The detection of socio-economic impacts of protected area creation

Alison Specht[1]*, Jeaneth Machicao[2], Pedro Corrêa[2], Rodolphe Devillers[3], Yasuhisa Kondo[4], David Mouillot[5], Yasuhiro Murayama[6], Shelley Stall[7], Jamie Trammell[8], Danton Vellenich[2]

[1]University of Queensland, Australia; [2]University of São Paulo, Brazil; [3]French National Research Institute for Sustainable Development (IRD); [4]Research Institute for Humanity and Nature, Japan; [5]University of Montpellier, France; [6]National Institute of Information and Communications Technology, Japan; [7]American Geophysical Union, USA; [8]University of Southern Oregon, USA.

* correspondent for the paper

Abstract

This paper discusses a project aimed at detecting whether protected areas (PAs) influence the socio-economic well-being of adjacent communities. The Belmont Forum funded PARSEC project is using satellite images and deep learning algorithms to predict socio-economic conditions. In this paper we show our on-going work for the selection of PAs, development of methodology using deep learning to detect socio-economic indicators from remote sources, facilitation in data management and the approach used to handle a complex inter-disciplinary and trans-national team. We note the challenges in selecting case studies with examples from Australia, Brazil, Japan and the USA, and meshing remote sensed data with census data. We discuss the advantages of good data management for the individual and for the project and some simple steps to make this easy.

1 Introduction

The impact of global changes on the world's ecosystems, as well as the repercussions for human societies, is a major concern for the research and policy agenda. Protected areas (PAs) can contribute directly or indirectly to socioeconomic outcomes such as food security and development opportunities for nearby human communities in this uncertain future.

This topic is being addressed through the Belmont Forum-funded PARSEC project (www.parsecproject.org), with funding partners from France, the USA, Brazil, Japan, and associates in Australia and the UK. The method we are using is based on recent research on how satellite images and deep learning can be used to predict socioeconomic conditions (Jean et al., 2016; Yeh et al., 2020).

As a desired outcome from this project, the Belmont Forum expects not only a good scientific product but also a workflow and outputs that are transparent, open, and reproducible (i.e. it is FAIR, Wilkinson et al., 2016). To this end, the PARSEC project has two components, a team of synthesis scientists (the synthesis strand) and a team of data scientists (the data strand). The synthesis strand is working on the subject of this paper, while the data strand is developing tools to support active international collaboration and good data management practices. These two strands are designed to learn from each other: by the end of the project the domain scientists should be better equipped to practice good data management, and the data scientists better able to speak with and respond to researcher priorities. In this paper we discuss some initial successes and challenges using examples from Australia, Brazil, Japan and the USA.

2 Methods

2.1 Towards the detection of socio-economic indicators

The synthesis strand is essentially a working group of the FRB-CESAB, the French synthesis centre, and led by David Mouillot of the University of Montpellier. This strand is engaged with the core business of detecting the socio-economic effects of the protected

areas. The work can be summarised in two steps: (i) to identify areas with PAs for which satellite images before and after their establishment exist; and (ii) to use novel techniques of deep learning and artificial intelligence to model the socioeconomic outcomes of the effects of creation of those protected areas on their settlements.

2.1.1 Site Selection

The primary concern for the synthesis strand to date has been site selection. We are prioritising protected areas (PAs) that are listed globally in the IUCN World Database of Protected Areas (WDPA, https://www.protectedplanet.net/), and we need PAs that will allow us to obtain satellite images before and after their establishment, ideally those PAs established after 1990 and before 2015. The PAs selected will have adjacent, small towns nearby that are relatively unaffected by strong influences like big cities and mining activity. For validation, 'mirror' or reference sites are also being selected (Figure 1). We have stratified the selection according to ecoregions using Olson, et al. (2001) for the terrestrial sites, and Spalding, et al. (2007) for the marine.

