Second Latin America and Caribbean SCIENTIFIC DATA MANAGEMENT WORKSHOP

FEBRUARY 10th and 24th 2021

9:00am (UTC - 3:00)

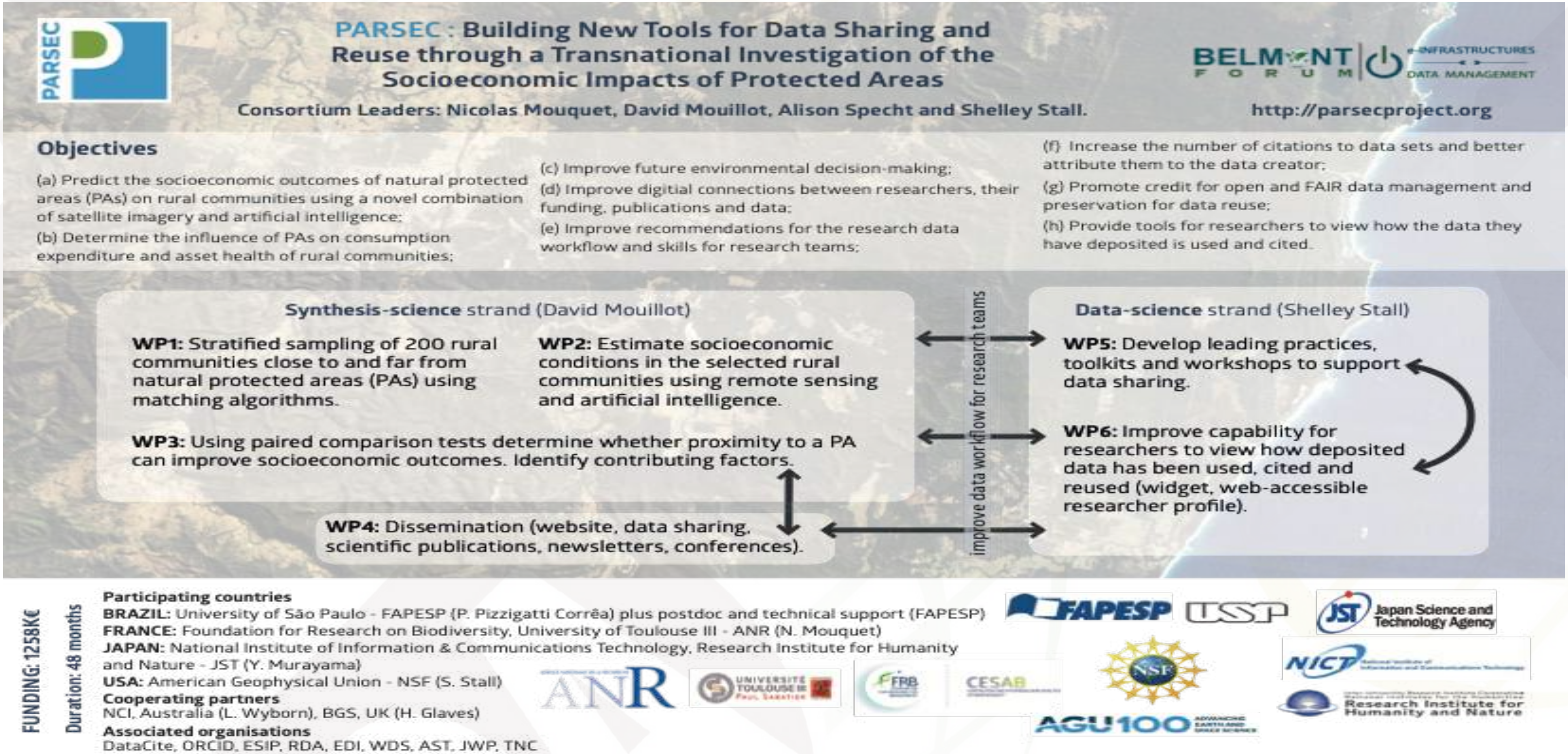
Panel 1 - Scientific data management in Health and the Environment

PARSEC : Building New Tools for Data Sharing and Reuse through a Transnational Investigation of the Socioeconomic Impacts of Protected Areas

https://parsecproject.org

Prof. Dr. Pedro Luiz Pizzigatti Corrêa - pedro.correa@usp.br
Digital Systems and Computer Engineering Department
Escola Politécnica da Universidade de São Paulo - EPUSP
Big Data and Data Science Research Group of EPUSP wds.poli.usp.br

PARSEC Project



PARSEC : Building New Tools for Data Sharing and Reuse through a Transnational Investigation of the Socioeconomic Impacts of Protected Areas

Consortium Leaders: Nicolas Mouquet, David Mouillot, Alison Specht and Shelley Stall.

