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

Kendra Walker, Lauren Sharwood,
Atahualpa Ayala, Robert Heilmayr

May 2025

Abstract

1 Background and summary

There is increasing need for timely and consistent maps of agricultural systems to understand impacts of policies and agricultural activities on the environment as well as impacts of environmental and climate change on agricultural production. To serve such purposes, agricultural maps need to be not only detailed in their ability to represent agricultural practices and crop types across space but also to convey dynamic processes across time to inform policy-relevant questions such how and where change occurs along agricultural frontiers [1], how agricultural practices affect the environment, and how food systems can adapt to climate and economic change shocks or long term change [2]. While time series maps of agriculture are available in increasing detail for industrialized areas such as the United States and Europe, such maps are less comprehensive for most developing countries, where agricultural frontiers are most active and food insecurity most prevalent. Although there are several existing 10 m resolution crop maps at global scale as well as for specific regions of the developing south, these maps are generally either very generalized in their representation of croplands as common industrial prototypes or very localized snapshot products.

Paraguay provides an interesting landscape for crop mapping because it contains large industrialized agricultural landscapes similar to those found in Europe, as well as large areas of smallholder activity similar to those observed in other developing nations. Global products that target specific industrial crops, such as the world soy map [3], or the ESA world cereal map [4], do a good job of capturing the industrial crops in Paraguay, but do not show much of the smallholder crop areas as cropland. While this might be expected for these maps, since many of the crops grown by smallholders are not the targeted crops, similar results are seen with more general global crop maps (e.g. [5]), as well as with regional maps that include agriculture (e.g. [6, 7]). Producing

a map that can capture agricultural change in Paraguay requires not only the ability to differentiate cropping systems and crop types, but to do so across a diverse landscape with a distinct rainfall gradient. The eastern region of the country, where the vast majority of crops are grown, is humid with abundant rainfall, while the western Chaco region contains semi-arid landscape. Although a small percentage of Paraguay's crops are grown in the west, the rapid rate of deforestation of the Chaco region for agriculture (especially cattle ranching, but increasingly crops as well) has been the topic of much research [8, 9, 10, 11, 12]. To understand processes of change related to agriculture, maps need to provide detail in not only agricultural categories but also other land cover types for context.

Publicly available satellite data from the well-established Landsat and Sentinel missions are ideal for long-term analyses and monitoring and have been found effective in distinguishing various land covers as well as agricultural systems down to crop type when combined with machine learning methods. Class distinction can be maximized by reducing input noise through processes such as those used in Harmonized Landsat/Sentinel products [13] and time series smoothing [14], leveraging the full time series to produce phenological features [15, 16], using edge detection and segmentation methods to locate fields and minimize pixel-level noise [17, 16, 18], using a robust classifier such as random forest [19, 8], training the classifier with location- and system-specific data [20], and applying post-classification probabilistic techniques to stabilize time-series observations [21, 22]. By combining these methods in a data-cube framework with the leverage of high-performance computing, we are able to produce maps that distinguish 34 different land cover classes, including several crop species, across a diverse landscape within both large commercial and smallholder production systems.

The resulting product is presented here as a set of national maps of Paraguay from 2017-2024 at 10 m resolution that distinguish 34 land cover categories. The focus of these maps is to depict agricultural land and processes, but other land covers are differentiated both to facilitate processing and to provide contextual information. Along with crops, accuracy in mapping forest cover was prioritized to facilitate forest-cover change analyses. The maps are validated with both a ground-based sample focused on crop type and a stratified random sample focused on assessing the ability to map both industrial cropland and the more rare smallholder fields.

2 Methods

2.1 data inputs

To produce annual crop maps of Paraguay at 10 m resolution, we ingested and processed all available Sentinel-2 and Landsat data for the country, (excluding Landsat-7 after 2017, due to its drifting trajectory [23]). In recent years, this comprises about 80 Landsat and 100 Sentinel-2 views at each location per

year, or around 7000 scenes for all of Paraguay per year and more than 50,000 scenes for the 2017-2024 period. 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. The complete processing workflow is shown in 1. All processing tools are available and described in more detail at [pytuyau].

