Package ‘spatialsample’

June 17, 2022

Title Spatial Resampling Infrastructure
Version 0.2.0
Description Functions and classes for spatial resampling to use with the ‘rsample’ package, such as spatial cross-validation (Brenning, 2012) <doi:10.1109/IGARSS.2012.6352393>. The scope of ‘rsample’ and ‘spatialsample’ is to provide the basic building blocks for creating and analyzing resamples of a spatial data set, but neither package includes functions for modeling or computing statistics. The resampled spatial data sets created by ‘spatialsample’ do not contain much overhead in memory.

License MIT + file LICENSE
URL https://github.com/tidymodels/spatialsample,
https://spatialsample.tidymodels.org

BugReports https://github.com/tidymodels/spatialsample/issues
Depends R (>= 3.4)
Imports dplyr (>= 1.0.0), ggplot2, glue, purrr, rlang (>= 1.0.0), rsample (>= 0.9.0), sf, tibble, tidyselect, units, vctrs (>= 0.3.6)

Suggests covr, gifski, knitr, lwgeom, modeldata, rmarkdown, testthat (>= 3.0.0), tidyr, vdiffr, yardstick

VignetteBuilder knitr

Config/Needs/website tidyverse/tidytemplate
Config/testthat/edition 3
Config/testthat/parallel true
Encoding UTF-8
LazyData true

RoxygenNote 7.2.0

NeedsCompilation no

Author Julia Silge [aut, cre] (<https://orcid.org/0000-0002-3671-836X>), Michael Mahoney [aut] (<https://orcid.org/0000-0003-2402-304X>), RStudio [cph, fnd]
**Description**

This method provides a good visualization method for spatial resampling.

**Usage**

```r
## S3 method for class 'spatial_rset'
autoplay(object, ..., alpha = 0.6)

## S3 method for class 'spatial_block_cv'
autoplay(object, show_grid = TRUE, ..., alpha = 0.6)
```

**Arguments**

- `object` A `spatial_rset` object or a `spatial_rsplit` object. Note that only resamples made from `sf` objects create `spatial_rset` and `spatial_rsplit` objects; this function will not work for resamples made with non-spatial tibbles or data.frames.
- `...` Options passed to `ggplot2::geom_sf()`.
- `alpha` Opacity, passed to `ggplot2::geom_sf()`. Values of alpha range from 0 to 1, with lower values corresponding to more transparent colors.
- `show_grid` When plotting `spatial_block_cv` objects, should the grid itself be drawn on top of the data? Set to FALSE to remove the grid.

**Details**

The plot method for `spatial_rset` displays which fold each observation is assigned to. Note that if data is assigned to multiple folds (which is common if resamples were created with a non-zero radius) only the "last" fold for each observation will appear on the plot. Consider adding `ggplot2::facet_wrap(~ fold)` to visualize all members of each fold separately. Alternatively, consider plotting each split using the `spatial_rsplit` method (for example, via `lapply(object$splits, autoplot)`).
Value

A ggplot object with each fold assigned a color, made using `ggplot2::geom_sf()`.

Examples

```r
boston_block <- spatial_block_cv(boston_canopy, v = 2)
autoplot(boston_block)
lapply(boston_block$splits, autoplot)
```

---

### boston_canopy

*Boston tree canopy and heat index data.*

**Description**

A dataset containing data on tree canopy coverage and change for the city of Boston, Massachusetts from 2014-2019, as well as temperature and heat index data for July 2019. Data is aggregated to a grid of regular 25 hectare hexagons, clipped to city boundaries. This data is made available under the Public Domain Dedication and License v1.0 whose full text can be found at: [https://opendatacommons.org/licenses/pddl/1-0/](https://opendatacommons.org/licenses/pddl/1-0/).

**Usage**

`boston_canopy`

**Format**

A data frame (of class `sf`, `tbl_df`, `tbl`, and `data.frame`) containing 682 records of 22 variables:

- **grid_id** Unique identifier for each hexagon. Letters represent the hexagon’s X position in the grid (ordered West to East), while numbers represent the Y position (ordered North to South).
- **land_area** Area excluding water bodies
- **canopy_gain** Area of canopy gain between the two years
- **canopy_loss** Area of canopy loss between the two years
- **canopy_no_change** Area of no canopy change between the two years
- **canopy_area_2014** 2014 total canopy area (baseline)
- **canopy_area_2019** 2019 total canopy area
- **change_canopy_area** The change in area of tree canopy between the two years
- **change_canopy_percentage** Relative change calculation used in economics is the gain or loss of tree canopy relative to the earlier time period: (2019 Canopy-2014 Canopy)/(2014 Canopy)
- **canopy_percentage_2014** 2014 canopy percentage
- **canopy_percentage_2019** 2019 canopy percentage
**change_canopy_absolute**  Absolute change. Magnitude of change in percent tree canopy from 2014 to 2019 (% 2019 Canopy - % 2014 Canopy)

**mean_temp_morning**  Mean temperature for July 2019 from 6am - 7am

**mean_temp_evening**  Mean temperature for July 2019 from 7pm - 8pm

**mean_temp**  Mean temperature for July 2019 from 6am - 7am, 3pm - 4pm, and 7pm - 8pm (combined)

**mean_heat_index_morning**  Mean heat index for July 2019 from 6am - 7am

**mean_heat_index_evening**  Mean heat index for July 2019 from 7pm - 8pm

**mean_heat_index**  Mean heat index for July 2019 from 6am - 7am, 3pm - 4pm, and 7pm - 8pm (combined)

**geometry**  Geometry of each hexagon, encoded using EPSG:2249 as a coordinate reference system (NAD83 / Massachusetts Mainland (ftUS)). Note that the linear units of this CRS are in US feet.

