Automatic processing, quality assurance and serving of real-time weather data

Matthew Williams[a], Dan Cornford[a], Lucy Bastin[a], Richard Jones[a], Stephen Parker[a]

[a]*Knowledge Engineering Group, Aston University, Birmingham, B4 7ET, UK*

Abstract

Recent advances in technology have produced a significant increase in the availability of free sensor data over the Internet. With affordable weather monitoring stations now available to individual meteorology enthusiasts, a reservoir of real time data such as temperature, rainfall and wind speed can now be obtained for most of the world. Despite the abundance of available data, the production of usable information about the weather in individual local neighbourhoods requires complex processing that poses several challenges.

This paper discusses a collection of technologies and applications that harvest, refine and process this data, culminating in information that has been tailored toward the user. In this instance, this allows a user to make direct queries about the weather at any location, even when this is not directly instrumented, using interpolation methods provided by the INTAMAP project. A simplified example illustrates how the INTAMAP web processing service can be employed as part of a quality control procedure to estimate the bias and residual variance of user contributed temperature observations, using a reference standard based on temperature observations with carefully controlled quality. We also consider how the uncertainty introduced by the interpolation can be communicated to the user of the system, using UncertML, a developing standard for uncertainty rep-

resentation.

Key words: User-contributed data, UncertML, INTAMAP, data quality, interpolation

1. Introduction

The term ‘mashup’ in Web development refers to the combination of different services and data into a single integrated tool. This paper discusses a mashup in which weather data from hundreds of individual sensors is harvested, refined and processed using several interoperable standards, to provide information that has been customised to a user’s requirements. To support the practical use of this data, streamlined interfaces have been developed that provide access for small footprint devices, e.g. mobile phones. The combination of these technologies results in a tool capable of navigating seemingly complex data and providing answers to highly specific queries such as “What is the temperature in my garden right now?” and “Will the roads be icy on my way home?”.

Section 2 introduces the mashup architecture with an overview of the data flow. Section 3 details the harvesting process and the interface to the data. Section 4 notes the importance of uncertainty propagation through the system, and describes the methods and standards used to achieve this. Section 5 discusses the refining and processing stages that occur as part of the INTAMAP interpolation service [1]. Section 6 describes a technique used to estimate the uncertainty of the user-contributed data, using the INTAMAP service, and Section 7 gives more detail on client applications that use the framework to gather information that has been tailored for them. Finally, we gather conclusions and insights in Section 8.

[1] http://www.intamap.org

2. Overview

The system discussed in this paper provides access to user-contributed weather data through open standards. Wrapping Weather Underground data with an interoperable interface allows more structured access than presently available. The system also provides a mechanism for estimating the uncertainty and bias of the Weather Underground data; providing users with more detailed information.

The interfaces used within the system employ the latest technologies from the Open Geospatial Consortium (OGC). The OGC is a standards organisation that develop and maintain XML standards for geospatial services. Specifically, a Sensor Observation Service (SOS) (Na and Priest, 2007) interface provides an access layer to the underlying weather data. A SOS interface provides the basic create, update, retrieve and delete functionality, commonly associated with databases, for sensor-observed data. Data can be filtered spatially, temporally or by specific attribute values. The uncertainty estimation process is provided by the INTAMAP (INTerpolation and Automated MAPping) project. INTAMAP is a Web Processing Service (WPS) (Schut, 2007), providing near real-time interpolation of sensor data (Williams et al., 2007). The WPS interface is more abstracted than the SOS, providing a loose framework within which any arbitrary process may reside. Data communicated between the services and clients is encoded using the Observations & Measurements (O&M) (Cox, 2007) standard. O&M provides a common encoding for all sensor-observed data. However, the properties of an observation within O&M are flexible, allowing the integration of other XML specifications. Specifically this system integrates UncertML, a language for quantifying uncertainty (Williams et al., 2009). UncertML [2] is

[2] `http://www.uncertml.org`

a relatively new XML vocabulary and is currently under discussion within the OGC. Embracing the open standards laid out by the OGC results in a collection of loosely-coupled, autonomous, services. These design criteria underpin the philosophy behind Service Oriented Architectures (SOAs) (Erl, 2004, 2005).

Each of the components depicted in Figure 1 provides specific functionality that combines to produce a usable system. This section gives a brief overview of the main components, while Sections 3 – 7 investigate the finer details.

The system components can be logically divided into three groups: data acquisition, processing services and client applications. The data is acquired from the Weather Underground Web site and stored in a database (Step 1). Access to the data is provided by a SOS, (discussed in Section 3.2.2), which is essentially a Web Service providing simple insertion and retrieval methods for observation data. The observations returned by the SOS are encoded in the O&M schema, as discussed in Section 3.2.1.

Steps 2-5 cover the processing and correction of the data. Processing of the data is handled by a WPS, a standardised interface for publishing geospatial processes. The WPS used here was developed by the INTAMAP project. It provides bleeding-edge interpolation methods through a WPS access layer, and is discussed in greater detail in Section 5. Section 6 outlines a Matlab application that utilises INTAMAP and the SOS interface to estimate uncertainties on the user-contributed data collected from Weather Underground.

Step 6 is the stage at which data is actually consumed or updated by client applications using the processing and access components, and these applications are discussed in Section 7. The whole system demonstrates the benefits of INTAMAP and of the interoperable infrastructure to which INTAMAP lends itself.

