Package 'popRF'

July 26, 2021

Type Package

Title Random Forest-Informed Population Disaggregation

Version 1.0.0

Maintainer Maksym Bondarenko <mb4@soton.ac.uk>

Description Disaggregating census-based areal population counts to finer gridded population surfaces using Random Forest algorithm to determine the target area weights (see _Stevens, et al._ (2015) <doi:10.1371/journal.pone.0107042>).

URL https://github.com/wpgp/popRF

BugReports https://github.com/wpgp/popRF/issues

Imports doParallel, parallel, raster, methods, stats, foreach, gdalUtils, randomForest, quantregForest, plyr

Depends R (>= 3.2.0)

License GPL-3

Encoding UTF-8

RoxygenNote 7.1.1

NeedsCompilation no

Author Maksym Bondarenko [aut, cre, cph] (<https://orcid.org/0000-0003-4958-6551>), Jeremiah J Nieves [aut] (<https://orcid.org/0000-0002-7423-1341>), Forrest R. Stevens [aut], Andrea E. Gaughan [aut], Chris Jochem [ctb] (<https://orcid.org/0000-0003-2192-5988>), David Kerr [ctb], Alessandro Sorichetta [ctb] (<https://orcid.org/0000-0002-3576-5826>)

Repository CRAN

Date/Publication 2021-07-26 07:10:02 UTC

R topics documented:

popRF *Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-Sensed and Ancillary Data.*

Description

Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-Sensed and Ancillary Data.

Usage

```
popRF(pop, cov, mastergrid, watermask, px_area, output_dir, cores=0,
quant=FALSE, set_seed=2010, fset=NULL, fset_incl=FALSE,
fset_cutoff=20, fix_cov=FALSE, check_result=TRUE, verbose=TRUE,
log=FALSE, ...)
```

Arguments

pop	Character vector containing the name of the file from which the unique area ID and corresponding population values are to be read from. The file should contain two columns comma-separated with the value of administrative ID and population without columns names. If it does not contain an absolute path, the file name is relative to the current working directory.
cov	A nested list of named list(s), i.e. where each element of the first list is a named list object with atomic elements. The name of each named list corresponds to the 3-letter ISO code of a specified country. The elements within each named list define the specified input covariates to be used in the random forest model, i.e. the name of the covariates and the corresponding, if applicable and local, path to them. If the path is not a full path, it is assumed to be relative to the current working directory. Example for Nepal (NPL):

```
list(
    "NPL"=list(
                "covariate1" = "covariate1.tif",
                "covariate2" = "covariate2.tif"
               )
   )

## $NPL
## $NPL$covariate1
## [1] "covariate1.tif"
##
## $NPL$covariate2
## [1] "covariate2.tif"
```

mastergrid — A named list where each element of the list defines the path to the input mastergrid(s), i.e. the template gridded raster(s) that contains the unique area IDs as their value. The name(s) corresponds to the 3-letter ISO code(s) of a specified country(ies). Each corresponding element defines the path to the mastergrid(s). If the path is local and not a full path, it is assumed to be relative to the current working directory. Example:

```
list(
    "NPL" = "npl_mastergrid.tif"
   )
```

watermask — A named list where each element of the list defines the path to the input country-specific watermask. The name corresponds to the 3-letter ISO code of a specified country. Each corresponding element defines the path to the watermask, i.e. the binary raster that delineates the presence of water (1) and non-water (0), that is used to mask out areas from modelling. If the path is local and not a full path, it is assumed to be relative to the current working directory. Example:

```
list(
    "NPL" = "npl_watermask.tif"
   )
```

px_area — A named list where each element of the list defines the path to the input raster(s) containing the pixel area. The name corresponds to the 3-letter ISO code of a specified country. Each corresponding element defines the path to the raster whose values indicate the area of each unprojected (WGS84) pixel. If the path is local and not a full path, it is assumed to be relative to the current working directory. Example:

```
list(
    "NPL" = "npl_px_area.tif"
   )
```

```
## $NPL
## [1] "npl_px_area.tif"
```

output_dir — Character vector containing the path to the directory for writing output files. Default is the temp directory.