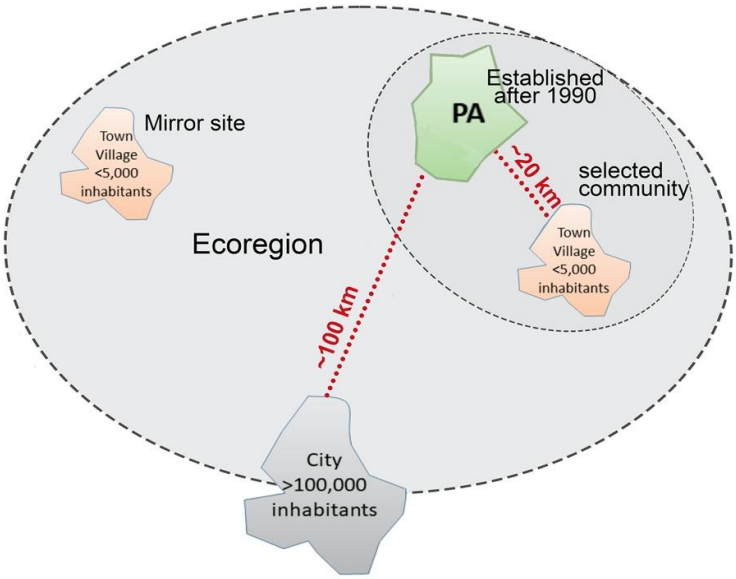

Figure 1: PARSEC criteria for site selection.

2.1.2 Satellite Data Availability

To conduct a time-series analysis for change detection using remote means, satellite images are preferably required a few times before the creation of a protected area, and several times after its creation. Reference sites also need to have such a time series. Effects may not be seen for some time after the 'impact'.

For a global project such as this, the availability of images is variable across space and time. Some countries have better coverage than others, and of course changes in satellites over time affects the quality of the images available. Some images are expensive, and this project does not have a large budget for acquiring them. Image type determines their spatial resolution, from $1m^2$ for some modern satellite images (often only commercially obtainable) to $70m^2$ resolution for old Landsat MSS images.

As an example for this paper, we have chosen a marine protected area in Jurien Bay, Western Australia, and the small town next to it (Figure 2). This PA was created between 2003 and 2005 (depending on source). In the satellite image before its creation (1985), the little village can clearly be observed. Over time the small town has clearly grown, but whether its growth is due to the creation of the PA, and whether the economic circumstances of the people living there have changed, is the question to be determined.