BELMONT FORUM | e-INFRASTRUCTURES DATA MANAGEMENT

http://parsecproject.org

Objectives

(a) Predict the socioeconomic outcomes of natural protected areas (PAs) on rural communities using a novel combination of satellite imagery and artificial intelligence;

(b) Determine the influence of PAs on consumption expenditure and asset health of rural communities;

(c) Improve future environmental decision-making;

(d) Improve digitial connections between researchers, their funding, publications and data;

(e) Improve recommendations for the research data workflow and skills for research teams;

(f) Increase the number of citations to data sets and better attribute them to the data creator;

(g) Promote credit for open and FAIR data management and preservation for data reuse;

(h) Provide tools for researchers to view how the data they have deposited is used and cited.

Synthesis-science strand (David Mouillot)

WP1: Stratified sampling of 200 rural communities close to and far from natural protected areas (PAs) using matching algorithms.

WP2: Estimate socioeconomic conditions in the selected rural communities using remote sensing and artificial intelligence.

WP3: Using paired comparison tests determine whether proximity to a PA can improve socioeconomic outcomes. Identify contributing factors.

WP4: Dissemination (website, data sharing, scientific publications, newsletters, conferences).

improve data workflow for research teams

Data-science strand (Shelley Stall)

WP5: Develop leading practices, toolkits and workshops to support data sharing.

WP6: Improve capability for researchers to view how deposited data has been used, cited and reused (widget, web-accessible researcher profile).

FUNDING: 1258K€

Duration: 48 months

Participating countries

BRAZIL: University of São Paulo - FAPESP (P. Pizzigatti Corrêa) plus postdoc and technical support (FAPESP)

FRANCE: Foundation for Research on Biodiversity, University of Toulouse III - ANR (N. Mouquet)

JAPAN: National Institute of Information & Communications Technology, Research Institute for Humanity and Nature - JST (Y. Murayama)

USA: American Geophysical Union - NSF (S. Stall)

Cooperating partners

NCI, Australia (L. Wyborn), BGS, UK (H. Glaves)

Associated organisations

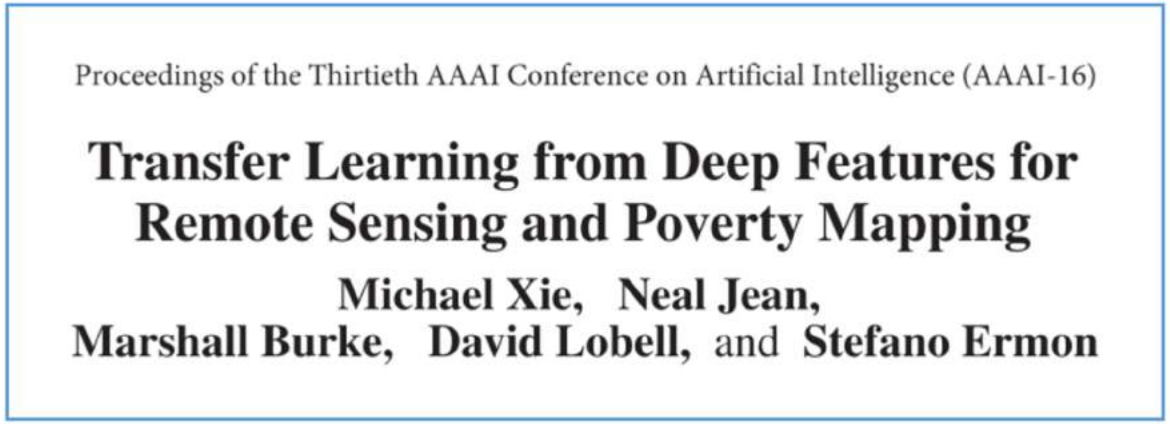
DataCite, ORCID, ESIP, RDA, EDI, WDS, AST, JWP, TNC

FAPESP
USP
JST Japan Science and Technology Agency
ANR
UNIVERSITÉ TOULOUSE III PAUL SABATIER
FRB
CESAB
NSF
NICT
AGU100
Research Institute for Humanity and Nature

IA models

State-of-the-art deep-learning & poverty prediction

[Xie, et al, 2016]

