ON DEMAND ENHANCEMENT OF EARTH SYSTEM DATA CUBES WITH HIGH-RESOLUTION SOCIOECONOMIC DATA STREAM

Guido Kraemer[1] in cooperation with Miguel D. Mahecha[1], Fabian Gans[2], Markus Reichstein[2]

[1]Remote Sensing Centre for Earth System Research, Leipzig University, Talstr. 35, Leipzig, Sachsen 04103
[2]Max Planck Institute for Biogeochemistry, 07745 Jena

24/07/2020

Abstract

Many subsystems of the Earth are constantly monitored in space and time with a large number of different data streams (e.g. gridded climate data, biophysical parameters of the land surface, or of aquatic bodies etc.). Interoperability among these data streams can be achieved via data cube approaches that allow efficient implementations of user-defined workflows. Today the human-environment nexus is affecting all aspects of the Earth system functioning and should be considered in any environmental analysis. In fact there is also a rapidly growing wealth of socioeconomic datasets available enabling scientists to address human-environment interactions in a rapidly changing world. However, although these datasets are often freely available, they are not yet available as part of any given data cube and often come e.g. as annual shape files at some administrative units. It is therefore inconvenient to work jointly with gridded environmental data and socioeconomic vector data. In this pilot we explore novel opportunities to spatially disaggregate coarse socioeconomic data with novel Machine Learning methods. The idea is that data integrals are maintained, while the spatial detail is realistically represented. Our aim is to implement algorithms that allow us to enhance any given Earth system data cube with socioeconomic datastreams enabling the joint analyses of societal, biospheric and atmospheric datasets through a unified interface. Allowing the formation of joint data cubes containing biospheric, atmospheric, and social data is the aim. Integrated cubes of this kind will provide standardized interfaces for managing, accessing, and analyzing these data streams jointly through Jupyter notebooks.

I. *Introduction*

What is the scientific context? The unprecedented availability of data streams describing different facets of the Earth now offers fundamentally new avenues to understand Earth system processes. At the same time, human activity causes unprecedented changes to the Earth, especially the biosphere. The total impact is far from being understood and will likely increase in the future. In the last decades humanity has made huge progress toward better lives for all human beings (Kraemer et al. 2020a): Poverty and hunger have been reduced, despite a globally increasing population. Resolving basic problems of humanity has caused an enormous increase in human population causing even more pressure on the Earth's ecosystems. Understanding the interactions between humans and the planet requires the integration of socioeconomic datasets with biospheric and atmospheric datasets.

What is the data-challenge you face? Several practical hurdles, especially the lack of data interoperability, limit the joint potential of these data streams. In this particular case, the issue is that spatiotemporally gridded data cubes have become a standard in the analysis of the climate system and many land-surface processes, while most socioeconomic data sets are provided a coarse temporal resolutions via irregular spatial shapes that typically represent the administrative reporting unit. However such coarse spatial patterns obscure where human activity really plays a role and therefore needs to be downscaled to the grid cells and time-points where the activity really plays a role.

What is state-of-the-art? The Earth system data lab approach (Mahecha et al. 2020) has successfully leveraged data cubes to easily analyse many covariates along any dimension (space, time, variables, models etc) and was successfully used to analyze e.g. high-dimensional system trajectories (Kraemer et al. 2020b). The approach will be further developed in TA2 of the NFDI4Earth and make a wealth of biospheric and atmospheric data available, as well as providing a unified interface and toolbox to access and analyze these data. However, while this system is apt to integrate new global socioeconomic datasets that are derived from satellite remote sensing e.g. the the Global Urban Footprint (Esch et al. 2018), it is not fit to deal with integrating administrative data streams. The Euro-Stats database[1] or World Development Indicators[2], for instance, contain a large wealth of socioeconomic indicators which provide a very valuable source for researchers that are, however, practically not co-interpretable within the ESDL if one wants to preserve the high spatial detail. However, latest advances in machine learning (for instance via Gaussian Processes) make it effectively possible to disaggregate such data. Impressive examples from e.g. Malaria research has proven that one can - with appropriate geospatial covariates - draw high-resolution images of infections that are not only smooth ín space, but also conserve count numbers when re-aggregated. The aim is to elaborate and implement a strategy to automate such a process that shall allow us to make as many socioeconomic data streams accessible and cointerpretable from a common data cube framework as implemented in the ESDL and extend existing toolboxes for jointly analyzing these datasets.

