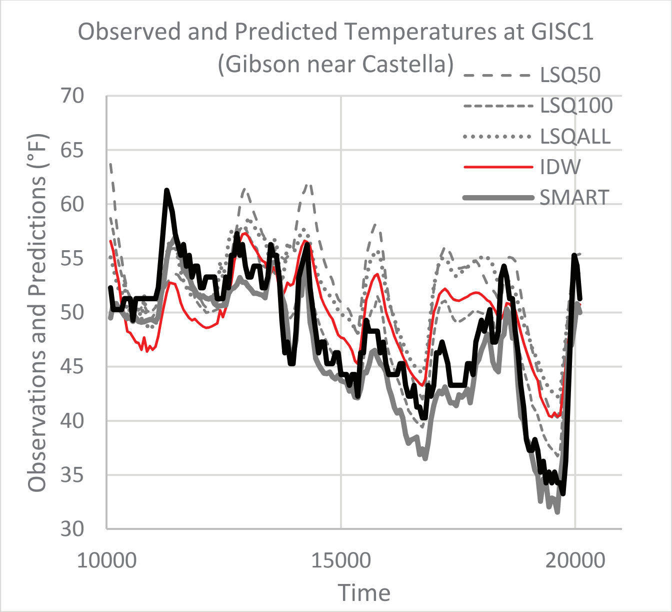

Figure 4: Predictions for GISC1 from the Interpolators

Table 4: MSE by QCD for GISC1

	SMART	IDW	LSQ50	LSQ100	LSQALL
Q	6.613	34.528	90.706	19.340	62.990
S	5.202	22.287	26.250	7.112	25.041
V	8.808	9.733	23.103	4.959	18.688
ALL	6.839	20.570	40.070	9.169	31.544
QCD Counts: Q=39, S=68, V=61, Total=168					

B. *Identification of Problem Sites*

We now look at the coefficients and MSE values for the site-to-site SMART mappings and use these to identify "bad" sites. We extracted temperature observations from the MADIS Mesonet subset for all of December 2015 and bounded between 38.5°N and 42.5°N latitude and -124.5°W and -119.5°W longitude, covering California north of Sacramento and overlapping a portion of Oregon, Nevada and the Pacific Ocean.

1) Sites for Which Mappings Could Note Be Determined

There were 36 sites for which site-to-site mappings could not be determined. We examined these sites individually and the results were not surprising, with each site exhibiting one or more of the following issues: there was little or no data during the training period, there were large gaps in reporting, or there was a very narrow range in reported data. Many of these sites could have readily been identified by other means, and we intend to filter sites in the future based on measures such as temporal completeness [1][2]. Still, their handling is non-trivial. Some of these sites began reporting data during the testing period and some of that data could have been useful depending on the application. A potential disadvantage of our SMART method is that it will disregard such sites until a new training period has passed in which there is sufficient data to evaluate the site. In favor of our approach, one could argue that it is prudent to hold out data from such a site until the site can be proven to produce quality data.

2) Outlier Sites

For each site we averaged the *l.a* values for mappings to that site. Even though there were outliers in the individual *l.a* values, we used the arithmetic mean for the site averages. A more robust measure such as trimmed mean or median could be used to mitigate the impact of outliers if necessary. For those sites having an average *l.a* value on the extreme negative side, the observations are far lower than most other sites. For those having an average *l.a* value on the extreme positive site, then the corresponding site has observations far greater than most other sites. At first we may consider such sites questionable and should seek a reasonable explanation for their behavior. It is possible that a high mountaintop site will experience lower temperatures than all or most other sites. And, it is possible that a low desert site will experience higher temperatures than most other sites. But if the differences are extreme, particularly in comparison to other similarly-located sites, then such a site should be considered bad / erroneous, and it should be held out from further analysis until the associated problems are rectified. From observation of the apparent outliers we set our cutoff for *l.a* as $(-56.73, 51.33)$, where any site having an average l.a value outside this interval is labeled "bad".

For the average *l.b* coefficients we found a cutoff interval of $(0.16, 1.71)$. For this dataset, we do not expect negative l.b values since there should be a positive correlation between sites, at least nearby sites. We do not expect *l.b* values of zero or near zero since that would signify that a site has non-changing observations. And, we don't expect large values for *l.b* since the would signify ranges for a site outside the norm of that for other sites.

