CO-049

DATA QUALITY WEB SERVICE FOR SENSOR NETWORKS

HECKE A., ANDERS K.H.
Carinthia University of Applied Science, VILLACH, AUSTRIA

BACKGROUND AND OBJECTIVES

The Carinthian University of Applied Science, Department of Geoinformation (CUAS) is in acitve research cooperation with the Carinthian Government, Department of Hydrography (AKL18). This cooperation focuses on semi-automatic validation of sensor data collected by automatic weather stations. AKL 18 operates a Carinthia-wide sensor network including over 150 automatic weather stations that collect hydrographical, meteorological and nivological data for different environmental parameters (water level, air temperature, precipitation, snow height,…) in a temporal resolution of 15 minutes and submit it to an internal, central server. Multiple factors (malfunction/breakdown of sensors, calibration issues, transmission failure, willful, external influence,…) lead to incorrect data that, till now, are validated and corrected in a time-consuming manual process. Goal of the cooperation is the development of a web-based service that implements methods of a defined theoretic system of rules and enables the user to detect erroneous data and later on reconstructs data for the missing time slots. Hereby, a special focus lies on spatial methods that enable the handling of errors like illustrated in Figure 1.

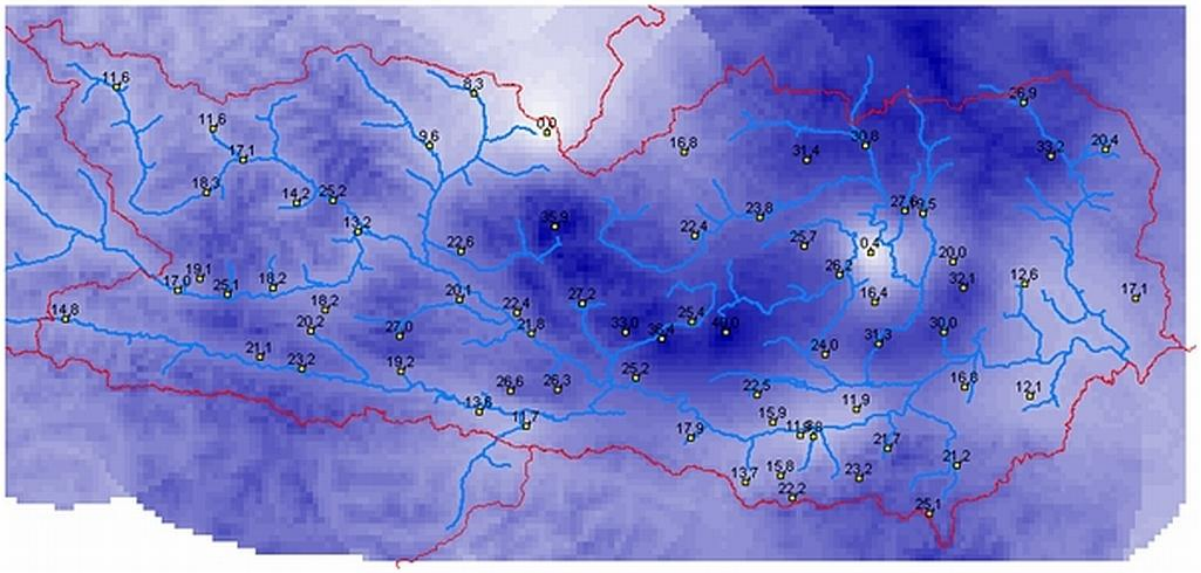

Figure 1 AKL18 sensor network; raster created by Kriging interpolation of precipitation data; the "island" with small values in the central Eastern region is caused by a sensor malfunction

