XPolaris: Retrieving Soil Data from POLARIS

L. Moro Rosso[†], A. de Borja Reis, A. Correndo[†] & I. Ciampitti[†]

Department of Agronomy, Kansas State University, Manhattan, KS-66506
[†]Contact: lhmrosso@ksu.edu; correndo@ksu.edu; ciampitti@ksu.edu.

Compiled on 2021-06-17

This tutorial intends to assist potential users interested in soil raster data of the United States for research purposes. All data is sourced from the POLARIS database, a 30-meter probabilistic soil series map of the contiguous United States (Chaney et al., 2016, 2019). POLARIS represents an optimization of the Soil Survey Geographic (SSURGO) database, circumventing issues of spatial disaggregation, harmonization, and spatial gaps. It was constructed using a machine learning algorithm, the Disaggregation and Harmonisation of Soil Map Units Through Resampled Classification Trees (DSMART-HPC) (Odgers et al., 2014). Moreover, the available `.tif` format makes POLARIS an easy option to obtain soil data for sites located in the US.

Table of contents

1 Introduction **2**
1.1 `exkansas` locations 2

2 Location areas **3**
2.1 `xplot` function 3

3 Downloading images **4**
3.1 `ximages` function 4

4 Extracting soil data **5**
4.1 `xsoil` function 5

References **5**

1 Introduction

The `XPolaris` package (version 1.0.1) was released in `CRAN` on June 17 of 2021. In addition, a development version of `XPolaris` is hosted in `GitHub`, containing basic usage instructions in a `README.md` file. The following R code must be executed in the order it is presented.

```
# Installation may ask for updating some packages
install.packages('XPolaris')
library(XPolaris)
```

Users can retrieve soil data for the sample coordinates provided in this tutorial, or use them as a reference to build their own set of geographical locations. Besides geographical locations, it is possible to define what soil parameters and depths are needed. There is no limit on the amount of data requested by the user. However, the image download depends on internet connection and large objects can surpass the memory limit of the R environment and/or machine. `XPolaris` is sourced by the POLARIS database and the three package functions were tested across different operating systems and R versions. Functions have the following purposes:

1. `xplot`: generates a map displaying the rater images from which locations will be retrieved;
2. `ximages`: downloads the images from the POLARIS database to the user's local machine;
3. `xsoil`: extracts the soil data from raster images and creates a `data.frame` object.

More information on functions usage can be found in the `XPolaris_documentation.pdf` file, and also using the help within R (e.g., `?xplot`). Although users do not need to edit function codes, they are available and commented on `GitHub`. In this approach, `XPolaris` serves beginner R programmers looking for US soil data, but also experienced collaborators willing to improve future versions of the package. If there are any problems with `XPolaris` installation and use, please report here.

1.1 exkansas locations

Location information must be stored in a `data.frame` object with three columns, named: `ID`, `lat`, and `long`. Latitude (`lat`) and longitude (`long`) coordinates must be expressed in decimal degrees, and location IDs must be unique. The locations example from Kansas were saved in the `exkansas` object, coming along with `XPolaris`. However, the example data was also saved as the `exkansas.csv` file and uploaded to the repository. These locations might help users testing the code, facilitating the preparation of other location `data.frame`. The locations "Scandia" and "Belleville" are close in space, to illustrate that the code only downloads a minimum number of images to extract the soil information (proximal locations may fall in the same raster files).

The following code assigns `exkansas` to the `df.loc` object.

```
df.loc <- exkansas
```

This is how `df.loc` should look:

```
##           ID     lat     long
## 1    Scandia 39.8291 -97.8458
## 2 Belleville 39.8158 -97.6720
## 3     Ottawa 38.5398 -95.2446
```