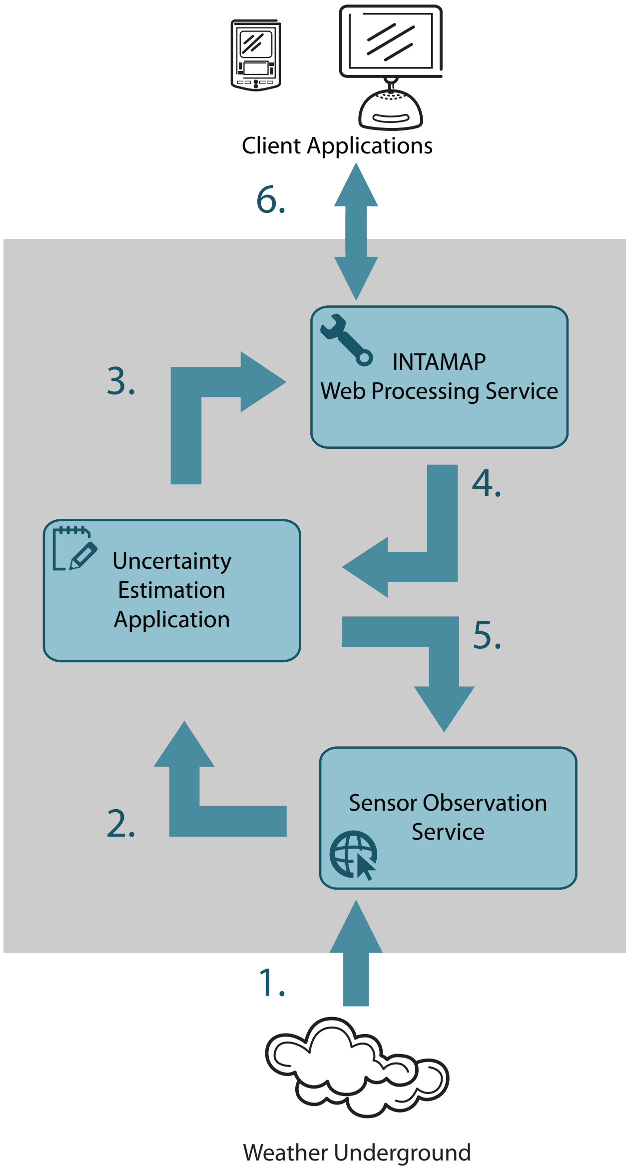

Figure 1: An overview of the system architecture shows the flow of data from the Weather Underground Web site to the end-user client application. A SOS provides an interoperable interface to the data. Uncertainty of the user-contributed data is estimated using the INTAMAP service, and used to update observations. The uncertainty (in this case, the prediction variance) of the final interpolated map is also conveyed to the client.

3. Data acquisition, storage and access

The system outlined in the previous section revolves around user-contributed data. All data used within this system is weather data, specifically temperature values in degrees Celsius. However, the software and statistical methods discussed have general applicability and might be used with a variety of datasets, including other weather variables such as pressure, soil contamination measurements, bird sightings (transformed into density maps) or disease reports from monitoring networks.

3.1. Weather Underground

Weather Underground[3] is an online community of weather enthusiasts providing up-to-the-minute information about current weather conditions around the globe. Under its surface lies a vast repository of freely available weather data recorded by thousands of individual weather stations. This data is proprietary to Weather Underground Inc. and may be used for non-commercial purposes provided that the source is clearly acknowledged. Commercial use, however, is not permitted without advance written consent [4]. For this experiment we used a subset of data gathered from the Weather Underground repositories.

Each of the contributing stations on Weather Underground has a 'current conditions' XML file which is updated each time the station sends a new set of observations. However, this XML file does not conform to any recognised XML Schema standard, severely hindering third party consumption. Supplementing the 'current conditions' file is a 'historic observations' file containing all previous data; however, this is formatted in Comma Separated Values format, which

[3] http://www.wunderground.com

[4] http://www.wunderground.com/members/tos.asp

obstructs interoperability. Furthermore, access to the data is hidden behind a series of Web pages that offer no interoperable API, and limited querying functionality. Section 3.2 discusses how we solved these problems by providing an interoperable infrastructure to the Weather Underground data.

While user-contributed data is vast in quantity, it may vary drastically in quality. Issues such as quality of sensing equipment and location of sensor will affect the accuracy and precision of any observed values. Quantifying these uncertainties probabilistically allows more informed and sophisticated processing, for example through a Bayesian framework (Gelman et al., 2003). Weather Underground currently does not provide any uncertainty information with the observation data, and so Section 6 outlines a technique for estimating these uncertainties using interpolation. The reference level for this technique is based on temperature measurements from the UK's Met Office[5], which have well-characterised uncertainty.

3.2. Interoperable Weather Underground infrastructure

This section discusses solutions to several important issues with Weather Underground data, namely:

- no recognised interoperable standard for describing observation data,
- no interoperable interface to query and access the data, and
- no quantified uncertainty information.

These are issues which are likely to arise with many user-contributed data networks, so these solutions could be adapted to many other contexts.

[5] http://www.metoffice.gov.uk

3.2.1. Observations & Measurements

Weather Underground data does not conform to a recognised XML standard, and is therefore cumbersome and difficult to integrate into existing standards-compliant software. For the purpose of the system outlined in Section 2, the Observations & Measurements (O&M) standard was adopted. O&M was developed and agreed by the OGC, and is a conceptual model and encoding for describing observations (Cox, 2007). The conceptual model outlined in the O&M specification is perfectly suited to describing data recorded at weather stations, and consequently is ideal for encoding data from the Weather Underground. The base of the model can be broken down into a *feature of interest*, i.e. the observation target (which usually includes a geospatial component), and an *observed result*. Further information is captured within other properties, some of which are detailed below:

observedProperty the phenomenon for which the result describes an estimate.

procedure a description of the process used to generate the result, typically described using the Sensor Model Language (Botts and Robin, 2007).

resultQuality quality information about the observed value. This is pertinent to the third issue outlined in Section 3.2.

Utilising the O&M language as a transportation device lays the foundations of an interoperable weather data exchange platform. To build on these foundations we employ another OGC standard, the Sensor Observation Service.

3.2.2. Sensor Observation Service

With the standard closed interface, access to and subsequent processing of the Weather Underground data is difficult. Providing an open, XML-based, API opens up this wealth of information for consumption by standards-compliant software. The Sensor Observation Service (SOS) standard (Na and Priest, 2007) complements O&M by providing a series of methods for accessing observation data. The SOS is a Web Service which outputs requested observations in the form of an O&M instance document. By utilising the OGC Filter encoding specification (Vretanos, 2005), complex queries can be performed, filtering by time, space, sensor or phenomenon.

The SOS employed in this system was built around the 52 North SOS implementation[6]. Currently, no existing SOS implementation provides the functionality to serve observations with attached uncertainties. For the purposes of this system, therefore, we developed an extension of the 52 North SOS that allows uncertainty to be included in the SOS output through the use of UncertML. This extension provides the functionality to describe observation errors by a variety of means; as statistics (variance, standard deviation etc), as a set of quantiles, or as probability distributions. The generated UncertML is inserted into the O&M **resultQuality** property. UncertML is discussed in detail in the following section.