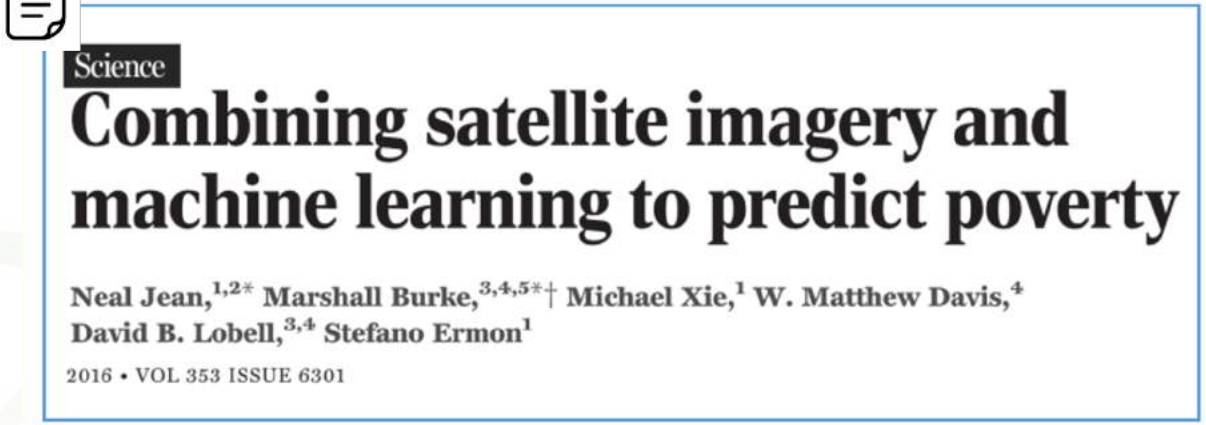


Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)

Transfer Learning from Deep Features for Remote Sensing and Poverty Mapping

Michael Xie, Neal Jean,
Marshall Burke, David Lobell, and Stefano Ermon

[Jean, et al, 2016]



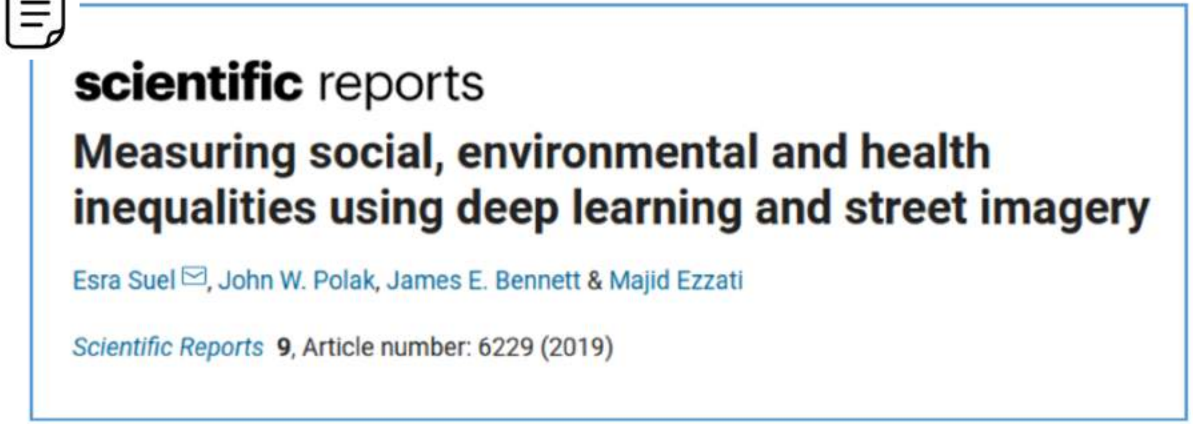
Science

Combining satellite imagery and machine learning to predict poverty

Neal Jean,^{1,2*} Marshall Burke,^{3,4,5*†} Michael Xie,^{1} W. Matthew Davis,^{4}
David B. Lobell,^{3,4} Stefano Ermon^{1}

2016 • VOL 353 ISSUE 6301

[Suel, et al, 2019]



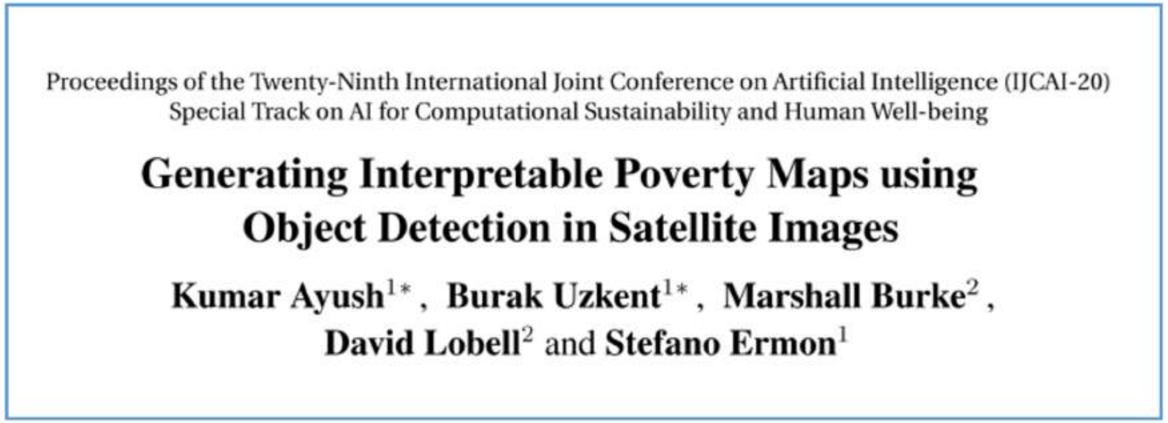
scientific reports

Measuring social, environmental and health inequalities using deep learning and street imagery

Esra Suel, John W. Polak, James E. Bennett & Majid Ezzati

Scientific Reports **9**, Article number: 6229 (2019)