cores — Integer vector containing an integer. Indicates the number of cores to use in parallel when executing the function. If set to 0 (`max_number_of_cores -1`) will be used based on as many processors as the hardware and RAM allow. Default is `cores = 0`.

quant — Logical vector indicating whether to produce the quantile regression forests (TRUE) to generate prediction intervals. Default is `quant = TRUE`.

set_seed — Integer, set the seed. Default is `set_seed = 2010`

fset — Named list containing character vector elements that give the path to the directory(ies) containing the random forest model objects (.RData) with which we are using as a "fixed set" in this modeling, i.e. are we parameterizing, in part or in full, this RF model run upon another country's(ies') RF model object. The list should have two named character vectors, "final" and "quant", with the character vectors corresponding to the directory paths of the corresponding folders that

hold the random forest model objects and the quantile regression random forest model objects, respectively. Numerous model objects can be in each folder "./final/" and "./quant/" representing numerous countries with the understanding that the model being run will incorporate all model objects in the folder, e.g. if a model object for Mexico and

`fset_incl` Logical vector indicating whether the RF model object will or will not be combined with another RF model run upon another country's(ies') RF model object. Default is `fset_incl` = FALSE

`fset_cutoff` Numeric vector containing an integer. This parameter is only used if `fset_incl` is TRUE. If the country has less than `fset_cutoff` admin units, then RF popfit will not be combined with the RF model run upon another country's(ies') RF model object. Default is `fset_cutoff` = 20.

`fix_cov` Logical vector indicating whether the raster extent of the covariates will be corrected if the extent does not match mastergrid. Default is `fix_cov` = FALSE.

`check_result` Logical vector indicating whether the results will be compared with input data. Default is `check_result` = TRUE.

`verbose` Logical vector indicating whether to print intermediate output from the function to the console, which might be helpful for model debugging. Default is `verbose` = TRUE.

`log` Logical vector indicating whether to print intermediate output from the function to the log.txt file. Default is `log` = FALSE.

`...` Additional arguments:
`binc`: Numeric. Increase number of blocks sugesting for processing raster file.
`boptimise`: Logical. Optimize total memory requires to processing raster file by reducing the memory need to 35%.
`bsoft`: Numeric. If raster can be processed on less then `cores` it will be foresed to use less number of `cores`.
`nodesize`: Minimum size of terminal nodes. Setting this number larger causes smaller trees to be grown (and thus take less time). See `randomForest` for more details. Default is `nodesize` = NULL and will be calculated as `length(y_data)/1000`.
`maxnodes`: Maximum number of terminal nodes trees in the forest can have. If not given, trees are grown to the maximum possible (subject to limits by nodesize). If set larger than maximum possible, a warning is issued. See `randomForest` for more details. Default is `maxnodes` = NULL.
`ntree`: Number of variables randomly sampled as candidates at each split. See `randomForest` for more details. Default is `ntree` = NULL and `ntree` will be used `popfit$ntree`
`mtry`: Number of trees to grow. This should not be set to too small a number, to ensure that every input row gets predicted at least a few times. See `randomForest` for more details. Default is `ntree` = NULL and `ntree` will be used `popfit$mtry`.
`proximity`: Logical vector indicating whether proximity measures among the rows should be computed. Default is `proximity` = TRUE. See `randomForest` for more details.
`const`: Character vector containing the name of the file from which the mask will be used to constraine population layer. The mask file should have value `0`

as a mask. If it does not contain an absolute path, the file name is relative to the current working directory.

Details

This function produces gridded population density estimates using a Random Forest model as described in *Stevens, et al. (2015)* doi: 10.1371/journal.pone.0107042. The unit-average log-transformed population density and covariate summary values for each census unit are then used to train a Random Forest model (doi: 10.1023/A:1010933404324) to predict log population density. Random Forest models are an ensemble, nonparametric modeling approach that grows a "forest" of individual classification or regression trees and improves upon bagging by using the best f a random selection of predictors at each node in each tree. The Random Forest is used to produced grid, i.e. pixel, level population density estimates that are used as unit-relative weights to dasymetrically redistribute the census based areal population counts. This function also allows for modelling based upon a regional parameterisation (doi: 10.1080/17538947.2014.965761) of other previously run models as well as the creation of models based upon multiple countries at once (doi: 10.1016/j.compenvurbsys.2019.01.006). This function assumes that all data is unprojected and is in the WGS84 coordinate system.

