2008

Techniques for Computing Fitness of Use (FoU) for Time Series Datasets with Applications in the Geospatial Domain

Lei Fu
University of Nebraska

Leen-Kiat Soh
University of Nebraska, lsoh2@unl.edu

Ashok Samal
University of Nebraska - Lincoln, asamal1@unl.edu

Published in *Geoinformatica* 12 (2008), pp. 91-115; doi: 10.1007/s10707-007-0025-0

Submitted October 4, 2006, revised January 19, 2007; accepted February 8, 2007; published online May 5, 2007; posted in the University of Nebraska-Lincoln Digital Commons October 2010.

Techniques for Computing Fitness of Use (FoU) for Time Series Datasets with Applications in the Geospatial Domain

Lei Fu, Leen-Kiat Soh, and Ashok Samal

Department of Computer Science and Engineering
University of Nebraska–Lincoln, Lincoln, NE, USA

Abstract
Time series data are widely used in many applications including critical decision support systems. The goodness of the dataset, called the Fitness of Use (FoU), used in the analysis has direct bearing on the quality of the information and knowledge generated and hence on the quality of the decisions based on them. Unlike traditional quality of data which is independent of the application in which it is used, FoU is a function of the application. As the use of geospatial time series datasets increase in many critical applications, it is important to develop formal methodologies to compute their FoU and propagate it to the derived information, knowledge and decisions. In this paper we propose a formal framework to compute the FoU of time series datasets. We present three different techniques using the Dempster-Shafer belief theory framework as the foundation. These three approaches investigate the FoU by focusing on three aspects of data: data attributes, data stability, and impact of gap periods, respectively. The effectiveness of each approach is shown using an application in hydrological datasets that measure streamflow. While we use hydrological information analysis as our application domain in this research, the techniques can be used in many other domains as well.

Keywords: Fitness of Use (FoU), Dempster-Shafer belief theory, time series data

1 Introduction

Time series datasets contain the value of a certain variable as a function of time. In the geospatial domain, such datasets are used to store measurements for many natural and human events at different points along the time line. Over the years, many time series datasets have been collected in many domains such as meteorology, agriculture, biology, ecology, hydrology and oceanography. Large amounts of these time series datasets are now universally becoming available and being widely used (e.g., water data [1]). In many of these domains the datasets are used to compute various pieces of information and to derive knowledge that is used as thebasis for making critical decisions. For example, the trends in flow

of streams and amount of precipitation are often used to determine the severity of droughts [2]. This information is then used to decide the amount and type of support that can be provided to farmers and ranchers.

In most applications, it is commonly assumed that the datasets are perfect and without any blemish. This assumption is, of course, not true. The data is merely a representation of a continuous reality both in space and time. It is difficult to measure the values of a continuous space and time variable with infinite precision. Limitations are also the result of inadequate human capacity, sensor capabilities and budgetary constraints. The discrepancy therefore exists between the reality and the datasets that are derived to represent it. It is critical to capture the degree of this discrepancy and incorporate it in the analysis to obtain a more accurate picture of the temporal processes. This measure of quality of a dataset is clearly a function of the purpose for which it is used. If the dataset is being used to analyze the long-term behavior of a temporal process, e.g., global warming, then it is important to have the measurements over a long period of time without significant gaps. Similarly if a process is very dynamic, e.g., streamflow of a river, then it is necessary to measure its parameters at a finer temporal scale.

This aspect of the datasets is often called its *Fitness of Use* (FoU). For a given application, this value varies among the datasets. Information derived from high-FoU datasets is more useful and accurate for the users of the application than that from low-FoU datasets. The challenge is to develop appropriate methods to evaluate this FoU measure for a dataset in order to gain information on how the dataset can be used or how appropriate the dataset is for a particular application [3]. This challenge and other aspects of goodness of datasets have been identified as reasons for failure in many data warehouse projects [4],[5].

In this paper, we focus on computing the FoU of dynamic datasets that are quite common and are widely used in many critical decision support systems. Our approach is to develop efficient methods to evaluate the FoU of time series data using rules derived for specific applications. While we use a specific application from the geospatial domain to illustrate our techniques, they are applicable in many different domains. Specific examples of these types of datasets include measurements of precipitation and other weather related parameters, streamflow, amount of water in lakes and reservoirs, crop yield, or virtually, any dataset that records measurement of the same quantity over a long period of time at regular intervals We use an information theoretic approach to compute the FoU of a dataset. The Dempster-Shafer belief theory [6] is used as the basis for our approach, in which the FoU is represented as a range of possibilities and integrated into one value based on the information from multiple sources. Dempster-Shafer is the primary alternative to Bayes theory [7], [8], which requires previously known probability determinations for data fusion.

The paper is organized as follows. In Section 2, we review the related research work in computing fitness of use. In Section 3, we define and formulate the problem of computing the FoU for time series data. We describe the Dempster-Shafer belief theory and how it applies to our problem. Three solution approaches-heuristic analysis, temporal variability analysis, and time series analysis-and algorithms for computing the FoU are described in Section 4. In Section 5, we describe the implementation and results of using the FoU computation in a geospatial application domain. Finally, we give a summary and describe directions for future work in Section 6.

2 Related Work

The focus of this paper is not on the computation of traditional data quality, but on their fitness of specific applications. Specifically, we seek to incorporate the FoU in knowledge

discovery and data mining (KDD) approaches and apply this to geospatial datasets. In this section, we first differentiate between traditional data quality and the FoU measure. Then, we discuss approaches to computing the FoU in geospatial time series datasets.

Though related, traditional data quality and the FoU are different. Previous studies addressed data quality as a multi-dimensional concept [9]-[13]. Data quality is commonly evaluated along five major dimensions [3]: (1) accuracy that measures the conformity of the recorded value with the actual value, (2) precision which is the degree of detail that can be recorded, (3) resolution which is the level of detail that can be represented, (4) consistency which means that the representation of the data is in the same format, and (5) completeness which implies that all values for a certain variable are recorded. Furthermore, data quality may also include data accessibility, appropriate amount of data, data believability, data completeness, data representation, ease of manipulation, data interpretability, data reputation, data security, and data timeliness [11]. In some GIS-based analysis, the quality of geospatial data has been evaluated with uncertainty as the basis of quality [14]. Data quality is therefore expressed as the degree of discrepancy between the data in GIS and the data in the geographic reality. In [15],[16], the quality of geospatial data was investigated by detecting the errors and inconsistencies in spatial datasets.

On the other hand, the definition of the FoU measures the "quality" of data from the viewpoint of its suitability for a specific application. For example, a value with precision of three decimal points is more accurate than the same value rounded to two decimal points. However, both values may have the exactly the same FoU in an application, in which the values are used as integers. In [17], the FoU was derived from the aspects of metadata quality, data availability, data resolution, data collection methods, data classification methods, and cost of accessing data. In [18], the authors studied the FoU obtained from multiple sources with different levels of reliability. Compared to these studies, our work is focused on metadata quality and considers a single data source, e.g., streamflow measurements.

Considering the complex nature of geospatial time series datasets and the limitation of methods used to capture the data, different models have been proposed to model errors and inconsistencies, on which the FoU is based. In [19],[20], the probability theory was used to evaluate the uncertainty in the capture of geospatial data. The probability theory is especially powerful when dealing with uncertainty that is introduced by randomness (e.g., digitization of a point using GIS [14]). However, the appropriateness of probability theory becomes questionable when dealing with the uncertainty that is caused by fuzziness [21] (e.g., the creation of a map using remote sensing classification techniques [14]). The possibility theory, on the other hand, shows efficiency in calculation of uncertainty that is introduced due to fuzziness but is less efficient in dealing with uncertainty caused by randomness. In [22], the possibility theory was used in FoU evaluation. In [23], the authors studied the uncertainty using various mathematical models, e.g., normal or Gaussian distributions. Such uncertainty analysis is not used in our study because we focus only on the data but not the quality of the instruments used to capture data.

KDD techniques are designed to extract interesting information and knowledge from vast amounts of data and can play a major role in affecting the FoU of the data. Steps in a typical KDD application are: (1) data gathering, (2) data storage, (3) data retrieval, (4) data mining, and (5) knowledge delivery [24]. During the data gathering stage, the FoU can be affected by the manual entry (e.g., duplicate data entry, error of measurement equipment). At the data storage stage, inappropriate data model or structure and untimely updates may lower the FoU. At the data retrieval stage, the problems include computational constraints and inefficient use of memory. Data mining can be used to compute the FoU [25],[26]. For

example, mining of association rules were employed to detect, quantify, explain and correct data quality deficiencies in databases. In [27], a machine learning algorithm, namely C4.5, was used to design a data auditing generator. The data auditing generator can systematically generate and pollute an artificial benchmark that is used to audit a database. Statistical methodologies (e.g., standard statistical tests) for detecting abnormal or missing data and profiling the quality of a dataset have been developed [28],[29]. In [30], a module was developed to capture data quality problems in database systems by performing data extraction, data transformation and data loading. In [31], a data quality model was developed to improve the query performance of database management systems.