4. Propagating uncertainty through a series of interoperable services

Uncertainty exists within all data measured by sensors, and the magnitude of this uncertainty increases greatly in the case of user-contributed data. Issues such as poor quality measuring equipment, ill-positioned sensors and observation operator errors all contribute to unreliable measurements. Processing this data through models, such as interpolation, propagates these uncertainties, and this is a particularly important consideration in the case of spatially-referenced

[6]http://52north.org/

data, where recorded sensor location may also be unreliable Heuvelink (1998). In order to optimally utilise any data (for example, within a decision making support tool) users require as complete a numerical description of its uncertainties as possible.

Traditionally, environmental models and decision support tools have been implemented as tightly-coupled, legacy software systems (Rizzoli and Young, 1997). When migrating to a loosely-coupled, interoperable framework, as discussed here, a language for describing and exchanging uncertainty is essential. UncertML, a language capable of describing and exchanging probabilistic representations of uncertainty, was used throughout this system.

4.1. UncertML overview

UncertML is an XML language capable of quantifying uncertainty in the form of various statistics, probability distributions or series of realisations. This section provides a brief overview of UncertML; for a complete guide we refer the user to Williams et al. (2009).

All uncertainty types discussed here (e.g., the `Statistic`, the `Distribution` and the `Realisations`) inherit from the `AbstractUncertaintyType` element (Figure 2). This allows all types to be interchanged freely, giving an abstract notion of ‘uncertainty’, whether it be described by summary statistics, density functions or through a series of simulations. It should be noted that the scope of UncertML does not extend to issues covered by other XML schemata including units of measure and the nature of the measured phenomena. This separation of concerns is deliberate, and allows UncertML to describe uncertainty in a broad range of contexts.

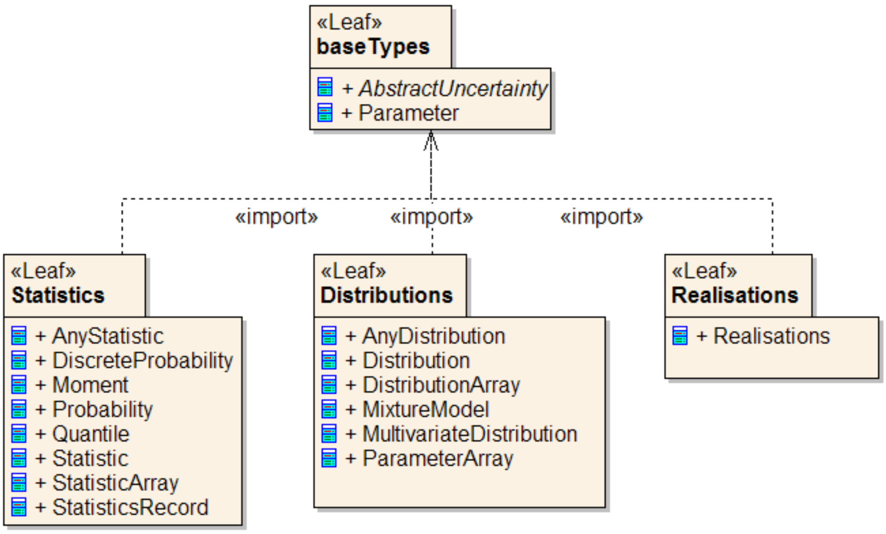

Figure 2: An overview of the UncertML package dependencies.

```
<un:Statistic definition="http://dictionary.uncertml.org/
    statistics/mode">
    <un:value>34.67</un:value>
</un:Statistic>
```

Listing 1: A `Statistic` describing the mode value of a random variable.

4.1.1. *Statistics*

Most statistics are described using the `Statistic` type in UncertML. As with all types in UncertML, the `Statistic` references a dictionary via the `definition` attribute. It is this semantic link, combined with a `value` property, that enables a single XML element to describe a host of different statistics. Listing 1 shows an UncertML fragment describing the statistic ‘mode’.

UncertML also provides two aggregate statistic types. The `StatisticsRecord` is used to group numerous different statistics and the `StatisticsArray` is a concise method for encoding values of the same statistic type. Aggregates may be used within one another, i.e. a `StatisticsArray` of `StatisticsRecord`s and

```xml
<un:Distribution definition="http://dictionary.uncertml.
  org/distributions/gaussian">
  <un:parameters>
    <un:Parameter definition="http://dictionary.uncertml
      .org/distributions/gaussian/mean">
      <un:value>34.564</un:value>
    </un:Parameter>
    <un:Parameter definition="http://dictionary.uncertml
      .org/distributions/gaussian/variance">
      <un:value>67.45</un:value>
    </un:Parameter>
  </un:parameters>
</un:Distribution>
```

Listing 2: A Gaussian `Distribution` with mean and variance parameters.

vice versa.

4.1.2. Distributions

Within UncertML, parametric distributions are syntactically similar to statistics. However, semantically, distributions provide a complete description of a random variable and are therefore an integral component. The `Distribution` type in UncertML is used to describe any parametric distribution; the addition of ‘parameters’ instead of a single value differentiates the `Distribution` from the `Statistic` (Listing 2).

A `DistributionArray` allows multiple distributions to be encoded concisely. Types for describing mixture models and multivariate distributions also exist.

4.1.3. Realisations

In some situations, a user may not be able to simply represent the uncertainties of the data with which they are working. In such a situation, a sample from the random quantity might be provided, allowing uncertainty to be described implicitly. Within UncertML this is achieved using the `Realisations` type.

4.2. Propagating UncertML through interoperable services

UncertML was integrated into several key areas throughout the system outlined in Section 2. Firstly, the access and storage of the user-contributed data is handled by an extended (i.e., ‘uncertainty-enabled’) implementation of the 52 North Sensor Observation Service (Section 3). Secondly, the INTAMAP Web Processing Service, which provides advanced interpolation methods in an automatic context, can utilise UncertML-encoded information. The only mandatory input to INTAMAP is a collection of observations encoded in the Observations & Measurements schema. Where observation errors are known, they are encoded as UncertML and included in the O&M instance. In this system the observations came directly from the UncertML-enabled SOS. Thirdly, the output of the INTAMAP service is an UncertML document including any propagated uncertainties. Client applications are then able to produce visualisations of the predictions and accompanying uncertainty.