For the average *l.mse* values by site, the outliers generally overlap with the outliers for the *l.a* and *l.b* values. However, we did find multiple sites that had an outlier *l.mse* value despite not having outlier *l.a* or *l.b* values, so it is worthwhile to check these as well. A cutoff interval of $(1.58, 31.82)$ was used.

After determining the outlier sites from the site averages of the *l.a*, *l.b* and *l.mse* values, we then examined each outlier site. In every case, there was at least one of the following logical explanations as to why the site was an outlier including: multiple apparent series rather than one, constant or near constant values, many outliers, large outliers, range of values much wider than the normal range, wild variation, gaps or sparse data during training period, sporadic reporting, or generally doesn't follow the expected trend. Figure 5 and Figure 6 show two interesting examples.

Even though values fell within a reasonable range during the training period for Site 24, a greater number of apparent outliers occurred subsequently in the month. The mappings appear to have identified a problem that became an even greater problem.

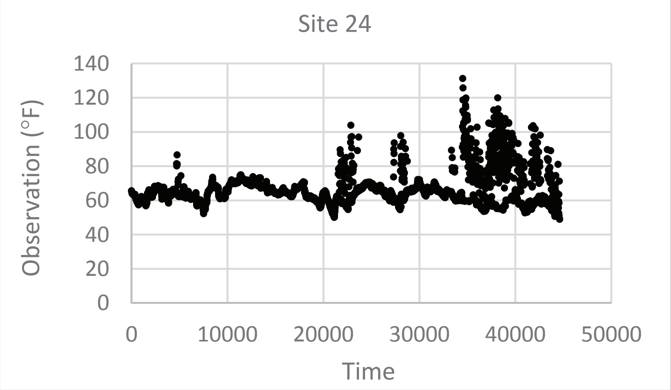

Figure 5. Site 24 from MADIS Subset showing Errors During Training Period that Become Greater Errors Later in the Month

There are two apparent series within the single series for Site 26. It is unclear how this happened and we would not otherwise have assumed this to be possible. Perhaps two sites' data are being combined into one site.

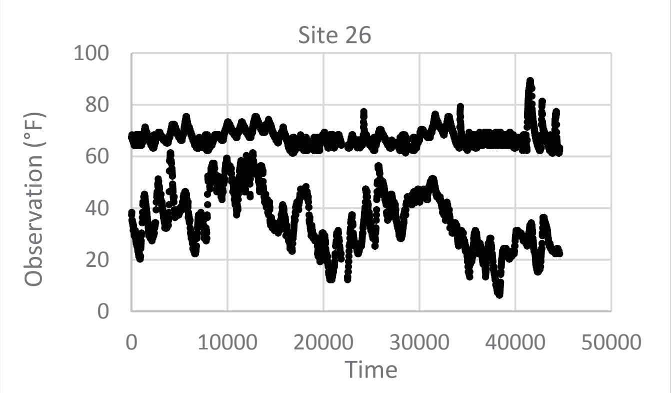

Figure 6. Site 26 from MADIS Subset Erroneously Presents Two Apparent Data Streams within One Data Stream

3) Bad Metadata

Next we investigate another type of error – bad metadata. In this case, sites may produce valid observations but their location or timestamps or other metadata may be incorrect. In turn, when used in models that depend on this metadata, results should be considered erroneous and the site should be considered bad until the metadata can be corrected.

a) Bad Location Metadata

In order to develop a method for identifying bad location and timestamp metadata, we again needed a representative ground truth dataset. For this purpose, we used the Mesoscale Analysis and Prediction System (MAPS) and the Rapid Update Cycle (RUC) Surface Assimilation Systems (MSAS/RSAS) dataset from NOAA for air temperatures in the Northern California area for the entire month of December 2015 [28][29]. The MSAS/RSAS data provides estimated surface observations for grid points with 8-mile spatial and one-hour temporal resolution. Since one-hour temporal resolution was not sufficient for our analysis and was not representative of the frequency of reporting of site-based sensor observations, we used bi-cubic splines to interpolate down to 1-minute temporal resolution. We treated the individual grid points as sites.