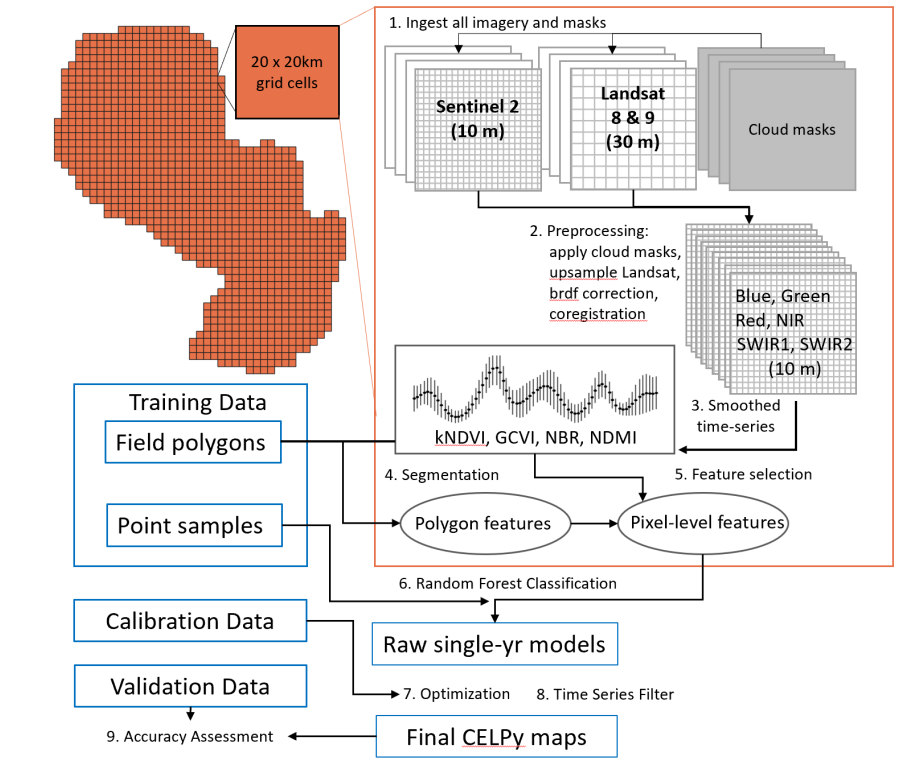

Figure 1: Processing overview

2.2 Preprocessing

Our preprocessing pipeline includes tasks to normalize input imagery in following with the harmonized Landsat/Sentinel project [13]. This includes:

cloud masking: We used Landsat collection-2 level-2 and Sentinel-2 level 2A products, both of which are already atmospherically corrected. Both include a native cloud mask (the "SCL" asset for Sentinel-2 and the "cloud_qa" asset for Landsat), which we applied to each band to remove clouds.

Landsat upsampling: The 30 m Landsat imagery was upsampled to 10 m to match the Sentinel-2 imagery using cubic convolution.

BRDF normalization: We applied Bidirectional Reflectance Distribution Function (BRDF) normalization to adjust for spectral differences between sensors as well as between pixels at different positions of the viewing track [24].

co-registration: To ensure geographic alignment, each image was co-registered to a representative Landsat-8 image using open source AROSICS software [25].

2.3 Feature processing

The full pre-processed datacubes were converted to smoothed time-series observations prior to feature generation. To maximize the power of the time series observations, raw band values were converted to spectral 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)[26], Enhanced Vegetation Index (EVI2)[27], kernel-adjusted Normalized Difference Vegetation Index (kNDVI)[28], and Woody Index (WI)[29]. During preliminary steps of the modeling 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 [30] 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 interrupting the primary cropping season (roughly Nov-Mar).

Table 1: Indices used to build model features

index	equation	source
Normalized Burn Index	**NBR** = $\frac{NIR - SWIR2}{NIR + SWIR2}$	
Normalized Difference Moisture Index	**NDMI** = $\frac{NIR - SWIR1}{NIR + SWIR1}$	[26]
Green Chlorophyll Vegetation Index	**GCVI** = $\frac{NIR}{Green} - 1$	
Kernel Normalized Difference Vegetation Index	**kNDVI** = $\left(tanh\left(\frac{NIR - Red}{NIR + Red}\right)^2\right)$	[28]
Enhanced Vegetation Index 2	**EVI2** = $2.5 * \frac{NIR - Red}{1 + NIR + 2.4 * Red}$	[27]
Woody Index	**WI** = $\begin{cases} 0.001 & \text{if (SWIR1+Red)} > 0.5 \\ 1 - \frac{SWIR1+Red}{0.5} & \text{otherwise} \end{cases}$	[29]

