Galarus and Angryk *Open Geospatial Data, Software and Standards* (2016) 1:2
DOI 10.1186/s40965-016-0003-2

Open Geospatial Data, Software and Standards
a SpringerOpen Journal

ORIGINAL ARTICLE Open Access

Spatio-temporal quality control: implications and applications for data consumers and aggregators

Douglas E. Galarus[1*] and Rafal A. Angryk[2]

Abstract

Background: Quality control for real-time spatio-temporal data is often presented from the perspective of the original owner and provider of the data, and focuses on general techniques for outlier detection or uses domain-specific knowledge and rules to assess quality. The impact of quality control on the data aggregator and redistributor is neglected. As sensor networks proliferate, multiple providers can distribute and redistribute the same original sensor data. Relationships between providers become complex, with data acquired from original and third-party sources. One provider may acquire data from another, and so forth, resulting in larger data sets with value-added components such as quality indicators, but with costs such as increased lag between original observation times and (re)distribution times.

Methods: The focus of this paper is to define and demonstrate quality control measures for real-time, spatio-temporal data from the perspective of an aggregator to provide tools for evaluation and comparison of overlapping, real-time, spatio-temporal data providers and for assessment and optimization of data acquisition, system operation and data redistribution. We define simple measures that account for temporal completeness and spatial coverage. The measures and methods developed are tested on real-world data and applications.

Results: Our results show that these simple measures combine to form methods that are useful in comparing providers and identifying patterns in data which can then be exploited to optimize system performance relative to bandwidth, and to assess the impact of provider quality control mechanisms.

Conclusion: The simple measures presented demonstrate the utility of quantifying data quality from the perspective of the data aggregator and redistributor.

Keywords: Data quality, Load shedding, Data stream processing, Spatial-temporal data, Data cleaning

Introduction

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. As such, 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, overlapping providers. Increased attention must be given to quality control from the perspective of the aggregator and disseminator of data, and to the impact of quality control on their processes and products.

"Quality" is inherently subjective and dependent on the user and use of the data. Quality control is an exercise in measuring the quality of data, assessing it for a given use, and applying the results to that use. For instance, measures such as accuracy, precision, timeliness, reliability, etc., can be formulated and used by data consumers to determine which data is "good" and which data is "bad" relative to their applications. For real-time applications, data that is not timely (i.e., data that is stale or old when it first becomes available) may be of little use even if it is accurate where-as it may be useful for other applications that are not time-sensitive. Data may be accurate in

* Correspondence: dgalarus@montana.edu
[1]Department of Computer Science, Western Transportation Institute, Montana State University, Bozeman, MT 59717-4250, USA
Full list of author information is available at the end of the article

terms of representing real-world conditions such as ambient air temperature, yet that data becomes unusable if the metadata associated with it such as location or time are incorrect. Having access to provider quality control measures and to data and metadata that can be used to formulate quality control measures is critical to successful use by consumers.

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 data provider may present quality control indicators at the sensor level while another flags at the station level, leaving uncertainty as to which of multiple sensor readings is in question. Aggregating such data into a uniform and cohesive offering is a challenge, as is the task of selecting which providers should be used from multiple, overlapping offerings.

Spatial-temporal data, used in the absence of quality control measures, will likely yield questionable or poor results. Because of these challenges, we must investigate ways to aggregate and derive quality control measures from provided data including sensor observations and timestamps not only corresponding to the original observation, but also to the times at which the data is made available, processed and redistributed. The best approach to improving the quality of data is to start at the source – the sensors. But, we must recognize and work with what is within our control. As aggregators of data from sensor networks controlled by other agencies, we make the best of what they give us and ideally add value to this data. In all likelihood, we will have no control over the content, format and distribution mechanisms used by the providers. We might not even have a direct mechanism for reporting problems to the provider and seek resolution. What we can do is implement our own quality control mechanisms and use them to optimize the performance of our systems.

For instance, we can evaluate the spatial-temporal coverage of provider data in the presence of multiple, overlapping providers and in light of bandwidth and processing constraints. We can seek answers to questions of whether to include data from one provider relative to others. For example, what do we gain in terms of spatial 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? We can evaluate the impact of quality control processes implemented on our systems and in provider systems. We can compare providers to determine overlap, and determine which data we should use based on quality control measures, and we can determine spatial and temporal gaps in the data we are provided.

Quite often, sensor-level quality control processes utilize domain-specific, rule-based systems or general outlier detection techniques to flag "bad" values. For instance, NOAA's Meteorological Assimilation Data Ingest System (MADIS) [1] applies a range of [−60 °F, 130 °F] to its validity check for air temperature observations [2] while the University of Utah's MesoWest [3] uses the range [−75 °F, 135 °F] in their quality control checks for air temperature [4]. These ranges are intended to represent the possible air temperature values that could be observed in real world conditions, at least within the coverage area of the given provider. If an observation falls outside the range, then the provider will flag that observation as having failed the range test and the observation will for all practical purposes be considered "bad". Obviously range tests aren't perfect checks. For instance, the record high United States temperature would fail MADIS's range test, although it would pass MesoWest's test. Both MADIS and MesoWest employ a suite of tests to observations that go beyond their simple range tests. "Buddy" tests are used to compare observations at a given point to neighboring observations. MADIS uses Optimal Interpolation in conjunction with cross-validation to measure the conformity of an observation to its neighbors [2]. MesoWest uses multivariate linear regression to estimate observations [5]. A real observation is compared to the estimate for its location and if the deviation between estimated and observed is high, then the real observation is flagged as questionable.

These approaches help to assess the accuracy of the given observation, yet quality and performance in general needs 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. Formulating such problems is challenging too because, by definition, we generally first view "quality" as being more subjective than "quantity". Our challenge is to express quality in quantifiable terms.

In this paper, we present specific spatio-temporal quality control measures, applicable to a wide variety of spatio-temporal provider data distribution mechanisms. We present practical methods using these quality control measures, and demonstrate their utility.

We do not attempt to correct erroneous data or improve collection at the source. Others state correctly that correction at the source is the best way to improve data quality [6]. Our objective in this paper is to make the most of the data from providers as-is. We do not perform outlier detection or otherwise attempt to assess accuracy, precision or other direct quality measures on

individual sensors. Instead, we use provider quality control descriptors to label "bad" data. In separate work, we tackle the problem of identifying "bad" data [7, 8]. We do not directly address system or network performance or present a distributed approach which would directly interact with sensors in the field. Building on prior work [9], we optimize measures such as coverage relative to bandwidth and scheduling downloads of provider data. Our interest is that of data aggregator/consumer, and we work within the relevant constraints of what can and cannot be controlled from this role.

The rest of this paper is organized as follows: Section 2 provides background from a real life domain and related work, and sets the stage for our approach which is presented in Section 3. In Sections 4 and 5 we present our experimental results and analyze performance. In Section 6 we present conclusions and future work.

Background

Motivation

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. It is critical that we display high-quality information, yet characterizing the quality of the data remains a challenge.

The WeatherShare system [10] 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 in the neighborhood of $70,000, it is unrealistic to expect pervasive coverage of the roadway from RWIS alone. WeatherShare aggregates Caltrans RWIS data along with weather data from other third-party aggregation sources such as MADIS [1] and MesoWest [3] to present a unified view of current weather conditions from approximately 2000 stations within California. A primary benefit of the system is far greater spatial coverage of the state, particularly roadways, relative to the Caltrans RWIS network alone.

We have implemented automated quality control procedures for identification of "bad" data with limited success in the WeatherShare system principally to assess sensor accuracy for Caltrans RWIS. Quality control indicators from MADIS and MesoWest have been considered for use from time to time, but differences across these providers have limited our usage of these indicators. To date, we have not made a concerted effort to reconcile these differences. In recent work [9], we have started to formally quantify and evaluate the impact of quality control, and optimize the performance of these systems relative to quality control measures. Prior to this effort, we have not attempted to evaluate these providers side-by-side for use in our application. Both provide a vast amount of data covering our area of interest (California) and beyond. Taking a greedy (more is better) approach, we have chosen to include data from both systems. As a result, we consume a lot of bandwidth and present data that is redundant and sometimes conflicting. We would like to determine if we truly need data from both providers or if we can get by with just one. We may even need to determine whether we can acquire data from providers closer to the source, if not the original owners of the sensors. MADIS uses MesoWest as a provider for some but certainly not all of its data. As such, we have loosely made the following general assumptions:

Assumption 1: MADIS is a superset of MesoWest.
Assumption 2: MesoWest is more-timely than MADIS.

It is important to recognize that these are only assumptions since, until the work presented in this paper and in [9], we have not rigorously compared these providers side-by-side. However, these assumptions do seem reasonable based on the provider relationship between the two systems and based on what we've casually observed in the performance of our own systems. It should also be noted that the evaluation we present is relative to our application and coverage area and should not be taken as an all-encompassing assessment of the two systems. Both MADIS and MesoWest provide data outside our geographic area of interest, and include data from sensors that we have not made use of in our applications. However, the methodology and measures we develop in this paper are generally applicable.

Figure 1 shows MesoWest stations and status and Fig. 2 shows MADIS stations and status in the WeatherShare system. Table 1 shows station counts by provider, including Caltrans, in the System. The greater number of stations shown for MesoWest versus MADIS is principally due to a choice of giving MesoWest stations greater priority for display in WeatherShare. This design-decision was made based on the assumptions above and to address the problems of overlapping and sometimes conflicting data. In subsequent systems we have chosen to include all observations from both providers, and have observed greater coverage of the state. (See the Western States One-Stop-Shop for Rural Traveler Information [11] and the Caltrans Aviation WeatherShare [12]). Because of the priority we have given the

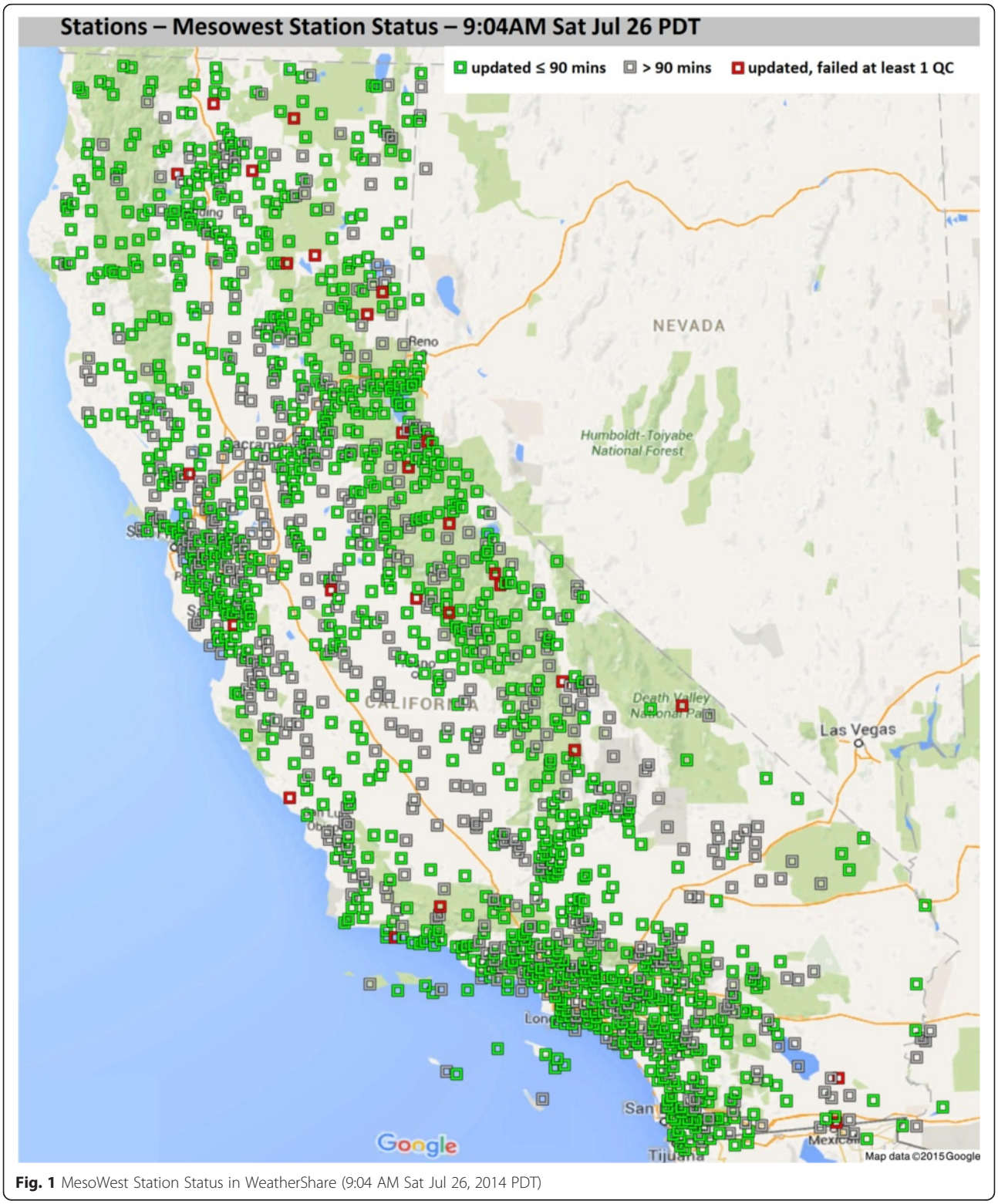

Fig. 1 MesoWest Station Status in WeatherShare (9:04 AM Sat Jul 26, 2014 PDT)

