

GTAPShape: A flexible approach to spatially disaggregating national land endowments for CGE modeling

Micah Cameron-Harp¹, Nelson Villoria¹, Jayson Beckman²

mcameronharp@ksu.edu, nvilloria@ksu.edu

¹Kansas State University, Department of Agricultural Economics

²USDA, Economic Research Service

Author's note: Cameron-Harp and Villoria acknowledge funding from the USDA Economic Research Service Cooperative agreement # 58-3000-2-0087 and from USDA National Institute of Food and Agriculture (NIFA) under Hatch project number S1072 and competitive grant number 2022-10683. The findings and conclusions in this publication are those of the author(s) and should not be construed to represent any official USDA or U.S. Government determination or policy.

Abstract

This paper presents `gtapshape`, an R package that allows the user to flexibly disaggregate the national endowments used in the computational general equilibrium (CGE) models based on the GTAP-AEZ framework. By allowing the user to specify the set of subnational boundaries in the form of a shapefile, `gtapshape` allows for a richer understanding of how within-country heterogeneity impacts the results of CGE models. `gtapshape`'s modular strategy also allows for fast updating of the database as new sources of data become available. `gtapshape` is fully written in R and hosted in GitHub as free and open software. This should facilitate its incorporation into specialized workflows

1 Background

The first release of the Global Trade Analysis Project (GTAP) land use and land cover (LULC) database significantly advanced our ability to investigate the land use implications of global market dynamics by allowing spatially explicit modeling of subnational markets (Hertel et al., 2009). Nicknamed GTAP-AEZ, the LULC database divided countries' land endowment into Agroecological Zones (Ramankutty et al., 2007) using spatially explicit data on agricultural production (Monfreda et al., 2009) and forest cover (Sohngen et al., 2009).

Versions 9 and 10 of the GTAP-AEZ database are thoroughly documented (Baldos, 2017; Baldos & Corong, 2020), providing a detailed account on how to produce the database directly using spatially explicit data on land use and land cover. Since its creation, the GTAP-AEZ model has been used in various contexts to analyze the impacts of biofuels (Hertel et al., 2010), climate policy (Golub et al., 2013) trade policy (Villoria et al., 2022), among others (Hertel et al., 2010).

However, dis-aggregating to AEZs may not align with the patterns of spatial heterogeneity for important environmental outcomes. Villoria et al. (2022) demonstrate that the AEZ boundaries group together portions Brazil's Cerrado and Amazonia, two distinct biomes with active land conversion to soybeans, thus making less effective the analysis of biome-specific policies. Moreover, the significant spatial heterogeneity in the consequences of land use changes for greenhouse gas emissions, biodiversity, and water quality do not conform to the AEZ boundaries (Myers et al., 2000; Paustian et al., 2017; Swan et al., 2020; Zimmerman et al., 2008)..

Instead of relying on AEZs, the newly created *gtapshape* R package allows users to create GTAP databases with user-specified sub-national delineations of land areas. This increased flexibility allows modelers to address a broader set of research questions pertaining to the land-environment-energy nexus. Our approach allows users to flexibly disaggregate land rental rates

using spatially explicit production and land cover data along with national level prices. This allows us to determine how to divide land rents among the subnational units chosen.

In Baldos (2017) and Baldos and Corong (2020), a database of land use and land cover data is created using Agro-ecological zones as the sub-national boundaries dividing up countries' land area. Our approach relies on similar spatially explicit data on agricultural production and land use but builds in several important advances. First, as stated previously, the *gtapshape* package allows the user to specify these sub-national boundaries. This means *gtapshape* is able to produce LULC databases, compatible with GTAP CGE models, for a wider variety of research contexts. For example, a user interested in the effects of global trade policy on species extinction could use a shapefile of biodiversity hotspots as the sub-national boundaries. Second, we allow the user to specify the year of national-level production data to use from the FAO when constructing the database. As such, *gtapshape* can produce databases representing a broader range of years (FAO data from 2011-2022 are included in the package) than the set of years released by the GTAP center. We have also added the option to use the most recent data on gridded production, CROPGRID (citation) which provides an important update to the work of Monfreda et al. (citation). *gtapshape*'s major effort has been to harmonize the large set of inputs needed to synthesize a GTAP LULC database so that the creation of new datasets is extremely fast. By simply selecting a new year from the FAOSTAT database, or a new set of subnational boundaries (such updated AEZs), or whether to use CROPGRID or Monfreda, the user can obtain a fully working GTAP database in HAR format in a matter of minutes¹. *gtapshape* is completely written

¹ Compiling an entire database takes ~90 seconds in a Lenovo ThinkPad laptop with an Intel Core Ultra 7 (12 Cores) processor and 32 GB of RAM, running Windows 11.

in R, based on publicly available datasets, and hosted in GitHub under a MIT License² and therefore can be “forked” and used by any interested party within the limits of the license.

In the next section, we provide instructions for installing the *gtapshape* package and an introduction to its core functionalities. Then, we provide additional detail on the geo-spatial data underpinning the package and how it was processed to facilitate the dis-aggregation routines. The pre-processing of the underlying spatial data is documented using fully reproducible R vignettes. The programming principle of the package is that any addition occurs at the beginning of the building process, and then a set of routines homogenizes all the inputs so as to minimize the need to modify programs due to ad-hoc naming conventions.

Finally, we provide three example uses of the package to demonstrate its capabilities of *gtapshape*. First, we compare how land rents differ between databases produced using two standard sets of sub-national boundaries built into the *gtapshape* package: the 18 AEZs and 14 World Wildlife Fund terrestrial ecoregions. Second, we produce updated datasets using more recent land use data representing agricultural production in 2020 as opposed to the data from 2000 used to produce versions 9 and 10 of the GTAP-LULC database (Baldos, 2017; Baldos & Corong, 2020). Finally, we create a time series of databases to illustrate how the package can be used to create the data necessary for dynamic CGE models.

² The MIT License on GitHub allows users to freely use, copy, modify, merge, publish, distribute, sublicense, and sell copies of the software, as long as the original copyright notice and license text are included in all copies or substantial portions of the software

2 Introduction to the *gtapshape* package

gtapshape is an R package designed for the flexible aggregation of land use and land cover data specifically tailored for use with the GTAP Agro-Ecological Zones (AEZ) model described in Hertel et al. (2008). This package provides tools to efficiently manipulate and aggregate spatial and country-level data on global land use and land cover to the GTAP regions and sectors. The following instructions for installing the package and an introduction to its capabilities are also available in the README document for the package at the package GitHub repository: <https://github.com/nvilloria/gtapshape>. Note, the user will need to have version a current version of R (R Core Team, 2017) to install the package and utilize its features.

