A FRAMEWORK FOR SPATIOTEMPORAL WEATHER SENSOR DATA QUALITY

by

Douglas Edward Galarus

A prospectus submitted in partial fulfillment

of the requirements for the degree

of

Doctor of Philosophy

in

Computer Science

MONTANA STATE UNIVERSITY

Bozeman, Montana

June 2015

(revised August 2015)

DEDICATION

The work of my PhD is dedicated to my family in gratitude for their patience during the time that I worked on it. It has been especially hard for my small children to understand why I am constantly needing time to do "school work", and their lives to date have (Michael) or nearly have (Lucy) wholly overlapped with the time I've been working on the PhD. My family is my number one priority and I've tried to make as much possible time for them, but wish I could have made more time during these past years. Perhaps the greatest reward upon completion of this task will be having more time to spend with them.

Thank you Lucy and Michael for allowing me time to concentrate the work of my PhD. Thank you Jana for carrying the load while I took care of work and school, and for your never-ending support. Thank you Rachelle for being an inspiration by returning to school yourself and for your encouragement. And thank you Dad. Even though you're gone now, I think of you every day and wish we had more time together and that you could see me complete this thing that I started in your presence. To all of you – please realize that I am doing this for you and for us.

Time is perhaps the greatest challenge we face in life, and I've found that most of my struggles stem from shortages of time. It seems fitting that the work of my PhD at least partially overlaps with challenges related to time.

ACKNOWLEDGEMENTS

The topic I am researching comes from Ian Turnbull at Caltrans. Ian's mantra of "accurate, timely and reliable" echo in every effort that we have teamed up on. Without Ian's requirements for quality on our projects, my interests likely would have fallen elsewhere. Ian is the engineer's engineer, and it has been a pleasure working with him.

Sean Campbell, also from Caltrans, has also been inspiring on this effort. We have shared many conversations regarding not only quality of weather sensor data, but also that of CCTV images, chain control messages, CMS messages, and just about every other type of traveler information available. Although not used directly on this project, much of the data that helped to inspire this project would not be available without the work that Sean has done to make it available. I hope that something practical from this work will trickle back to Sean and make his work easier.

My colleagues at the Western Transportation Institute current and past also deserve acknowledgement, particularly Dan Richter and Leann Koon. Thanks for being sounding boards for ideas, for implementing some of the ideas, and for helping to understand the applications of this work to transportation.

I acknowledge my advisor, Dr. Rafal Angryk, for continuing to be my principal advisor after leaving Montana State University for Georgia State. I am appreciative for the time you've spent guiding me to the goal. My PhD committee has also provided some helpful comments on this work.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

ABSTRACT

In this prospectus, we investigate the impact of various data quality factors on the problem of determining data quality for observations from a given weather sensor data stream. Our problem is an offshoot of various research and development projects conducted at the Western Transportation Institute for the California Department of Transportation (Caltrans) in relation to their Road-Weather Information Systems (RWIS). It has been challenge for Caltrans to assess the quality of the data from their RWIS units, in addition to field calibration and ground-truthing.

We have generally employed an approach of using data from third-party providers for comparison against the RWIS data. However, this approach is dependent on multiple quality factors which one of our project champions characterizes as the trio of accuracy, timeliness and reliability. Is the data correct? I.e., is it an accurate measure of the real condition it represents? Do we receive it in a timely fashion? I.e., can we obtain an observation soon after it was recorded in the field? And, can we reliably gain access to this data? For instance, do the communication networks and computer systems that provide the data perform reliably or are they prone to outages? All of these quality factors can have an impact on quality control processes that we implement.

Specifically in our comprehensive exam presentation we will investigate the example application of ordinary-kriging to the problem of assessing and quantifying the quality of individual sensor observations. On the surface, ordinary-kriging appears to be a good choice for this task. Given a set of known observations and associated locations, we can estimate values for additional locations using ordinary-kriging as an interpolator. In a typical application of ordinary-kriging, one would estimate values at locations for which observations are not known. In our application, we will hold-out our known observation and compare it to the prediction given by ordinary-kriging. Further, since ordinary-kriging provides confidence intervals, our comparison can be expressed in terms of these confidence intervals, and if the observation falls outside of the confidence interval, it might be rejected as being bad. However, as we will demonstrate, the quality factors mentioned above all can contribute performance challenges to our application of ordinary-kriging.

As our proposed research thesis, we will present the components of a general framework for spatiotemporal sensor data quality.

PROBLEM STATEMENT AND SIGNIFICANCE

Problem Statement

The specific problem we address in our research is that of determining the quality of spatiotemporal weather sensor data. For a variety of reasons, such data may not reflect the specific conditions they are intended to measure and represent. At a first glance, one might think that this task is synonymous with outlier detection. While outlier detection does play a prominent role, the challenges go beyond identifying outliers. A sensor might become "stuck" and produce the same output over an extended period of time. This output might conform to other nearby observations and fall within an acceptable range of values, yet it does reflect actual conditions. Similarly, a sensor may drift, reporting values further and further from correct values over time. And, a sensor may report correct values, but its clock may be incorrect, resulting in incorrect timestamps. Similarly, an incorrect location may be associated with a site, and that site's sensor readings may be associated with a location far away.

Our over-arching goal is to define multidimensional data quality measures, use these measures to determine the impact of data quality on state of the art algorithms for assessing data quality, and develop a new algorithm to assess data quality in light of poor poor and uncertain data quality.

Problem Significance

The amount and availability of data from sensor networks has grown rapidly in recent years due to increased computing power, greater coverage and bandwidth of communication networks, and reduced storage costs, as well as reduced costs for sensing equipment. The types of monitoring have expanded from environmental sensing, industrial monitoring and control, as well as traffic monitoring, to the monitoring of household appliances, power consumption and control of household heating/cooling. The evolving "Internet of Things" will surely make even more data available for new applications from numerous, multiple providers. As such, increased attention must be given to quality control from not only the perspective of data owner but also that of the aggregator and disseminator of data, and to the impact of quality control on their processes and products.

Quality control measures, if included at all, are generally presented from the perspective of the original data provider, with a focus on sensor accuracy, precision and other measures assessing the direct performance of the sensor. Differing quality control measures and policies from providers yield further challenges to data aggregators. For instance, one weather data provider may present quality control indicators at the sensor level while another flags at the station level indicating that something is wrong with the station but leaving uncertainty as to which of multiple sensor readings are in question.

Aggregating such data into a uniform and cohesive data feed is a challenge, as is the task of selecting which providers should be used from multiple, overlapping offerings.

Because of these challenges, we must investigate ways to develop and apply quality control measures for data from providers as well as for our own data. While the best approach to improving quality of data is to start at the source, the sensors, we must recognize and work with what is within our control. As aggregators of data from sensor networks controlled by other entities, we have no control over the sensor data and the feeds that deliver the data to us. Still, we can use provider quality control measures and apply our own to best utilize data provided to us from other sources.

For instance, we can evaluate the spatiotemporal coverage of provider data in the presence of multiple, overlapping providers. We can seek answers to questions of whether to include data from one provider relative to others. For instance, what do we gain in terms of coverage by using data from one provider versus two, and what is the cost in terms of bandwidth and storage? What is the overlap in data from multiple providers? Does it improve spatial and temporal coverage?

Quite often, sensor-level quality control processes utilize domain-specific, rule-based systems or general outlier detection techniques to flag “bad” values. Yet quality and performance in general need to be assessed in further dimensions that account for spatial and temporal aspects of applications. For instance, we may want to maximize visual “coverage” of a map displayed in a web application at “critical usage times” with “good” data values while working within limited bandwidth. Such problems involve

multiple, conflicting objectives, making them challenging to solve. And formulating such problems is challenging too because, by definition, we generally first view “quality” as being more subjective than “quantity”. Thus, our challenge is to express quality in quantifiable terms.

Multiple quality dimensions must play a role in the necessarily complex techniques of assessing sensor accuracy, creating a chicken and egg problem. If we assess the quality of an observation by comparison to neighboring observations, then we must also account for the quality of those neighboring observations. Obviously, if neighboring observations are not accurate, then our comparisons to them for accuracy will lack validity. But the problem goes beyond accuracy of the observations. We must also account for timeliness and reliability. And we must account for accuracy of metadata. If an observation is incorrectly located or if the timestamps of observations are incorrect, then we may be comparing against observations that are correct, but are distant in terms of time, location or both.

The consequences of ignoring data quality are summarized by the well-known adage, “Garbage in, garbage out.” How can we trust our applications and models if the inputs are bad? In turn, how can we assess our data for quality so that we can be confident in its use?

BACKGROUND

Our Work at WTI with Caltrans

Since 2003, the Western Transportation Institute (WTI) at Montana State University (MSU), in partnership with the California Department of Transportation (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. As such, it is critical that we display quality information, yet assessing the quality of the data remains a challenge.

The WeatherShare system [1] was 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 each deployment costing as much as $70,000 or more, it is unrealistic to expect pervasive coverage of the roadway from RWIS alone.

In Phase 1 and Phase 2, WeatherShare aggregated Caltrans RWIS data along with weather data from other third-party aggregation sources such as NOAA's Meteorological Assimilation Data Ingest System (MADIS) [2] and the University of Utah's Mesowest [3] to present a unified view of current weather conditions from approximately 2,000

stations within California. A primary benefit of the system was far greater spatial coverage of the state, particularly roadways, relative to the Caltrans RWIS network. A desired secondary benefit of the system was the ability to compare RWIS readings with those of other nearby sensors to assess accuracy. Formal, automated quality control procedures were implemented in the WeatherShare system to assess sensor accuracy for not only Caltrans RWIS, but also all other sensor readings stored in the system.

The Phase 1 and Phase 2 systems achieved the primary goal of enhancing Caltrans personnel's ability to assess current and prospective short-term weather conditions and act accordingly. An emphasis was placed on creating an easy to read and understand at a glance interface. The system did not fully achieve the secondary goal of increasing Caltrans ability to assess RWIS sensor accuracy in an efficient manner guided by automated procedures. This apparent shortcoming stems from both limitations in implemented procedures, as well as unrealized potential to deliver this information through an easy to use and informative interface.

We implemented automated quality control procedures for identifying of "bad" data with limited success in the WeatherShare system principally to assess sensor accuracy for Caltrans RWIS and also for all other sensor readings stored in the system. Quality control indicators from providers such as MADIS were considered for use from time to time, but differences across providers limited our usage of these indicators. Within the scope of the WeatherShare project, we have not made a concerted effort to reconcile these differences, let alone to formally quantify and evaluate the impact of

quality control, and optimize the performance of these systems relative to quality control measures.

Ian Turnbull, Chief ITS Engineer at Caltrans District 2 in Redding was the original project champion for the WeatherShare project, and has been involved with all subsequent, related projects. It is from Ian that we are given the directive of providing "accurate, timely and reliable" data. Our efforts have generally handled this directive informally as qualitative rather than quantitative. Given our experience with these projects, we recognize the need for more formal, quantitative handling of multi-dimensional quality measurement. To date, such measurement has been elusive for a variety of reasons. For instance, sensors are not uniformly distributed through the state of California, yet we desire at a high level a uniform presentation of sensor data in a spatiotemporal sense in certain displays.

Figure 1 shows the Current Air Temperature layer from Caltrans' Aviation Weathershare (aviation.weathershare.org) system. Notice that approximately 150 observations are shown, and these observations visually cover most of the state. It is not viable to show all observations at this zoom level despite there being potentially 1000 or more recent observations available at a given point in time. The map would be too cluttered to read, and there would also be application performance issues associated with display overhead. Yet it is desirable to select a subset sufficient to show representative conditions throughout the entire state.

Figure 2 shows the Current Air Temperature layer in proximity to Los Angeles. At this zoom level it is apparent that the spatial distribution of sensors is not uniform. There are many sensors reporting observations from downtown Los Angeles while there are relatively few along Interstate 15 and Interstate 40 in proximity to Barstow. Notice also apparent bad data. There is an 11 degree reading reported near Oceanside along the coast which is obviously incorrect because no neighboring sites present readings that cold. By similar reasoning can we also conclude that the 35 degree reading near Palm Desert is incorrect? In this case, even though other readings near that location are higher and some are much higher, the observation may well be correct since it was taken near the top of a mountain.

Bad readings, inconsistent reporting of observations, and non-uniform distribution of sensors make it challenging to present an "accurate, timely and reliable" depiction of current conditions across the entire state. Yet it is important to present "quality" data in a timely manner so users can recognize changing weather conditions such the passing of a cold front, which in turn might cause icy roadways, or icing conditions in the air. Strong winds are problematic for both surface transportation and aviation, and precipitation, especially when combined with below freezing temperatures present significant hazards.

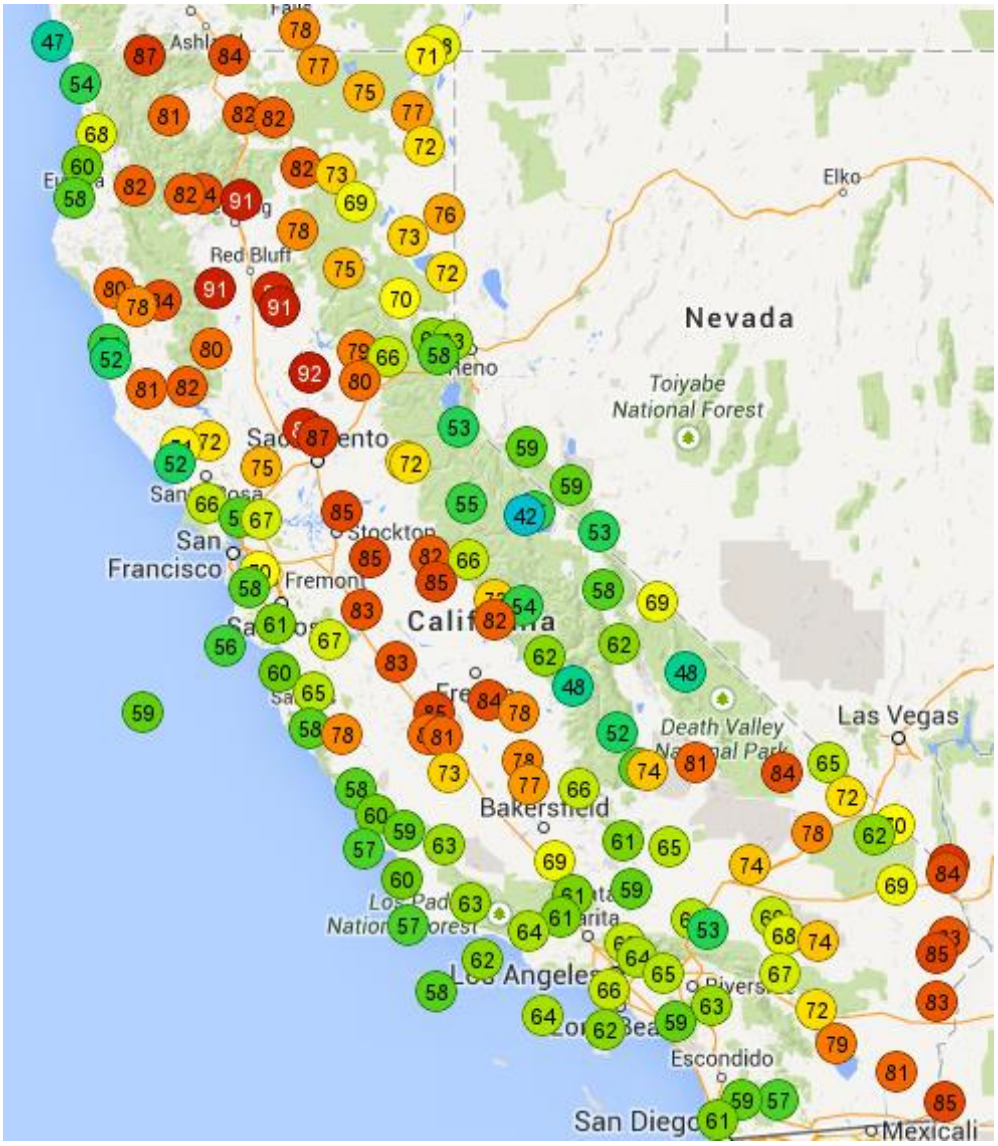

Figure 1. Current Temperature Layer from aviation.weathershare.org

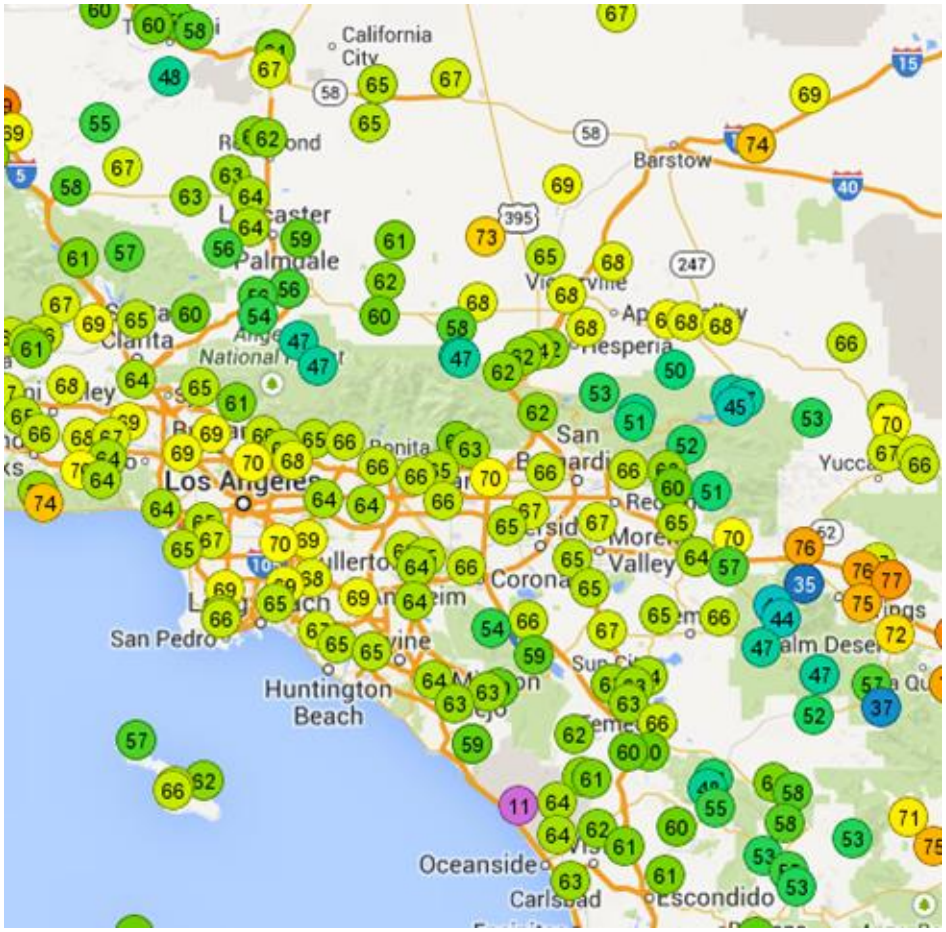

Figure 2. Current Temperatures zoomed to Los Angeles in aviation.weathershare.org

WeatherShare, now in its Third Phase, is set to become the official repository and internal access point for Caltrans RWIS data. This new emphasis presents an opportunity and challenge for revisiting the assessment of quality of the Caltrans RWIS data. The fundamental challenge of quality assessment here is related to that of the map displays above, but places greater emphasis on individual sites and sensors. For instance, sensors at a given site may be linked to safety warning systems or may be used by operations personnel to make key decisions regarding chain control or road closures. Thus, if the sensors are not presenting "accurate, timely and reliable" information, the safety of the traveling public is at risk. In such situations, we cannot simply discard observations from the site as bad, but must recognize that there is a problem so that it can be corrected.

