# Package 'lightsf'

October 19, 2025

**Type** Package

**Title** A Curated Collection of Georeferenced and Spatial Datasets

**Version** 0.1.0

**Maintainer** Ingrid Romero Pinilla <ingridpinilla11@gmail.com>

**Description** Provides a diverse collection of georeferenced and spatial datasets from different domains including urban studies, housing markets, environmental monitoring, transportation, and socio-economic indicators.
The package consolidates datasets from multiple open sources such as Kaggle, chopin, spData, adespatial, and bivariateLeaflet.
It is designed for researchers, analysts, and educators interested in spatial analysis, geostatistics, and geographic data visualization.
The datasets include point patterns, polygons, socio-economic data frames, and network-like structures, allowing flexible exploration of geospatial phenomena.

**License** GPL-3

**URL** https://github.com/roming20/lightsf,
https://roming20.github.io/lightsf/

**BugReports** https://github.com/roming20/lightsf/issues

**Encoding** UTF-8

**LazyData** true

**Suggests** ggplot2, dplyr, testthat (>= 3.0.0), knitr, rmarkdown

**RoxygenNote** 7.3.2

**Config/testthat/edition** 3

**VignetteBuilder** knitr

**NeedsCompilation** no

**Author** Ingrid Romero Pinilla [aut, cre]

**Depends** R (>= 3.5.0)

**Repository** CRAN

**Date/Publication** 2025-10-19 13:10:02 UTC

# Contents

---

`afcon_poly` *Spatial Patterns of Conflict in Africa (1966–1978)*

---

### Description

This dataset, 'afcon_poly', is a data frame summarizing spatial patterns of conflict across 42 African countries between 1966 and 1978. The dataset was originally used in Anselin (1995) to study spatial autocorrelation in political conflict. It excludes South West Africa, Spanish Equatorial Africa, and Spanish Sahara. The dataset includes centroid coordinates, country names, and the total number of recorded conflicts during this period.

### Usage

```
data(afcon_poly)
```

### Format

A data frame with 42 observations and 5 variables:

**x** Longitude coordinate of the country centroid (numeric)

**y** Latitude coordinate of the country centroid (numeric)

**totcon** Total number of conflicts recorded, 1966–1978 (numeric)

**name** Name of the country (factor with 42 levels)

**id** Numeric country identifier (numeric)

### Details

The dataset consists of 42 observations (countries) and 5 variables.

The dataset name has been kept as ‘afcon_poly‘ to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the ‘lightsf‘ package and assists users in identifying its specific characteristics. The original content has not been modified in any way.

### Source

Data taken from the **spData** package version 2.3.4.

### References

Anselin, L. (1995). Local Indicators of Spatial Association—LISA. *Geographical Analysis*, 27(2), 93–115.

---

## `atropellados_pts` *Georeferenced Pedestrian Car Collisions (2015, Santiago de Chile)*

---

### Description

This dataset, atropellados_pts, is a data frame containing information on pedestrian car collisions that occurred in Santiago de Chile in 2015. Each record includes the geographical coordinates of the accident, location description, and the number of victims categorized by severity (fatal, serious, less serious, and minor).

### Usage

```
data(atropellados_pts)
```

### Format

A data frame with 1,841 observations and 8 variables:

**X** Longitude coordinate of the accident (numeric)

**Y** Latitude coordinate of the accident (numeric)

**Ubicacion** Location description of the accident (character)

**Fallecidos** Number of fatalities (integer)

**Graves** Number of serious injuries (integer)

**MenosGrave** Number of less serious injuries (integer)

**Leve** Number of minor injuries (integer)

**Accidentes** Total number of accidents at the location (integer)

## Details

The dataset name has been kept as 'atropellados_pts' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the lightsf package and assists users in identifying its specific characteristics. The original content has not been modified in any way.

## Source

Data taken from Kaggle: https://www.kaggle.com/datasets/sandorabad/georeferenced-car-accidents-santiago
select=AtropellosGS2015.csv

---

# `auckland_poly` *Infant Mortality in Auckland, New Zealand (1977–1985)*

---

## Description

This dataset, 'auckland_poly', is a data frame containing information on infant mortality in census area units (CAUs) of Auckland, New Zealand. The dataset has 167 rows, each corresponding to a CAU, and 4 columns with geographic coordinates and mortality-related statistics. It is often used in spatial epidemiology studies and in demonstrations of spatial analysis methods.