5. INTAMAP

Providing weather information that has been tailored toward the user relies on either *knowing* the weather at the user’s location, or, more frequently, *predicting* the weather at the user’s location using observed data at known locations. This process of prediction is typically called interpolation. The INTAMAP (INTeroperability and Automated MAPping) project provides an open interface to complex geostatistical algorithms (Williams et al., 2007). Combining an interoperable interface and *automated* interpolation methods allows INTAMAP to be accessed by inexperienced geostatistical users.

INTAMAP uses, as an interface, the interoperable framework provided by the OGC's Web Processing Service (WPS) specification. This framework supplies a formal structure that enables the description of any geostatistical process through its inputs and outputs. INTAMAP has a single mandatory input - a series of observations encoded in the Observations & Measurements standard. However, several other *optional* inputs exist to allow the user to customise the work flow. Using these options, a user can, for example, specify the prediction locations using Geography Markup Language (GML) (Portele, 2007), or request exceedance probabilities using UncertML. Ultimately, however, the capacity of INTAMAP to automate many choices is what makes the service accessible. For example, if users supply the bare minimum inputs, without specifying an algorithm or supplying a GML-encoded spatial domain for their results, the service will select the most appropriate interpolation algorithm based on the statistical characteristics of the input observations, and will automatically calculate the extent and resolution of the output maps, based on their spatial arrangement. This allows users to easily test and explore INTAMAP's capabilities, and refine their requirements as they learn more about the options offered. A typical output of INTAMAP is the mean (predicted value) and prediction variance (a measure of uncertainty), encoded in UncertML, at a single location, at several locations or over a regular grid. Complementing the Web Processing Service is an Application Programming Interface (API) written in Java. This API handles XML writing and parsing, allowing INTAMAP to be integrated into existing Java applications with very few lines of code. Tools within the API also allow the creation, where applicable, of GeoTiff files to visualise the results.

Behind the WPS interface lies an interpolation engine written in the statistical language 'R'[7]. Several differing interpolation methods are available, catering for a range of scenarios. Automap (Hiemstra et al., 2008) provides an automatic implementation of Ordinary Kriging. For contexts where the data contains extreme values, or "hot spots", a Copula Kriging method (Kazianka and Pilz, 2009) is provided. A third method, Projected Spatial Gaussian Process (PSGP) (Ingram et al., 2008) addresses two issues:

- the cubic growth in computational complexity for likelihood based inference in Gaussian process models (model-based geostatistics) which limits their application to smallish data sets of less than 2000 observations;
- the inability of most geostatistical methods to deal with non-Gaussian errors on observations, or non-linear sensor models.

The first point makes PSGPs particularly useful when tackling large datasets (more than 2000 observations). However, it is the second point that enables the PSGP method to propagate the observation errors within the user-contributed data. INTAMAP is able to select an appropriate interpolation method for a specific dataset using several criteria; data characteristics (e.g., the presence of extreme values); time constraints; and the presence or absence of quantified uncertainties on the observations.

6. Using INTAMAP to estimate observation error on user-contributed data

The data obtained from Weather Underground is submitted by a range of users, who will apply differing levels of quality control to their data, and site their sensors in a wide variety of locations and exposures. In contrast, weather

[7] http://www.r-project.org

data collected by professional meteorological services undergoes rigorous quality control, and is collected under standardised conditions, including specification of the instrument housing and height, the surrounding enclosure and the exposure of the site (Oke, 1982). When instruments (and in particular the thermometers which we consider here) are sited in urban areas, their readings are likely to be strongly affected by the micro-climates that exist around buildings. These micro-climates, which can particularly affect readings from easily-accessible monitoring locations such as domestic homes and gardens, are largely related to changes in thermal storage and associated radiative balance (World Meteorological Organization, 1983). It is also quite possible that some instruments might not be correctly screened from direct radiation, or are attached to walls that are themselves exposed. In the following section we explore how statistical methods, based on using the INTAMAP web service, can be used in a simplistic manner to estimate the observation bias and residual observation variance in these user-contributed data. We note that the methods applied here are intended to be illustrative. Therefore they often employ rather simplistic assumptions, which will be discussed later.

In order to address the issue of bias in the Weather Underground data, we need to determine a reference level or standard. In this work we use temperature observations from the Met Office synoptic observing network, (denoted T_{MO}), which were obtained from the British Atmospheric Data Centre. Hourly temperature data were obtained at 203 synoptic stations covering the UK for the 27th of May 2009. This day was chosen because it was relatively challenging to the simplifying assumptions made in the analysis. A warm front was crossing the UK from the west, with clearer conditions over northern Scotland, thus the weather situation was complex, with cloudy skies over most of the UK, a situation that might be expected to minimise any biases due to micro-climatic effects, but clearer skies over the north and east of Britain which could show significant biases. The Weather Underground temperature data (denoted T_{WU}) was also obtained for the same period, and the observations closest in time to the hourly synoptic data were selected for each site, so long as they were within 15 minutes of the synoptic observation time.

A gross outlier removal method excluded all observations outside the range -25^oC to $+30^oC$ which is climatologically reasonable. The aim of the outlier removal is to remove outliers in the Weather Underground data that are the result of instrument failure, transmission errors and other processes which produce very implausible observations. Visualising the resulting data reveals no further clearly defined outliers. After this selection around 500 Weather Underground stations were available for each hour.

A more sophisticated treatment of outliers is possible, and ultimately desirable, for automated preprocessing and quality control of user-contributed data. Several detailed reviews on the topic offer and evaluate techniques which will be of value for further development of such systems. These include algorithm comparison and benchmarking exercises for interpolating noisy data, such as the Spatial Interpolation Comparison (EUR, 2003, 2005), and more detailed considerations of spatial outliers (points whose values are particularly unusual in the context of their local spatial neighbourhoods) (Shekhar et al., 2003; Chawla and Sun, 2006). Spatial outliers are especially important in the context of automated decision support because of the capacity of 'false positive' values to trigger alerts and the opposing need to capture genuine extreme events (Sharma et al., 1999; Pilz and Spock, 2008). A number of studies have considered how existing statistical methods to detect clusters and spatial outliers might be extended for auto-

Algorithm 1 Outline of the simple bias estimation algorithm applied to the Weather Underground data.

