CELPy: A detailed land cover dataset for Paraguay from 2017 to 2024

Kendra Walker, Lauren Sharwood,
Atahualpa Ayala, Robert Heilmayr

Mar 2025

Abstract

1 Background and summary

Outline of paragraphs:

1. General motivating problem statement: Review of land cover maps with crop species mapping - emphasize that much prior work has focused in developed country contexts (e.g. Europe, US). When crop mapping does occur in developing countries / South America, it may be biased towards industrial crops (e.g. Mapbiomas, Southern Cone project)
2. More specific motivating problem statement: Intro to Paraguay - review existing map products and highlight why another product is valuable. Highlight importance of Eastern paraguay for crop production, where many crops are grown in smallholder production systems.
3. General statement of solution: What are the new methods / sensors we draw upon to more accurately map land cover in paraguay? Provide a few background refs to highlight the potential of our methods (e.g. importance of spectral phenologies for crop differentiation, segmentation). Mention combination of targeted, multi-year field campaigns, and stratified random sample for rigorous validation.
4. Description of data product: Provide a summary of our maps. What years / geographic scope / classes? What is overarching accuracy?

2 Methods

2.1 data inputs

To produce annual crop maps of Paraguay at 10 m resolution, we ingest and process all Sentinel-2 and Landsat data for the country, (excluding Landsat-7

after 2017, due to its drifting trajectory [cite]). In recent years, this comprises about XX Landsat and XX Sentinel-2 views at each location per year (which equals around 5000 total scenes per year). We handle this large volume of data through a processing pipeline on a high-performance computing cluster that utilizes standardized products from a cloud-based STAC (SpatioTemporal Asset Catalog) within a gridded structure to facilitate parallelization. Processing tools are available and described in more detail at [pytuyau].

2.2 Preprocessing

Preprocessing tasks include image normalization in following with the harmonized Landsat/Sentinel project [1]. This includes:

cloud masking:

BRDF processing: to adjust for spectral differences that arise from atmospheric conditions as well as differences that have been noted along the sensor tracks [2].

co-registration: to ensure geographic alignment, each image is co-registered to a representative Landsat-8 image using open source AROSICS software [3].

2.3 Feature processing

All features are derived from spectral indices combining bands of the normalized imagery. We selected indices found in the literature to perform well in distinguishing crops in similar settings. These indices (defined in Table 2.3) include the Green Chlorophyll Vegetation Index (GCVI), Normalized Burn Ratio (NBR), Normalized Difference Moisture Index (NDMI)[4], Enhanced Vegetation Index (EVI2)[5], kernel-adjusted Normalized Difference Vegetation Index (kNDVI)[6], and Woody Index (WI)[7]. During preliminary steps of the modelling process, EVI2 and NBR were found to not increase performance beyond that achieved with the other four indices and were thus removed from our model. Each of the remaining four indices (GCVI,kNDVI,NBR & NDMI) contributes information from a different band (Green, Red, SWIR1 & SWIR2, respectively), normalized with NIR. These four indices were computed for all images and run through a dynamic time-series smoothing function [8] to reduce noise and condense the input data into evenly spaced observations at 10-day intervals. Datasets for annual modeling were constructed from images from 1-July to 30-June to avoid breaking apart the primary cropping season (roughly Nov-Mar).

Table 1: Indices used to build model features

index	equation	source
Normalized Burn Index	$\mathbf{NBR} = \frac{NIR - SWIR2}{NIR + SWIR2}$	
Normalized Difference Moisture Index	$\mathbf{NDMI} = \frac{NIR - SWIR1}{NIR + SWIR1}$	[4]
Green Chlorophyll Vegetation Index	$\mathbf{GCVI} = \frac{NIR}{Green} - 1$	
Kernel Normalized Difference Vegetation Index	$\mathbf{kNDVI} = \left(tanh\left(\frac{NIR - Red}{NIR + Red}\right)^2\right)$	[6]
Enhanced Vegetation Index 2	$\mathbf{EVI2} = 2.5 * \frac{NIR - Red}{1 + NIR + 2.4 * Red}$	[5]
Woody Index	$\mathbf{WI} = \begin{cases} 0.001 & \text{if (SWIR1+Red)} > 0.5 \\ 1 - \frac{SWIR1+Red}{0.5} & \text{otherwise} \end{cases}$	[7]