Value

Raster* object of gridded population.

Author(s)

Maksym Bondarenko mb4@soton.ac.uk, Jeremiah J. Nieves J.J.Nieves@liverpool.ac.uk, Forrest R. Stevens forrest.stevens@louisville.edu, Andrea E. Gaughan ae.gaughan@louisville.edu, David Kerr dk2n16@soton.ac.uk, Chris Jochem W.C.Jochem@soton.ac.uk and Alessandro Sorichetta as1v13@soton.ac.uk

References

- Stevens, F. R., Gaughan, A. E., Linard, C. & A. J. Tatem. 2015. Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-Sensed and Ancillary Data. PLoS ONE 10, e0107042 doi: 10.1371/journal.pone.0107042
- L. Breiman. 2001. Random Forests. Machine Learning, 45: 5-32. doi: 10.1023/A:1010933404324
- Gaughan, A. E., Stevens, F. R., Linard, C., Patel, N. N., & A. J. Tatem. 2015. Exploring Nationally and Regionally Defined Models for Large Area Population Mapping. International Journal of Digital Earth, 12(8): 989-1006. doi: 10.1080/17538947.2014.965761
- Sinha, P., Gaughan, A. E, Stevens, F. R., Nieves, J. J., Sorichetta, A., & A. J. Tatem. 2019. Assessing the Spatial Sensitivity of a Random Forest Model: Application in Gridded Population Modeling. Computers, Environment and Urban Systems, 75: 132-145. doi: 10.1016/j.compenvurbsys.2019.01.006

Examples

```
## Not run:

library("popRF")
```

```
pop_table <- list("NPL"="/user/npl_population.csv")

input_cov <- list(
                "NPL"=list(
                           "cov1" = "covariate1.tif",
                           "cov2" = "covariate2.tif"))


input_mastergrid <- list("NPL" = "npl_mastergrid.tif")
input_watermask  <- list("NPL" = "npl_watermask.tif")
input_px_area    <- list("NPL" = "npl_px_area.tif")

res <- popRF(pop=pop_table,
             cov=input_cov,
             mastergrid=input_mastergrid,
             watermask=input_watermask,
             px_area=input_px_area,
             output_dir="/user/output",
             cores=4)

# Plot populataion raster
plot(res$pop)

# Plot Error via Trees
plot(res$popfit)


## End(Not run)
```

---

popRFdemo *Function to demo the popRF package using WorldPop input data.*

Description

This function allows the user to generate a population layer using the WorldPop geospatial covariates and subnational census-based population estimates for 230 countries. All input datasets use a geographical coordinate system (GCS) with WGS 1984 datum (EPSG:4326) in Geotiff format at a resolution of 3 arc-second (0.00083333333 decimal degree, approximately 100m at the equator). Mastergrid of sub-national administrative unit boundary was rasterised by CIESIN.

Following covariates will be downloaded and used to disaggregat population (2020 year) from census units into grid cells.

- subnational_admin_2000_2020.tif - sub-national units provided by nationalEAs
- esaccilc_dst011_2015.tif - Distance to ESA-CCI-LC cultivated area edges 2015.
- esaccilc_dst040_2015.tif - Distance to ESA-CCI-LC woody-tree area edges 2015.
- esaccilc_dst130_2015.tif - Distance to ESA-CCI-LC shrub area edges 2015.
- esaccilc_dst140_2015.tif - Distance to ESA-CCI-LC herbaceous area edges 2015.
- esaccilc_dst150_2015.tif - Distance to ESA-CCI-LC sparse vegetation area edges 2015.