Remove gross outliers from the Weather Underground data
Randomly split the Met Office data into training and validation sets
for hour = 1 to 24 **do**
    Use the psgp method on the INTAMAP system to predict $\hat{T}_{WU}$ using T_{MO} with a variance estimated to be $0.36^o C^2$
    Compute $\delta T_{WU} = T_{WU} - \hat{T}_{WU}$
end for
Compute $T_{WU}^{bias} = \mathrm{E}[\delta T_{WU}]$
Compute $T_{WU}^{var} = \mathrm{var}[\delta T_{WU}]$

mated systems (Patil and Taillie, 2003; Brenning and Dubois, 2008) while recognising the influence of heterogeneous covariates (Goovaerts and Jacquez, 2004). This body of work offers some robust solutions for future quality control Web Services; however, for this simple exploratory example, such treatment was not deemed necessary.

The basic idea of this analysis is that we employ the INTAMAP interpolation system to predict the temperature at the Weather Underground locations, based on the Met Office synoptic station observations, which we assume are unbiased. In order to withhold a set of observations for validation of our approach the synoptic station data is split into two halves using random sampling. One half is used for prediction at the Weather Underground locations and the other half retained for validation. Since random sampling is used for the locations of the training and validation sets, it is possible that the results could be sensitive to this partition; however, a sensitivity analysis reveals that the results shown in the paper are stable with respect to this partition, presumably because 100 stations is a sufficiently large number to attain reasonable coverage of Britain. A summary of the overall approach is shown in Algorithm 1. The approach is very simplistic,

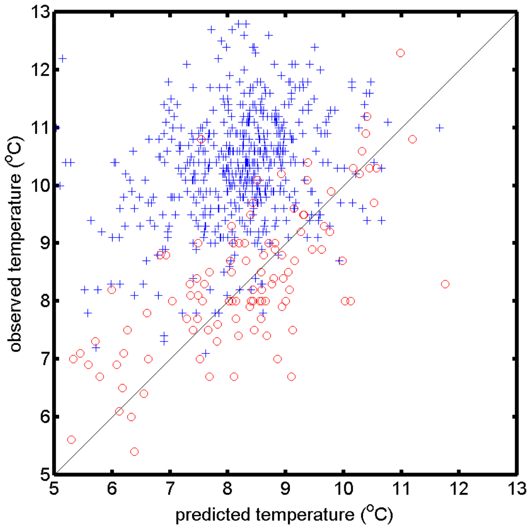

Figure 3: Predicted versus observed temperatures for Weather Underground (blue crosses) and Met Office (red circles) stations at 09:00 on the 27th May 2009.

but illustrates well the dangers of using uncorrected user-contributed data.

Figure 3 shows a plot of predicted versus observed temperatures. It is well known that temperatures are extremely sensitive to elevation, particularly in locations such as Britain (Cornford and Thornes, 1996). Therefore, prior to all interpolation a linear trend in both x,y and elevation is removed. The trend model is estimated using least squares methods, which is strictly not appropriate here due to the correlated residuals, but does allow the INTAMAP service to be used without modification. A more refined version could employ universal kriging or regression kriging (Hengl et al., 2007), however for this illustration the differences are likely to be small. The typical lapse rates estimated for the period examined range from 3.5 to $5.1^oC/km$, and the inclusion of the lapse rates improves the estimation of the variograms in the interpolation process as might be expected. The residual process is spatially correlated and a variogram is fitted in

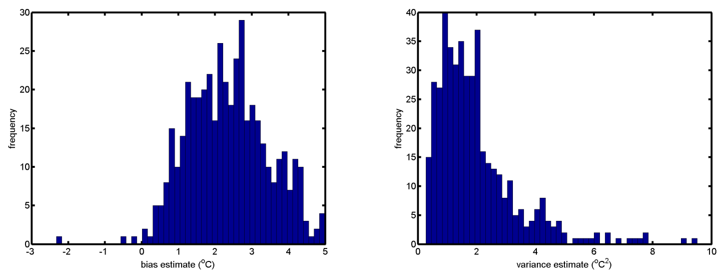

Figure 4: Histograms of the estimated bias (left) and residual variance (right) for Weather Underground temperatures for 27th May 2009.

the INTAMAP system with range parameters typically between 100 and 200 km, sill variances typically 2^oC^2 and nuggets typically 0.5^oC^2, this varying with time of day. The average minimum distance between Met Office stations in the training data is ~ 40 km making spatial prediction of the regression residuals using kriging appropriate. The predictions are based on the training set of Met Office stations, and are made at both Weather Underground and Met Office validation set locations. It is immediately clear that the Weather Underground stations are significantly biased, being typically some 2^oC warmer than might be expected (the mean bias is 2.34^oC and the standard deviation is 1.09^oC). The validation set of Met Office stations remains essentially unbiased. The scatter is reduced for the Met Office stations compared to earlier work which ignored the effect of elevation. The scatter for the Weather Underground stations is larger, and is not significantly changed by the addition of elevation as a predictor, suggesting that there might be other factors affecting these which are not connected to elevation.

Looking at the statistics of the bias and residual variance based on these predictions, on average the Weather Underground stations are significantly positively biased (although not all are), and many have rather large residual variances (Figure 4). The positive bias might be expected – Weather Underground stations

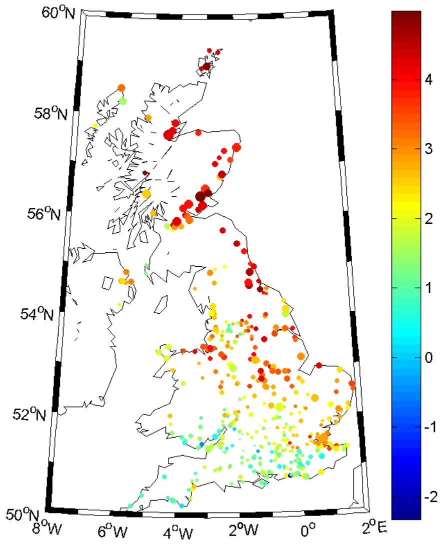

Figure 5: Mapping the estimated bias (colour) and residual variance (symbol size) for Weather Underground temperatures for 27th May 2009.

are often sited in urban areas, since they are often in the owners' gardens, which tend to be more sheltered and closer to large buildings than the standard Met Office enclosures. Figure 4 shows that while many Weather Underground stations are significantly biased, some are not biased at all with respect to the synoptic station measurements. This emphasises the degree of variability in the estimated biases – a single bias estimate for the whole Weather Underground station network would not be sufficient. The same pattern can be seen in the variance.

