

EGU22-3855

<https://doi.org/10.5194/egusphere-egu22-3855>

EGU General Assembly 2022

© Author(s) 2022. This work is distributed under the Creative Commons Attribution 4.0 License.



CGC: an open-source Python module for geospatial data clustering

Ou Ku¹, Francesco Nattino¹, Meiert Grootes¹, Emma Izquierdo-Verdiguier², Serkan Girgin³, and Raul Zurita-Milla³

¹Netherlands eScience Center, Science Park 140, 1098 XG Amsterdam, The Netherlands

²Institute of Geomatics, University of Natural Resources and Life Science (BOKU), 1190, Vienna, Austria

³Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, PO Box 217, 7500 AE, Enschede, the Netherlands

With the growing ubiquity of large multi-dimensional geodata cubes, clustering techniques have become essential to extracting patterns and creating insights from data cubes. Aiming to meet this increasing need, we present Clustering Geodata Cubes (CGC): an open-source Python package designed for partitional clustering of geospatial data. CGC provides efficient clustering methods to identify groups of similar data. In contrast to traditional techniques, which act on a single dimension, CGC is able to perform both co-clustering (clustering across two dimensions e.g., spatial and temporal) and tri-clustering (clustering across three dimensions e.g., spatial, temporal, and thematic), as well as of subsequently refining the identified clusters. CGC also entails scalable approaches that suit both small and big datasets. It can be efficiently deployed on a range of computational infrastructures, from single machines to computing clusters. As a case study, we present an analysis of spring onset indicator datasets at continental scale.