[Ayush, et al, 2016]

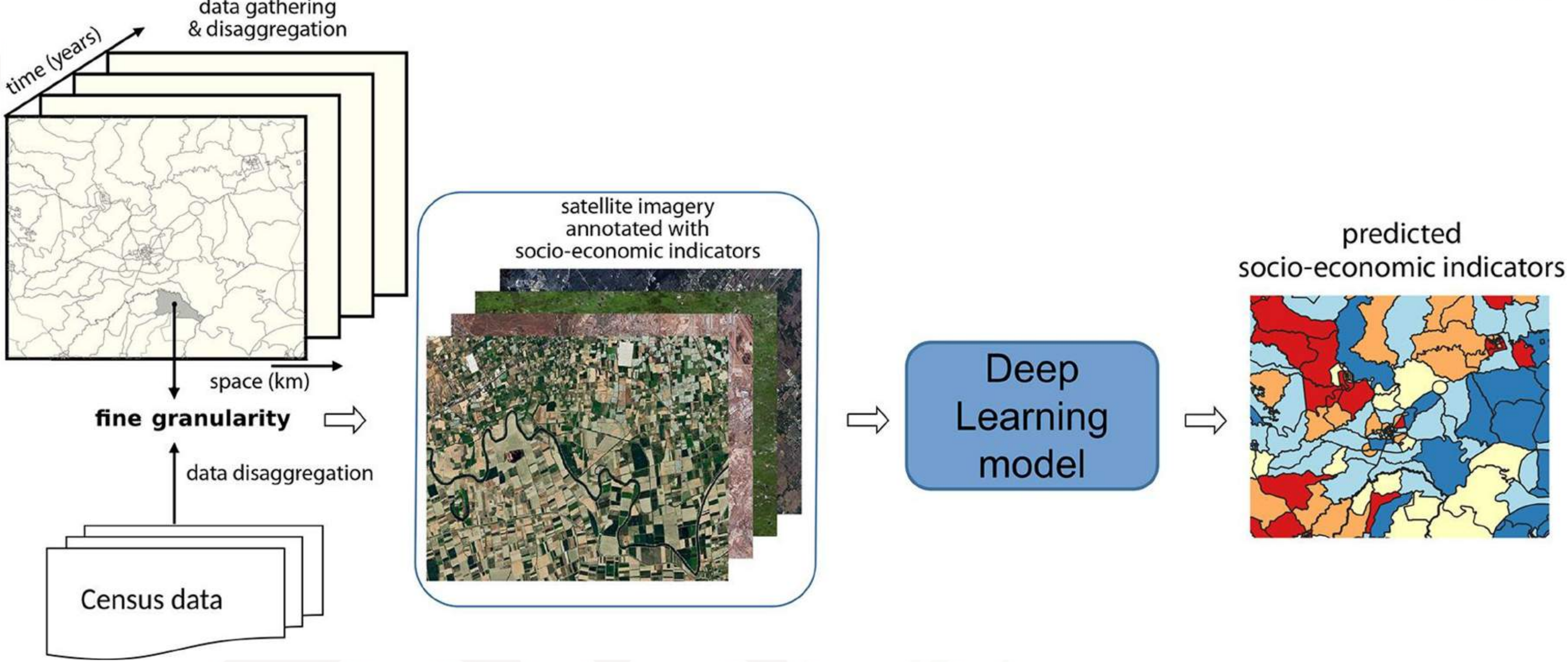


Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20)
Special Track on AI for Computational Sustainability and Human Well-being

Generating Interpretable Poverty Maps using Object Detection in Satellite Images

Kumar Ayush^{1*}, Burak Uzkent^{1*}, Marshall Burke^{2},
David Lobell^{2} and Stefano Ermon^{1}

General Methodology



data gathering
& disaggregation

time (years)

space (km)