2.1 Installation and usage

First, install the package from the repository on GitHub using the following code in R from the package README document:

```
## Install devtools if you don't already have it
install.packages("devtools")
## Install gtapshapeagg from GitHub
devtools::install_github("nvilloria/gtapshape", build_vignettes = TRUE)
```

Next, we provide a convenient function ‘`build.dbase.from.sf()`’ which processes all the data needed to create the land use and land cover headers and sets needed to split national land rents into subnational land rents. The ‘`build.dbase.from.sf()`’ function takes four inputs: An R simple features file (Edzer Pebesma, 2018; Edzer Pebesma & Roger Bivand, 2023) with the desired subnational boundaries (18 Agro-ecological zones and 14 World Wildlife Fund terrestrial ecoregions are included in the package), the year for which the FAO data on production, prices and area harvested should be processed (2011-2022 are included in the package), the source of the gridded data on global crop production (from either the Monfreda et al. (2008) or Tang et al. (2024) datasets provided in the package), and the name of the har file that will be generated with the

physical data on land use and land cover needed by the GTAP AEZ mode. Executing the following code from the package README document will create a GTAP-LULC database with the default settings: the included shapefile of the 18 Agro-ecological Zones, FAOSTAT data from 2017, the Monfreda et al. (2008) crop production rasters, and three output files named “gtaplulc.har”, “gtaplulc-sets.har”, and “gtaplulc-map.har.”

```
library(gtapshape)
## Build the land use and land cover headers and sets needed to split
## subnational land rents into 18 AEZS:
aez18 <- build.dbase.from.sf()
```



There are three outputs produced by the prior code block. The file titled “gtaplulc.har” contains the primary, processed data at the sub-national level. It contains the following seven headers:

1. QCR8: Crop production (MT) for the 8 gtap crop categories (pdr, wht, gro, v_f, osd, c_b, pfb, ocr).
2. VCR8: Value of crop production (1000 USD) for the 8 gtap crop categories.
3. HARV: Harvested area (ha) for the 8 gtap crop categories.
4. QLV3: Livestock production (heads) for the 3 gtap livestock sectors (ctl, rmk, wol).
5. VLV3: Livestock output value (1000 USD) for the 3 gtap livestock sectors.
6. LAND: Land cover area (ha).
7. RTMB: Timber land rents (USD Million).

The second output from the “build.dbase.from.sf()” command, the “gtaplulc-sets.har” file produced by the previous code block, is a har file defining the set elements of the database and has the following headers (sets):

1. REG: GTAP regions (defaults to the 160 regions in the GTAP database V11c (Aguiar et al., 2022)).
2. SUBN: Subnational boundaries (defaults to 18 AEZs)
3. CRP8: The eight GTAP crop categories (pdr, wht, gro, v_f, osd, c_b, pfb, and ocr)
4. CRP9: CRP8 + forest products (frs)
5. LCOV: Seven land cover categoris (Forest, SavnGrasslnd, Shrubland, Cropland, Pastureland, Builtupland, Otherland)

Finally, the last output from the “`build.dbase.from.sf()`” command is a GTAP Aggregation Template text file based on the GTAP Database V11c (Aguiar et al., 2022), that can be used to aggregate the GTAP database. We provide another free package, named *gtapshapeagg*³, which uses this aggregation template and the other two outputs to split land rents by the choice of subnational boundaries, the 18 Agro-Ecological Zones in this example. For a more in-depth explanation of how the gridded data are transformed into data specific to each country and subnational area, please refer to the package vignette titled “`build.gtapdatabase.for.year.and.from.sf`” (for an index of the vignettes included in the package, type “`browseVignettes('gtapshape')`” in the R console. In the next section, we describe the underlying spatial data used by the package and how it was processed.

³ See <https://github.com/nvilloria/gtapshapeagg> for installation instructions and worked examples on how to split the land rents in the standard GTAP database using the land use headers produced by *gtapshape*. See also the appendix.

3 Data

The data included in the *gtapshape* package is pre-processed to reduce the size of the package and the computational resources required to execute its functions. However, we provide a series of vignettes with the package detailing how the gridded spatial data are processed and combined to create the resulting database in case a user would like to use an alternative dataset or replicate the steps we used to create the package. The gridded, raster data used in each subsection below are downloaded when the ‘getrawdata’ function in the package is executed. Note, some of the pre-processing described here mirrors the processes described by Baldos (2017) in "Development of GTAP version 9 Land Use and Land Cover database for years 2004, 2007 and 2011." However, *gtapshape* provides several facilities for updating that original work or using alternative data sources. In figures 1 and 2, we provide an illustration of the additional flexibility allowed by the *gtapshape* package.