2.3.1 pixel-level features

Pixel-level features were generated directly from the smoothed time series for each index as well as from the polygon features described in 2.3.2. 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 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,Edge,Area,APrEf,WetStd	(all 4 in segmentation training)	6
Ancillary	Forest Strata (4 biome flags)	NA	4

entries with strikeout indicate that these were removed from the final model following optimization tests.
[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 (Ext), distance from the field edge (Dist), and probability of falling on the border of a crop field (Edge), and three field-level features: field area (Area), field homogeneity in the primary growing season (WetStd), and the field's area-to-perimeter efficiency (APrEf). Field homogeneity is estimated as the standard deviation across pixels in a given polygon of the average November/December GCVI value (after the time-series smoothing step for each pixel. The APrEf is $perimeter/4\sqrt{area}$ and is helpful in identifying polygons that were not fully segmented and likely contain multiple smaller fields. While an APrEf value of 1.0 indicates a perfect square, values up to 2.5 are often observed for single fields, as shown in Fig. 2, as fields are rarely actually square. We find APrEf values beyond 2.5 to be suggestive of multiple fields, with values greater than 3.0 almost never observed for single fields.

The field segmentation process was 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 [31] based on work by Waldner and Diakogiannis [32]. 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 [33, 32]. 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 [34, 32]. Tanimoto loss was used during training to update parameter values [35, 32]. 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 [32] 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. The four biomes are: The Dry Chaco, covering most of western Paraguay, the Humid Alantic Forest, covering most of eastern Paraguay, an intermediate sub-humid region comprising palm savannas and flood plains, and a smaller area of Cerrado.

2.4 Classification schema

Our full model includes 36 classes, summarized in Table 4, with details for each class in the Look-up Table provided with the dataset. The 36 classes include 14 crop classes and 22 non-crop classes. Of the 14 crop classes, six (soy, corn, wheat,

Figure 2: Illustration of Area-Perimeter Efficiency (APrEF) measure for segmentation data to identify undersegmented polygons. Values close to one indicate fields that approximate squares while values up to 2.5 are considered normal for single fields. The highlighted polygon is clearly multiple fields, with an APrEF value of 6.75.

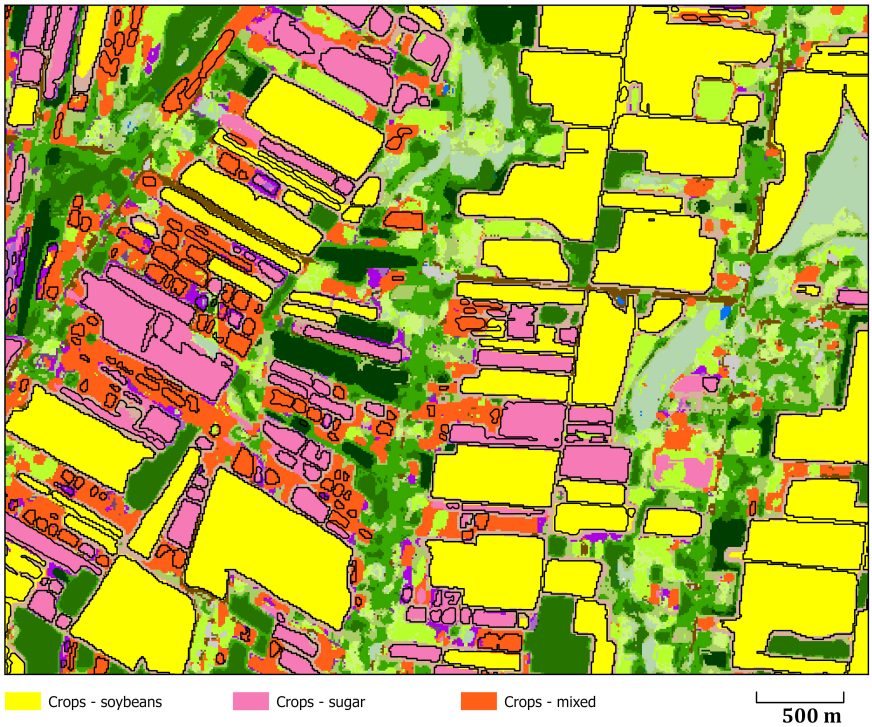

Figure 3: Example of field segmentation (black lines) overlaid on classified pixels, showing difficulty in segmenting very small mixed crop pixels. All colors not shown in the legend are non-crop vegetation.