Figure 5 shows the spatial distribution of both the estimated bias (colour) and variance (size) at the Weather Underground sites where data was available for the full 24-hour study period. There are interesting patterns in this plot, but it is rather difficult to ascribe these to specific causes – they might be related to meteorological conditions, social differences in the locations of instruments and

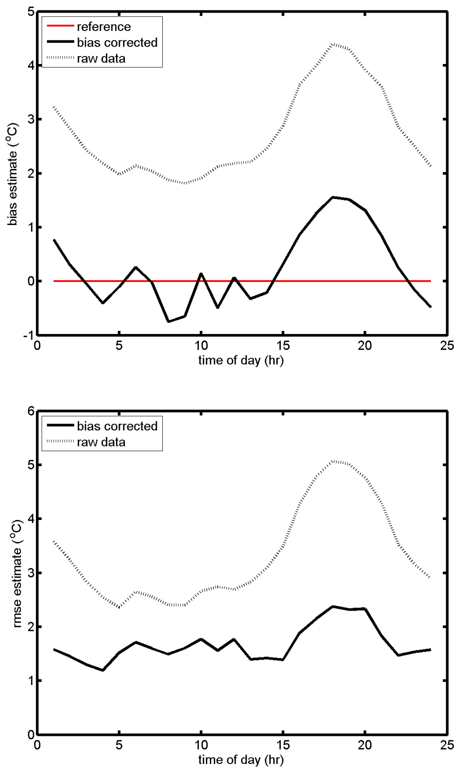

Figure 6: Testing the bias correction, using corrected and raw Weather Underground data to predict at the withheld Met Office stations. Prediction bias (left) and root mean square error (right) for 27th May 2009.

local environment, or, most likely, a combination of the above. It should also be noted that the bias correction will be most reliable when the Met Office stations are close to the Weather Underground stations, due to the use of a random field model. If this method for bias estimation were to be used in a more serious application, further developments of the model would be required and more extensive model validation would be necessary to ensure the robustness of the results.

Such a bias-corrected set of observations from Weather Underground could have two important advantages, as follows.

Firstly, it would allow Weather Underground data to be used as standardised data inputs in a wide variety of application domains - for example;

- monitoring climate change;
- numerical weather model data assimilation streams
- mapping surface air temperature to explore vegetation growth in the UK.

- with the caveats that to make full use of the data a more complete characterisation of the micro-meteorological environment of the stations would be required. There might be some concern that such processed data would not be suitable for monitoring climate change, because the bias correction is based on the reference stations (the Met Office network). However this network is carefully quality controlled and represents the best estimate we have of surface climate change. An interesting point for future analysis would be to monitor how the bias and variance changes with changing climate – do the micro-climatic effects change as climate changes? If these data were to be used in a climate change setting it is important that a more rigorous error analysis and propagation should be performed. In the data assimilation context the corrected measurements would have realistic error variances, which would down-weight the impact of less representative observation locations, but still allow the observations to be used. If further predictors were available, the variance in the observations might be explained as a bias dependent on, for example, local site characteristics. This would allow a further bias correction in each observation and increase the information content (in a variance / entropy reduction sense) making the observation more useful for data assimilation.

Secondly, it would allow Weather Underground users to establish the bias and uncertainties in their observations, which could help identify siting problems and lead to improved instrument location practice amongst amateur weather recorders. Figure 6 shows the effect of the bias correction. Here the INTAMAP interpolation service is employed twice for each hour of Weather Underground observations - once correcting for bias and using the estimated variance (from the procedure described above), once using the raw data. As expected, the predictions at the Met Office test locations (i.e., the validation data locations which were not used in the bias estimation at all) are almost totally unbiased if the Weather Underground data is bias corrected, and the root mean square prediction error is greatly improved using the bias correction and variance estimates. Note that there remains a time-varying signal in the bias correction which indicates that, unsurprisingly, the time-stationary bias model is probably too simplistic.

We note that the approach described herein is an initial attempt to address the uncertainty in user-contributed data, and has several potentially significant limitations:

- we do not account for external variables and their influence on surface air temperature, other than elevation;
- we treat the bias and variance as being constant in time;
- we do not fully utilise the uncertainty in the predictions from the INTAMAP system in computing the bias and variance;
- *spatial* outliers are not explicitly identified or removed in this instance;
- we do not iterate the algorithm to further improve the performance.

In further work it would be possible to develop a more complete Bayesian framework for estimating the uncertainties on this user-contributed data (particularly

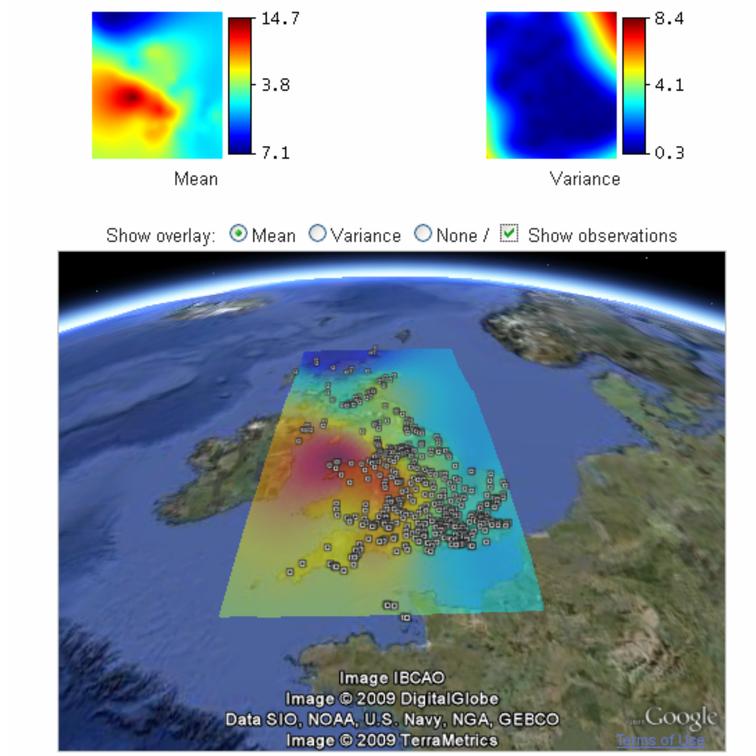

Figure 7: Using the INTAMAP system to interpolate temperature data from Weather Underground for 15:00, 27th May 2009. Note that the PSGP method was used to account for the estimated bias and variance in the observations.