- esaccilc_dst160_2015.tif - Distance to ESA-CCI-LC aquatic vegetation area edges 2015.
- esaccilc_dst190_2015.tif - Distance to ESA-CCI-LC artificial surface edges 2015.
- esaccilc_dst200_2015.tif - Distance to ESA-CCI-LC bare area edges 2015.
- esaccilc_dst_water_100m_2000_2012.tif - ESA-CCI-LC inland waterbodies 2000-2012.
- coastline_100m_2000_2020.tif - Distance to coastline 2000-2020.
- dst_roadintersec_100m_2016.tif - Distance to OSM major road intersections.
- dst_waterway_100m_2016.tif - Distance to OSM major waterways.
- dst_road_100m_2016.tif - Distance to OSM major roads.
- px_area.tif - Grid-cell surface areas.
- srtm_slope_100m.tif - SRTM-based slope 2000 (SRTM is Shuttle Radar Topography Mission).
- srtm_topo_100m.tif - SRTM elevation 2000.
- viirs_100m_2016.tif - VIIRS night-time lights 2015 (VIIRS is Visible Infrared Imaging Radiometer Suite).
- wdpa_dst_cat1_100m_2017.tif - Distance to IUCN strict nature reserve and wilderness area edges 2017.
- dst_bsgme_100m_2020.tif - Distance to predicted built-settlement extents in 2020.

All downloaded files will be saved into subdirectory `covariates`.

Usage

```
popRFdemo(project_dir,
          country="NPL",
          cores=0,
          quant=TRUE,
          ftp=TRUE,
          verbose=TRUE,
          log=TRUE, ...)
```

Arguments

`project_dir`	Path to the folder to save the outputs.
`country`	character. ISO of the country (see country codes). Default one is NPL (Nepal)
`cores`	is a integer. Number of cores to use when executing the function. If set to 0 (`max_number_of_cores -1`) will be used based on as many processors as the hardware and RAM allow. Default is `cores` = 0.
`quant`	If FALSE then quant will not be calculated
`ftp`	is logical. TRUE or FALSE: flag indicating whether FTP or HTTPS of WorldPop data hub server will be used. Default is `ftp` = TRUE.
`verbose`	is logical. TRUE or FALSE: flag indicating whether to print intermediate output from the function on the console, which might be helpful for model debugging. Default is `verbose` = TRUE.

log — is logical. TRUE or FALSE: flag indicating whether to print intermediate output from the function on the log.txt file. Default is log = FALSE.

... — Additional arguments:
binc: Numeric. Increase number of blocks sugesting for processing raster file.
boptimise: Logical. Optimize total memory requires to processing raster file by reducing the memory need to 35%.
bsoft: Numeric. If raster can be processed on less then cores it will be foresed to use less number of cores.
nodesize: Minimum size of terminal nodes. Setting this number larger causes smaller trees to be grown (and thus take less time). See randomForest for more details. Default is nodesize = NULL and will be calculated as length(y_data)/1000.
maxnodes Maximum number of terminal nodes trees in the forest can have. If not given, trees are grown to the maximum possible (subject to limits by nodesize). If set larger than maximum possible, a warning is issued. See randomForest for more details. Default is maxnodes = NULL.
ntree Number of variables randomly sampled as candidates at each split. See randomForest for more details. Default is ntree = NULL and ntree will be used popfit$ntree
mtry Number of trees to grow. This should not be set to too small a number, to ensure that every input row gets predicted at least a few times. See randomForest for more details. Default is ntree = NULL and ntree will be used popfit$mtry.

Value

Raster* object of gridded population surfaces.

References

- Global spatio-temporally harmonised datasets for producing high-resolution gridded population distribution datasets doi: 10.1080/20964471.2019.1625151.
- WorldPop (www.worldpop.org - School of Geography and Environmental Science, University of Southampton; Department of Geography and Geosciences, University of Louisville; Departement de Geographie, Universite de Namur) and Center for International Earth Science Information Network (CIESIN), Columbia University (2018). Global High Resolution Population Denominators Project - Funded by The Bill and Melinda Gates Foundation (OPP1134076) doi: 10.5258/SOTON/WP00649.

Examples

```
## Not run:
popRFdemo(project_dir="/home/user/demo",
          country="NPL",
          cores=0)

## End(Not run)
```

Index