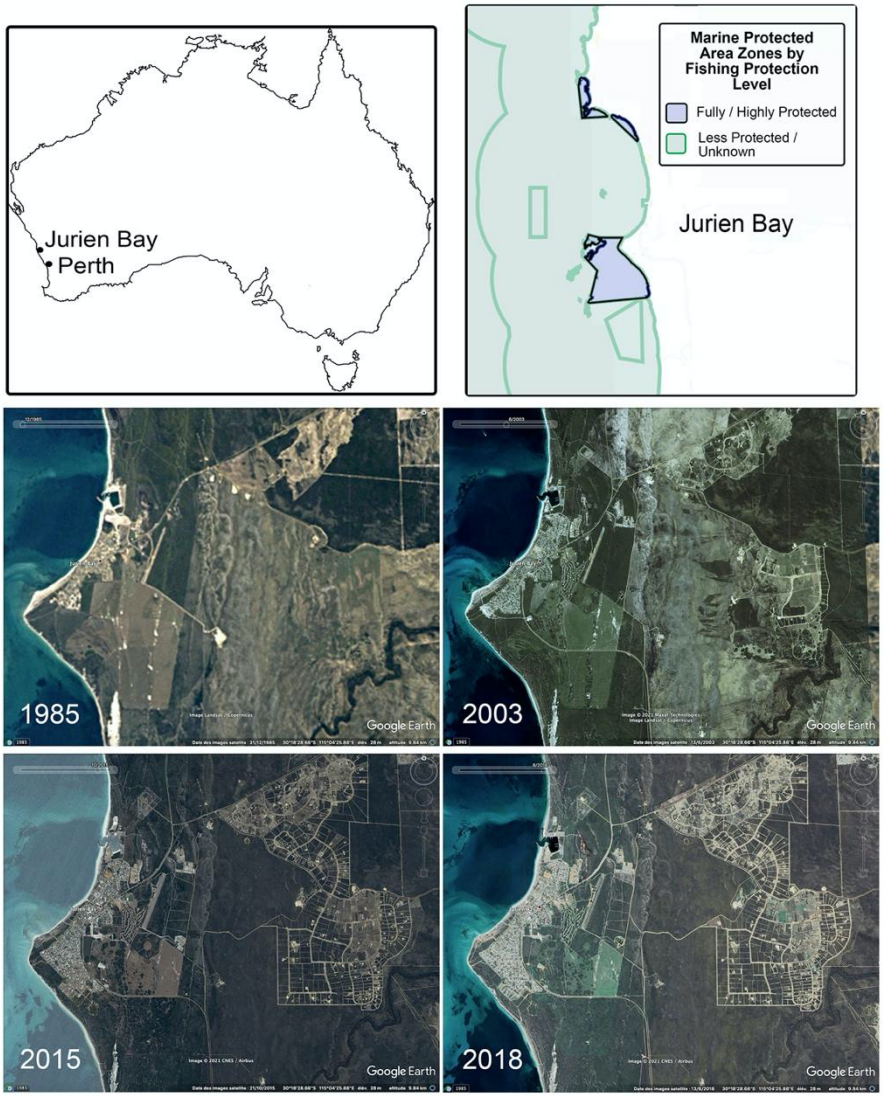

Figure 2: Satellite data before and after the creation of the marine protected area in Jurien Bay, Western Australia.

In many situations there are a lack of signals from the satellite image that are socioeconomically meaningful. One of the inspirations for this project was successful in finding that in East Africa (Jean, et al., 2016), and a few others have shown some success including Yeh, et al. (2020), and Ayush, et al. (2021).

2.1.3 Workflow

Census data from each country and international (harmonised) sources will be spatially aggregated and interpreted through satellite imagery using visually detectable socioeconomic indicators. Using a deep learning model the team aims to develop predicted socioeconomic indicators that may well be used for decision support (Figure 3). Moreover, within the deep learning methodology the satellite images that belong to a selected site will be split for training, test and validation, which are used for the deep learning model.

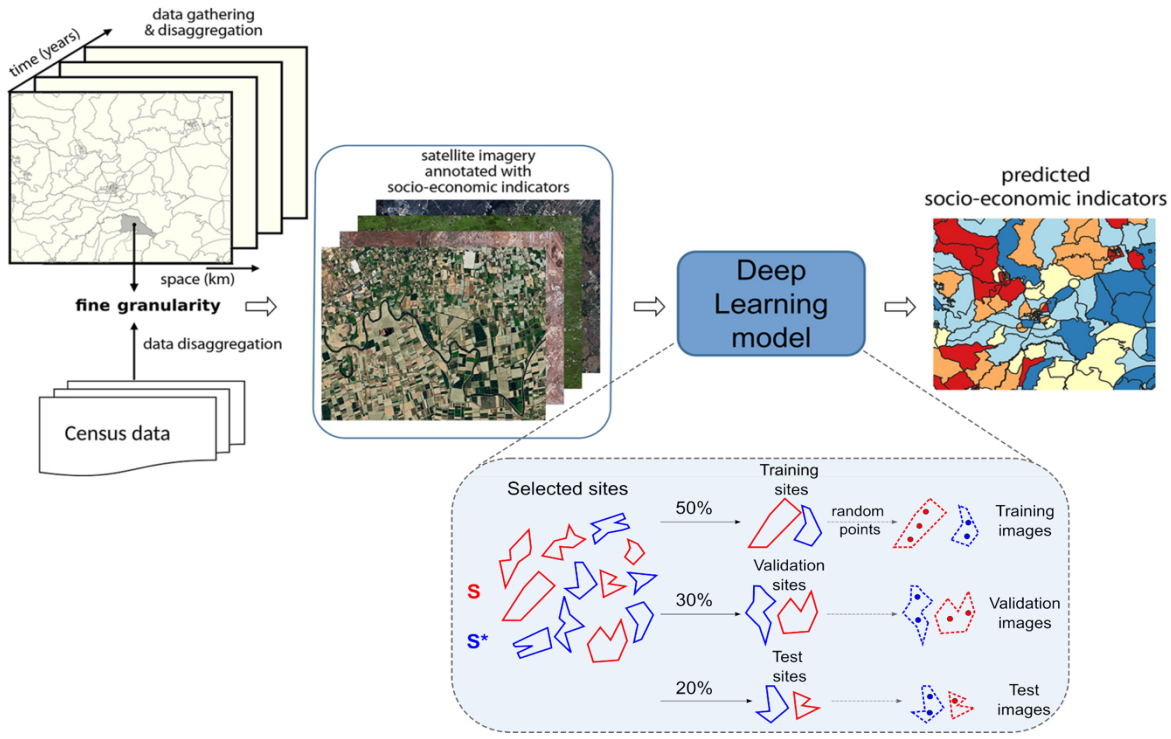

Figure 3: The proposed workflow. First, the dataset preparation, where the remote sensing image data and census data are acquired and aggregated into the image-labelled dataset. Second, neural network architecture will be used to train these data. Third, the configuration of experiments is set up using data sets split for training, validation and testing. The sites selected (S: sites (villages and towns) near to PAs and S*: the mirror sites without PAs) are used for validation of the experiments in order to analyse the effect. Diagram based on Machicao, et al., 2020.