MesoWest data over that from MADIS, the data we show in Fig. 1, Fig. 2 and Table 1 is not representative of the coverage from these providers, particularly the relative number of stations from each. This data only shows a snapshot in time, which isn't necessarily representative of coverage at other points in time.

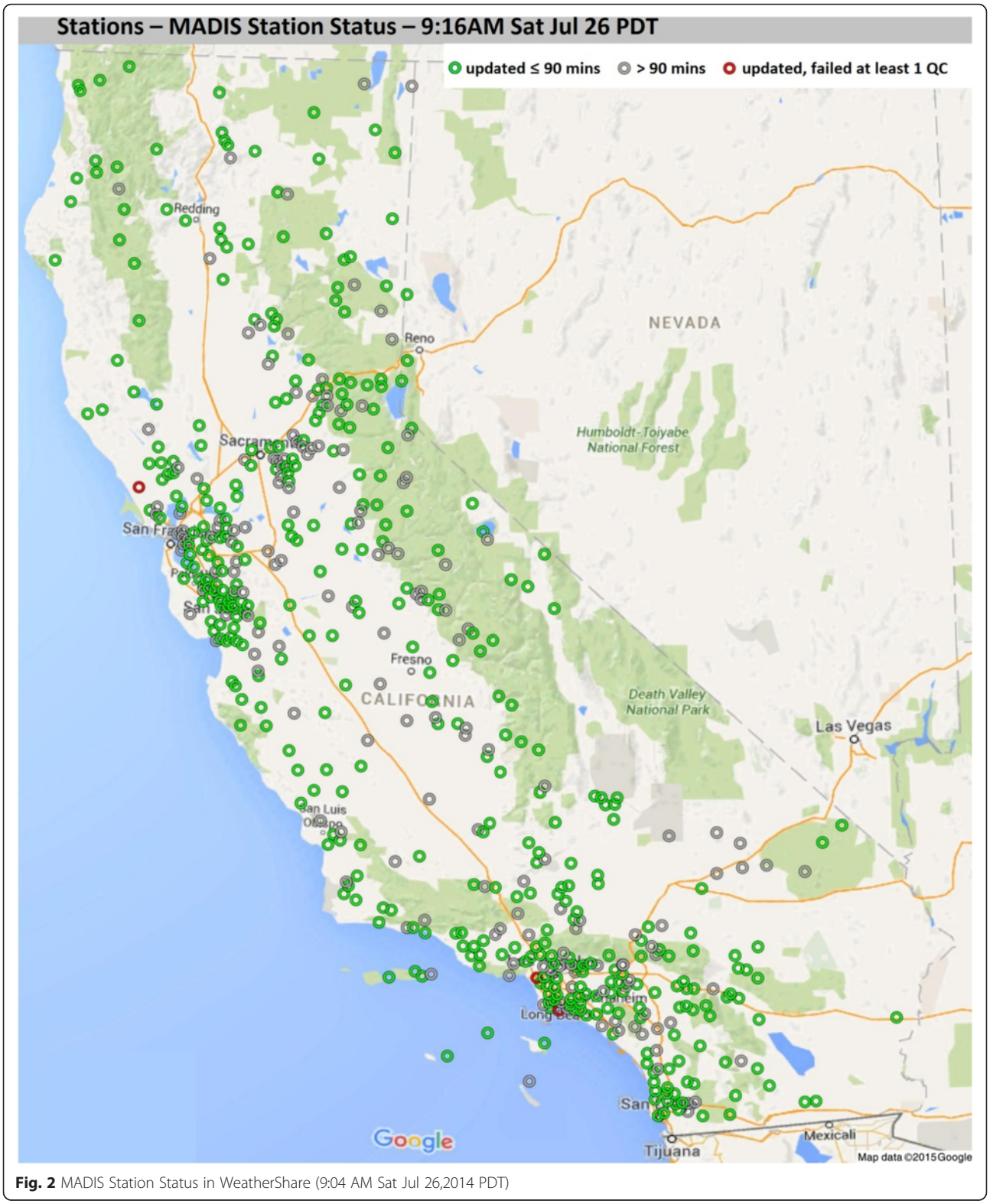

Fig. 2 MADIS Station Status in WeatherShare (9:04 AM Sat Jul 26,2014 PDT)

Several key questions come to mind regarding our use of the MADIS and MesoWest data: 1) What is the benefit in using data from both systems versus just one? For instance, should we use MesoWest in place of MADIS or vice-versa, or should we continue to use both? 2) Which of the two systems provide greater coverage of

Table 1 Status and Station Counts by Provider in WeatherShare (9:04 AM SAT JUL 26, 2014 PDT)

Source	Total	Up to date	Outdated
Caltrans RWIS	107	49	58
MADIS	690	429	261
Mesowest	2474	1158	1316
TOTAL	3271	1636	1635

California? 3) Which of the two systems provides more timely information? Further questions come to mind regarding the use of data from an individual provider such as MADIS: 1) If we depend on MADIS quality control measures to filter out bad data, what is the impact on the performance of our system? 2) What schedule should we follow in downloading the MADIS data so-as to ensure levels of performance while reducing or perhaps minimizing the overall amount of data consumed? In this paper we address these questions and lay the groundwork for finding answers to subsequent quality-related questions.

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 to date have handled these attributes 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 throughout the state of California, yet we desire at a high level a uniform presentation of sensor data in a spatio-temporal sense.

Figure 3 shows the Current Air Temperature layer from Caltrans' Aviation Weathershare [12]. 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 with 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. It is desirable to select a subset sufficient to show representative conditions throughout the entire state.

Figure 4 shows the Caltrans Aviation WeatherShare 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 the apparent bad data. There is an 11 degree reading reported near Oceanside along the coast which is obviously incorrect.

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 as 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, presents significant hazards.

Literature review

Data quality from the perspective of the consumer is presented subjectively in [13] as a comprehensive framework of data quality attributes. Batini, et al. [14] present a more recent survey and summary of data quality dimensions from the literature, and point out varying definitions for dimensions such as timeliness and completeness. Luebbers, et al. [15] develop data mining tools to assist in data quality assessment, and present a definition for data auditing to include measurement and improvement of data quality. Bisdikian, et al. [16] present overlap and differences between Quality of Data and Quality of Information (QoI). While these papers are useful in general terms, they do not include specific, comprehensive measures that can be applied to our spatio-temporal challenges.

Devillers, et al. [17] provide a comprehensive review of spatial data quality, including treatment of temporal aspects, and distinction 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 from Bédard and Vallière [18] which presented six characteristics of external quality for geospatial databases: definition, coverage, lineage, precision, legitimacy, and accessibility. The work given in [19] is relevant because it presents sources of uncertainty in spatial-data mining, and these sources can also be viewed as sources of data quality problems. These sources provide general guidance to us but do not provide specific implementations that address our spatio-temporal situation.

Data cleaning is presented in the context of pull-based and push-based data acquisition in [20], along with a model-based approach to outlier/anomaly detection. Ives, et al. [21] present an adaptive query system for systems integrating overlapping data sources, including query optimization, while Sofra, et al. [22] investigate the trade-offs between accuracy and timeliness of

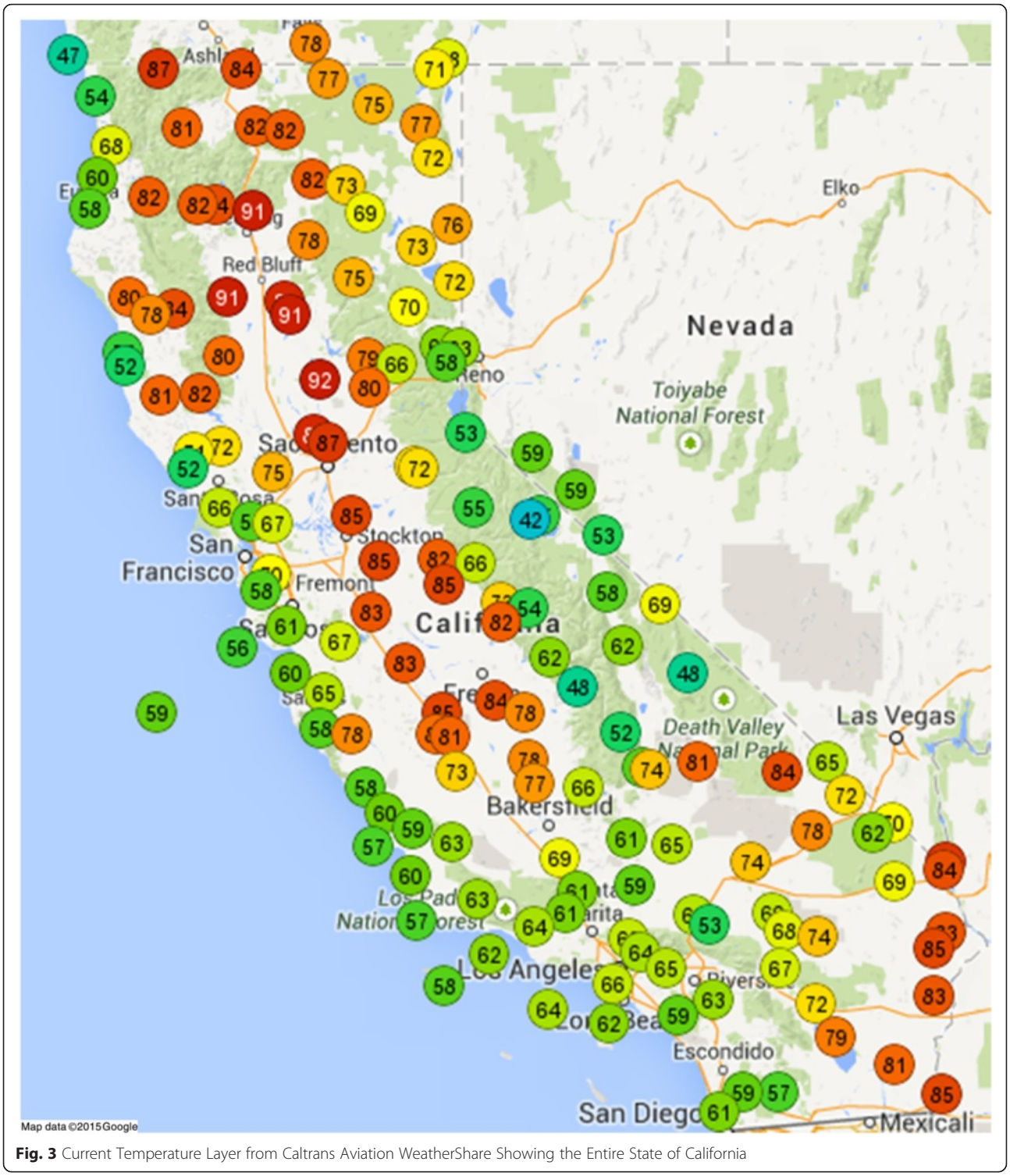

Fig. 3 Current Temperature Layer from Caltrans Aviation WeatherShare Showing the Entire State of California

information acquired in a data aggregation network. Also from the networking domain, the work presented by Charbiwala, et al. in [23] focuses on rate control guided by Quality of Information (QoI) measures. They indicate that such efforts are highly application-dependent. Another network-related publication [24] presents four components of data quality: accuracy, consistency, timeliness and completeness. Timeliness is expressed principally as a network phenomenon. Fugini, et al. [25] define completeness, currency, internal consistency, timeliness, importance, source reliability and confidentiality for cooperative web information