fine granularity

data disaggregation

Census data

satellite imagery
annotated with
socio-economic indicators

Deep
Learning
model

predicted
socio-economic indicators

Types of Data Used and Generated

Raw Data: Inputs

Ground Truth Data
Local surveys

Satellite images

Processed Data: Outputs

Global gridded socio-economic information

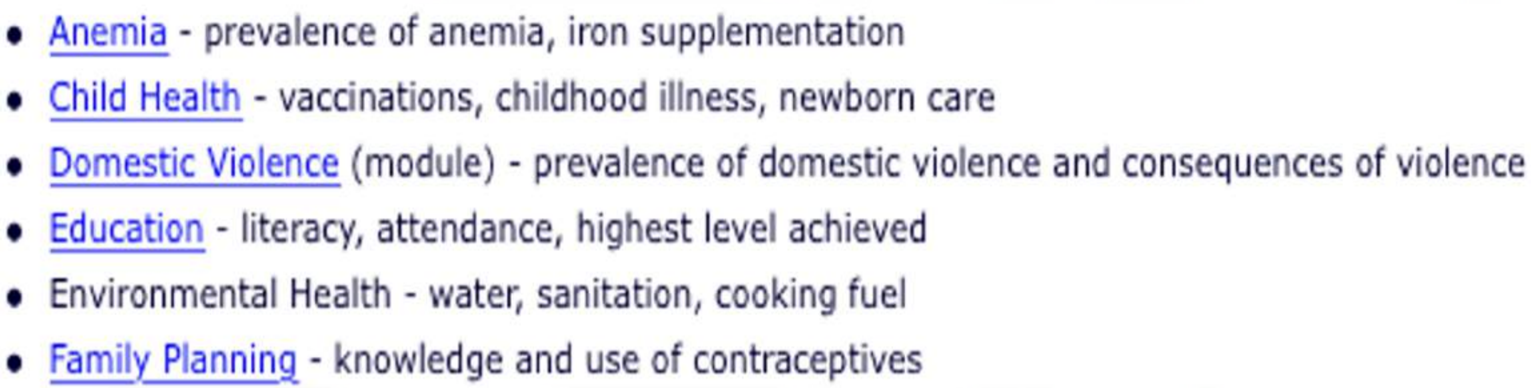
Socio-economic dynamics in our study sites

Raw data

The DHS Program
Demographic and Health Surveys

USAID
FROM THE AMERICAN PEOPLE

+30 0000 household surveys every 5 years in +90 countries



- Anemia - prevalence of anemia, iron supplementation
- Child Health - vaccinations, childhood illness, newborn care
- Domestic Violence (module) - prevalence of domestic violence and consequences of violence
- Education - literacy, attendance, highest level achieved
- Environmental Health - water, sanitation, cooking fuel
- Family Planning - knowledge and use of contraceptives

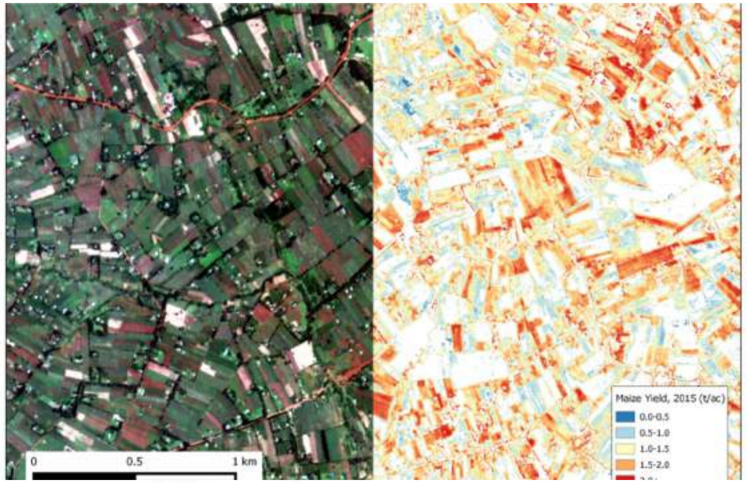
No restriction access for academic research
Can be stored locally or downloaded when necessary

Raw data

Satellite images

Many options

- Both free and fees required
- Various Resolutions
- Various time series
- Usability can be an issue for some datasets