2.2 Towards optimising data management and developing tools for researchers

The data scientists, led by Shelley Stall of the American Geophysical Union, are working in parallel with the research team. They are guided by the Belmont Forum's Data and Digital Object Management Plan (DDOMP: https://bit.ly/3yPsQII) and Bishop et al., (2020). The Belmont Forum requires its funded projects to ensure they are open and reproducible, and the data strand aims to be an exemplar. This strand is charged with anticipating and responding to the data management needs of the synthesis strand and engaging actively with the wider research community. This includes the digital logistics of communication through to delivery of resultant data in a CoreTrustSeal registered repository (Lin et al., 2020). It also requires the actions taken in the project to be re-traceable and reportable at any time.

Communication and feedback from the wider community is achieved via conference participation (e.g. the Research Data Alliance, CoData, ESIP, the European, Japanese and American Geophysical Unions), round table discussions and training workshops as well as refereed articles. The engagement of a range of members from both strands in these activities is an important component of the sociological health of the whole team.

3 Results

3.1 Selected PAs

The first task has been to identify areas with PAs for which we could obtain satellite images before and after their establishment, and these are illustrated for four countries, the USA, Australia, Brazil, and Japan (Figure 4). As mentioned before, our main source for suitable protected areas is the WDPA, but we validate them with national information (e.g. *Ministério do Meio Ambiente* in Brazil, and CAPAD in Australia). In the USA there are many protected

areas, but most of them, especially on the east coast, were established before 1990, so they were immediately excluded. On the west coast, there are suitable PAs, but villages that are not affected by other influences are hard to find. In Australia many parks were created in the 1990s, which is very promising, and in a draft selection last year in Western Australia we were able to identify some potential sites, but again adjacent settlements and the absence of disturbance factors become really significant in many cases. The same is true in Brazil, where most of the suitable parks are in relatively isolated regions (in the north west), and we have the same challenge.

In Japan, the filtering criteria were adjusted due to the small distance between cities. Despite this, the PAs that fell within the project's temporal and spatial envelope were those found on the Ogasawara Islands (Figure 4), so a different approach is being explored for Japan. The selection of suitable PAs remains a challenge and we have learned a lot during this process, not just about the choice of parks themselves.