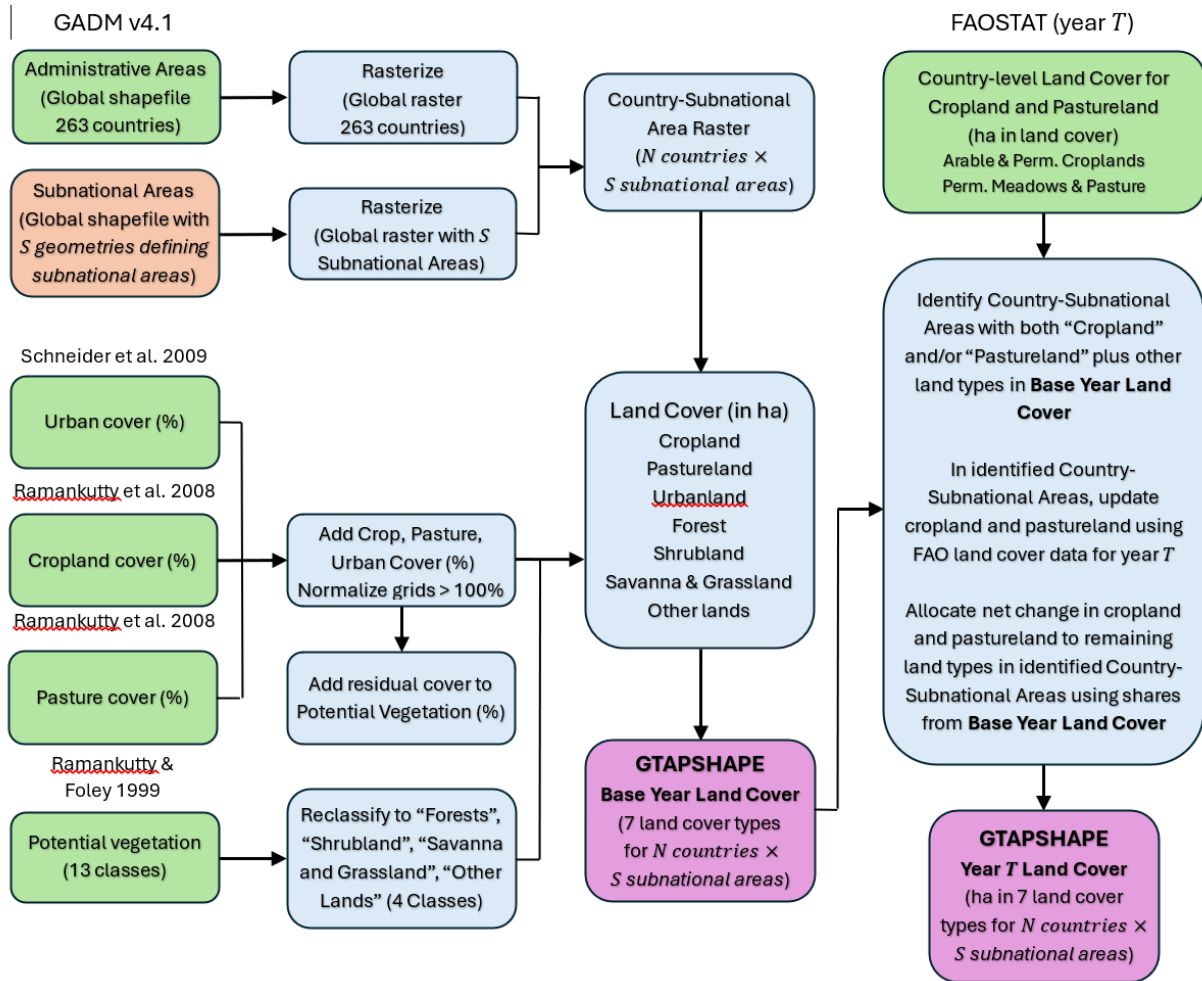


Figure 1: Creation of the land cover data by the *gtapshape* package. Note, the user specifies the shapefile of subnational areas and the year, T , of FAO data.

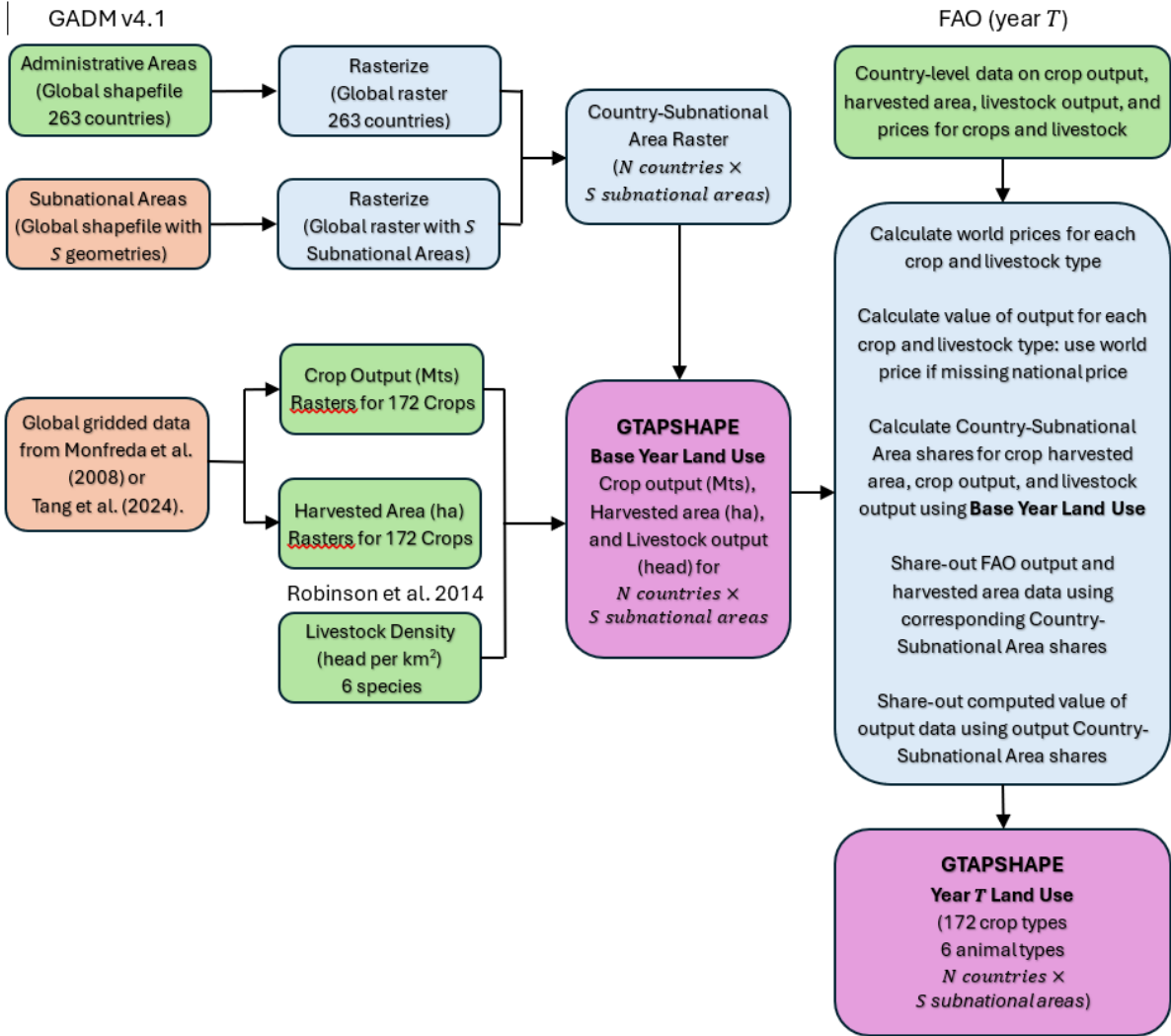


Figure 2: Creation of the land use data by the *gtapshape* package. Note, the user specifies the shapefile of subnational areas, the crop production data to use (either from Monfreda et al. (2008) or Tang et al. (2024)), and the year, T , of FAO data.

3.1 Downloading the raw data

The data inputs in figures 1 and 2 are available for download in a zipped file hosted in the GTAP Center server at Purdue University. These data can be downloaded using the function ``getrawdata()'``, as described below:

```
getrawdata()
```

Building the database from GIS layers requires downloading a very large compressed file (~3.5 GB zipped, 16.6 GB unzipped). This may take in excess of an hour, and it is probably impractical with a slow internet connection. The file has the underlying rasters and shape files used to split national land markets. These data are necessary only if there is a need to change a GIS layer. Otherwise, these data are not needed. Please refer to the package vignettes for documentation of how the underlying rasters and shapefiles were processed.