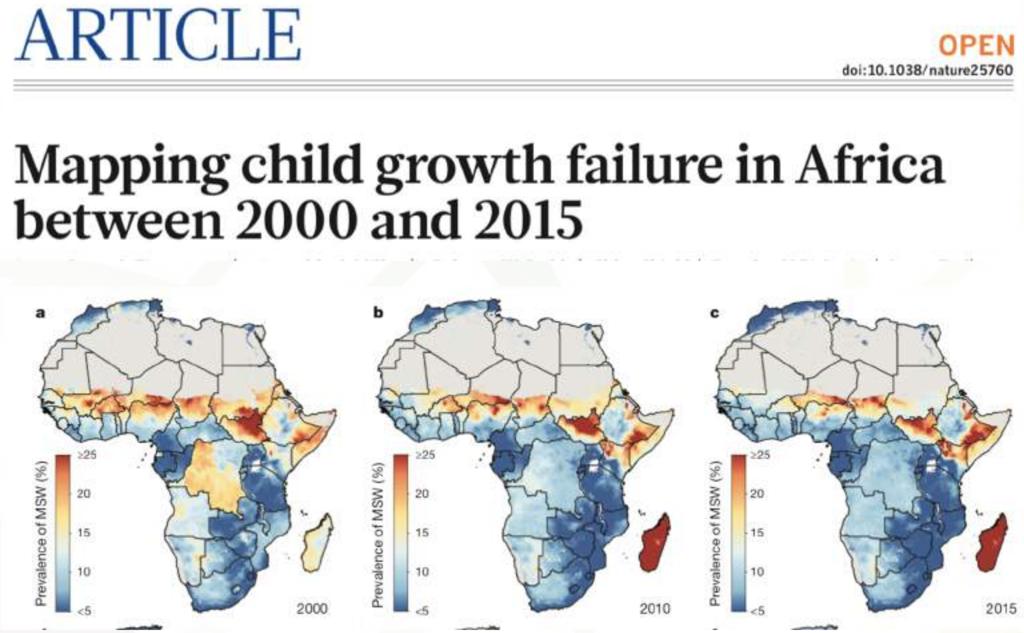
Processed data

PARSEC

BELMONT FORUM

Global gridded socio-economic information

LSMS
Living Standards Measurement Study



ARTICLE

OPEN

doi:10.1038/nature25760

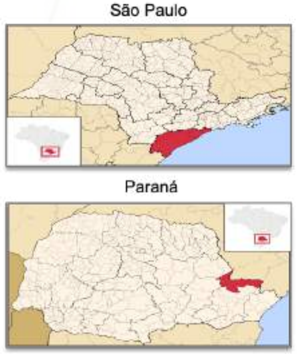
Mapping child growth failure in Africa between 2000 and 2015

Only some information/data have gridded at the regional scale

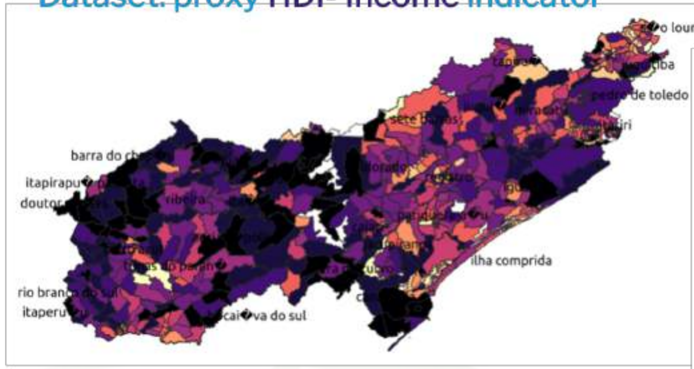
We plan to grid many of them globally and make them available