In general, the series from near sites should be similar and the series from far sites should be less similar. The *l.mse* values from our SMART mappings provide an indication of similarity, and we expect a positive correlation between *l.mse* and the distance between sites, with the lowest *l.mse* values corresponding to the nearest sites. A plot of distance versus *l.mse* for all grid points with respect to a grid point near Dunsmuir shows this relationship. See Figure 7.

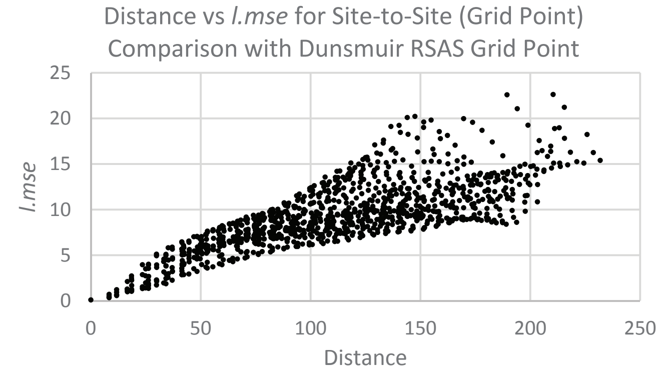

Figure 7. Distance versus *l.mse* for Site-to-Site Comparison of RSAS Grid Point near Dunsmuir with other Grid Points

To investigate a mislocated site, we changed the location metadata of the Dunsmuir grid point to that of a grid point 190 miles to the south in Sacramento and computed distances accordingly. As expected, the near points no longer have the least *l.mse* values. Points between 175 and 210 miles away show the least *l.mse* values, indicating that the site is mislocated. See Figure 8.

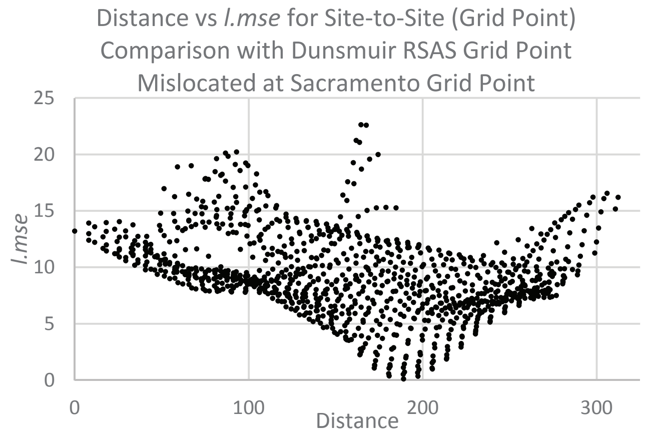

Figure 8. Distance versus *l.mse* for Site-to-Site Comparison of RSAS Grid Point near Dunsmuir with other Grid Points in which Distance is Compute as if the Dunsmuir site was mislocated 190 Miles to the South in Sacramento

The same general positive correlation holds for the MADIS data, albeit with outliers and greater variability. These outliers are likely due to mislocated sites or errors within the series for individual sites, and variation. Differences in reporting times and frequencies may further contribute to variation in the data.

We have known that GISC1 (Gibson near Castella) was mislocated by MADIS at (38.56556, -121.485) in downtown Sacramento. Weathershare users have identified this site as mislocated. The incorrect location data persisted for years in the MADIS feed but was finally updated in 2016, separate from and subsequent to the analysis presented here. MADIS corrected GISC1's location to (41.022, -122.399) in 2016, 175 miles to the north near the Caltrans Gibson Maintenance yard and near the town of Castella. For the December 2015 data, the mislocation of GISC1 is readily apparent in the plot of distance versus *l.mse*. See Figure 9. Not only are the values for *l.mse* high for small distances from the mislocated site, the lowest *l.mse* values correspond to sites approximately 175 miles away which are in proximity to the correct location. There is no apparent indication why/how this site was mislocated.