Quality control for automatically collected data is an integral part of every sensor network operator. Most of the time these operators are public organizations (Deutscher Wetterdienst (DWD), Zentralanstalt für Meteorologie und Geodynamik (ZAMG), AKL 18,…) that enforce rather complex quality control mechanism. These mechanism often include several steps and incorporate manual work too [1][5]. Beside public organizations, agricultural institutions research in the field of sensor data validation. First approaches for the development of such systems exist [2][3], the main focus is the optimization of harvest and to correlate plant grow to the predominant meteorological situation. Especially the high operating expense of manual validation accelerates the demand for automatic methods to detect errors in time series and to deal with these errors in an appropriate way. The following chapters document the concept and implementation of our prototypical tool that enables automatic data validation. The presented results arise from the research cooperation as well as further projects that focus on the topics sensor networks and sensor data validation. One of these further projects is Sensors4All that is funded by the federal ministry for science and research under the scope of Sparkling Science. Sparkling Science is intended to give schools a chance to contribute in a real research project and to especially encourage pupils for engineering disciplines.

APPROACH & METHODS

The first phase of the research cooperation can be divided into 3 main parts. The first part includes the development of a general concept and methodology for automatic validation of sensor data. The second part focuses on the compilation of a theoretic system of rules that is used to document methods for error detection and correction based on various environmental parameters. Finally the third phase contains the development of a first prototype, applying the findings during the first two phases. This prototype currently is in a testing phase in a testing environment as well as in the productive environment.

The general concept considers the integration of the prototype into the existing system at AKL18. For this reason we use the TSTP (Time Series Transfer Protocol) – server developed by Aquaplan (www.aquaplan.de) for time series handling in the overall system. Figure 2 illustrates the overall system architecture of eco components (depicted in yellow) developed in the current research cooperation. The TSTP server is used as an interface for the time series that are stored as simple files. It enables HTTP-GET and HTTP-POST access via a small API to read/write time series data. Another important component in the overall system is the software Callisto that is developed by Aquaplan too. This software fetches data in a proprietary format via FTP and writes it to the actual time Series via TSTP. Furthermore Callisto is now used to trigger the automatic evaluation of newly arriving data. Therefore we use the functionality of Callisto to start a predefined script when new data is imported.