Study Case of Brazil Team – Vale do Ribeira – SP

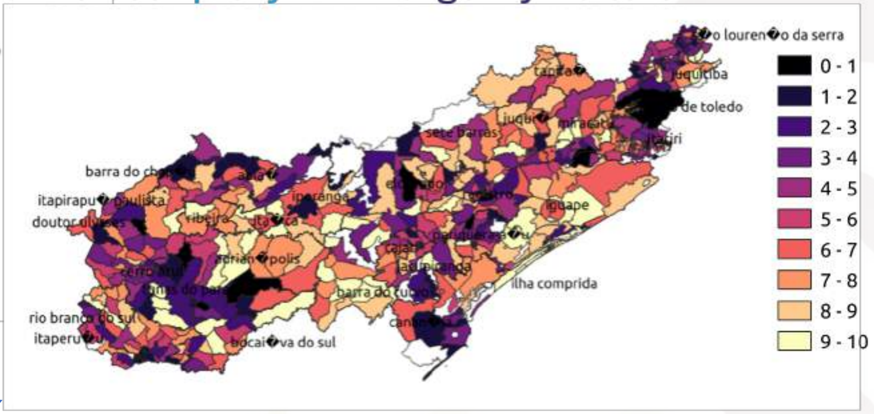
Ubication



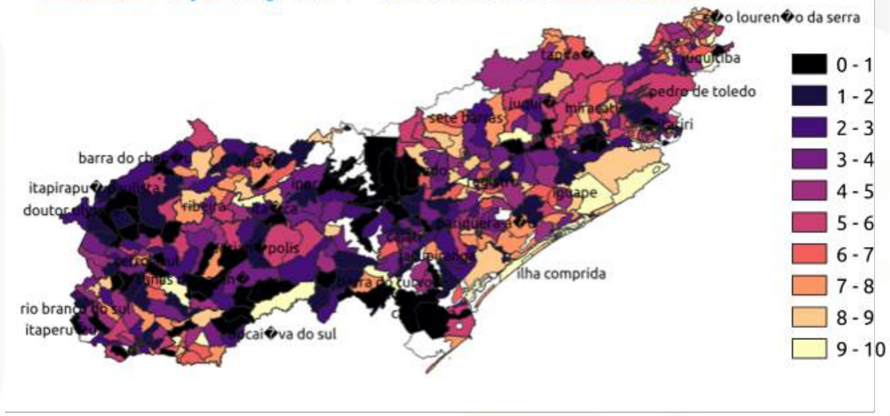
Dataset: proxy HDI- Income indicator

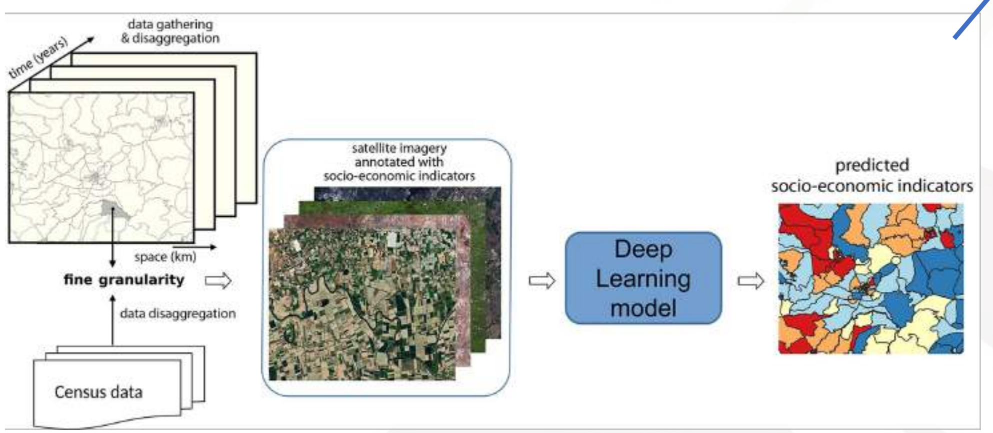


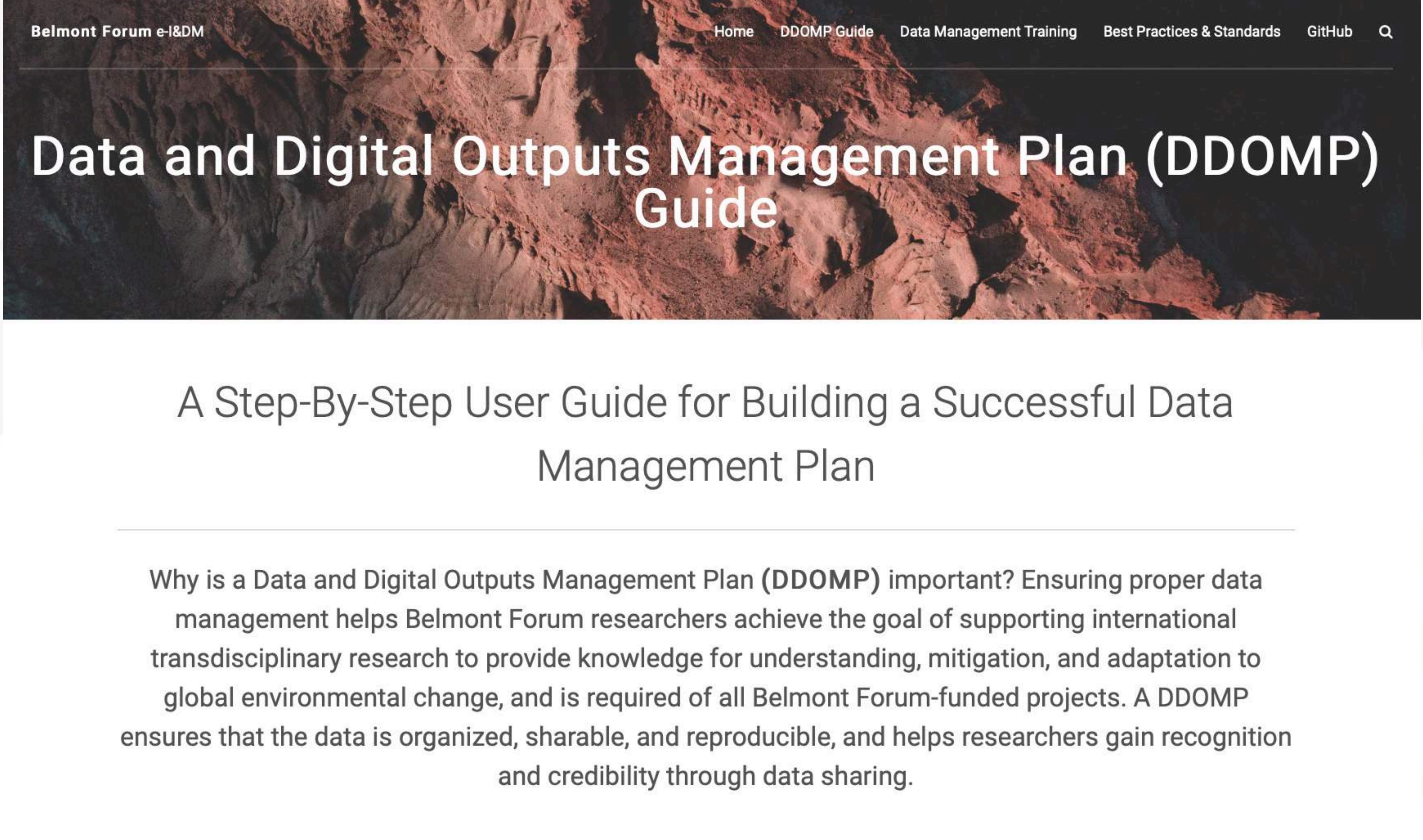
Dataset: proxy HDI- Longevity indicator



Dataset: proxy HDI- Education indicator




data gathering & disaggregation
time (years)
space (km)
fine granularity
data disaggregation
Census data
satellite imagery annotated with socio-economic indicators
Deep Learning model
predicted socio-economic indicators



Data and Digital Outputs Management Plan (DDOMP) Guide

A Step-By-Step User Guide for Building a Successful Data Management Plan

Why is a Data and Digital Outputs Management Plan **(DDOMP)** important? Ensuring proper data management helps Belmont Forum researchers achieve the goal of supporting international transdisciplinary research to provide knowledge for understanding, mitigation, and adaptation to global environmental change, and is required of all Belmont Forum-funded projects. A DDOMP ensures that the data is organized, sharable, and reproducible, and helps researchers gain recognition and credibility through data sharing.