## Usage

```
data(auckland_poly)
```

## Format

A data frame with 167 observations and 4 variables:

**Easting** Easting coordinate (numeric)

**Northing** Northing coordinate (numeric)

**Deaths.1977.85** Number of infant deaths between 1977 and 1985 (numeric)

**Under.5.1981** Population under age 5 in 1981 (numeric)

## Details

In addition to the 'auckland_poly' data frame, the original source also provides two related spatial objects: 'auckland.nb', a neighbour list of CAUs based on contiguity, and 'auckpolys', a polylist object representing polygon boundaries. These are not included here, but can be generated from the original dataset using spatial analysis workflows.

The dataset name has been kept as 'auckland_poly' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the 'lightsf' package and assists users in identifying its specific characteristics. The original content has not been modified in any way.

## Source

Data taken from the **spData** package version 2.3.4.

## `bacprodxy_pts` *Bacterial Production Sampling Points in Lake St. Pierre (2005)*

### Description

This dataset, bacprodxy_pts, is a data frame containing the geographical coordinates (longitude and latitude) of 25 sampling locations where bacterial production was measured in Lake St. Pierre (Québec, Canada). The samples were collected on August 18, 2005.

### Usage

```
data(bacprodxy_pts)
```

### Format

A data frame with 25 observations and 2 variables:

**Longitude** Longitude coordinate of the sampling point (numeric)

**Latitude** Latitude coordinate of the sampling point (numeric)

### Details

The dataset name has been kept as 'bacprodxy_pts' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the lightsf package and assists users in identifying its specific characteristics. The original content has not been modified in any way.

### Source

Data taken from the **adespatial** package version 0.3-28

## `baltimore_pts` *Housing Sales in Baltimore, Maryland (1978)*

### Description

This dataset, 'baltimore_pts', is a data frame containing housing sales data and property characteristics for Baltimore, Maryland, in 1978. It has been widely used in spatial econometrics and hedonic regression studies. Each row corresponds to a house, including sale price, structural attributes, lot size, and geographic coordinates (X, Y) on the Maryland grid (projection type unknown).

### Usage

```
data(baltimore_pts)
```

## Format

A data frame with 211 observations and 17 variables:

**STATION** Census tract station identifier (integer)

**PRICE** House sale price (numeric)

**NROOM** Number of rooms (numeric)

**DWELL** Dwelling type indicator (numeric)

**NBATH** Number of bathrooms (numeric)

**PATIO** Presence of patio (numeric indicator)

**FIREPL** Presence of fireplace (numeric indicator)

**AC** Presence of air conditioning (numeric indicator)

**BMENT** Presence of basement (numeric indicator)

**NSTOR** Number of stories (numeric)

**GAR** Presence of garage (numeric indicator)

**AGE** Age of the dwelling (numeric)

**CITCOU** City/county indicator (numeric)

**LOTSZ** Lot size (numeric)

**SQFT** Interior square footage (numeric)

**X** X coordinate (numeric)

**Y** Y coordinate (numeric)

## Details

The dataset consists of 211 observations (houses) and 17 variables.

The dataset name has been kept as ‘baltimore_pts‘ to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the ‘lightsf‘ package and assists users in identifying its specific characteristics. The original content has not been modified in any way.

## Source

Data taken from the **spData** package version 2.3.4.

## `boston_pts` *Boston Housing Data with Geographic Coordinates*

### Description

This dataset, `boston_pts`, is a data frame containing information on housing values and neighborhood characteristics in the Boston area. It is based on the classic dataset by Harrison and Rubinfeld (1978), corrected for minor errors and augmented with the latitude and longitude of the observations. Gilley and Pace also note that the `MEDV` variable is censored, with values at or over USD 50,000 set to USD 50,000.