**Details**

Note that this dataset is in the EPSG:2249 (NAD83 / Massachusetts Mainland (ftUS)) coordinate reference system (CRS), which may not be installed by default on your computer. Before working with boston_canopy, run:

- `sf::sf_proj_network(TRUE)` to install the CRS itself
- `sf::sf_add_proj_units()` to add US customary units to your units database

These steps only need to be taken once per computer (or per PROJ installation).

**Source**


---

**spatial_block_cv**  Spatial block cross-validation

**Description**

Block cross-validation splits the area of your data into a number of grid cells, or "blocks", and then assigns all data into folds based on the blocks their centroid falls into.
**spatial_block_cv**

**Usage**

```r
spatial_block_cv(
  data,
  method = c("random", "snake", "continuous"),
  v = 10,
  relevant_only = TRUE,
  radius = NULL,
  buffer = NULL,
  ...
)
```

**Arguments**

- **data**: An object of class sf or sfc.
- **method**: The method used to sample blocks for cross validation folds. Currently supports "random", which randomly assigns blocks to folds, "snake", which labels the first row of blocks from left to right, then the next from right to left, and repeats from there, and "continuous", which labels each row from left to right, moving from the bottom row up.
- **v**: The number of partitions of the data set.
- **relevant_only**: For systematic sampling, should only blocks containing data be included in fold labeling?
- **radius**: Numeric: points within this distance of the initially-selected test points will be assigned to the assessment set. If NULL, no radius is applied.
- **buffer**: Numeric: points within this distance of any point in the test set (after radius is applied) will be assigned to neither the analysis or assessment set. If NULL, no buffer is applied.
- **...**: Arguments passed to `sf::st_make_grid()`.

**Details**

The grid blocks can be controlled by passing arguments to `sf::st_make_grid()` via `. . . .`. Some particularly useful arguments include:

- **cellsize**: Target cellsize, expressed as the "diameter" (shortest straight-line distance between opposing sides; two times the apothem) of each block, in map units.
- **n**: The number of grid blocks in the x and y direction (columns, rows).
- **square**: A logical value indicating whether to create square (TRUE) or hexagonal (FALSE) cells.

If both `cellsize` and `n` are provided, then the number of blocks requested by `n` of sizes specified by `cellsize` will be returned, likely not lining up with the bounding box of `data`. If only `cellsize` is provided, this function will return as many blocks of size `cellsize` as fit inside the bounding box of `data`. If only `n` is provided, then `cellsize` will be automatically adjusted to create the requested number of cells.
Value

A tibble with classes spatial_block_cv, spatial_rset, rset, tbl_df, tbl, and data.frame. The results include a column for the data split objects and an identification variable id.

References


Examples

spatial_block_cv(boston_canopy, v = 3)

spatial_buffer_vfold_cv

V-Fold Cross-Validation with Buffering

Description

V-fold cross-validation (also known as k-fold cross-validation) randomly splits the data into V groups of roughly equal size (called "folds"). A resample of the analysis data consists of V-1 of the folds while the assessment set contains the final fold. These functions extend rsample::vfold_cv() and rsample::group_vfold_cv() to also apply an inclusion radius and exclusion buffer to the assessment set, ensuring that your analysis data is spatially separated from the assessment set. In basic V-fold cross-validation (i.e. no repeats), the number of resamples is equal to V.

Usage

spatial_buffer_vfold_cv(
  data,
  radius,
  buffer,
  v = 10,
  repeats = 1,
  strata = NULL,
  breaks = 4,
  pool = 0.1,
  ...
)

spatial_leave_location_out_cv(
  data,
  group,
  ...)
spatial_buffer_vfold_cv

v = NULL,
radius = NULL,
buffer = NULL,
...
)

Arguments

data  A data frame.
radius Numeric: points within this distance of the initially-selected test points will be
assigned to the assessment set. If NULL, no radius is applied.
buffer Numeric: points within this distance of any point in the test set (after radius is
applied) will be assigned to neither the analysis or assessment set. If NULL, no
buffer is applied.
v The number of partitions of the data set.
repeats The number of times to repeat the V-fold partitioning.
strata A variable in data (single character or name) used to conduct stratified sam-
ppling. When not NULL, each resample is created within the stratification variable.
Numeric strata are binned into quartiles.
b breaks A single number giving the number of bins desired to stratify a numeric stratifi-
cation variable.
pool A proportion of data used to determine if a particular group is too small and
should be pooled into another group. We do not recommend decreasing this
argument below its default of 0.1 because of the dangers of stratifying groups
that are too small.
... Not currently used.
group A variable in data (single character or name) used to create folds. For leave-
location-out CV, this should be a variable containing the locations to group
observations by, for leave-time-out CV the time blocks to group by, and for
leave-location-and-time-out the spatiotemporal blocks to group by.