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 (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. Crops generally observed in smallholder systems, such as beans, peanuts, and tobacco, were added to the mixed-crop class due to inadequate ground data to create individual training classes. In the default model, this mixed-crop category was expanded to include the three other smallholder crops (cassava, sesame, and horticulture). Soy, corn, and wheat were also combined into a single class in the default model because they commonly occur on the same field at different times in the year, and efforts to identify a single "main" crop from available high-resolution imagery are complicated by different cropping calendars with two to four cycles per year. Thus, while training data were collected for 14 crop classes, these classes were condensed into ten classes in the default model.

Edges around homogeneous crop fields were included as a separate category to avoid these areas all being classified as mixed crop and thus impeding quantification of smallholder systems. Pixels observed to be homogeneous crop, but within 10 m of a field boundary, were targeted included in the crop_edge class. Even with this additional class, early models were found to exaggerate crop area in other areas 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 modeling 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. The distinct regional differences between Chaco in western Paraguay and wetter biomes in eastern Paraguay, where the original training data were collected, also drove the inclusion of additional forest classes to better represent shrub forests and other dry 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.

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

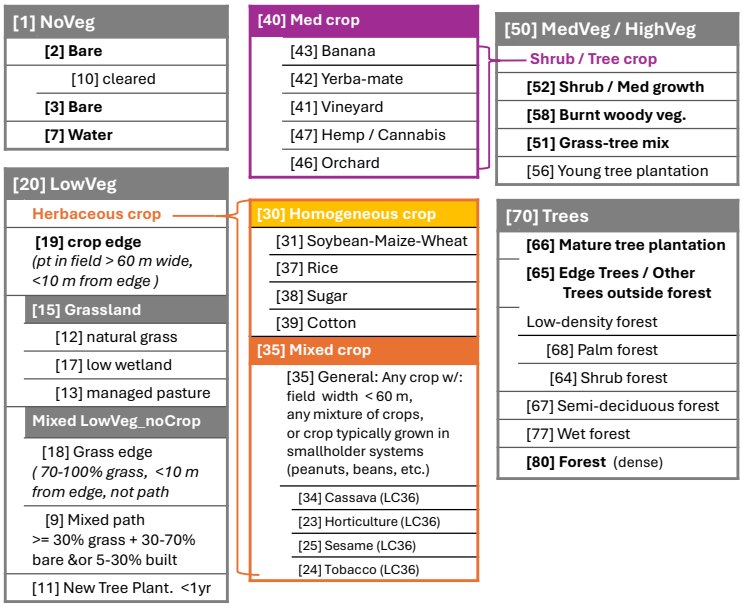

Figure 4: CELPy classes. Classes with white text are for aggregation purposes and are not included in the final maps. A Look-up Table is provided with the dataset with the full set of classes and alternative groupings

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: Data sets for training (tr), calibration (cal) validation (val)

use	name	N	description	source
tr	SegChips	1026	Digitized polygons in 1100 1 km x 1 km chips to train segmentation model	our data
tr/cal	Ground_E2022	1287	Ground observations of crop type in E.Py 2022 primary growing season	our data
tr/cal	Ground_W2024	527	Ground observations of crop type in W.Py 2024 primary growing season	our data
tr/cal	MAG	4500	Industrial crops (soy, rice, sugar) from government data	MAG
tr/cal	HighRes_active	11062[a]	Multi-year obs. of landscape in WorldView and Google Earth selected via active learning methods	our data
tr/cal	Chaco_farm	583	Multi-year, multi-crop farm portfolio in W.Py	??
tr/cal	Chaco_lc	100	Ground visits of underrepresented landscape in W. Paraguay	Yann
cal	DistrictSamp	5100	Multi-year observations in Google Earth of 300 random points in 17 sample districts, for smallholder map	our data
val	Ground_E2024	2500	Ground observations of crop type in E.Py 2024 primary growing season	our data
val	Ground_W2024	855	Ground observations of crop type in W.Py 2024 primary growing season	our data
val	HighRes_rand		Multi-year observations in Google Earth of stratified random sample	our data

[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 used 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 Crop filter

Individual fields were homogenized to a single crop (or mixed crop) class based on the segmentation data and rings of mixed crop around higher vegetation are removed based on the following rules:

- **witin segmented crop polygons (with 20 m buffer applied):** All pixels classified as crop were reclassified to the majority crop class for that polygon. If a pixel was classified as mixed crop but was in a polygon with a different majority crop, it was reclassified to crop_edge if within 20 m of the field perimeter (shown in Fig.5a, and the majority crop if deeper within the field.