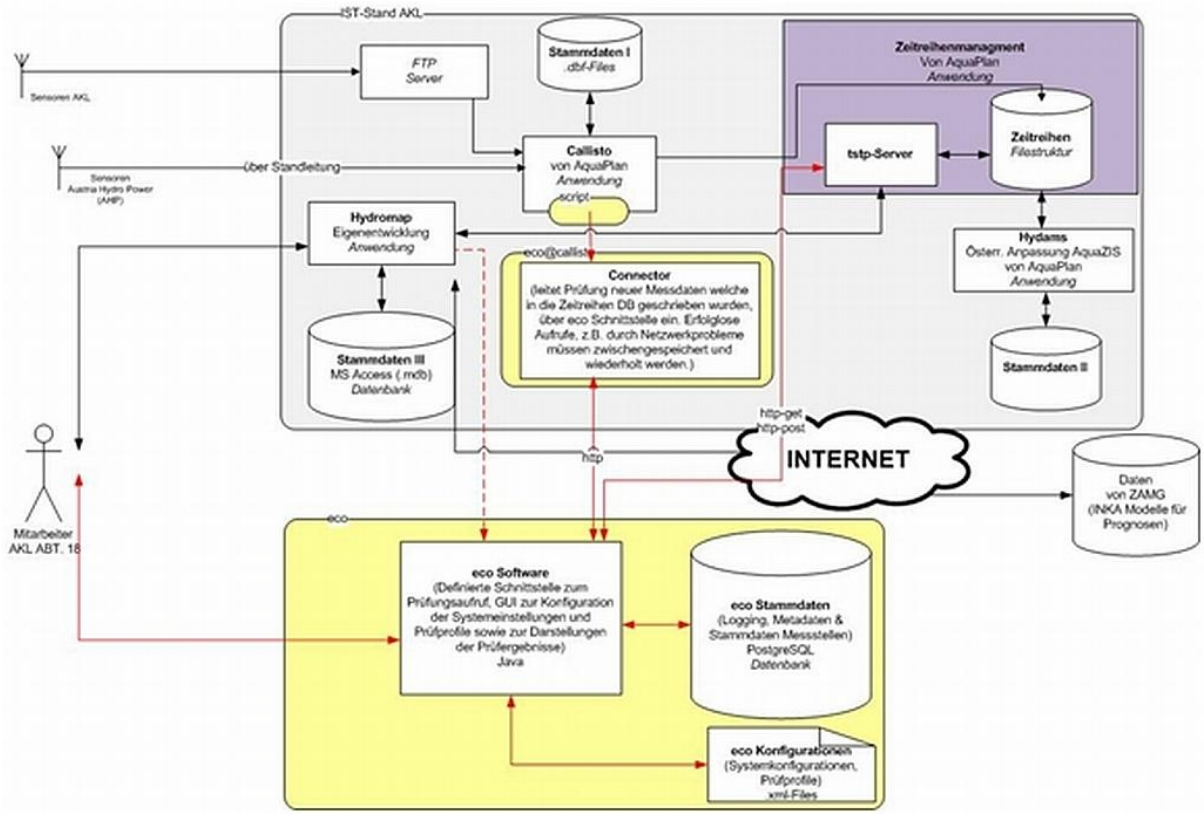

Figure 2 Overall system at AKL18 including the new eco components (yellow)

To configure the system, the user himself has the option to define evaluation profiles that are implemented as simple XML Files with a defined structure, see Figure 3 for an example. Each profile consists of a sequence of rules where every rule has a method for error detection and one for error correction (that is only used if an error is detected). Furthermore, certain methods need additional parameters that are also stored in the profile

```
<?xml version="1.0" encoding="UTF-8"?>
<akl>
    <profile type ="automatic" phenomenon="AirTemperature" create="01/06/2010 12:15" modified="25/06/2010 12:15">
        <rules>
            <rule d_id="LimitsCheck" c_id="SetLuecke" >
                <parameters>
                    <parameter id="upperLimit" value="35.0"/>
                    <parameter id="lowerLimit" value="-45.0"/>
                </parameters>
            </rule>
            <rule d_id="StepChange" c_id="SetLuecke" >
                <parameters>
                    <parameter id="timeSpan" value="15"/>
                    <parameter id="relativeThreshHold" value="10.0"/>
                    <parameter id="absoluteThreshHold" value="10.0"/>
                </parameters>
            </rule>
            <rule d_id="SimpleNeighbourCheck" c_id="InterpolateTimeSeries"/>
        </rules>
    </profile>
</akl>
```

Figure 3 Sample evaluation profile for the environmental parameter air temperature

The simplest rule could consist of the detection method LimitsCheck, with an upper threshold of 40°C and a lower threshold of -40°C as parameters. As method for correction, SetLuecke (set gap) is used. Applying this rule on a time series would mean that each observation that has a value exceeding the two thresholds would be set to the predefined Lücke (gap) value, the TSTP equivalent for NULL values. The method CompareCorrelatingEnvironmentalParameter for example doesn't need further parameters to be defined in the XML evaluation profile. In fact the environmental parameter on which it is used defines the implementation (e.g. when precipitation occurs, the data for global radiation and humidity is checked; as a matter of fact in case of precipitation global radiation will be rather small and humidity rather high).

The theoretic system of rules documents methods for error detection and correction based on environmental parameters. Therefore, existing approaches are used [3][4][5] and extended by new ones, due to the fact that in literature the main focus is set to the parameters air temperature and precipitation. The current version (compare Figure 4) of the compilation documents the detection methods

• *LimitsCheck* (value between upper and lower threshold)

• *StepChange* (gradient within accepted threshold)

• *CompareCorrelatingEnvironmentalParameter* (checking correlating parameter like humidity and global radiation for precipitation)

• *Flatlinercheck* (checks alteration in the time series, e.g. not the same air temperature for the whole day)

• *SimpleNeighborCheck* (compares the actual observation to the median of surrounding stations)

• *SpatialRegression* (compares the actual observation to the value derived by the spatial regression with the other stations)

• *CompareModeledData* (compares the actual observation to a modeled value of e.g. a now casting system provided by the ZAMG)

Furthermore the following correction methods are documented

• *OnlyProtocol* (no handling, except documentation)

• *SetLuecke* (erroneous observation set to NULL)