`gtapshape-raw-data.zip` will be downloaded to the current working directory where R is open (i.e., `getwd()`). In order to run the vignettes discussed in the rest of this section, the user must unzip the data in a folder named `raw_data`. The commands in the vignettes will look for the data subdirectories within this folder.

3.2 *Global raster of country boundaries*

To determine the crop production or livestock production in a given country-subnational area combination, we use the gridded data on 263 country boundaries provided in version 4.1 of the Database of Global Administrative Areas (GADM) database (*Database of Global Administrative Areas*, 2022). Executing the following code will open the vignette documenting how the `gadm\_rast.tif` file included in the `inst/GADM` folder of the *gtapshape* package is created from the original GADM data downloaded with “getrawdata” function.

```
vignette('GADM.country.raster', package = 'gtapshape')
```



3.3 *Creating the shapefile of 18 Agro-Ecological Zones*

Executing the following code will open the vignette which describes how we created the 18 Agro-Ecological Zones (AEZ) shapefile included in this package. We follow the procedure used to create version 11 of the GTAP-AEZ Land Use and Land Cover database described in Baldos and Corong (2020). This procedure relies on data on the length of the growing period

(LGP) and the thermal climate available from version 4 of the GAEZ database (FAO & IIASA, 2025).

```
vignette('create.18.aez.shapefile', package = 'gtapshape')
```



3.4 *Land cover data circa the year 2000*

The following code will open the vignette describing how we generate data on the 7 land cover types present in the GTAP Land Use and Land Cover datasets used in the *gtapshape* package.

```
vignette('land.cover', package = 'gtapshape')
```



The land cover data are created using publicly available raster data depicting the distribution of cropland and pastureland, potential vegetation classes, and urban areas. The global, gridded data on the distribution of cropland and pastureland are described in Ramankutty et al. (2008), and depict the fraction of each grid cell's area categorized as cropland or pastureland at a 5 arc minute resolution. The potential vegetation data are from Ramankutty and Foley (1999) and depict the global coverage of 13 different potential vegetation classes. For our analyses, we collapse the original 13 classes into 4 categories: "Forests", "Shrubland", "Savanna + Grassland", and "Other Lands" as in Baldos (2017). The urban land cover data from Schneider et al. (2009) are packaged as a categorical raster indicating the raster cells which are urban at a 500 meter resolution.

3.5 *Download FAOSTAT data on production and land cover for user-specified year*

One of the advantages of the *gtapshape* package is that the user specifies the year of FAO data used to update the GTAP database. The vignette called in the following code block illustrates

how the *FAOSTAT* package (Michael et al., 2025) is used to download the raw data on agricultural production and harvested area for years 2010 through 2022.

```
vignette('faostat.data', package = 'gtapshape')
```



3.6 *Updating land cover using data downloaded from FAOSTAT*

This vignette explains how the gridded data on land cover for year 2000 (from the “land.cover” vignette) is used to estimate the shares of land covers other the FAOSTAT values for croplands and pastures. The procedures in this vignette update the land cover values for the user-specified year of FAOSTAT data.

```
vignette('Country_level_landcover_ts', package = 'gtapshape')
```



3.7 *Gridded data on crop and livestock production*

The two vignettes called in the following code block describe how we generate land use data in the GTAP Land Use and Land Cover datasets used in the *gtapshape* package. The data are created using publicly available raster data depicting global crop yields and harvested area for 172 crops and 4 livestock species. The first vignette includes code for processing the crop data representing production circa the year 2000 from Monfreda et al. (2008) and livestock production for the year 2005 from Robinson et al. (2014). The second vignette goes over the processing of the CROPGRIDS crop area data depicting production in the year 2020 from Tang et al. (2024).

```
#Crop production in 2000 and livestock production in 2005 vignette  
vignette('land.use', package = 'gtapshape')  
#Crop production in 2020 vignette  
vignette('CROPGRIDS.convert.to.dataframes', package = 'gtapshape')
```



3.8 *Creating national level rents from forestry activities*

The final vignette detailing pre-processing of data is called by the following code and details how the national timber land rents included in the package are created. The raw data used in this vignette contain the forest rental rates included in Lee et al. (2008).

```
vignette('preprocessing.of.forest.rents', package = 'gtapshape')
```



4 Using the *gtapshape* package

In this section, we provide three illustrative examples demonstrating the novel capabilities of *gtapshape*, its improvements on past datasets, and some potential applications. The code necessary for generating the results depicted in each example is provided within the text and is also included in the ReadMe document for the *gtapshape* package.

4.1 *Agro-Ecological Zones versus World Wildlife Fund Ecoregions*

In this first example, we demonstrate the core utility of the *gtapshape* package and its implications for analyses of the impacts of global land use changes. The primary advantage of *gtapshape* is its ability to create GTAP-LULC databases with user-specified sub-national divisions of land use and land cover data that are compatible with the GTAP-AEZ model. To illustrate this, we first construct a GTAP-LULC database using the 18 Agro-Ecological Zones as a baseline for comparison using the following code:

```
library(gtapshape)
## Build the land use and land cover headers and sets needed to split
## subnational land rents into 18 AEZs:
aez18 <- build.dbase.from.sf()
```



Next, we create a second GTAP-LULC database using the same year of FAOSTAT data and gridded crop production rasters, but using an alternative shapefile of sub-national boundaries depicting the 14 Terrestrial Ecoregions used by the World Wildlife Fund (Olson et al., 2001):

```
system.time(biome14 <- build.dbase.from.sf(subnat_bound_file="biomes14",  
                                           file="gtaplulc-biome14-2017.har"))
```



Note, all that is required to produce this new dataset is the new shapefile to be used. In the previous code block, we specify that the “biomes14” shapefile included with the package should be used when creating the resulting “gtaplulc-biomes14-2017.har” file. Note, as the default values for the year and crop_rasters arguments of the “build.dbase.from.sf” wrapper function are 2017 and the Monfreda et al. (2008) data, respectively, the two databases rely on the same underlying spatial data on crop production to spatially allocate identical FAOSTAT data. However, as we illustrate in Figure 3, there are significant differences in the resulting GTAP-LULC databases.