To summarize, the focus of our work is in computing the FoU of a dataset and not the traditional aspect of data quality. Many previous studies have investigated geospatial data by only considering a single dimension, e.g., scale incompatibility or error detection, or by analyzing multiple dimensions separately. Our study evaluates the fitness of use as a whole by making use of an information fusion approach to combine the impact from all the related data aspects/attributes into a single measurement. For example, data gap (interruptions in measurements) is an important data attribute in FoU evaluation, which was considered in [32] and [33]. However, those studies did not provide a systematic approach for combining data gap with other related factors (as suggested by [34]). Specifically, our research investigates the FoU by incorporating the impact from data gap, data stability factors, and other data attributes. This is particularly important for analyzing geospatial time series data, where gaps in measurement may be a significant issue. Our proposed approach is based on the Dempster-Shafer belief theory which has been used by Eastman for uncertainty management for decision support tools [35].

3 Problem Formulation

The central problem we address in this paper is to determine the FoU of time series datasets for applications. Assume that we are given a set of time series datasets, $S = \{S_1, S_2,...,S_n\}$. A dataset S_i may consist of many types of information including (and not limited to) spatial coordinates, metadata about the dataset, denoted by aux_i, and the actual time series data, denoted by tS_i.

The metadata for a dataset includes the type of information being recorded (e.g., precipitation, or discharge of water in a stream), the period of record, and the frequency of measurement. Thus,

$$\text{aux}_i = \langle \text{type}_i, \text{tb}_i, \text{te}_i, \text{int}_i \rangle\rangle,$$

where tb_i and te_i denote the beginning and the ending time stamps for the measurements, and int_i is the interval at which the measurements are made. Other metadata such as the type and age of recording device can also be added.

The time series data in a dataset consist of a sequence of measurements,

$$tS_i = \langle m_{i,1}, m_{i,2}, \ldots, m_{i,p} \rangle.$$

Each measurement stores both the time the measurement was taken and the actual value recorded by the sensor. Thus, each measurement is given by:

$$m_{i,j} = \langle t_{i,j}; v_{i,j} \rangle.$$

We assume that the measurements in the dataset are maintained in chronological order. Therefore,

$$t_{i,j} < t_{i,k} \text{ for } j < k.$$

Furthermore, the first and last measurement times should match the period of record stored in the metadata,

$$\text{tb}_i = t_{i,1} \text{ and } \text{te}_i = t_{i,p}.$$

For our research, we consider datasets that store measurements at regular intervals. In general, different datasets record values at different intervals. For example, some datasets will be based on daily observations, while others at weekly or monthly intervals. The work reported here can be extended to the case when the measurements are taken at irregular intervals, but this is beyond the scope of this paper.

The problem of finding the suitability of a dataset for a given application is to define a function for the FoU that computes the fitness of use of a dataset described above. The function FoU maps a time series dataset S_i to a normalized value between 0 and 1:

$$\text{FoU}(S_i, A) = [0,1],$$

where S_i is a single dataset and A is the intended application of the data. The application A is represented in the form of domain knowledge that describes how goodness of a dataset is viewed. We use a set of rules that specify this information. Thus,

$$A = \{R_1, R_2, \ldots, R_d\},$$

where R_i is a domain rule to describe the goodness of a dataset, and d is the number of rules. Therefore, our *FoU* function is defined with respect to an application domain. Different applications can use different rules for goodness and derive different FoU values for the same dataset.

4 Approach

The challenge in computation of the fitness of use for datasets is to identify the dimensions that are important in the application, to formalize the domain rules, and finally to combine them to derive an overall quality indicator. The important dimensions must be representative and collectively be able to define the expected variance in the datasets. The fundamental issue is the fact that the natural processes are inherently variable and hence their measurements are uncertain. Furthermore, the FoU of a dataset can be measured along different dimensions. For example, if the data is recorded at a daily interval, the resulting dataset can be viewed to have higher quality than one that is recorded weekly. The reason for this assessment is that more precise analysis can be obtained from daily data than weekly data. One can computationally derive weekly data from daily data,. However, this can lead to bias and trends that do not accurately reflect reality. Therefore, there is a loss of information and hence it is not possible, in general, to derive daily data from weekly data. Similarly, if there are missing records in the dataset, then the length of the period for which there are no measurements is a (negative) factor in the quality of the dataset.

In this paper, we propose three different approaches to compute the fitness of use of datasets. They differ in the way in which the quality is modeled and quantified. In all three approaches however, we use the Dempster-Shafer belief theory [6] as the basis to derive a

composite quality indicator. The three proposed approaches are heuristic analysis, temporal variability analysis, and predicted error estimates using time series analysis. Briefly, the heuristic analysis uses commonsense heuristics (rules of thumb) taking into account consistency, length, recency, resolution (e.g., spatial or temporal), completeness, and noise to evaluate the FoU of datasets. The temporal variability analysis considers a coefficient of variation to measure the stability of values at the same periodic time points over an interval of measurements, with the assumption that high stability implies high FoU. The time series approach uses a regression model to estimate the maximum predicted error, which is then used as an inversely proportional estimate of the FoU of datasets.

The general application on which the FoU is defined is decision support systems to providing information to users on the "goodness" of the measurement stations. Information on the goodness of the measurement stations allow users to, for example, decide how to add new stations or make use of existing stations to extrapolate values for regions not covered by the stations. This general application can thus be plugged into most GIS-based applications (e.g., hydrological analysis and drought assessment).

In the following subsections, we first briefly describe the Dempster-Shafer belief theory and how it can be used to compute the FoU of datasets. Then we discuss the three proposed approaches.

4.1 Dempster-Shafer belief theory

There are several limitations of applying the traditional Bayes probability theory to represent the full scope of uncertainty [36]. For example, the probability of an outcome occurring is measured based on how often that particular outcome occurs in a series of trials with respect to other outcomes in the long run. However, in most uncertain scenarios, the probability of an outcome is not known as (1) we do not have the frame of all possible outcomes, especially in a non-discrete domain, and (2) we do not have enough data to ascertain the frequency of an outcome. Thus, uncertainty is often represented with subjective probability. The Dempster-Shafer belief theory addresses this limitation. Instead of a single probability value associated with an event or measurement, the theory allows an interval-valued probability to be associated with the event. There are several advantages of the Dempster-Shafer belief theory. First, it does not require that the individual elements follow a certain probability. In other words, Bayes' theorem considers an event to be either true or untrue, whereas Dempster-Shafer allows for unknown states [37]. This characteristic makes Dempster-Shafer belief theory a powerful tool for the evaluation of risk and reliability in many real applications when it is impossible to obtain precise measurements/results from real experiments. In addition, Dempster-Shafer Belief Theory provides a framework to combine the evidence from multiple sources and does not assume disjoint outcomes [31]. Also, Dempster-Shafer's measures are not necessarily less accurate than Bayesian methods, and in fact reports have shown that it can sometimes outperform Bayes' theorem [38], [39].

The two central ideas of Dempster-Shafer belief theory are: (a) obtaining degrees of belief from subjective probabilities for a related question, and (b) Dempster's rule for combining such degrees of belief when they are based on independent items of evidence. For a given proposition, *P* and given some evidence, we derive a confidence interval, defined by an interval of probabilities within which the true probability lies within a certain confidence. This interval is defined by the *belief* and *plausibility* supported by the evidence for the given proposition. The lower bound of the interval is called the *belief* and measures the strength of the evidence in favor of a proposition. The upper bound of the interval is called the *plausibility*. It brings together the evidence that is compatible with the proposition and is

not inconsistent with it. The values of both belief and plausibility range from 0 to 1. The belief function (bel) and the plausibility function (pl) are related by:

$$\text{pl}(P) = 1 - \text{bel}(\bar{P})$$

where $\bar{P}$ is the negation of the proposition P. Thus, bel ($\bar{P}$) is the extent to which evidence is in favor of $\bar{P}$. It should also be noted that it is not necessary that pl(P) + pl($\bar{P}$) = 1. Likewise, it is not necessary that bel (P) + bel ($\bar{P}$) = 1.