2.3.1 pixel-level features

Pixel-level features were generated directly from the smoothed time series for each index. The full pixel-based feature set used in this set of models include the maximum, minimum, average, median, amplitude, and standard deviation and coefficient of variation for each index for the full year, the wet season (Nov to Mar), and the dry season (May to Sept), as well as the phenological sequence represented by the smoothed value from the 20th day of each month. The kNDVI value at the peak of the wet season was also used. Other more complex phenological variables, such as the rate of greenup, rate of senescence, length of season, and date of peak of season were tested but found to not improve the model. Because all index values were smoothed in the time-series processing, measures commonly used to reduce noise, such as using percentile averages, were not deemed necessary. Through preliminary testing, other redundant pixel-based features were removed. The final set of pixel-level features used in the models presented here is shown in Table 2.3.1.

Table 2: Features used in model

	stats	index[a]	numvars
Annual stats	Max,Min,Amp,Avg,~~CV~~,Std	kNDVI, GCVI, NDMI, NBR	20
Wet season	Max,Min,~~Amp~~,~~Avg~~,CV,~~Std~~	kNDVI, GCVI, NDMI, NBR	16
Dry season	Max,Min,~~Amp~~,~~Avg~~,CV,~~Std~~	kNDVI, GCVI, NDMI, NBR	16
Monthly	20th of each month	kNDVI, GCVI, NDMI, NBR	48
Pheno	posv_wet (value from peak of season)	kNDVI	1
Polygon	Ext,Dist,Bounds,Area,APrEf,WetStd	(all 4 in segmentation training)	6
Ancillary	Forest Strata (4 biome flags)	NA	4

[a] after time series smoothing

2.3.2 Polygon features (Field segmentation)

The model includes six features that were estimated from a field segmentation process. These include three pixel-level features: probability that a pixel is within a crop field (crop extent), distance from the field edge (if crop extent is not zero), and probability of falling on the border of a crop field, and three field-level features: field area, field homogeneity (the standard deviation of the average Nov/Dec GCVI value (after smoothing) for all pixels in a given polygon), and the field's area-to-perimeter efficiency. The perimeter efficiency measure captures the deviation from a square and is helpful in identifying polygons that likely contain multiple smaller fields.

The field segmentation process is trained with digitized field boundaries and corresponding bimonthly observations from the smoothed time series of the four spectral indices. Field boundary training data was created for 1026 1 km x 1 km sample chips, of which 826 were randomly sampled across Paraguay (700 in eastern Paraguay and 100 in western Paraguay) and 200 were added in areas with more smallholder farming activity. Within these chips, all crop fields were manually digitized in ArcGIS Pro, using high-resolution imagery (from ESRI basemap, Google Earth, or Maxar) along with spectral profiles from Planet NICFI monthly NDVI mosaics to determine presence of crop. We used our processed 10 m Sentinel-2 imagery to inform boundary decisions, as this is the same resolution as the intended model output.

The digitized training chips and associated time-series data were converted to PyTorch training data, with 32 image filters applied to artificially augment the dataset. This set was in turn used to to train a convolutional neural network and run model inference on a GPU on a high-performance computing cluster. We built the model using Cultionet, an open-source image segmentation library [9] based on work by Waldner and Diakogiannis [10]. Cultionet employs Unet 3+ architecture, commonly used in medical image segmentation, which utilizes deep supervisions and skip connections to take in full-scale semantic information from input images [11, 10]. The model also incorporates a single encoder and three parallel decoders, where one decoder learns the crop-extent prediction and the other two decoders learn the auxiliary tasks of contour detection and distance map estimation, capturing shape and boundary information [12, 10]. Tanimoto loss was used during training to update parameter values [13, 10]. We used 15 epochs for model training, after finding model loss, validation loss, crop class F score, and boundary class F score statistics to stabilize around this number in preliminary testing. The output is a three band inference composite containing the three pixel-level polygon features (crop extent probability, distance to border and border probability).