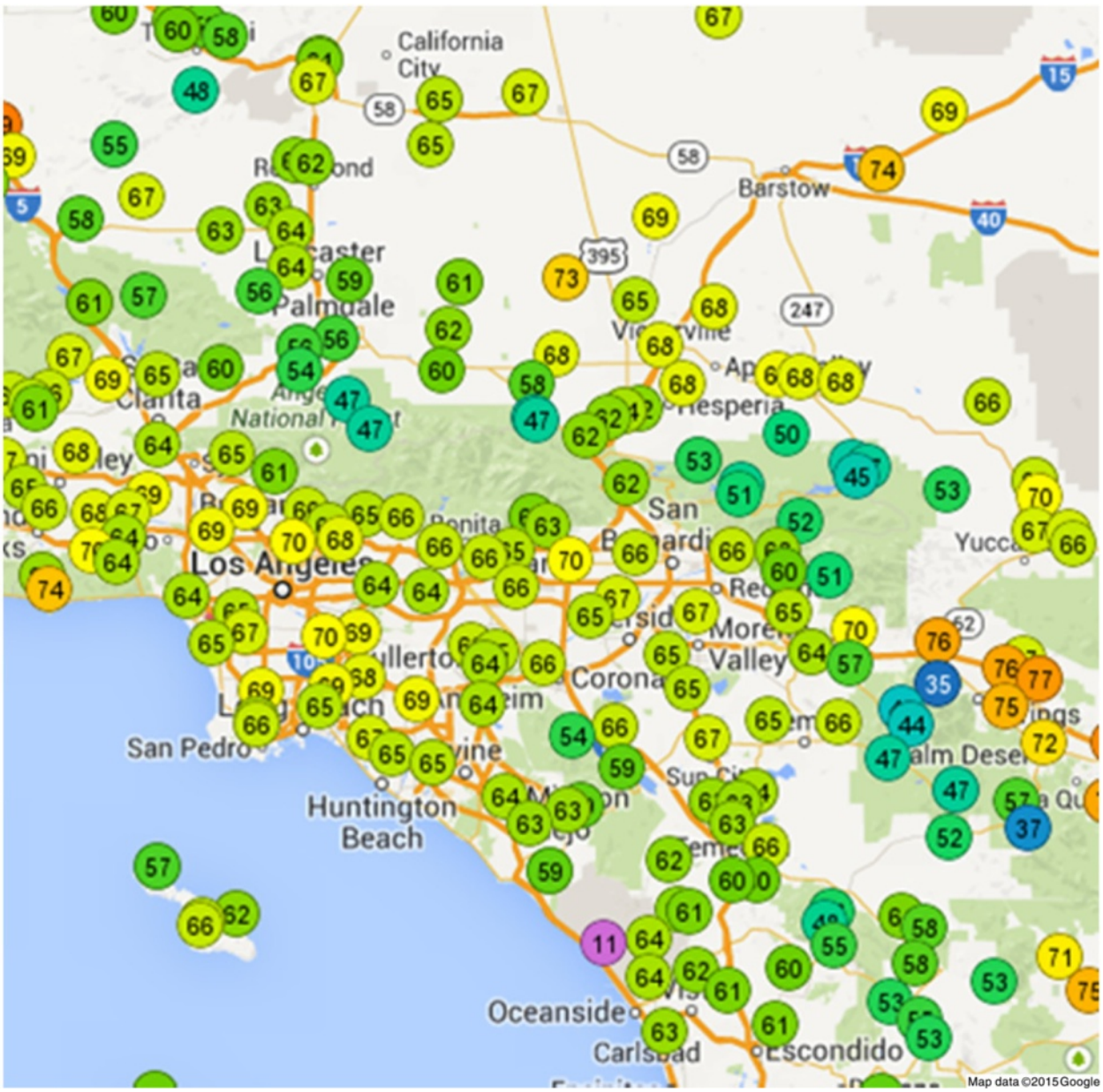

Fig. 4 Current Temperature Layer from Caltrans Aviation WeatherShare in Proximity to Los Angeles

systems. The definitions provided here are general and useful conceptually but require definition specific to spatio-temporal data to be of direct help to us.

The closest work in relation to ours is presented Klein, et al. in [26–30], 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, and all are presented in relation to sensor data streams. The chief missing component in these works relative to ours is an accounting for the spatial aspect. For Klein, et al., data is considered and managed as individual streams. In our work, it is important to not only consider data from individual sites and sensors but the collective of all sites and sensors and their interrelationships. For instance, data from one site may be in error while that from another nearby site may be good. The latter ("good") could be used in place the former ("bad") in many of our applications. Specific measures are presented by Klein, et al. and are of use as examples, while some such as completeness have apparent short-comings for our application which we address in this paper. We wish to

evaluate overlapping data providers, and there is no direct mechanism to do so here. In subsequent work [31, 32], Klein, et al. incorporate their data quality measures into a larger middleware architecture named GINSENG, intended for performance monitoring and control of sensor networks. The specific measures used are similar to those presented by Klein in prior work.

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]. The work in [35] presents load-shedding for spatio-temporal data streams, but does not specifically address quality control measures. Other relevant work regarding load-shedding for data streams can be found in [36–41]. Of these [38], from Nehme, et al. appears most relevant due to its spatio-temporal setting, presenting a clustering approach to load-shedding in which moving objects that are similar in terms of location, time, direction and speed are clustered, and data from individual members of the cluster can be dropped with the representatives of the cluster summarizing them. Our applications do not directly involve moving objects. The added complexity necessary for clustering does not benefit us as a data consumer since we do not have the opportunity select individual data elements for download. Our choices are all-or-nothing relative to providers and their feeds at given publication times.

Jeung, et al. [42] present an automated metadata generation approach that includes a probabilistic measure of data quality. In [43] Hossain, et al. dynamically assess three quality attributes for the detection and identification of human presence in multimedia monitoring systems, where-as Rodríguez and Riveill [44] present data quality in relation to e-Health monitoring systems. Crowd-sourced citizen science [45] and volunteered geographic information [46–48] efforts attract data quality research for obvious reasons. When the public assists in collecting data, the benefits of public collection must be weighed against the potential for poor quality submissions. These efforts do indicate that the benefits of public participation outweigh the drawbacks while leaving open paths for further research. In [45], Kelling, et al. tackle the problem of quality with analysis both of data submission and subsequent observer variation. Goodchild, et al. call upon existing data quality standards such as the US Spatial Data Transfer Standard [49] and the Content Standard for Digital GeoSpatial Metadata [50] while demonstrating the open-ended nature of quality assurance for volunteered geographic information. Barron, et al. [47] reference the ISO 19113 [51], ISO 19114 [52] and ISO 19157 [53] standards while pointing out that data quality for volunteered geographic information projects such as OpenStreetMap (OSM) [54] depends on the user's purpose. In turn they present a framework tailored to "fitness for purpose" with six different categories of purpose and 25 measures within those categories, all specific to OSM. Ballatore, et al. [48] investigate "conceptual quality" using OSM, and indicate wider applicability than that of [47].

While these sources demonstrate ongoing interest and need for related research, none of these approaches directly addresses quality control for spatio-temporal data for our consumer/aggregator situation.

Methods

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. We focus our approach on information available to the consumer of sensor data from a provider. While bounding the scope of our interests, we are cognizant of the complex system through which sensor readings are provided to us. See Fig. 5.

A. Definitions

1) Observations

We first define 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 represent an *original observation* o as a 4-tuple, $o = (s, t, l, v) = (o_s, o_t, o_l, o_v)$, consisting of the source (station/sensor), (original) timestamp, location, and a sensor value. We represent a *provided observation* ω as a 3-tuple, $\omega = (\tau, o, \phi,) = (\omega_\tau, \omega_o, \omega_\phi)$, consisting of the provider timestamp, an original observation, and quality control indicators for the observation from the provider. The provider timestamp indicates the time at which the observation is made available by the provider. The quality control indicators are a set of provider-generated assessments of the quality of the observation. Specific definition of these indicators is provider-dependent.

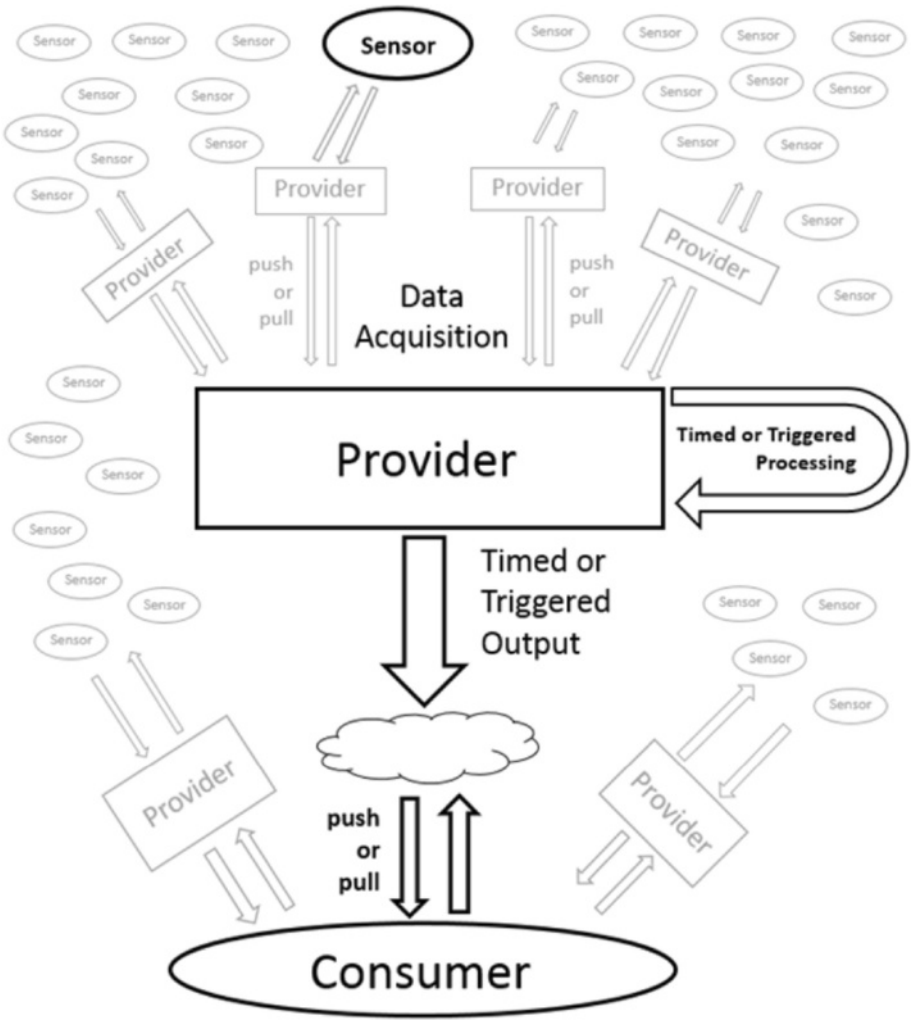

Fig. 5 Consumer, Provider and Sensor Relationships

2) Provider Distribution Mechanisms

We intend 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.

