Package 'gridsample'

November 30, 2016

Title Tools for Grid-Based Survey Sampling Design

Version 0.1.3

Author Dana R. Thomson (University of Southampton), Nick W. Ruktanonchai (University of Southampton), Forrest R. Stevens (University of Louisville), Marcia Castro (Harvard University), Andrew J. Tatem (University of Southampton)

Maintainer Nick Ruktanonchai <nrukt00@gmail.com>

Description Multi-stage cluster household surveys are commonly performed by governments and programs to monitor population demographic, social, economic, and health outcomes. In these surveys, communities are sampled in a first stage of sampling from within subpopulations of interest (or strata), households are sampled in a second stage of sampling, and sometimes individuals are listed and further sampled within households. The first stage of sampling, where communities of sample populations are defined, are called Primary Sampling Units (PSUs) while the households are secondary sampling units (SSUs). Census data are typically used to select PSUs within strata. If census data are outdated, inaccurate, or not available at fine enough scale, however, gridded population data can be used instead. This tool selects PSUs within user-defined strata using gridded population data, given desired numbers of sampled households within each PSU. The population densities used to create PSUs are drawn from rasters such as the population data from the WorldPop Project (http://www.worldpop.org.uk). PSUs are defined within a stratum using a serpentine sampling method, and can be set to have a certain ratio of urban and rural PSUs, or to be evenly distributed across a coarse, user-defined grid.

Depends R (>= 3.2.3)

License MIT + file LICENSE

Encoding UTF-8

LazyData true

Imports rgdal, raster, data.table, rgeos, geosphere, sp, spatstat, methods, maptools

RoxygenNote 5.0.1.9000

Suggests knitr, rmarkdown

NeedsCompilation no

Repository CRAN

Date/Publication 2016-11-30 15:22:40

R topics documented:

`gs_mode` *Most common stratum calculator*

Description

For each cell in the coarse user-defined grid (only specified if `cfg_sample_spatial == TRUE`), this function calculates the stratum that occurs most often within each cell.

Usage

```
gs_mode(rast)
```

Arguments

`rast`	`data.table` object. This `data.table` where each cell that lies within a larger grid cell is represented as a row. For each row, the variable `grid_id` is the ID of the cell from the coarser grid, `sampled` denotes whether a cell has been sampled, `stratum` defines the stratum each cell lies within, and `raster_index` is a unique value for each cell in the raster.

Value

Vector of values representing the stratum that occurs most often within a given subset of the raster.

Author(s)

Forrest R. Stevens, <forrest.stevens@louisville.edu>

gs_rasterize *Rasterize polygon layer*

Description

This function creates a raster layer that adopts values from a defined field in a polygon layer, using `rasterize` from the raster package. This function also converts values to binary if desired, where all zero values are recorded as zero, and all non-zero values are recorded as one. This function also saves the output raster in the working directory.

Usage

```
gs_rasterize(input_features, output_raster, template_raster, binary = FALSE,
  field = "ID", overwrite = FALSE, format = "GTiff")
```

Arguments

`input_features`	SpatialPolygons* object. Name of input shapefile layer. Should be a SpatialPolygons object.
`output_raster`	Character. Desired name of output raster layer.
`template_raster`	Raster* object. Raster with desired characteristics (resolution, extent) of output raster.
`binary`	logical. If `TRUE`, any non-zero values will be converted to one.
`field`	character. Name of variable that output raster should inherit.
`overwrite`	logical. Defines whether to overwrite if `output_raster` already exists.
`format`	character. Desired format of output raster file.

Value

Vector of values representing the stratum that occurs most often within a given subset of the raster.

Author(s)

Forrest R. Stevens, <forrest.stevens@louisville.edu>

gs_sample *GridSample sampling algorithm*

Description

The gs_sample algorithm creates primary sampling units (PSUs) for multi-stage cluster household surveys based on gridded population data. Typical complex survey design is supported with input of, a raster of urbanized areas, and a raster of study strata. Each of these rasters need to be in an identical projection and have an identical grid resolution.