For many of our experiments to date related to our PhD work, we have focused on the Caltrans District 2 Dunsmuir RWIS, referred to subsequently as CTDUN. See Figure 3. We have focused our attention in these experiments to ambient air temperature. CTDUN sits near the northern extreme of the Sacramento Canyon. Starting north of Redding and extending nearly to Mt. Shasta, the Sacramento Canyon progresses along Interstate 5 from an elevation of approximately 500 feet to over 3900 feet near the base of Mt. Shasta. Dunsmuir sits at approximately 2450 feet. Mt. Shasta has an elevation of over 14,000 feet and the Trinity Alps, to the west of Dunsmuir, rise to as much as 9000 feet. Separate from Mt. Shasta, the area to the north of Dunsmuir is arid relative to the Canyon area to the south, and rain forest is present near the coast less than 100 miles to the west. With varying terrain and climate, the spatial neighborhood of CTDUN is far from

homogeneous making it an excellent site for testing and evaluation of quality assessment techniques. See Figure 4.

Figure 3. Photo of the Caltrans Dunsmuir RWIS Tower

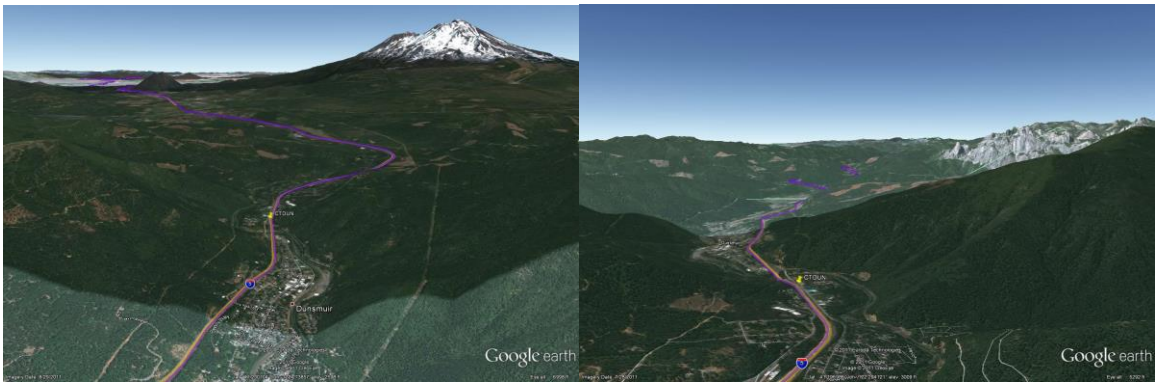

Figure 4. Surrounding Area Facing North and South from CTDUN

In January 2010, there was great variation in temperature at CTDUN. This variation is generally periodic, with an expected diurnal effect, but is also influenced by weather systems, particularly in the winter. There are some interesting non-periodic episodes within this time period. Specifically, there is a time in which the temperature hovers slightly above freezing from 1/20/2010 to approximately 1/25/2010. See Figure 5. Are readings from this period erronous? Several things are suspect. First, the readings approach but never cross below freezing. Second, the range of the data during this period is approximately 8 days, and there are several sub-periods that span more than 24 hours in which there is almost no variation. It is challenging to determine if this data is erroneous when looking at this time series alone. In subsequent figures, we compare to readings for the same period for the other nearby Caltrans sites.

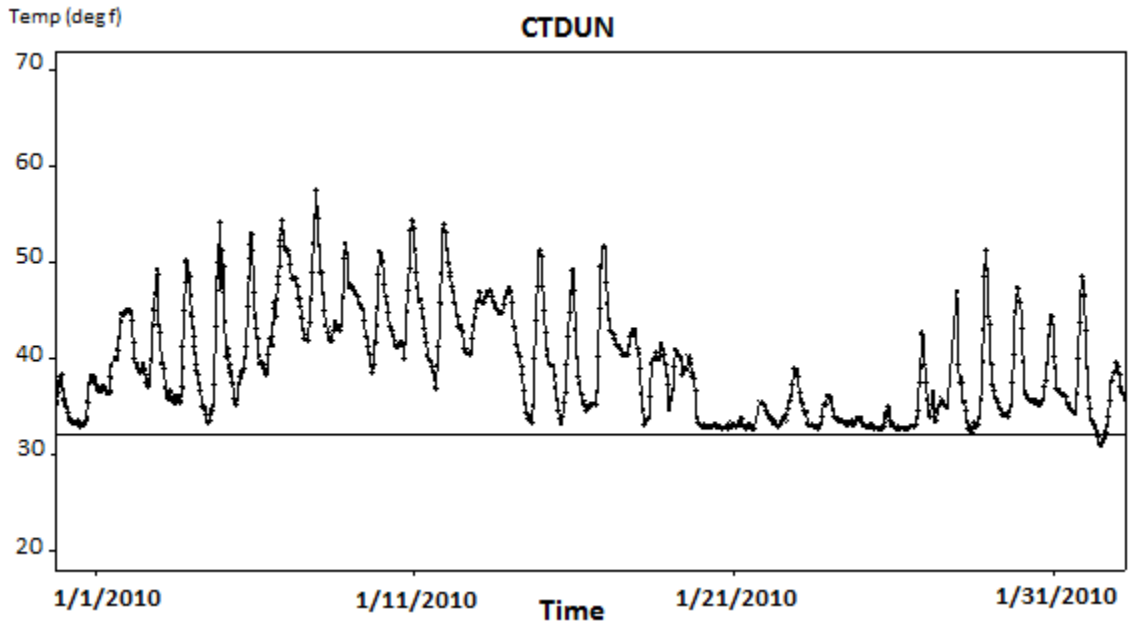

Figure 5. Air Temperature at CTDUN in January 2010

Black Butte Summit sits north of CTDUN and is the location of another Caltrans site, CTBBS. Readings from CTBBS generally agree with those from CTDUN. Note however the absence of data for this site between 1/20/2010 and 1/24/2010, overlapping the questionable data from CTDUN. The cause of this gap is unknown. It could be a consequence of communication problems and/or problems with equipment at the site. Or, it could be that data for this period were removed due to failure of Level 1 (range) quality checks. See Figure 6.

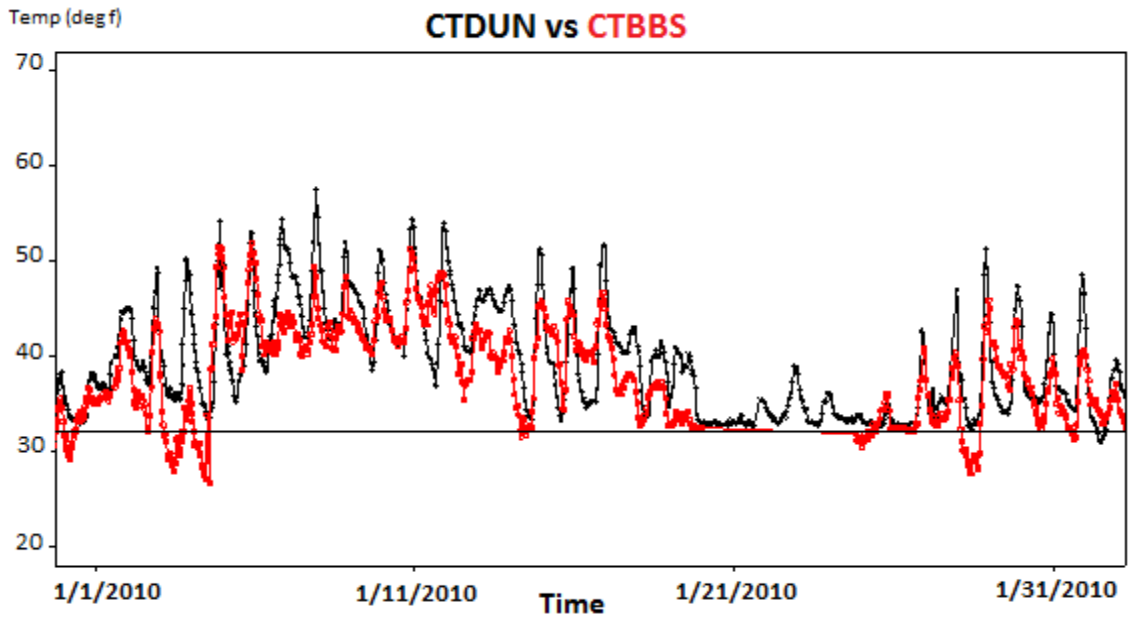

Figure 6. Air Temperature at CTDUN and CTBBS in January 2010

CTSNS is located on Snowman Summit, south of Mt. Shasta. While close to CTDUN, there is significant variation in terrain between the two sites. Still, there is general agreement between the two sites. Note that there are two periods from approximately 1/20/2010 to 1/31/2010 in which data was not available from the site, again overlapping the questionable period from CTDUN. Note further a period immediately prior to 1/21/2010 in which the data shows very little variation, similar to that observed subsequently at CTDUN. See Figure 7.

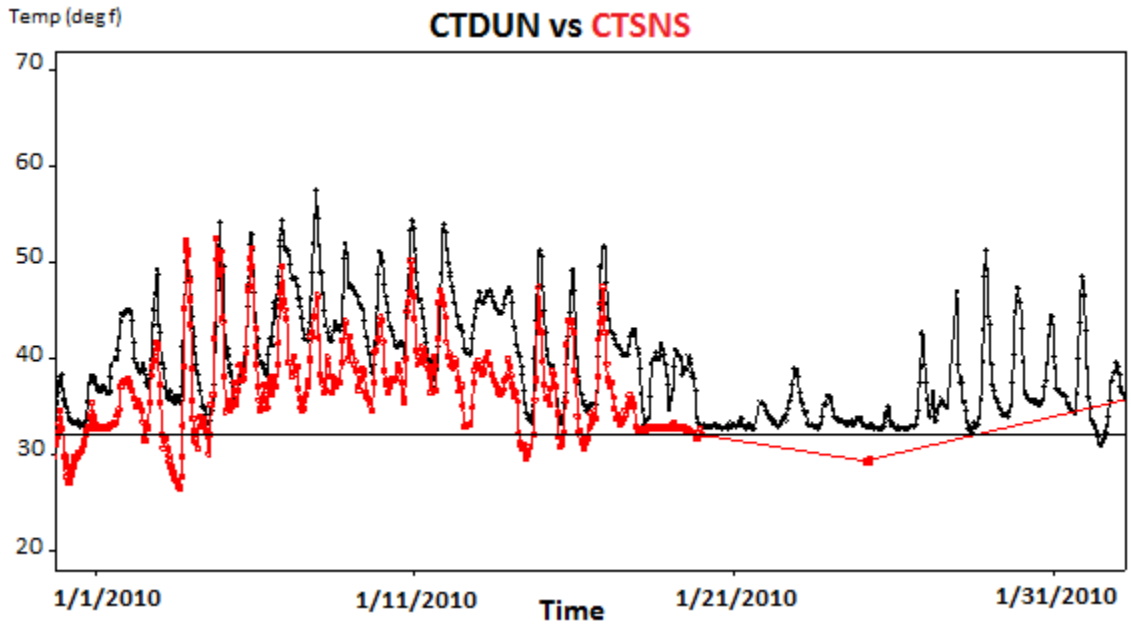

Figure 7. Air Temperature at CTDUN and CTSNS in January 2010

CTVOL, another Caltrans site located south of CTDUN at Vollmers also shows agreement with CTDUN, although it does appear to generally be slightly warmer than CTDUN. Note several apparently erroneous readings of 32 degrees on approximately 1/12/2010 and 1/13/2010. See Figure 8.

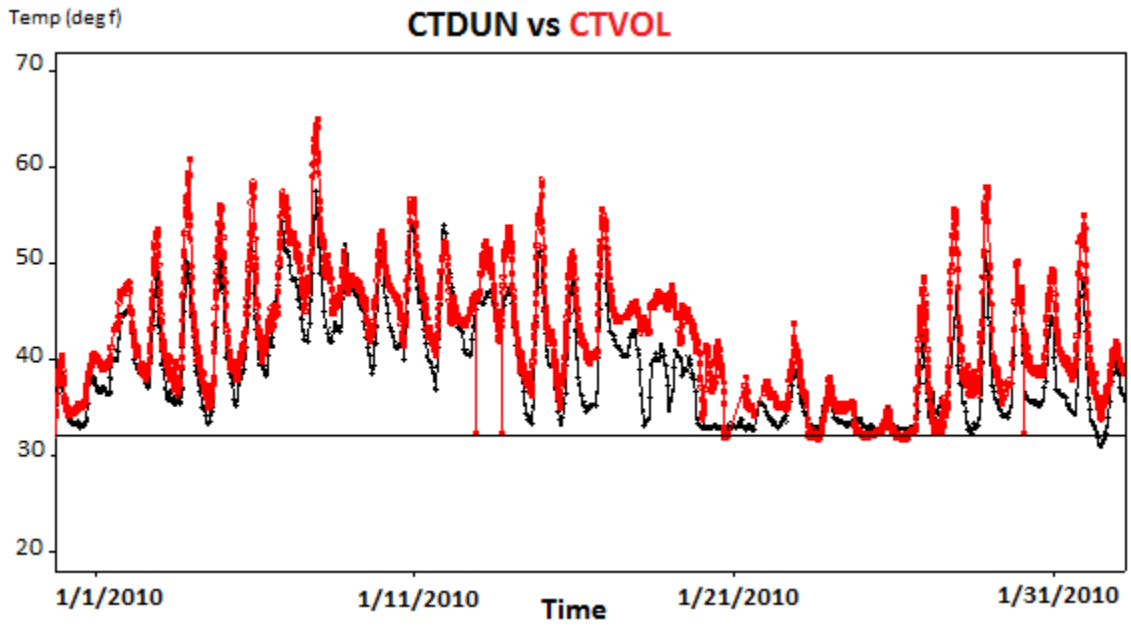

Figure 8. Air Temperature at CTDUN and CTVOL in January 2010

From these ad-hoc comparisons we can begin to see some of the challenges in assessing spatiotemporal data quality. As expected, there is (co)variation in data between neighboring sites, yet the data does not match perfectly. Neighboring sites have their own issues with accuracy, timeliness and reliability. And, there is not always an observation with which to compare directly.

Weather Data Provider Approaches

For help in addressing the challenges of spatiotemporal data quality assessment, we look first and foremost to the weather community, including providers we have used in our systems.

MADIS

NOAA's Meteorological Assimilation Data Ingest System (MADIS) is an online database of real time and archived weather data including sensor readings over 34,000 sites in North America, including Hawaii and Central America [2], [4]. MADIS implements automated quality control processing for data in its meteorological surface dataset and states that, "considerable evidence exists that the retention of erroneous data, or the rejection of too many good data, can substantially distort forecast products." [5] Data is stored with quality control flags indicating the results from various quality control checks, and this information is available to all users of the data.

For their meteorological surface dataset, MADIS implements three levels of automated quality control [6]. The "Level 1" quality control checks are also referred to as "validity checks" or "range checks". They check that a sensor reading is within a range of predetermined values indicating the "tolerance limits" of a value. The range for air temperature in degrees Fahrenheit is given as [-60, 130]. There are three "Level 2" quality control checks: "internal consistency," "temporal consistency," and "statistical

spatial consistency." "Temporal consistency," also sometimes referred to as a "delta test," compares the change in a value over a period of time against a reference maximum value. For instance, if the air temperature at a site changes by more than 35 degrees Fahrenheit within an hour, then the current reading is flagged as failing. The "internal consistency" checks are somewhat more complex and can include cross-sensor comparison. For instance, if the dewpoint temperature exceeds air temperature at a site, then both are flagged as having failed the internal consistency test. There are other more complex internal consistency checks for readings such as sea level pressure that operate on the recent historical values for a sensor. Level 2 "statistical spatial consistency" checks are used in conjunction with Level 3 spatial checks, described below, and the other quality control checks. The statistical spatial consistency check flags an observation as failed if it has failed any other quality control check for 75% of the time during the previous 7 days. It continues to be marked as fail until its failure rate decreases to below 25% for a weekly period.

The Level 3 quality control check is referred to as the "spatial consistency" or "buddy" check, and is a variant of the Optimal Interpolation (OI) technique [7]. For a given site and observation, an interpolated value is determined for that site using neighboring sites and excluding the site being analyzed. If the difference between the actual value and the interpolated value is "small," then the site is considered to be in agreement with its neighbors and it passes the spatial consistency check. However, if the difference is not small, then the interpolation and analysis is repeated with one of the

neighboring observations removed. If the removal of the neighbor results in a small difference between the interpolated and observed value, then the observation is flagged as "good" and the neighboring observation is flagged as "bad". This process is repeated by replacing the neighbor that was removed and removing another neighbor. If none of the removals results in a small difference between actual and interpolated, then the original observation is flagged as bad.

Where possible, the spatial consistency check uses uniform selection of neighbors determined by the nearest in each of eight directional sectors around the site being checked. Further, Rapid Update Cycle Surface Assimilation Systems (RSAS) fields [8] are used as background grids. The RSAS grids provide interpolated values at up to one hour frequency. The differences between the RSAS interpolated values and the observed values are used to perform the interpolation in the OI analysis. Problems such as temperature inversions, in which terrain may have a significant impact on gradients associated with conditions, are noted.

Clarus

The *Clarus* initiative was established in 2004 by USDOT Federal Highway Administration Road Weather Management Program and the Intelligent Transportation Systems Joint Program Office to "reduce the impact of adverse weather conditions on surface transportation users." [9] Specifically, Clarus was developed to collect

atmospheric and pavement observations from state-owned road weather information systems (RWIS) environmental sensor stations (ESS) in near real time.

The *Clarus* System provided ESS data from participating states and Canadian provinces. *Clarus* implemented ten quality control algorithms including a Barnes Spatial Test and an IQR (Inter-Quartile Range) test. The *Clarus* team indicated that there were no identified algorithms for pavement and subsurface readings and that other algorithms are lacking in which other background data sets such as radar are used, and a plan was made by the team to make improvements to existing algorithms. They further noted that many of the parameters used in quality control algorithms are not presented with the data, leaving users with limited information to determine the nature of failure.

The Barnes Spatial Test, is based on the Barnes Spatial Interpolation Scheme [10], which uses a Gaussian inverse distance weighting scheme to interpolate values over a two dimensional area using known readings within that area. Sites included DOT RWIS sites as well as aviation Automated Weather Observing System (AWOS) / Automated Surface Observing System (ASOS) sites.

The Clarus team indicated a need to add more coastal and marine observations to aid in checks of coastal stations. Further, they questioned whether Optimal Interpolation should be used instead of Barnes Spatial Analysis since it might work better in mountain areas. Neither method incorporates terrain data, but Optimal Interpolation may better account for terrain by using directional sectors to include neighbors in all directions. Another recommendation made by the Clarus team was to develop an “overall

observation confidence value" that uses the results of all quality control flags to create a single value that describes the quality of the observation.

The Barnes Interpolation Scheme is used as a basis for the Barnes Spatial Quality Control test, as applied by the Oklahoma Mesonetwork [11], which uses one pass of the Barnes Interpolation Scheme to estimate values for each observation as follows:

- All stations within the radius of influence except the station for the observation being evaluated are used to estimate a value for the observation. Stations for which there are other quality control failures are excluded.
- Basic statistics including the mean, median and standard deviation are calculated for the observations of the stations within the radius of influence.
- A z-value is calculated using the difference between the observed and estimated value and the standard deviation of the observations:

$$\Delta = \frac{z_e - z_o}{\sigma}$$

- Flags are set as follows:
    - "Suspect" flag for differences exceeding twice the standard deviation.
    - "Warning" flag for differences exceeding three times the standard deviation.
    - No flag if fewer than six observations were used to determine the estimated value.

- A predetermined threshold for the standard deviation is used to avoid flagging stations when the standard deviation is small.

Mesowest

The Utah Mesonet (Mesowest) uses multivariate linear regression [12] to incorporate elevation into an interpolation model and subsequent quality control checks for temperature, dewpoint and pressure. The model uses *(x,y,z)* coordinates to represent station location, including elevation, along with the temperature observation *T* at that point, as:

$$\hat{T} = T_0 + \frac{\delta T}{\delta x}\Delta x + \frac{\delta T}{\delta y}\Delta y + \frac{\delta T}{\delta z}\Delta z$$

The solution is:

$$\begin{bmatrix} n & \sum x & \sum y & \sum z \\ \sum x & \sum x^2 & \sum xy & \sum xz \\ \sum y & \sum xy & \sum y^2 & \sum yz \\ \sum z & \sum xz & \sum yz & \sum z^2 \end{bmatrix} \begin{bmatrix} T_0 \\ \frac{\delta T}{\delta x} \\ \frac{\delta T}{\delta y} \\ \frac{\delta T}{\delta z} \end{bmatrix} = \begin{bmatrix} \sum T \\ \sum Tx \\ \sum Ty \\ \sum Tz \end{bmatrix}$$

It is noted by Mesowest that time could also be included in the model but that there are associated non-linearity issues that cause it to add no value. The regression is run over a six hour period and observations are compared to values interpolated from the regression. If the difference between predicted and observed exceeds 10 degrees Fahrenheit, the observation is flagged. The justification for using such a long time period

for the regression is that shorter time intervals might be more dramatically affected by real weather conditions, resulting in greater natural departures in observations from the model. Further note that in MesoWest feed data, a single flag is used to represent data quality for a site rather than individual observations. Thus, users of the feed cannot directly determine which sensors may have failed quality control based on the flag alone since multiple sensor values are typically associated with a site.