3) Individual Site/Sensor Quality Measures

We first present quality measures relative to an individual site/sensor. These measures form a basis upon which aggregates over time and space can be developed. In this paper, we will use quality control indicators from the provider to assess accuracy/correctness. Luebbers, et al. [15] present accuracy/correctness as answering the question: "*Does the data reflect the entities' real world-state?*" Our reason for using provider quality control indicators in this paper is that we want to assess the impact of such indicators relative to the temporal and spatial quality measures that we will subsequently define. For instance, we are interested in the impact of requiring the use of observations that have "passed" provider quality control versus using any and all observations, including those that have not passed quality control or have not been quality-checked. Our assumption is that quality control assessment may be at least in part a batch process, not run immediately upon receipt of each individual observation. In separate work we

investigate the short-comings of provider quality control and present alternate approaches [7, 8].

The first measure we define is lag. For lag, we use a measure similar to that used for timeliness by others including Klein and Lehner [27] with the caveat that we are principally interested in lag relative to a data provider. For a provided observation $\omega = (\tau, o, \phi,)$ where $o = (o_s, o_t, o_l, o_v)$, we define

$$provider\_lag(\omega) = \tau - o_t.$$

We can also define lag in more general terms for use of an original observation at time t as

$$lag\_at\_use(o, t) = t - o_t.$$

Lag is the difference between the time when an observation occurs and when it becomes available from the provider. When using lag, lesser values are better than greater values, with a lag of 0 being the best that can be theoretically achieved. We note however that it is possible to have computed values of lag that are less than zero if clocks are not synchronized across the system. In fact, it is unrealistic to expect that clocks are synchronized across a large and complex multiagency system. In subsequent work we will tackle the problems of bad metadata, including bad timestamps and unsynchronized clocks, as well as incorrect location data. In [17], Devillers, et al. refer to these issues as relating to temporal accuracy and positional accuracy.

The second measure we define is temporal completeness. The general intent is an indication of how well a time interval is covered by observations. Completeness, or window completeness, is defined in [27] and [28] as the ratio of the number of "originally measured, not-interpolated" values to the containing (time) window size. This could be accomplished by way of a rate. For example, a station might provide 4 observations per hour. Unfortunately, this isn't 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 mins apart. Instead, we define (temporal) completeness in terms of lag. Let O be a set of original observations. We define the current/most-recent observation at time c as:

$$current(O, c) = \arg\max_{o \in O}\{o_t : o_t \leq c\}$$

If we assume a time interval $I = \{a, a+1, a+2, \ldots, b\}$, specified with discrete (seconds, minutes or similar) units, then we can define a number of aggregate measures of lag, $lag\_completeness(O, I)$ for a time interval I based on lag relative to the current/most-recent observation at the points in time within the interval:

$$lag\_completeness(O, I) = \sum_{t \in I} lag(current(O, t), t)$$

$$lag\_completeness(O, I) = \frac{\sum_{t \in I} lag(current(O, t), t)}{|I|}$$

$$lag\_completeness(O, I) = max_{t \in I} lag(current(O, t), t)$$

Note that we measure completeness here in terms of time, and that lesser values are better. The second measure, which is an average over time, is similar to granularity as defined by Klein, et al. in [27]. More elaborate measures using decay and autocorrelation are possible, as well as continuous measures. Such measures are more informative than a simple rate because they provide indications of the age of observations over time. Since the measures above are defined in terms of sets of observations, these measures can be applied to sets that are restricted based on provider quality control indicators. For instance, we might restrict our attention to only the observations that have fully "passed" provider quality control. Doing so can help us assess the impact of provider quality control.

The last measure we define is (spatial) coverage. In [17] Devillers, et al. restate a characteristic provided by Bédard and Vallière [18] that coverage is a measure that "evaluates whether the territory and the period for which the data exists, the 'where' and 'when' meet user needs." Note that while stated in general, this definition is important because it addresses coverage using both spatial and temporal aspects. We seek to define a measure of coverage that is both spatial and temporal in nature.

Our definitions for temporal completeness above apply to sets of observations from a single site/sensor. These definitions can be directly extended to sets of observations from multiple sources. For instance, we can compute lag and completeness for observations from locations within a cell in a spatial grid. In turn, we can define aggregates that include both spatial and temporal aspects of our data. We may do this if we allow "coverage" of the cell to come from multiple sites within the cell.

Assume a discrete time interval $I = \{a, a+1, a+2, \ldots, b\}$. Let G be the geographic area of interest. Assume a partition $\{G_1, G_2, \ldots, G_n\} : G = \cup_{i=1}^{n} G_i,\ \forall_{i \neq j, i,j \in \{1,\ldots n\}} G_i \cap G_j = \varnothing$ of the geographic area of interest. Let O be a set of observations from this geographic region. Partition O as $P = \{O_1, O_2, \ldots, O_n\} : \quad O_i = \{o \in O : o_l \in G_i\}$. Then the following measures can be used to describe spatial coverage relative to the spatial partition $\{G_1, G_2, \ldots, G_n\}$:

$$lag\_coverage(P, I, O) = \frac{\sum_{i=1}^{n} completeness(O_i, I)}{n}$$
$$lag\_coverage(P, I, O) = min_{i=1}^{n} completeness(O_i, I)$$
$$lag\_coverage(P, I, O) = Median(\{completeness(O_i, I) \mid i = 1..n\})$$
$$lag\_coverage(P, I, O) = Q_1(\{completeness(O_i, I) \mid i = 1..n\})$$
$$lag\_coverage(P, I, O) = Q_3(\{completeness(O_i, I) \mid i = 1..n\})$$
$$lag\_coverage(P, I, O) = Percentile_{90}(\{completeness(O_i, I) \mid i \in \{1, ..., n\}\})$$
$$lag\_coverage(P, I, O, c) = |\{i : i \in \{1, ..., n\}, completeness(O_i, I) > c\}|$$

We use the prefix "lag" to help recall that these measures are based on lag and are ultimately measures of time.

B. Applications

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 have passed all provider quality control checks. We can use the measures defined above to investigate a number of interesting problems:

1) Comparison of Providers
   Using criteria such as coverage, we can compare providers directly. Using a measure of coverage from above, we can compare one provider to another and we can compare the impact of combining data from providers. The comparison can be made using overall measures or grid-based measures.
2) Coverage of Maps and Gap Analysis
   If we partition our data into a geographic grid, we can use our coverage measures to assess overall coverage of a region and identify gaps in coverage. For instance, we could use a grid to assess coverage of a map such as that shown in Fig. 3, and determine gaps where coverage is less than that desired for individual cells. We can also attempt to determine parameters to provide a given level of coverage such as age of data – we might determine, for instance that in order to provide coverage of the map, we need to show data up to 90 mins old versus data that is less than 60 mins old.
3) Impact of Quality Control
   Using the coverage measures defined earlier, we can investigate the impact of provider quality control. Assuming that quality control, at least in part, is conducted via batch process on the provider, we assume there will be a delay between the time in which an observation first becomes available and when that observation is assessed for quality. If the provider supplies observations both prior to and following quality control assessment, then this impact can be analyzed. This analysis can be done for the overall data set, by individual site/sensor, or by groupings such as a geographic grid.
4) Optimal Download Schedule and/or Load Shedding
   Using criteria such as coverage, we can attempt to optimize relative to processing time, bandwidth consumed, etc. For instance, we might choose to consume enough data to maintain a certain level of coverage, but not consume excessive data if doing so results in little change in coverage. Further, there may be optimal times at which to consume data, corresponding to internal processes of the provider such as data import and quality control assessment schedules, and relative to our own needs such as coverage of a specific geographic area.

Results and discussion part I – multiple provider comparison

In this section we apply our measures to several multi-provider challenges we face on the various Weathershare projects using data from MADIS [1] and MesoWest. [3] The MesoWest and MADIS datasets provide an interesting opportunity to apply these measures because:

- The datasets overlap, with MesoWest providing data to MADIS.
- The datasets are provided using different, file-based distribution mechanisms. MesoWest provides a single file that is updated approximately every 15 mins, and is not cumulative. MADIS provides files grouped by hour – all observations for a given hour go into the same file. Files are updated as new observations come into the system, and updated files are published roughly every five minutes. There is a great deal of redundancy in successive versions of a single hours' file.
- Both datasets provide quality control indicators although different approaches are used by each provider. MADIS provides quality control indicators at the sensor level, while MesoWest provides quality control indicators at the station level. MADIS also provides sensor observations at various stages in the quality assessment process. For instance, an observation may be distributed first in the absence of some of the quality control checks which run in batch and is subsequently updated when these checks are run. See Tables 2 and 3.

We restrict our attention to a grid consisting of two-hundred sixty-eight 0.5° Latitude x 0.5° Longitude cells which overlap with or are adjacent to California. This grid includes cells overlapping with the Pacific Ocean, Mexico, Nevada and Arizona. A finer grid or other, non-

uniform partition could also be used. We further restrict our attention to the time period spanning June 2014 GMT. During this time period, we downloaded and stored every MADIS file from the Mesonet subset when we detected that the file was updated, and kept separate copies corresponding to each update, and did similar for MesoWest. Downloading all data is the best case possible for these data sets in terms of coverage and completeness, but may not be feasible for normal operation because of the amount of data involved, particularly for MADIS. See Table 4.

For both data sets, the original files are compressed (gzip). Table 4 shows compressed file sizes and is representative of the size of the file downloads. From these files, temperature observations for our area and time period of interest were extracted. Note that these temperature observations make up a relatively small fraction of the data contained within these files since the spatial coverage of both MADIS and MesoWest exceeds our area of interest, and many observation types other than temperature are included. The One-Stop-Shop [11] provides coverage of multiple states, and all of the systems make use of sensors in addition to temperature, so there is added benefit in using the all-inclusive feeds from these providers.

For our experiment, we filter out data that has been flagged as bad by at least one of the provider-specific quality control checks. We include data for which some of the quality control checks may not have been applied and may still be pending. Our reason for doing this is that we want to determine the timeliest display of correct data that is possible, and if necessary, we can depend on quality control mechanisms implemented on our systems to further filter out bad data. For MADIS, this means that we include all observations have a QCD flag of “C”, “S”, “V” or “G”. See Table 5. For MesoWest, we include observations that have a QC flag of “0”, “2” or “9”. Note that we further exclude observations for which the observation time is subsequent to the provider time (timestamp of the file in which it is provided). There were a number of observations for which this occurred that had otherwise not failed quality control, including MesoWest’s “Suspect Time” check, and the inclusion of these observations would bias the results. Stated simply: these observations would appear to fall in the future relative to the time in which they were provided. The likely cause of this problem is discrepancies between system clocks, and there were observations with differences as great as 19.5 hours. See Table 6.

Table 2 MesoWest Quality Control Flags (QFLAG)

QFLAG		Description
-1	Suspect	One of the variables in an observation did not pass the "range checks".
0	Unknown	MesoWest Quality Control processes have not been applied to this observation.
1	Caution	Data in this observation did not pass one of the "statistical checks".
2	OK	Data has passed all MesoWest Quality Control processes successfully.
3	Suspect Time	The reported time of the observation appears suspect.
9	N/A	Station lacks significant data to run the multivariate linear regression analysis.