[1] https://ec.europa.eu/eurostat

[2] https://databank.worldbank.org/source/world-development-indicators

What is the vision if your challenge would be solved? Extending existing data cubes will promote interdisciplinary and data intensive research by facilitating access and processing of these datasets to researchers of different fields.

II. *Pilot description*

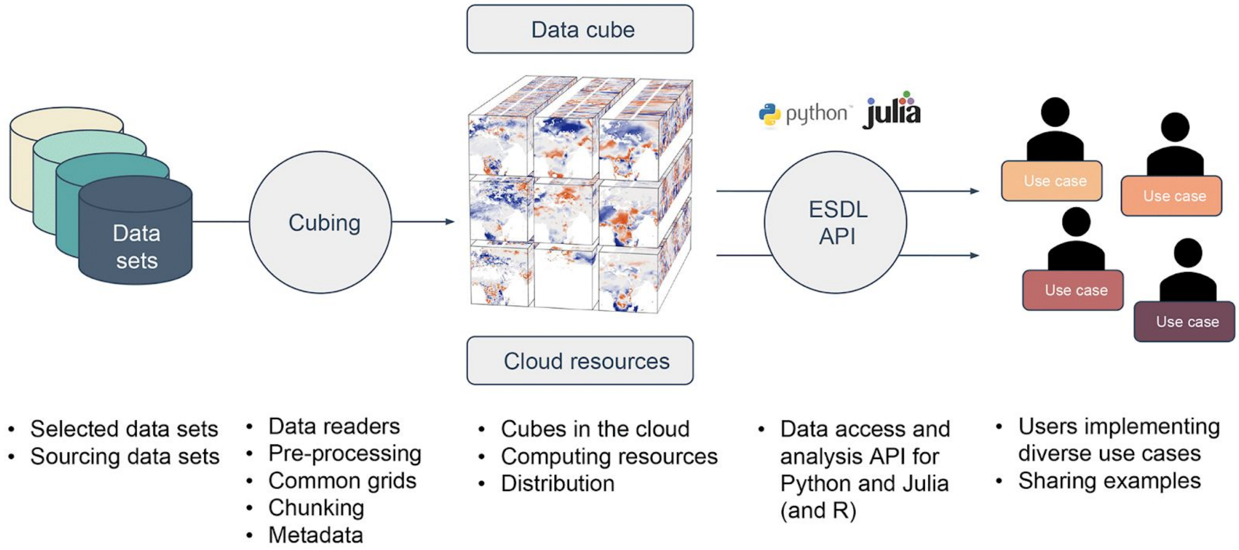


Figure 1: Selected data sets are preprocessed to common grids and saved in cloud-ready data formats (Zarr). Based on these cubed data sets, a global Earth system data cube can be produced that is either stored locally or in the cloud. Via appropriate application programming interfaces (APIs), users can efficiently access the ESDC in their native language. Users can fully focus on designing workflows for their research (from Mahecha et al. 2020).

What is the technological backbone you rely upon? The Earth System Data Lab (ESDL) is an integrated data and analytical hub that curates a multitude of data streams representing key processes of the different subsystems of the Earth in a common data model and coordinate reference system. This infrastructure was developed originally jointly with the European Space Agency (ESA), is open access and will serve as a key element in the TA 2 Facilitate as it enables researchers to apply their own workflow to the analysis-ready data cubes. Today, data streams included focus on the analysis of:

- Ecosystem states at the global scale in terms of relevant biophysical variables.
- Biosphere–atmosphere interactions as encoded in land fluxes of carbon, water and energy.
- Terrestrial hydrology.
- State of the atmosphere.
- Meteorological conditions.