The MesoWest methodology using multivariate linear regression is noted as useful in making longer term comparisons for observation versus predicted, and that it has helped to identify other errors such as mis-recording of station elevations. It is noted that the model may not account for small scale events. An enhancement uses the Barnes Spatial Interpolation Scheme to add differences between the regression fit and the model back into the model. The model is demonstrated primarily in Utah, which is characterized topographically by the Great Basin and the Wasatch Mountain Range, providing significant diversity in terrain, but applied to data stored by MesoWest, representing the entire United States and more.

PRISM

PRISM (Precipitation-elevation Regressions on Independent Slopes Model), developed at Oregon State University, accounts for elevation and general topographic impact on weather variation, creating a grid of estimated precipitation using station readings that fall within topographic facets [13]. PRISM is based on a key underlying phenomenon called the orographic effect: on a given mountain slope, precipitation generally increases with elevation. The system uses "facets", which are contiguous areas over which slope orientation is relatively constant, to group stations into regions for which a precipitation event will likely impact the entire area, with variation corresponding to the orographic effect.

Rather than compute a regression over all data, PRISM uses a simple linear regression for each grid cell *(i,j)* as follows:

$$P_{ij} = b_{0_{ij}} + b_{1_{ij}} E_{ij}$$

where the parameters are the intercept, slope and elevation of the cell. The values used for the regression are restricted to the stations within a predefined radius and within the same facet as the cell. If not enough cells are found to perform the regression, then the area is expanded to include other similar, adjacent facets. If no stations are available in similar facets, the nearest station is used in the calculation. PRISM indicates that 95% prediction intervals are determined using standard methods for calculation of prediction intervals for linear regression.

PRISM has been applied to measures other than precipitation, including temperature and snowfall. [14] Several meteorological phenomena are documented in which assumed variation with elevation may be violated. Temperature inversions in mountain valleys cause the normal relationship of temperature decreasing with elevation to be inverted, with colder temperatures pooled in a valley and higher temperatures above. It is also possible for coastal precipitation to occur in a layered fashion in which higher elevations, those above the layer, experience less precipitation than lower elevations. The PRISM model has been extended to include multiple layers to represent boundaries between such phenomena.

Although the basic regression model is simple, stations are weighted within the model relative to the grid location using distance, elevation, cluster, vertical-position, topographic-facet, coastal-proximity, topographic position and effective terrain weights. For instance, a "topographic index" grid is used to describe the height of a pixel relative to the surrounding terrain height.

PRISM is used for quality control of Snow Telemetry (SNOTEL) data, operating on the assumption of spatial consistency [16]. A station's data is withheld from the interpolation to estimate the value at the station's location. If there is a large difference between predicted and actual values, the actual value is suspect. In an approach similar to the Optimal Interpolation method described earlier, observations are removed individually from the model to recheck for consistency.

The PRISM group states that interpolation for daily temperature can be improved significantly if a predictive background grid is used, representing the long-term climatological temperature for the corresponding day or month. It is further noted that 30 year means are the only such grids available, and that there may be problems with patterns that deviate significantly from the mean. They indicate that if averages could be separated into classes such as wet days versus dry days, then the results would be even better. Obviously the PRISM model is heavily dependent on meteorological expertise.

WeatherShare

The (original Phase 1 and 2) WeatherShare System implemented a modified subset of the quality control procedures used by MADIS, incorporating Level 1, Level 2 and Level 3 quality control checks. [15] The Level 1 range check (-60°F to 130°F) for air temperature is used, as is the Level 2 delta check for changes of 35°F or more in one hour, as well as a check for no change over 24 hours.

WeatherShare Level 3 quality control consists of spatial consistency checks using a linear regression model similar to that used by Mesowest. This method uses the values surrounding a suspect station and attempts to predict an acceptable range of values for the value being checked. Level 3 quality control is currently implemented for air temperature only. WeatherShare applies Level 3 quality control in a two-step process:

Step 1: First, all temperature observations within the past 90 minutes for all current weather stations as well as the station location data (latitude, longitude and elevation) are used to calculate a linear regression model, and the residuals (observed value minus predicted value) are calculated. For every station with an absolute residual greater than 40, a further normality check is processed as Step 2.

Step 2: All nearby (within 25 miles) station temperature observations within the last 90 minutes are gathered. The resulting temperature observations as well as latitude, longitude, and elevation data are used to estimate a new multivariate linear regression model. With this new model, if the reported temperature data is different from the predicted regression value over the last 90 minutes by 10°F or more, the observation is flagged as "failed."

Tuning and selecting proper parameters for this method has proven non-trivial because there is no exact "right answer" to the choice of cutoff values. A trial-and-error approach was used to select values that balanced false positives with false negatives. The value of 40 degrees for the first step is large and not ideal, but was chosen to reduce the number of false positives. The need for a high tolerance certainly may be a consequence of a poor fit for the regression model across the entire state. However, the model has proven itself useful in discovering anomalies including the movement of portable weather stations in which their subsequent location was not reported correctly. Level 3 spatial quality control has been considered experimental throughout the project. The WeatherShare project team has been hesitant to implement spatial quality control checks

for other measures, particularly wind and precipitation, assuming that their variation will be greater than that of temperature.

Spatiotemporal Data Quality Literature

Data quality from the perspective of the consumer is presented subjectively in [16], as a comprehensive framework of data quality attributes. A more recent survey and summary of data quality dimensions from the literature is presented in [17], and points out varying definitions for dimensions such as timeliness and completeness. Data mining tools to assist in data quality assessment are developed in [18], and a definition for data auditing is presented to include measurement and improvement of data quality. Overlap and differences between Quality of Data and Quality of Information (QoI) are presented in [19]. While these papers are useful in general terms, they do not include specific, comprehensive measures that can be applied to our spatiotemporal situation.

A comprehensive review of spatial data quality is presented in [20], which includes some treatment of temporal aspects, and distinguishes between internal and external quality. Internal quality includes dimensions such as accuracy, completeness and consistency, while external quality is defined as fitness for use or purpose. They also cite and expand on prior work which presented six characteristics of external quality for geospatial databases: definition, coverage, lineage, precision, legitimacy, and accessibility. Sources of uncertainty in spatial-data mining are presented in [21], and can also be viewed as sources of data quality problems.

Data cleaning is presented in the context of pull-based and push-based data acquisition in [22], along with a model-based approach to outlier/anomaly detection. An

adaptive query system for systems integrating overlapping data sources is presented in [23], including query optimization, while the trade-offs between accuracy and timeliness of information acquired in a data aggregation network are investigated in [24]. Also from the networking domain, rate control guided by Quality of Information (QoI) measures is focused on in [25]. They indicate that such efforts are highly application-dependent. Four components of data quality: accuracy, consistency, timeliness and completeness are presented in another network-related publication [26] and timeliness is expressed principally as a network phenomenon. Completeness, currency, internal consistency, timeliness, importance, source reliability and confidentiality for cooperative web information systems are defined in [27].

The closest and best defined work in relation to ours is presented in [28][29][30][31][32], although this work is presented in relation to the transfer and management challenges of including quality control information in data streams and in optimal, quality-based load-shedding for data streams. Specific measures presented include accuracy, confidence, completeness, data volume and timeliness.

Quality of Service (QoS) is used for load-shedding in [33] while noting that conflicting objectives are common, and is also presented in the context of operator scheduling in [34]. Load-shedding for spatio-temporal data streams is presented in [35], but does not specifically address quality control measures. Other relevant work regarding load-shedding for data streams can be found in [36][37][38][39][40][41]. An automated metadata generation approach that includes a probabilistic measure of data quality is

presented in [42]. Three quality attributes for the detection and identification of human presence in multimedia monitoring systems are dynamically assessed in [43] Data quality in relation to e-Health monitoring systems is presented in [44].

None of these approaches directly addresses quality control for spatiotemporal data that is immediately applicable to our situation. However, the data quality attributes presented may be of benefit. As such, the methods used by the weather data providers appear to be state of the art.

APPROACH AND WORK DONE TO DATE

To date we have published work related to this study in three computer science conferences. We have also presented our applications at numerous transportation-related conferences. Following is a summary of the work presented at the computer science conferences. We have made a progression to high visibility conferences relevant to our work, culminating in a presentation at ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems 2014.

FLAIRS 2012

Our first published work related to our proposed dissertation topic was published in conjunction with the International Florida Artificial Intelligence Research Society Conference (FLAIRS) in 2012. [45] In this paper we demonstrated improved results by using a data-driven approach to inverse distance weighting versus a purely geographic approach.

We investigated air temperature observations using data from MADIS covering a rectangular region that includes all of California and portions of Oregon, Nevada, Idaho and Arizona spanning July 2001 through December 2010. Table 1 shows quality control descriptors associated with each sensor reading in the MADIS data set. The dataset provided additional detail indicating which tests were applied and which resulted in failure for the reading.

Table 1. MADIS Quality Control Descriptors

B	subjective bad
C	coarse pass, passed level 1
G	subjective good
Q	questioned, passed level 1, failed level 2 or level 3
S	screened, passed level 1 and level 2
V	verified, passed level 1, level 2 and level 3
X	Rejected/erroneous, failed level 1
Z	preliminary, no quality control check

Although not an emphasis of this study, we pre-processed data using a [-60° F, 130° F] range check, in conformance to the test used by MADIS. It is recognized that such preliminary checks and filters may be key to the performance of the more advanced (Level 3) spatial checks. In effect, this pre-processing removed all observations having an "X" quality control descriptor. There were some observations in the data set having quality control descriptors other than "X" which also failed this range test and these were removed too.

In this paper, we used the Barnes Spatial Interpolation Scheme (an inverse distance weighting method) for comparison and enhancement using data-driven techniques. In general, California offers an ideal setting for the evaluation of quality control procedures because of its geographic and meteorological diversity. California includes coastal areas, mountains, deserts, rain forests, and both the highest and lowest points in the contiguous 48 states. While the Barnes Spatial Interpolation Scheme is widely applied, it is also susceptible to the challenges of varying terrain.

For methods such as the Barnes Spatial Interpolation Scheme, our contention was that naïve assumptions of uniform spatial proximity could be replaced with time-series distance to indicate (dis)similarity of stations based on historical data. Further, stations could be grouped based on this same time-series distance or (dis)similarity measure to form a radius-based or nearest neighbor-based grouping, per station.

Time series distance can be computed in many different ways. In our experiment, we implemented an online approach, which continually updates the time series distance between stations as new observations are reported by using the sum-of-squares difference between each observation and the most recent observations from other stations.

Using this measure, we developed a method that we named Time-Series-Distance-Filter Interpolation (TSDFI) as a variation of the Barnes Spatial method by replacing the station to station geographic distance measure with our time series distance measure.

Both the Barnes Spatial method and our new TSDFI method use inverse distance weighting by way of a Gaussian function to interpolate values, creating a model of reported data. Both were implemented in an online fashion, with the dataset processed chronologically, modeling each sensor value in the dataset. The error of reported versus predicted was recorded for each sensor value and subsequently aggregated as mean-squared-error for comparison.

Over 233 million temperature observations from 2001-2010 and the corresponding predicted values using Barnes Spatial (Barnes) and the Time Series

Distance Filter Interpolation (TSDFI) methods were analyzed. The TSDFI method had an overall mean-squared error of less than half that of Barnes over the entire data set.

The mean-squared error for TSDFI is consistently less than that for Barnes Spatial over time by nearly a factor of two. Figure 9 shows peaks and troughs in the MSE for both methods, with peaks occurring in proximity to June of each year and troughs in proximity to December. Further investigation is necessary to determine if this is a consequence of normal seasonal variability in the underlying data and whether there is a need to account for such variability in the models. For instance, should there be separate summer and winter models?

The mean-squared errors for the Barnes and TSDFI methods yield promising results when grouped by observations according to the MADIS quality control descriptors. Recall that all readings flagged with quality control descriptor X, rejected/erroneous, were removed prior to application of the interpolation methods. For those readings that passed all three levels of MADIS quality control, labeled V, Table 2 shows that the mean-squared error for TSDFI is very small, and approximately one-third that of the Barnes Spatial method. For those flagged Q for questionable, having failed level 2 or level 3 in MADIS, the MSE for the TSDFI method is very high and comparable to that of Barnes Spatial. This is not problematic since the high MSE is attributable to questionable and likely erroneous sensor readings rather than model error, and would be indicative of such readings. Results corresponding to other MADIS quality control

descriptors show similar results and appear to indicate that the TSDFI method may discriminate valid versus erroneous readings relatively well.

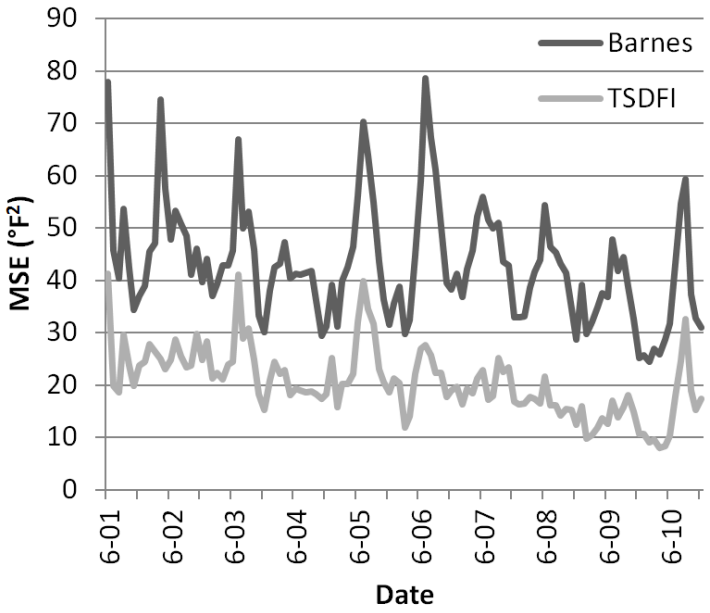

Figure 9. MSE ($°F^2$) Over Time for Barnes versus TSDFI

Table 2. MSE for Barnes versus TSDFI by MADIS QC Indicator

MADIS QC Indication		**Count**	**Barnes MSE**	**TSDFI MSE**
subjective bad	B	300	55.82	60.54
coarse pass, passed level 1	C	439,630	109.88	51.57
subjective good	G	206,873	23.01	15.52
questioned, passed level 1, failed level 2 or level 3	Q	13,709,191	246.13	158.28
screened, passed level 1 and level 2	S	21,337,164	38.93	16.32
verified, passed level 1, level 2 and level 3	V	197,415,497	25.86	8.01
Rejected/erroneous, failed level 1	X	*not counted*	-	-
preliminary, no quality control check	Z	28,862	49.12	22.42
		Overall	**40.17**	**17.70**

By comparing TSDFI error to Barnes Spatial error for individual stations, we can speculate on reasons for large differences, including those that may be attributable to quality control problems for stations. Most stations have modest MSE values (less than 100) for both Barnes and TSDFI. There are some that have large MSE for both. See Figure 10 and Figure 11. Those with a large MSE for both measures are stations that are likely producing erroneous data. Those with a low MSE for our TSDFI distance and a high MSE for Barnes Spatial are likely incorrectly located sites. Note that variation in terrain may contribute to other differences in MSE between the two methods for stations that are correctly located and producing correct data.

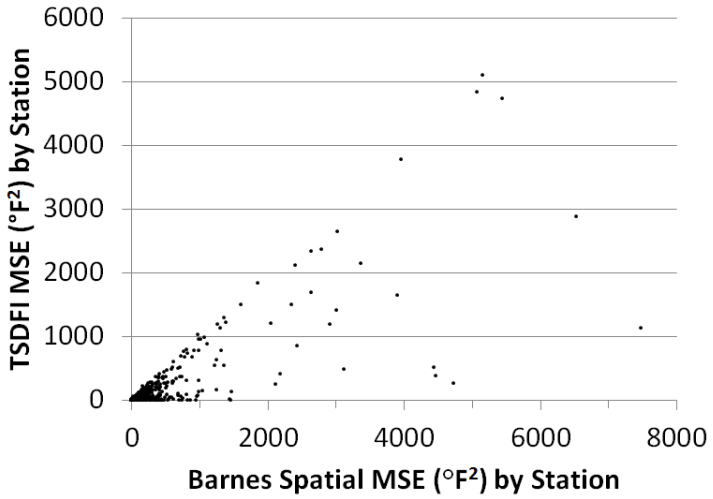

Figure 10. Plot of Barnes Spatial and TSDFI Mean-Squared Errors by Station

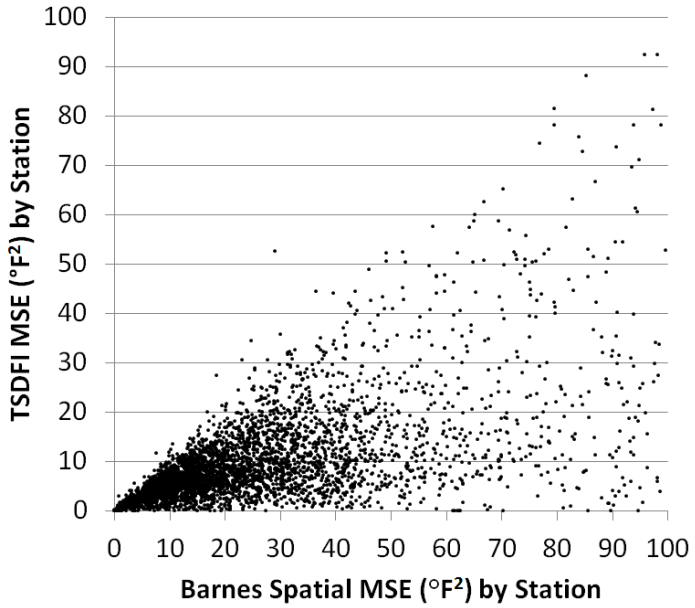

Figure 11. Collapsed Plot of Barnes Spatial and TSDFI MSE by Station

Consider station F2988, for which the reported location is 37.07°N, 119.03°W. MADIS reports an elevation near zero for this station, which is inconsistent with a latitude/longitude of a point in the Sierra Nevada mountain range. The MSE for Barnes Spatial for this station is 89.321. The MSE for TSDFI for this station is 1.513. This indicates that other stations in proximity to this station do not match it well while there are other stations which are better matches, which further seems to indicate that the reported station location is incorrect. However, there are only 22 readings for this station, and this discrepancy may be a consequence of lack of data. For station SCNC1, the MSE for Barnes Spatial is 5,155.291 and the MSE for TSDFI is 5,097.405. This station is located on San Clemente Island, which is over 60 miles from the California coast. While

there are several other stations located on San Clemente Island, there are very few additional stations in proximity. Either this station is problematic in general or there are not enough stations in proximity for comparison using either method. In fact, the closest station in terms of time series distance is station F1426, which is located at Camp Pendleton, north of San Diego and over 60 miles away.

The data-driven method showed sufficient promise to merit subsequent research and development into related methods for quality assessment. TSDFI reduces overall model error in comparison to Barnes Spatial by grouping stations based on similarity of sensor time series and weighting them accordingly rather than by using spatial distance. There is further room for improvement by optimizing parameters for these models, including the potential to vary parameters on a per-station basis. And, it may be worthwhile to investigate a hybrid method that combines both time-series distance and spatial distance. Elevation might also be accounted for directly using similar approaches.

It would also be desirable to investigate varying time periods for both time series distance calculation and prediction. In our investigation we used data from June 2001 through December 2010. It is important to determine how much data is necessary to develop a model that is applicable year-round and to subsequent years. We suspect that seasonal patterns will have an impact on performance.

