Propagation and Provenance of Probabilistic and Interval Uncertainty in Cyberinfrastructure-Related Data Processing

P. Pinheiro da Silva[1], **A. Velasco**[2], **M. Ceberio**[1], **C. Servin**[1], **M. Averill**[2], **L. Longpré**[1], and **V. Kreinovich**[1]

Departments of [1]Computer Science and [2]Geological Sciences
University of Texas, El Paso, TX 79968, USA, vladik@utep.edu

Abstract

In the past, communications were much slower than computations. As a result, researchers and practitioners collected different data into huge databases located at a single locations such as NASA and US Geological Survey. At present, communications are so much faster that it is possible to keep different databases at different locations, and automatically select, transform, and collect relevant data when necessary. The corresponding cyberinfrastructure is actively used in many applications; what is mostly lacking is the *propagation* of uncertainty via this cyberinfrastructure. In the first part of our talk, we describe algorithms for this uncertainty propagation.

When the resulting accuracy is too low, it is desirable to produce the *provenance* of this inaccuracy: to find out which data points contributed most to it, and how an improved accuracy of these data points will improve the accuracy of the result. For probabilistic uncertainty, it is mainly the question of fast algorithmic implementation of the known formulas. For interval uncertainty, we had to also derive the formulas.

The above results mainly deal either with the *probabilistic* situations, when we either know the probability distributions of different measurement errors (and different errors are independent), or with *interval* situations, when we only know the upper bounds on the measurement errors. Probabilistic estimates tend to *underestimate* the resulting error—since in reality, different measurement errors are correlated (e.g., they have the same systematic error components). Interval estimates tend to *overestimate* because they are based on—often unrealistic—worst-case scenarios. It is thus desirable to combine these estimates to get more realistic error bounds. We describe several such combination methods, their mathematical justifications, and their successful use in processing geospatial data.

References

[1] Longpré, L., and Kreinovich, V.: "How to efficiently process uncertainty within a cyberinfrastructure without sacrificing privacy and confidentiality", In: Nedjah, N., Abraham, A., and de Macedo Mourelle, L. (Eds.): *Computational Intelligence in Information Assurance and Security*, Springer-Verlag, 2007, pp. 155–173.