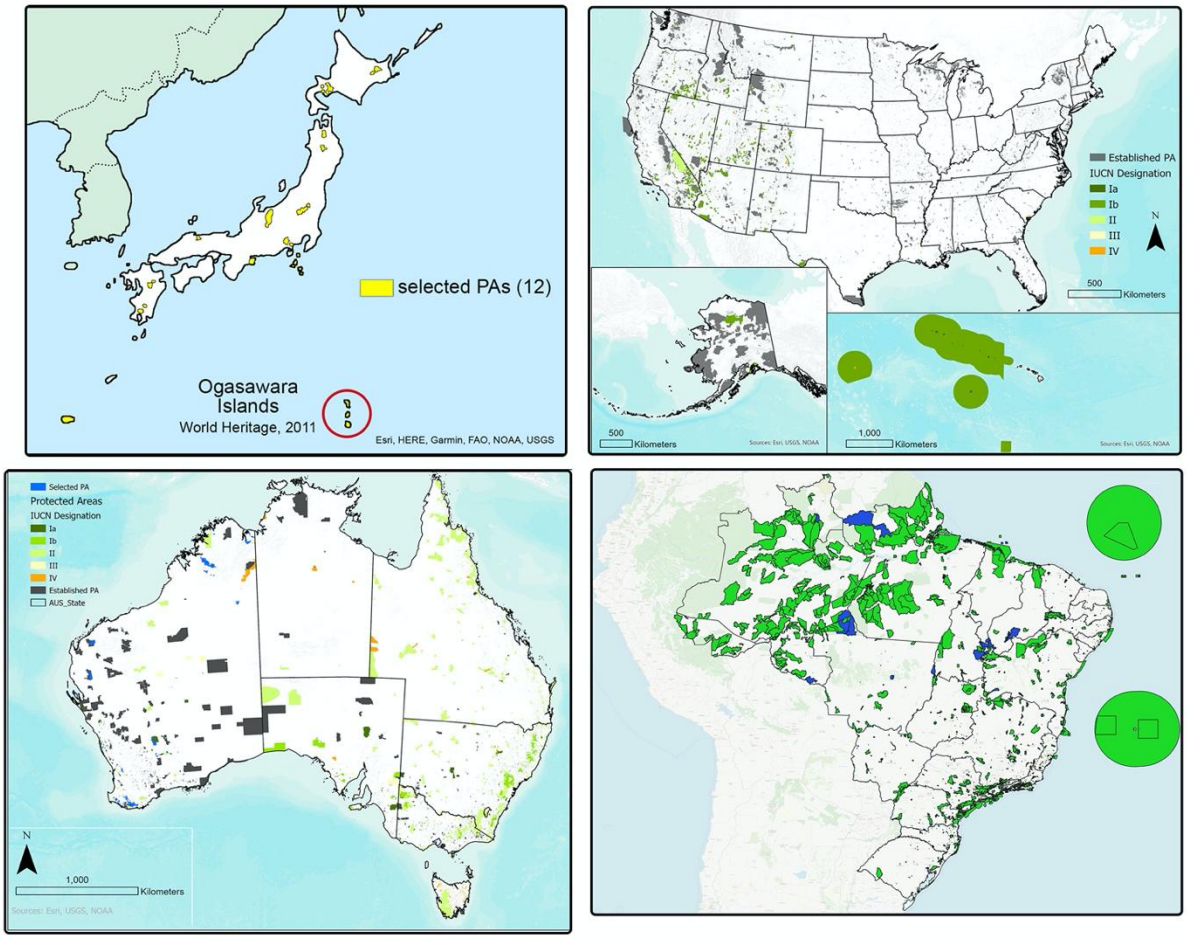

Figure 4: Maps from Australia, Brazil and the USA showing protected areas that fall within our criteria (green) and selected PAs for Western Australia and Brazil (blue). For Japan a range of potential PAs are shown in yellow, but only the PA on the Ogasawara Islands (highlighted) falls within the project criteria.

3.2 Test of remote sensing of socio-economic parameters

In an experiment to test the use of machine learning to discover indices of socio-economic well-being from remote sources we took Google Street View (GSV) images as our 'remote' source using an approach developed by Suel et al. (2019) in a very comprehensive case of use of GSV for the city of London.

As described more fully in Machicao et al. (subm.), we aimed to evaluate the prediction of income indicators using street view images, through a case study conducted in the area Vale do Ribeira, located in the southeast of Brazil. This area is a semi-rural area with a large proportion occupied by protected areas. We collected 2010 census data from the Brazilian Institute of Geography and Statistics (IBGE) and Google Street View imagery of 30 regions within the Vale do Ribeira for our images. As the images returned from GSV are panoramic, we requested four picture orientations (0°, 90°, 180° and 270°) to cover the image view completely. The computation for the analysis involved extraction by a pre-trained convolutional neural network (CNN) whose goal is to classify the images to detect socio-economic indices (Figure 5).