Gridded socioeconomic data are, so far, provided only in scattered repositories. Central data collections e.g. Eurostats and related collections instead, curate data at the level of certain administrative units, but can be queried very efficiently.

What is the innovation compared to the status quo? The question of how to achieve an expansion to socioeconomic data streams remains unclear. In this pilot we will explore one avenue to integrate socioeconomic datasets into the ESDL using spatial disaggregation. This will enhance the resolution of the socioeconomic datasets and make the resulting gridded datasets accessible though the same interfaces as the already existing biospheric and atmospheric datasets.

Which are the standards and interoperability approaches used in the pilot's context? The ESDL uses well documented and reproducible workflows for the creation of the data cubes, the data cubes are stored in the open Zarr formats, and the analysis toolboxes leverage open source programming languages, which can be extended easily by the users.

What is the proposed solution? We offer an approach to automatically (on the fly) disaggregate spatial data based on appropriate covariates. The reason that covariates need to be exchangeable is that for each novel analysis, one may require independence among used variables and hence the user needs to have the flexibility to exclude certain covariates. The methods of choice is the method used by Law et al. (2018) for the case of Malaria, but we will screen different relevant methods approaches, e.g. Appel & Pebesma (2020), Keil et al. (2013), and Nandi et al. (2020). At the end we provide a software package to automate the process and first proof-of-concept applications.

III. *Relevance for the NFDI4Earth*

What are expected users and stakeholders and how do they benefit? This pilot allows scientists from human geography, climate science, economy, sociology, demography, and biology to jointly create and later analyze a large variety of different datastreams relevant to understand the human-environment nexus. It then allows them to access the data remotely through cloud hosted data cubes in the Zarr format or to analyze the data remotely through Jupyter notebooks. This provides a set of tools for infrastructure providers to make data and computing capacities available to the public. The NFDI TA 2 Facilitate will receive a tool to enrich the central data cubes on demand.

What is the potential for other sub-branches in the Earth System Sciences? We break the barriers between scientific silos and enable interdisciplinary research. On the long-term, an enhanced Earth system data cube will provide the tools for analyzing global interactions between society and the biosphere to better measure and understand the interactions and minimize their impacts on the environment.

What elements of FAIR are particularly addressed? The ESDL fully complies with the FAIR principles: Datasets are made freely available and easily identifiable through a doi. The data is stored in the open Zarr format, containing rich metadata and easily readable with many softwares commonly used in the analysis of Earth observation data. This also makes the data easily interoperable with any other data that can be read through these tools. The data gathering and preprocessing is well documented and can be fully reproduced from a git repository.

What aspects of the research data life cycle are particularly addressed? This pilot will facilitate the processing and the analysis of data by providing interfaces and analysis toolboxes. It will facilitate the access and reuse of data using open data formats that allow the storage and access of datasets in the cloud.

Are there particular contributions that help the NFDI4Earth to engage with? The open data formats and toolboxes will allow entities, such as NFDI4Earth, to easily provide users access to a wide variety of datasets and computing capabilities for their analysis. It also facilitates teaching, because less thought has to be spent on setting up a proper computational environment for the students. The integration with widely used programming languages allows an easy interoperability with external data sources.