Source [56]:

Table 3 MADIS Quality Control Descriptors (QCD)

QC Descriptor		Description
Z	Preliminary	no QC
C	Coarse pass	passed level 1
S	Screened	passed levels 1 and 2
V	Verified	passed levels 1, 2, and 3
X	Rejected/erroneous	failed level 1
Q	Questioned	passed level 1, failed 2 or 3
G	Subjective good	“accept” list overrides other QC
B	Subjective bad	“reject” list overrides other QC

level 1 = validity; level 2 = internal consistency, temporal consistency, statistical spatial consistency checks;
level 3 = spatial consistency check
Source [2]:

Table 4 Size of Download Data

	# files	Size (Gigabytes)
MADIS Mesonet	20862	168.14 GB
Mesowest	2879	1.98 GB

A. Application of Quality Control Measures

For each cell in the grid, we compute 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

Table 5 MADIS Temperature Observations by QCD

QC Flag	Original count	# Used in our experiment
S	6910240	6910240
G	14149	14149
V	23333	23333
Q	7889	
X	6706	
TOTAL	6,962,317	6,947,722

Table 6 MESOWEST Temperature Observations by FLAG

QC flag	Original count	Obs_Time > File_Time	# Used in our experiment
-1	259,762	340	
0	207,582	9949	197633
1	623,163	1103	
2	8,495,685	4474	8,491,211
3	11,308	11307	
9	280,748		280748
TOTAL	9,878,248		8,969,592

corresponding to the cell. Thus, the most recent observation from any site within the cell will be counted as the current observation for the cell at a given point in time. Our reason for doing this is that 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. Data is analyzed for only observations that have not failed the respective system's quality control process. Note that for both providers this includes data that may not have had all quality control checks applied.

We use the first measure of lag_coverage presented as an indication of spatio-temporal coverage. This measure allows us to examine various lag thresholds/cutoffs.

B. Results

Table 7 shows cell counts and percents for (average) lag_coverage by provider. Figure 6 shows (average) lag_coverage for observations from MesoWest. Figure 7 shows (average) lag_coverage for observations from MADIS. Figure 8 shows (average) lag_coverage for observations from MesoWest and MADIS, combined.

Table 7 Cell Counts and Percents for Average Lag_Coverage by Provider

	MesoWest		Madis		Combined	
Avg Lag ≤ 10 min	0	0.0 %	2	0.7 %	4	1.5 %
Avg Lag ≤ 20 min	30	11.2 %	135	50.4 %	139	51.9 %
Avg Lag ≤ 30 min	147	54.9 %	174	64.9 %	180	67.2 %
Avg Lag ≤ 40 min	180	67.2 %	195	72.8 %	202	75.4 %
Avg Lag ≤ 50 min	203	75.7 %	208	77.6 %	217	81.0 %
Avg Lag ≤ 60 min	218	81.3 %	221	82.5 %	231	86.2 %
Avg Lag ≤ 90 min	231	86.2 %	226	84.3 %	234	87.3 %
90 min < Avg Lag	10	3.7 %	8	3.0 %	7	2.6 %
No Coverage	27	10.1 %	34	12.7 %	27	10.1 %

1) Coverage of Maps and Gap Analysis
   There are apparent gaps in coverage, for which there is no coverage from either provider or for which the timeliness could be improved. This includes the desert area east and northeast of Los Angeles and San Diego. There are also several locations in the Central Valley as well as northern California in which the timeliness could be improved. Note that these are rural areas, and communication challenges likely impact the timeliness of data transmission. While there is some coverage of grids primarily overlapping the Pacific Ocean, there are grids that are not covered at all. These grids could be excluded from subsequent analysis. If using a 90 mins cutoff for lag, a majority of the map will be covered by either of the providers individually and for both combined.
2) Comparison of Providers
   If we target a 90 mins or less lag for display of an observation, then MesoWest does provide greater coverage than MADIS by five cells: 231 cells (86.2 %) versus 226 cells (84.3 %). However, MADIS holds an advantage for lower cutoffs. For observations with a 60 mins or less lag, MADIS covers 221 grids (82.5 %) while MesoWest covers 218 grids (81.3 %). For observations with a 20 mins or less lag, MADIS covers 135 grids (50.4 %), while MesoWest only covers 30 grids (11.2 %).
   If we combine both data sets, we see further improvement. Over two-thirds of the grids will be covered with an average lag of 30 mins or less and nearly 90 % of the grids will be covered by an average lag of 90 mins or less. There will only be a 1.1 % reduction in coverage if we reduce our cutoff to 60 mins. This small reduction in coverage would allow us to download less MADIS files. MesoWest adds coverage at several maritime cells that MADIS does not cover, as well as a cell in Arizona near the southeast border with California. More importantly it adds coverage for a cell adjacent to Interstate 10 east of Barstow. MADIS improves the timeliness of data across nearly the entire map. There are a few exceptions including a cell northwest of Bakersfield in which MesoWest improves the average lag_coverage by over 50 mins. Another cell located east of San Diego is improved by over 40 mins.
   Our use of data from both providers does appear to be justified. If we were to include data from just one of the providers, MADIS would generally provide more timely data while MesoWest would provide coverage of some cells that MADIS does not cover.

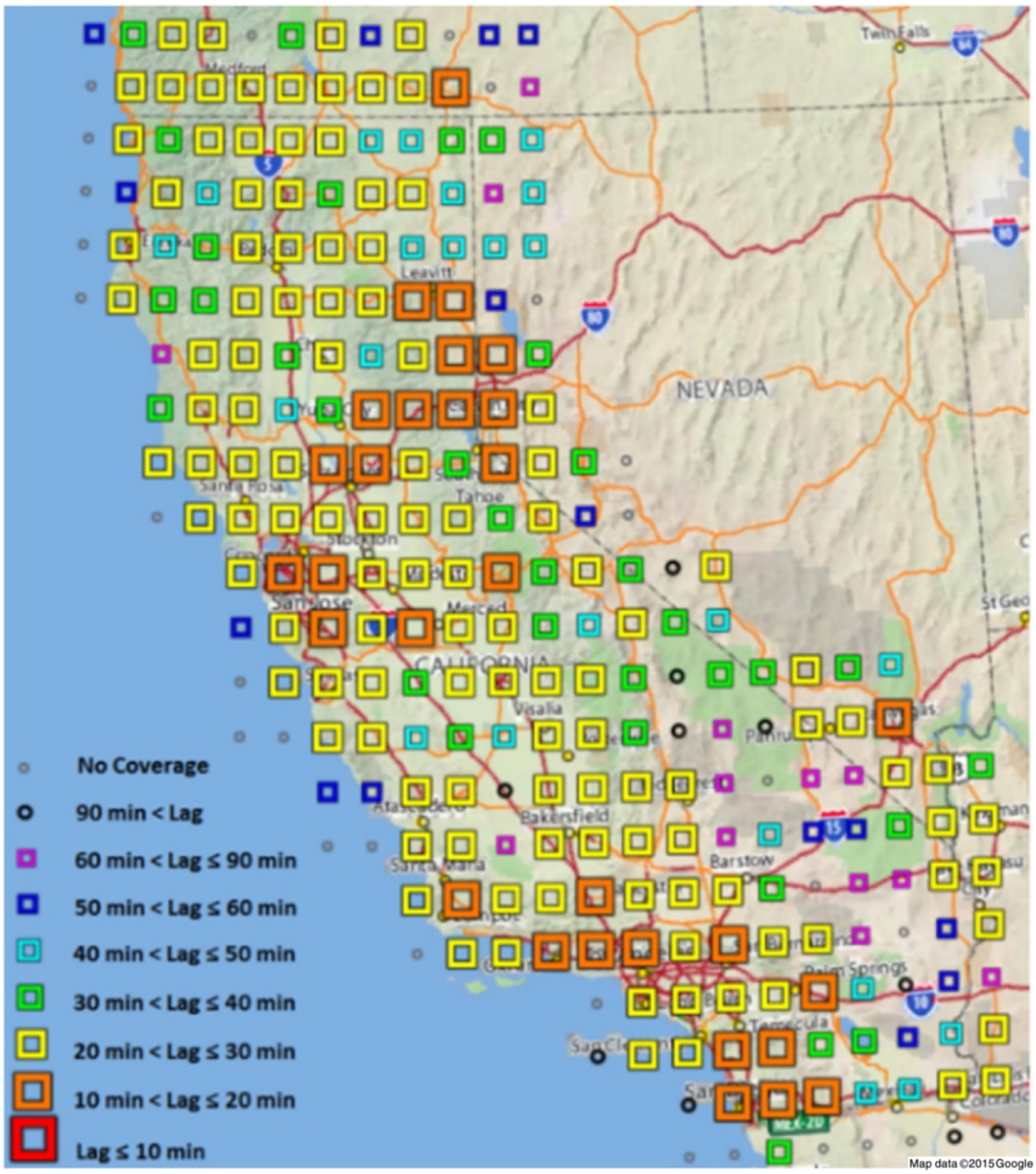

Fig. 6 Coverage by MesoWest

As such, the providers are complementary for our application. If we were most concerned with the overall size of our data downloads, we could possibly use MesoWest data alone, but we would be sacrificing timeliness of data in the process.

Results and discussion part II – single provider optimization

In this section we apply our measures to several single-provider challenges we face on the various Weathershare projects using data from MADIS [1]. The MADIS

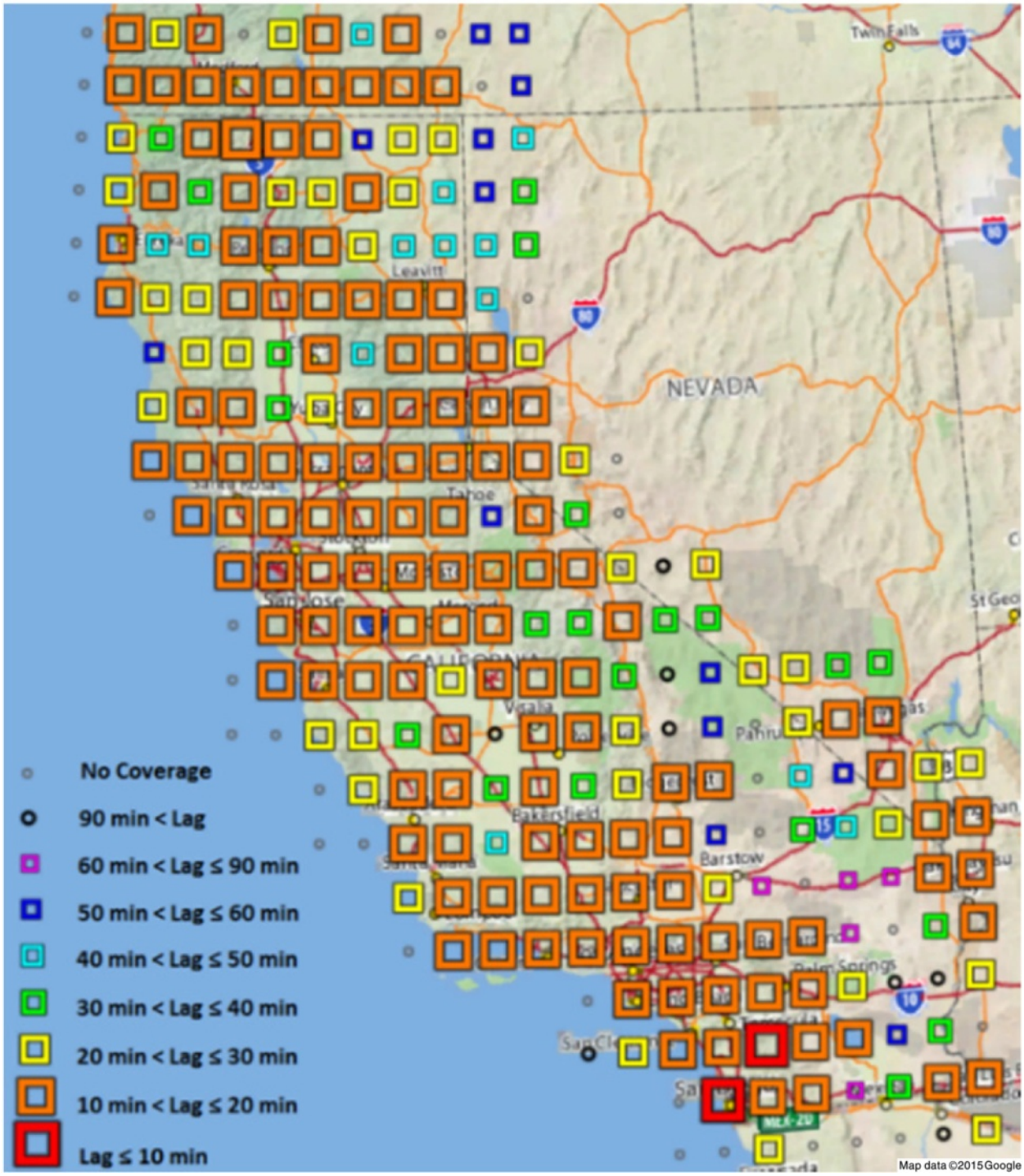

Fig. 7 Coverage by MADIS

dataset provides an interesting opportunity to apply these measures.

