Methodology

Robert Schlegel

27 March 2017

Question

Given the many possible ways to group (cluster) synoptic climate data, which would be the most appropriate for this investigation, and why?

Background

Initially we had decided upon the use of self organizing maps (SOMs) as this is one method that has been employed in climate science for similar applications around the world (e.g. Cavazos (2000), Hewitson and Crane (2002), Morioka et al. (2010)), and within South Africa (Sang et al., 2008). One of Sarah's students, Peter Gibson (Gibson, Perkins-Kirkpatrick, et al. (2016), Gibson, Uotila, et al. (2016)), has used SOMs for climate oriented inquiry, too. The calculation of SOMs in R turned out not to be as difficult as anticipated due to a package (Wehrens and Buydens, 2007) with which one could quite easily make the calculations for synoptic climatologies. However, after a literature review on the topic I began to accept that perhaps SOMs were not the best choice for our analysis. And that clustering of some other sort would be better. K-means clustering was the alternative to SOMs proposed at the outset of this branch of the research and there is certainly quite a bit of literature to draw on for this (e.g. Corte-Real et al. (1998), Burrough et al. (2001), Kumar et al. (2011)).

Rationale

The reason I finally came to this conclusion is that one of the primary characteristics that set SOMs apart from other clustering techniques is that the nodes (clusters) in a SOM are spatially reliant on each other. Meaning that a SOM excels at visualizing/ quantifying which nodes (clusters) are most similar to each other, and why. We are not interested in how similar air-sea states may be with one another during coastal marine heatwaves (MHWs) but rather how dissimilar they may be. It is not necessary to identify the gradient that may exist between these synoptic states during events. Therefore it is better to use a clustering technique that is not designed to quantify the gradient between what we endeavour to find as distinctly separate groupings. So if SOMs are no longer to be used, then what shall replace them? K-means clustering, MDS (Multi-dimensional scaling), PCA (Principal component analysis), and hierarchical clustering are the techniques that come to mind most quickly as answers to this question. I think we can rule out PCA because the goal of this paper is not to quantify how much of a particular variable(s) may explain the synoptic air-sea states we observe, but to classify the different synoptic air-sea states during events into discrete clusters. Additionally, I think we can also rule out MDS because we are not interested in the distance between synoptic states (and by extension, clusters of these synoptic states) caused by which variables (pixels in this case), but primarily the clusters themselves. That leaves us with two choices as I see it. I think there are pros and cons for K-means clustering and hierarchical clustering. Hierarchical clustering is appealing to me because the dendrogram this method produces is visually very clear, but it does not produce clusters in the more generally sense. Literature for the use of hierarchical clustering in climate science exists (e.g. Unal et al. (2003)), but not in the same apparent abundance as K-means and increasingly in SOMs.

Proposed methodology

This leads us to our talking points for the Skype session tomorrow (2017/03/27). Regardless of which clustering technique we decide upon, the distinction between (SOM) nodes proposed in Johnson (2013) that only as many nodes (clusters) should be included as are significantly dissimilar from one another is appealing to me. I think we should use that to determine our number of clusters. Furthermore, both Eric and AJ have noted the necessity that the metrics for the MHWs for each node be calculated as well so as to allow for a more meaningful classification/ interpretations of the cluster results. I shall do so. Ramos (2001) looked at classifying areas of the Mediterranean based on rainfall data and used a synthesis of K-means and hierarchical clustering. Ambroise et al. (2000) applied hierarchical clustering to the results of SOMs, which seems excessive. Regardless, my point is that we don't have to use just one technique as the precedent for a synthesis of clustering methods exists.

Whilst delving further into the philosophy of clustering I found this paper: Jain (2010). It was remarkably useful in clearing up some common terms that I've seen used as well as giving some advice on why to use which clustering or classification methods. One point that is stressed is that partitional algorithms (K-means) are considered superior to hierarchical algorithms because they do not progressively divide (or agglomerate) data into cluster. Instead they create clusters simultaneously considering all of the available data. Another way of dealing with high dimensional climate data (like synoptic images) is via fuzzy c-means clustering, which allows data points to exist within multiple clusters (e.g. McBratney and Moore (1985)). Either way, due to the highly dimensional nature of these data, it is necessary to use an algorithm that is able to find groupings of points within the overall collection of points for each synoptic image. Meaning, the clustering algorithm used must be able to see localised patterns apart from the overall noise of the air-sea state that likely does not set any one day of the year apart from any other. CLIQUE is one such algorithm (Road and Jose, 1998). However, there is no one best choice, and often many algorithms should be tried as some may fit better with the data in question than others. Jain (2010) recommends trying a few algorithms and picking the one that produces the most distinct clusters. Furthermore, they reiterate that clustering is an exploratory tool, and only suggests hypotheses.

Because the underlying variability in the Agulhas/ Benguela systems is generally known, we are not necessarily looking for anomalous signals, even though we are looking at air-sea states during coastal MHWs. Instead we want to cluster the signals that may be found to occur during MHWs into their own clusters. And to do so as delicately as possible, removing as much of the massive amount of noise that needs to be sifted through. Which is why one thing that needs to still be investigated is what is the effect on clustering when the study area is made smaller. Specifically, removing the Agulhas retroflection from the synoptic images. I don't think that the Agulhas retroflection above the Southern Ocean is going to be having as direct of an impact on the coast of South Africa as the adjacent air-sea pixels however, the retroflection always features very heavily in all of the nodes (SOMS) and hierarchical clusters I have created thus far.