The term *Frame of Discernment* (FOD) consists of all hypotheses for which the information sources can provide evidence. This set is finite and consists of mutually exclusive propositions that span the hypotheses space. For a finite set of mutually exclusive propositions (θ) the set of possible hypotheses is its power set (2^{θ}), i.e., the set of all possible subsets including itself and a null set. Each of these subsets is called a focal element and is assigned a confidence interval [belief, plausibility].

Based on the evidence, we first assign a probability mass to each focal element. The masses are *probability-like* in that they are in the range [0, 1] and sum to 1 over all hypotheses. However, they represent the belief assigned to a focal element. In most cases, this basic probability assignment is derived from the experience and rules provided by some experts in the application domain.

Given a hypothesis, H, its belief is computed as the sum of all the probability masses of the subsets of H as follows:

$$\text{bel}(H) = \sum_{e \subset H} m(e),$$

where $m(e)$ is the probability mass assigned to the subset e. The probability mass function distributes the values on subsets of the frame of discernment. Only to those hypotheses, for which it has direct evidence, are assigned non-zero values. Thus, the Dempster-Shafer belief theory allows for having a single piece of evidence supporting a set of multiple propositions being true.

If there are multiple sources of information, we can derive probability mass functions for each data source. These mass values are then combined using Dempster's Combination Rule to derive joint evidence to support a hypothesis from multiple sources. Given two basic probability assignments, m_A and m_B for two independent sources (A and B) of evidence in the same frame of discernment, we can compute the joint probability mass, m_{AB}, according to Dempster's Combination Rule:

$$m_{AB}(C) = \frac{\sum_{A \cap B = C} m(A)*m(B)}{1 - \sum_{A \cap B = \varnothing} m(A)*m(B)}$$

Furthermore, the rule can be repeatedly applied for more than two sources sequentially, and the results are order-independent. That is, combining different pieces of evidence in different sequences yields the same results.

Finally, to determine the confidence in a hypothesis H being true, we multiply belief by plausibility:

$$\text{confidence}(H) = \text{bel}(H) \cdot \text{pl}(H)$$

Thus, the system is highly confident about a hypothesis being true if it has high belief and plausibility for that hypothesis being true.

For all of our approaches, we use three discrete FoU outcomes of the datasets: suitable (s), marginal (m), and unsuitable (u). Thus,

$$\theta = \{s,m,u\}$$

and the frame of discernment is:

$$FOD = 2^{\theta} = \{\varnothing, \{s\}, \{m\}, \{u\}, \{s,m\}, \{s,u\}, \{m,u\}, \{s,m,u\}\}.$$

4.2 Heuristic Analysis

As described before, the assignment of probability masses can be done in different ways. In our first approach, we use a set of domain heuristics for this purpose and then use the combination rule to compute the fitness of use of the datasets. The heuristics can be based on common sense knowledge or can be based on expert feedback. We use the following criteria to judge the FoU of geospatial time series datasets for general GIS applications.

Consistency—A dataset is consistent if it does not have any gaps. A consistent dataset has a higher fitness value.

Length—The period of record for the dataset is also an important factor in the quality. Longer periods of record generally imply higher fitness value.

Recency—Datasets that record more recent observations are considered to be of higher fitness value.

Temporal Resolution—Data are recorded at different time scales (sampling periods). For example, the datasets can be recorded daily, weekly or monthly. Sometimes data are recorded daily, but summarized weekly or monthly. Depending on the application higher or lower resolution may be better. The higher resolution had more information, but is noisier as well. This is also called the granularity [40].

Completeness—A data record may have many attributes, e.g., time, location, and one or more observations. For example, a weather station may record the daily high and low temperature, daily precipitation, moisture level, etc. A dataset is complete if all the relevant attributes are recorded. Incomplete datasets are considered to be inferior [40].

Noise—All datasets have some noise due to many different factors. They include severe weather conditions during data collection, human factors, and inaccuracy of measurement devices. All these factors may lead to lower FoU values for a dataset.

For each of the above criteria, we can define one or more heuristics to determine the subjective probability mass for different data quality values. We specify the heuristics in the form of rules as follows:

$$C_1(S_i) \wedge C_2(S_i) \wedge \ldots \wedge C_n(S_i) \rightarrow \text{mass}(S_i, \{\text{qtype}\}) = m$$

where C_i specifies a condition of the dataset, C_j (S_i) evaluates to true if the condition C_j holds for the dataset S_i, and mass(S_i, {qtype}) denotes the mass of evidence that the dataset S_i contributes to the FoU outcome types in {qtype}. The symbol $\wedge$ is used for logical conjunction (and) and the symbol $\rightarrow$ is used for logical implication (implies). We say that the rule is triggered or fires [41] if all the conditions are met. When the rule fires, we evaluate the right-hand side of the rule, which assigns a value m to the probability mass for a given set of outcome types, which in this case can be described in the following formulation: {qtype} $\subseteq$ {suitable, marginal, unsuitable}.

Applying a set of rules as defined above to dataset S_i thus yields a set of masses for different combinations of outcome types. These masses are then combined using the Dempster's Combination Rule, as discussed in Section 4.1, to yield a coherent set of masses for each element of FOD. We then further reduce the result by considering only the singletons: {suitable}, {marginal}, and {unsuitable}, which allows us compute the belief, plausibility, and confidence values on only these three outcome types.

It should be noted that the criteria and the corresponding rules should not be treated as universal. For example, if the historical context is of great importance in an application, the recency of the dataset will be of less significance than its "oldness." The rules can then be adjusted accordingly.

4.3 Temporal variability analysis

In addition to the generally recognized factors that affect the FoU, as described in the previous section, we note that the stability of the datasets is also important. Many natural processes follow periodic patterns, the periodicity guided by factors including diurnal, seasonal or annual fluctuations. Other than exceptional events (e.g., severe floods), measurements at the same point time in a cycle should be similar (i.e., without much variation). An unstable dataset with random variations will generally not be useful in developing predictive models. To illustrate, Table 1 shows the measurements of two stream gauges during the same period on the same date of the year. Station No. 1 (or dataset S_i) has stable measurements while the second station shows dramatic changes. Thus it is more difficult to derive a concrete pattern from the measurements for Station No. 2 (or dataset S_j). We deem the FoU of S_j to be lower than that of S_i.

To capture the above patterns of goodness, we use the temporal variability of the datasets. Suppose that a geospatial time series S_i has the following measurements, as previously defined in Section 3:

$$tS_i = \langle m_{i,1}, m_{i,2}, \ldots, m_{i,p} \rangle, \text{ and } m_{i,j} = \langle t_{i,j}, v_{i,j} \rangle$$

Suppose that the measurements are collected periodically at a regular interval. For example, measurements can be collected three times a day, at 0900 h, 1200 h, and 1600 h, for 31 days between 2006-01-01 and 2006-01-31; or collected once a day, for 365 days, for 10 years between 1991 and 2001. The period of record for the first dataset is 31 days while it is 10 years for the second dataset. Suppose that we define the *periodicity* of a data series as the time between two measurements collected at the same spatial location at the same time mark. Thus, in our examples, the periodicity of the first data set is 1 day, while the period of the second data set is 365 days. Given this notion of periodicity, we can compute the average value of all measurements at each particular time mark over the entire period of record. In this case, we can compute the average of 31 measurements at 0900 h, the average of 31 measurements

Table 1. Parts of records of two stream flow gauges (i.e., of two datasets, S_i and S_j) showing different patterns

Time stamp	Station no. 1 (S_i)	Station no. 2 (S_j)
1978-11-01	5.0	22.0
1979-11-01	5.0	25.0
1980-11-01	5.3	85.0
1981-11-01	5.2	70.0
1982-11-01	5.1	75.0

Datasets are obtained from USGS.

at 1200 h and 31 measurements at 1600 h for the first data set. Similarly we can compute the average of ten measurements for each day of the week for the second data set. Formally, $ts_i = \langle m_{i,1}, m_{i,2}, \ldots, m_{i,\mathrm{p}} \rangle$ can be re-written as:

$$tS_i = \langle m_{i,1}, m_{i,2}, \ldots, m_{i,\mathrm{period}}, m_{i,\mathrm{period}+1}, m_{i,\mathrm{period}+2}, \ldots, m_{i,2*\mathrm{period}}, m_{i,2*\mathrm{period}+1}, \ldots, m_{i,k*\mathrm{period}} \rangle$$

such that $t_{i,k*period} - t_{i,1} = \mathrm{int}_i$.Given the above representation, we can derive the periodic mean at each time mark j as:

$$\mathrm{mean}_{i,j} = \frac{\sum_{p=0}^{k} m_{i,p*period+j}}{k}$$

Likewise, we can derive the periodic variance for the time marks j as:

$$\mathrm{var}_{i,j} = \frac{k \sum_{p=0}^{k} m^2_{i,p*period+j} - \left(\sum_{p=0}^{k} m_{i,p*period+j} \right)^2}{k\,(k-1)}$$