The two maps in Figure 3 portray the cropland area in 1,000s of hectares at the GTAP region by sub-national area level. The top map in Figure 3 contains the results from the first database produced using the 18 Agro-Ecological Zones, and the bottom map contains the results for the database produced using the 14 WWF Terrestrial Ecoregions. Notice, while a country’s total land endowment is identical between the databases, the choice of sub-national boundary determines the number of categories that land endowment is divided amongst and their distribution. As a result, any subsequent analysis of the impacts of land use changes will have markedly different implications. For example, consider the difference between the Agro-Ecological Zones and WWF Ecoregions in North America. The AEZs divide the eastern United States into three AEZs along a north-south gradient, while the WWF Terrestrial Ecoregions divide the same region into inland and coastal regions. If we are interested in examining the effects of changes in cropland area on greenhouse gas mitigation efforts, the AEZ borders may provide

additional and meaningful heterogeneity in the outcomes. For example, the potential greenhouse gas sequestration benefits from agricultural practices like no-till and covering cropping are highly correlated with latitude due to its dependence on soil temperatures (Paustian et al., 2017; Swan et al., 2020).

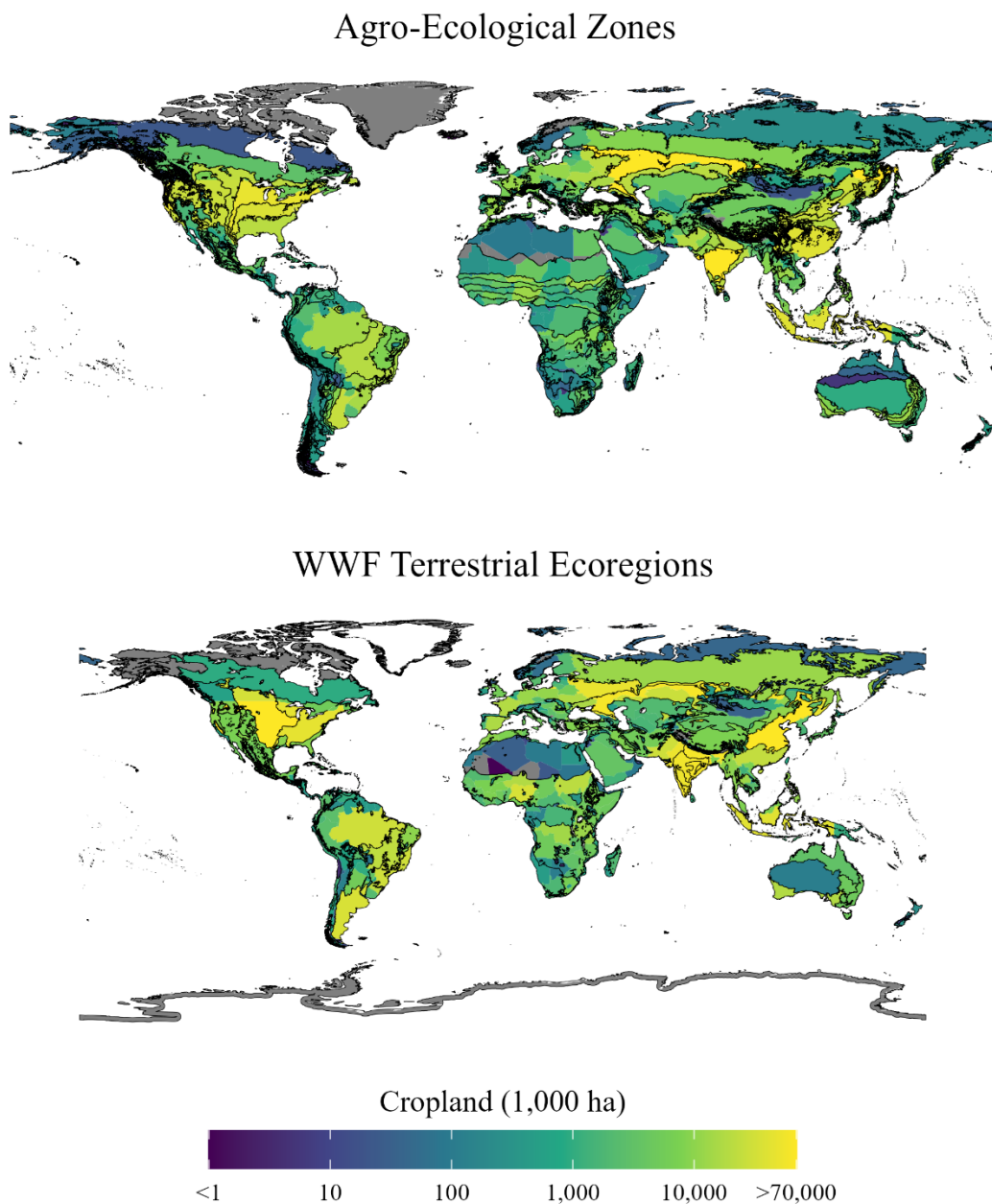


Figure 3: Distribution of cropland by choice of subnational boundaries used to spatially disaggregate national land cover values. The top figure depicts the results for the 18 Agro-

Ecological Zones and the bottom figure depicts the results when the 14 World Wildlife Fund Terrestrial Ecoregions are used to create a GTAP-LULC dataset. The datasets used to generate both figures are created using FAOSTAT data from 2017 and the Monfreda et al. (2008) gridded data on the global distribution of crop production.

4.2 *Monfreda et al. (2008) versus Tang et al. (2024) CROPGRIDS land use data*

Next, we demonstrate how the *gtapshape* package can create LULC datasets that incorporate changes in global crop production between 2000 and 2020. Versions 9 and 10 of the Land Use and Land Cover databases released by GTAP include the years 2004, 2007, 2011, and 2014 (Baldos, 2017; Baldos & Corong, 2020), but rely on gridded production data which reflect the global distribution of crop production in the year 2000 (Monfreda et al. 2008). We address this limitation in the *gtapshape* package by allowing the use of the CROPGRIDS gridded data on global crop production for the year 2020 when disaggregating national values (Tang et al., 2024). In doing so, we allow the user to create more accurate LULC databases which reflect recent changes in the distribution of crops and the resulting impacts on subnational land rents.

