Discovering relationships in climate-vegetation dynamics using satellite data

Christina Papagiannopoulou CHRISTINA.PAPAGIANNOPOULOU@UGENT.BE
Willem Waegeman WILLEM.WAEGEMAN@UGENT.BE
Depart. of Mathematical modelling, Statistics and Bioinformatics, Ghent university, Belgium

Mathieu Depoorter MATHIEU.DEPOORTER@UGENT.BE
Niko E. C. Verhoest NIKO.VERHOEST@UGENT.BE
Laboratory of Hydrology and Water Management, Ghent University, Belgium

Wouter A. Dorigo WOUTER.DORIGO@GEO.TUWIEN.AC.AT
Depart. of Geodesy and Geo-Information, Vienna University of Technology, Austria

Diego G. Miralles DIEGO.MIRALLES@UGENT.BE
Depart. of Earth Sciences, VU University Amsterdam, the Netherlands
Laboratory of Hydrology and Water Management, Ghent University, Belgium

Keywords: time series analysis, feature construction, shapelets

Abstract

Advances in satellite Earth observation have resulted in the development of consistent global historical records of environmental and climatic variables, forming enormous amounts of multivariate time series. In this work we present a novel machine learning framework for detecting relationships between climatic time series and vegetation indices. Our pipeline consists of several components, including data fusion from various databases, time series decomposition techniques, feature construction methods and predictive modeling. Experimental results indicate that with this pipeline it is possible to detect patterns that express relationships in large-scale climate data.

Earth Observation (EO) satellite data provide a wealth of information about the dynamics of our planet in recent decades. Independent sensors on different platforms monitor vegetation, soils, oceans and atmosphere, collecting optical, thermal, or gravimetry information (e.g. (Su et al., 2011)). Composite records of environmental and climatic variables now span up to 30 years, enabling the study of climate-vegetation interactions over multi-decadal scales. Such records can be interpreted as long multivariate time series with different spatial and temporal resolutions. Due to their volume and complexity, the resulting datasets pose important challenges with respect to data processing and data analysis, leading to a need for developing novel methods.

Vegetation is a major player in the global climate system by affecting the water, the energy and the carbon cycles. Plants alter climate through the transfer of water vapor from land to atmosphere, direct effects on the surface net radiation, exchange of carbon dioxide with the atmosphere, or changes in roughness length affecting wind speed and direction (Bonan, 2008; McPherson et al., 2007). Given the crucial role of vegetation in climate, and the influence of climate on vegetation dynamics, understanding how vegetation will respond to projected climatic changes is crucial to narrow down the uncertainty in the predictions of global warming. A first and necessary step, however, is to investigate the sensitivity of vegetation to past-time climate variability.

This abstract is based on (Papagiannopoulou et al., 2016a) which is under review. In the original article, we present a novel framework for finding climatic drivers that affect vegetation. We consider a data-

Appearing in *Proceedings of Benelearn 2016*. Copyright 2016 by the author(s)/owner(s).

