Hydrol. Earth Syst. Sci. Discuss., 8, C2541–C2547, 2011
www.hydrol-earth-syst-sci-discuss.net/8/C2541/2011/

Hydrology and Earth System Sciences Discussions

Interactive comment on "Catchment classification by runoff behaviour with self-organizing maps (SOM)" *by* R. Ley et al.

R. Ley et al.

leyrita@uni-trier.de

Received and published: 30 June 2011

The Authors would like to thank the reviewer for the comments and corrections, which will help us to improve the paper.

C: We may expect more technical advices for hydrologists who would like to apply SOM. There are several parameters and we do not know how they were fixed. Drawbacks and advantages of SOM could be also listed (it is not clear why SOM should be used instead of one classical clustering approach; in particular there is still one question about the number of clusters to be considered)

A: A detailed technical advice is beyond the scope of this paper and would exceed the extend of this paper. There are many detailed technical advices in literature, e.g. Kohonen (2001), Kaski (1997) and Vesanto et al. (2000), who provide technical advice for SOM trained with their toolbox for Matlab, used for this study. Properties that distinguish SOM from other data mining tools are that it is numerical, non-parametric, no assumptions about data distribution or shapes of the clusters are necessary, they are insensitive against a small portion of missing data, topology preserving, they represent graded relationships and allow visualizations of structures in high dimensional data sets (Kaski, 1997). Furthermore, SOMs trained with the batch algorithm are reproducible and provide insights and interpretations of resulting clusters. There are only a few parameters to fix before training a SOM: the distance measure, a neighbourhood function, the way of normalization and the size of the SOM, i.e. the number and composition of neurons. A drawback of clustering with SOMs in the way we do it in this study is that we need hierarchical clustering to define cluster borders. This is only necessary, if we train the SOM with more neurons as clusters and cluster border are not easy to see on the u-matrix, as in our case. We define the number of clusters to be considered by analysing the dendrogram of the hierarchical clustering. So the number of clusters is a result of the cluster process, we don't have to predefine a number of cluster. We will concentrate advances of SOM in the introduction and add on page 3050, line 25: "... preserved as much as possible. Properties that distinguish SOM from other data mining tools are that it is numerical, non-parametric, insensitive against a small portion of missing data, it represents graded relationships, provides visualizations of structures in high dimensional data sets, needs no assumptions about data distribution or cluster shapes, and may find unexpected structures in the data (Kaski, 1997). There are only a few parameters to fix before training a SOM: the distance measure, a neighbourhood function, normalization and the size of the SOM." In section 2.3.4 (SOM) we'll eliminate on page 3057/58 line 25ff: "Some properties that ... in the data (Kaski, 1997)."

C: P 3051: The authors should justify the choice of the signature indices and, in connection to these variables, the reasons for looking for clusters. Indeed the relevance of a classification depends on the final objective and on the use of the analysis. In

HESSD

8, C2541–C2547, 2011

this paper we may suspect that SOM are established to study spatial variability of flood characteristics at regional scale.

A: The aim of the paper is to classify catchments at a regional scale by their response behaviour with a view to regionalisation and to apply SOMs for catchment classification. Clustering of catchments by their response behaviour is the first step to do this. Catchment runoff response behaviour covers a wide range of aspects, described here by event runoff coefficients (ERCs) and flow duration curves (FDCs). From this data we calculated a huge amount of indices, which describe, seen by themselves, important aspects of runoff behaviour. A correlation analysis showed very high correlations between many of these indices. High correlated indices (spearman's rank coefficient > 0.8) don't bring new insights in the analysis and therefore were excluded from analysis. On the other hand we considered an even distribution of indices with respect to season, high and low flow and the importance of indices. An explanation of the selection process will be added to the revised version.

C: P 3055: Why was the method suggested by Merz et al. (2006) modified? Parameters for the extraction was fixed (peak flow with direct runoff > 2 * baseflow and a maximal time window of 24 hours, threshold fixed to 5 mm to identify rainfall event). How were these parameters defined? For large basin (> 1000 km2) the duration may exceed 24 hours. Is there any bias in the selection (underestimation of the related runoff coefficients)? In addition there is certainly a time lag between the rainfall event and the answer of the basin. The time windows are not identical for runoff and rainfall. How are they defined to ensure consistency?

A: The method used to calculate event runoff coefficients was developed for catchments in whole Austria, with a much larger gradient of landscapes than in Rhineland-Palatinate. Some parameters of this method were defined by an iterative process to fit best for the studied catchments. For catchments in this study we modified these parameters to optimize them for catchments in Rhineland Palatinate. To verify runoff coefficients, calculated with this method, we compared them with manually calculated coef-

HESSD
8, C2541–C2547, 2011

ficients of different catchments. Because of different methods for baseflow-separation automatically calculated runoff coefficients are often slightly higher than manually calculated (linear baseflow-separation). Mean difference between manually and automatically calculated runoff coefficients is about 0.05. The time window of 24 h is used only for identification of peak flows as the highest flow in this period. The duration of an event is calculated by minimizing the direct flow before and after a peak flow (start and end of an event) and considers different travel time ranges of the catchments. The time window for rainfall was defined as time before start and end of an event with help of a characteristic time scale for each event. In the revised version of the paper we will describe this more clearly.

C: Excluding events due to snow melt is relevant to study link between rainfall-runoff relations. The selection does not consider as well events with low runoff coefficients in late summer (due to drought conditions). Why is it important to exclude them since these events describe the basin behavior and the dependency to actual conditions (objective mentioned P 3056)?

