Quality Control from the Perspective of a near-Real-Time, Spatial-Temporal Data Aggregator and (re)Distributor

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

Rafal A. Angryk
Department of Computer Science
Georgia State University
Atlanta, GA 30302

ABSTRACT
Quality control for near-real-time spatial-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. The focus of this paper is to define and demonstrate quality control measures for real-time, spatial-temporal data from the perspective of the aggregator to provide tools for assessment and optimization of 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.

Categories and Subject Descriptors
H.2.8 [**Database Management**]: Database Applications – *data mining, spatial databases and GIS.*

General Terms
Algorithms, Measurement, Timeliness, Completeness, Coverage, Reliability.

Keywords
Data Quality; Data Stream Processing, Load-Shedding, Spatial-Temporal Data

1. INTRODUCTION
With the proliferation of sensor networks and the evolving "Internet of Things" as well as the ease of aggregating and redistributing data from multiple 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 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. Spatial-temporal data, used in the absence of quality control measures, will yield questionable or poor results. We must investigate ways to derive quality control measures from provided data including sensor observations and timestamps which account for spatial and temporal aspects of applications.

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

Scope: We do not attempt to correct erroneous data or improve collection at the source. Others state correctly that correction at the source is the best way to improve data quality. [1] Our objective in this paper is to make the most of the data from a provider 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 to problem of identifying "bad" data. [2][3] 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.

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

2. 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 web-based systems such as WeatherShare (http://www.weathershare.org) 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 they display quality information.

WeatherShare aggregates Caltrans RWIS data along with weather data from other third-party aggregation sources such as NOAA's Meteorological Assimilation Data Ingest System (MADIS) (http://madis.noaa.gov/) to present a unified view of current weather conditions from approximately 2000 stations within California. There are several key questions regarding the use of the MADIS data: 1) What is the impact of using MADIS quality control measures to filter out bad data on the performance of our systems? 2) What schedule should we follow in downloading the MADIS data so-as to ensure levels of performance while

SIGSPATIAL '14, November 04-07 2014, Dallas/Fort Worth, TX, USA

Copyright 2014 ACM 978-1-4503-3131-9/14/11…$15.00
*http://dx.doi.org/10.1145/2666310.2666426*INTRODUCTION

minimizing the amount of data consumed? 3) How does the MADIS data compare to that from other providers such MesoWest? Can we use MesoWest in place of MADIS, or should we continue to use both? In this paper we address the first two questions and lay the groundwork for answering questions such as the third, which will be addressed in future work.

Literature Review: Data quality from the perspective of the consumer is presented subjectively in [4], as a comprehensive framework of data quality attributes. [5] presents a more recent survey and summary of data quality dimensions from the literature, and points out varying definitions for dimensions such as timeliness and completeness. [6] presents overlap and differences between Quality of Data and Quality of Information (QoI). While these papers are useful in general, they do not include specific, comprehensive measures that can be applied to our spatial-temporal situation.

[7] provides a comprehensive review of spatial data quality, includes some treatment of temporal aspects, and distinguishes between internal and external quality. The authors also cite and expand on prior work which presented six characteristics of external quality for geospatial databases. [8] 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.

The closest work in relation to ours is presented in [9][10][11][12], addressing 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.

None of these approaches directly addresses quality control for spatial-temporal data that is immediately applicable to our situation.

3. OUR APPROACH

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.

3.1 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, $\omega = (\tau, o, \varphi,) = (\omega_\tau, \omega_o, \omega_\varphi)$, 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.

3.2 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.

3.3 Quality Measures

We first present quality measures relative to an individual site / sensor and extend these measures to form a basis for aggregates over time and space. In this paper, we use provide quality control indicators to assess accuracy. In separate work we present alternate approaches for assessment of accuracy. [2][3]

First we define lag. We use a measure similar to timeliness in [9] 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 it becomes available from the provider. For a provided observation $\omega = (\tau, o, \varphi,)$ where $o = (o_s, o_t, o_l, o_v)$, we define

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

The second measure we define is temporal completeness, which indicates how well a time interval is covered by observations. Window completeness is defined in [9] and [10] 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 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 minutes apart. Instead, we define (temporal) completeness using lag. Let O be a set of original observations. We define the current observation at time c as

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

If we assume a time interval I, then we define:

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

This measure is similar to granularity in [9]. 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 define (spatial) coverage. [7] 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.