Given the set of means and variances for all time marks in a period, we can further compute the coefficient of variation at each time mark j:

$$\mathrm{cov}_{i,j} = \frac{\sqrt{\mathrm{var}_{i,j}}}{\mathrm{mean}_{i,j}}$$

The temporal variability of the dataset S_i can then be defined as the average value of coefficient of variation for all time marks:

$$\bar{c}(S_i) = \frac{\sum_{j=1}^{\mathrm{period}} \mathrm{cov}_{i,j}}{\mathrm{period}}$$

We can then use heuristics to assign probability masses to the different outcomes based on the value of $\bar{c}$. For example, to assign probability masses to the outcomes, we divide the temporal variability into three ranges: the upper (largest) one-third, the middle one-third and the lower (smallest) one-third. For each range, we define one or more heuristics to determine the probability mass for different FoU values. The heuristics are specified in the form of rules as follows:

$$(\bar{c}(S_i) \text{within range } k \rightarrow \text{mass } (S_i, \{\text{qtype}\}) = m$$

where $\bar{c}(S_i)$ is the average coefficient of variation of the dataset S_i, and the range k is one of the three ranges mentioned above. For a given dataset S_i, we evaluate the right hand side of the above rule and assign a value m to the probability mass for a given type (suitable, marginal, or unsuitable). We can also combine these probability masses with those described in Section 4.2 using Dempster's Combination Rule.

4.4 Time series analysis (predicted error estimate)

Another approach to determine the goodness of a data set is to apply a regression model to the data and estimate the maximum predicted error. The maximum error will be inversely proportional to the fitness of the dataset for use in our application. Ordinary regression analysis assumes that the error variance is the same for all observations. When the er-

ror variance is not constant, the data are said to be *heteroscedastic*, and ordinary least-square estimate methods are inaccurate [42]. More efficient use of the data and more accurate prediction error estimates can be made by models that take the heteroscedasticity into account. Time series data are often autocorrelated and hence are *heteroscedastic*. It is particularly so in hydrological datasets which are strongly correlated to seasonal trends. In such cases, generalized autoregressive conditional heteroscedasticity models are more appropriate since they correct for serial correlation.

Autoregressive models augment the regression model with an autoregressive model for the random error to account for the autocorrelation of the errors. The model is given by:

$$y_t = \beta_0 + x_t\beta_1 + \nu_t,$$

where x and y are the independent (regressor) and dependent (response) variables and are functions of time, β_0 and β_1 are the intercept and the slope in a linear model, and ν_t is the autocorrelated error given by:

$$\nu_t = -\varphi_i\nu_{t-1} - -\varphi_2\nu_{t-2} - \ldots - -\varphi_m\nu_{t-m} + \varepsilon_t$$

where φ_i 's are the model parameters and ε_t is the error estimate that is independent and is normally distributed with a zero mean.

By simultaneously estimating the regression coefficients β_0 and β_1 and the autoregressive error model parameters φ_i 's, we can predict the upper and lower confidence limits with a fixed probability value (say 0.98).

This approach can be used to estimate the maximum error value for gaps in the time series. Typically, a time series consists of a set of observations made at a succession of equally spaced points in time. However, in practical applications, some measurements may not be recorded due to various reasons, such as equipment malfunction or human errors. The missing data records or gaps impact the quality of the dataset. The degree of the impact is dependent on the nature of the datasets; some gaps may heavily impact the quality of dataset while some other gaps may have minor or even no impact on the quality of dataset, if we can accurately predict the values at the gaps. To illustrate, Table 2 shows two datasets with the same gap (the gap period from 1957-10-03 to 1957-11-04). In the case of the first dataset, the gap has little impact. It is very likely that the stream is dry during the gap period if it was dry both before and after the gap period. On the other hand it is not clear what the value should be during the gap period for the second dataset since stream volumes before and after the gap period are not consistent.

We use the upper and lower confidence limits of the predicted values using an autoregressive model to guide the quality of the dataset. Figure 1 shows an example of the confi-

Table 2. Parts of records of two stream flow gages (i.e., two datasets, S_i and S_j) showing different patterns

Time stamp	Station no.1 (S_i) 06875500	Station no. 2 (S_j) 06453600
1957-10-01	0.0	122.0
1957-10-02	0.0	125.0
1957-11-05	0.0	85.0
1957-11-06	0.0	70.0
1957-11-07	0.0	75.0

Datasets are obtained from USGS

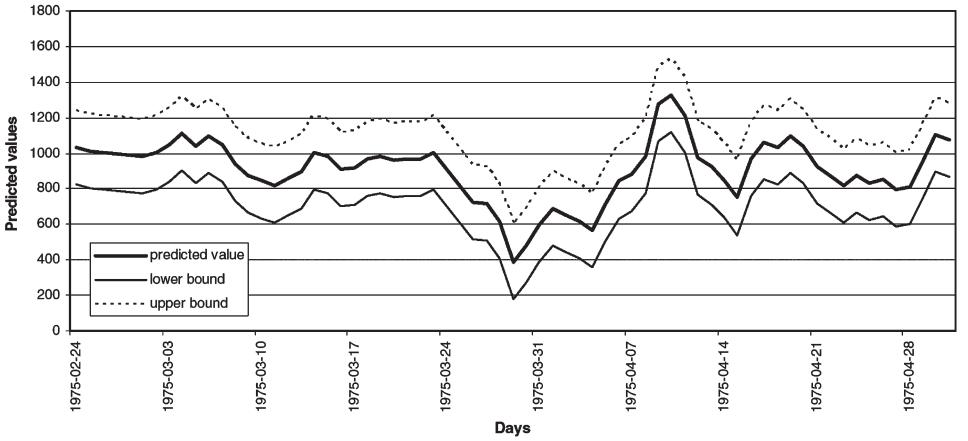

Figure 1. The predicted value and the lower and upper bounds on it using the autoregressive model. The dataset corresponds to the USGS stream gauge 06462000

dence limits using autoregressive analysis. The blue line represents the predicted values of the corresponding dataset. The upper bound shows the maximum values the prediction can reach whereas the lower bound shows the minimum values the prediction can reach.

We use the maximum range of the confidence interval during the gaps as a guide to the FoU the dataset. We then use heuristics to assign probability masses to the different outcomes based on the value of confidence interval. To assign probability masses to the outcomes, we divide the confidence interval into three ranges: the upper (largest) one-third, the middle one-third and the lower (smallest) one-third. For each of the three ranges, we define one or more heuristics to determine the probability mass for different FoU values. The heuristics are specified in the form of rules as follows:

$$(\text{confidence\_interval } (S_i)\text{within range } r) \rightarrow \text{mass}(S_i, \{\text{qtype}\}) = m$$

where confidence_interval(S_i) is the maximum range of the confidence interval computed using the predicted error estimate defined above of a given dataset S_i, and the range r is one the three ranges mentioned above. For a given dataset, S_i, we evaluate the right hand side of the above rule and assign a value m to the probability mass for a given type (suitable, marginal, or unsuitable). Using Dempster's Combination Rule, we can also combine these probability masses with the other heuristics (described in Section 4.2 and 4.3) to get an overall estimate of the FoU.

5 Implementation and results

5.1 Application domain

The focus of this paper is to describe an approach that can be used to compute the fitness of use of time series data for geospatial applications. As described in Section 4, FoU is a function of a dataset and a given application. To illustrate the use of the approach, we choose the hydrological analysis as our application domain. The water cycle is a complex process that includes precipitation, surface water runoff, groundwater flow, and evapo-

transpiration (directly from the surface of the water or indirectly via plants). In order to understand each component of the water cycle, an elaborate sensor network has been put into place. For example, the National Weather Service has been collecting data about precipitation at various locations for over a hundred years. The US Geological Survey (USGS) maintains thousands of stream gauges to measure the flow of water in streams. State and local agencies also maintain a large network of groundwater monitoring wells to measure the water level at a variety of locations. For example, there are 273 stream gauges, 251 weather stations and 4,160 groundwater monitoring wells in the state of Nebraska, many of which have long periods of record. To illustrate our approach, we will apply the Dempster-Shafer belief theory to determine the FoU of the datasets from the 274 stream gages.

Hydrological data processes, while being periodic, are inherently difficult to model due to noise and multiple contributing factors. Local hydrological processes, for example, are affected by the global climate cycles. In some datasets the recorded measurements are estimates. For example, during winter months, the instruments in streams may be inoperable because of ice. Some components of the water cycle are measured at regular intervals while others are measured intermittently. Some stations record data every hour, others once a month and still others a few times a year. Furthermore, water-related monitoring stations are distributed in space. The overlap between different networks is inconsistent both spatially and temporally. While at any given point in space and time, there may be only a small number of measurements (if any), there are sufficient measurements in space to obtain an overall picture of the state of the hydrological resources.