From the three raster outputs, we derived the three field-level measures by extracting polygon vectors for each field. For vector extraction, we used a simple contour method in which the boundary raster was subtracted from the extent raster plus 1 and all values above a user-defined threshold were converted to a single value. We tested an alternative watershed segmentation method [10] with the theory that it could capture smaller fields due to the user-defined seed

parameter that influences which fields are captured. In preliminary tests, however, we found this watershed method to overestimate crop in areas without crop, particularly wetlands, and used the contour method as our final vectorization method. Our final set of vectors was found to miss many very small fields. This is partially because of an intrinsic minimum mapping unit (MMU) of 30 m x 30 m due to the requirement of Cultionet that a crop encompasses at least three pixels to be detected (as both crop extent and the border pixels need to be registered). Fields smaller than this mmu were grouped for digitization, following a set of rules to try to capture individual parcels as best possible. The shifting configuration of smallholder fields throughout the year likely makes these tiny fields and even many smallholder fields larger than the MMU difficult to capture with an annual segmentation model.

2.3.3 biome data

The Western Chaco area is much drier than the Eastern Atlantic Forest region. Thus, in addition to the pixel and polygon features described above, we included one ancillary layer comprising the four major biome regions in Paraguay to inform landscape variability.

2.4 Land cover model

Our full model includes 35 classes, summarized in Table 1, with details for each class in the Look-up Table provided with the dataset. The 35 classes include 15 crop classes and 20 non-crop classes. Of the 15 crop classes, six (soy, corn, wheat, sugar, rice, and cotton) are annual crops frequently grown on large fields, five (banana, yerba-mate, grapes, orchard fruits, and hemp/cannabis) are perennial crops, and three (mandioca (cassava), sesame, and horticulture), are generally grown in smallholder systems. A final mixed-crop category was included to represent situations common to smallholder systems, where multiple crops are observed in a pixel. Three additional crops (beans, peanuts, and tobacco) were added to the mixed-crop class because they were observed too infrequently in the ground data to create reliable training data and could not be distinguished in high-resolution imagery. While training data were collected for all 15 crop classes, these classes were condensed into ten classes in the default model by combining the three smallholder crops with the mixed crops and by combining soy, corn, and wheat into a single class. The latter were combined because they commonly occur on the same field at different times in the year, and different cropping calendars with two to four cycles per year complicate efforts to identify a single "main" crop from available high-resolution imagery.

Crop points were positioned at least 10 m from the edge of the field when possible to facilitate the mapping of crop types. To ensure representation of edge pixels in the model, an additional crop_edge class was created where pixels within 10 m of a field boundary were targeted specifically. With mixed vegetation included in only two categories, mixed-crop and crop_edge, early models were found to exaggerate crop area where mixtures of vegetation and bare soil

occur, particularly along paths and at the edges of pastures. To compensate, these two mixed classes (paths and grass_edge) were also included in the training data. Other non-crop land cover types were determined in a similar iterative manner, with many added during preliminary modelling stages to provide comparison for non-crop classes often confused with crops. For example, early models showed high confusion between perennial crops such as bananas and yerba-mate and other types of natural vegetation such as palm forest and different natural shrub compositions, thus inspiring the addition of these non-crop classes. Distinct regional differences between the Chaco in western Paraguay and the wetter biomes in eastern Paraguay, where the original training data were collected, also drove the inclusion of additional forest classes to more accurately represent shrub forest and other drier forests. These forest classes were informed by supplementary ground data provided by (cite Yann), which we then augmented with points that appeared similar in WorldView images.