### Usage

```
data(boston_pts)
```

### Format

A data frame with 506 observations and 20 variables:

**TOWN** Town name (factor with 92 levels)

**TOWNNO** Town number (integer)

**TRACT** Census tract number (integer)

**LON** Longitude (numeric)

**LAT** Latitude (numeric)

**MEDV** Median value of owner-occupied homes in USD 1,000s (numeric, censored at 50)

**CMEDV** Corrected median value of owner-occupied homes (numeric)

**CRIM** Per capita crime rate by town (numeric)

**ZN** Proportion of residential land zoned for lots over 25,000 sq.ft. (numeric)

**INDUS** Proportion of non-retail business acres per town (numeric)

**CHAS** Charles River dummy variable (factor: "0" = not bounded, "1" = bounded)

**NOX** Nitric oxides concentration (parts per 10 million, numeric)

**RM** Average number of rooms per dwelling (numeric)

**AGE** Proportion of owner-occupied units built prior to 1940 (numeric)

**DIS** Weighted distances to five Boston employment centers (numeric)

**RAD** Index of accessibility to radial highways (integer)

**TAX** Full-value property-tax rate per `$10,000` (integer)

**PTRATIO** Pupil-teacher ratio by town (numeric)

**B** Proportion of Black residents, defined as 1000(Bk - 0.63)^2 (numeric)

**LSTAT** Percentage of lower status of the population (numeric)

## Details

The dataset consists of 506 observations and 20 variables, including socio-economic, environmental, and housing characteristics. Geographic coordinates (longitude and latitude) are provided for spatial analysis. Related data objects include `boston.utm`, a matrix of tract point coordinates projected to UTM zone 19, and `boston.soi`, a sphere of influence neighbors list.

The dataset name has been kept as `boston_pts` to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the `lightsf` package and assists users in identifying its specific characteristics. The suffix `pts` indicates that the dataset includes spatial point information. The original content has not been modified in any way.

## Source

Data taken from the **spData** package version 2.3.4

---

# `coffee_poly` *World Coffee Production Data*

---

## Description

This dataset, `coffee_poly`, is a tibble containing estimates of global coffee production by country. The data represent thousands of 60 kg bags of coffee produced in 2016 and 2017. It is intended for teaching purposes only and not for research use.

## Usage

```
data(coffee_poly)
```

## Format

A tibble with 47 observations and 3 variables:

**name_long** Country name (character)

**coffee_production_2016** Coffee production in 2016, in thousands of 60 kg bags (integer)

**coffee_production_2017** Coffee production in 2017, in thousands of 60 kg bags (integer)

## Details

The dataset consists of 47 observations (countries) and 3 variables, including the country name and production values for two years. The data provide a simple example of tabular international production figures that can be used in spatial and non-spatial analyses.

The dataset name has been kept as `coffee_poly` to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the `lightsf` package and assists users in identifying its specific characteristics. The suffix `poly` indicates that the dataset can be linked to polygon boundaries for mapping. The original content has not been modified in any way.

### Source

Data taken from the **spData** package version 2.3.4

---

`columbus_poly` *Columbus Neighborhood Data (1980)*

---

### Description

This dataset, `columbus_poly`, is a data frame containing socioeconomic and housing characteristics for 49 neighborhoods in Columbus, Ohio, based on 1980 data. The dataset is widely used in spatial econometrics and geographic analysis.

### Usage

```
data(columbus_poly)
```

### Format

A data frame with 49 observations and 22 variables:

**AREA** Area of the neighborhood (numeric)

**PERIMETER** Perimeter of the neighborhood (numeric)

**COLUMBUS.** Identifier variable (integer)

**COLUMBUS.I** Identifier variable (integer)

**POLYID** Polygon ID (integer)

**NEIG** Neighborhood ID (integer)

**HOVAL** Housing value (numeric)

**INC** Household income (numeric)

**CRIME** Crime rate (numeric)

**OPEN** Open space (numeric)

**PLUMB** Plumbing quality (numeric)

**DISCBD** Distance to central business district (numeric)

**X** X coordinate of centroid (numeric)

**Y** Y coordinate of centroid (numeric)

**AREA** Area variable (numeric, duplicated)

**NSA** Neighborhood spatial attribute A (numeric)

**NSB** Neighborhood spatial attribute B (numeric)

**EW** East/West indicator (numeric)

**CP** Central place indicator (numeric)

**THOUS** Thousands of dollars (numeric)

**NEIGNO** Neighborhood number (numeric)

**PERIM** Perimeter variable (numeric, duplicated)

### Details

In addition to the attributes, the original dataset also included a polygon list of neighborhood boundaries, a centroid matrix, and a neighbor list object, although these are not part of `columbus_poly`. The matrix `bbs` is deprecated but retained in other packages for compatibility.