The quality of individual monitoring stations is different and can be viewed as inversely proportional to the noise, incompleteness, irregularity, and inaccuracy found in a dataset. For example, the FoU from a stream gauge that records daily streamflow with a 50-year record without any gaps will be deemed higher than a stream gauge that has only 30 years of weekly measurements with a 5-year gap in the middle.

5.2 Geographic Scope

In our study, we consider the datasets that store surface water measurements in the state of Nebraska [43]. There are a total of 273 surface water measurement sites in the state, many with records that go back to over 50 years. The measurements are recorded daily. Many stations have gaps in their record reducing the quality of the data. The datasets are obtained from US Geological Service (http://water.usgs.gov). Each dataset consists of series of records that consists of the measuring station ID, geographic coordinates of the station, the date of measurement and the streamflow measurement in cubic feet per second. Table 3 shows part of a sample dataset.

5.3 FoU using Heuristic Approach

Based on the criteria defined in Section 4.2, we have designed two rulebases (i.e., Rulebase 1 and Rulebase 2) to compute the quality of the time series datasets. In the two rulebases, the criteria length and consistency are measured with different temporal resolutions. In Rulebase 1 we use a yearly time scale in the specification of rules. In contrast, the rules in Rulebase 2 use a monthly time scale. The temporal resolution affects the data consistency since it is determined by the number of gaps in the data record. In Rulebase 1, the gaps are coarser (i.e., yearly), while they are finer (i.e., monthly) in Rulebase 2. There are a total of 21 rules in Rulebase 1 and 30 rules in Rulebase 2. Sample rules that match the records for Station 6445500 (period of record from 1936-10-1 to 1944-1-31 with no gap) are given below:

$$[\text{surfacewater}(S_i) \wedge \text{daily}(S_i) \wedge \neg\text{ten\_year\_old}(S_i) \wedge \neg\text{current}(S_i) \wedge 1940\text{s}(S_i) \wedge \neg\text{gap}(S_i)] \rightarrow \text{mass}(S_i, \{\text{suitable}\}) = 0.380$$

$$\left[\begin{array}{l}\text{surfacewater}(S_i) \wedge \text{daily}(S_i) \wedge \neg\text{ten\_year\_old}(S_i) \wedge \neg\text{current}(S_i) \wedge 1940\text{s}(S_i) \wedge \neg 1950\text{s}(S_i) \\ \wedge\ [\text{record\_length}(S_i) < 500] \wedge \neg\text{gap}(S_i)\end{array}\right] \rightarrow \text{mass}(S_i, \{\text{unsuitable}\}) = 0.770$$

The first rule states that if the station (1) is a surface water station, (2) records measurements daily with an interval of record less than 10 years, (3) is not current, (4) has data period overlapping 1940s, and (5) has no gap, then the assertion that "the FoU recorded for the station will be *suitable*" will have a probability mass of 0.38, a rather low value. Similarly, the second rule assigns the probability mass of the assertion that the "FoU for the station is *unsuitable*." Thus for the Station 6445500, we will have the following probability mass assignments.

$$\text{mass}(6445500, \{\text{suitable}\}) = 0.380$$
$$\text{mass}(6445500, \{\text{unsuitable}\}) = 0.770$$

After the individual masses are assigned, the overall quality of each dataset is obtained by combining them using Dempster's Combination Rule as explained in Section 4.1. Table 4 shows the FoU classification results for all the datasets using Rulebase 1 and Rulebase 2.

Results show that Rulebase 1 classifies 270 datasets as good and three as unsuitable while Rulebase 2 classifies 183 datasets as good and 85 datasets as unsuitable. We analyzed the results using common sense reasoning and input by domain experts and determined that the Rulebase 2 to be more reasonable. For example, Rulebase 1 classifies the Station 6445500 (illustrated in this section) as *suitable* while Rulebase 2 classifies it as *unsuitable*. As suggested by common sense knowledge and confirmed by domain experts, the FoU of station 6445500 should be considered *unsuitable* as suggested by Rulebase 2.

Table 3. Some measurement records of surface water Station 06455500

Source	ID	Latitude	Longitude	Date	Value (ft^3/s)	Type
USGS	06455500	42°27′23.47″	103°04′07.75″	1946-10-01	1.9	daily
USGS	06455500	42°27′23.47″	103°04′07.75″	1946-10-02	1.9	daily
USGS	06455500	42°27′23.47″	103°04′07.75″	1946-10-03	1.9	daily
USGS	06455500	42°27′23.47″	103°04′07.75″	1946-10-04	1.9	daily
USGS	06455500	42°27′23.47″	103°04′07.75″	1946-10-05	1.9	daily

Table 4. Summary of FoU computation using Rulebase 1 and Rulebase 2

		Rulebase 1	Rulebase 2
Fitness of use	Suitable	270	183
	Marginal	0	5
	Unsuitable	3	85

Table 5. FoU of some sample stations using Rulebase 2

ID	Recency		Granularity	Consistency (gap)	Fitness of use		
	Start date	End date			Suitable	Marginal	Unsuitable
6453500	1949-10-1	1994-9-30	Daily	yes (no gap)	**0.962**	0.000	0.000
6459200	1962-10-1	1981-9-30	Daily	yes (no gap)	0.101	**0.331**	0.014
6445500	1936-10-1	1944-1-31	Daily	yes (no gap)	0.015	0.000	**0.770**

Bold values correspond to the classifications of the stations.

Table 5 shows some sample stations classified using Rulebase 2. The FoU of Station 6445500 is considered unsuitable with a high degree of confidence (0.770) since it has low recency and a short record. The FoU of Station 6453500 is suitable with almost 100% confidence as it has a long record of measurements (about 45 years) and is quite recent (1994). The FoU of Station 6459200 is marginal with a confidence value of 0.331, and suitable with a confidence value of 0.101. Thus, the system decides that the station is marginal in its fitness value.

5.4 FoU based on temporal variability analysis

In Section 4.3, dataset stability is defined based on temporal variability of a dataset S_i using the average of coefficient of variance:

$$\overline{c}(S_i) = \frac{\sum_{j=1}^{\text{period}} \text{cov}_{i,j}}{\text{period}}$$

In our implementation, we set *period* to 365 days, with each time mark being the particular day of a year. The mean is then taken over k periods (or k years), where k is defined as in $t_{i,k*\text{period}} - t_{i,1} = \text{int}_i$, as presented in Section 4.3. Essentially, for this implementation, we compute the mean and variance of streamflow measurements for day 1 of a year, the mean and variance of streamflow measurements for day 2 of a year, and so on. We also compute the coefficient of variation for each day, and arrive at the average as above. Thus, in this implementation, c is the *average value of the daily variations*.

Figure 2 shows a distribution of surface water stations against the average daily variations. Based on the distribution, we divide the variations into three ranges: (a) upper one third ($c > 2$), (b) the middle one-third ($0{:}75 < c < 2$), and the lower (smallest) one-third

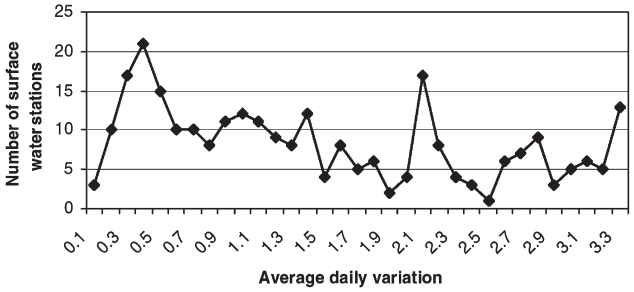

Figure 2. Distribution of surface 25 water stations against the average daily variations

Table 6. FoU of some sample stations using Rulebase 2 and temporal variability analysis

ID	Recency		Granularity	Length	$\bar{c}$	Fitness of use		
	Start date	End date				Suitable	Marginal	Unsuitable
6796973	1982-10-1	1992-09-30	daily	5,478	1.10	0.099	0.000	**0.496**
6459200	1962-10-1	1981-09-30	daily	6,940	0.11	**0.500**	0.061	0.003
6460900	1958-03-1	1974-10-02	daily	6,060	0.83	0.011	**0.207**	0.048

Bold values correspond to the classifications of the stations.

($0 < c < 0.75$). Of the 273 datasets, 91 datasets have c values lower than 0.75, 96 datasets have c values between 0.75 and 2, and 86 datasets have c values higher than 2.

