

UncertML: an XML schema for exchanging uncertainty.

Matthew Williams¹, Dan Cornford¹, Lucy Bastin¹, Ben Ingram¹

¹School of Engineering and Applied Science, Aston University, Birmingham, UK
Tel. +44 (0) 121-204-3451
Email: williamw@aston.ac.uk

KEYWORDS: uncertainty, interoperability, WebServices, SensorML, GeoWeb

1. Introduction

Authors from Burrough (1992) to Heuvelink et al. (2007) have highlighted the importance of GIS frameworks which can handle incomplete knowledge in data inputs, in decision rules and in the geometries and attributes modelled. It is particularly important for this uncertainty to be characterised and quantified when GI data is used for spatial decision making. Despite a substantial and valuable literature on means of representing and encoding uncertainty and its propagation in GI (e.g., Hunter and Goodchild 1993; Duckham et al. 2001; Couclelis 2003), no framework yet exists to describe and communicate uncertainty in an *interoperable* way. This limits the usability of Internet resources of geospatial data, which are ever-increasing, based on specifications that provide frameworks for the 'GeoWeb' (Botts and Robin 2007; Cox 2006).

In this paper we present UncertML, an XML schema which provides a framework for describing uncertainty as it propagates through many applications, including online risk management chains. This uncertainty description ranges from simple summary statistics (e.g., mean and variance) to complex representations such as parametric, multivariate distributions at each point of a regular grid. The philosophy adopted in UncertML is that all data values are inherently uncertain, (i.e., they are random variables, rather than values with defined quality metadata).

2. Use Cases

Most data contains uncertainty, arising from sources including measurement error, observation operator error, processing/modelling errors, or corruption. Processing this uncertain data, typically through models (which typically also have errors), propagates the uncertainty. The ability to optimally utilise data relies on a complete description of any uncertainty. Below are three example use cases.

2.1 Error in sensor observations

Many of the observations used in geostatistical modelling arise from sensors, which typically exhibit specific error characteristics. With increased availability of sensor observation data on the Internet, understanding these characteristics is fundamental to meaningful processing. SensorML (Botts and Robin 2007) is an Open Geospatial Consortium (OGC)-standard language that describes the lineage of an observation through the physical processes of a sensor system. Currently, it relies on simple representations of error (e.g., tolerance to two standard deviations). UncertML provides an explicit structure to describe these (potentially complex) sensor error characteristics. Harnessing the expressive capabilities of Geography Markup Language (GML), UncertML may be used directly within a SensorML document, promoting reusability; a key ingredient for interoperability (Erl 2004).

2.2 Interpolation results

The Sensor Web Enablement (SWE) (Botts et al. 2006) is an OGC-run initiative which provides XML and Web Service standards to enable the interchange of heterogeneous sensor web information on the Internet. As published sensor networks proliferate, the importance of processing chains which can consume the published observations can only increase. An example of such processing is the production of interpolated maps from discrete point observations; a context specifically addressed by the authors as part of the INTAMAP (INTeroperability and Automated MAPping) project (Williams et al. 2007). INTAMAP will provide a fully automated interpolation Web Processing Service which can be consumed by informed clients. However, for the results to have real value they must describe the uncertainties inherent in the sensor data, as well as those introduced by the interpolation process.

2.3 Processing chains

UncertML is relevant to any processing activity involving an explicit model; e.g., a simulator which maps a set of input values to output values. As simulators imperfectly describe the real systems they represent, they must account for system uncertainties. Methods for the probabilistic treatment of simulators are being developed (O'Hagan 2006) that will in future encourage more careful specification of model error and of its propagation through processing chains. We envisage such processing chains to be composed of processing applications connected in a distributed system using Web Service interfaces.

3. Requirements

Each UncertML uncertainty type can be encoded for a variety of spatial domains (Point, Line, Polygon and Grid). Over these domains, users may request summary statistics (e.g., Mean, Variance, Quantiles), or detailed descriptions such as a full parametric distribution, or a set of n samples from that distribution. UncertML provides hard-typed parametric distributions (e.g., GaussianDistribution) to improve usability, but also allows users to construct distributions at run-time, by defining the component variables, parameters and affordances. UncertML will also enable the use of mixture models by chaining generic distributions, effectively offering a universal density model. These approaches are consistent with our priorities of flexibility, usability, performance, portability and extensibility.

4. UncertML design

4.1 GML inheritance and common properties

GML is an established XML framework for describing the real world (Portele 2007). This framework can be extended through GML application schema to allow detailed description of real world phenomena while conforming to a standardised structure. This flexibility has led to widespread adoption of GML. UncertML utilises the benefits of GML by ensuring base types inherit from the GML *AbstractFeatureType*, via the UncertML abstract base *Feature*, *Uncertainty* (Figure 1). This inheritance allows UncertML types to be seamlessly integrated into any current XML language that understands the GML feature model.

Each UncertML uncertainty type has a set of common properties: *variables*, *parameters*, an *encoding block* that stores the associated values and a number of *affordances* (permitted operations for the distribution, with a description of required inputs and outputs, e.g., *getMoments*).

4.2 Relation to SWE Common

The Sensor Web Enablement initiative is very relevant to UncertML, thus careful attention has been paid to existing SWE technologies. For example, the SWE Common namespace (Cox 2006; Botts et al. 2006) provides a dictionary with common variable types (e.g., temperature, atmospheric pressure) which can be used with UncertML.

