28*th* IUGG Conference on Mathematical Geophysics, June 7-11, 2010, Pisa, Italy
Session 7: Quantifying the uncertainty in Earth systems

Comparing the performance of different expert elicitation models using a cross-validation technique

Franco Flandoli[1], Enrico Giorgi[2], Willy P. Aspinall[3], Augusto Neri[2]

[1] *Dip.to di Matematica Applicata, Università di Pisa, Pisa, Italy*
[2] *Istituto Nazionale di Geofisica e Vulcanologia, Sezione di Pisa, Pisa, Italy*
[3] *Dept. of Earth Sciences, University of Bristol, and Aspinall & Associates, Tisbury, UK*

Keywords: *expert elicitation; cross-validation*

The problem of ranking and weighting experts' performances when their judgments are being elicited for decision support is considered. Different methods exist to score experts such as the Cooke Classical model, Equal Weights, and Single Best Expert models. The purpose of our research is twofold: 1) to introduce a new model for providing optimal point-wise estimates, named the Expected Relative Frequency (ERF) model, and 2) to use a cross-validation technique to compare different scoring models' performances on real data. The new ERF model gives positive value to an expert who provides central values close to the known values. In contrast to a self-referencing conventional comparison of models, a cross-validation technique is adopted here. A single round of cross-validation involves partitioning a sample of seed items into two complementary subsets, computing expert weights from one subset (called the training set), and validating the result on the other subset (called the validation set or testing set). To control variability, average over multiple rounds of cross-validation is performed. Here, the cross-validation analysis is applied to five elicitation datasets from different expert judgment problems. For the comparison step of the cross-validation method it is necessary to introduce the concept of reward and the Cooke Classical model, the new ERF model, Equal Weights and the single Best Expert performances are compared with respect to three specific reward indices, called Calibration, Informativeness, and Expected Accuracy. These measure both an expert's ability to assess uncertainty and his accuracy in point-wise value estimation. A detailed comparison of average rewards and the probability of success of one method over another is given for the case studies. Results show that, for any specific reward index, there is only a limited probability that one method is better than another and, even in the average, no single method emerges as best for all the forms of reward. However, certain tendencies are observed: the Cooke Classical model is generally most suitable for assessing uncertainties, while the new ERF model results are preferable if the criterion is central value estimation accuracy, alone; the performance of the Equal Weights model is generally quite good but rarely, if ever, optimal, and Best Single Expert and individual single experts are ordinarily outperformed by any of the models based on pooled opinions.