Assume a time interval I and a geographic area of interest G. Assume a partition $\{G_1, G_2, \dots, G_n\}$ of G. Let O be a set of observations from this geographic region. Partition O as $P = \{O_1, O_2, \dots, O_n\}: O_i = \{o \in O : o_l \in G_i\}$. Then measures such as the following can be used to describe spatial coverage relative to the spatial partition $\{G_1, G_2, \dots, G_n\}$:

$$lag\_coverage(P, I, O) = \frac{\sum_{i=1}^{n} completeness(O_i, I)}{n}$$

$$lag\_coverage(P, I, O) = min_{i=1}^{n} completeness(O_i, I)$$

$$lag\_coverage(P, I, O, c) = |\{i: i \in \{1, \dots, n\}, completeness(O_i, I) > c\}|$$

4. EXPERIMENTAL RESULTS

4.1 Data

We test our measures by direct application to several challenges we face on the various Weathershare projects, using data from MADIS.[13]

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. [14][15] 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 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 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 restrict 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.

4.2 Use 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 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.

4.3 Results

4.3.1 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 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.

4.3.2 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. 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.

4.3.3 Download Schedule

In Figure 1, 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 doesn't 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 1.

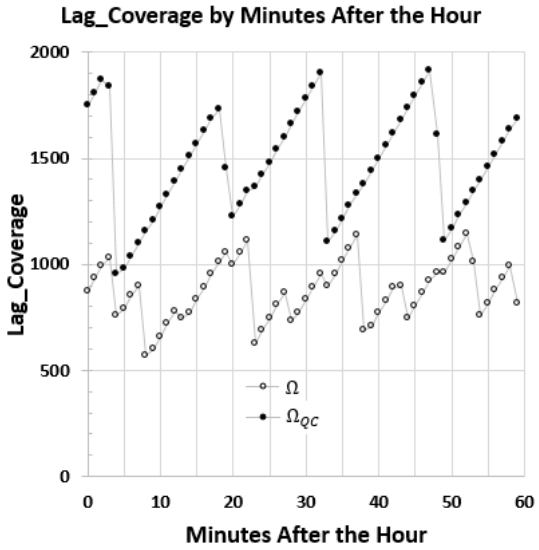

Figure 1: Average Lag_Coverage at each Minute During the Hour

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 2.

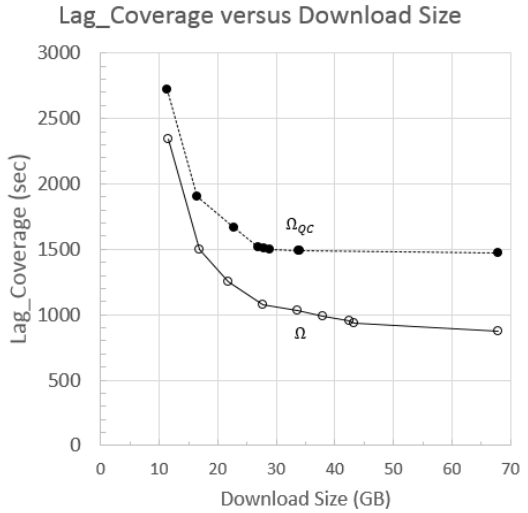

Figure 2: Lag_Coverage versus Download Size for Optimal Download Schedules

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.

5. CONCLUSIONS AND FUTURE WORK

The simple measures we present 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 help 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 spatial-temporal data from third-party data providers.

In future work we will use these measures to analyze data from multiple providers with overlapping data. Given two data providers with similar and overlapping but non-identical offerings of spatial-temporal data, we are interested in determining if data from one provider can be used in lieu of that from the other or if both are necessary.

We also intend to investigate methods for detecting bad metadata. In terms of spatial and temporal attributes, we will identify data for which the timestamps or the locations are incorrect. The approaches we used in this paper, combined with methods we developed in [2] and [3] provide a foundation we can build upon for this task.

6. ACKNOWLEDGMENTS

The California Department of Transportation (Caltrans) sponsored WeatherShare and other related projects. We acknowledge Ian Turnbull and Sean Campbell from Caltrans, Daniell Richter and other WTI staff for their work and support 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.

REFERENCES

[1] R. D. De Veaux and D. J. Hand, "How to lie with bad data," *Stat. Sci.*, pp. 231–238, 2005.

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

[3] 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.

[4] 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.

[5] 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.

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

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

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

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

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

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

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

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

[14] "MADIS Meteorological Surface Quality Control." [Online]. Available: http://madis.noaa.gov/madis_sfc_qc.html.

[15] "MADIS Quality Control." [Online]. Available: http://madis.noaa.gov/madis_qc.html.