• *InterpolateFromTimeseries* (linear interpolation)

• *InterpolateFromNeighbors* (median of neighboring stations)

• *UseRegression* (the derived spatial regression value is used)

• *UseModeledData* (e.g. uses the modeled value)

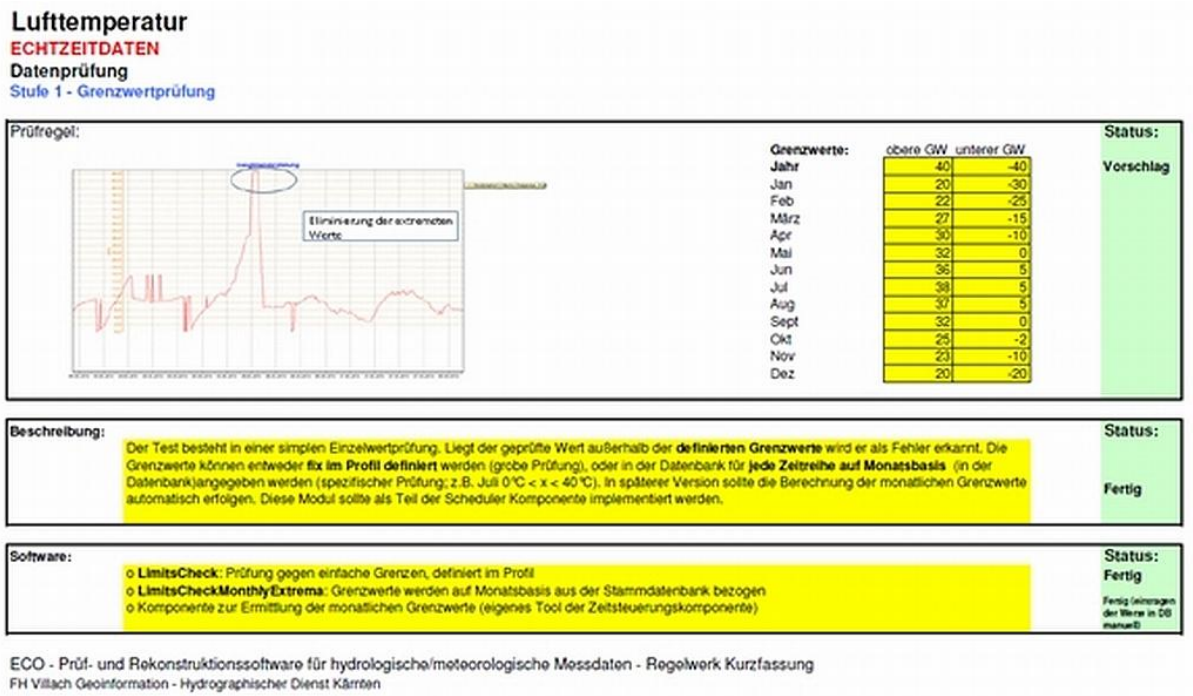

Figure 4 Sample detection method of the system of rules compilation; describing LimitsCheck for air temperature

The first prototype for the given requirements is already implemented, now enabling automatic evaluation of data. It consists of three components ecoEvaluationServer, ecoDatabase and ecoSmallClient. EcoEvaluationServer is a Java servlet-based web application that is deployed in an Apache Tomcat Servlet Container. It is the main component that is triggered when new data arrives; it loads the time series data as well as the corresponding evaluation profile. It logs the evaluation results and finally writes corrected or reconstructed data back to the TSTP server. The main entry point is implemented via a servlet that processes requests for a valid time series id and a time span that should be evaluated. If the user has defined an evaluation profile that is connected to the time series a new evaluation. This evaluation object implements the Java Runnable interface and added to a ThreadPoolExecutor object. In case that non-busy threads are available it is executed immediately, otherwise it is hold in a LinkedBlockingQueue till a nun-busy thread is available. The evaluation object itself works autonomously and reads/writes data by itself as required. Beneath the automatic evaluation the system design defines two further execution modes. The scheduled mode implies an additional scheduler component that enables the user do define evaluation schedules. This mode is used to process time-consuming evaluations that exceed the resources for automatic evaluation. Furthermore the mode can be used for evaluations that don't make perfectly sense in the automatic mode, e.g. the Flatliner Check for air temperature (check if air temperature was constant for the last 12 hours) or in case that metadata has to be updated on a regular basis (e.g. the monthly extremes for air temperature are updated for the last month on the 1st its successor). The manually-operated mode is used to analyze time series in an interactive way so that the user can individually change the correction behavior for each detected error. This mode will mostly be used to evaluate and to optimize the parameters used in detection methods so that the overall accuracy of the eco system is improved.