MADIS stores files by hour – all observations for a given hour go into the same file. Files are updated as new observations come into the system, and updated files are published roughly every five minutes. Files are named using the format YYYYMM_HH00, corresponding to the hour (GMT) of the included observations.

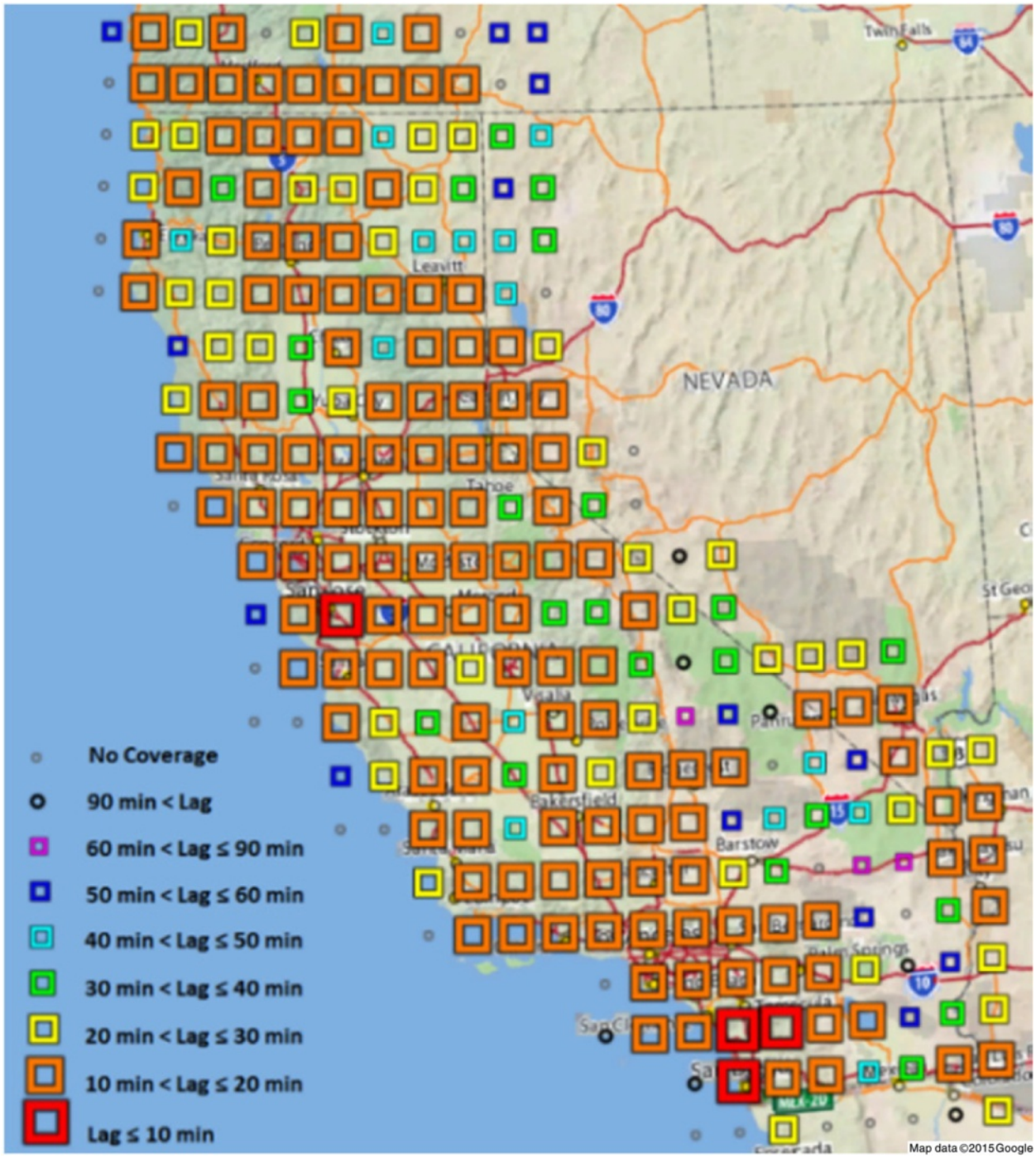

Fig. 8 Coverage by MADIS and MesoWest COMBINED

The files use the NETCDF format and are compressed using gzip. To include timestamp information for each update of an hourly file, we represent these files as $F_{h,t}$, where h is the hour represented by the file and t is the timestamp of the file. Then for each $\omega = (\tau, o, \phi,) \in F_{h',t'}$ where $o = (o_s, o_t, o_l, o_v)$, we have $\tau = t'$, $h' \leq o_t < h' + 60$ minutes. For instance, $F_{20140309\_1700,20140309\_1704}$ represents the file containing observations between 5 PM and 6 PM on March 9th, 2014 GMT and time-stamped at 5:04 PM. This is the first time in which observations

falling within that hour were made available and only observations with timestamps early in that hour were present at that time. Compressed, this file had a size of 14.8 KB. The next update to this particular hour's observations occurred at 5:08 PM, measuring 749 KB and we represent this file as $F_{20140309\_1700,20140309\_1708}$. The 20140309_1700 file was updated 25 more times in the next several hours with the final update, $F_{20140309\_1700,20140309\_1920}$, occurring at 7:20 PM and measuring 14.2 MB in size. If all 27 updates of this file were retrieved, over 256 MB of bandwidth would be consumed. Note that each file will contain only one copy of an individual observation, so there is no duplication within the files. However, subsequent file versions will contain observations that were included in prior versions as well as new observations, resulting in a considerable amount of duplication.

We use air temperature for this investigation. This is one of many variables provided in the MADIS MesoNet feed. The MADIS dataset provides multiple levels of quality control checks, dependent on the sensor type [2, 55]. The quality control checks are rule-based as described in [2]. For a given temperature observation, the MADIS quality control description (QCD) field indicates an assessment of the quality of the observation. A value of QCD = V indicates that the observation has been verified and has passed three levels of quality control checks. A value other than QCD = V can occur if either the observation has failed to pass one of the quality control checks or if less than three levels of quality control were checked for the observation. We use ϕ=QCD as the quality control indicator for the provided observation. Following are descriptions of the various quality control checks performed by MADIS on air temperature data. Note that the first three of these are domain-specific and rule-based:

- Level 1 Validity Check: If an air temperature observation falls between −60 F° and 130 F°, it passes this QC check. All values outside this range fail this QC check.
- Level 2 Temporal Consistency Check: If an air temperature observation differs from another air temperature observation at the same site by 35 F° or more within an hour, then the observation fails this QC check. Otherwise, it passes this QC check.
- Level 2 Internal Consistency Check: Checks consistency between readings from different sensors at the same site. For instance, dew point temperature cannot exceed the air temperature. If it does, both observations are flagged as failing this QC check.
- Level 2 Statistical Spatial Consistency Check: Using weekly statistics, if observations from a particular site/sensor has failed any QC check 75 % of the time during the past 7 days, observations will be marked as failing this QC check until the failure rate falls below 25 %.
- Level 3 Spatial Consistency ("Buddy") Check: Using Optimal Interpolation, if an observation differs significantly from its neighbors, then it fails this QC check. Cross-validation is used in conjunction with Optimal Interpolation to determine the conformity (or lack-there-of) for neighboring observations.

These QC checks are not necessarily performed at the same time. For instance, the Level 3 check might be applied in a timed, batch process rather than immediately when a new observation is acquired. A single original observation may result in multiple provided observations corresponding to times at which the containing hourly file is updated. The quality control QCD value may change as subsequent quality control checks are applied.

We restrict our attention to a grid consisting of fifty-six 1° Latitude x 1° Longitude cells which overlap with California. This grid includes cells overlapping with the Pacific Ocean, Mexico, Nevada and Arizona. A finer grid or other, perhaps non-uniform partitions could also be used. There are sensors located in all of these cells. See Fig. 9. We further restrict our attention to the time period between 3/5/2014 16:22 GMT and 3/17/2014 17:19 GMT. During this time 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. Downloading all data corresponds to the best case possible for this data set in terms of coverage and completeness, but may not be feasible for normal operation because of the massive amount of data involved.

A. Application of Quality Control Measures

For each cell in the grid, we compute completeness as the average lag (in seconds) of data within the cell over all time units within the period for which we collected data.

$$lag\_completeness(O,I) = \frac{\sum_{t\in I} lag(current(O,t),t)}{|I|}$$

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 at a given point in time. Our reason for doing this is that 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 then assess

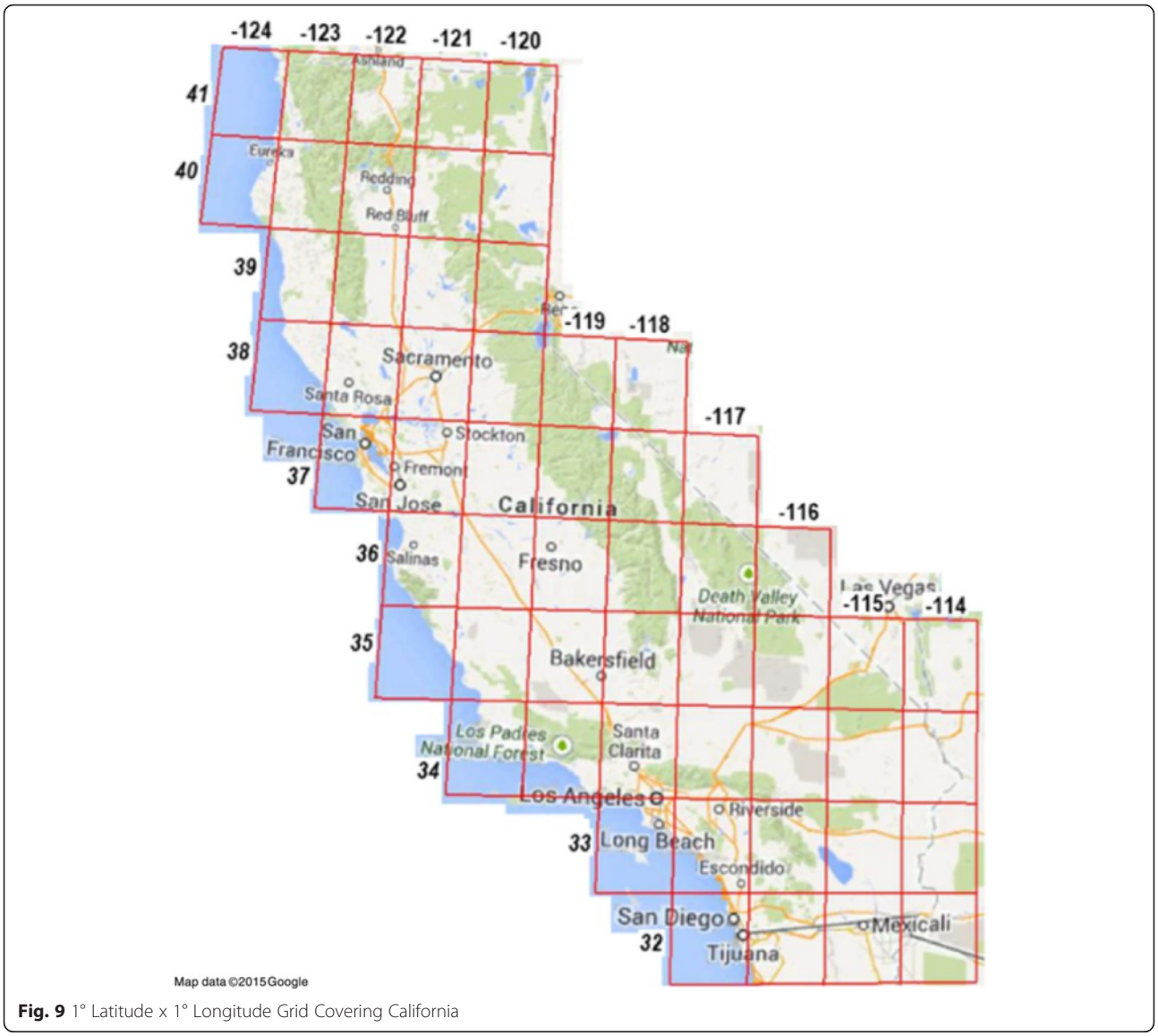

Fig. 9 1° Latitude x 1° Longitude Grid Covering California

coverage using summary statistics over all the cells. Data is analyzed for all data versus QC-passed data.