ACM SIGSPATIAL IWGS 2013

Our second published work related to our proposed dissertation topic was published in conjunction with the 2013 ACM SigSpatial International Workshop on GeoStreaming (IWGS). [46] In this paper we formed neighborhoods using various distance measures and experimented with "robust" techniques to trim extreme data that would otherwise have an adverse effect on the distance measures. In turn, we used inverse distance weighting to compute interpolated values for comparison against original observations.

For our experiment, we focused on neighborhoods for the Caltrans District 2 Dunsmuir RWIS (CTDUN). We restricted our attention in this experiment to ambient air temperature. We used data exclusively from MADIS for our experiment since it provides quality control descriptors which can be used as a basis for comparison. See Table 1. Note that the MADIS metadata for this site includes incorrect location data that is approximately 0.63 miles to the south of the actual CTDUN site, and we use the MADIS-provided location in calculations requiring location. This deviation is relatively small compared to deviations we have observed from third-party aggregators, including some of several hundred miles or more for certain sites.

For neighborhood formation, we restricted our attention to stations within a bounding box defined between 43.5° N and 31° N latitude and -126.5° W and -112° W longitude. This bounding box includes all of California, all of Nevada, southern Oregon

and portions of Utah, Idaho and Arizona. This area yields observations from 10,294 stations between July 2001 and December 2012. Not all stations report through this entire time period. Over this period, over 506,000,000 temperature observations were recorded for these stations in MADIS.

Table 3 summarizes the training and test breakdowns of the data used for our experiment. We used data from 2001 to 2011 for neighborhood formation. We then tested these neighborhoods on the 2001-2011 data and on 2012 data. We did this using all data in the data set. We also tested separately with neighborhoods formed using MADIS QCD V-only data, indicating that it has passed all quality control tests. We did this to determine if we can process all data and find comparable results to that from preprocessed, verified data. Table 4 and Table 5 show counts by quality control descriptor for the CTDUN site and for all observations in our area of interest.

Table 3. Training and Test Sets

QCD	Train/Test	Test
all	2001-2011	2012
V only	2001-2011	2012

Table 4. CTDUN Observations by Quality Control Descriptor and Train/Test Group

QCD	2001-2011	2012
Q	2327	1138
S	14279	42
V	83010	30933
X	15	38
Z	6	

Table 5. Overall Observations by Quality Control Descriptor and Train/Test Group

QCD	2001-2011	2012
B	300	
C	929,390	
G	957,295	177,621
Q	24,360,030	6,166,404
S	33,063,308	517,959
V	341,216,751	98,893,671
X	341,898	85,279
Z	44,133	
Total	**400,913,105**	**105,840,934**

We restricted our attention further by using observations only from stations reporting at least 10,000 observations during the 2001 to 2012 time period, reporting observations prior to and including 2012 and with reported latitude and longitude that each change by no more than 0.1°. The latter check is made because each observation is reported with a location, and station locations change over time. For mobile stations, such change makes sense. For other stations, this change occurs because station metadata has been updated. We choose to only use stations with relatively consistent location reports. This reduces our pool of stations to a count of approximately 3200.

We then proceeded to define neighborhoods using data from the training sets and apply these neighborhoods to estimate CTDUN observations in the test sets. Deviations from the actual observations are used to evaluate performance. Our methods are defined in detail in the subsequent section.

We computed multiple station to station distances for the subsequent formation of neighborhoods. The distances were defined as follows:

- *Great Circle Distance*, $G(s_d, s_i)$, is the great circle distance, in miles, between station s_d and station s_i. Use of this distance results in a method equivalent to the Barnes Spatial Test[47][10][11].
- *Direct Data Distance*, $D(s_d, s_i)$, is the root mean-squared error when using the x values as a direct predictor of the y values. This may also be viewed loosely as the Euclidean distance between the time-series associated with the two stations.
- *Shift Distance*, $S(s_d, s_i)$, is similar to the Direct Data Distance, but first shifts the values in the independent series so the mean matches that of the dependent series. This is done because stations that are close in proximity but differ in elevation may follow similar patterns in temperature change, and the station at a higher elevation will generally have lower temperatures. Formally, the Shift Distance is the root mean-squared error when using the x values added to the difference between the mean of the y values and the mean of the x values as a predictor of the y values.
- *Least Squares Regression Distance*, $L(s_d, s_i)$, is similar to Shift Distance, but allows for a more general mapping between the series for the two stations. Least Squares Regression Distance is the root mean-squared error when using the least

squares regression predictor of the y values by the x values as a predictor of the y values.

We further defined variants of the data-related distances by operating on a *trimmed* subset of observation pairings. The intent of the *trimmed* subsets is to remove the extreme differences between the x and y values, particularly those corresponding to outliers from either set. See Figure 12. A value $\delta = 0.1$ was used in the experiments presented here, for retention of the middle 80% of the data. This value was chosen as large enough to exclude the most extreme differences in the data while small enough to include the predominant relationship between the two observation sets.

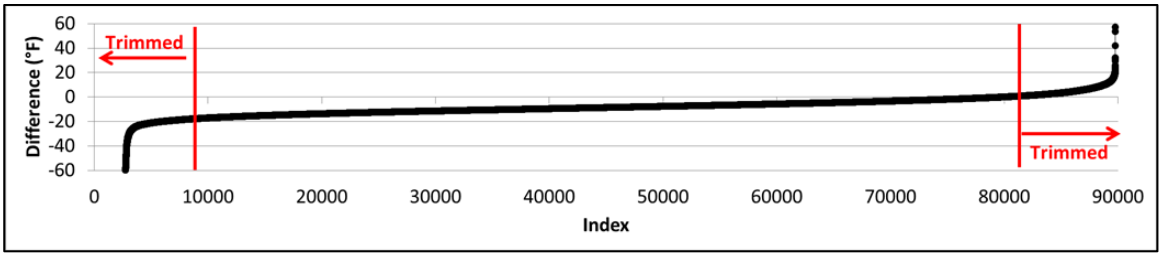

Figure 12. Sorted and Trimmed Differences Between WSHC1 and CTDUN.

Trimming yields three new distance measures based on the measures presented above:

- *Direct Trimmed Distance*, $DT(s_d, s_i)$
- *Shift Trimmed Distance*, $ST(s_d, s_i)$
- *Least Squares Regression Trimmed Distance*, $LT(s_d, s_i)$

Neighborhood membership was defined using weights. The greater a weight, the greater influence a member will have on calculations based on the neighborhood. A Gaussian weighting function is used in the Barnes Spatial Test and we followed a similar approach here where

$$W_{dist}(s_d, s_i) = e^{\frac{-d^2}{\sigma^2}}, \text{ where } d = dist(s_d, s_i)$$

for a given distance function, $dist$, and constant σ. Using the relationship between this Gaussian weighting function and that of the density function for the standard normal distribution, a value of σ may be selected relative to the distance function $dist$ such that a given percentage of the influence of the weights will correspond to a range of the nearest N stations as follows:

$$\sigma = \frac{D}{2}$$

where

$$D = \arg\min_{d \in R^+} |\{s_i \in S : s_i \neq s_d, dist(s_d, s_i) \leq d\}| \leq N.$$

For our experiment, we used $N = 100$ corresponding to the nearest 100 stations relative to each distance function $dist$. Given the large number of overall stations, this was a reasonable choice since it distributes weight over a relatively small subset of stations while including enough stations to account for missing data. The Barnes Spatial Test commonly uses a 70 mile radius hard cutoff. We choose instead to distribute weights

by station count to provide an approximately even weighting across the various methods compared in our study.

Estimates were formed using the weighted mean over the observations for a given neighborhood. A second estimate corresponds to the trimmed mean, which is computed as the weighted mean after removing a percentage of the upper and lower data. Figure 13 shows an example in which the trimmed mean and the original mean differ due to an extreme value. The intent of the trimmed neighborhoods is to remove extreme values which otherwise could adversely impact the weighted mean.

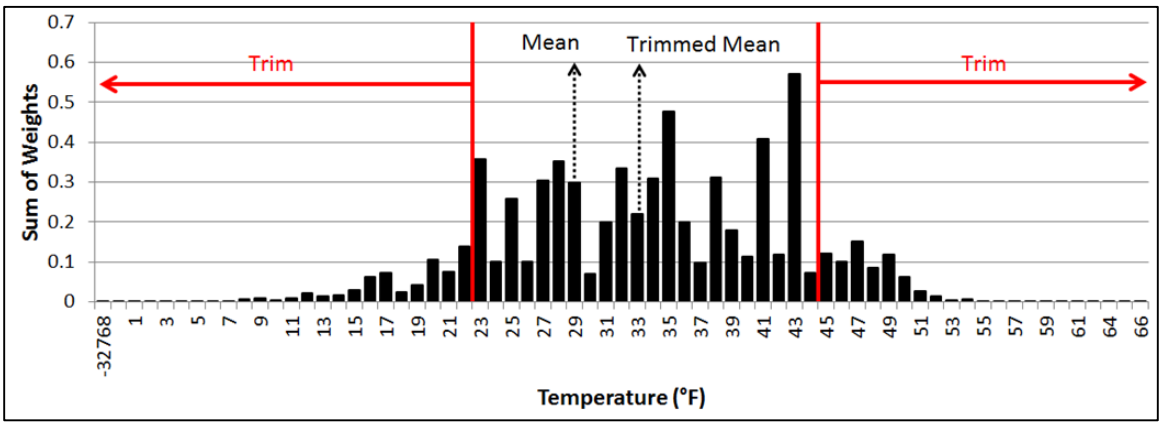

Figure 13. Estimation by Mean and Trimmed Mean

A value of $\varepsilon = 0.1$ was used in the experiments presented in this paper, corresponding to retention of the middle 80% of the weighted values at a given time. This value was chosen as large enough to exclude the most extreme differences in the data while small enough to include a majority of predicted values.

Given an estimate $NE(o)$ or $TNE(o)$ for observation o, the error is computed as the squared difference between estimate and actual. For a set of observations such as all observations for a given station, overall error is computed as the root-mean-squared error, the square root of the mean of the errors over all observations in the set.

It is desirable that the error for "good" readings to be small and the error for "bad" readings be large so that good readings can be distinguished from bad by way of the computed error. If an observation from a station is erroneous, it likely will not agree with observations from stations in its neighborhood, and the error will be large. If an observation from a station is valid, then it likely will agree with observations from stations in its neighborhood, and the error will be small.

Since the MADIS data set includes quality control descriptors, we grouped our results by these descriptors to compare relative to the MADIS assessment of quality for individual observations. The MADIS quality control descriptors were further used to assess the robustness of the neighborhood definition process by comparing the results when neighborhoods are formed using all data versus neighborhoods formed using only data that has been verified by MADIS as having passed all quality control checks. Table 6 summarizes the methods presented for neighborhood formation and subsequent estimation. Results were presented for each of these methods.

Table 6: Neighborhood Formation and Estimation Methods

Method		Neighborhood Membership Distance Measure	Trimmed Neighborhood Estimation?
	G	Great Circle (Barnes Spatial Test)	N
	Gϵ	Great Circle	Y
	D	Direct Data	N
	Dϵ	Direct Data	Y
	DT	Direct Data Trimmed	N
	DTϵ	Direct Data Trimmed	Y
	S	Shift	N
	Sϵ	Shift	Y
	ST	Shift Trimmed	N
	STϵ	Shift Trimmed	Y
	L	Least Squares Regression	N
	Lϵ	Least Squares Regression	Y
	LT	Least Squares Regression Trimmed	N
	LTϵ	Least Squares Regression Trimmed	Y

With neighborhoods formed using all data from 2001-2011, the LTϵ method shows the least RMSE for observations from 2001-2011 with the V, S or Z MADIS QCD descriptors. See Table 7. The STϵ and DTϵ methods showed comparable although slightly higher RMSE for these groups. The G method (Barnes Spatial Test) yielded the worst RMSE for the V, S and Z QCD groups. The DT, ST and LT groups all resulted in relatively high RMSE for the V group. It is unclear why this happened. For the X QCD group, all methods yielded a high RMSE. For the Q QCD group, all methods yielded a moderate RMSE.

Table 7. Results (RMSE °F) for Neighborhoods formed with All Data, 2001-2011

Neighborhood Formation and Estimation Method	MADIS QCD				
	Q	**S**	**V**	**X**	**Z**
G	7.35	**6.30**	**9.84**	568.54	**10.09**
Gϵ	6.10	4.84	4.76	568.70	2.32
D	7.04	4.02	3.84	565.41	4.33
Dϵ	7.06	4.02	3.81	565.50	3.78
DT	8.54	4.96	9.75	565.72	2.65
DTϵ	7.34	3.46	3.21	565.87	2.44
S	7.90	4.80	4.49	565.42	5.39
Sϵ	8.03	4.85	4.52	565.44	5.17
ST	7.71	3.96	7.50	565.34	2.65
STϵ	7.41	3.19	3.00	565.29	2.07
L	7.71	4.70	4.44	565.52	5.80
Lϵ	7.77	4.67	4.42	565.57	5.22
LT	7.69	3.99	7.53	565.37	2.48
LTϵ	7.40	**3.16**	**2.98**	565.35	**2.03**

With neighborhoods formed using all data from 2001-2011, the LTϵ method shows the least RMSE for observations from 2012 with the QCD V descriptor. See Table 8. The STϵ and DTϵ methods showed comparable but higher RMSE for this group. It is unclear why, but the DT method showed the highest RMSE for this group. For the QCD S group, the G method showed the best RMSE while the Sϵ method showed the worst. The DTϵ, LTϵ and STϵ methods, along with the DT method showed results for the QCD S group comparable to the G method. For the QCD X group, all methods yielded a high RMSE. For the QCD Q group, all methods yielded a moderate RMSE with the exception of the DT method, which yielded a high RMSE.

Table 8. Results (RMSE °F) for Neighborhoods formed with All Data, 2012

Neighborhood Formation and Estimation Method	MADIS QCD			
	Q	**S**	**V**	**X**
G	6.50	**3.70**	4.90	569.83
Gϵ	6.36	3.83	4.97	569.91
D	6.92	5.62	4.22	564.65
Dϵ	7.02	5.58	4.15	564.77
DT	51.30	4.06	**34.20**	575.33
DTϵ	6.54	3.86	3.28	565.93
S	9.28	7.61	6.30	561.98
Sϵ	9.48	**7.68**	6.46	562.06
ST	8.96	4.93	5.10	561.13
STϵ	6.88	4.16	3.13	564.92
L	9.24	7.31	6.11	562.64
Lϵ	9.63	7.24	6.33	562.95
LT	8.71	4.75	4.80	561.31
LTϵ	6.88	4.04	**3.09**	564.98

When using MADIS QCD V-only data for neighborhood formation, we observed similar results. See Table 9 and Table 10. (Note that the G and Gϵ methods are shown here for comparison, However, since neighborhood formation for those methods is based solely on geographic location, results are identical to that for inclusion of all data.) The LTϵ method again shows the least RMSE for both the 2001-2011 and the 2012 data sets for the V group. The DTϵ and STϵ methods showed comparable but higher RMSE. And, the Dϵ, Sϵ and Lϵ methods showed comparable results. Using the V-only data for neighborhood formation removes outliers and appears to alleviate the need for trimming when forming neighborhoods. Phrased differently, since the results here are comparable although slightly better than those for the DTϵ, STϵ and LTϵ for neighborhoods formed

using all data, it appears that it is possible to achieve comparable results in the absence of prior QCD indicators by using trimming when forming neighborhoods.

Table 9. Results (RMSE °F) for Neighborhoods formed with V-only Data, 2001-2011

		MADIS QCD				
		Q	**S**	**V**	**X**	**Z**
Neighborhood Formation and Estimation Method	**G**	7.35	**6.30**	9.84	568.54	10.09
	G€	6.10	4.84	4.76	568.70	2.32
	D	8.02	5.29	10.68	565.68	2.56
	D€	7.31	3.31	3.05	565.80	2.36
	DT	8.31	5.51	**11.29**	565.63	2.64
	DT€	7.33	3.41	3.15	565.77	2.41
	S	7.31	4.22	7.88	565.80	2.30
	S€	7.10	3.19	2.98	565.75	1.82
	ST	7.42	4.34	8.22	565.60	2.37
	ST€	7.24	3.15	2.93	565.59	1.86
	L	7.24	4.29	8.11	565.93	1.93
	L€	7.05	3.22	3.02	565.94	1.66
	LT	7.42	4.39	8.45	565.65	2.18
	LT€	7.24	**3.14**	**2.92**	565.64	1.80

Table 10. Results (RMSE °F) for Neighborhoods formed with V-only Data, 2012

		MADIS QCD			
		Q	**S**	**V**	**X**
Neighborhood Formation and Estimation Method	**G**	6.50	3.70	4.90	569.83
	G€	6.36	3.83	4.97	569.91
	D	31.38	3.83	20.32	571.27
	D€	6.46	3.66	3.10	565.96
	DT	45.42	3.89	**29.73**	573.92
	DT€	6.45	3.71	3.17	566.00
	S	8.77	4.31	4.82	562.62
	S€	6.57	3.55	2.92	565.69
	ST	8.17	**4.75**	4.57	561.73
	ST€	6.70	3.55	2.89	565.47
	L	8.42	4.07	4.42	563.02
	L€	6.63	**3.36**	2.90	565.82
	LT	8.00	4.48	4.28	561.80
	LT€	6.72	3.47	**2.87**	565.50

Overall, trimmed estimates appear to be very beneficial versus untrimmed estimates. In combination with neighborhoods formed using trimmed distances, the associated methods consistently produced better results than those that did not apply trimming to both aspects of the problem, particularly the Barnes Spatial Test. Using trimming alone in neighborhood formation and not in estimation sometimes produced poor results. Even in the case where data used to form neighborhoods is preprocessed to remove outliers, there is benefit in using trimming to form neighborhoods and for estimation. Combining both methods provides results that are consistently better than geographic neighborhoods.

Our approach is resistant to outliers, as demonstrated by results with all data considered for neighborhood formation. Our approach is resistant to missing data, as demonstrated by its performance in light of varying station observation frequency and timing. Our approach appears to be resistant to spatiotemporal variation as demonstrated by performance over a test period (2012) that was not included in the training data used to form neighborhoods. Our approach appears to be resistant to bad metadata by not depending on station location metadata. And, it is not limited to one or several members per neighborhood. Members contribute with varying influence by way of weights assigned to members.

In regard to resistance to bad metadata, we can directly apply our distance measures to identify and confirm stations with bad location data. Station GISC1, “Gibson

near Castella", is located by MADIS in downtown Sacramento. From our own experience we know that Gibson is near Castella which is approximately 20 miles south of Dunsmuir. We have received several "error reports" in WeatherShare indicating that observations from this station are not always consistent with other stations in downtown Sacramento. By Trimmed Least Squares Regression (LT) distance, this station is the 8th "closest" station to CTDUN. However, by Great Circle Distance (G), it is not even in the top 250 nearest stations. This discrepancy confirms user observations that the station is mis-located. We believe that we can implement automated processes to identify these mis-located stations and even approximate their correct locations.

In all cases examined, it does appear that trimming is a beneficial and perhaps necessary step for estimation using the neighborhoods. Trimming is also useful for neighborhood formation when all data is considered and is not preprocessed to remove outliers. This approach appears useful as a replacement procedure to remove outliers and does not depend on domain-specific knowledge such as is applied in the rule-based MADIS QC approach.

There certainly are many adjustments that can be made to this approach. For instance, an alternate robust method for estimation using neighborhoods consistent with our approach would be to use a weighted median. Such an approach may be desirable since it should be even less susceptible to large weights at the extremes.

Since our study only looked at neighborhoods for one challenging site, additional study is merited to determine if parameters can be selected to perform well for

neighborhoods for a larger set of stations or if parameters must be optimized for each site. In prior work we presented online algorithms for application to quality control [45]. To make the work presented in this paper truly applicable, it must be adapted for online use in a manner that is computationally feasible and that provides similar if not better quality control performance. We are confident that this is possible and intend to investigate it further. And, we intend to further investigate and implement mechanisms to proactively identify incorrectly located stations and provide guidance regarding their correct locations.