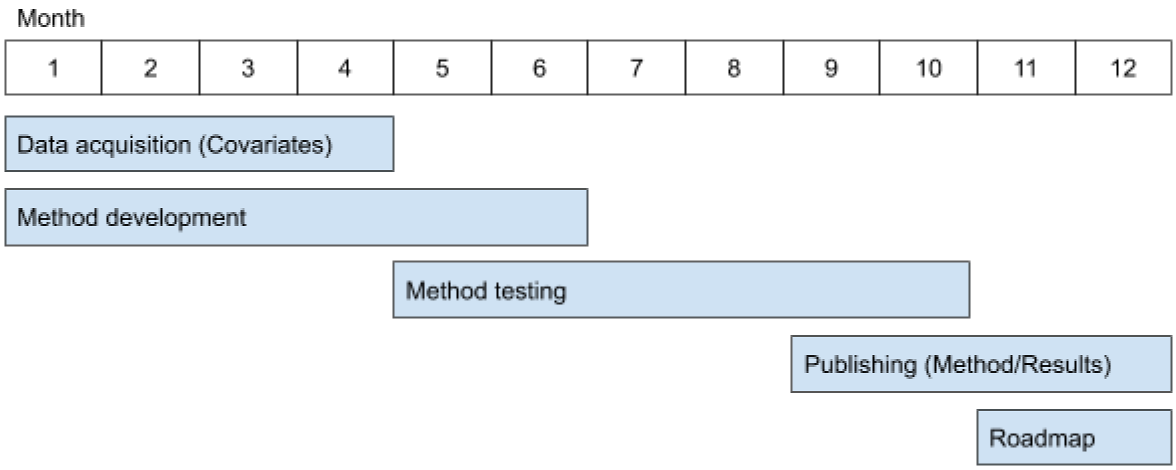
IV. *Deliverables*

Technical operability of the pilot: A Julia or Python package to automatically enhance an existing Earth system data cube with a wide range of socioeconomic data streams.

Roadmap: A description of how future socioeconomic data sets with spatiotemporal extent can be integrated on a regular basis with existing climatological and biospheric data cubes and converted into interoperable “Analysis Ready Data Cubes”. The roadmap will particularly report on the difficulties that the pilot faced and recommend future approaches for achieving interoperability between data streams representing human and physical geographical aspects.

V. *Work Plan & Requested funding*

Funding: We request 1PJ to be be able to implement the technical work proposed here
Milestone plan:



References

Law, H. C., Sejdinovic, D., Cameron, E., Lucas, T., Flaxman, S., Battle, K., & Fukumizu, K. (2018). Variational learning on aggregate outputs with Gaussian processes. Advances in Neural Information Processing Systems (pp. 6081-6091).

Mahecha, M. D., Gans, F., Brandt, G., Christiansen, R., Cornell, S. E., Fomferra, N., Kraemer, G., Peters, J., Bodesheim, P., Camps-Valls, G., Donges, J. F., Dorigo, W., Estupinan-Suarez, L. M., Gutierrez-Velez, V. H., Gutwin, M., Jung, M., Londoño, M. C., Miralles, D. G., Papastefanou, P., & Reichstein, M. (2020). Earth system data cubes unravel global multivariate dynamics. Earth System Dynamics, 11(1), 201–234. https://doi.org/10.5194/esd-11-201-2020

Kraemer, G., Reichstein, M., Camps-Valls, G., Smits, J., & Mahecha, M. D. (2020a). The Low Dimensionality of Development. Social Indicators Research. https://doi.org/10.1007/s11205-020-02349-0

Kraemer, G., Camps-Valls, G., Reichstein, M., & Mahecha, M. D. (2020b) Summarizing the state of the terrestrial biosphere in few dimensions. Biogeosciences, 17(9), 2397–2424. https://doi.org/10.5194/bg-17-2397-2020

Esch, T., Bachofer, F., Heldens, W., Hirner, A., Marconcini, M., Palacios-Lopez, D., Roth, A., Üreyen, S., Zeidler, J., Dech, S., & Gorelick, N. (2018). Where We Live—A Summary of the Achievements and Planned Evolution of the Global Urban Footprint. Remote Sensing, 10(6), 895. https://doi.org/10.3390/rs10060895

Appel, M., & Pebesma, E. (2020). Spatiotemporal multi-resolution approximations for analyzing global environmental data. Spatial Statistics, 38, 100465. https://doi.org/10.1016/j.spasta.2020.100465

Keil, P., Belmaker, J., Wilson, A. M., Unitt, P., & Jetz, W. (2013). Downscaling of species distribution models: A hierarchical approach. Methods in Ecology and Evolution, 4(1), 82–94. https://doi.org/10.1111/j.2041-210x.2012.00264.x

Nandi, A. K., Lucas, T. C. D., Arambepola, R., Gething, P., & Weiss, D. J. (2020). disaggregation: An R Package for Bayesian Spatial Disaggregation Modelling. ArXiv:2001.04847 [Stat]. http://arxiv.org/abs/2001.04847