The following code creates two LULC databases that use identical subnational boundaries (18 AEZs) and FAOSTAT land use and land cover data from 2022. For the first database, we specify that the Monfreda et al. (2008) gridded crop production data should be used when allocating national data to each of the subnational areas:

```
system.time(aez18_cropgrid_2022 <- build.dbase.from.sf(subnat_bound_file="aez18",
  crop_rasters = "monfreda",
  year=2022,
  file="gtaplulc-cropgrid-2022.har"))
```

But, for the second database, we specify that the CROPGRIDS gridded data should be used instead (Tang et al., 2024):

```
system.time(aez18_cropgrid_2022 <- build.dbase.from.sf(subnat_bound_file="aez18",
  crop_rasters = "cropgrids",
  year=2022,
  file="gtaplulc-cropgrid-2022.har"))
```

We then plot the difference in the value of production across all 8 GTAP crop commodities between the two databases in Figure 4 to illustrate the consequences of using the more recent gridded data. Notice in Figure 4, using the more recent gridded data significantly alters the how the value of crop production is distributed among AEZs for many of the largest crop-producing countries. In China, India, the United States, and Brazil, for example, using the CROPGRIDS data causes large changes in how the value of crop production is allocated across AEZs.

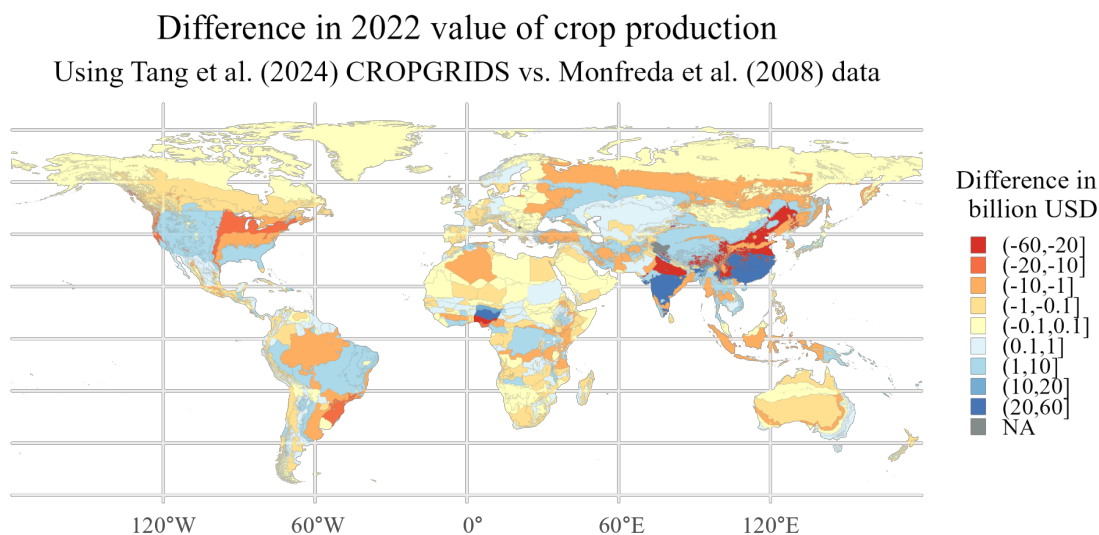


Figure 4: Difference in the 2022 total value of crop production between databases generated using the Monfreda et al. (2008) and CROPGRIDS (Tang et al., 2024) gridded data. Data are displayed as the difference in billions of US dollars and are calculated by subtracting the value from the Monfreda et al. (2008) database from the CROPGRIDS database.

4.3 Creating a time series of GTAP-LULC datasets

Our final example case demonstrates another capability of the *gtapshape* package to capture changes in land use and land cover over time. In contrast to the last example which focused on the ability to choose the underlying gridded crop production data (depicted in Figure 2), here we illustrate the significance of being able to specify the year of FAOSTAT data used to create a LULC database (the choice of year T in Figures 1 and 2). In addition, we provide an example of how the accompanying *gtapshapeagg* package is used to disaggregate national land rent values

by subnational area. First, we use a loop to create a series of LULC databases for all years of FAOSTAT data included in the package, 2011 to 2022, using the 18 Agro-Ecological Zones for the subnational boundaries and the Monfreda et al. (2008) gridded crop production data:

```
for (y in 2011:2022) {  
  aez18_cropgrid <- build.dbase.from.sf(subnat_bound_file="aez18",  
                                       crop_rasters = "monfreda",  
                                       year=y,  
                                       file=paste0("gtaplulc-monfreda-aez18-", y, ".har"))  
}
```

With the series of LULC databases we just created, we could examine trends in land use and land cover at the AEZ level. Instead, we will demonstrate how the 2011-2022 LULC databases from *gtapshape* can be employed to disaggregate land rents with the accompanying *gtapshapeagg* package and depict changes in land rents over time. In the next code block, we install the *gtapshapeagg* package which splits the national level rents from an official release of the GTAP database and then run the “`setup_gtapshapeagg()`” command to check it installed correctly. Note, the *gtapshapeagg* package compiles and executes TABLO-generated programs using GEMPACK (Horridge, J. et al., 2018). As such, users must have GEMPACK installed and hold a current source-code license to use the *gtapshapeagg*.. The final command in the following code block requires a password as it unzips an official release of the version 11c GTAP database Model V6⁴ (Aguilar et al., 2022):

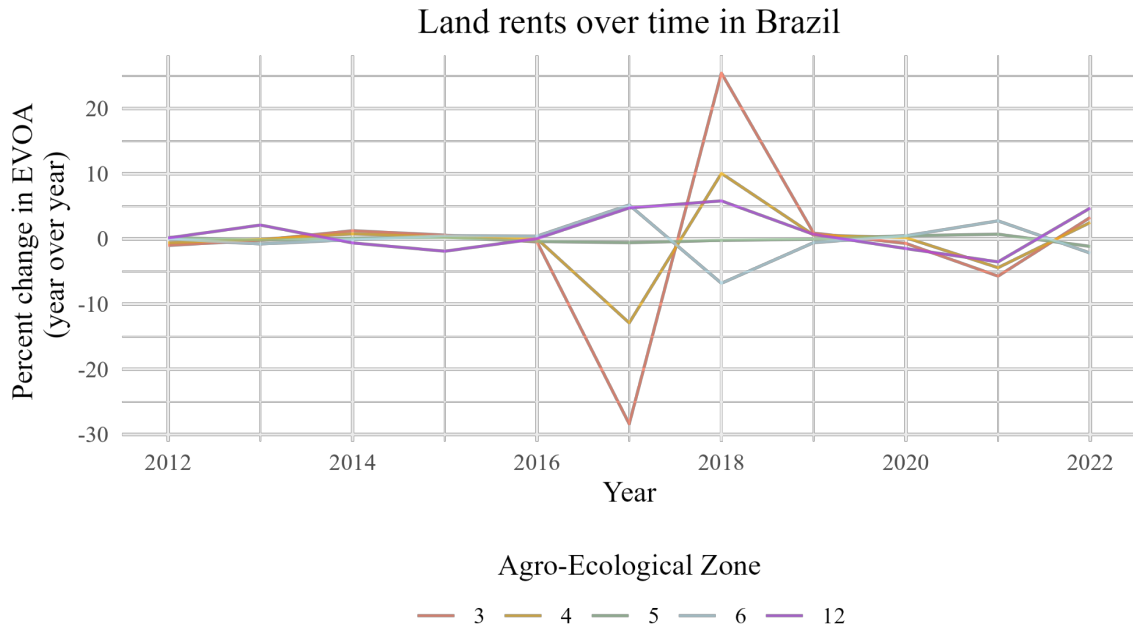
⁴ We include the password-protected data to ensure reproducibility, which among other things, facilitates the review process, but all is needed is to put in the current folder a properly licensed GTAP database. While in principle any GTAP database should work, we have tested the land rent splitting code only on the GTAP database V11c, GTAP Model V6. Any change in set names, for example, will make these routines to halt.

```
## Install and load gtapshapeagg package:
devtools::install_github("nvilloria/gtapshapeagg", build_vignettes = TRUE)
library(gtapshapeagg)

## Check that necessary gempack programs are in the working directory:
setup_gtapshapeagg()

## Unzip the GTAP database, password required:
unzip_gtapdata()
```