- **outside segmented crop polygons (with 20 m buffer applied):** Pixels classified as crops within smallholder areas that did not receive segmentation data were assumed to be very small fields (that do not segment well) and assigned to the mixed-crop class. Smallholder areas are defined as areas where average field size is smaller than 5 ha, and identified with a 100-pixel (1-ha) moving window on the field size output from the segmentation data. Some of these pixels are corrected to non-crop in the following step.

- **within bands of (likely non-crop) mixed vegetation along class edges:** Rings of misclassified mixed crop resulting from mixtures of high vegetation and grass were removed by reclassifying mixed-crop pixels to shrub if adjacent to higher, non-crop vegetation and not also adjacent to 3x3 pixel block of mixed-crop, as illustrated in Fig.5b.

- **smallholder crops – For the "CELpy_smallholder_2022" product quantifying smallholder crops only**: all remaining mixed crop and crop in fields smaller than 5 ha (or sugar smaller than 3 ha) were reclassified to smallholder crop based on the following rules: 1) pixels classified as crop but outside of segmented polygons (with 20 m buffer) are assigned to the smallholder crop class. 2) Crop pixels within segmented polygons are assigned field sizes based the polygon area or the area/APrEf if APrEf is greater than 2.5 (to correct for errors of incomplete segmentation of smaller fields). Crop pixels with resulting field areas smaller than 5 ha (or 3 ha for sugar) were reclassified to the smallholder class.

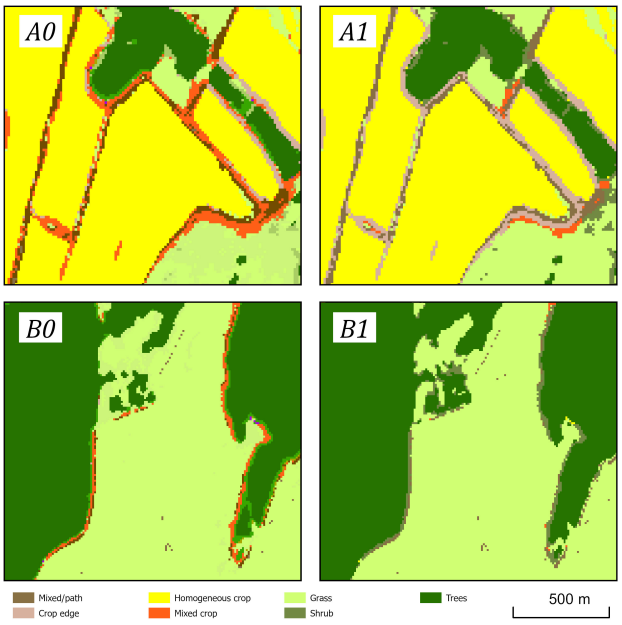

Figure 5: Examples of A) Mixed crop to crop edge filter within segmented polygons, B) Mixed crop to shrub filter around the edge of trees. A0 and B0 are pre-filtered results while A1 and B1 are the final products.

2.6.2 Temporal filter

For the seven time-series products, an additional temporal filter was applied to further harmonize the data and restrict transitions to plausible categories in the following cases:

- **generally stable classes sensitive to inter-annual climatic variance:** To prevent abrupt switching between forest types or grass types, pixels classified as grass or tree were assigned the majority sub-class for that category for the 7-yr sequence.
- **generally stable categories sensitive to compositional changes:** some classes are easily confused with other classes at certain stages of growth. Easily confused sets include medium crops and tree plantation, banana and palm forest, and sugar and palm forest. Pixels classified as one of these classes were assigned the majority class for the set for the 7-yr sequence.
- **generally stable categories sensitive to positional inconsistencies:** Although positional differences in input imagery were corrected for in our pre-processing steps, slight differences can still result in different classifications for edge pixels. This was commonly observed in pixels along water, which can switch between water, wet grassland, wet medium vegetation,

or wet forest depending on slight positional differences in the imagery. A pixel classified as one of these wet classes was assigned the majority class for the 7-yr time-series unless the classified class persisted for at least three years in a row (suggesting true change). Similarly, the changes between palm savannas (grass-tree mix) and palm forests are likely due to small positional or climatic differences rather than actual change, thus the same rule was applied for these classes.