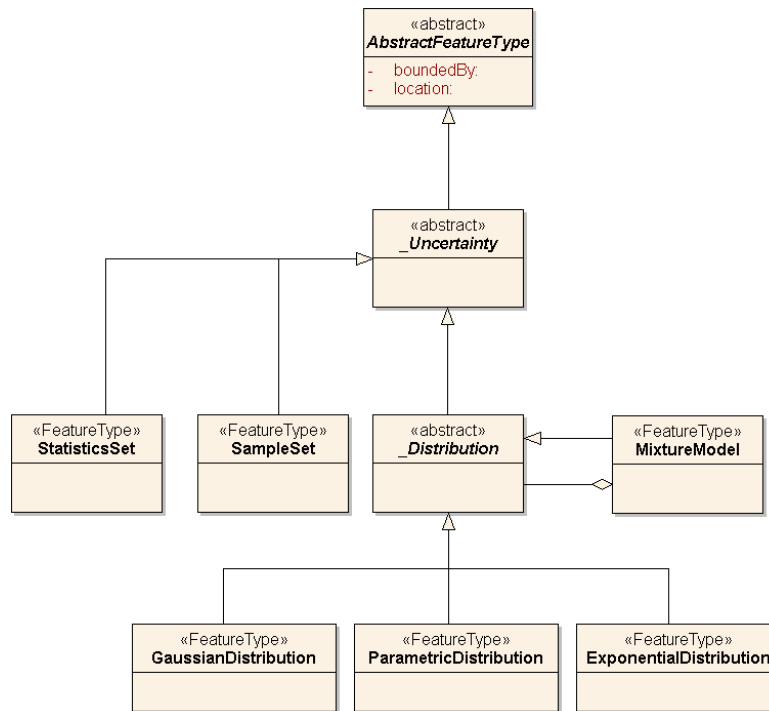


Figure 1. UML diagram displaying the inheritance hierarchy of UncertML.

5. Discussion and future directions

UncertML is currently in version 1, which has flexible support for continuous random variables. This representation is suitable for many environmental variables; however future releases will see added support for discrete and categorical random variables and possibly fuzzy representations. We note that UncertML is designed to encode *any* type of uncertainty. Whilst the above use cases highlight application to attribute uncertainty, location uncertainty is plainly an issue for geospatial data. We envisage that future iterations of GML might replace coordinates in the geometry types with *Uncertainty* types.

In the rare situation of complete knowledge the (Dirac) delta distribution can support deterministically known values. This has the advantage that one must explicitly state that the value is known with certainty (a strong statement). If one merely wishes to provide a value, then a mean, mode or median can be given without a dispersion measure, making the user aware that this is only a characterisation of the distribution's first moment, which should be treated with due care.

An example application in which UncertML is used to inform interpolations of European raditation data, via a Web Processing Service, is presented. In further work we will develop an extensive API to support UncertML, which we hope will form the basis for many schemata dealing with uncertain objects or phenomena. We also plan to extend the work to conditional

distributions and stochastic processes, providing a mechanism to communicate uncertain models as well as results.

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.

References

- Botts M, Percivall G, Reed C, and Davidson J (2006) OGC Sensor Web Enablement: Overview and High Level Architecture. *Technical report, OGC*.
- Botts M and Robin A (2007) OpenGIS Sensor Model Language (SensorML) Implementation Specification.
- Burrough P (1992) Development of intelligent geographical information systems. *International Journal of Geographical Information Science* **6** pp497-513.
- Couclelis H (2003) The Certainty of Uncertainty: GIS and the Limits of Geographic Knowledge. *Transactions in GIS* **7**(2) pp165–175.
- Cox S (2006) Observations and Measurements. *Technical report 05-087, OGC*.
- Duckham M, Mason K, Stell J, and Worboys M (2001) A formal approach to imperfection in geographic information. *Computers, Environment and Urban Systems* **25** pp89–103
- Erl T (2004) *Service-Oriented Architecture: A Field Guide to Integrating XML and Web Services*. Prentice Hall, USA.
- Heuvelink G, Brown J and van Loon E (2007) A probabilistic framework for representing and simulating uncertain environmental variables. *International Journal of Geographical Information Science* **21**(5) pp1-11.
- Hunter G and Goodchild M (1993) Managing Uncertainty in Spatial Databases: Putting Theory into Practice. *Journal of the Urban and Regional Information Systems Association* **5**(2) pp55-62.
- O'Hagan A (2006) Bayesian analysis of computer code outputs: a tutorial. *Reliability Engineering and System Safety*, **91** pp1290-1300.
- Portele C (2007) OpenGIS Geography Markup Language (GML) Encoding Standard. *Technical report, OGC*.
- Williams M, Cornford D, Ingram B, Bastin L, Beaumont A, Pebesma E, and Dubois G (2007) Supporting interoperable interpolation: the INTAMAP approach. In *International Symposium on Environmental Software Systems*, Prague, May 2007.

Biographies

Matt Williams is a PhD student at Aston University looking at describing uncertainty using an interoperable framework (UncertML). He also maintains an interest in geospatial Web Service specifications including WFS, WMS & WPS.

Dan Cornford is a senior lecturer in Computer Science at Aston University. After a PhD in Spatial Statistics and Applied Meteorology at Birmingham University he moved to work in machine learning, focussing on application of principled probabilistic methods to complex systems, with particular emphasis on data assimilation and inference in stochastic processes.

Lucy Bastin is a Lecturer in GIS at Aston University. After a PhD on urban plant metapopulations, and research into fuzzy classification / uncertainty visualisation at Leicester University, she worked as a GIS software developer. Her research interests include spatial epidemiology and automatic interpolation via WPS.

Ben Ingram is a Research Fellow at Aston University working on the INTAMAP project. His interests lie in the development of geostatistical methods for treating large data sets, with emphasis on parallel treatments and sequential approaches.