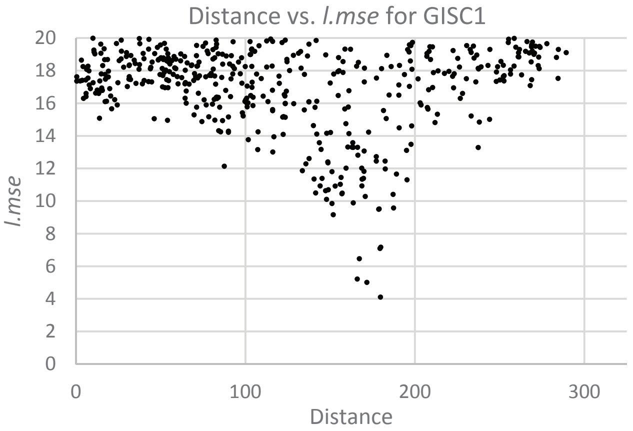

Figure 9. Distance versus *l.mse* for Site-to-Site Comparison of GISC1 (Gibson near Castella) Site from the MADIS Subset with other MADIS Sites. This site was mislocated in Sacramento 175 miles to the south of its correct location by MADIS.

Several other mislocated sites are identified readily by comparing distance and *l.mse*, and subsequent corrections by MADIS confirm that these sites were mislocated. For these sites there are clues as to the nature of the original error. MADIS mislocated Site WVVCA at (40.680, -120.83), near Eagle Lake 105 miles to the east of its correct location. MADIS subsequently relocated this site to (40.680, -122.830) along State Route 299, near Weaverville. The WVV likely stands for Weaverville. The incorrect longitude was -120.83 and the correct longitude was -122.830. There may have been a transcription error for this site. Site KLHM - Lincoln Regional Airport - was mislocated by MADIS at (38.91, -120.65), 37.6 miles to the east of its correct location, and was subsequently relocated to (38.910, -121.350) at the Lincoln Regional Airport. Given that the latitudes correspond and the longitudes differ, transcription error is again suspected.

Our approach for identifying mislocated sites does work, as confirmed by sites known to have been mislocated by MADIS. Approximate, correct locations may be identified by way of the location of the sites corresponding to the least *l.mse* values. It is important to identify these in a grouped, robust fashion rather than simply use the site with the least *l.mse* value since that site too may be mislocated.

b) Bad Timestamp Metadata

The last type of error we investigate is that of bad timestamps. We believe that many if not most of the sites have data for which the timestamps are not synchronized with the correct time. In order to investigate this possibility, we again use the RSAS data from December 2016. Here we make an assumption that clocks are synchronized and timestamps are correct for sites represented by grid points in the RSAS data. We will plot *l.mse* versus *q.axis* and investigate the data relative to constant *q.axis* values, particularly *q.axis*=0. We use *l.mse* rather than distance because of problems with mislocation of sites as demonstrated in the previous section and with the intent of being able to better discern similar sites. For low *l.mse* values, *q.axis* values should be near zero. And, the *q.axis* values should generally be centered around *q.axis*=0. This pattern holds in Figure 10 for the site represented by the grid point near Dunsmuir, relative to the other grid points. There are outlier values for larger *l.mse*. but the general pattern still holds.

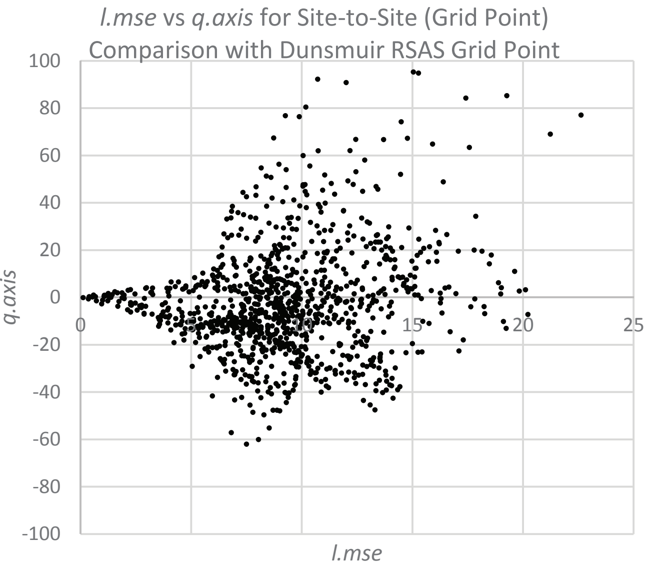

Figure 10. *l.mse* versus *q.axis* for Site-to-Site Comparison of RSAS Grid Point near Dunsmuir with other Grid Points Assuming Synchronized Clocks

Next, we shift the Dunsmuir timestamps by -15 minutes to represent a clock that is 15 minutes fast. For the modified data as compared to that from the other grid points and without modifying timestamps for the other grid points, we see a similar pattern, but the plot is shifted and now falls such that the *l.mse* values are approximately centered around *q.axis*=-15, matching the shift in the timestamps. See Figure 11. Similarly, if we shift the Dunsmuir timestamps by +15 minutes to represent a clock that is 15 minutes slow, we would see a corresponding shift of +15 minutes.