ACM SIGSPATIAL 2014

In our third published paper [48], we investigated the impact of provider quality control on several real-time applications. The relationship between data provider and consumer can be complex. Sensor readings may pass through multiple providers before reaching the provider from whom we acquire data. A single sensor reading might be included in data feeds from numerous providers. Providers may acquire data from other providers at varying times and through varying methods. Providers may apply their own processing to convert data to common units and formats or to perform quality assessment. In turn, they may provide data at varying times and through a wide variety of distribution mechanisms. As a consumer of such data, we may be privy only to information that can be inferred from the direct data feed. Yet we need to recognize the complexity of the overall system, and realize that the path from the sensor to us may be far from direct. See Figure 14. Consumer, Provider Sensor Relationships.

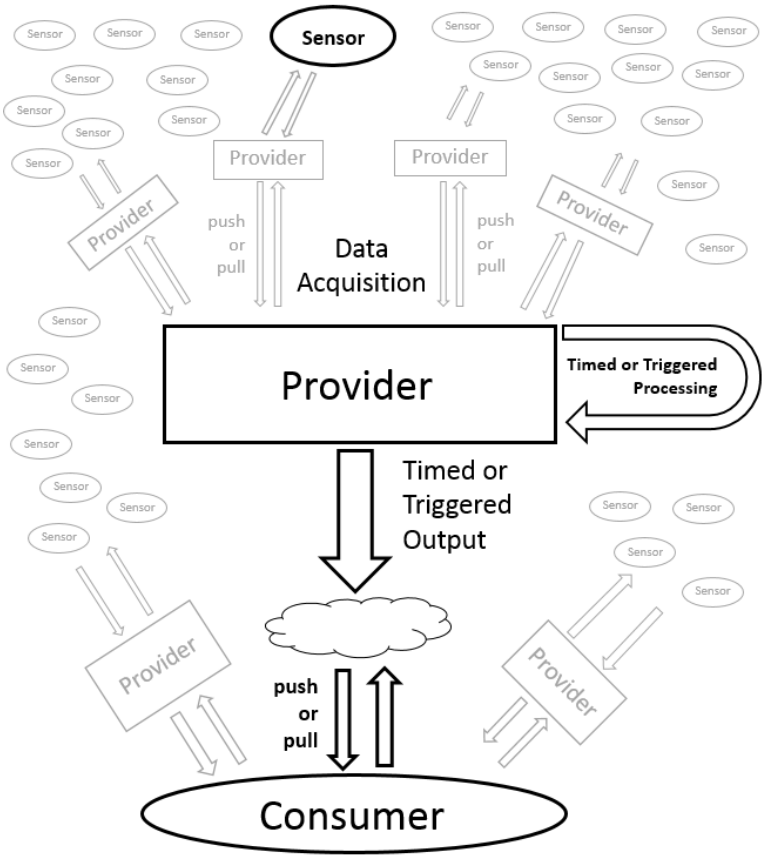

Figure 14. Consumer, Provider Sensor Relationships

We focused our approach on information available to the consumer of sensor data from a provider. While bounding the scope of our interests, we were cognizant of the complex system through which sensor readings are provided.

We first defined two types of observations to distinguish between an (original) observation recorded directly by a sensor in the field and a (provided) observation from a provider. The key distinction is the timestamps, although conversion of units and format may yield further differences.

We intended that our approach be applicable to a variety of general provider distribution mechanisms, whether they be push- or pull-oriented relative to the consumer. This includes single site/sensor streams and aggregate streams, as well as files. As implied by our definition of provider observations, we require that a timestamp be included or readily attainable to indicate the precise time at which the provider makes each observation available. For instance, the timestamp could be the modification time for a published file.

We first presented quality measures relative to an individual site / sensor and extended these measures to form a basis for aggregates over time and space. In this paper, we used provider quality control indicators to assess accuracy. In prior work we presented alternate approaches for assessment of accuracy. [45][46]

First we defined lag. We use a measure similar to timeliness in [29] with the caveat that we are principally interested in lag relative to a data provider. Lag is the

difference between the time when an observation occurs and when it becomes available from the provider.

The second measure we defined was temporal completeness, which indicates how well a time interval is covered by observations. Window completeness is defined in [29] and [30] as the ratio of the number of "originally measured, not-interpolated" values to the containing (time) window size. For example, a station might provide 4 observations per hour. This is not very informative – the result for a burst of 4 successive observations one minute apart within an hour is the same as that for 4 observations spaced 15 minutes apart. Instead, we defined (temporal) completeness using lag as the average lag over a time interval.

This measure is similar to granularity in [29]. Alternative measures such as a sum or maximum and more elaborate measures using decay and autocorrelation are possible. These measures are more informative than a simple rate because they provide indications of the age of observations over time. Our measure is defined in terms of sets of observations and can be applied to sets that are restricted based on provider quality control indicators. For instance, we may restrict our attention to observations that have fully "passed" provider quality control. Doing so can help us assess the impact of provider quality control.

Last, we defined (spatial) coverage. [20] restates a characteristic from Bedard and Valliere where coverage is a measure that "evaluates whether the territory and the period

for which the data exists, the ‘where’ and ‘when’ meet user needs.” This is important because it addresses both spatial and temporal aspects.

We can compute lag and completeness for observations from locations within a cell in a spatial grid and define aggregates that include both spatial and temporal aspects of our data. For instance, we can compute the temporal completeness for individual cells in a grid and then compute the average temporal completeness over all grids.

We tested our measures by direct application to several challenges we face on the various Weathershare projects, using data from MADIS.[2]

MADIS stores files by hour – all observations for a given hour go into the same file. Each file contains only one copy of an individual observation, so there is no duplication within the files. Subsequent file versions contain observations that were included in prior versions as well as new observations, resulting in duplication. MADIS provides multiple levels of quality control checks. [6][49] A single original observation may result in multiple provided observations corresponding to times at which the containing hourly file is updated. The quality control value may change as subsequent quality control checks are applied.

We restricted our attention to a grid consisting of fifty-six 1° Latitude x 1° Longitude cells that overlap with California. This grid includes cells overlapping the Pacific Ocean, Mexico, Nevada and Arizona. A finer grid or non-uniform partitions could also be used. There are sensors located in all of these cells. We use air temperature for this investigation. We further restricted our attention to the time period between 3/5/2014

16:22 GMT and 3/17/2014 17:19 GMT. During this period, we downloaded and stored every MADIS file from the Mesonet subset as the file was updated, and kept separate copies corresponding to each update.

For each cell in the grid, we computed completeness as the average lag (in seconds) of data within the cell over all time units within the period for which we collected data. We compute over the set of all observations within a cell as if they are from a single source corresponding to the cell. The most recent observation from any site within the cell will be counted as the current observation for the cell since we desire to cover the map in a fashion that gives equal attention to each cell, and does not over-represent cells containing many sensors. We assess coverage using summary statistics over all the cells. Data is analyzed for all data versus QC-passed data.

Let Ω represent a set of provider observations ω satisfying a set of restrictions on location and time. Then let Ω_{QC} represent the subset of Ω that has passed all provider quality control checks.

If we use all data as-is, including data that has not passed quality control, 75% of the cells show an average lag of less than 15 minutes (900 seconds). The greatest average lag is nearly 45 minutes (2700 seconds). If we only use data that has passed quality control, 75% of the cells show an average lag no more than 24 minutes (1440 seconds). The greatest average lag is 66 minutes (3960 seconds). In general, there is a 10 minute or greater additional lag for using data that has passed provider quality control versus using all data. This lag is suspected to be due to batch processing of quality control. In the

extreme case (41 minutes), the lag is likely attributable to a higher proportion of bad data in that cell and/or delayed communication.

Recognizing that dependency on provider quality control results in a 10 minute or greater lag penalty, it does seem best to implement quality control mechanisms in our system so long as they can be implemented in a timely manner.

We can look at the results from individual cells to better assess the timely coverage of the map and determine where gaps in coverage exist. For both the Ω and Ω_{QC} datasets there are eight outliers greater than Q3 +1.5 IQR. Seven of these occur in low-population desert areas, with five overlapping the Nevada border near Death Valley, and another two in the Southern-most portion of California, east of Los Angeles and San Diego. One cell corresponds to a low-population coastal area approximately half way between San Francisco and Los Angeles. The latter is also an outlier in terms of the difference between the Ω and Ω_{QC} averages, with a difference of over 41 minutes. This extreme value indicates that the cell does not include sensors that report observations passing quality control in a timely manner, due to bad data and/or slow reporting. Awareness of this deficiency allows us to better focus on things we can control such as our download schedule.

In Figure15, we show lag by minute (average over all cells) for both the Ω and the Ω_{QC} data sets. There are several apparent patterns. For the Ω dataset, the least lag occurs at 8 minutes after the hour. As a result, if we were to make just one download, it would be optimal to do this at 8 minutes after the hour. There are other times with low lag

including 23, 38, and 54 minutes after the hour. And there are further good times including 44, 59 and 4 minutes after the hour and several others, with an apparent 15 minute period. We attribute this pattern to different schedules for data import, batch output and other batch processing. For the Ω_{QC} data set, the pattern is clearer, and does not correspond exactly to that for the Ω data set. 4, 20, 33 and 49 yield local best times. We speculate that there is a batch process that runs approximately every 15 minutes, and an optimal download schedule should take this into account. See Figure 15.

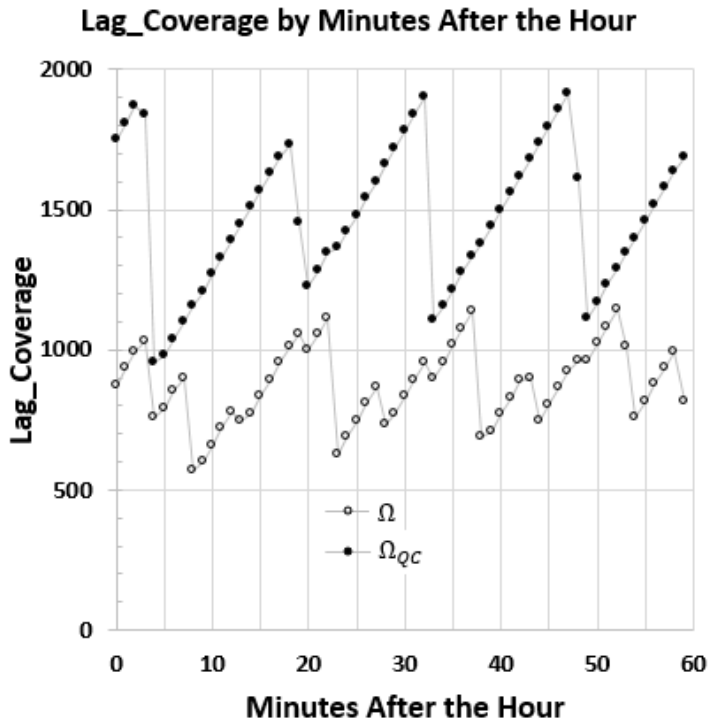

Figure 15. Average Lag_Coverage at each Minute during the Hour

For the **Ω** data set, it is debatable whether more than four download times would improve coverage sufficient to merit the added bandwidth. For the Ω_{QC} data set, the optimal schedule for four downloads yields coverage that is only 45 seconds greater than the best possible, yet it requires less than half the bandwidth. There is little reason to do more than four downloads per hour, since the additional bandwidth required to do so results in little improvement in Lag_Coverage. See Figure 16.

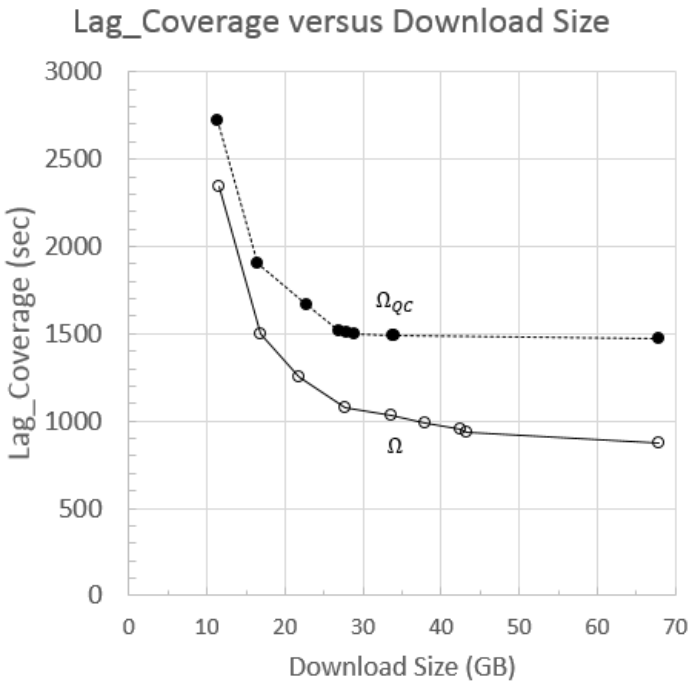

Figure 16. Lag_Coverage versus Download Size for Optimal Download Schedules

By restricting ourselves to only the file for the current hour and the prior hour we can get comparable results. For the Ω dataset, the {8,23,38,54} schedule yields a

Lag_Coverage of 1082.2 seconds, which is less than 2 seconds greater than that for the same schedule when downloading all new files at those times. However, the overall download size will be only 4.9 GB, as compared to 27.8 GB. For the Ω_{QC} dataset, the {4,20,33,49} schedule yields a Lag_Coverage of 1517.2 seconds, which is also less than 2 seconds worse than the same schedule when downloading all new files at those times.The download size is 4.2 GB compared to 26.9 GB. These results are even better when compared against downloading all files at all times, which would consume 67.8 GB.

The simple measures we presented in this paper were demonstrated as useful in helping to solve complex problems related to bandwidth/load-shedding relative to visual coverage of a map with data acquired from a third-party provider. These measures helped to reveal underlying patterns related to acquisition, processing and provision of data by the provider. These measures can be implemented in a simple manner and are applicable to a wide variety of situations for consumers of spatiotemporal data from third-party data providers.

BACKGROUND FOR NEXT PAPER: KRIGING

To further demonstrate the challenges of spatiotemporal data quality assessment, we present in this section a standard application of kriging, the popular spatial interpolation method. Kriging appears to be naturally applicable to the task of data quality assessment. However, as we will demonstrate, it is easily impacted by multiple data quality dimensions and, as such, its applicability is hindered unless data quality issues in the inputs are addressed.

Kriging and Optimal Interpolation, mentioned previously in conjunction with MADIS quality control, 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. [50]

Kriging is appealing for the task of data quality assessment not only because it provides “best linear unbiased predictions” in the form of a weighted average of neighboring sites, but also because it provides error estimates in the form of “kriging variance”. Intuitively, we can hold out an observation and use kriging to provide an estimate at the location of the observation. Then we can compare the observation to the estimate in terms of difference either directly or measured by standard deviations relative

to the kriging variance. If the difference, in absolute value, is “large”, then we might reject the observation as being “bad”.

Next we present a derivation of ordinary kriging in order to show the assumptions and potential pitfalls. The derivation presented here uses the same notation and follows the steps found in [51] and originally in [52] with minor formatting and formulaic changes, as well as added commentary. See [53] for a similar but more concise derivation as well as the derivation in [54] for alternatives.

Let Y be a random field over a space S. Let $[s_1, s_2, \dots, s_n]$ be points in S, and let $y = [y_1, y_2, \dots, y_n]$ be values sampled from Y at locations $[s_1, s_2, \dots, s_n]$. Let s_0 be a location at which we wish to the estimate the value of the field, $Y_0 = Y(s_0)$. Let $w = [w_1, w_2, \dots, w_n]$ be weights and let our estimator take the form:

$$\hat{Y}(s_0) = \hat{Y}_0 = \sum_{i=1}^{n} w_i y_i = [w_1 \quad \dots \quad w_n] \begin{bmatrix} y_i \\ \dots \\ y_n \end{bmatrix} = w'y$$

which is the estimate at s_0.

Assume $E(Y_i) = E\big(Y(s_i)\big) = \mu$, an unknown mean. Further assume that $Var(Y_0) = \sigma^2$. Let $C_{ij} = Cov\big(Y_i, Y_j\big)$ and let $c_i = Cov(Y_i, Y_0)$. We want to find $\hat{Y}_0 = w'y$ such that

1) $E\big(\hat{Y}_0\big) = E(Y_0)$ (un-biased). This is satisfied if $\sum_{i=1}^{n} w_i = 1$ since $E(Y_i) = \mu, \forall i$.
2) The prediction variance $\sigma_\varepsilon^2 = E\left[\big(Y_0 - \hat{Y}_0\big)^2\right] = Var\big(Y_0 - \hat{Y}_0\big)$ is minimized.

Then the mean-squared-error of the estimate is:

$$\sigma_\varepsilon^2 = Var\left(Y_0 - \hat{Y}_0\right) = Var(Y_0) + Var\left(\hat{Y}_0\right) - 2Cov\left(\hat{Y}_0, Y_0\right)$$

$$= \sigma^2 + Var\left(\sum_{i=1}^{n} w_i y_i\right) - 2Cov\left(\sum_{i=1}^{n} w_i y_i\,, Y_0\right)$$

$$= \sigma^2 + \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j Cov\left(Y_i, Y_j\right) - 2\sum_{i=1}^{n} w_i Cov(Y_i, Y_0)$$

$$= \sigma^2 + \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j C_{ij} - 2\sum_{i=1}^{n} w_i c_i$$

We now minimize $Var\left(Y_0 - \hat{Y}_0\right)$ such that $\sum_{i=1}^{n} w_i = 1$ using a Lagrange multiplier.

1) Form the Lagrangian: $L = Var\left(Y_0 - \hat{Y}_0\right) + 2\lambda\left(\sum_{i=1}^{n} w_i - 1\right)$

$$= \sigma^2 + \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j C_{ij} - 2\sum_{i=1}^{n} w_i c_i + 2\lambda\left(\sum_{i=1}^{n} w_i - 1\right)$$

2) Take the partial derivatives of L with respect to the w_i's and λ, set them to 0, and solve.

$$\begin{pmatrix} 2\sum_{j=1}^{n} w_j C_{1j} - 2c_1 + 2\lambda = 0 \\ \vdots \\ 2\sum_{j=1}^{n} w_j C_{nj} - 2c_n + 2\lambda = 0 \\ 2\lambda\left(\sum_{i=1}^{n} w_i - 1\right) = 0 \end{pmatrix} \leftrightarrow \begin{pmatrix} \sum_{j=1}^{n} w_j C_{1j} + \lambda = c_1 \\ \vdots \\ \sum_{j=1}^{n} w_j C_{nj} + \lambda = c_n \\ \sum_{j=1}^{n} w_j + 0 = 1 \end{pmatrix}$$

$$\begin{bmatrix} C_{11} & \cdots & C_{1n} & 1 \\ \vdots & \ddots & \vdots & \vdots \\ C_{n1} & \cdots & C_{nn} & 1 \\ 1 & \cdots & 1 & 0 \end{bmatrix} \begin{bmatrix} w_1 \\ \vdots \\ w_n \\ \lambda \end{bmatrix} = \begin{bmatrix} c_1 \\ \vdots \\ c_n \\ 1 \end{bmatrix} \leftrightarrow \begin{bmatrix} C & 1 \\ 1' & 0 \end{bmatrix} \begin{bmatrix} w \\ \lambda \end{bmatrix} = \begin{bmatrix} c \\ 1 \end{bmatrix}$$

Then $w = \begin{bmatrix} w \\ \lambda \end{bmatrix} = \begin{bmatrix} C & 1 \\ 1' & 0 \end{bmatrix}^{-1} \begin{bmatrix} c \\ 1 \end{bmatrix}$ yields the weights and the sum of squared error is $\sigma_\varepsilon^2 = Var(Y_0 - \hat{Y}_0) = \sigma^2 + w' \begin{bmatrix} c \\ 1 \end{bmatrix}$

Note several things about the derivation of ordinary kriging:

- Distance between points is not used directly. It may be implicit in the covariance, but not necessarily.
- Time is not accounted for, nor is it discounted. Typical usage assumes observations were made at the same time. It is possible that the covariance could account for differences in times of observations.
- Covariance is not characterized other than that it exists between the relevant points.
- The only assumption on variance is that it is known at the point for which the prediction will be made.
- The strongest assumption is that the mean of the random field across the space is constant. This is a strong assumption and may require a transformation of the data.