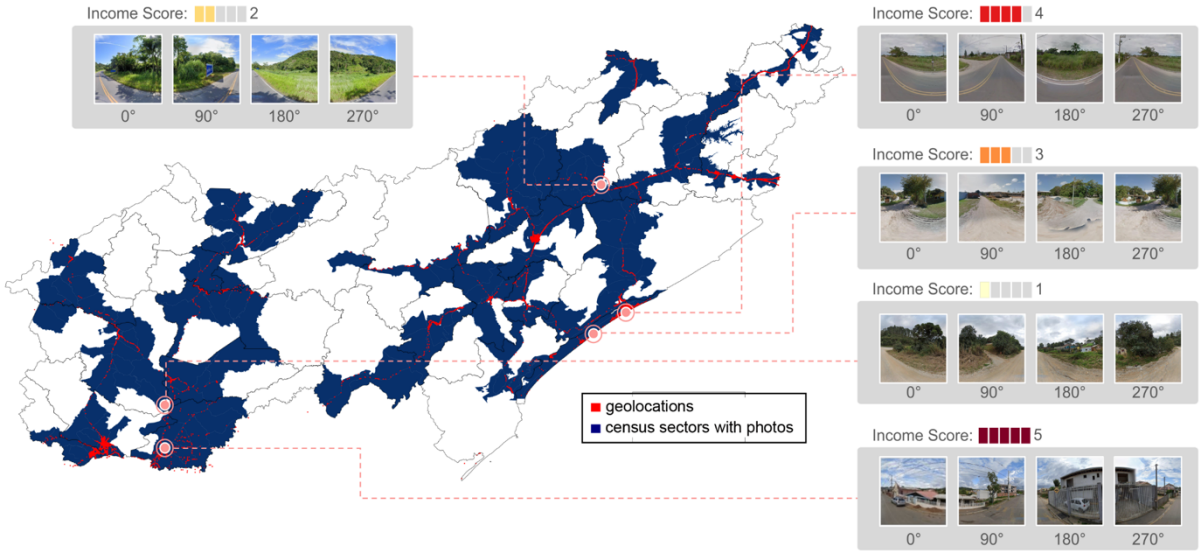

Figure 5: Deep Learning method applied to the prediction of a socioeconomic (income) indicator on the Vale do Ribeira (southeast of Brazil) using Google Street view images. The census areas without colour are mostly occupied by sparsely settled rural or protected areas. Source: (Machicao subm.)

The observed income score was compared spatially with the predicted values from the census (Figure 6). The best performance (80%) was obtained for the higher income region of the predicted cases, while it decreased uniformly for other regions. The result was inherently biased by the number of images in GSV, as the higher income regions had more available images (had more visible infrastructure) than lower income areas.

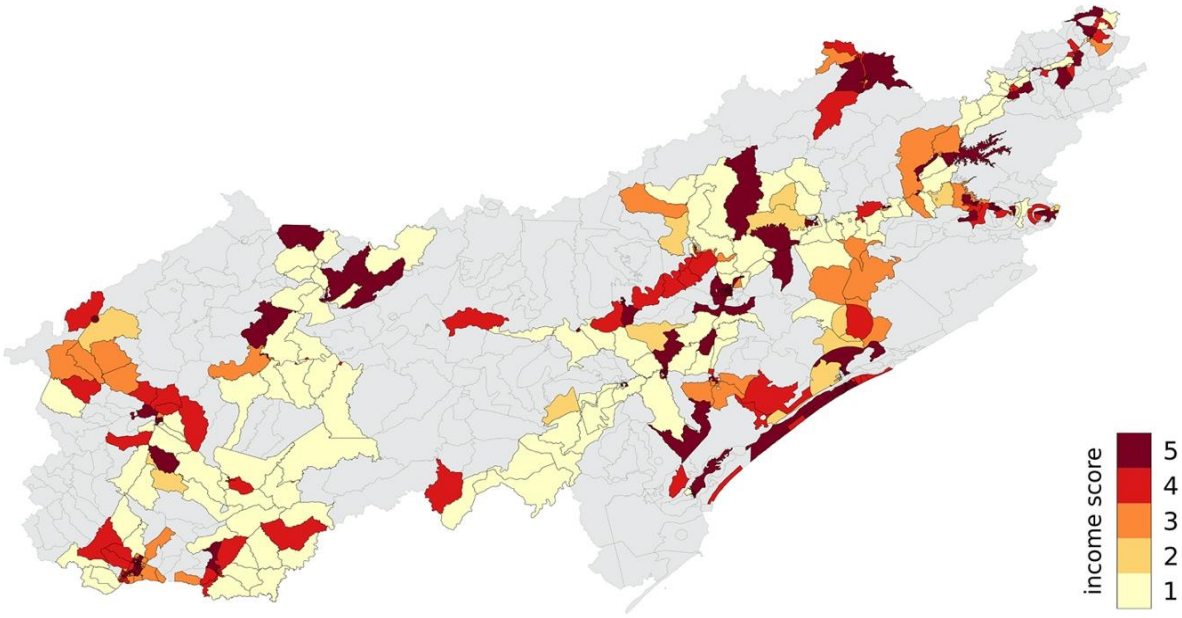

Figure 6: Results of the prediction of income level using GSV images on the Vale do Ribeira. This plot shows the performance of the trained network to predict income score. The colour scale runs from 1 to 5, with 1 representing the lowest income and 5 representing the highest. Source: (Machicao subm.)