- **probable noise from unmasked clouds:** Pixels classified as no or low vegetation for a single year, surrounded by a forest classification for the prior and subsequent years, were assumed to be noise from unmasked clouds and reclassified to match the following year.
- **illogical growth sequences:** Pixels classified as trees (not plantation or crop) that were classified as no vegetation or low vegetation in the previous year (and not corrected with the rule above)) were reclassified as medium vegetation.

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 **RH**: Conceptualization, Funding acquisition, Resources, Supervision, Writing - Original Draft

6.3 Competing interest

The authors declare no competing interests.

References

[1] José A. Marengo, Juan C. Jimenez, Jhan-Carlo Espinoza, Ana Paula Cunha, and Luiz E. O. Aragão. Increased climate pressure on the agricultural frontier in the eastern amazonia–cerrado transition zone. *Scientific Reports*, 12(1):457, 2022.

[2] Krisha Lim, Bruno Wichmann, Martin K. Luckert, and Peter Läderach. Impacts of smallholder agricultural adaptation on food security: evidence from africa, asia, and central america. *Food Security*, 12(1):21–35, 2020.

[3] Xiao-Peng Song, Matthew C. Hansen, Peter V. Potapov, Bernard Adusei, Jeffrey Pickering, Marcos Adami, Andre Lima, Viviana Zalles, Stephen V. Stehman, Carlos M. Di Bella, Maria C. Conde, Esteban J. Copati, Lucas B. Fernandes, Andres Hernandez-Serna, Samuel M. Jantz, Amy H. Pickens, Svetlana Turubanova, and Alexandra Tyukavina. Massive soybean expansion in south america since 2000 and implications for conservation. *Nature Sustainability*, 4(9):784–792, 2021.

[4] Kristof Van Tricht, Jeroen Degerickx, Sven Gilliams, Daniele Zanaga, Marjorie Battude, Alex Grosu, Joost Brombacher, Myroslava Lesiv, Juan Carlos Laso Bayas, Santosh Karanam, et al. Worldcereal: a dynamic open-source system for global-scale, seasonal, and reproducible crop and irrigation mapping. *Earth System Science Data*, 15(12):5491–5515, 2023.

[5] D. Zanaga, R. Van De Kerchove, D. Daems, W. De Keersmaecker, C. Brockmann, G. Kirches, J. Wevers, O. Cartus, M. Santoro, S. Fritz, M. Lesiv, M. Herold, N.-E. Tsendbazar, P. Xu, F. Ramoino, and O. Arino. Esa worldcover 10 m 2021 v200, 2022. Accessed at: https://doi.org/10.5281/zenodo.7254221.

[6] MapBiomas. Mapbiomas trinational atlantic forest project – collection 2, version 1 of the annual coverage and land use series, 2022. Accessed at: https://gee-community-catalog.org/projects/mapbiomas.

[7] J. Graesser, R. Stanimirova, K. Tarrio, E. J. Copati, J. N. Volante, S. R. Veron, S. Banchero, H. Elena, D. de Abelleyra, and M. A. Friedl. Temporally-consistent annual land cover from landsat time series in the southern cone of south america. *Remote Sensing*, 14(16), 2022.

[8] M. Baumann, C. Israel, M. Piquer-Rodriguez, G. Gavier-Pizarro, J. N. Volante, and T. Kuemmerle. Deforestation and cattle expansion in the paraguayan chaco 1987-2012. *Regional Environmental Change*, 17(4):1179–1191, 2017.

[9] E. Da Ponte, M. Garcia-Calabrese, J. Kriese, N. Cabral, L. P. de Molas, M. Alvarenga, A. Caceres, A. Gali, V. Garcia, L. Morinigo, M. Rios, and A. Salinas. Understanding 34 years of forest cover dynamics across the paraguayan chaco: Characterizing annual changes and forest fragmentation levels between 1987 and 2020. *Forests*, 13(1), 2022. Da Ponte, Emmanuel Garcia-Calabrese, Monserrat Kriese, Jennifer Cabral, Nestor Perez de Molas, Lidia Alvarenga, Magali Caceres, Arami Gali, Alicia Garcia, Vanina Morinigo, Luis Rios, Macarena Salinas, Alejandro Perez de Molas, Lidia Florencia/0000-0001-7649-0585; Da Ponte, Emmanuel/0000-0002-5354-0364 1999-4907.