Once properly installed, we use another loop to split the national land rents into the subnational areas (18 Agro-Ecological Zones) and then create a final aggregated GTAP-AEZ database. These code to perform these two steps uses the “splitlr” and “createdat” commands from the *gtapshapeagg* package and can be found in Appendix A. In Figure 5, we display the resulting land rents (EVOA) for the five largest Agro-Ecological Zones in Brazil to illustrate the resulting dynamics. Notice, in Figure 5, that land rents change over time even though we used the same Monfreda et al. (2008) gridded crop production data to produce all of the LULC databases. By allowing the user to specify the year of FAOSTAT land use and land cover data, *gtapshape* can depict the within-country shifts in land rents caused by annual changes in land cover, prices, and production quantities.



5 Conclusion

The outcomes of land use changes are often heterogeneous in space, varying significantly within a single country. As such, CGE analyses with a single land endowment at the national scale may provide an inaccurate depiction of expected land use changes and the resulting effects. The *gtapshape* package builds on the approach pioneered by the GTAP-AEZ model by allowing for even greater flexibility in how subnational land areas are defined. The user of the *gtapshape* package decides the subnational boundaries that comprise a country's land endowment. As a result, the user can model a greater variety of land use contexts wherein land use constraints or outcomes do not align with the Agro-Ecological Zones. Furthermore, by allowing the user to choose the underlying gridded production data and year of land use and land cover data, the *gtapshape* package gives modelers the ability to examine land use changes in a dynamic context.

6 References

- Aguiar, A., Chepeliev, M., Corong, E., & Van Der Mensbrugghe, D. (2022). The Global Trade Analysis Project (GTAP) Data Base: Version 11. *Journal of Global Economic Analysis*, 7(2), 1–37. <https://doi.org/10.21642/JGEA.070201AF>
- Baldos, U. L. (2017). *Development of GTAP 9 Land Use and Land Cover Data Base for years 2004, 2007 and 2011* (GTAP Research Memoranda Series) [GTAP Research Memoranda Series]. GTAP Research Memoranda. <https://doi.org/10.21642/GTAP.RM30>
- Baldos, U. L., & Corong, E. (2020). *Development of GTAP 10 Land Use and Land Cover Data Base for years 2004, 2007, 2011 and 2014* (GTAP Research Memoranda Series) [GTAP Research Memoranda Series]. GTAP Research Memoranda. <https://doi.org/10.21642/GTAP.RM36>
- C. J. Michael, Markus Gesmann, Filippo Gheri, Paul Rougieux, & Sebastian Campbell. (2025). *FAOSTAT* (Version 2.4.0) [Dataset]. <https://gitlab.com/paulrougieux/faostatpackage>
- Database of Global Administrative Areas* (Version 4.1). (2022). [Dataset]. <https://gadm.org/data.html>.
- Edzer Pebesma. (2018). Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*, 10(1), 439–446.
- Edzer Pebesma & Roger Bivand. (2023). *Spatial Data Science: With Applications in R* (1st ed.). Chapman and Hall/CRC. <https://doi.org/10.1201/9780429459016>
- FAO & IIASA. (n.d.). *Global Agro Ecological Zones version 4 (GAEZ v4)* [Dataset].
- Golub, A. A., Henderson, B. B., Hertel, T. W., Gerber, P. J., Rose, S. K., & Sohngen, B. (2013). Global climate policy impacts on livestock, land use, livelihoods, and food security.

- Proceedings of the National Academy of Sciences*, 110(52), 20894–20899.
<https://doi.org/10.1073/pnas.1108772109>
- Hertel, T. W., Golub, A. A., Jones, A. D., O'Hare, M., Plevin, R. J., & Kammen, D. M. (2010). Effects of US Maize Ethanol on Global Land Use and Greenhouse Gas Emissions: Estimating Market-mediated Responses. *BioScience*, 60(3), 223–231.
<https://doi.org/10.1525/bio.2010.60.3.8>
- Hertel, T. W., Lee, H.-L., Rose, S., & Sohngen, B. (2008). *Modeling Land-use Related Greenhouse Gas Sources and Sinks and their Mitigation Potential* (GTAP Working Paper) [GTAP Working Paper]. GTAP Working Paper. <https://doi.org/10.21642/GTAP.WP44>
- Hertel, T. W., Rose, S., & Tol, R. S. (2009). Land Use in Computable General Equilibrium Models: An Overview. In *Economic Analysis of Land Use in Global Climate Change Policy* (pp. 3–35). Routledge.
- Horridge, J., Jerie, M., Mustakinov, D., & Schiffmann, F. (2018). *GEMPACK manual, GEMPACK Software*.
- Lee, H.-L., Hertel, T. W., Rose, S., & Avetisyan, M. (2008). *An Integrated Global Land Use Data Base for CGE Analysis of Climate Policy Options* [GTAP Working Papers]. Center for Global Trade Analysis, Department of Agricultural Economics, Purdue University.
<https://EconPapers.repec.org/RePEc:gta:workpp:2603>
- Monfreda, C., Ramankutty, N., & Foley, J. A. (2008). Farming the planet: 2. Geographic distribution of crop areas, yields, physiological types, and net primary production in the year 2000. *Global Biogeochemical Cycles*, 22(1), 2007GB002947.
<https://doi.org/10.1029/2007GB002947>