3.3 Data management practices

The data strand has established a common set of resources for the transnational and transdisciplinary team. This starts with support for effective group function.

Communication tools used are primarily email and Slack (the latter less well used by the members), with regular zoom meetings held within country teams, strands, between the postdocs in the project (Brazil and France) and around particular sub-projects. Twice a year the whole group meets and given this project has mainly existed in the time of Covid19, this has, with one exception, been remote. Additional opportunities have been grasped at conferences with regular co-located meetings, for example, at RDA Plenaries since the start of the project.

Google Drive is being used for document sharing (including meeting notes) and as a temporary repository for data files. Google Drive is connected to a dedicated space on Open Science Framework (OSF, https://osf.io/), which in turn is integrated with Amazon Web Service (AWS, https://aws.amazon.com/). Software development with GitHub works well and is commonly used by the developers, with jupyter notebooks and R Markdown some of the most popular forms of recording code and versioning. As recommended, we have identified a repository for final data storage, the Environmental Data Initiative (EDI, https://environmentaldatainitiative.org/), which has become a partner in the project. We have a project library on Zotero for our bibliographic records. We have our own community on Zenodo to publish workshop materials, presentations, short reports, and software (https://bit.ly/2U1iCpj). Zenodo provides a digital object identifier (with versioning capability) and citation for each object published there. More details about this organisation can be found in Stall (2021).

Another big contribution of the data strand has been listening to the synthesis science team and getting them to think about a process during their workflow that would facilitate their data management goals and open-access practice. This has evolved into a simple checklist for the researcher which can be applied to any research team, not just PARSEC (Stall et al., 2021). Everyone on the team has their ORCID profile, ensuring links to CrossRef and DataCite are activated. Every quarter the team member reviews their ORCID profile to check that it is updated. Every week the datasets created and used are ideally recorded using

spreadsheets provided on Google Drive, and any workflow or provenance details added. Monthly we update conference presentations and posters in a spreadsheet on Google Drive (cross-checking with Zenodo), and each team member is charged with ensuring data sets, images and software used are preserved and the publications that referred to them are recorded.

4 Discussion

We have presented some of the challenges and preliminary results for the PARSEC project, a project involving forty people from different disciplines, countries, languages and cultural backgrounds. Having two components, the data strand and the synthesis strand, has been useful to divide the goals and organize the knowledge flow.

The selection of protected areas with the criteria we require to address the core project question has proved challenging. Part of the reason for the project was to test the generality of the approach of Jean et al. (2016), and it was expected that there would be different situations in each country. Adjustments for variation in population density, from high (Japan) to low (some parts of Brazil and Australia) and conflicting land-use activities remain to be solved.

Blending census information with remote imagery, however, looks promising, assuming a good series of repeated images can be obtained. The experiment with Google Street View (GSV) was a direct attempt to test this, and it highlighted some useful limitations for which we need to plan. These include repeatability of the method. In the case of GSV the attractiveness of its ubiquitous nature throughout the world is offset by, even for one extraction event, the high temporal variability in the images, anywhere between 2011 and 2019 for the Vale do Ribeira study, while the census data were available for 2010 only. We continue to conduct small experiments that will be useful for choosing the best data management and the appropriate configuration for deep learning architecture.

The work of the data strand aims to ensure transparency across the team and encourage and support group work practices. Following the tasks outlined in the checklist has greatly eased the onerous reporting process for Belmont and the country funders (all of which have additional separate annual reporting requirements). The regular attention to recording activity ensures appropriate credit for data providers as well as team members when publishing their data, code or articles of any sort. The data strand has had many outputs which have largely been stimulated by the fortnightly meetings and high activity (conferences, presentations posters) in the data science community.

The desired outcome is not only to have a good science project, but also to have a workflow and a product that is transparent, open and reproducible.

Acknowledgements

This research is product of the PARSEC group funded by the Belmont Forum as part of its Collaborative Research Action (CRA) on Science-Driven e-Infrastructures Innovation (SEI), the São Paulo Research Foundation (FAPESP), Brazil, the Agence National de Recherche (ANR) France, the Japan Science and Technology Agency (JST), the National Science Foundation (NSF) of the United States of America, and the synthesis centre CESAB of the French Foundation for Research on Biodiversity. In addition, the author's organisations have provided valuable support for each participant. Special thanks to Vitor Dias Souza, an undergraduate student at the University of São Paulo, for help with transcription.