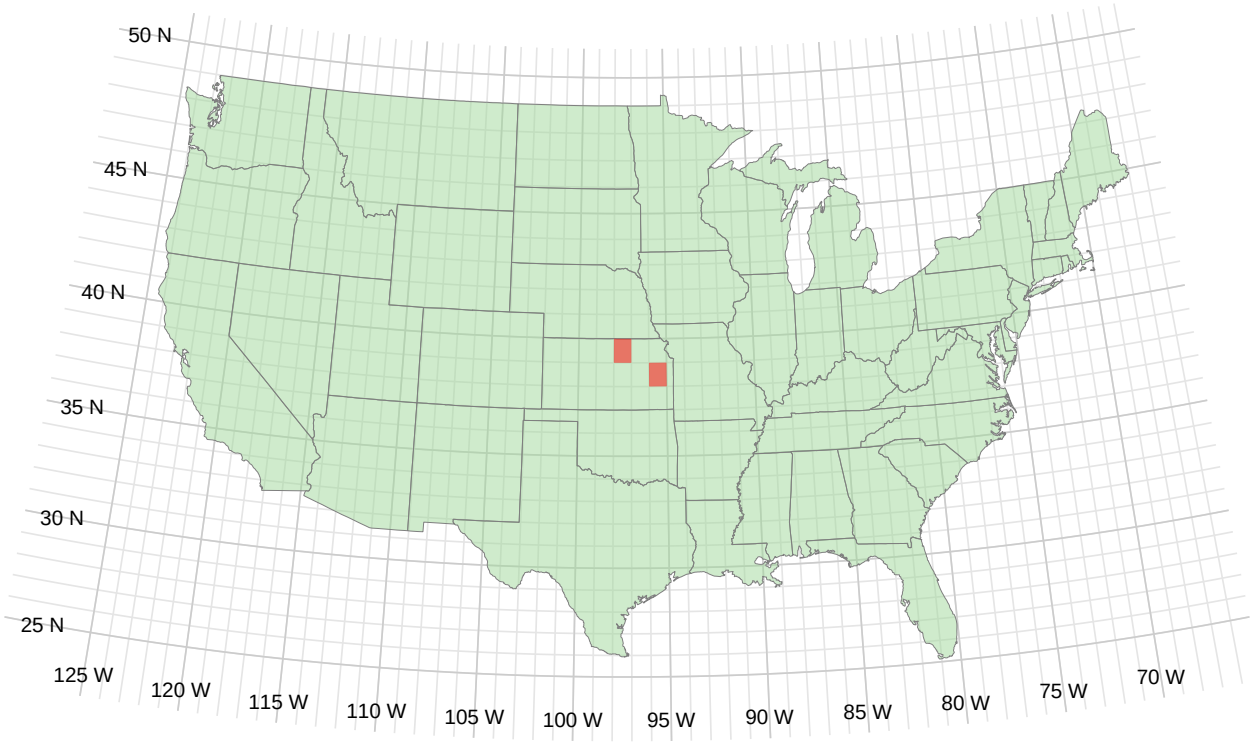
2 Location areas

The POLARIS database is composed of `.tif` images covering an area of one-by-one lat-long degrees. Images are organized by folders with the following hierarchy: variable (soil parameters), statistics (e.g., mean), and layer (soil depth). The layer folder contains several `.tif` images named, for example, as `lat2324_lon-81-80.tif`. In this case, that image covers an area from the parallel 23 to 24, and from the meridian -81 to -80. The `xplot` function was designed to identify the minimum number of "squares" for covering all locations and displaying the information in a plot. The function creates a new directory, called `POLARISOut`, for storing the plot as a `.jpg` image.

2.1 `xplot` function

The `locations` argument must be a `data.frame`, similar to the `df.loc` object. The `localPath` argument uses a temporary directory as default, but users can use any other path, such as the current working directory (easily accessed with the `getwd()` function).

```
xplot(locations = df.loc,
      localPath = getwd())
```



If a location falls in the square border (round lat-long coordinates) the code identifies a minimum number of images to be downloaded. This is done by choosing an image already requested for other coordinates (if applicable). Also, once an image is downloaded into the user's machine, the same image will not be downloaded again. Details on functions rationale are presented in `GitHub`.

3 Downloading images

Each raster image represents a given summary statistics, soil variable, and layer depth. Table 1 presents a list of all the soil variables available in the database (with respective codes). The five statistics options are as follows: `mean`, `mode`, median (`p50`), five (`p5`), and 95 (`p95`) percentiles. The six depths of soil layers are as follows: `0_5`, `5_15`, `15_30`, `30_60`, `60_100`, and `100_200`. The function `ximages` will create the image path, and download the `.tif` files in the `POLARISOut` directory.

Table 1: Variable codes from the POLARIS database with respective output units.

