SPACE AND ARTIFICIAL INTELLIGENCE

Online Conference, September 4th, 2020
Organized by CLAIRE and ESA, in association with ECAI2020

Clustering geo-data cubes

Names and affiliations of the authors: Raul Zurita-Milla (1),
Emma Izquierdo-Verdiguier (2), Serkan Girgin (1),
Francesco Nattino (3), Ou Ku (3),
Meiert W. Grootes (3), Romulo Goncalves (4)
(1) Faculty ITC, University of Twente, Enschede, NL,
(2) Institute of Geomatics, BOKU, Vienna, AT,
(3) Netherlands eScience Center, Amsterdam, NL,
(4) Helmholtz Centre Potsdam, GFZ, Potsdam, DE

Designated speaker: Raul Zurita-Milla

Earth observation sensors deliver ever-expanding collections of geospatial data at multiple resolutions (spatial, temporal and thematic or spectral). Efficient tools to extract knowledge from these collections are currently missing. Here we present the first release of Clustering geo-Data Cubes (CDC), a Python package to cluster geospatial data cubes by explicitly considering their dimensionality. CDC has three main hallmarks: 1/ it is based on innovative co- and tri-clustering methods that identify groups of pixels with similar spatio-temporal and/or thematic information by simultaneously considering all the dimensions of the data. This overcomes a major limitations of traditional clustering approaches, which analyze each dimension separately; 2/ it provides refined clusters by re-grouping the results obtained from co- and/or tri-clustering. These refined clusters better capture the patterns present in the data and represent a more automatic approach to analyze geospatial data cubes because the number of clusters is automatically chosen via an optimization procedure; and 3/ it allows users to run tasks efficiently by either using NumPy's threading capabilities or Dask's parallel computing power. Hence, CDC is a scalable package that can analyze both small and big geospatial data cubes. These hallmarks are showcased through several case studies.