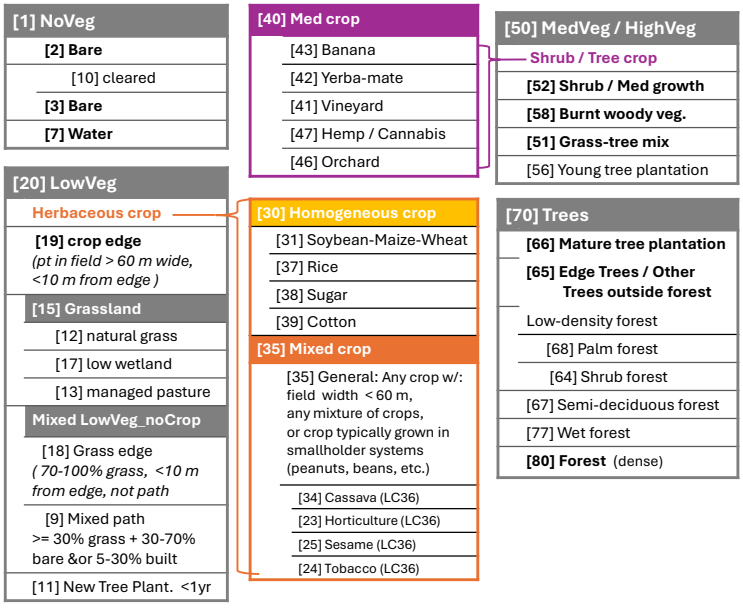

Figure 1: CELPy classes. Classes with white text are for aggregation purposes and are not included in the final maps. A Look-up Table is provided with the dataset with the full set of classes and alternative groupings

2.4.1 Land cover training data

The single-year training data available for the model (Table 2.4.1) consists of 1878 samples from the ground focused on crop-type identification, 4500 field point samples from large commercial fields from the Ministry of Agriculture, 583 point samples from a farm in western Paraguay growing a diverse portfolio of crops, 100 ground samples from another area of western Paraguay that was underrepresented in our other datasets, and more than 11000 points with classes interpreted from high resolution imagery (WorldView or similar imagery in Google Earth). XXXXX of the single-year points were viewed at all temporal points available in Google Earth and interpolated at unobserved points in time based on a set of logical rules to produce a multi-year training set.

training data from ground sample: Ground sampling was conducted during the primary growing season (Dec-Mar) and was focused on identification of crop type, especially targeting areas with diverse smallholder activity. Ground samples for training were selected opportunistically in that they were confined to points observable from roads navigable in the wet season. As points were viewed from the edge of the road, they were also confined to locations where the land cover class 15 m away could be seen and interpreted unambiguously (i.e. away from class edges).

training data from high-resolution imagery: High-resolution imagery was used to interpret land covers that could be assessed with confidence at the available resolution. Crops can usually be identified as crop, but crop types cannot be assessed with confidence unless specific conditions occur in the image (such as conspicuous harvest patterns). Most non-crop land covers in our model can be assessed with high confidence by a well-trained observer. For a portion of the landscape in western Paraguay that was less familiar to us, we procured georeferenced images from (cite Yann) to understand the landscape and more confidently label classes.

Table 3: Training data source

	N	focus	source
Ground sample 2022	1287	crop type in E.Py 2022 primary growing season	our data
Ground sample 2024	527	crop type in W.Py 2024 primary growing season	our data
MAG (around 2021)	4500	industrial crops (soy, rice, sugar)	MAG
HighRes	11062[a]	multi-year obs. of landscape in WorldView and Google Earth	our data
Chaco_farm	583	multi-year, multi-crop farm portfolio in W.Py	??
Chaco_other	100	ground visits of underrepresented W.Py landscape	Yann

[a] for single year. Multiyear distribution shown in Table 4.