References

Ayush, K., Uzkent, B., Tanmay, K., Burke, M., Lobell, D., Ermon, S., 2021. Efficient Poverty Mapping using Deep Reinforcement Learning. arXiv:2006.04224 [cs]. http://arxiv.org/abs/2006.04224

Bishop, B., Gunderman, H., Davis, R., Lee, T., Howard, R., Samors, R., Murphy, F., Ungvari, J., 2020. Data Curation Profiling to Assess Data Management Training Needs and Practices to Inform a Toolkit. Data Science Journal 19, 4. https://doi.org/10.5334/dsj-2020-004

Jean, N., Burke, M., Xie, M., Davis, W.M., Lobell, D.B., Ermon, S., 2016. Combining satellite imagery and machine learning to predict poverty. Science 353, 790–794. https://doi.org/10.1126/science.aaf7894

Lin, D., Crabtree, J., Dillo, I., Downs, R.R., Edmunds, R., Giaretta, D., De Giusti, M., L'Hours, H., Hugo, W., Jenkyns, R., Khodiyar, V., Martone, M.E., Mokrane, M., Navale, V., Petters, J., Sierman, B., Sokolova, D.V., Stockhause, M., Westbrook, J., 2020. The TRUST Principles for digital repositories. Scientific Data 7, 1–5. https://doi.org/10.1038/s41597-020-0486-7

Machicao, J., Corrêa, P., Ferraz, K., Vellenich, D.F., David R., Mabile, L., Stall, S., Specht, A., O'Brien, M., Meneguzzi, L., Ometto, J., Santos, S., subm. A deep-learning method for the prediction of socio-economic indicators from street-view imagery using a case study from Brazil. Data Science Journal.

Machicao, J., Jarry, R., Vellenich, D. F., Ometto, J. P., Ferraz, K., Deps, N., Penteado, M. S. X., Stall, S., Specht, A., Mabile, L., Chaumont, M., Corrêa, P., David, R., 2020. Evaluation of deep-learning methods to understand the prediction of socio-economic indicators from remote sensing imagery. https://doi.org/10.5281/zenodo.4280070

Olson, D.M., Dinerstein, E., Wikramanayake, E.D., Burgess, N.D., Powell, G.V.N., Underwood, E.C., D'amico, J.A., Itoua, I., Strand, H.E., Morrison, J.C., Loucks, C.J., Allnutt, T.F., Ricketts, T.H., Kura, Y., Lamoreux, J.F., Wettengel, W.W., Hedao, P., Kassem, K.R., 2001. Terrestrial Ecoregions of the World: A New Map of Life on Earth. BioScience 51, 933. https://doi.org/10.1641/0006-3568(2001)051[0933:teotwa]2.0.co;2

Spalding, M.D., Fox, H.E., Allen, G.R., Davidson, N., Ferdaña, Z.A., Finlayson, M., Halpern, B.S., Jorge, M.A., Lombana, A., Lourie, S.A., Martin, K.D., McManus, E., Molnar, J., Recchia, C.A., Robertson, J., 2007. Marine Ecoregions of the World: A Bioregionalization of Coastal and Shelf Areas. BioScience 57, 573–583. https://doi.org/10.1641/B570707

Stall, S., 2021. PARSEC: A FAIR Data Use Case with 40 Researchers, 6 Countries, and one Data Management Plan. https://doi.org/10.5281/zenodo.4978466

Stall, S., Specht, A., Corrêa, P.L.P., David, R., Edmunds, R., Mabile, L., Machicao, J., O'Brien, M., Wyborn, L., Vellenich, D.F., Miyairi, N., Murayama, Y., 2021. PARSEC DDOMP Workbook Checklist. https://doi.org/10.5281/ZENODO.4909851

Suel, E., Polak, J.W., Bennett, J.E., Ezzati, M., 2019. Measuring social, environmental and health inequalities using deep learning and street imagery. Sci Rep 9, 6229. https://doi.org/10.1038/s41598-019-42036-w

Wilkinson, M.D., Dumontier, M., Aalbersberg, Ij.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., 't Hoen, P.A.C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B., 2016. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018. https://doi.org/10.1038/sdata.2016.18

Yeh, C., Perez, A., Driscoll, A., Azzari, G., Tang, Z., Lobell, D., Ermon, S., Burke, M., 2020. Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. Nat Commun 11, 2583. https://doi.org/10.1038/s41467-020-16185-w