- There is no assumption of specific underlying distributions. Particularly, the derivation does not assume underlying Gaussian processes – i.e., that the field is a Gaussian Random Field. If the variables are Gaussian, then additional results follow. For instance, if a Gaussian Random Field is assumed, then the Best Linear Predictor will also be the Best Predictor (linear or non-linear). Conversely, the Best Linear Predictor can perform poorly in non-Gaussian situations. See [55].

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 as the sole parameter, and is determined by the data set as a whole. For the sake of example, we follow this same approach here, modeling covariance using data from the present time window. To accomplish this we implemented a fitter/solver for the estimation of covariance functions using the following popular forms. See [56] for similar representations of these and additional functions for semi-variograms rather than covariance.

Exponential: $C(h) = ce^{-kh}$

Gaussian: $C(h) = ce^{-kh^2}$

Spherical: $C(h) = c(1 - 1.5(kh) + 0.5(kh)^3)$

We used the GSL (http://www.gnu.org/software/gsl/) non-linear optimization code to fit data to these functions and determined that an exponential distribution provided a reasonable fit to observed covariance.

We implemented ordinary kriging on the MADIS data set from 2014 and ran estimates for the CTDUN site to compare against actual observations. We used a 90 minute time window and 60 mile distance window to restrict our attention to "close" observations. We used 10 mile bins to compute estimates of covariance by distance fit to an exponential distribution. We also restricted our use of data from stations other than CTDUN to observations having QCD=V from MADIS, indicating that these observations have passed all of the MADIS quality control tests and, as such, may be considered "good". Note that there was an outage at CTDUN which spanned the middle of year.

The following plots show estimates (red) versus actual (black). The estimates follow the general trend of the CTDUN observations. However, in some cases they are off by as much as 10 degrees. The estimates are not very smooth relative to the actual observations. See Figure 17 and Figure 18.

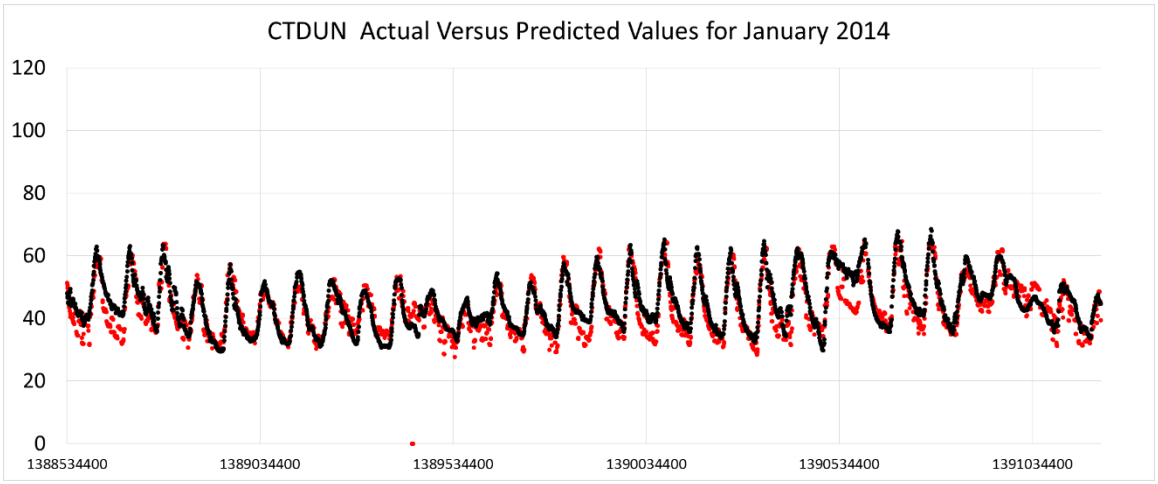

Figure 17. Actual Values versus those Predicted by Kriging for January 2014

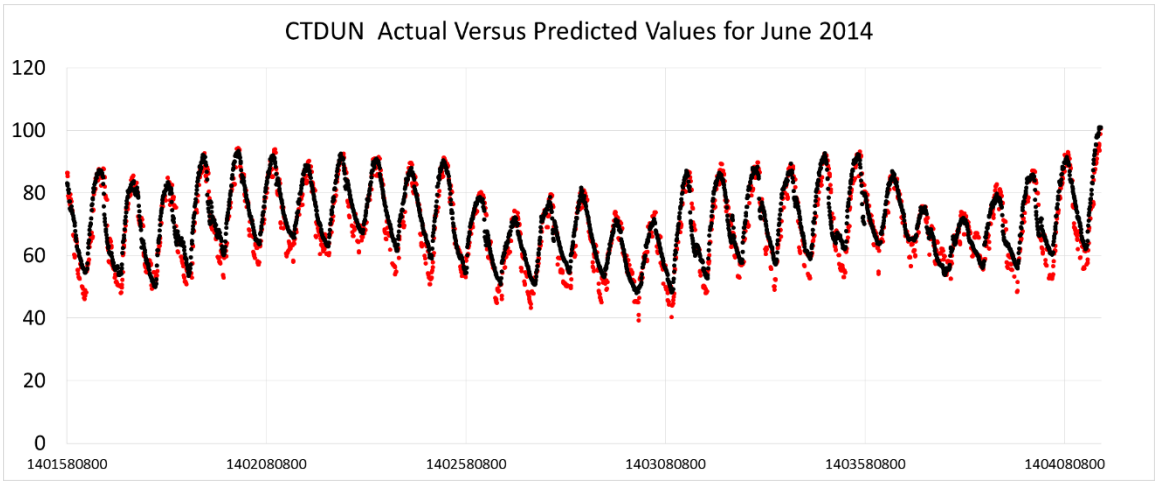

Figure 18. Actual Values versus those Predicted by Kriging for June 2014

The following one-day graph shows how the estimates jump at various points in time. We will subsequently demonstrate why this occurs. See Figure 19.

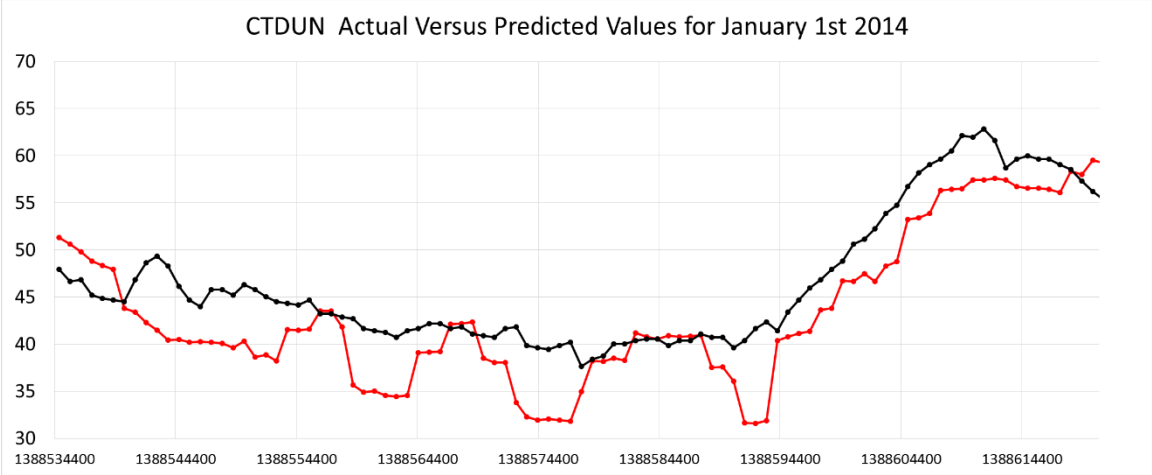

Figure 19. Actual Values versus those Predicted by Kriging for January 1st, 2014

Using the computed kriging variance, we compute z-values at each data point to show a comparison of the actual observation to the predicted value. These values vary dramatically and nearly reach +/- 4 standard deviations for observations that do not otherwise appear to be suspect.

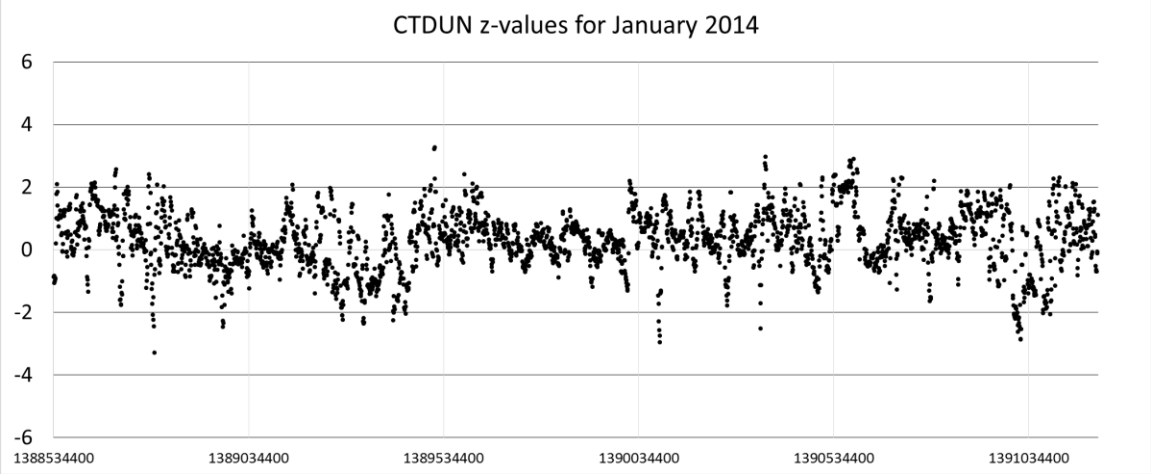

Figure 20. z-values Computed with Kriging Variance for January 2014

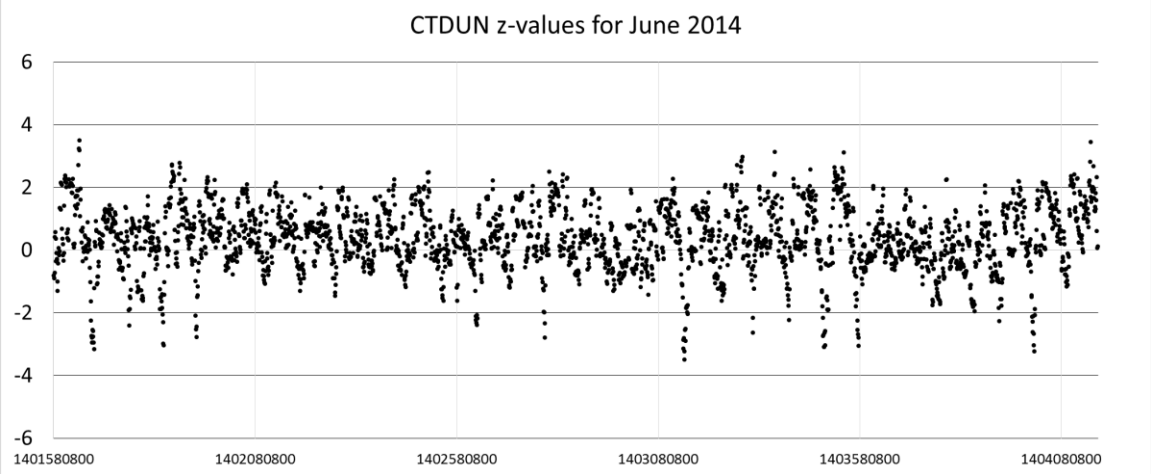

Figure 21. z-values Computed with Kriging Variance for June 2014

Looking at the weights associated with individual stations is helpful in understanding the behavior of ordinary kriging in our experiment. The first two of the following plots show station locations relative to CTDUN. The first plot shows all stations within 60 miles that were used in the experiment and the cumulative weights assigned to them throughout the experiment including negative weights. Those shown without a fill color are cumulative negative. See Figure 22.

As might be expected, the sites nearest to CTDUN have the greatest weight. And, there is a shadowing effect in which near sites shadow further sites. Notice that one site to the northwest shadows another site that is adjacent but further from CTDUN. Note also that one site to the north-northeast accumulates the greatest weight since it is closest to CTDUN.

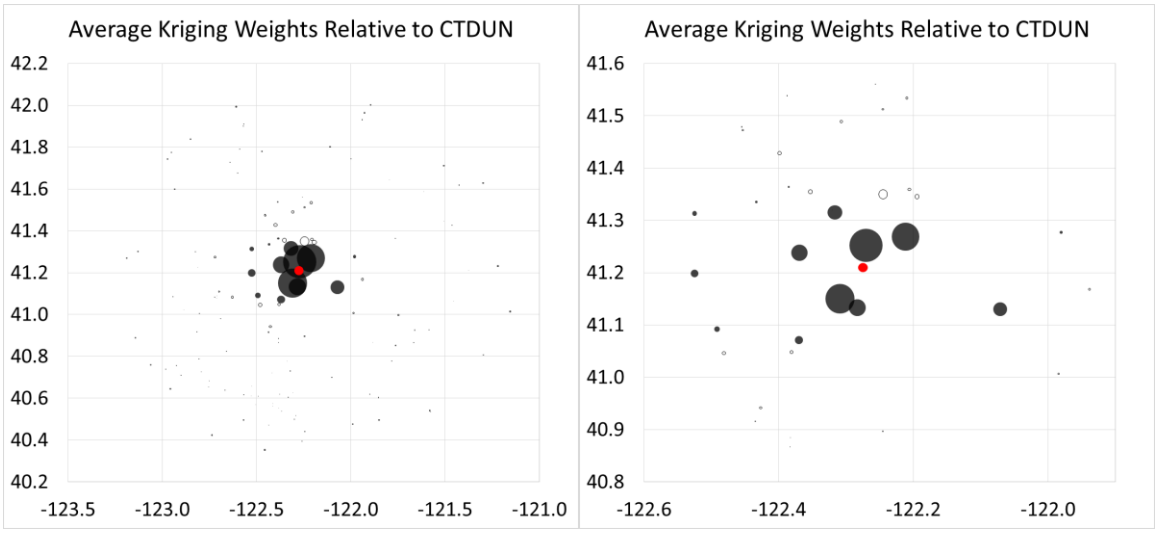

Figure 22. Average Kriging Weights of Neighboring Sites to CTDUN over 2014. The second chart is a zoomed in view of the first.

Here we look at weights assigned at two different times in the experiment to show how the weights can change. At time 1388558160 (epoch time), the station nearest to CTDUN gets a much larger weight than other sites. At time 1388559060, this nearest station is not present because its data falls outside our 90 minute time window. Other sites, particularly one to the northeast are assigned greater weights. (Recall that the weights are constrained to sum to one at each iteration.) See Figure 23.

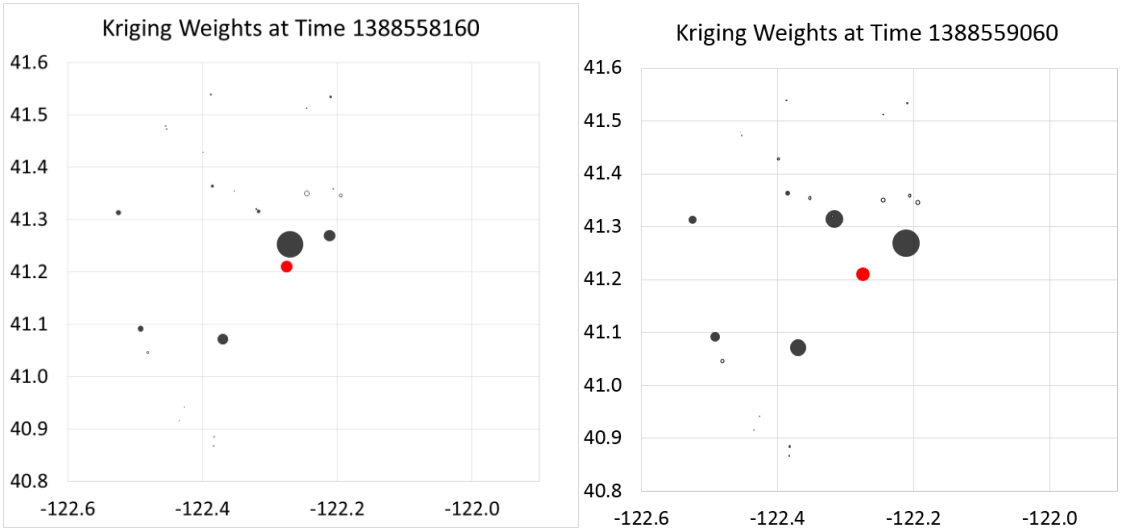

Figure 23. Kriging Weights of Neighboring Sites for Two Time Instances

Station UP636 is closest to CTDUN. As shown in the following plot, the weights assigned to this station are generally 0.5 or higher, although there are some cases in which they are less than 0.3. This station is typically given half of the overall weighting or more. Note that this station reports infrequently, only about half of the time CTDUN reports. There are numerous times when its data will fall outside our selected time window. See Figure 24.

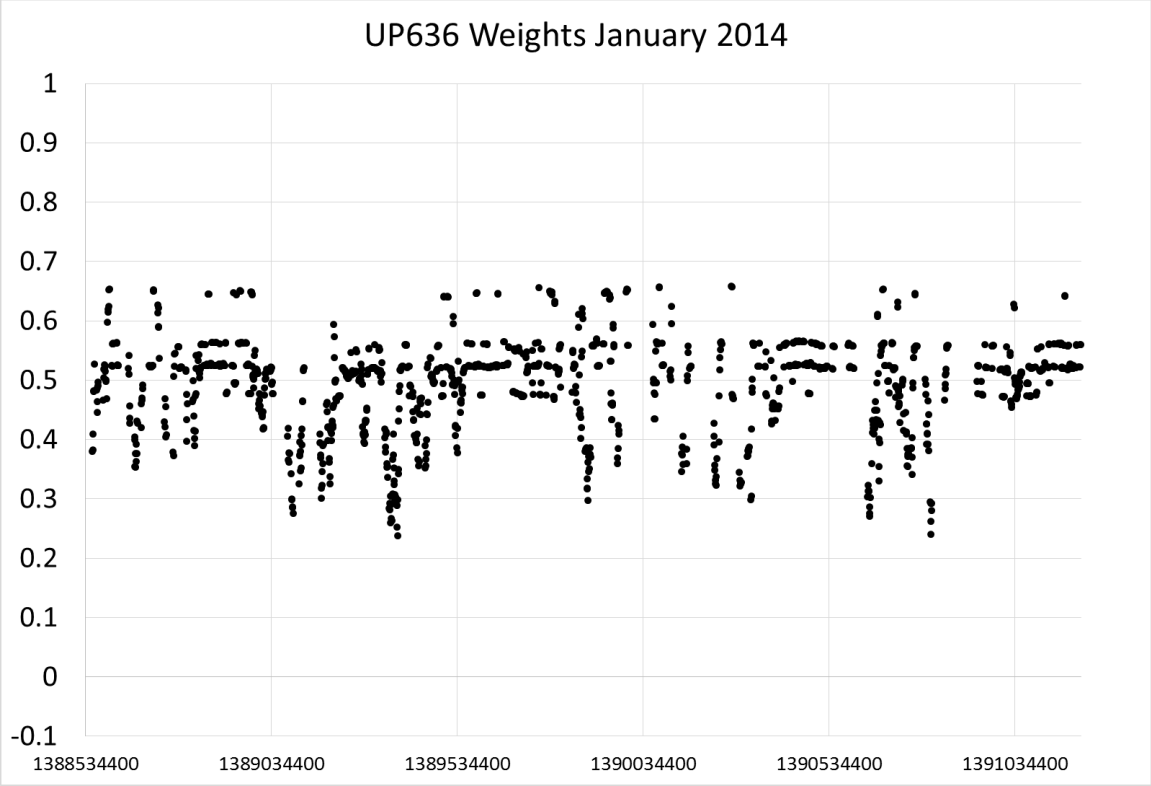

Figure 24. Kriging Weights for Site UP636 in January 2014

SDFC1 is located to the north-northeast of CTDUN and is typically shadowed by UP636. Not only is the weight for this station typically near zero, but quite often it is negative. See Figure 25.