Some sample rules to assign probability masses based on the above classification are given below:

$$\text{UpperThird}\ (S_i) \rightarrow \text{mass}\ (S_i, \{\text{unsuitable}\}) = 0.7$$
$$\text{MiddleThird}\ (S_i) \rightarrow \text{mass}\ (S_i, \{\text{marginal}\}) = 0.5$$
$$\text{LowerThird}\ (S_i) \rightarrow \text{mass}\ (S_i, \{\text{suitable}\}) = 0.7$$

The predicate *UpperThird* returns *true* if the average daily variation (c) belongs to the upper one-third range and returns *false* otherwise. Similarly, the predicates *MiddleThird* and *LowerThird* return *true* when the average daily variation (c) is in middle and lowest thirds, respectively. It should be noted that these assignments are determined experimentally and can be adjusted based on domain expertise.

We then use Dempster-Shafer belief theory to combine the probability masses, which are assigned according to the average daily variations (c), with the rules included in the heuristic approach. Sample rules used for assignment of probability masses in temporal variability analysis have been described above. The following example illustrates how we combine the probability mass assignment rules in the heuristic approach with those in the temporal variability analysis. In the example, the station 6454100 has the record period from 1957-10-1 to 1991-9-30 (daily record); the record length is 12,418 days; no gap period; and the c value is 0.35. Based on the rules defined in Section 4.2, using Dempster's Combination Rule, we assign the probability mass to the station as follows:

$$\begin{bmatrix} \text{surfacewater}\ (S_i) \wedge \text{daily}\ (S_i) \wedge \text{twentyplus\_years\_old}\ (S_i) \wedge \neg\text{current}\ (S_i) \wedge \text{1990s}\ (S_i) \wedge \neg\text{gap}\ (S_i) \\ \wedge\ [10{,}000 < \text{record\_length}\ (S_i) < 15{,}000] \wedge \text{LowerThird}\ (S_i) \end{bmatrix}$$
$$\rightarrow \text{mass}\ (S_i, \{\text{suitable}\}) = 0.986$$

Table 6 shows the FoU for some sample stations using this approach. For example, station 6796973 has a length of record of 5,478 and is the most recent among the three datasets. However, its average daily variation is extremely high and results in its FoU being assigned *unsuitable*. Station 6460900's average daily variation is moderately high, and even though it

Table 7. Summary of FoU computation using heuristic and temporal variability study

		Heuristic	Temporal Variability
Fitness of use	Suitable	183	184
	Marginal	5	3
	Unsuitable	85	86

Table 8. Sample stations that change classes after temporal variability analysis is added to the system with Rulebase 2

ID	Recency		Consistency (gap)	Length	$\bar{c}$	Fitness of use					
	Start date	End date				Suitable		Marginal		Unsuitable	
						Old	New	Old	New	Old	New
6796973	1982-10-1	1992-9-30	no gap	5,478	1.10	**0.324**	0.099	0.003	0.000	0.137	**0.496**
6459200	1962-10-1	1981-9-30	no gap	6,940	0.11	0.101	**0.500**	**0.331**	0.061	0.014	0.003
6803170	1989-8-16	2001-3-30	has gap	4,398	0.85	**0.308**	0.190	0.003	0.001	0.158	**0.276**

has a suitable length of record, it is considered to have *marginal* FoU. The confidence in the classification is also weak (0.207) as the system considers it to be possibly *suitable* (0.114). Comparing Stations 6796973 and 6460900, we see that the system tends towards the length of records and the average daily variation.

Table 7 compares the results using Heuristic Analysis and the results from Temporal Variability Analysis. We see that with the inclusion of the temporal variability analysis, the number of suitable stations increases from 183 to 184, the number of unsuitable stations increases from 85 to 86, and the number of average datasets decreases from 5 to 3.

Table 8 shows the old and new confidence values for three stations whose classifications changed. The FoU for Station 6796973 changes from *suitable* to *unsuitable* since it has a large variability. The confidence in classification is also higher, from 0.342 to 0.469. This clearly demonstrates the impact of the temporal variability analysis on the classification. Station 6459200, on the other hand, benefits from a low temporal variability and has its FoU changed from *unsuitable* to *marginal*. Station 6803170 also changes from *suitable* to *unsuitable* and compared to Station 6796973, it has a relatively low confidence value 0.276 in the classification.

5.5 FoU based on time series analysis

In time series analysis, we explicitly model the impact of gaps on the fitness of use. As described in Section 4.4, we use the *confidence interval* of the gap to determine the data quality. If it is small, it means that the error in the predicted value for a measurement in the gap is low. Consequently, the gap has only a minor impact on the FoU. On the other hand, if the confidence interval is large, then the error in the predicted value will be high and the gaps will have a significant impact on the FoU. In our dataset, we had 25 stations that have

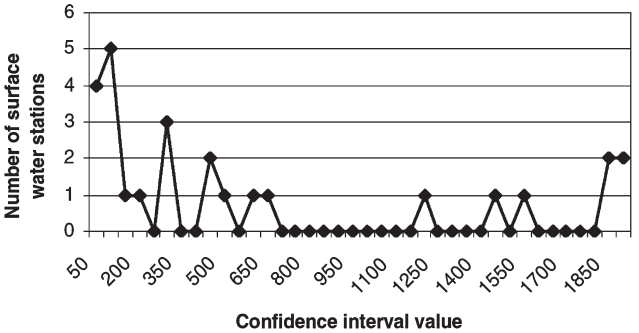

Figure 3. Distribution of surface 6 water stations against the confidence interval values

Table 9. FoU of some sample stations using Rulebase 2 and time series analysis

ID	Recency		Record length	Confidence interval	Fitness of use		
	Start date	End date			Suitable	Marginal	Unsuitable
6770478	1983-10-1	1989-6-27	2,045	1,800	0.000	0.000	**0.959**
6804900	1990-8-28	2001-9-30	4,051	33	**0.251**	0.001	0.182
6811000	1954-10-1	1969-9-30	5,499	4	0.031	**0.410**	0.040

Bold values correspond to the classifications of the stations.

gaps of various sizes. Figure 3 shows a distribution of the confidence intervals. Based on the distribution, we divide the confidence interval into three ranges: (a) large (*confidence interval* >1,000), (b) medium (250 < *confidence interval* ≤ 1,000) and (c) small (*confidence interval* ≤ 250). Using this scheme, 7 stations have *large* confidence interval, 7 have *medium* confidence interval and 11 have *small* confidence interval.

Sample rules to assign probability masses are given below:

$$\text{Large ConfidenceInterval } (S_i) \rightarrow \text{mass } (S_i, \{\text{unsuitable}\}) = 0.7$$
$$\text{Medium ConfidenceInterval } (S_i) \rightarrow \text{mass } (S_i, \{\text{marginal}\}) = 0.5$$
$$\text{Small ConfidenceInterval } (S_i) \rightarrow \text{mass } (S_i, \{\text{suitable}\}) = 0.7$$

The predicates *LargeConfidenceInterval*, *MediumConfidenceInterval*, and *SmallConfidenceInterval* have obvious meanings. These assignments are determined experimentally and can be adjusted based on domain expertise. To reduce the subjectivity of this approach, one may elicit knowledge from a group of experts using one of three group-appropriate techniques: brainstorming, consensus decision making, and the nominal group technique [44].

Similar to our implementation that incorporates temporal variability analysis, we use Dempster-Shafer belief theory to combine the probability masses, which are assigned according to confidence intervals, with the rules included in the heuristic approach. The following example illustrates how we combine the probability mass assignment rules in the heuristic approach with those in the time series analysis. In the example, the station 6462500 has the record period from 1948-1-1 to 1994-9-30 (daily recorded); the record length is 16,709 days; there is one gap event from 1975-9-30 to 1976-10-1; and the *confidence interval* value is 86. Based on the rules defined in Section 4.2, we assign the probability mass to the station as follows:

$$\begin{bmatrix} \text{surfacewater } (S_i) \wedge \text{daily } (S_i) \wedge \text{twentyplus\_years\_old } (S_i) \wedge \neg\text{current } (S_i) \wedge \text{1990s } (S_i) \\ \wedge \text{ oneyear\_gap } (S_i) \wedge [15{,}000 < \text{record\_length } (S_i) < 20{,}000] \wedge \text{SmallConfidenceInterval } (S_i) \end{bmatrix}$$
$$\rightarrow \text{mass } (S_i, \{\text{suitable}\}) = 0.986$$

Table 10. Summary of FoU computation using heuristic and time series analysis

Fitness of use	Heuristic analysis	Time Series analysis
Suitable	13	11
Marginal	1	1
Unsuitable	11	13

Table 11. Sample stations that change classes after time series analysis are added to the system with Rulebase 2

ID	Recency		Length	Confidence interval	Fitness of use (old)			Fitness of use (new)		
	Start date	End date			Suitable	Marginal	Unsuitable	Suitable	Marginal	Unsuitable
6852000	1948-10-1	1994-9-30	8035	274	**0.372**	0.002	0.129	0.197	0.001	**0.286**
6681999	1969-10-1	1990-9-30	7305	1438	**0.287**	0.009	0.154	0.039	0.001	**0.603**

Table 12. Average confidence values in the FoU Classes

Fitness of use	Approaches		
	Heuristics analysis	Temporal variability analysis	Time series analysis
Suitable	0.83	0.80	0.79
Average	0.27	0.24	0.21
Unsuitable	0.60	0.53	0.56

Table 9 shows some examples of the classification using time series analysis. Station 6770478 has a large confidence interval and hence is classified as *unsuitable* with close to 100% confidence. In contrast, Station 6811000 is classified as *marginal* even though the data records are very old because it has a very low confidence interval. Station 6804900, received a *suitable* FoU rating, since it is more recent and has a relatively small confidence interval. However, it does not receive a high confidence (only 0.251) in the classification.