The final set of points used to train the models was generated based on a set of balancing rules. For each year, a targeted minimum of 200 samples for each of the 36 land cover classes was included, although this was not always possible for more rare classes. The class with the maximum prevalence (forest) was allotted the maximum sample size, with other classes allotted sample sizes

in proportion to their estimated prevalence in a preliminary map. Mixed crop and other mixed classes were allotted a sample three times their prevalence, as determined optimal in testing. The training sample for each year in the time series is balanced independently based on the available samples for that year. All samples are then combined into the multi-year training set. A separate 2021/22-only training set was used for the 2022 smallholder map to maintain consistency with other work.

2.5 Classification

We use a random forest classifier to predict 32 classes from a feature space comprised of the variables in Table 2.3.1.

2.6 Post-Model filters

2.6.1 agricultural filter

to remove smallholder rings where vegetation changes, to fill fields with majority crop

2.6.2 temporal filter

to remove illogical transitions likely due to cloud noise

3 Data Records

Seven-year time series from random forest model trained with samples from 2017-2024, with post-model temporal filter applied across series:

- CELpy2024 - Landcover map of Paraguay for primary 2024 growing season
- CELpy2023 - Landcover map of Paraguay for primary 2023 growing season
- CELpy2022 - Landcover map of Paraguay for primary 2022 growing season
- CELpy2021 - Landcover map of Paraguay for primary 2021 growing season
- CELpy2020 - Landcover map of Paraguay for primary 2020 growing season
- CELpy2019 - Landcover map of Paraguay for primary 2019 growing season
- CELpy2018 - Landcover map of Paraguay for primary 2018 growing season

Combined Landcover map of Paraguay for 2024 using five previous years to inform shifting cropping systems

- CELpy_AgSystems_2024

Stand-alone map for 2022 focused on quantifying smallholder agriculture (trained with samples from 2021-2022 only, with filter applied to classify all fields <5 ha (besides industrial sugar) as smallholder):

- CELpy_smallholder_2022

Lookup Table (.csv table) with classification values, names, and descriptions for each class, along with suggested re-groupings for different purposes:

- CELpyLUT

4 Technical Validation

Our most recent crop maps are validated with a holdout sample of XXX points visited on the ground during the main growing season of 2023 (Nov 2022 - Mar 2023) and XXXX points visited the ground during the main growing season of 2024 (Dec 2023 - Mar 2024). While crop-type mapping requires ground sampling to determine crop types, we can distinguish between presence of crops and other land covers with high certainty using high-resolution imagery. Our ground points are thus supplemented with XXXX random points viewed with high resolution imagery on Google Earth at multiple points in time. Sample points were used for annual map validation if they were observed during that year or could be confidently classified for that year bases on other observations and adherence to a set of rules. For some areas that were not well represented in Google Earth for a critical time period, we acquired WorldView images from Maxar to fill gaps and improve spatial representation of the sample.

5 Usage Notes

optional additional technical notes about how to access or process the data

6 Code availability

Our full processing pipeline is available at: (pytuyau) This pipeline makes use of several open-source packages maintained by Jordan Graesser:

- geowombat – raster I/O and general utilities with Xarray/Dask/Rasterio `https://github.com/jgrss/geowombat`
- satsmooth - time series reconstruction `https://github.com/jgrss/satsmooth`
- cultionet – crop segmentation with a neural network `https://github.com/jgrss/cultionet`
- eostac – STAC download and radiometric normalization `https://github.com/jgrss/eostac`

Additional tools are available at: `https://github.com/klwalker-sb/LUCinSA_helpers`

A guide to the processing flow on an HPC environment is available at: `https://klwalker-sb.github.io/LUCinLA_stac/`

6.1 Acknowledgements

This material is based upon work supported by the National Aeronautics and Space Administration under Grant No. 80NSSC20K1489 issued through the Land Cover and Land Use Change Program. The authors thank staff at Paraguay's Ministerio de Agricultura y Ganadería and the Servicio Nacional de Catastro for data access. In addition, we thank Keyla Morales and Conner Edwards for her help in digitizing field boundaries and labeling imagery for validation, Ryan Ashraf for help with model-building and optimization, and Camila Berger for help collecting field data.