- Monfreda, C., Ramankutty, Navin, & Hertel, Thomas W. (2009). Global Agricultural Land Use Data for Climate Change Analysis. In *Economic Analysis of Land Use in Global Climate Change Policy* (pp. 33–49). Routledge.
- Myers, N., Mittermeier, R. A., Mittermeier, C. G., Da Fonseca, G. A. B., & Kent, J. (2000). Biodiversity hotspots for conservation priorities. *Nature*, 403(6772), 853–858. <https://doi.org/10.1038/35002501>
- 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(11), 933. [https://doi.org/10.1641/0006-3568\(2001\)051\[0933:TEOTWA\]2.0.CO;2](https://doi.org/10.1641/0006-3568(2001)051[0933:TEOTWA]2.0.CO;2)
- Paustian, K., Easter, M., Brown, K., Chambers, A., Eve, M., Huber, A., Marx, E., Layer, M., Stermer, M., Sutton, B., Swan, A., Toureene, C., Verlayudhan, S., & Williams, S. (2017). *Field- and farm-scale assessment of soil greenhouse gas mitigation using COMET-Farm* (pp. 341–359). American Society of Agronomy, Crop Science Society of America, Soil Science Society of America. <https://doi.org/10.2134/agronmonogr59.c16>
- R Core Team. (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.
- Ramankutty, N., Evan, A. T., Monfreda, C., & Foley, J. A. (2008). Farming the planet: 1. Geographic distribution of global agricultural lands in the year 2000. *Global Biogeochemical Cycles*, 22(1), 2007GB002952. <https://doi.org/10.1029/2007GB002952>

- Ramankutty, N., & Foley, J. A. (1999). Estimating historical changes in global land cover: Croplands from 1700 to 1992. *Global Biogeochemical Cycles*, 13(4), 997–1027. <https://doi.org/10.1029/1999GB900046>
- Ramankutty, N., Hertel, T., Lee, H.-L., & Rose, S. K. (2007). Global agricultural land-use data for integrated assessment modeling. In *Human-Induced Climate Change: An Interdisciplinary Assessment* (pp. 252–265). Cambridge University Press. <https://doi.org/10.1017/CBO9780511619472.025>
- Robinson, T. P., Wint, G. R. W., Conchedda, G., Van Boeckel, T. P., Ercoli, V., Palamara, E., Cinardi, G., D'Aiotti, L., Hay, S. I., & Gilbert, M. (2014). Mapping the Global Distribution of Livestock. *PLoS ONE*, 9(5), e96084. <https://doi.org/10.1371/journal.pone.0096084>
- Schneider, A., Friedl, M. A., & Potere, D. (2009). A new map of global urban extent from MODIS satellite data. *Environmental Research Letters*, 4(4), 044003. <https://doi.org/10.1088/1748-9326/4/4/044003>
- Sohngen, B., Tennity, C., Hnytka, M., & Meeusen, K. (2009). Global Forestry Data for the Economic Modeling of Land Use. In *Economic Analysis of Land Use in Global Climate Change Policy* (pp. 49–72). Routledge. <http://ideas.repec.org/p/gta/workpp/2602.html>
- Swan, A., Easter, M., Chambers, A., Brown, K., Williams, S., Creque, J., Wick, J., & Paustian, K. (2020). *COMET-Planner Carbon and greenhouse gas evaluation for NRCS conservation practice planning*. United States Department of Agriculture Natural Resources Conservation Service and Colorado State University.
- Tang, F. H. M., Nguyen, T. H., Conchedda, G., Casse, L., Tubiello, F. N., & Maggi, F. (2024). CROPGRIDS: A global geo-referenced dataset of 173 crops. *Scientific Data*, 11(1), 413. <https://doi.org/10.1038/s41597-024-03247-7>

Villoria, N., Garrett, R., Gollnow, F., & Carlson, K. (2022). Leakage does not fully offset soy supply-chain efforts to reduce deforestation in Brazil. *Nature Communications*, 13(1), 5476. <https://doi.org/10.1038/s41467-022-33213-z>

Zimmerman, J. B., Mihelcic, J. R., & Smith, A. J. (2008). Global Stressors on Water Quality and Quantity. *Environmental Science & Technology*, 42(12), 4247–4254. <https://doi.org/10.1021/es0871457>

Appendix A

The following code uses the *gtapshapeagg* package to split the land rents contained in the version 11 GTAP database (Aguilar et al., 2022) using the LULC databases created for 2011-2022 in section 4.3 and then aggregates the resulting data using the mapping file included in the *gtapshape* package.

```
##Split land rents in GTAP database using each of the LULC datasets from 2011-2022:
for (y in 2011:2022) {
  ## The first command, splitlr, allocates the national land rent data
  ## contained in the GTAP database released by the GTAP center using a
  ## LULC database produced by gtapshape.
  splitlr(
    ## This argument specifies the data file with LULC by subnational boundaries,
    ## regions and products, created by gtapshape for a specific year:
    landdat = paste0("gtaplulc-monfreda-aez18-", y, ".har"),
    ## Argument specifying are the LULC sets from gtapshape:
    landsets = paste0("gtaplulc-monfreda-aez18-", y, "-sets.har"),
    ## Argument indicating the file containing the standard GTAP database sets:
    stdgtapsets = system.file("GTAPv11c", "gsdgset11cMV6.har", package = "gtapshapeagg"),
    ## Argument indicating the standard GTAP database:
    stdgtapdata = system.file("GTAPv11c", "gsdgdat11cMV6.har", package = "gtapshapeagg"),
    ## Argument defining the directory for the outputs, landgtapsets and landgtapdat:
    dir= paste0("AEZ18v11c_", y)
  )
  ## The second command, createdat, uses a mapping file to aggregate specific
  ## GTAP commodity groups and regions, as well as populate the parameters
  ## specific to subnational-areas used by the GTAP-AEZ model.
  g <- createdat(
    ## Argument indicating the mapping file created by gtapshape:
    mapfile = paste0("gtaplulc-monfreda-aez18-", y, "-map.txt"),
    ## Argument indicating the file containing the sets created by the
    ## splitlr command above:
    setfile = paste0("./AEZ18v11c_", y, "/landgtapsets.har"),
    ## Argument indicating the disaggregated GTAP database created by the
    ## splitlr command above:
    datfile = paste0("./AEZ18v11c_", y, "/landgtapdat.har"),
    ## File containing standard GTAP parameters:
    stdprm = system.file("GTAPv11c", "gsdgpar11cMV6.har", package = "gtapshapeagg"),
    ## Name of the directory where the final database will be saved:
    dir = paste0("My_18AEZagg_", y)
  )
}
```