The dataset name has been kept as `columbus_poly` to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the `lightsf` package and assists users in identifying its specific characteristics. The suffix `poly` indicates that the dataset can be linked to polygon boundaries. The original content has not been modified in any way.

### Source

Data taken from the **spData** package version 2.3.4

## `conafchile_pts` *Georeferenced Forest Fires in Chile (2016–2017 Season)*

### Description

This dataset, ‘conafchile_pts‘, is a data frame containing georeferenced forest fire records and associated characteristics between July 1, 2016, and June 30, 2017. The dataset includes detailed information such as location, administrative codes, fire causes, vegetation affected, and surface area impacted. The data were compiled by CONAF and correspond to forest fires recorded in Chile.

### Usage

```
data(conafchile_pts)
```

### Format

A data frame with 5,234 observations and 30 variables:

**X** Index of the fire record (integer)

**temporada** Fire season (character, e.g., "2016-2017")

**codreg** Region code (integer)

**codprov** Province code (integer)

**codcom** Commune code (integer)

**ambito** Institutional scope (character, e.g., "Conaf")

**numero** Fire identification number (numeric)

**nombre_inc** Name of the fire incident (character)

**utm_este** UTM Easting coordinate (numeric)

**utm_norte** UTM Northing coordinate (numeric)

**inicio_c** Location of ignition (character)

**combus_i** Initial fuel type (character)

**causa_gene** General cause code (numeric)

**causa_espe** Specific cause code (character)

**pino_0010** Surface with pine (0–10 years old) affected (numeric)

**pino_11_17** Surface with pine (11–17 years old) affected (numeric)

**pino_18** Surface with pine (18+ years old) affected (numeric)

**eucalipto** Surface with eucalyptus affected (numeric)

**otras_plan** Surface with other plantations affected (numeric)

**total_plan** Total surface of plantations affected (numeric)

**arbolado** Surface of woodland affected (numeric)

**matorral** Surface of shrubland affected (numeric)

**pastizal** Surface of grassland affected (numeric)

**total_veg** Total surface of vegetation affected (numeric)

**agricola** Surface of agricultural land affected (numeric)

**desechos** Surface of waste material affected (numeric)

**total_otra** Total surface of other land use affected (numeric)

**sup_t_a** Total affected surface area (numeric)

**long** Longitude or projected coordinate (numeric)

**lat** Latitude or projected coordinate (numeric)

### Details

The dataset name has been kept as ‘conafchile_pts‘ to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the lightsf package and assists users in identifying its specific characteristics. The suffix ‘pts‘ indicates that the dataset contains georeferenced point data. The original content has not been modified in any way.

### Source

Data taken from Kaggle: `https://www.kaggle.com/datasets/sandorabad/georeferenced-forestfires-2017-chile`

## `countries_pts` *Countries Latitude-Longitude Dataset*

### Description

This dataset, countries_pts, is a data frame containing information on 245 countries, including their names and geographical coordinates (latitude and longitude). It provides a simple reference for mapping and spatial analysis.

### Usage

```
data(countries_pts)
```

## Format

A data frame with 245 observations and 4 variables:

**country** Country code or identifier (character)

**latitude** Latitude of the country (numeric)

**longitude** Longitude of the country (numeric)

**name** Country name (character)

## Details

The dataset name has been kept as 'countries_pts' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the lightsf package and assists users in identifying its specific characteristics. The original content has not been modified in any way.

## Source

Data taken from Kaggle: `https://www.kaggle.com/datasets/arviinndn/countries`

---

# `cyclehire_pts` *Cycle Hire Stations in London*

---

## Description

This dataset, `cyclehire_pts`, is an `sf` object containing point locations of cycle hire stations across London. Each observation represents a hire point with information about its name, area, number of available bikes, and number of empty docking slots at the time of data collection.