This relationship appears to hold for real sites and observations from the MADIS dataset. We find that the plots of *l.mse* versus *q.axis* for many sites are centered at or near *q.axis*=0. However, some sites clearly do not show this relationship, and we think these sites have timestamps that are not synchronized with the majority of the other sites. For example, site CFRC1 is located in the Shasta National Forest near where the Trinity River joins the upper portion of Trinity Lake, approximately 24.5 miles southwest of Dunsmuir. When we plot *l.mse* versus *q.axis* for site mappings to CFRC1, we see that they are centered at a *q.axis* value of 40 or greater. We believe the time offset for timestamps from this site relative to others is large. See Figure 12.

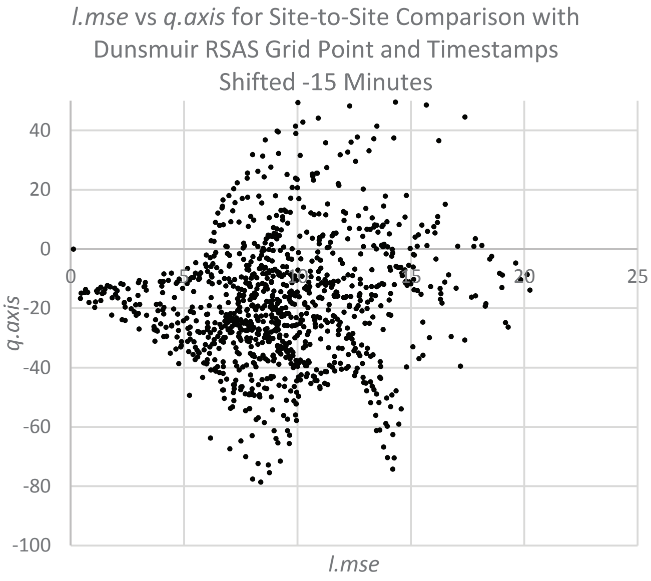

Figure 11. *l.mse* versus *q.axis* for Site-to-Site Comparison of RSAS Grid Point near Dunsmuir with other Grid Points Assuming the Dunsmuir Clock is 15 Minutes Fast (timestamps -15) and All Other Clocks are Correct

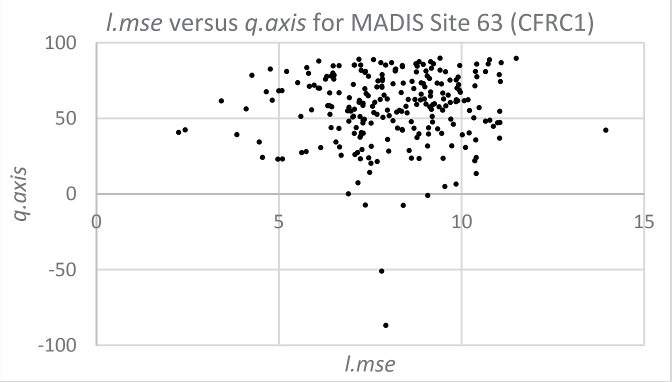

Figure 12. *l.mse* versus *q.axis* for Site-to-Site Comparison of CFRC1 from the MADIS Subset with other MADIS Sites

We have found numerous other examples of sites in which we are confident that the timestamps are incorrect, and we have found a great deal of variability in this data. Ultimately, we believe that there are both large and small errors in the timestamps across the entire MADIS dataset.

V. Conclusions and Future Work

The SMART method presented in this paper shows promise for identifying bad sites that provide spatial-temporal data. It produces simple (linear) site-to-site mappings and error estimates for the mappings that can be used to identify erroneous observations and sites. The coefficients of the mappings and associated performance measures can be compared across all sites and outlier values of these parameters correspond to bad sites. Robust methods such as Least Trimmed Squares are used to produce robust mappings that associate similar sites in light of errors.