Usage

```
gs_sample(population_raster, strata_raster, urban_raster,
  desired_cell_size = NA, cfg_hh_per_stratum, cfg_hh_per_urban,
  cfg_hh_per_rural, max_psu_size = Inf, min_pop_per_cell, cfg_pop_per_psu,
  cfg_sample_rururb = FALSE, cfg_sample_spatial = FALSE,
  cfg_sample_spatial_scale, output_path, sample_name)
```

Arguments

population_raster	Raster* layer. Input population raster for PSU creation. Values should be number of people in each pixel.
strata_raster	Raster* layer. IRaster that defines the stratum that each pixel lies within. Generally created by rasterizing shapefile of polygons that indicate strata.
urban_raster	Raster* layer. IRaster of urbanized areas. where a cell value of 1 indicates urban cells and 0 indicates rural cells.
desired_cell_size	numeric. Desired cell size (in square kilometers) for output raster of PSUs. Defaults to NA, which yields an output raster at the same resolution as population_raster.
cfg_hh_per_stratum	numeric. Target household sample size per stratum. In a non-stratified sample, this is the total sample size of households. In a stratified sample, this is the household sample size per stratum.
cfg_hh_per_urban	numeric. Number of households expected to be selected per urban PSU during survey fieldwork.
cfg_hh_per_rural	numeric. Number of households expected to be selected per rural PSU during survey fieldwork.
max_psu_size	numeric. Maximum allowed geographic size of a given PSU in square kilometers. Defaults to infinity.

`min_pop_per_cell`
: numeric. Minimum population in a raster cell required for it to be considered for sampling. Cells with less than this value will be excluded from the sample. Defaults to 0, therefore including all cells.

`cfg_pop_per_psu`
: numeric. Target minimum population per PSU.

`cfg_sample_rururb`
: logical. Oversample rural/urban areas if one domain does not meet the target sample size per stratum? Default is `FALSE`.

`cfg_sample_spatial`
: logical. Oversample to ensure that at least one PSU is found within a larger grid with cell size defined by user? Default is `FALSE`.

`cfg_sample_spatial_scale`
: If `cfg_sample_spatial == TRUE`, this defines the cell size in the larger grid where the algorithm will ensure at least one PSU is located in each larger grid cell. Defined in square kilometers.

`output_path`
: character. Output folder name.

`sample_name`
: character. Name of output PSU raster and shapefile.

Details

A number of sampling features are optional. Oversampling in urban/rural areas, oversampling to be spatially representative, and stratification are not required. At a minimum, the user generates a simple random sample of PSUs in a study area by inputting a `population_raster`, defining the study area boundary as one stratum with strata_raster, defining the output shapefile parameters `output_path` and `sample_name`, and configuring the parameters `cfg_hh_per_stratum`, `cfg_hh_per_urban`, `cfg_hh_per_rural`, and `cfg_pop_per_psu`. See the "Stratification", "Urban/rural domains", "Spatial sampling", and "PSU size and framework" sections for additional information.

Value

Shapefile of household survey primary sampling unit (PSU) boundaries

Stratification

To stratify the sample, define strata geographic boundaries with `strata_raster`, and specify the sample size per strata with `cfg_hh_per_stratum`. For example, if a national sample will have 10,000 households from 5 provinces, then `cfg_hh_per_stratum = 2000`. The parameter `cfg_hh_per_stratum` is the minimum sample size to generate representative population statistics. In some surveys, strata are represent urban/rural populations within administrative units. If this is the case, then `strata_raster` should include the boundaries of urban and rural sampling areas within each administrative area, and `cfg_hh_per_stratum` should reflect the correct sample size per stratum - for example, a national sample of 10,000 households from each urban and rural areas in 5 provinces would have `cfg_hh_per_stratum = 1000`.