We can represent the files as they become available using sequence notation: $\mathcal{F} = \langle F_{h_1,t_1}, F_{h_2,t_2}, ..., F_{h_n,t_n} \rangle$. As mentioned previously, it is not practical in our production system to download every file when it becomes available. We did so for a relatively short period of time for the purpose of the analysis presented here, but doing so in our production system would require excessive bandwidth with potentially little gain in coverage and completeness, and we would find ourselves constantly processing new versions of files. As such, we want to determine optimal download schedules and decide what data we can ignore. Specifically, we want to determine a download schedule to be carried out within every hour. For instance, starting on the hour, we could download files every 15 mins. We represent this schedule as: $\{0, 15, 30, 45\}$. Our reason for choosing hourly download schedules is for ease of implementation and that we recognize that the provider follows hourly schedules as well. Other time periods could be handled in a similar fashion.

In conjunction with determining an optimal schedule, we wish to determine if there is an optimal data age constraint. At present we have chosen to show data on our maps that is no more than 90 mins old. Our rationale for this cutoff is that if data is older than 90 mins, then conditions may have changed significantly. However, we have chosen not to use a lesser cut-off out of concern that we would sacrifice

coverage of the map. We can formalize our 90 mins age restriction in conjunction with a schedule and the sequence of updated files as follows:

Let a download schedule S be represented by $\{s_1, \ldots, s_n | s_i \in \{0, 1, \ldots, 59\}\}$. Then at a schedule time $s' \in S$, we download the following files:

$$\left\{F_{h',t'} : h' = hour\left(h'\right), s' - \left(h' + 60\ minutes\right) \leq 90\ minutes, t' = argmax_{t:t \leq s'}\left\{F_{h',t}\right\}\right\}$$

Using the schedule {0, 15, 30, 45} and a 90 mins cutoff, we would have the following example behavior: At time 2014-03-10 17:45, we would download the most recent 20140310_1700 and 20140310_1600 files. At time 2014-03-10 18:15, we would download the 20140310_1800, 20140310_1700 and 20140310_1600. At most, we will download three files using this 90 mins cutoff. If we use a 60 mins cutoff, however, we will only need to download two files.

B. Results

1) Impact of Provider QC

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 quality control checks.

Table 8 shows summary statistics indicating overall coverage as demonstrated by (temporal) completeness of individual cells in seconds. Table 9 shows summary statistics for the differences in completeness of cells in seconds between all data and data that passed MADIS quality-control.

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 mins (900 s). The greatest average lag is nearly 45 mins (2700 s). If we only use data that has passed quality control, 75 % of the cells show an average lag no more than 24 mins (1440 s). The greatest average lag is 66 mins (3960 s). In general, there will be a 10 mins or greater additional lag for using data that has passed quality control versus using all data. This lag is suspected to be due to batch processing of quality control checks. In the extreme case (41 mins), the lag may also be attributable to a higher proportion of bad data in that cell and perhaps delayed communication.

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

2) Coverage of Maps/Gap Analysis

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. Not surprisingly 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 mins. This extreme value indicates that this area does not include sensors that report observations passing quality control in a timely manner, either due to bad data or slow reporting or both. There are some observations from this cell that pass QC. Further investigation would be needed to assess the cause. See Table 10.

Table 8 Overall Coverage as Demonstrated by Summary Statistics for Completeness of Cells in Seconds

Set	Ω	Ω_{QC}
Min	554.8	1195.4
Q1	637.9	1224.4
Median	697.7	1264.7
Q3	895.7	1429.3
Max	2673.7	3959.2

Table 9 Differences in completeness of cells in seconds between all data and data that passed MADIS Quality-Control

Min	399.9
Q1	526.0
Median	589.7
Q3	620.5
Max	2477.1

Table 10 Outlier Cells in Terms of Completeness (in seconds)

Cell		*Lag_Completeness (sec)*		*Difference*
Latitude	Longitude	Ω	Ω_{QC}	*sec*
37	−117	*1360.6*	*1814.4*	*453.8*
36	−116	*1369.4*	*1815.2*	*445.8*
35	−121	*1482.1*	*3959.2*	*2477.1*
38	−118	*1508.9*	*2139.2*	*630.3*
36	−117	*1666.9*	*2452.8*	*785.9*
35	−116	*1715.7*	*2212.0*	*496.3*
32	−115	*2548.4*	*3157.2*	*608.8*
34	−115	*2673.7*	*3295.2*	*621.5*

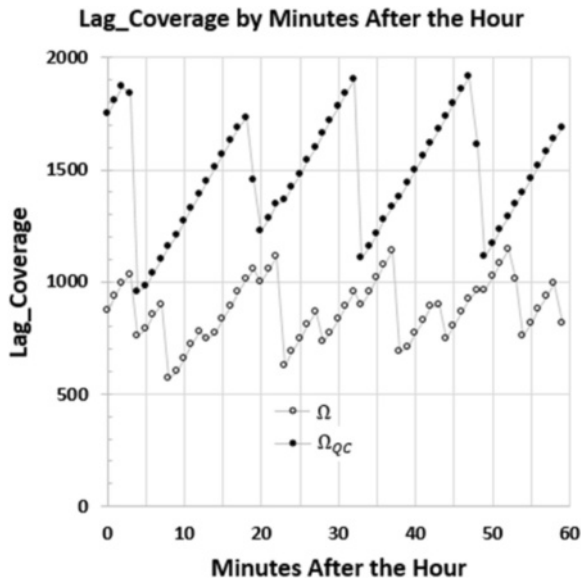

Fig. 10 Average Lag_Coverage at each Minute During the Hour

There isn't a lot we can do about this other than perhaps adjust our own download and processing schedules accordingly. However, awareness of this deficiency allows us to better focus on things we can control such as our download schedule.

3) Optimal Download Schedule/Load Shedding
It isn't practical to download all data as soon as it becomes available (approximately every five minutes). There is too much redundancy in the data, which would result in excessive bandwidth consumption. However, downloading all data in this fashion over a short period of time can help us in determining optimal download schedules.

In Fig. 10 we show lag by minute (average over all cells) for both the Ω and the Ω_{QC} data sets. There are some apparent patterns. For the Ω dataset, the least lag occurs at 8 mins after the hour. As a result, if we were to make just one download, it would be optimal to do this at 8 mins after the hour. There are other times with low lag including 23, 38, and 54 mins after the hour. And there are further good times including 44, 59 and 4 mins after the hour and several others, with an apparent, approximately 15 mins period. We attribute this pattern to different schedules for data coming into the system as well as batch output and other batch processing. For the Ω_{QC} data set, the pattern is clearer, and doesn't correspond exactly to that for the Ω data set. 4, 20, 33 and 49 yield local best times. Otherwise, there is a subsequent lag that corresponds directly to time elapsed. It appears that there is a batch process that runs approximately every 15 mins, and an optimal download schedule should take this into account.

For the Ω data set, optimal schedules at an increasing number of times yield improved coverage, but also increased bandwidth. It is debatable whether more than four download times would improve coverage sufficient

Table 11 Optimal Download Schedules for the Ω Data Set

Schedule	Size (GB)	Lag_Coverage (sec)
{8}	11.5	2339.4
{8,38}	16.9	1499.8
{8,23,44}	21.8	1250.6
{8,23,38,54}	27.8	1080.6
{8,23,38,44,54}	33.6	1029.7
{4,9,23,38,44,54}	37.9	989.6
{4,9,23,28,38,44,54}	42.5	957.1
{4,8,14,23,28,38,44,54}	43.3	933.7
{Download all Files}	67.8	872.8

Table 12 Optimal Download Schedules for the Ω_{QC} Data Set

Schedule	Size (GB)	Lag_Coverage (sec)
{4}	11.3	2725.9
{4,33}	16.5	1904.2
{4,33,49}	22.7	1665.2
{4,20,33,49}	26.9	1515.5
{4,5,20,33,49}	27.9	1506.3
{4,5,20,23,33,49}	28.9	1499.4
{4,5,20,23,33,48,49}	33.8	1492.8
{4,5,19,20,23,33,48,49}	34.0	1486.7
{Download all Files}	67.8	1474.4

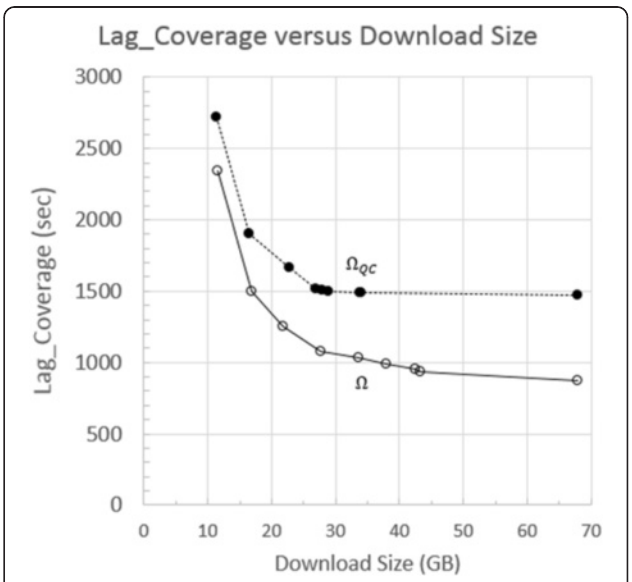

Fig. 11 Lag_Coverage versus Download Size for Optimal Download Schedules

Table 13 Measures for Downloads of Only the Most Recent File

Ω Schedule = { 8,23,38,54 }		Ω_{QC} Schedule = { 4,20,33,49 }	
Size (GB)	Lag coverage (sec)	Size (GB)	Lag coverage (sec)
4.9 GB	1082.2	4.2GB	1517.2

to merit the added bandwidth. See Table 11. For the Ω_{QC} data set, the optimal schedule for four downloads yields coverage that is only 45 s greater than the best possible, yet it requires less than half the bandwidth. See Table 12. There doesn't seem to be 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 Fig. 11.

We can make further improvements by recognizing that the lag_coverages correspond to observations that are less than an hour old. If we restrict ourselves to files containing data observed within the last hour, then we can reduce the download size even further. In fact, we can restrict ourselves to only the file for the current hour and the prior hour and get comparable results. See Table 13. For the Ω dataset, the {8,23,38,54} schedule yields a Lag_Coverage of 1082.2 s, which is less than 2 s 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 the 27.8 GB when downloading all new files. For the Ω_{QC} dataset, the {4,20,33,49} schedule yields a Lag_Coverage of 1517.2 s, which is also less than 2 s worse than the same schedule when downloading all new files at those times. The download size is 4.2 GB, which is far less than the 26.9 GB required to download all new files for the same schedule. These results look even better when compared against downloading all files at all times, which would consume 67.8 GB. See Table 13.