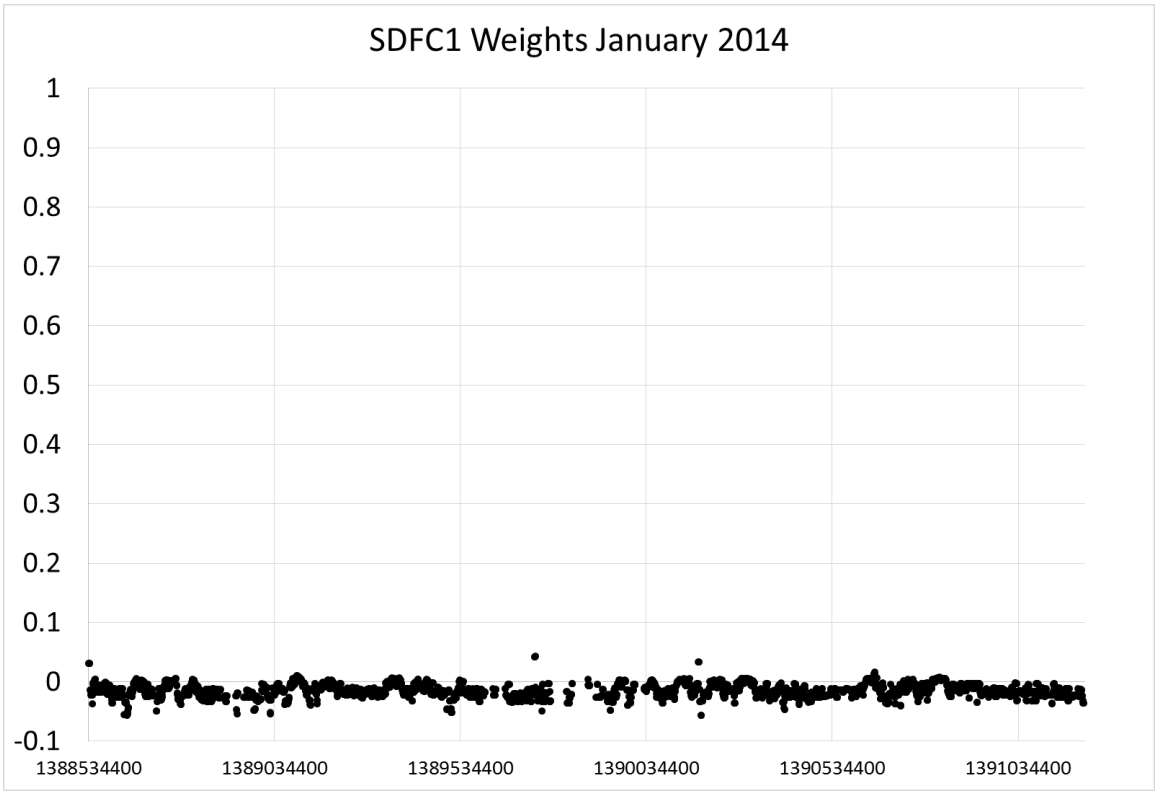

Figure 25. Kriging Weights for Site SDFC1 in January 2014

CTSNS, another Caltrans site, is located to the east northeast of CTDUN. Notice how dramatically the weight assigned to this site changes over time. It ranges from nearly 0.5 to below 0.1. See Figure 26. The presence or absence of data from UP636 is the chief cause of this. When UP636 is present, CTSNS gets a smaller weight. When UP636 is absent, CTSNS gets a greater weight. Other nearby stations also influence these weights.

Between UP636 and CTSNS, site SDFC1 is shadowed in every instance and never receives a large weighting.

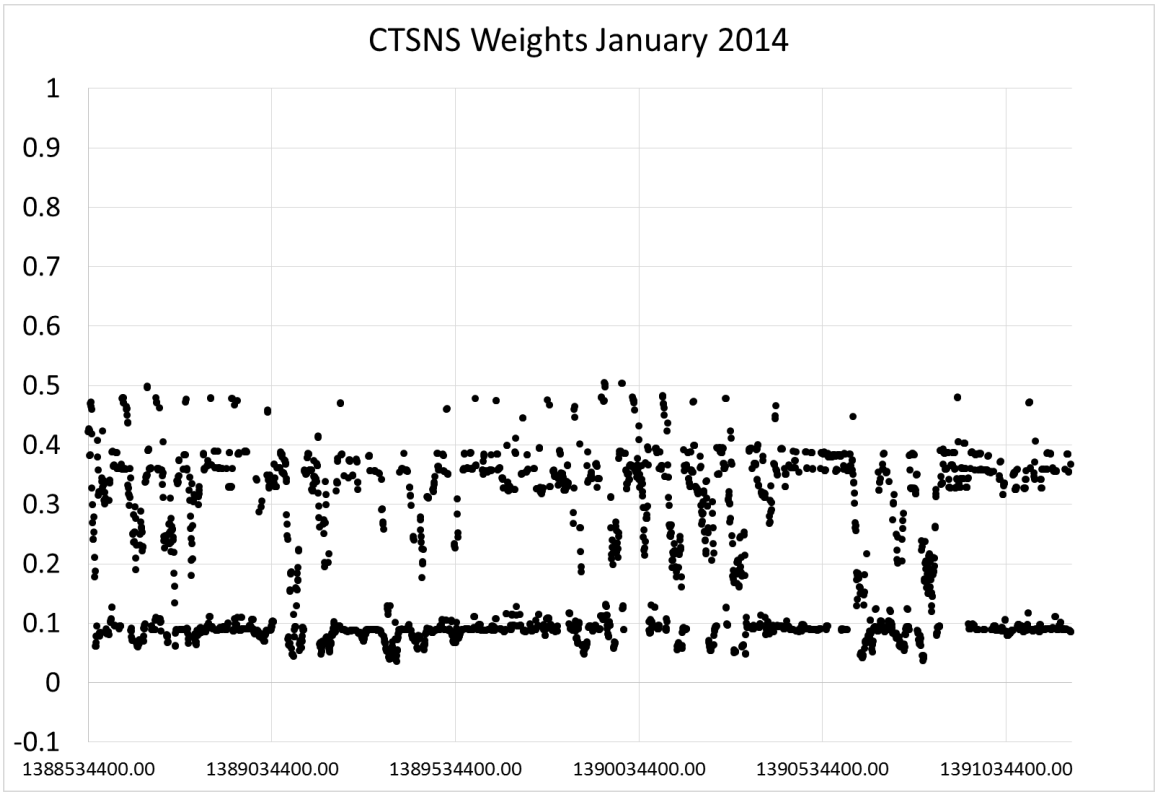

Figure 26. Kriging Weights for Site CTSNS in January 2014

Overall weight statistics for sites within 10 miles of CTDUN are shown in Table 11. Again we note the variation and interplay of the weights. Recall that for the standard approach to using Kriging, a covariance model is created to be a function of distance alone. Thus, distance and the modeled covariance function are the chief contributing factors in determining these weights at any point in time. As we see above, even for these nearby sites, the nearer sites dominate the further sites and this coupled with the variation in the modeled covariance function can cause a dramatic variation in weights. And, the nearest site heavily influences all others

Table 11. Statistics for Kriging Weights of Neighboring Sites to CTDUN over 2014

Station	Count	Avg Weight	Min Weight	Max Weight	Latitude	Longitude	Distance	Bearing
UP636	12238	0.511367	0.144055	0.818061	41.25223	-122.27020	2.93	4.58
UP595	14572	0.337887	0.062921	0.603455	41.15014	-122.30880	4.50	203.22
CTSNS	21761	0.206364	0.022080	0.602063	41.26900	-122.21140	5.24	38.87
MSCAS	7479	0.209584	0.032939	0.445401	41.23806	-122.36920	5.28	291.57
GRDC1	9354	0.183027	0.025298	0.515004	41.13306	-122.28310	5.33	184.70
MSAC1	20890	0.058754	-0.044714	0.278170	41.31533	-122.31660	7.60	343.37
KMHS	21525	0.000088	-0.043755	0.685917	41.32000	-122.32000	7.96	342.82
SDFC1	19876	-0.023037	-0.073381	0.064931	41.35000	-122.24500	9.80	9.05

We now show a table of weights assigned to UP636 over a relatively short period of time. Notice here that two separate UP636 observations are used, and each is used 6 times. As such, each observation gets older and older relative to the CTDUN observation it is being compared to. However, in the first case, the weights increase with age, giving

more weight to older and presumably less-well-correlated observations. Age of observations is not taken into account other than perhaps indirectly by way of the covariance model computation, so the weights are an artifact of the presence or absence of other observations and the modeled covariance.

Table 12. Kriging Weights for UP636 over two Time Intervals

CTDUN Time	UP636 Time	Offset	Weight
1388553660	1388553600	-60	0.380073
1388554560	1388553600	-960	0.382506
1388555460	1388553600	-1860	0.408989
1388556360	1388553600	-2760	0.481481
1388557260	1388553600	-3660	0.482639
1388558160	1388553600	-4560	0.527640
1388564460	1388564400	-60	0.483469
1388565360	1388564400	-960	0.446177
1388566260	1388564400	-1860	0.463602
1388567160	1388564400	-2760	0.486636
1388568060	1388564400	-3660	0.490731
1388568960	1388564400	-4560	0.497237

A review of the literature related to kriging yields some insight into the challenges we have identified and their potential mitigation. [57] presents a general method for detection of outliers based on jack-knifing [58], which is similar to cross-validation. Similar to the approach implemented by MADIS using Optimal Interpolation, this method involves individually removing each data point, producing an estimate via a model at the location of the removed point, and computing the difference between estimated and observed values. In turn, large differences may signify an outlier. The authors then indicate that point kriging is applicable to this situation. It was noted that the

determination of covariance is subject to outliers, and that robust methods may be applied to help mitigate this problem. For instance, trimmed means may be used where the highest and lowest 10% of observed values are removed prior to computation of the mean. Challenges in selecting the general covariance model and subsequent parameters are also discussed. The authors reference techniques from [59] for recommendations in the event that jack-knifing does not work. These include: transformation of the raw data to yield nearer-to-normal distributions, fitting a drift in the data using a method such as universal kriging, or de-clustering the data by treating observations falling near each other as collective observations. (Simple kriging assumes a known, constant mean. Ordinary kriging assumes an unknown, constant mean. Universal kriging assumes the mean follows an underlying polynomial trend.) They proceed to indicate that this method was not practical in their study due to the large amount of data, and that the use of robust methods to compute the covariance may suffice. Note that the authors present an alternate method based on IRF-k [60], with the intent of having a fully automated method. They indicate, however, that point kriging provides the advantage of "deeper insight" due to the necessity of having an analyst determine the covariance model. They recommend the use of point kriging when there is no significant drift and the use of IRF-k when there is significant drift. In general, they indicate that both methods perform comparably.

The authors of [61] also indicate that selection of a covariance model is difficult to automate, and chose not to present kriging results in their study because they were similar to those from distance-weighting approaches. They did, however, indicate that

distance-weighting approaches did not account for variations in elevation, which is correlated positively with precipitation and inversely with temperature. They also mention that de-trended inverse distance weighting methods produced good results for their models of snow water equivalent.

In general, kriging carries a large computational cost due to the necessity of performing matrix inversion for every point to be estimated. Computational costs result from the number of elements in the matrix. If all data observations are used, then this matrix may be large. Kriging using local neighborhoods [62][63] is an alternative that uses only the nearest neighbors for individual predictions. In [64] the authors point out that having a local neighborhood for each location carries its own computation (and storage) burden, which would make computation on a fine grid computationally challenging too, and indicate the discontinuity of prediction and error surfaces where different neighborhoods are used is also a problem.

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. [65] 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. [66]

In [67], the authors investigated data whether they should estimate covariance for each day versus use a single covariance model for all days for modeling daily temperature and precipitation. They found that the best interpolation results came from

the single covariance estimation for all days, attributing its success to having greater statistical certainty by virtue of using a greater amount of data. They also indicate that kriging variance is not a true estimate of uncertainty, and that it is better to perform an ensemble of stochastic simulations. They note that doing so is computationally intensive.

In [68], the authors indicate that a key feature of kriging is that intra-site distances are accounted for in determining weights. Sites that are close to each other will be down-weighted relatively because of assumed correlation and sites far apart from each other will be given greater weight relatively because of greater statistical independence. They indicate that kriging is of even more benefit when attempting to merge separate but spatially overlapping data sets such as rain gauge readings and radar measurements. For this application, they chose to generate a covariance model at each time step since they treated accumulation periods as non-overlapping.

In [69], the authors point out computational complexity as the main disadvantage of kriging, and present an approach of dividing large data sets into sub-models which they compute separately and merge subsequently. They indicate that the advantages of kriging outweigh the computational challenges. For the purposes of modeling covariance, rather than treat space-time as an $\mathbb{R}^3$ space, they separate space and time as $\mathbb{R}^2 \times T$. In turn, they compute spatial covariance separate from temporal covariance. Ultimately, "correlation decay" is calculated based on the individual spatial and temporal covariance functions and spatial and temporal distance. Additionally they note that temporally old

and/or spatially isolated observations, while providing benefit in the absence of other data, contribute greater uncertainty to the model.

In [70], the authors compared kriging to least squares with data samples of size less than 50 and found that kriging performed no better than least squares in predicting an overall mean or a pointwise fit. However, kriging estimates were less variable. In other words, when least squares performed poorly, sometimes it performed a lot worse than kriging.

In [71], the authors found that for the development of digital elevation models, kriging performed better than other interpolation methods such as inverse distance weighting when sampling density was low and spatial structure of elevation was high. When there was little spatial structure due to elevation, inverse distance weight performed better. When sampling density was high, all methods performed similarly.

Interestingly, in [72], the authors found different results for interpolation methods to estimate air quality when monitor density was high. They found similar performance for the interpolation methods when monitor density was low. They investigated spatial averaging, nearest neighbor, inverse distance weighting and kriging. Note that corrections were made in their study for incorrect or unexpected location metadata. Several stations were recognized as being located by latitude / longitude in counties different than their ids indicated, and were relocated to the correct county by adjusting either the latitude or the longitude, but not both, assuming one was incorrect. The exact locations could not be fully determined. Several other monitors were collocated, and averaging was used to

consider them as a single site/monitor. For kriging, covariance models were determined for separate, regions covering the United States. The different interpolation methods were compared by pairwise scatterplots against each other. Overall, they indicated that kriging produced the most realistic estimates.

The authors of [73] found that kriging and Laplacian smoothing splines generally performed better than a number of other methods including inverse distance weighting for prediction of soil pH.

The authors of [74] points out the strengths of kriging including determination of weights using covariance models as well as the configuration of the data. They show examples using soil salinity data as well as vegetation cover.

The authors of [75] found a lack of spatial autocorrelation in precipitation data during the dry season, which caused kriging results equivalent to simple averaging of data. They did find kriging useful for rainy season data, and also used kriging variances to help identify areas with good spatiotemporal data coverage. Data quality issues played a significant role in their study, requiring a large amount of manual preprocessing effort.

In modeling air temperature and precipitation for the Czech Republic with data spanning a nearly 50 year period, the authors of [76] devoted considerable effort to quality control, particularly methods that could be automated. Comparison of time series of a candidate site and its neighbors via pairwise comparison was used as their first level check, comparison of interquartile ranges to series pairs or differences between series was used as their second level check, and modeling and comparing predicted versus

actual values using interpolation methods such as inverse distance weighting and kriging constituted their third level check. They indicated that neighboring stations could be selected using either distance or correlation. Correlations could be computed either on original series or on first differences. They selected up to eight neighboring stations with high correlation coefficients and a distance limit of 300 km and an elevation difference of 500 m. They observed that a higher number of outliers for air temperature were found in summer than in winter, and more outliers were found in morning and evening than at noon. They also applied a process of homogenization where they standardized neighboring station values using the average and standard deviation of the candidate station.

In the context of visualizing data quality, the authors of [77] demonstrate that data quality at a given location changes over time. While data quality could certainly vary over time, we believe the variation observed is likely due more to uncertainty and variation in neighboring data.

While working with human-reported rainfall reports from Australia, the authors of [78] note the short-comings of kriging and other interpolation methods in the presence of low station density and compromised quality of observations at neighboring stations.

From our literature review and our experiment with ordinary kriging, we make the following observations:

- We want an unbiased method that results in usable z-values with mean zero and low variance.

- We recognize the absolute error or mean-squared error may be more practical than a z-value.
- We want a method that produces “smooth” results in terms of estimated values and also z-values.
- We want a method that is robust, and produces a good result even in the presence of poor data quality.
- In an experiment that we did not present here, we filtered stations using some initial measures of quality. This resulted in worsened results because it introduced a bias, but made improvements in terms of the spread of the results. As such, there is promise in filtering out poor quality data and sites, yet we still must account for the weaknesses of using a standard approach to ordinary kriging.
- The approach we implemented did not incorporate temporal information beyond using data within an arbitrary time window.

Some things that we have idenfied for subsequent research on kriging include:

- We should compute and use real station to station co-variances / correlation based on time lag. This will be key to making kriging viable here. We could also try estimating a covariance function based on both distance and time but the approach of using station to station co-variances should be better.
- We may need to transform the data.

    - The assumption of a constant mean may have been violated, so normalization would help.
    - We could also attempt to remove the periodic elements.
- We intend to develop and test for real-time implementation.
- We should investigate using non-QC-checked data as input.
- We may need to investigate robust techniques for the necessary estimates above.

Our published work in relation to inverse-distance weight approaches demonstrates both the need and viability of the items above.

REMAINING CONTRIBUTIONS

While research has been done to compare kriging, inverse distance weighting, multivariate linear regression, and other interpolation techniques, we are unaware of any published, comprehensive development and comparison of these techniques applied to the determination of data quality in light of multidimensional challenges presented by spatiotemporal data. The weather data providers cited earlier, including our own efforts, provide perhaps the most comprehensive presentation of these techniques as applied to assessment of data accuracy, and represent the state of the art. However, each provider has focused on their preferred technique and has generally only addressed the accuracy of observations. We will continue with our development of a multidimensional data quality approach to account for data quality measures including accuracy, timeliness and reliability of original observations, as well as accounting the accuracy of metadata including location and timestamp. We intend also to continue the development of these measures in light of complex provider networks and distribution mechanisms. Our focus is on neighborhood based, “buddy” checks.

FRAMEWORK & TIMELINE

Our over-arching goal is to define multidimensional data quality measures and use these measures to determine the impact of data quality on state of the art algorithms for assessing data quality. In turn, we enhance these algorithms to operate better in light of poor and uncertain data quality.

Our specific objectives include the following:

- Further develop data quality measures for assessing spatiotemporal data quality.
- Further analyze kriging and develop a kriging-based technique that is robust in light of data-quality issues, similar to what we have already done with inverse-distance weighting based techniques.
- Analyze multivariate linear regression based techniques and develop a related technique that is robust in light of data quality issues, similar to what we have already done with inverse-distance weighting based techniques. We already have done preliminary investigation on this.
- Evaluate the three state of the art techniques (inverse distance weighting, kriging, and multivariate linear regression) in real-time scenarios. This includes using data that has not been quality checked by providers.

    - Included will be a practical analysis of the data and computational challenges in light of the reality of a large amount of streaming data. For instance, is historical data needed? How much computational effort and network bandwidth is required? Etc.
    - Included will be an analysis of not only how to assess accuracy of the primary data, but also that of the metadata such as locations and timestamps.
- Present the results of our analysis in the form of a general framework that can be used to assess spatiotemporal data quality. See Figure 27.

Our original intent was to complete this work and our dissertation by the end of Spring Semester 2016. It is more likely that we will finish in Fall Semester 2016 or Spring Semester 2017.