Urban/rural domains

If urban/rural populations are not part of the stratification scheme, then they are often treated as a sub-domain. Sub-domains represent important sub-populations for which representative statistics are generated from the survey data, and thus each sub-domain should meet the minimum sample size set for each stratum. If either the urban/rural sub-domain does not include enough households to generate population statistics, then the sub-domain is oversampled. To implement this step with `gs_sample`, set `cfg_sample_rururb = 1`. In practice, rural areas are often more difficult and expensive to visit, and thus a greater number of households might be sampled from rural PSUs than urban PSUs. This is why the user may specify different numbers of households to be sampled from each urban PSUs (`cfg_hh_per_urban`) and rural PSUs (`cfg_hh_per_rural`); if the same number of households will be sampled from all PSUs, then configure both of these parameters with the same value. Note, the number of PSUs that will be generated in each stratum is `cfg_hh_per_stratum` divided by some number between `cfg_hh_per_urban` and `cfg_hh_per_rural`.

Spatial sampling

To select a sample that is both representative of the population and of space, set `cfg_sample_spatial = 1` and specify `cfg_sample_spatial_scale`, the spatial scale at which the sample should be representative. The spatial scale should be meaningful; for example, it will facilitate small area estimates with limited statistical error for the administrative units below the administrative units used to stratify the sample. Determining an appropriate spatial scale might take trial and error. If the study area has large regions of sparse population, a typical non-spatially representative sample will follow a distribution similar to the population and have large areas without a PSU. In this case, the user might need to increase the spatial resolution of the sample, or force the algorithm to generate more PSUs in each stratum by increasing `cfg_hh_per_stratum` and/or reducing `cfg_hh_per_urban` and `cfg_hh_per_rural`.

PSU size and fieldwork

Three additional parameters can be configured to deal with idiosyncrasies of gridded population data and improve feasibility of fieldwork. The user can set a maximum geographic size of PSU in square kilometers, `max_psu_size`. We recommend choosing a size that can feasibly be visited on foot during one day. The user might also specify which cells are included in the sample frame with `min_pop_per_cell`. Selection of a sensible value is highly dependent on the specific gridded population dataset being used, and the scale of the input data (eg 200m grid cells). Finally, the cell size of the output raster can be specified with `desired_cell_size`, which can be modified to account for the expected accuracy of the input gridded population datasets.

Examples

```
require(raster)
poprast <- raster(ncols=50,nrows=50,xmx=10,xmn=9,ymn=9,ymx=10,
crs=CRS("+proj=longlat +datum=WGS84 +no_defs +ellps=WGS84 +towgs84=0,0,0"),
vals=runif(2500,0,1))
stratarast <- raster(ncols=50,nrows=50,xmx=10,xmn=9,ymn=9,ymx=10,
crs=CRS("+proj=longlat +datum=WGS84 +no_defs +ellps=WGS84 +towgs84=0,0,0"),
vals=c(rep(1,times=1250),rep(2,times=1250)))
urbanrast <- poprast > 0.9
gs_sample(population_raster = poprast,
```

```
            strata_raster = stratarast,
            urban_raster = urbanrast,
            desired_cell_size = NA,
            cfg_hh_per_stratum = 20,
            cfg_hh_per_urban = 2,
            cfg_hh_per_rural = 2,
            min_pop_per_cell = 0.01,
            cfg_pop_per_psu = 10,
            cfg_sample_rururb = FALSE,
            cfg_sample_spatial = FALSE,
            cfg_sample_spatial_scale = 100,
            output_path=tempdir(),
            sample_name="Example")
```

`gs_zonal_raster` *Zonal statistics calculator*

Description

This function calculates zonal statstics across a raster layer, for each polygon in a rasterized polygon layer.

Usage

```
gs_zonal_raster(x, z, stat = "mean", digits = 1, na.rm = TRUE, ...)
```

Arguments

`x`	Raster* layer. The layer that zonal statistics should be calculated from.
`z`	Raster* layer. A rasterized version of the zonal layer.
`stat`	character. Name of statistic used to calculate a value across each polygon in `z`. ex. "mean", "sum".
`digits`	numeric. Number of significant digits in zonal statistic output.
`na.rm`	logical. Defines whether to remove `NA`
`...`	Other variables

Value

Vector of values representing the calculated statistic for each polygon, sorted by the order of polygons in the polygon layer.

Author(s)

Forrest R. Stevens, <forrest.stevens@louisville.edu>

Index