EcoDatabase is implemented via a PostgreSQL database server and used to store additional meta-information of the system. This includes information on the time series and the corresponding weather station like, operator, position, age, type and so on. Furthermore ecoDatabase is used to store the user-defined relation of time series and evaluation profile as well as the whole logging information like execution of evaluation objects and by that detection and corrections for erroneous observations.

EcoSmallClient is a small Java console application that pipes the triggering action of Callisto Software to the ecoEvaluationServer by translating the original java console invocation into a HTTP GET call. In future this component will also be used to back up calls that have not been processed as the whole system relies on a working network that cannot be guaranteed to be available all the time.

RESULTS

The first prototype is now evaluated in the productive environment. Therefore, a default profile for each of the parameters air temperature, precipitation and water level was created and as evaluation for the

corresponding environmental parameters configured. So at the moment about 400 time series are evaluated when new data arrives and the effects on robustness and performance are observed. Figure 5 illustrates the result of an evaluation, where detected errors are linearly interpolated. Though, the green curve represents the corrected time series, whereas the red curve represents the original time series.

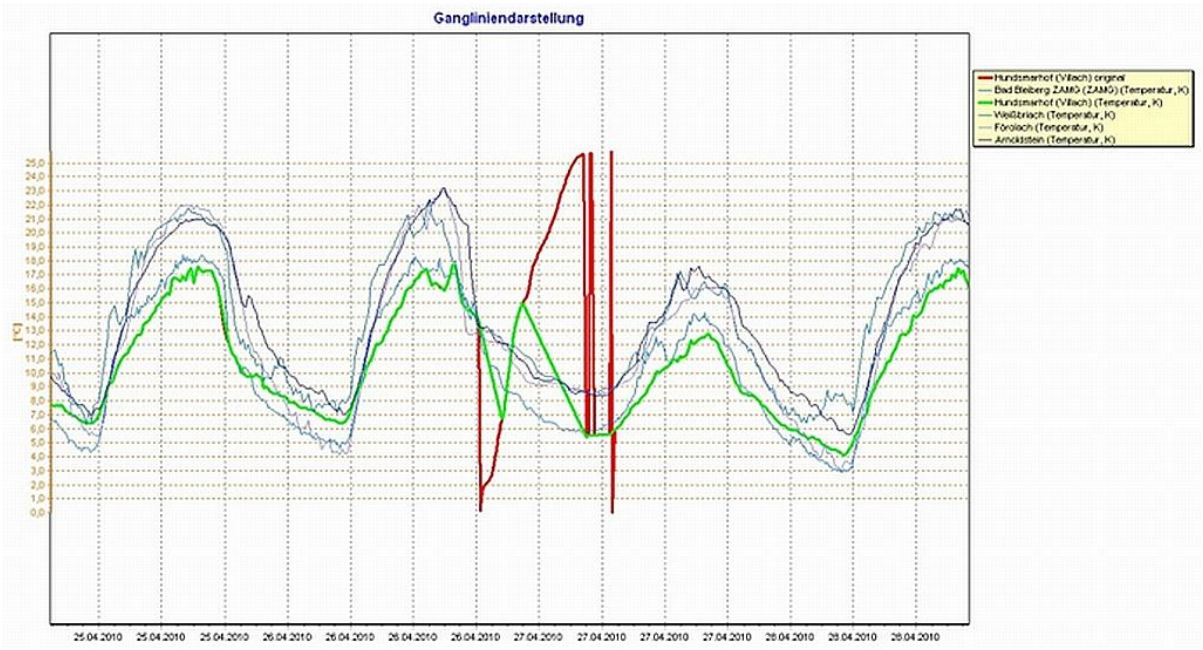

Figure 5 Comparison of an original time series and the newly created, corrected one; the values represented by the red curve where detected as errors and linearly interpolated to the green curve