Conclusions and future work

The relatively simply measures we present in this paper were demonstrated as useful in helping to answer complex problems related to visual coverage of a map with data acquired from overlapping, third-party providers. They allow for comparison of the providers, and they allow for optimization related to bandwidth/load-shedding. These measures help to reveal underlying patterns related to acquisition, processing and provision of data by a provider. They can be implemented in a simple manner and are generally applicable to a wide variety of situations for consumers of spatio-temporal data from third-party data providers.

In future work we intend to investigate methods for detecting what loosely may be described as bad metadata. In terms of spatial and temporal attributes, we will specifically attempt to identify data for which the timestamps or the locations are incorrect. The approaches we used in this paper, combined with methods we developed in [7, 8] and [9] provide a foundation we can build upon for this task.

Competing interests
The authors declare that they have no competing interests.

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

Author details
[1]Department of Computer Science, Western Transportation Institute, Montana State University, Bozeman, MT 59717-4250, USA. [2]Department of Computer Science, Georgia State University, Atlanta, GA 30302-5060, USA.

Received: 2 December 2015 Accepted: 4 January 2016
Published online: 03 March 2016

References

1. NOAA, "Meteorological Assimilation Data Ingest System (MADIS)." [Online]. Available: http://madis.noaa.gov/. [Accessed: 26-Dec-2015].
2. NOAA, "MADIS Meteorological Surface Quality Control." [Online]. Available: https://madis.ncep.noaa.gov/madis_sfc_qc.shtml. [Accessed: 26-Dec-2015].
3. U. of Utah, "MesoWest Data." [Online]. Available: http://mesowest.utah.edu/. [Accessed: 26-Dec-2015].
4. U. of Utah, "MesoWest Data Variables." [Online]. Available: http://mesowest.utah.edu/cgi-bin/droman/variable_select.cgi. [Accessed: 26-Dec-2015].
5. Splitt ME, Horel JD. 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: Amer. Meteor. Soc; 1998. p. 359–62.
6. De Veaux RD, Hand DJ. How to Lie with Bad Data. Stat Sci. 2005;20(3):231–8.
7. DE Galarus, R A Angryk, and JW Sheppard, "Automated Weather Sensor Quality Control.," FLAIRS Conf., pp. 388–393, 2012. ISBN number: ISBN 978-1-57735-558-8.
8. Galarus DE, Angryk RA. Mining robust neighborhoods for quality control of sensor data. Proc 4th ACM SIGSPATIAL Int Work GeoStreaming - IWGS. 2013;13:86–95.
9. DE Galarus and RA Angryk, "Quality Control from the Perspective of the Real-Time Spatial-Temporal Data Aggregator and (re)Distributor," in ACM SIGSPATIAL '14, 2014.
10. WTI/MSU, "The WeatherShare System." [Online]. Available: http://www.weathershare.org/. [Accessed: 29-Dec-2015].
11. WTI/MSU, "The Western States One-Stop-Shop for Rural Traveler Information." [Online]. Available: http://oss.weathershare.org/. [Accessed: 29-Dec-2015].
12. WTI/MSU, "Caltrans Aviation WeatherShare." [Online]. Available: http://aviation.weathershare.org/. [Accessed: 29-Dec-2015].
13. Wang RY, Strong DM. Beyond accuracy: What data quality means to data consumers. J Manag Inf Syst. 1996;12(4):5–33.
14. Batini C, Cappiello C, Francalanci C, Maurino A. Methodologies for data quality assessment and improvement. ACM Comput Surv. 2009;41(3):16.
15. Luebbers D, Grimmer U, Jarke M. Systematic development of data mining-based data quality tools. Proc 29th Int Conf Very large data bases. 2003;29:548–59.
16. C Bisdikian, R Damarla, T Pham, and V Thomas, "Quality of information in sensor networks," in 1st Annual Conference of ITA (ACITA'07), 2007.
17. Devillers R, Jeansoulin R. Fundamentals of Spatial Data Quality, Chapter 2. Spatial Data Quality: Concepts. Newport Beach: Wiley-ISTE; 2010.
18. Y Bédard and D Vallière, "Qualité des données à référence spatiale dans un contexte gouvernemental," *Université Laval, Quebec*, 1995, p. 53.
19. W Shi, S Wang, D Li, and X Wang, "Uncertainty-based spatial data mining," Proc. Asia GIS Assoc. Wuhan, China, pp. 124–135, 2003.

20. S Sathe, T G Papaioannou, H Jeung, and K Aberer, "A survey of model-based sensor data acquisition and management," Managing and Mining Sensor Data, Springer 2013, pp. 9–50. ISBN: 978-1-4614-6309-2.
21. Ives ZG, Florescu D, Friedman M, Levy A, Weld DS. An adaptive query execution system for data integration. ACM SIGMOD Rec. 1999;28(2):299–310.
22. N Sofra, T He, P Zerfos, BJ Ko, K-W Lee, and KK Leung, "Accuracy analysis of data aggregation for network monitoring," MILCOM 2008 - 2008 IEEE Mil. Commun. Conf., pp. 1–7, 2008. ISBN: 978-1-4244-2676-8.
23. ZM Charbiwala, S Zahedi, Y Kim, YH Cho, and MB 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.
24. Hermans F, Dziengel N, Schiller J. Quality estimation based data fusion in wireless sensor networks. MASS'09 IEEE 6th Int Conf Mob Adhoc Sens Syst. 2009;2009:1068–70.
25. M Fugini, M Mecella, P Plebani, B Pernici, and M Scannapieco, "Data quality in cooperative web information systems," Personal Communication. citeseer. ist. psu. edu/fugini02data. html, 2002. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.18.9821&rep=rep1&type=pdf. [Accessed: 26-Dec-2015].
26. A Klein and G Hackenbroich, "How to Screen a Data Stream." [Online]. Available: http://mitiq.mit.edu/ICIQ/Documents/IQ%20Conference%202009/Papers/3-A.pdf Conference 2009/Papers/3-A.pdf. [Accessed: 26-Dec-2015].
27. Klein A, Lehner W. How to Optimize the Quality of Sensor Data Streams. Proc 2009 Fourth Int Multi-Conference Comput Glob Inf Technol. 2009;00:13–9.
28. A Klein, "Incorporating quality aspects in sensor data streams," Proc. {ACM} first {Ph.D.} Work. {CIKM}, pp. 77–84, 2007. ISBN: 978-1-59593-832-9.
29. Klein A, Lehner W. Representing Data Quality in Sensor Data Streaming Environments. J Data Inf Qual. 2009;1(2):1–28.
30. A Klein, HH Do, G Hackenbroich, M Karnstedt, and W Lehner, "Representing data quality for streaming and static data," Proc. - Int. Conf. Data Eng., pp. 3–10, 2007. ISBN: 978-1-4244-0832-0.
31. Z Jerzak, A Klein, and G Hackenbroich, "GINSENG data processing framework," in Reasoning in Event-Based Distributed Systems, Springer Berlin Heidelberg, 2011, pp. 125–150.
32. O'donovan T, Brown J, Büsching F, Cardoso A, Cecílio J, Furtado P, Gil, A Jugel, W-B Pöttner, U Roedig, and others, The GINSENG system for wireless monitoring and control: Design and deployment experiences. ACM Trans Sens Networks. 2013;10(1):4.
33. N Tatbul, "Qos-driven load shedding on data streams," XML-Based Data Manag. Multimed. Eng. 2002 Work., pp. 566–576, 2002.
34. Carney D, Çetintemel U, Rasin A, Zdonik S, Cherniack M, Stonebraker M. Operator scheduling in a data stream manager. VLDB. 2003;29:838–49.
35. Mokbel MF, Xiong X, Aref WG, Hambrusch SE, Prabhakar S, Hammad MA. PLACE: a query processor for handling real-time spatio-temporal data streams. Proc Thirtieth Int Conf Very Large Data Bases. 2004;30:1377–80.
36. Babcock B, Datar M, Motwani R. Load shedding for aggregation queries over data streams. Proc - Int Conf Data Eng. 2004;20:350–61.
37. B Babcock, M Datar, and R Motwani, "Load shedding in data stream systems," Data Streams: Models and Algorithms. Springer, 2007. pp. 127–147.
38. RV Nehme and EA Rundensteiner, "ClusterSheddy: Load shedding using moving clusters over spatio-temporal data streams," 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," Proceeding VLDB '03 Proc. 29th Int. Conf. Very large data bases, Vol. 29 pp. 309–320, 2003.
40. N Tatbul, U Çetintemel, and S Zdonik, "Staying fit: Efficient load shedding techniques for distributed stream processing," Proc. 33rd Int. Conf. Very Large Data Bases, pp. 159–170, 2007. ISBN: 978-1-59593-649-3.
41. Tatbul N, Zdonik S. Window-aware load shedding for aggregation queries over data streams. Proc 32nd Int Conf Very Large Data Bases. 2006;6:799–810.
42. H Jeung, S Sarni, I Paparrizos, S Sathe, K Aberer, N Dawes, TG Papaioannou, and M. Lehning, "Effective Metadata Management in Federated Sensor Networks," SUTC 2010 - 2010 IEEE Int. Conf. Sens. Networks, Ubiquitous, Trust. Comput. UMC 2010 - 2010 IEEE Int. Work. Ubiquitous Mob. Comput., pp. 107–114, 2010. ISBN: Print ISBN: 978-1-4244-7087-7.
43. Hossain MA, Atrey PK, El Saddik A. Modeling and assessing quality of information in multisensor multimedia monitoring systems. ACM Trans Multimed Comput Commun Appl. 2011;7(1):1–30.
44. CCG Rodríguez and M Riveill, "e-Health monitoring applications: What about Data Quality?," 2010. [Online]. Available: http://ceur-ws.org/Vol-729/paper2.pdf.
45. Kelling S, Fink D, La Sorte FA, Johnston A, Bruns NE, Hochachka WM. Taking a 'Big Data'approach to data quality in a citizen science project. Ambio. 2015;44(4):601–11.
46. Goodchild MF, Li L. Assuring the quality of volunteered geographic information. Spat Stat. 2012;1:110–20.
47. Barron C, Neis P, Zipf A. A Comprehensive Framework for Intrinsic OpenStreetMap Quality Analysis. Trans GIS. 2014;18(6):877–95.
48. A Ballatore and A Zipf, "A conceptual quality framework for volunteered geographic information," in Spatial Information Theory, Springer International Publishing CY - Cham 2015, pp. 89–107.
49. USGS, "Spatial Data Transfer Standard (SDTS)." [Online]. Available: http://mcmcweb.er.usgs.gov/sdts/. [Accessed: 28-Dec-2015].
50. F (Federal G. D. Committee), "Content Standard for Digital Geospatial Metadata." [Online]. Available: http://www.fgdc.gov/metadata/csdgm/. [Accessed: 28-Dec-2015].
51. "ISO 19113:2002, Geographic information-Quality principles." 2002.
52. "ISO 19114:2003, Geographic information – Quality evaluation procedures." 2003.
53. "ISO 19157:2013, Geographic information – Data quality." 2013.
54. OpenStreetMap Foundation, "OpenStreetMap," Open Database License (ODbL). 2013.
55. NOAA, "MADIS Quality Control." [Online]. Available: https://madis.ncep.noaa.gov/madis_qc.shtml. [Accessed: 26-Dec-2015].
56. U. of Utah, "MesoWest Quality Control Flags Help Page." [Online]. Available: http://mesowest.utah.edu/html/help/key.html. [Accessed: 26-Dec-2015].