Benefits of a Workbook

Supports the iterative / flexible needs of the research team
Provides guidance on “what data”, “where to store”, “what to track”
Provides guidance on when actions are taken
Gives method for

- what to do **during** the project
- how to **preserve** your digital objects for **publication and sharing**

Details to follow

Checklist for your team to make it “super simple”
Validation task for the PI to ensure **compliance** and **consistency**

DDOMP Checklist – Team Resources

- Material development and temporary storage location
  - Google Drive
- Team communications and information decimation tools
  - Email, Slack
- Dataset storage location during the project
  - Open Science Framework (AWS integration)
- Software development platform
  - GitHub
- Data preservation (including derived products) repository
  - Environmental Data Initiative
- Software preservation repository
  - Zenodo
- Training, workshop material preservation repository
  - Zenodo

PARSEC

- PIs – 4
- Country Leaders – 6
- Funders – 4
- Researchers – 30
- Languages - 4

PARSEC Data and Digital Output Management Plan and Workbook

Further details can be found on the process and methods used for PARSEC:

Stall, Shelley, Specht, Alison, Corrêa, Pedro Luiz Pizzigatti, David, Romain, Edmunds, Rorie, Mabile, Laurence, Machicao, Jeaneth, O'Brien, Margaret, Wyborn, Lesley. (2020). PARSEC Data and Digital Output Management Plan and Workbook. Zenodo. 10.5281/zenodo.3891426

Use your DMP or DDOMP to make your own Checklist.

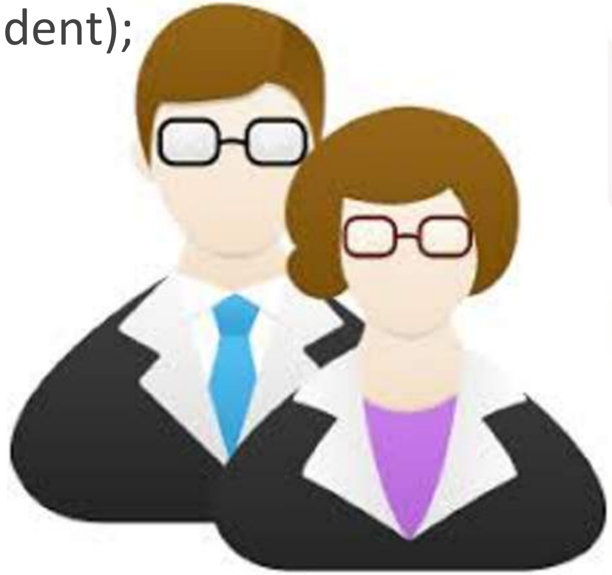
Special thanks to PARSEC Brazilian team

Researchers:

- Profa. Dra. **Katia** Maria Paschoaletto Micchi de Barros Ferraz: (ESALQ/USP);
- Dr. **Jean** Pierre Henry Balbaud Ometto: Instituto Nacional de Pesquisas Espaciais (INPE);
- Dra. Marina **Jeaneth** Machicao Justo - (EPUSP), (postdoc fellowship PARSEC);
- Dra. **Solange** Maria Dos Santos: (Scielo);
- Dr. **Silvio** Marchini - (ESALQ/USP);
- Eng. **Danton** Ferreira Vellenich (EPUSP), (TT fellowship PARSEC and Master Student);
- Prof. Dr. **Pedro** Luiz Pizzigatti Corrêa: (EPUSP) (Country Leader).

Support of Brazilian Institute of Geography and Statistics (IBGE):

- Dra. **Nadya** Maria Deps (IBGE);
- MSc. **Miguel** Suarez Xavier Penteado (IBGE).



Second Latin America and Caribbean SCIENTIFIC DATA MANAGEMENT WORKSHOP

FEBRUARY 10th and 24th 2021

9:00am (UTC - 3:00)

Panel 1 - Scientific data management in Health and the Environment

PARSEC : Building New Tools for Data Sharing and Reuse through a Transnational Investigation of the Socioeconomic Impacts of Protected Areas

https://parsecproject.org

Prof. Dr. Pedro Luiz Pizzigatti Corrêa - pedro.correa@usp.br
Digital Systems and Computer Engineering Department
Escola Politécnica da Universidade de São Paulo - EPUSP
Big Data and Data Science Research Group of EPUSP wds.poli.usp.br