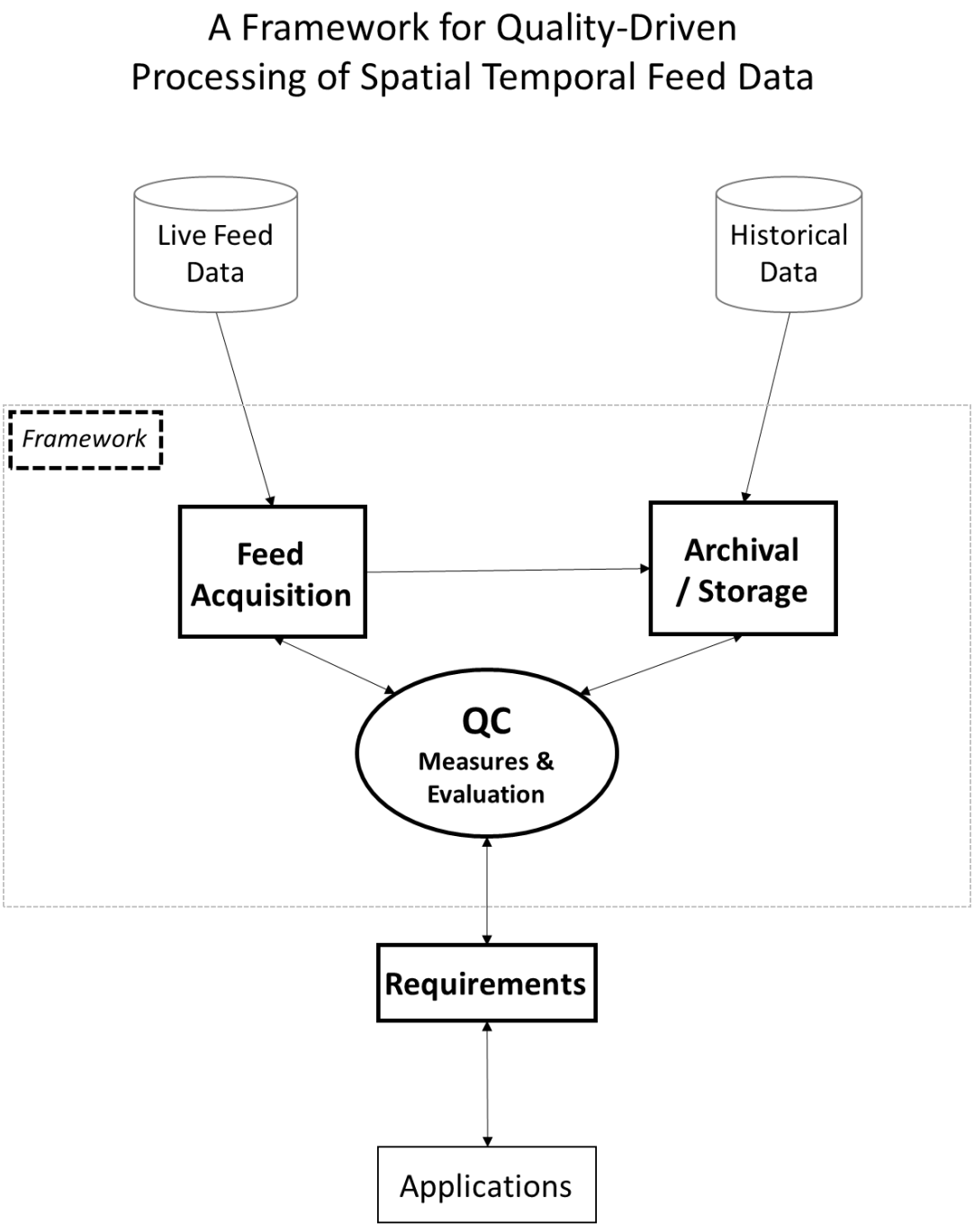

Figure 27. A Framework for Quality-Driven Processing of Spatiotemporal Data

ADDENDUM

My committee requested that I add further documentation to address several issues raised by the committee on my prospectus. This section documents these issues and proposed resolution.

We Need an Artificial Data Set for Verification (generate labeled data for supervised verification). Include sensor drift. Use as evaluation or use for model building.

We have identified a real data set and a method for generating artificial data for this purpose:

- RSAS/MSAS Grids from the National Weather Service is a real data set similar to the data set we have been analyzing.
    - The rasters provide fields representing extrapolated weather conditions one hour apart. While a product of actual weather observations, these rasters can be treated as / ground-truth. In reality, they are not error-free. However, they will suffice for our purposes.
- Fractional Brownian Surfaces can be generated and used to represent terrain and weather patterns.
    - Fractional Brownian surfaces as functions of 2D space (x,y) can to represent terrain.
    - Fractional Brownian surfaces as functions of 3D space (x,y,t) to represent changing spatial-temporal phenomenon (weather).
    - Flow patterns can be represented by moving through the x-y plane in the weather surface.
    - Periodic (diurnal and seasonal) changes can be introduced by adding periodic functions to the surfaces.
- By using these fields as ground-truth, we can evaluate and develop algorithms in a controlled environment.
    - We'll select points within the grids to represent sites.
    - We'll select neighborhoods in a similar fashion and using a systematic approach: near versus far away, balanced on all sides versus unbalanced.
    - We'll introduce errors into the data.
        - Bad observation values.
        - Drifting observation values.
        - Bad metadata: location and time.
        - Errors at a single site versus errors at multiple sites versus systemic errors.
    - We'll introduce other aspects found in real sensor networks:
        - Varying reporting time and frequency.
        - Network outages.
    - We'll evaluate / develop algorithms relative to neighborhoods:
        - How well do they work for interpolation given various configurations of neighbors?

- What is the impact of errors on performance? Are the methods robust in light of errors? What can be done to make them more robust?
- What impact do various quality measures have on performance? Accuracy? Timeliness? Reliable?
- How can the methods be used to identify errors?

Markov Random Fields and Linear Gaussian Models were mentioned as methods to generate random fields for the purposes described above. The real data set and method for generating artificial data sets above should suffice for our purposes, and should be representative of the type of data we are working with.

We Need to Demonstrate Computer Science Relevance

In order to accomplish this, we intend to develop our own new algorithm that can be used to assess quality of spatial temporal data in light of the quality issues in the data. To do so we intend to demonstrate how the popular methods are affected by quality issues (accuracy, timeliness, reliability, …) and how and when we can overcome these issues with our new algorithm. The data sets and procedures described above will be used to accomplish this. The committee expressed interest in the potential to accomplish such a results and, in turn, it to the problem of identifying problems in the current sensor feeds we have been examining. We will do this.

It was requested that the following be included/addressed:

- The Time Element: This will be included as described above and in the associated algorithms.
- Demonstrable Impact by way of assessing the quality and diagnosing problems with real sensor data. This will be done.
- Computing Intensive: By virtue of the data we are dealing with and its spatial temporal nature, it will be computing intensive.

Algorithm Development and Evaluation (Addressing Computer Science Relevance)

See above for specifics development and evaluation in a controlled environment.

We intend to conduct a comparative analysis of the following relative to various types of errors and neighborhood configurations:

- The standard algorithms:
  - Inverse Distance Weighting
  - Kriging
  - Least Squares Regression
- The standard algorithms enhanced to resist errors. For instance, Kriging with a cross-validation hold-out scheme.
- The standard algorithms used to identify errors.

- A new algorithm that we intend to develop that will incorporate the strengths of the standard algorithms relative to errors.

We intend to evaluate based on the following criteria:

- Interpolation accuracy in light of data errors and neighborhood configurations.
- Computational effort (time) required.
- Data required.
- Ability to identify erroneous observations, erroneous metadata and problem sites.

REFERENCES CITED

[1] "The WeatherShare System." [Online]. Available: http://www.weathershare.org/.

[2] "Meteorological Assimilation Data Ingest System (MADIS)." [Online]. Available: http://madis.noaa.gov/.

[3] U. of Utah, "MesoWest Data." [Online]. Available: http://mesowest.utah.edu/.

[4] "MADIS Meteorological Surface Integrated Mesonet Data Providers." [Online]. Available: https://madis.ncep.noaa.gov/mesonet_providers.html.

[5] "NOAA's National Weather Service Field Systems Operations Center Test and Evaluation Branch : Meteorological Assimilations Data Ingest System (MADIS)." [Online]. Available: http://www.nws.noaa.gov/ops2/ops24/madis.htm.

[6] "MADIS Meteorological Surface Quality Control." [Online]. Available: https://madis.ncep.noaa.gov/madis_sfc_qc.shtml.

[7] 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.

[8] "NOAA MSAS/RSAS." [Online]. Available: http://msas.noaa.gov/.

[9] F. H. Administration, "Success Stories : Clarus." [Online]. Available: http://www.its.dot.gov/clarus/.

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

[11] 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.

[12] 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.

[13] C. Daly, R. P. Neilson, and D. L. Phillips, "A statistical-topographic model for mapping climatological precipitation over mountainous terrain," *J. Appl. Meteorol.*, vol. 33, no. 2, pp. 140–158, 1994.

[14] C. Daly, G. H. Taylor, and W. P. Gibson, "The PRISM approach to mapping precipitation and temperature," in *Proc., 10th AMS Conf. on Applied Climatology*, 1997, pp. 20–23.

[15] D. Richter, S. Wang, and D. Galarus, "WeatherShare Phase 2 Final Report," *Mont. State Univ.*, 2009.

[16] 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.

[17] 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.

[18] D. Luebbers, U. Grimmer, and M. Jarke, "Systematic development of data mining-based data quality tools," in *Proceedings of the 29th International Conference on Very Large Data Bases (VLDB)*, 2003, pp. 548–559.

[19] C. Bisdikian, R. Damarla, T. Pham, and V. Thomas, "Quality of information in sensor networks," in *1st Annual Conference of ITA (ACITA'07)*, 2007.

[20] R. Devillers, R. Jeansoulin, and others, *Fundamentals of spatial data quality*. ISTE London, 2006.

[21] W. Shi, S. Wang, D. Li, and X. Wang, "Uncertainty-based spatial data mining," *Proc. Asia GIS Assoc. Wuhan, China*, pp. 124–135, 2003.

[22] S. Sathe, T. G. Papaioannou, H. Jeung, and K. Aberer, "A survey of model-based sensor data acquisition and management," in *Managing and Mining Sensor Data*, Springer, 2013, pp. 9–50.

[23] Z. G. Ives, D. Florescu, M. Friedman, A. Levy, and D. S. Weld, "An adaptive query execution system for data integration," in *ACM SIGMOD Record*, 1999, vol. 28, no. 2, pp. 299–310.

[24] N. Sofra, T. He, P. Zerfos, B. J. Ko, K.-W. Lee, and K. K. Leung, "Accuracy analysis of data aggregation for network monitoring," in *MILCOM 2008 - 2008 IEEE Military Communications Conference*, 2008, pp. 1–7.

[25] Z. M. Charbiwala, S. Zahedi, Y. Kim, Y. H. Cho, and M. B. Srivastava, "Toward quality of information aware rate control for sensor networks," in *Fourth International Workshop on Feedback Control Implemenation and Design in Computing Systems and Networks*, 2009.

[26] F. Hermans, N. Dziengel, and J. Schiller, "Quality estimation based data fusion in wireless sensor networks," in *MASS'09. IEEE 6th International Conference on Mobile Adhoc and Sensor Systems, 2009.*, 2009, pp. 1068–1070.

[27] M. Fugini, M. Mecella, P. Plebani, B. Pernici, and M. Scannapieco, "Data quality in cooperative web information systems," *Pers. Commun. citeseer. ist. psu. edu/fugini02data. html*, 2002.

[28] A. Klein and G. Hackenbroich, "How to Screen a Data Stream."

[29] A. Klein and W. Lehner, "How to optimize the quality of sensor data streams," in *ICCGI'09. Fourth International Multi-Conference on Computing in the Global Information Technology, 2009.*, 2009, pp. 13–19.

[30] A. Klein, "Incorporating quality aspects in sensor data streams," in *Proceedings of the ACM first Ph. D. workshop in CIKM*, 2007, pp. 77–84.

[31] A. Klein, H.-H. Do, G. Hackenbroich, M. Karnstedt, and W. Lehner, "Representing data quality for streaming and static data," in *IEEE 23rd International Conference on Data Engineering Workshop, 2007*, 2007, pp. 3–10.

[32] A. Klein and W. Lehner, "Representing data quality in sensor data streaming environments," *J. Data Inf. Qual.*, vol. 1, no. 2, p. 10, 2009.

[33] N. Tatbul, "Qos-driven load shedding on data streams," in *XML-Based Data Management and Multimedia Engineering—EDBT 2002 Workshops*, 2002, pp. 566–576.

[34] D. Carney, U. Çetintemel, A. Rasin, S. Zdonik, M. Cherniack, and M. Stonebraker, "Operator scheduling in a data stream manager," in *Proceedings of the 29th International Conference on Very Large Data Bases-Volume 29*, 2003, pp. 838–849.

[35] M. F. Mokbel, X. Xiong, W. G. Aref, S. E. Hambrusch, S. Prabhakar, and M. A. Hammad, “PLACE: a query processor for handling real-time spatio-temporal data streams,” in *Proceedings of the 30th International Conference on Very Large Data Bases (VLDB)*, 2004, pp. 1377–1380.

[36] B. Babcock, M. Datar, and R. Motwani, “Load shedding for aggregation queries over data streams,” in *Proceedings. 20th International Conference on Data Engineering, 2004.*, 2004, pp. 350–361.

[37] B. Babcock, M. Datar, and R. Motwani, “Load shedding in data stream systems,” in *Data Streams*, Springer, 2007, pp. 127–147.

[38] R. V Nehme and E. A. Rundensteiner, “ClusterSheddy: Load shedding using moving clusters over spatio-temporal data streams,” in *Advances in Databases: Concepts, Systems and Applications*, Springer, 2007, pp. 637–651.

[39] N. Tatbul, U. Çetintemel, S. Zdonik, M. Cherniack, and M. Stonebraker, “Load shedding in a data stream manager,” in *Proceedings of the 29th international conference on Very Large Data Bases (VLDB)*, 2003, pp. 309–320.

[40] N. Tatbul, U. Çetintemel, and S. Zdonik, “Staying fit: Efficient load shedding techniques for distributed stream processing,” in *Proceedings of the 33rd International Conference on Very Large Data Bases (VLDB)*, 2007, pp. 159–170.

[41] N. Tatbul and S. Zdonik, “Window-aware load shedding for aggregation queries over data streams,” in *Proceedings of the 32nd international conference on Very Large Data Bases (VLDB)*, 2006, vol. 6, pp. 799–810.

[42] H. Jeung, S. Sarni, I. Paparrizos, S. Sathe, K. Aberer, N. Dawes, T. G. Papaioannou, and M. Lehning, “Effective metadata management in federated sensor networks,” in *2010 IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing (SUTC)*, 2010, pp. 107–114.

[43] M. A. Hossain, P. K. Atrey, and A. El Saddik, “Modeling and assessing quality of information in multisensor multimedia monitoring systems,” *ACM Trans. Multimed. Comput. Commun. Appl.*, vol. 7, no. 1, p. 3, 2011.

[44] C. C. G. Rodríguez and M. Riveill, “e-Health monitoring applications: What about Data Quality?,” 2010. [Online]. Available: http://ceur-ws.org/Vol-729/paper2.pdf.

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

[46] D. E. Galarus and R. A. Angryk, “Mining robust neighborhoods for quality control of sensor data,” in *Proceedings of the 4th ACM SIGSPATIAL International Workshop on GeoStreaming - IWGS ’13*, 2013, pp. 86–95.

[47] P. A. Pisano, J. S. Pol, A. D. Stern, B. C. Boyce, and J. K. Garrett, “Evolution of the US Department of Transportation Clarus Initiative: Project status and future plans.,” in *Preprints, 23rd Conf. on Interactive Systems (IIPS) for Meteorology, Oceanography, and Hydrology, San Antonio, TX, Amer. Meteor. Soc. A*, 2007, vol. 4.

[48] D. E. Galarus and R. A. Angryk, “Quality Control from the Perspective of a near-Real-Time, Spatial-Temporal Data Aggregator and (re) Distributor,” in *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2014.

[49] “MADIS Quality Control.” [Online]. Available: http://madis.noaa.gov/madis_qc.html.

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

[51] J. Graham, “Ordinary Kriging (Ch. 5.5 - Bailey & Gatrell).” [Online]. Available: http://www.math.umt.edu/graham/stat544/ordkrige.pdf.

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

[53] N. Cressie and C. K. Wikle, *Statistics for spatio-temporal data*. John Wiley & Sons, 2011.

[54] G. Bohling, “KRIGING.” [Online]. Available: http://people.ku.edu/~gbohling/cpe940/Kriging.pdf.

[55] M. L. Stein, *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media, 2012.

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

[57] A. Bárdossy and Z. W. Kundzewicz, “Geostatistical methods for detection of outliers in groundwater quality spatial fields,” *J. Hydrol.*, vol. 115, no. 1, pp. 343–359, 1990.

[58] F. Mosteller, “The jackknife,” *Rev. l’Institut Int. Stat.*, pp. 363–368, 1971.

[59] A. G. Journel, “Nonparametric estimation of spatial distributions,” *J. Int. Assoc. Math. Geol.*, vol. 15, no. 3, pp. 445–468, 1983.

[60] G. Matheron, “The intrinsic random functions and their applications,” *Adv. Appl. Probab.*, pp. 439–468, 1973.

[61] S. R. Fassnacht, K. A. Dressler, and R. C. Bales, “Snow water equivalent interpolation for the Colorado River Basin from snow telemetry (SNOTEL) data,” *Water Resour. Res.*, vol. 39, no. 8, 2003.

[62] P. Goovaerts, *Geostatistics for natural resources evaluation*. Oxford university press, 1997.

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

[64] L. Hartman and O. Hössjer, “Fast kriging of large data sets with Gaussian Markov random fields,” *Comput. Stat. Data Anal.*, vol. 52, no. 5, pp. 2331–2349, 2008.

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

[66] 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.

[67] M. R. Haylock, N. Hofstra, A. M. G. Klein Tank, E. J. Klok, P. D. Jones, and M. New, “A European daily high-resolution gridded data set of surface temperature and precipitation for 1950--2006,” *J. Geophys. Res. Atmos.*, vol. 113, no. D20, 2008.

[68] S. A. Jewell and N. Gaussiat, “An Assessment of Kriging Based Rain-Gauge--Radar Merging Techniques,” *Q. J. R. Meteorol. Soc.*, 2015.

[69] P. Lorkowski and T. Brinkhoff, “Towards Real-Time Processing of Massive Spatio-temporally Distributed Sensor Data: A Sequential Strategy Based on Kriging,” in *AGILE 2015*, Springer, 2015, pp. 145–163.

[70] J. P. Hughes and D. P. Lettenmaier, “Data requirements for kriging: estimation and network design,” *Water Resour. Res.*, vol. 17, no. 6, pp. 1641–1650, 1981.

[71] V. Chaplot, F. Darboux, H. Bourennane, S. Leguédois, N. Silvera, and K. Phachomphon, “Accuracy of interpolation techniques for the derivation of digital elevation models in relation to landform types and data density,” *Geomorphology*, vol. 77, no. 1, pp. 126–141, 2006.

[72] D. W. Wong, L. Yuan, and S. A. Perlin, “Comparison of spatial interpolation methods for the estimation of air quality data,” *J. Expo. Sci. Environ. Epidemiol.*, vol. 14, no. 5, pp. 404–415, 2004.

[73] G. M. Laslett, A. B. McBratney, P. Pahl, and M. F. Hutchinson, “Comparison of several spatial prediction methods for soil pH,” *J. Soil Sci.*, vol. 38, no. 2, pp. 325–341, 1987.

[74] M. A. Oliver and R. Webster, “Kriging: a method of interpolation for geographical information systems,” *Int. J. Geogr. Inf. Syst.*, vol. 4, no. 3, pp. 313–332, 1990.

[75] I. Westerberg, A. Walther, J.-L. Guerrero, Z. Coello, S. Halldin, C.-Y. Xu, D. Chen, and L.-C. Lundin, “Precipitation data in a mountainous catchment in Honduras: quality assessment and spatiotemporal characteristics,” *Theor. Appl. Climatol.*, vol. 101, no. 3–4, pp. 381–396, 2010.

[76] P. Štěpánek, P. Zahradníček, and P. Skalák, “Data quality control and homogenization of air temperature and precipitation series in the area of the Czech Republic in the period 1961--2007,” *Adv. Sci. Res.*, vol. 3, no. 1, pp. 23–26, 2009.

[77] M. Ward and J. Zheng, “Visualization of spatio-temporal data quality,” in *GIS LIS-International Conference*, 1993, vol. 2, p. 727.

[78] N. R. Viney and B. C. Bates, “It never rains on Sunday: the prevalence and implications of untagged multi-day rainfall accumulations in the Australian high quality data set,” *Int. J. Climatol.*, vol. 24, no. 9, pp. 1171–1192, 2004.