Variable	Description	Output units
`ph`	Soil pH in water	–
`om`	Soil organic matter	%
`clay`	Clay	%
`sand`	Sand	%
`silt`	Silt	%
`bd`	Bulk density	g cm^{-3}
`hb`	Bubbling pressure (Brooks-Corey)	kPa
`n`	Measure of the pore size distribution (van Genuchten)	–
`alpha`	Scale param. inversely prop. to mean pore diameter	kPa^{-1}
`ksat`	Saturated hydraulic conductivity	cm hr^{-1}
`lambda`	Pore size distribution index (Brooks-Corey)	–
`theta_r`	Residual soil water content	m^3 m^{-3}
`theta_s`	Saturated soil water content	m^3 m^{-3}

3.1 `ximages` function

This function repeats some of the routines from `xplot` (e.g., identifying the squares), which makes the functions independent. It also has `locations` as an argument (`data.frame` with location coordinates). Vectors with character elements must be used for `variables`, `statistics`, and `layerdepths`. The user must choose parameters of interest and inform them within vectors for each argument. It is important the codes are consistent with Table 1. The `localPath` argument also uses a temporary directory as default, but here we used `getwd()` for all functions. The function output is a `tibble` with local image paths and location coordinates for retrieving soil data.

```
df.out <- ximages(locations = df.loc,
                  variables = c('ph','om','clay','silt','sand'),
                  statistics = c('mean'), # Only the mean will be downloaded
                  layersdepths = c('0_5','5_15','15_30'), # Depth layers (cm)
                  localPath = getwd())
```

Because POLARIS is a "probabilistic" soil series map, the database is not limited to a point estimate of a given variable. For each location, POLARIS estimates a distribution of values. This distribution is then summarized in the `statistics` options (summary statistics). If the `variables`, `statistics`, or `layersdepths` codes are not correct, the function will not retrieve the images.

4 Extracting soil data

This section imports the raster (`.tif`) images into the environment and extracts soil information for the coordinates in `df.loc`. It will also arrange the data and export a `.csv` file called `xsoil_data` (`POLARISOut` directory). The function removes the `log` scale from the original units (e.g., `om` is exported in %). Users might need to execute further unit conversions based on their needs.

4.1 `xsoil` function

The final function of this coding tutorial was designed to extract soil information from raster data, organize, and export a table (`.csv` file) into the `POLARISOut` directory. The main argument is called `ximages_output`, which was assigned to the `df.out` object in the previous code section.

```
xsoil(ximages_output = df.out,
      localPath = getwd())
```

```
## # A tibble: 9 x 10
##   statistics layersdepths ID           lat  long    ph    om  clay  silt  sand
##   <fct>      <fct>        <chr>      <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 mean       0_5          Scandia     39.8 -97.8  5.61  2.38  24.6  66.4  7.11
## 2 mean       0_5          Belleville  39.8 -97.7  5.61  2.38  24.6  66.4  7.11
## 3 mean       5_15         Scandia     39.8 -97.8  5.69  2.18  26.7  64.8  6.58
## 4 mean       5_15         Belleville  39.8 -97.7  5.69  2.18  26.7  64.8  6.58
## 5 mean       15_30        Scandia     39.8 -97.8  6.01  1.73  33.7  58.6  5.07
## 6 mean       15_30        Belleville  39.8 -97.7  6.01  1.73  33.7  58.6  5.07
## 7 mean       0_5          Ottawa      38.5 -95.2  6.16  1.92  22.4  65.6  8.84
## 8 mean       5_15         Ottawa      38.5 -95.2  6.17  1.81  24.9  63.0  8.68
## 9 mean       15_30        Ottawa      38.5 -95.2  6.27  1.46  34.4  53.3  8.01
```

References

Chaney, N. W., Minasny, B., Herman, J. D., Nauman, T. W., Brungard, C. W., Morgan, C. L. S., McBratney, A. B., Wood, E. F., & Yimam, Y. (2019). POLARIS soil properties: 30-m probabilistic maps of soil properties over the contiguous united states. *Water Resources Research*, *55*(4), 2916–2938. https://doi.org/https://doi.org/10.1029/2018WR022797

Chaney, N. W., Wood, E. F., McBratney, A. B., Hempel, J. W., Nauman, T. W., Brungard, C. W., & Odgers, N. P. (2016). POLARIS: A 30-meter probabilistic soil series map of the contiguous united states. *Geoderma*, *274*, 54–67. https://doi.org/https://doi.org/10.1016/j.geoderma.2016.03.025

Odgers, N. P., Sun, W., McBratney, A. B., Minasny, B., & Clifford, D. (2014). Disaggregating and harmonising soil map units through resampled classification trees. *Geoderma*, *214-215*, 91–100. https://doi.org/https://doi.org/10.1016/j.geoderma.2013.09.024