Details

When radius and buffer are both NULL, spatial_buffer_vfold_cv is equivalent to rsample::vfold_cv() and spatial_leave_location_out_cv is equivalent to rsample::group_vfold_cv().

References


Examples

```r
boston_vfold <- spatial_buffer_vfold_cv(
  boston_canopy,
  buffer = 500,
  radius = NULL
)
data(ames, package = "modeldata")
ames_sf <- sf::st_as_sf(ames, coords = c("Longitude", "Latitude"), crs = 4326)
ames_neighborhoods <- spatial_leave_location_out_cv(ames_sf, Neighborhood)
```

---

**spatial_clustering_cv**  
*Spatial or Cluster Cross-Validation*

**Description**

Spatial or cluster cross-validation splits the data into V groups of disjointed sets using k-means clustering of some variables, typically spatial coordinates. A resample of the analysis data consists of V-1 of the folds/clusters while the assessment set contains the final fold/cluster. In basic spatial cross-validation (i.e. no repeats), the number of resamples is equal to V.

**Usage**

```r
spatial_clustering_cv(
  data,
  coords,
  v = 10,
  cluster_function = c("kmeans", "hclust"),
  radius = NULL,
  buffer = NULL,
  ...
)
```

**Arguments**

- **data** A data frame or an sf object (often from `sf::read_sf()` or `sf::st_as_sf()`), to split into folds.
- **coords** A vector of variable names, typically spatial coordinates, to partition the data into disjointed sets via k-means clustering. This argument is ignored (with a warning) if data is an sf object.
- **v** The number of partitions of the data set.
- **cluster_function** Which function should be used for clustering? Options are either "kmeans" (to use `stats::kmeans()`) or "hclust" (to use `stats::hclust()`). You can also provide your own function; see Details.
radius Numeric: points within this distance of the initially-selected test points will be assigned to the assessment set. If NULL, no radius is applied.

buffer Numeric: points within this distance of any point in the test set (after radius is applied) will be assigned to neither the analysis or assessment set. If NULL, no buffer is applied.

... Extra arguments passed on to \texttt{stats::kmeans()} or \texttt{stats::hclust()}.

Details

The variables in the \texttt{coords} argument are used for k-means clustering of the data into disjointed sets, as outlined in Brenning (2012), or for hierarchical clustering of the data. These clusters are used as the folds for cross-validation. Depending on how the data are distributed spatially, there may not be an equal number of points in each fold.

You can optionally provide a custom function to \texttt{cluster\_function}. The function must take three arguments:

- \texttt{dists}, a \texttt{stats::dist()} object with distances between data points
- \texttt{v}, a length-1 numeric for the number of folds to create
- ... to pass any additional named arguments to your function

The function should return a vector of cluster assignments of length \texttt{nrow(data)}, with each element of the vector corresponding to the matching row of the data frame.

Value

A tibble with classes \texttt{spatial\_clustering\_cv}, \texttt{spatial\_rset}, \texttt{rset}, \texttt{tbl\_df}, \texttt{tbl}, and \texttt{data.frame}. The results include a column for the data split objects and an identification variable \texttt{id}. Resamples created from non-sf objects will not have the \texttt{spatial\_rset} class.

References


Examples

data(Smithsonian, package = "modeldata")
spatial\_clustering\_cv(Smithsonian, coords = c(latitude, longitude), v = 5)

smithsonian\_sf <- sf::st\_as\_sf(
    Smithsonian, 
    coords = c("longitude", "latitude"), 
    # Set CRS to WGS84 
    crs = 4326
)

# When providing sf objects, coords are inferred automatically
spatial\_clustering\_cv(smithsonian\_sf, v = 5)
# Can use hclust instead:
spatial_clustering_cv(smithsonian_sf, v = 5, cluster_function = "hclust")
Index

* datasets
  boston_canopy, 3

autoplot.spatial_block_cv
  (autoplot.spatial_rset), 2
autoplot.spatial_rset, 2

boston_canopy, 3

ggplot2::geom_sf(), 2, 3

rsample::group_vfold_cv(), 6, 7
rsample::vfold_cv(), 6, 7

sf::read_sf(), 8
sf::sf_add_proj_units(), 4
sf::st_as_sf(), 8
sf::st_make_grid(), 5
spatial_block_cv, 2, 4
spatial_buffer_vfold_cv, 6
spatial_clustering_cv, 8
spatial_leave_location_out_cv
  (spatial_buffer_vfold_cv), 6
stats::dist(), 9
stats::hclust(), 8, 9
stats::kmeans(), 8, 9