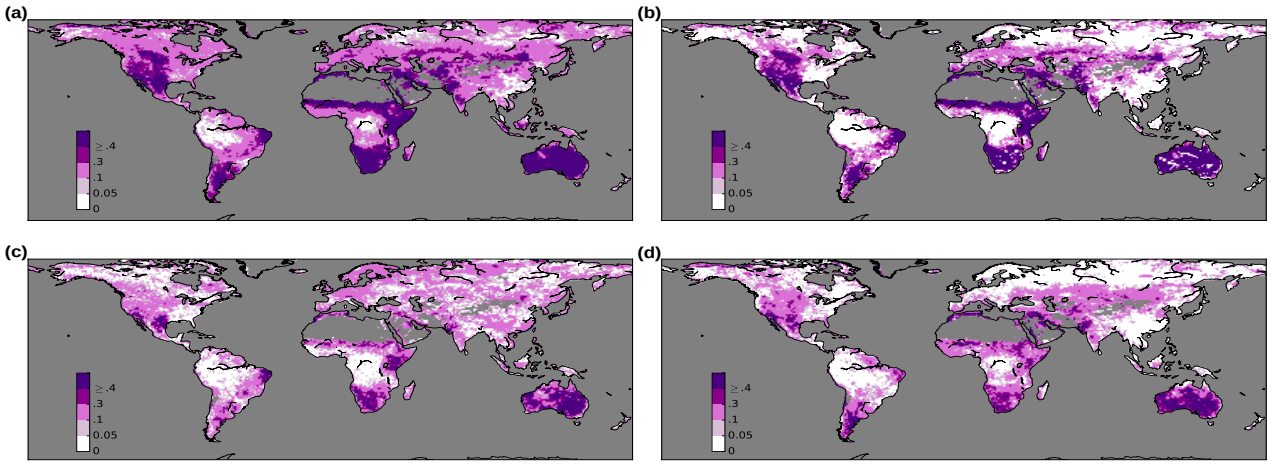


Figure 1. Model performance. Explained variance by different families of models. (a) Random forest approach, (b) Ridge regressions, (c) Pearson correlation for the most correlated feature to the target variable per pixel, (d) Autocorrelations.

driven approach, analyzing satellite time series that span the entire globe and three decades, resulting in a dataset of 130GB. We develop a pipeline that consists of several steps. In a first step, data from different sources were collected and merged into a single dataset that captures all needed information. In a second step, we reformulate the problem of discovering relationships in climate-vegetation dynamics as a machine learning problem, where vegetation is considered as the target time series, while climate information sources serve as predictor time series. We apply time series decomposition techniques to the target vegetation time series and the various predictor climatic time series to isolate seasonal cycles, trends and residuals. In this way unwanted correlations that originate from seasonal effects and trends are removed. Subsequently, we explore various techniques for constructing high-level features from climatic time series using techniques that are similar to shapelets (Ye & Keogh, 2009). We employ standard machine learning methods to search for shapelets that are predictive with respect to the residuals of vegetation time series. As shown in the experimental results, this approach allows us to discover novel insights w.r.t. climate-vegetation dynamics, moving beyond the state-of-the-art in this application domain.

Following the proposed pipeline, we end up with a dataset that has 5121 features generated on thirty-year time series. We analyze 13,097 land pixels independently, for which parallel implementations and a computer cluster can be employed. We use the random forest regressor and the model is assessed by means of five-fold cross-validation.

Figure 1a shows the result of the random forest model applied on the total number of features. As we can observe the model has better predictive skill in Australia, in the bulk of Africa and in a portion of North and South America. Compared to the other three approaches used as baseline methods, the results obtained by the random forest model are a big improvement. Ridge regression (Fig.1b) also performs well for the regions where the random forest yields a high R^2, in contrast to the other regions, where the map is mostly colored white. However, ridge regression in general leads to substantially worse results for almost all regions of the world. This result confirms that the relationships between the climatic variables are nonlinear. The filter approach (Fig.1c) based on Pearson correlations, which is the current state-of-the-art in modelling climate-vegetation dynamics, performs very poorly. Finally, in comparison with the autocorrelation method (Fig.1d), which intends to predict vegetation from the past vegetation observations, our approach gives better result in all regions. As such, we can conclude that an interplay between different variables is needed to model vegetation dynamics. More detailed results are presented in (Papagiannopoulou et al., 2016b) which is under review.

References

Bonan, G. (2008). Forests and climate change: forcings, feedbacks, and the climate benefits of forests. *science*, *320*, 1444–1449.

McPherson, R. A., Fiebrich, C. A., Crawford, K. C., Kilby, J. R., Grimsley, D. L., Martinez, J. E., Basara, J. B., Illston, B. G., Morris, D. A., Kloesel, K. A., et al. (2007). Statewide monitoring of the mesoscale environment: A technical update on the Oklahoma Mesonet. *24*, 301–321.

Papagiannopoulou, C., Depoorter, M., Miralles, D. G., Verhoest, N. E. C., Dorigo, W. A., & Waegeman, W. (2016a). Discovering relationships in climate-vegetation dynamics using satellite data. *Submitted in European Conference on Machine Learning and Principles and Practice of Knowledge Discovery (ECML/PKDD)*. Riva del Garda, Italy, (under review).

Papagiannopoulou, C., Miralles, D. G., Depoorter, M., Verhoest, N. E. C., Dorigo, W. A., & Waegeman, W. (2016b). Monthly vegetation dynamics primarily controlled by water availability. *Geophysical Research Letters, (in prep.)*.

Su, L., Jia, W., Hou, C., & Lei, Y. (2011). Microbial biosensors: a review. *Biosensors and Bioelectronics*, *26*, 1788–1799.

Ye, L., & Keogh, E. (2009). Time series shapelets: a new primitive for data mining. *Proc. of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '09* (p. 947). New York, New York, USA: ACM Press.