A: We don't exclude events in late summer with low runoff coefficients. In our data there are many very low runoff coefficients, especially in summer (Fig. 2c and 3a). To improve the automatically done event selection we exclude events with very low precipitation less than 5 mm and events with very low runoff peaks, lower than the seasonal mean runoff. Frequently these very small "events" are not separated accurate or caused only by low fluctuations of discharge. This exclusion affects "events" in summer and winter.

C: P 3058: Results are presented. The number of neurons has been chosen arbitrary. The authors should develop this part: on which basis was this number fixed? What do the authors mean by "the best choice"? What are the consequences of choosing a high number of neurons?

A: One of the few parameters to fix before training a SOM is the number of neurons.

HESSD

8, C2541–C2547, 2011

Unfortunately, there is not one “true” number of neurons, but many rules of thumbs in literature or software. For comparatively small data sets, as in this study, the relevance of this number is important for interpretation and depends on the aim of a SOM. If we want to cluster with a predefined number of clusters, a SOM can be trained with this number of neurons. But the analytical possibilities of such a SOM are small and quantisation error is high. Choosing a high number of neurons will reduce the quantization error of the SOM but clustering will be vague and not well interpretable. To get a graded clustering we chose SOM sizes smaller than the number of input vectors but larger than an estimated number of clusters. Result of a selection process between SOMs of different sizes is a SOM with a medium quantization error, no topologic error, many evident levels of clustering and much information to analyse clusters - the best choice for this study.

C: P 3060 and P 3080: Labels for Fig.4b and Fig.4c are inverted.

A: This will be corrected in the revised version.

C: P 3062: I am not familiar with SOM. Fig 4a shows 30 neurons located in a matrix with 11*9 units; Fig 4c shows 23 neurons located in a matrix with 6*5 units. I do not understand why the authors consider 25 neurons and the links between a and c since the sizes are different.

A: The U-matrix in Fig. 4a shows the 30 neurons of the SOM, indicated by numbers, and distances between two neurons as units between them, without numbers. Neurons and units of distances together built a 9*11 matrix while the neurons itself built a 5*6 matrix. Fig. 4c (neuron label) shows only the 30 neurons of the SOM and catchments assigned to neurons (BMU1), indicated by their catchment ID. With this visualisation we can assign catchments to clusters. Not all neurons are labelled with catchments, 5 neurons are interpolative, empty neurons. The 25 neurons labelled with up to four catchments is a first level of clustering. All visualisations of SOMs are linked by position: in each figure, a certain position corresponds to the same neuron.

HESSD

8, C2541–C2547, 2011

Interactive Comment

C: The range of Fig 5 is not the same as the range of Fig 4a.

A: This will be corrected in the revised version.

C: P 3064: Is it possible to locate the three borderline catchments on the map (Fig 6)? Please give more details on these basins (e.g. are they the most highly urbanized basins?).

A: The three borderline catchments will be located on the map in Fig. 6 by a striped signature, indicating the first and second cluster of these catchments. The three borderline catchments (catchments 1, 26, and 43) show no exceptional values for most of the physiographic and climatic catchment properties, urbanization is about 5%. They all have a more or less round shape and are above average flat; properties also apply to other not borderline catchments. Exceptional values for catchment properties of these catchments are: Catchment 1 shows a high mean field capacity, very low mean annual precipitation and a high amount of arable land. About 55 % of catchment 26 are forested and catchment 43 is comparatively flat with a high amount of grassland.

C: There is an overlap of 67% which corresponds to 30 basins among the 45. The overlap reaches 80% (36 basins among the 45) considering these catchments. The increase of 13% is not consistent with the number of borderline catchments (= 3 , P 3064).

A: There are 3 borderline catchments of clusters based on catchment response behaviour and 5 borderline catchments of clusters based on physical catchment properties. This information is missing in the paper. We will add this information to the end of Sect. 3.3 (Physical catchment properties) and also to Table 3 as numbers in brackets. Page 3067, line 25ff will be rewritten to: "In Sect. 3.2 we defined borderline catchments as catchments on the border of two catchments, belonging to both catchments. We identify 3 borderline catchments in clusters based on runoff behaviour and 5 borderline cases in clusters based on physical catchment properties. If we consider these 8 borderline catchments to their second cluster, we can improve the overlap between

HESSD

8, C2541–C2547, 2011

Interactive Comment

Full Screen / Esc

Printer-friendly Version

Interactive Discussion

Discussion Paper

the two pools of clusters from 67% to 84% (Table 3).”

C: P 3066-3067: In Table 3 the correspondence is established between clusters based on the number of shared basins. Is this criterion sufficient especially when the number is similar for different pairs of clusters?

A: For most of the clusters the number of shared basins allows a clear assignment with more than 4 shared catchments. Only clusters B and IV not clearly differentiate response behaviour to clusters by physical catchment properties and vice versa. Both share a small number of catchments to 4 clusters. Furthermore, clusters B and IV cover medium response behaviour and medium catchment properties, i.e. they show reasonable catchment properties. Considering 8 borderline catchments by assigning them to their second best cluster, cluster B shares 3 catchments with cluster IV (Table 3). Therefore we assign cluster B to cluster IV.

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., 8, 3047, 2011.

HESSD

8, C2541–C2547, 2011

Interactive Comment