where a reference data set is available), based on a spatio-temporal modelling approach, much like Kalman filtering (Kalman and Bucy, 1961). This ought to include as additional external inputs as many factors as possible that would help in explaining the variation in surface air temperatures, including elevation, distance to coast, urbanisation and a range of other micro-meterological factors.

Having estimated the bias and residual variance of the Weather Underground stations, we have exploited the ability of the PSGP method on the INTAMAP interpolation Web service to produce an interpolation for the whole of the UK. This interpolation used the Weather Underground data and accounted for the spatially varying bias and variance in the observations. The resulting interpolation, displayed on Google Earth, is shown in Figure 7. To our knowledge this is the first bias-corrected map of temperatures to be produced from user-contributed data at this level of detail.

7. Clients for using and contributing data

The framework developed here provided a basis for several interesting client applications to be developed. This section discusses two of these applications, demonstrating their operation.

7.1. Contributing data with a mobile device

The mobile client [8] was developed using Java Mobile Edition and can run on any device which supports this platform. Interpolation requests and map images are sent and received via the Internet using any available data connection supported by the device (e.g. WiFi, 3G). The client contains several features that have been simplified to allow operation on low-powered mobile devices, in addition to keeping the transferred data packets to a minimum.

The internal GPS receiver of a supported device is used to retrieve the longitude and latitude of the user. The client then downloads map images from OpenStreetMap on which the current location of the user is clearly pinpointed with a red marker.

The client can retrieve the latest temperature readings from the SOS using a simplified Web interface. This interface relies on HTTP GET requests rather than XML and returns comma separated values (x,y,z). Sacrificing some of the functionality provided by an XML interface allows a typical SOS response to be

[8]http://www.intamap.org/tryMobileClient.php

reduced in size from 2.1 Megabytes to 13 Kilobytes (a factor of 165). Only the observations that are within the boundaries of the current view are retrieved.

With a strong emphasis on user-contributed data, it is of course important to allow clients to upload information as well as access it. Therefore, users can also create and plot their own observations in addition to those retrieved from the SOS. A location can be chosen by either selecting a point on the map, using the current GPS coordinates of the device, or by entering the coordinates manually. Once the coordinates have been entered a temperature value is specified and the data is stored.

The user can submit interpolation requests to INTAMAP using the current data plotted on the screen. The client formats the data into an XML document which is then sent to a lightweight INTAMAP proxy. The response contains URLs to images representing the mean and variance of the interpolated data. These images can then be transparently placed over the existing map images.

The user can also inspect any given point on the interpolated map. A location is chosen using the cursor, and the client submits an interpolation request. The mean and variance values for that particular location are calculated by the server and returned to the client. Information regarding the chosen point is then displayed in a pop-up box.

7.2. Demonstrating INTAMAP using Google Earth

The INTAMAP project provides powerful interpolation methods through a simple XML interface. However, the overheads of the WPS interface mean it is not trivial to quickly realise the functionality of INTAMAP. For this reason a Web-based client application built around the Google Earth browser plugin was developed. The client, available at `http://www.intamap.org`, uses an HTML form to submit data to INTAMAP. Data should be formatted as comma separated x,y,z values. If the uncertainty of your data has been quantified as a standard deviation (perhaps using the technique outlined in Section 6) then this can be included as a fourth column. Google Earth works using latitude and longitude values, so if your data is projected into some coordinate system you must specify the EPSG code of that system. Clicking the 'interpolate' button sends the data to INTAMAP, resulting in two image overlays: the predicted values and the variance. The images seen in Figure 7 were generated using this Google Earth client.

8. Discussion and conclusions

This paper has demonstrated how integrating various technologies into a 'mashup' application provides a complex system, usable by the general public. Implementing a SOS interface provides a gateway into the system that can satisfy a variety of client applications. Due to the verbosity of XML payloads, simple service interfaces have been developed in parallel to enhance performance on small footprint devices. The individual components are chained, creating a collection of autonomous services which are loosely coupled to form a SOA.

UncertML provides quantification of uncertainties that arise as a result of the interpolation process. Utilising this information allows client applications to present realistic estimates which include uncertainty to answer the high-level questions posed in Section 1.

Many of the issues raised by the temperature information in this example are generic and will apply to all forms of user-contributed data: biases which can be partially explained by external variables and which differentially affect observations across time and space, a wide but heterogeneous network of sensors which sample at varying frequency, and a limited, but useful auxiliary set of reliable data which can be used to reference the uncertainty estimation. The interoperability challenges shown and solved here are also widespread; for example, the need to open up relatively impenetrable interfaces via standards-compliant mechanisms such as Sensor Observation Services, the wealth of data which can thus be exposed, and the huge value which can be added to it by relatively simple operations such as bias estimation.

As sensors become cheaper and people are increasingly connected to the Web it seems likely that user-contributed data will proliferate, and that the collection and use of this data could become a significant part of our environmental monitoring networks. Quality control and uncertainty assessment will therefore be crucial to the effective use of user-contributed data.

Acknowledgements

This work is funded by the European Commission, under the Sixth Framework Programme, by Contract 033811 with DG INFSO, action Line IST-2005-2.5.12 ICT for Environmental Risk Management. We are grateful to the Met Office and the British Atmospheric Data Centre for supplying the synoptic station data, and to the contributors to Weather Underground for free access to their data.

References

Botts, M., Robin, A., 2007. OpenGIS sensor model language (SensorML) implementation specification. OpenGIS standard 07-000, Open Geospatial Consortium Inc.
URL http://www.opengeospatial.org/standards/sensorml,[accessed_31_July_2010]

Brenning, A., Dubois, G., 2008. Towards generic real-time mapping algorithms for environmental monitoring and emergency detection. Stochastic Environmental Research and Risk Assessment 22, 601–611.