We presented examples from three data sets demonstrating the effectiveness of our SMART method in identifying bad observations and bad sites. Our method outperforms other popular interpolation methods in the presence of bad data and identifies bad sites.

In future work we intend to characterize the impact of the various components of spatio-temporal data quality including accuracy, precision, timeliness, reliability, completeness and coverage on interpolation-based methods including our SMART method.

VI. Acknowledgments

We acknowledge the California Department of Transportation (Caltrans) for its sponsorship of the WeatherShare project and other related projects. In particular, we acknowledge Ian Turnbull and Sean Campbell from Caltrans. We further acknowledge Daniell Richter and other staff at the Western Transportation Institute for their work on WeatherShare and other related projects. The work presented in this paper has been conducted subsequent to and separate from this prior work.

References

[1] D. E. Galarus and R. A. Angryk, “Quality Control from the Perspective of a near-Real-Time, Spatial-Temporal Data Aggregator and (re) Distributor,” *Proc. 22nd ACM SIGSPATIAL Int. Conf. Adv. Geogr. Inf. Syst.*, 2014.

[2] 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.

[3] R. Y. Wang and D. M. Strong, “Beyond accuracy: What data quality means to data consumers,” *J. Manag. Inf. Syst.*, vol. 12, no. 4, pp. 5–33, 1996.

[4] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, “Methodologies for data quality assessment and improvement,” *ACM Comput. Surv.*, vol. 41, no. 3, p. 16, 2009.

[5] R. D. De Veaux and D. J. Hand, “How to Lie with Bad Data,” *Stat. Sci.*, vol. 20, no. 3, pp. 231–238, 2005.

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

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

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

[9] M. A. Shafer, C. A. Fiebrich, D. S. Arndt, S. E. Fredrickson, and T. W. Hughes, “Quality assurance procedures in the Oklahoma Mesonetwork,” *J. Atmos. Ocean. Technol.*, vol. 17, no. 4, pp. 474–494, 2000.

[10] M. Limber, S. Drobot, and T. Fowler, “Clarus Quality Checking Algorithm Documentation Report,” techreport, 2010.

[11] J. Splitt, Michael E; 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].

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

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

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

[15] 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.

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

[17] D. E. Galarus and R. A. Angryk, “Quality Control from the Perspective of the Real-Time Spatial-Temporal Data Aggregator and (re)Distributor,” in *ACM SIGSPATIAL ’14*, 2014.

[18] R. A. Galarus, Douglas E.; Angryk, “A SMART Approach to Quality Assessment of Site-Based Spatio-Temporal Data,” in *ACM SIGSPATIAL ’16*, 2016.

[19] M. Hubert, P. J. Rousseeuw, and P. Segaert, “Multivariate functional outlier detection,” *Stat. Methods Appl.*, vol. 24, no. 2, pp. 177–202, 2015.

[20] M. Hubert, P. Rousseeuw, P. Segaert, and others, “Rejoinder to ‘multivariate functional outlier detection,’” *Stat. Methods Appl.*, vol. 24, no. 2, pp. 269–277, 2015.

[21] M. Hubert, J. Raymaekers, P. J. Rousseeuw, and P. Segaert, “Finding Outliers in Surface Data and Video,” *arXiv Prepr. arXiv1601.08133*, 2016.

[22] M. Tomczak, “Spatial Interpolation and its Uncertainty Using Automated Anisotropic Inverse Distance Weighting (IDW) - Cross-Validation/Jackknife Approach,” *J. Geogr. Inf. Decis. …*, vol. 2, no. 2, pp. 18–30, 1998.

[23] 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.

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

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

[26] M. F. Barnsley, R. L. Devaney, B. B. Mandelbrot, H.-O. Peitgen, D. Saupe, R. F. Voss, Y. Fisher, and M. McGuire, *The science of fractal images*. Springer Publishing Company, Incorporated, 2011.

[27] “NOAA MSAS/RSAS.” [Online]. Available: http://msas.noaa.gov/.

[28] NOAA, “MSAS/RSAS,” 2013. [Online]. Available: http://msas.noaa.gov/. [Accessed: 24-Sep-2016].

[29] NOAA, “The MSAS/RSAS Surface Analysis,” 2007. [Online]. Available: http://msas.noaa.gov/msas_descrip.html.