6.2 Author contributions

KW: Conceptualization, Methodology, Investigation, Formal analysis, Data Curation, Software, Supervision, Writing - Original Draft **LS**: Methodology, Data Curation **AA**: Investigation, Data Curation, Resources **JG**: Conceptualization,, Methodology, Software **RH**: Conceptualization, Funding acquisition, Resources, Supervision, Writing - Review & Editing

6.3 Competing interest

The authors declare no competing interests.

References

[1] Martin Claverie, Junchang Ju, Jeffrey G. Masek, Jennifer L. Dungan, Eric F. Vermote, Jean-Claude Roger, Sergii V. Skakun, and Christopher O. Justice. The harmonized landsat and sentinel-2 surface reflectance data set. *Remote Sensing of Environment*, 219:145–161, 2018.

[2] D. P. Roy, H. K. Zhang, J. Ju, J. L. Gomez-Dans, P. E. Lewis, C. B. Schaaf, Q. Sun, J. Li, H. Huang, and V. Kovalskyy. A general method to normalize landsat reflectance data to nadir brdf adjusted reflectance. *Remote Sensing of Environment*, 176:255–271, 2016.

[3] Daniel Scheffler, David Frantz, and Karl Segl. Spectral harmonization and red edge prediction of landsat-8 to sentinel-2 using land cover optimized multivariate regressors. *Remote Sensing of Environment*, 241:111723, 2020.

[4] B. C. Gao. Ndwi - a normalized difference water index for remote sensing of vegetation liquid water from space. *Remote Sensing of Environment*, 58(3):257–266, 1996.

[5] Zhangyan Jiang, Alfredo R. Huete, Kamel Didan, and Tomoaki Miura. Development of a two-band enhanced vegetation index without a blue band. *Remote Sensing of Environment*, 112(10):3833–3845, 2008.

[6] G. Camps-Valls, M. Campos-Taberner, A. Moreno-Martínez, S. Walther, G. Duveiller, A. Cescatti, M. D. Mahecha, J. Muñoz-Marí, F. J. García-Haro, L. Guanter, M. Jung, J. A. Gamon, M. Reichstein, and S. W. Running. A unified vegetation index for quantifying the terrestrial biosphere. *Science Advances*, 7(9), 2021.

[7] Eric A. Lehmann, Jeremy F. Wallace, Peter A. Caccetta, Suzanne L. Furby, and Katherine Zdunic. Forest cover trends from time series landsat data for the australian continent. *International Journal of Applied Earth Observation and Geoinformation*, 21:453–462, 2013.

[8] J. Graesser, R. Stanimirova, and M. A. Friedl. Reconstruction of satellite time series with a dynamic smoother. *Ieee Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:1803–1813, 2022.

[9] et.al Graesser, Jordan. Cultionet. `https://github.com/jgrss/cultionet`. Accessed: 2023-11-20.

[10] François Waldner and Foivos I. Diakogiannis. Deep learning on edge: Extracting field boundaries from satellite images with a convolutional neural network. *Remote Sensing of Environment*, 245:111741, 2020.

[11] Huimin Huang, Lanfen Lin, Ruofeng Tong, Hongjie Hu, Qiaowei Zhang, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, and Jian Wu. Unet 3+: A full-scale connected unet for medical image segmentation. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1055–1059, 2020.

[12] Balamurali Murugesan, Kaushik Sarveswaran, Sharath M Shankaranarayana, Keerthi Ram, Jayaraj Joseph, and Mohanasankar Sivaprakasam. Psi-net: Shape and boundary aware joint multi-task deep network for medical image segmentation. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 7223–7226, 2019.

[13] Foivos I. Diakogiannis, François Waldner, and Peter Caccetta. Looking for change? roll the dice and demand attention. *Remote Sensing*, 13(18), 2021.