Chawla, S., Sun, P., 2006. Slom: a new measure for local spatial outliers. Knowledge Information Systems 9 (4), 412–429.

Cornford, D., Thornes, J. E., 1996. A comparison between spatial winter indices and expenditure on winter road maintenance in Scotland. International Journal of Climatology 16, 339–357.

Cox, S., 2007. Observations and Measurements – Part 1 - Observation schema. OpenGIS standard 07-022r1, Open Geospatial Consortium Inc.
URL `http://www.opengeospatial.org/standards/om,[accessed_31_July_2010]`

Erl, T., 2004. Service-Oriented Architecture: a Field Guide to Integrating XML and Web Services. Prentice Hall PTR, Upper Saddle River, NJ, USA, 580pp.

Erl, T., 2005. Service-Oriented Architecture : Concepts, Technology, and Design. Prentice Hall PTR, Upper Saddle River, NJ, USA, 792pp.

EUR, 2003. Mapping radioactivity in the environment. Report on the Spatial Interpolation Comparison (SIC1997) exercise. Dubois G., Malczewski J., de Cort M. (Eds). Technical report, Office for Official Publications of the European Communities, Luxembourg, 268pp.

EUR, 2005. Automatic mapping algorithms for routine and emergency monitoring data. Report on the Spatial Interpolation Comparison (SIC2004) exercise. Dubois G. (Ed.). Technical report, Office for Official Publications of the European Communities, Luxembourg, 152pp.

Gelman, A., Carlin, J. B., Stern, H. S., Rubin, D. B., 2003. Bayesian Data Analysis (CRC Texts in Statistical Science), 2nd Edition. Chapman and Hall, London, 696pp.

Goovaerts, P., Jacquez, G., 2004. Accounting for regional background and population size in the detection of spatial clusters and outliers using geostatistical filtering and spatial neutral models: the case of lung cancer in Long Island, New York. International Journal of Health Geographics 3, 14.

Hengl, T., Heuvelink, G. B. M., Rossiter, D. G., 2007. About regression-kriging: From equations to case studies. Computers & Geosciences 33, 1301–1315.

Heuvelink, G., 1998. Error Propagation in Environmental Modelling with GIS. Taylor and Francis, London, 150pp.

Hiemstra, P. H., Pebesma, E. J., Twenhöfel, C. J. W., Heuvelink, G. B., 2008. Automatic real-time interpolation of radiation hazards: prototype and system architecture considerations. International Journal of Spatial Data Infrastructures Research 3, 58–72.

Ingram, B., Cornford, D., Csató, L., 2008. Robust automatic mapping algorithms in a network monitoring scenario. In: Atkinson, P. M., Lloyd, C. D. (Eds.), geoENV VII: Geostatistics for Environmental Applications. Springer, Netherlands, pp. 359–370.

Kalman, R., Bucy, R., 1961. New results in linear filtering and prediction theory. Transactions of the American Society of Mechanical Engineers, Journal of Basic Engineering, Series D 83, 95–108.

Kazianka, H., Pilz, J., 2009. Bayesian spatial modeling and interpolation using copulas. In: Proceedings StatGIS09: GeoInformatics for Environmental Surveillance, Milos, Greece. URL `http://www.math.uni-klu.ac.at/stat/Tagungen/statgis/2009/StatGIS2009_Kazianka_1.pdf,[accessed_31_July_2010]`

Na, A., Priest, M., 2007. OpenGIS Sensor Observation Service (SOS) encoding standard. OpenGIS standard 06-009r6, Open Geospatial Consortium Inc. URL `http://www.opengeospatial.org/standards/sos,[accessed_31_July_2010]`

Oke, T. R., 1982. The energetic basis of the urban heat island. Quarterly Journal of the Royal Meteorological Society 108, 1–24.

Patil, G., Taillie, C., 2003. Geographic and network surveillance via scan statistics for critical area detection. Statistical Science 18, 457–465.

Pilz, J., Spock, G., 2008. Why do we need and how should we implement bayesian kriging methods. Stochastic Environmental Research and Risk Assessment 22, 621–632.

Portele, C., 2007. OpenGIS Geography Markup Language (GML) encoding standard. OpenGIS standard 07-036, Open Geospatial Consortium Inc. URL `http://www.opengeospatial.org/standards/gml,[accessed_31_July_2010]`

Rizzoli, A. E., Young, W. J., 1997. Delivering environmental decision support systems: software tools and techniques. Environmental Modelling and Software 12, 237–249.

Schut, P., 2007. OpenGIS Web Processing Service 1.0.0. OpenGIS standard 05-007r7, Open Geospatial Consortium Inc. URL `http://www.opengeospatial.org/standards/wps,[accessed_31_July_2010]`

Sharma, P., Khare, M., Chakrabarti, S. P., 1999. Application of extreme value theory for predicting violations of air quality standards for an urban road intersection. Transportation Research

Part D: Transport and Environment 4 (3), 201–216.

Shekhar, S., Lu, C.-T., Zhang, P., 2003. A unified approach to detecting spatial outliers. GeoInformatica 7, 139–166.

Vretanos, P. A., 2005. OpenGIS filter encoding implementation specification. OpenGIS standard 04-095, Open Geospatial Consortium Inc.
URL http://www.opengeospatial.org/standards/filter,[accessed_31_July_2010]

Williams, M., Cornford, D., Bastin, L., Pebesma, E., 2009. Uncertainty markup language (UncertML). OpenGIS Discussion Paper 08-122r2, Open Geospatial Consortium Inc.
URL http://portal.opengeospatial.org/files/?artifact_id=33234,[accessed_31_July_2010]

Williams, M., Cornford, D., Ingram, B., Bastin, L., Beaumont, T., Pebesma, E., Dubois, G., 2007. Supporting interoperable interpolation: the INTAMAP approach. In: Swayne, D. A., Hrebicek, J. (Eds.), Proceedings International Symposium on Environmental Software Systems. Prague.

World Meteorological Organization, 1983. Guide to Meteorological Instruments and Methods of Observation. World Meteorological Organization 8, 5th Edition, Geneva, Switzerland, 681pp.