[10] C. Levers, M. Piquer-Rodríguez, F. Gollnow, M. Baumann, M. Camino, N. I. Gasparri, G. I. Gavier-Pizarro, Y. L. de Waroux, D. Müller, J. Nori, F. Pötzschner, A. Romero-Muñoz, and T. Kuemmerle. What is still at stake in the gran chaco? social-ecological impacts of alternative land-system futures in a global deforestation hotspot. *Environmental Research Letters*, 19(6), 2024.

[11] Yann Le Polain de Waroux, Rachael D Garrett, Robert Heilmayr, and Eric F Lambin. Land-use policies and corporate investments in agriculture in the gran chaco and chiquitano. *Proceedings of the National Academy of Sciences*, 113(15):4021–4026, 2016.

[12] M. F. Mereles and O. Rodas. Assessment of rates of deforestation classes in the paraguayan chaco (great south american chaco) with comments on the vulnerability of forests fragments to climate change. *Climatic Change*, 127(1):55–71, 2014.

[13] 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.

[14] Yang Shao, Ross S. Lunetta, Brandon Wheeler, John S. Iiames, and James B. Campbell. An evaluation of time-series smoothing algorithms for

land-cover classifications using modis-ndvi multi-temporal data. *Remote Sensing of Environment*, 174:258–265, 2016.

[15] Zunyi Xie, Yan Zhao, Ruizhu Jiang, Miao Zhang, Graeme Hammer, Scott Chapman, Jason Brider, and Andries B. Potgieter. Seasonal dynamics of fallow and cropping lands in the broadacre cropping region of australia. *Remote Sensing of Environment*, 305:114070, 2024.

[16] W. Y. Song, C. Wang, T. F. Dong, Z. H. Wang, C. X. Wang, X. D. Mu, and H. X. Zhang. Hierarchical extraction of cropland boundaries using sentinel-2 time-series data in fragmented agricultural landscapes. *Computers and Electronics in Agriculture*, 212, 2023.

[17] Jordan Graesser and Navin Ramankutty. Detection of cropland field parcels from landsat imagery. *Remote Sensing of Environment*, 201:165–180, 2017.

[18] F. Waldner, F. I. Diakogiannis, K. Batchelor, M. Ciccotosto-Camp, E. Cooper-Williams, C. Herrmann, G. Mata, and A. Toovey. Detect, consolidate, delineate: Scalable mapping of field boundaries using satellite images. *Remote Sensing*, 13(11), 2021.

[19] Mariana Belgiu and Lucian Drăgut. Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, 114:24–31, 2016.

[20] P. Defourny, S. Bontemps, N. Bellemans, C. Cara, G. Dedieu, E. Guzzonato, O. Hagolle, J. Inglada, L. Nicola, T. Rabaute, M. Savinaud, C. Udroiu, S. Valero, A. Bégué, J. F. Dejoux, A. El Harti, J. Ezzahar, N. Kussul, K. Labbassi, V. Lebourgeois, Z. Miao, T. Newby, A. Nyamugama, N. Salh, A. Shelestov, V. Simonneaux, P. S. Traore, S. S. Traore, and B. Koetz. Near real-time agriculture monitoring at national scale at parcel resolution: Performance assessment of the sen2-agri automated system in various cropping systems around the world. *Remote Sensing of Environment*, 221:551–568, 2019.

[21] S Parker Abercrombie and Mark A Friedl. Improving the consistency of multitemporal land cover maps using a hidden markov model. *IEEE Transactions on Geoscience and Remote Sensing*, 54(2):703–713, 2015.

[22] Txomin Hermosilla, Michael A. Wulder, Joanne C. White, Nicholas C. Coops, and Geordie W. Hobart. Disturbance-informed annual land cover classification maps of canada's forested ecosystems for a 29-year landsat time series. *Canadian Journal of Remote Sensing*, 44(1):67–87, 2018.

[23] Shi Qiu, Zhe Zhu, Rong Shang, and Christopher J. Crawford. Can landsat 7 preserve its science capability with a drifting orbit? *Science of Remote Sensing*, 4:100026, 2021.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

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

[32] 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.

[33] 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.

[34] 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.

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