Hydrol. Earth Syst. Sci. Discuss., 5, S2122–S2124, 2008
www.hydrol-earth-syst-sci-discuss.net/5/S2122/2008/
© Author(s) 2008. This work is distributed under
the Creative Commons Attribute 3.0 License.

Hydrology and Earth System Sciences Discussions

Interactive comment on "Bayesian objective classification of extreme UK daily rainfall for flood risk applications" *by* M. A. Little et al.

Anonymous Referee #2

Received and published: 9 December 2008

1. General comments

The paper addresses relevant scientific questions within the scope of the journal. It is well written, interesting and clear even to no experts in precipitation observations. However, in their research authors have employed statistical methods that are not suitable for the type of data to be analysed.

2. Specific comments

The main scientific issue to be discussed and solved by the authors is the suitability of the adopted statistical methods for their data.

In their paper, authors use Bayesian k-means clustering to analyse vectors composed

of 0-1 entries, that is, to analyse binary data. As authors clearly state at page 3044, in the basic k-means clustering algorithm, all data vectors belonging to a given cluster are implicitly assumed to follow the same spherical multivariate Gaussian distribution. That is, use of a sum of squares clustering criterion is equivalent to estimate a maximum likelihood model with spherical multivariate Gaussian distributions (Gordon 1999, pp. 65-66).

It is well known that the Gaussian model is suitable for modelling continuous data. When one has to deal with binary data, generally the Bernoulli distribution is the preferred probabilistic model. Thus, classical and Bayesian k-means clustering are not proper choices for classifying binary data.

Whenever data of this type have to be analysed through a clustering method, the most suitable solutions can be obtained using clustering methods based on latent class models. In a latent class model, the feature vector consists of p binary variables (exactly as for the UK rainfall data analysed in the paper). In the simplest latent class model, variables are assumed conditionally independent, and each variable is assumed to follow a Bernoulli distribution. Further models may be used to deal with more complex situations (see, for example, McLachlan and Peel 2000, p. 166; Magidson and Vermunt 2001).

Clustering methods based on latent class models allow to solve exactly the same problems addressed in the paper (finding a small set of representative vectors, a unique assignment to one of these vectors for each observed data vector, and the optimal number of clusters using model selection criteria such as the BIC). But they make use of models that are more suitable for the type of data to be analysed than the Gaussian one. Furthermore, through these methods it is possible both to check the validity of each assumption introduced in a model, and to choose the latent class model that best fits the data.

In summary, in order to improve the methodological exactitude of the paper, I strongly

HESSD

5, S2122–S2124, 2008

Interactive Comment

suggest to analyse the rainfall data through clustering methods based on latent class models.

3. Technical corrections

Page 3041: the meaning of MSLP is missing.

References

- Gordon A.D. (1999). *Classification*, second edition. Chapman & Hall/CRC, Boca Raton, Florida.

- Magidson, J., Vermunt, J.K. (2001). Latent class factor and cluster models, biplots and related graphical displays. *Sociological Methodology*. Vol. 31, pages 223-264.

- McLachlan, G., Peel, D. (2000). *Finite mixture models*. John Wiley & Sons, New York.

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., 5, 3033, 2008.