Table 10 lists the new results about the gap datasets. We find that FoU of two datasets changes from *suitable* to *unsuitable*. The change indicates that the gaps have significant impact on the FoU of these two datasets, as shown in Table 11. The classes of other datasets remain unchanged.

Table 11 shows the confidence values for the two datasets whose FoU has changed. Station 6852000 has a moderately large confidence interval. The system is not very decisive in the classification; it thinks that the FoU of the dataset is *unsuitable* with only 0.286 confidence and *suitable*with 0.197 confidence. The system believes that the station is either a *suitable* one or an *unsuitable* one. The confusion is caused by high values for the dataset in its recency, granularity, consistency, and length. The FoU of Station 6681999 also changes from *suitable* to *unsuitable* due to the relatively large confidence interval. In contrast with Station 6852000, Station 6681999 has a higher confidence value 0.603, which implies that the system is confident that the FoU is *unsuitable*.

5.6 Comparison of the approaches

Table 12 summarizes the confidence values for each of the three approaches. It shows the average confidence values assigned to classifications into *suitable*, *marginal*, and *unsuitable* categories. This value reflects the overall confidence in the assignment into classes. In general, the system is able to label a station with suitable FoU with high confidence (~0.80). It is able to label a station unsuitable FoU with lower confidence (~0.55). It is only marginally confident in labeling marginal FoU stations. This matches the experience of our domain experts. They are more confident in labeling a station as *suitable* or *unsuitable*, and not as confident in labeling stations that are in-between.

Table 13 shows the confusion matrix for the classification using the three approaches. The table shows that the three approaches are consistent with each other. For example, of the 183 stations labeled as *suitable* by the Heuristic Analysis, 171 were also labeled as *suitable* by Temporal Variability Analysis and 181 were labeled as *suitable* by the Time Series Analysis. Similarly, 87 stations labeled as *unsuitable* by Time Series Analysis, 85 were labeled as *unsuitable* by Heuristic Analysis and 75 were labeled as *unsuitable* by Temporal Variability Analysis. This provides another kind of validation for the three approaches.

Figure 4 shows the FoU of the surface water stations for the State of Nebraska. With respect to our decision support application for identifying regions with adequate or inad-

Table 13. The matrix of the classification using the three approaches

		Heuristic analysis			Temporal variability analysis			Time series analysis		
		Suitable	Marginal	Unsuitable	Suitable	Marginal	Unsuitable	Suitable	Marginal	Unsuitable
Heuristic analysis	Suitable	**183**	-	-	171	1	11	181	0	2
	Marginal	-	**5**	-	4	1	0	0	5	0
	Unsuitable	-	-	**85**	9	1	75	0	0	85
Temporal variability analysis	Suitable	171	4	9	**184**	-	-	169	4	11
	Marginal	1	1	1	-	**3**	-	1	1	1
	Unsuitable	11	0	75	-	-	**86**	11	0	75
Time series analysis	Suitable	181	0	0	169	1	11	**181**	-	-
	Marginal	0	5	0	4	1	0	-	**5**	-
	Unsuitable	2	0	85	11	1	75	-	-	**87**

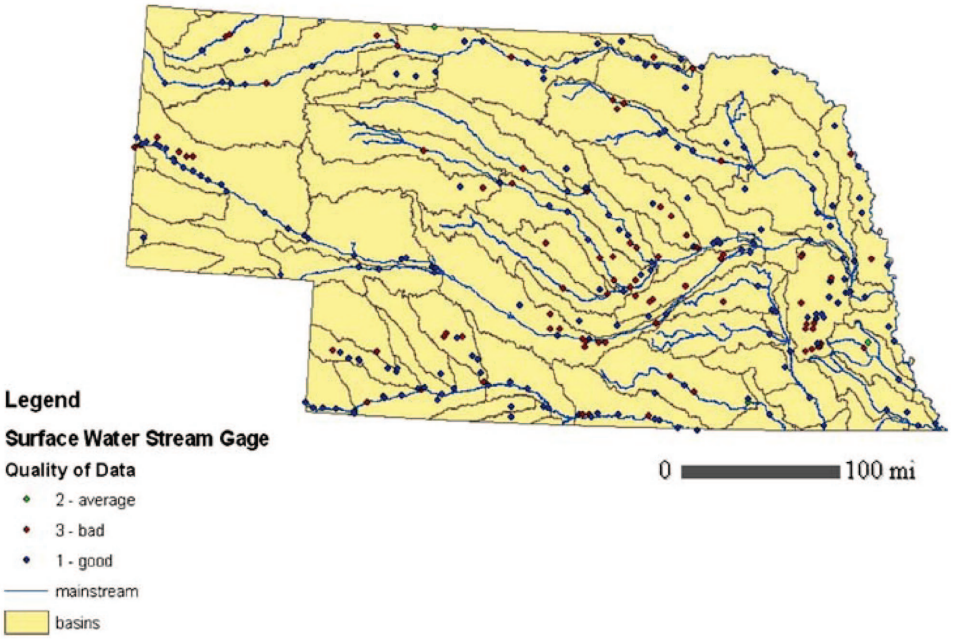

Figure 4. Distribution of surface water stations with FoU Labels

equate measurement of streamflow, the above approach achieves its objective. The map shows the major stream segments as well as the watersheds. The three kinds of FoU values are shown in different colors. The map shows that while there are a number of stations with unsuitable FoU, they are geographically distributed. Similarly, the suitable stations are also distributed, thereby facilitating accurate hydrological analysis based on these datasets. The results show that many of the unsuitable datasets are associated with measuring stations in the central and east-central parts of the state. This is important since this area is a part of the large agricultural belt of the state that relies heavily on irrigation. This information will be useful for both decision makers when they consider requests for irrigation permits. This is also important for decision makers who are responsible for the addition of new and removal of old measuring stations. This region would benefit from additional monitoring stations that would in the long run produce suitable datasets for hydrological analysis and for monitoring water resources.

6 Summary and conclusion

In our study, we have devised a framework to determine the FoU of geospatial time series datasets. As these datasets become widely available and get used in critical decision support systems, the significance of this problem grows even more. We define the FoU of a dataset as its suitability in a given application. This is in contrast with traditional data quality which is designed as an inherent characteristic of a given dataset. We have proposed three approaches (heuristic analysis, temporal variability analysis, and time series analysis) to compute the FoU. They have been evaluated with a specific application in the domain of hydrological analysis, i.e., to provide decision support for decision or policy makers in determining the adequacy or coverage of measuring stations. The heuristic approach

computes the FoU from the characteristics of a dataset such as data record starting time and ending time, data record gap periods. Applying this approach to our surface water datasets for the state of Nebraska, we found that fine measurement of FoU judging criteria helps improve the accuracy of FoU results. The temporal variability analysis approach computes the FoU of datasets by evaluating the fluctuation within datasets. Stable and long records help improve the FoU of a dataset. The time series analysis approach is a good method to evaluate how deeply record gaps impact the FoU of geospatial datasets, particularly for those datasets largely impacted by gaps.

While our approach has been tested in a specific domain, it is applicable in many other domains. In some cases (e.g., weather station records) some of the heuristics described here can be directly used. In other applications, one can develop heuristics (e.g. based on spatial data quality metrics) along the same lines as we have described here.