Finally, it has also been said that cluster analysis should be run on the air and sea data independent of each other, as well as dependent, to see how the results differ. It may be that the atmosphere plays a role in the formation of coastal MHWs much less often than the sea, and by isolating the variables, more telling patterns in either may emerge.

Summary

- Self-organizing maps (SOMs) is not the best clustering technique to use
- Other techniques such as MDS and PCA are likewise more sophisticated than is appropriate
- Therefore K-means or hierarchical clustering would be the best choice
- The quality of the results I have produced with the different clustering techniques thus far seemed best with hierarchical clustering, but Jain (2010) makes a very compelling argument for the use of K-means clustering
- That being said, not all K-means clustering algorithms are created equal and the appropriate one must be determined

- Because of this, K-means clustering appears to be the winning choice, but a synthesis could be performed
- The number of clusters used should be determined by using the maximum number of groups that are still statistically significantly different from one another (Johnson, 2013)
- The metrics for each MHW clustered into different groups needs to be shown in the results for better classification of groups
- The effect that the extent of the study area has on clustering needs to be investigated by shrinking the range of the lon/ lat in a meaningful way
- The effect of data resolution (e.g. 0.5 degree pixels) on clustering has already been investigated for daily climatologies, but must still be performed for the synoptic data during MHWs
- A good idea would be to run whichever cluster analysis we decide upon the daily synoptic climatology data and the synoptic data during MHWs together to see if they are clustered differently

References

Ambroise, C., S??ze, G., Badran, F., Thiria, S., 2000. Hierarchical clustering of self-organizing maps for cloud classification. Neurocomputing 30, 47–52. doi:10.1016/S0925-2312(99)00141-1

Burrough, P.A., Wilson, J.P., Van Gaans, P.F.M., Hansen, A.J., 2001. Fuzzy k-means classification of topo-climatic data as an aid to forest mapping in the Greater Yellowstone Area, USA. Landscape Ecology 16, 523–546. doi:10.1023/A:1013167712622

Cavazos, T., 2000. Using self-organizing maps to investigate extreme climate events: An application to wintertime precipitation in the Balkans. Journal of Climate 13, 1718–1732. doi:10.1175/1520-0442(2000)013<1718:USOMTI>2.0.CO;2

Corte-Real, J., Qian, B., Xu, H., 1998. Regional climate change in Portugal: precipitation variability associated with large-scale atmospheric circulation. International Journal of Climatology 18, 619–635. doi:10.1002/(SICI)1097-0088(199805)18:6<619::AID-JOC271>3.0.CO;2-T

Gibson, P.B., Perkins-Kirkpatrick, S.E., Renwick, J.A., 2016. Projected changes in synoptic weather patterns over New Zealand examined through self-organizing maps. International Journal of Climatology 36, 3934–3948. doi:10.1002/joc.4604

Gibson, P.B., Uotila, P., Perkins-Kirkpatrick, S.E., Alexander, L.V., Pitman, A.J., 2016. Evaluating synoptic systems in the CMIP5 climate models over the Australian region. Climate Dynamics 47, 2235–2251. doi:10.1007/s00382-015-2961-y

Hewitson, B.C., Crane, R.G., 2002. Self-organizing maps: Applications to synoptic climatology. Climate Research 22, 13–26. doi:10.3354/cr022013

Jain, A.K., 2010. Data clustering: 50 years beyond K-means. Pattern Recognition Letters 31, 651–666. doi:10.1016/j.patrec.2009.09.011

Johnson, N.C., 2013. How many enso flavors can we distinguish? Journal of Climate 26, 4816–4827. doi:10.1175/JCLI-D-12-00649.1

Kumar, J., Mills, R.T., Hoffman, F.M., Hargrove, W.W., 2011. Parallel k-means clustering for quantitative ecoregion delineation using large data sets, in: Procedia Computer Science. pp. 1602–1611. doi:10.1016/j.procs.2011.04.173

McBratney, A.B., Moore, A.W., 1985. Application of fuzzy sets to climatic classification. Agricultural and Forest Meteorology 35, 165–185. doi:10.1016/0168-1923(85)90082-6

Morioka, Y., Tozuka, T., Yamagata, T., 2010. Climate variability in the southern Indian Ocean as revealed by self-organizing maps. Climate Dynamics 35, 1075–1088. doi:10.1007/s00382-010-0843-x

Ramos, M.C., 2001. Divisive and hierarchical clustering techniques to analyse variability of rainfall distribution patterns in a Mediterranean region. Atmospheric Research 57, 123–138. doi:10.1016/S0169-8095(01)00065-5

Road, H., Jose, S., 1998. Automatic Subspace Clustering Mining of High Dimensional Applications for Data. Proceedings of the 1998 ACM SIGMOD international conference on Management of data 27, 94–105. doi:10.1145/276305.276314

Sang, H., Gelfand, A.E., Lennard, C., Hegerl, G., Hewitson, B., 2008. Interpreting self-organizing maps through space-time data models. Annals of Applied Statistics 2, 1194–1216. doi:10.1214/08-AOAS174

Unal, Y., Kindap, T., Karaca, M., 2003. Redefining the climate zones of Turkey using cluster analysis. International Journal of Climatology 23, 1045–1055. doi:10.1002/joc.910

Wehrens, R., Buydens, L.M.C., 2007. Self- and super-organizing maps in R: The kohonen package. Journal of Statistical Software 21, 1–19. doi:10.18637/jss.v021.i05