The example in figure 5 shows also one current problem of the prototype. Simple error correction functions like InterpolateFromTimeseries (linear interpolation) cannot reproduce the correct behavior of a sensor if the malfunction occurs too long relative to the sampling rate. Linear interpolation makes only sense for short time errors (one or two wrong samples). In general the corrected signal should correlate with correlated spatial neighbor sensors. The four blue curves in figure 5 are spatial neighbor sensors in the range of about 10 km distance, which are showing a high correlation to the green sensor curve. In the future we will use these neighbor sensors to detect errors and for the value correction.

CONCLUSION AND FUTURE PLANS

In this paper we have briefly described our work on a sensor evaluation web service, which can be adapted by the user by XML based sensor evaluation profiles. In the next phase a graphical user interface will be implemented, to enable configuration of the system in a client application. Two further evaluation modes will be implemented to allow scheduled evaluations of time series (e.g. once a day in the night for time-consuming operations) and manual, interactive ones too. The theoretic system of rules is under constant review and improvement, further methods will be implemented. In case of successful evaluation of the first prototype, other Austrian provinces are interested in the application of the system too.

Another focus in additional research might be de development of a standard-conform adapter for data exchange. The Open Geospatial Consortium (OGC) defines standards for sensor-related topics in the Sensor Web Enablement Initiative (SWE). An adapter for a Sensor Observation Service (SOS) would be of high interest. Also current investigations are dealing with general rule based systems like Jess [6], to enable a easy adaptable systems. In the moment the user has to define for every sensor an evaluation profile. In future the user should only define facts and possible solutions, but the system will decide by itself when it will use a certain method. Beside the rule system we are improving our error detection and correction method. We are working on robust outlier detectors and correction methods which are based on available neighbor sensors. Other research has to be done on the field of fast and efficient visualization of the sensor data to improve the manual detection of errors, which still will be needed to detect very special errors.

ACKNOWLEDGEMENT

Many thanks go to the Carinthian Government, Department of Hydrography for the financial support in the current research cooperation. Special thanks also to DI Johannes Moser, DI Christian Kopeinig and Christian Wernegger for the support in developing the theoretic system of rules, the discussions on system

architecture and integration as well as for the deployment. Additional thanks go to DI Dr. Gerald Gruber for the support through the department of Geoinformation and to our colleagues DI (FH) Stefanie Andrae and DI (FH) Alfred Wieser.

LITERATURE

[1] Deutscher Wetterdienst (DWD), (2009), Qualitätssicherung und Kontrolle beim DWD, Informationsfolder.

[2] Fröhlich Georg, (2001): Modellierung, Realisierung und Validierung eines offenen Managementsystems für agrarmeteorologische Messdaten, Dissertation an der TU München, Weihenstephan.

[3] Mateo Mark A. F., Leung Carson K., (2008): Design and Development of a Prototype System for Detecting Abnormal Weather Observation, Proceedings of the 2008 C³S²E conference, Montreal/Quebec.

[4] Reek Thomas, Doty Stephen I., Owen Timothy W., (1992): A Deterministic Approach to the Validation of Historical Daily Temperature and Precipitation Data from the Cooperative Network, Bulletin American Meteorological Society, Vol. 73, No.6.

[5] Vejen Flemming, Jacobsson Caje, Fredriksson Uwe, Moe Margareth, Adresen Lars, Hellsten Eino, Rissanen Pauli, Palsdottir Poranna, Arason Pordur, (2002): Quality Control of Meteorological Observations – Automatic Methods Used in the Nordic Countries, Climate Report, Norwegian Meteorological Institute, Oslo.

[6] www.jessrules.com (last accessed 14.02.2011)