References

1. J.L. Goodall, D.R. Maidment, and J. Sorenson. 2004. Representation of spatial and temporal data, in *ArcGIS, AWRA GIS and Water Resources III Conference*, Nashville, TN.
2. National Drought Monitor Center. http://drought.unl.edu/, last accessed January 29, 2007.
3. X. Yao. 2003. Research issues in spatio-temporal data mining, in *University Consortium for Geographic Information Science (UCGIS) Workshop on Geospatial Visualization and Knowledge Discovery*. Lansdowne, Virginia (white paper), November 18-20, 2003.
4. Meta Group. 1999. *Data Warehouse Scorecard*. Meta Group.
5. U. Grimmer and H. Hinrichs. 2001. A methodological approach to data quality management supported by data mining, in *Proceedings of the 6th International Conference on Information Quality* (IQ 2001).
6. G. Shafer. 1976. *A Mathematical Theory of Evidence*. Princeton University Press: Princeton, NJ.
7. E. Yudkowsky. An intuitive explanation of Bayesian reasoning, http://yudkowsky.net/bayes/bayes, last accessed January 12, 2007.
8. A. Gelman. 2004. *Bayesian Data Analysis*. CRC Press: Boca Raton, FL.
9. Y.W. Lee and D.M. Strong. Winter 2003-2004. Knowing-why about data processes and data quality, *Journal of Management Information Systems*, 20(3): 13-39.
10. R.Y. Yang, M.P. Ready, and H.B. Kon. 1995. Toward quality data: an attribute-based approach, *Decision Support Systems*, 12: 349-372.
11. L.L. Pipino, Y.W. Lee, and R.Y. Wang. April 2002. Data quality assessment, *Communications of ACM*, 45: 211-218.
12. D.P. Ballou and H.L. Pazer. 1985. Modeling data and process quality in multi-input, multi-output information system, *Management Science*, 31(2): 150-162.
13. K. Huang, Y.W. Lee, and R.Y. Wang. 1999. *Quality Information and Knowledge*. Prentice Hall: Upper Saddle River, NJ.
14. A.X. Zhu. 2004. Research issues on uncertainty in geographic data and GIS-based analysis, in *Research Agenda for Geographic Information Science*: 197-223.
15. M.P. Lynch and A.J. Saalfeld. 1985. Conflation: Automated map compilation—a video game approach, *Proceedings of Auto-Carto 7*, Falls Church, VA.
16. H. Foley, F. Petty, M. Cobb, and K.B. Shaw. 1997. Utilization of an expert system for the analysis of semantic characteristics for improved conflation in geographic information system, *Proceedings of the 10th International Conference on Industrial and Engineering Applications of AI*: 267-275, Atlanta, GA.
17. NCGIA. 1992. A research agenda for geographic information and analysis. *Technical Report* 92-7.
18. M.F. Goodchild and S. Gopal. 1990. *Accuracy of Spatial Databases*. Taylor and Francis: London.
19. M. Blakemore. 1983. Generalization and error in spatial databases, *Cartographica*, 21: 131-139.
20. N.R. Chrisman and M.K. Lester. 1991. A diagnostic test for error in categorical maps, Auto-Carto 10, *Technical Papers of the 1991 ACSM-ASPRS Annual Convention*, 6: 330-348, Baltimore, MD.

21. P.F. Fisher. 1991. Models of uncertainty in spatial data, *in* P.A. Longley, M.F. Goodchild, D.J. Maguire, and D.W. Rhind (editors), *Geographical Information System: Principles and Technical Issues*: 191-205. Wiley: New York.
22. A.X. Zhu. 1997. Measuring uncertainty in class assignment for natural resource maps using a similarity model, *Photogrammetric Engineering and Remote Sensing*, 63: 1,195-1,202.
23. S.C. Guptill and J.L. Morrison. 1995. *Elements of Spatial Data Quality*. Elsevier: Tarrytown, NY.
24. T. Dasu and T. Johnson. April 2002. AT&T Labs—Research SDM-2002, http://www.dataquality-research.com/index.html.
25. J. Hipp, U. Güntzer, and U. Grimmer. 2001. Data quality mining—making a virtue of necessity, *Proceedings of the 6th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery* (DMKD 2001): 52-57. Santa Barbara, California.
26. R. Srikant and R. Agrawal. 1995. Mining generalized association rules, *Proceedings of 21st VLDC Conference*.
27. D. Luebbers, U. Grimmer, and M. Jarke. 2003. Systematic development of data mining-based data quality tools, *Proceedings of the 29th VLDB Conference*, Berlin, Germany.
28. J. Theodore and D. Tamraparni. 1998. Comparing massive high-dimensional data sets, *Proceedings of ACM SIGKDD Conference*.
29. R.Y. Liu and K. Singh. 1993. A quality index based on data depth and multivariate rank tests, *Journal of the American Statistical Association*, 88(421): 252-268.
30. P. Vassiliadis, A. Vagena, S. Skiadopoulos, N. Karayannidis, and T. Sellis. 2000. Arktos: a tool for data cleaning and transformation in data warehouse environments, *IEEE Data Engineering Bulletin*, 23(4): 42-47.
31. R.Y. Wang, H.B. Kon, and S.E. Madnick. April 1993. Data quality requirements analysis and modeling, *Proceedings of the Ninth International Conference on Data Engineering*, Vienna, Austria.
32. B.K. Kahn, D.M. Strong, and R.Y. Wang. 2002. Information quality benchmark: product and service performance, *Communications of the ACM*, 45(4): 184-192.
33. Y.W. Lee, D.M. Strong, B.K. Kahn, and R.Y. Wang. 2002. AIMQ: A methodology for information quality assessment, *Information and Management*, 40(2): 133-146.
34. G. Shankaranarayanan and M. Ziad. 2003. Managing data quality in dynamic decision environment: An information product approach, *Journal of Data Management*, 14(4): 14-32.
35. J.R. Eastman. 2001. Uncertainty management in GIS: Decision support tools for effective use of spatial data, Chapter 18, *in* C. Hunsaker, M. Goodchild, M. Friedl, and E. Case (editors), *Spatial Uncertainty in Ecology: Implications for Remote Sensing and GIS Applications*: 379-390. Springer: New York.
36. K. Sentz and S. Ferson. April 2002. Combination of evidence in Dempster-Shafer belief theory, *SANDIA Technical Report*, SAND2002-0835, http://www.sandia.gov/epistemic/Reports/SAND20020835.pdf.
37. D. Konks and S. Challa. November 2005. An introduction to Bayesian and Dempster-Shafer data fusion, *DSTO-TR* 1436, Edinburgh, Australia, http://www.dsto.defence.gov.au/publications/2563/DSTO-TR-1436.pdf.
38. F. Cremer, E. den Breejen, and K. Schutte. October 1998. Sensor data fusion for antipersonnel land mine detection, *Proceedings of EuroFusion98*: 55-60.
39. J. Braun. 2000. Dempster-Shafer theory and Bayesian reasoning in multisensor data fusion, sensor fusion: architectures, algorithms and applications IV, *Proceedings of SPIE 4051*: 255-266.
40. G. Mihaila, L. Raschid, and M.E. Vidal. April 1999. Querying, "quality of data" metadata, *Proceedings of the Third IEEE Meta-Data Conference*, Bethesda, Maryland.
41. J.C. Giarratano and G.D. Riley. 2004. Expert systems: principles and programming, in *Principles and Programming*, 4th edition, Course Technology.
42. SAS Institute. 1999. *SAS/ETS User's Guide*, Version 8. SAS Publishing: Cary, NC.
43. L.-K. Soh, A. Samal, and W. Waltman. 2003. Watershed study: correlation analysis on seven watersheds in Nebraska. *Technical Report*, Department of Computer Science and Engineering, University of Nebraska.
44. K.L. McGraw and M.R. Seale. 2004. Knowledge elicitation with multiple experts: considerations and techniques, *Artificial Intelligence Review*, 2(1): 31-44.

Lei Fu received his M.Sc. degree in Computer Science from the University of Nebraska–Lincoln, Lincoln, NE, USA.

Leen-Kiat Soh received his B.Sc., M.Sc. and Ph.D. in Electrical Engineering from the University of Kansas, Lawrence, KS, USA. His research interests are in the areas of artificial intelligence, multiagent systems, computer-aided education, and computer science education. He is currently an Assistant Professor in the Department of Computer Science and Engineering at the University of Nebraska–Lincoln, Lincoln, NE, USA.

Ashok Samal received his B.Tech in Computer Science from the Indian Institute of Technology at Kanpur, India, and Ph.D. in Computer Science from the University of Utah, UT, USA. His research interests are image understanding, geospatial analysis, and computer-assisted learning. He is currently an Associate Professor in the Department of Computer Science and Engineering at the University of Nebraska–Lincoln, Lincoln, NE, USA.