## Usage

```
data(cyclehire_pts)
```

## Format

An `sf` object with 742 observations and 6 variables:

**id** Station identifier (integer)

**name** Name of the station (factor)

**area** Area of London where the station is located (factor with 121 levels)

**nbikes** Number of bikes available (integer)

**nempty** Number of empty docking slots (integer)

**geometry** Point geometry in XY coordinates (`sfc_POINT`)

## Details

The dataset name has been kept as `cyclehire_pts` to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the `lightsf` package and assists users in identifying its specific characteristics. The suffix `pts` indicates that the dataset contains point geometries. The original content has not been modified in any way.

## Source

Data taken from the **spData** package version 2.3.4

---

# `dc_poly` *Washington, D.C. Census Tract Data (ACS 2020)*

---

## Description

This dataset, ‘dc_poly‘, is an ‘sf‘ object containing population and median household income information for census tracts in Washington, D.C., based on the 2020 American Community Survey (ACS). It also includes spatial polygon geometries, allowing the data to be used directly for mapping and spatial analysis, such as creating choropleth maps of demographic and socioeconomic indicators.

## Usage

```
data(dc_poly)
```

## Format

An ‘sf‘ data frame with 206 observations and 5 variables:

**GEOID** Unique identifier for the census tract (character)

**NAME** Census tract name and jurisdiction (character)

**geometry** Polygon geometry representing the tract boundaries (‘`sfc_POLYGON`‘)

**B01003_001** Total population of the tract (numeric)

**B19013_001** Median household income of the tract (numeric, in USD)

## Details

The dataset consists of 206 observations (census tracts) and 5 variables. The geometry column contains polygon boundaries for each tract.

The dataset name has been kept as ‘dc_poly‘ to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the ‘lightsf‘ package and assists users in identifying its specific characteristics. The original content has not been modified in any way.

## Source

Data taken from the **bivariateLeaflet** package version 0.1.0

housing_pts *California Housing Prices (1990 Census)*

## Description

This dataset, ‘housing_pts‘, is a data frame containing information on median house prices for California districts, derived from the 1990 census. It includes geographic coordinates, demographic and housing characteristics, and district-level income and housing attributes. The dataset consists of 20,640 observations and 10 variables. Missing values may be present in some variables.

## Usage

```
data(housing_pts)
```

## Format

A data frame with 20,640 observations and 10 variables:

**longitude** Longitude coordinate of the district (numeric)

**latitude** Latitude coordinate of the district (numeric)

**housing_median_age** Median age of houses in the district (numeric)

**total_rooms** Total number of rooms in the district (numeric)

**total_bedrooms** Total number of bedrooms in the district (numeric)

**population** Population of the district (numeric)

**households** Number of households in the district (numeric)

**median_income** Median income in the district (numeric)

**median_house_value** Median house value in the district (numeric, in US dollars)

**ocean_proximity** Proximity of the district to the ocean (character string categories)

## Details

The dataset name has been kept as ‘housing_pts‘ to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of your package and assists users in identifying its specific characteristics. The suffix ‘pts‘ indicates that the dataset contains georeferenced point data. The original content has not been modified in any way.

## Source

Data taken from Kaggle: https://www.kaggle.com/datasets/camnugent/california-housing-prices

## `lightsf` *lightsf: Collection of georeferenced and spatial datasets from different domains*

### Description

Provides a diverse collection of georeferenced and spatial datasets from different domains including urban studies, housing markets, environmental monitoring, transportation, and socio-economic indicators. The package consolidates datasets from multiple open sources such as Kaggle, chopin, spData, adespatial, and bivariateLeaflet. It is designed for researchers, analysts, and educators interested in spatial analysis, geostatistics, and geographic data visualization. The datasets include point patterns, polygons, socio-economic data frames, and network-like structures, allowing flexible exploration of geospatial phenomena.

### Details

lightsf - Collection of georeferenced and spatial datasets from different domains.

Collection of georeferenced and spatial datasets from different domains.

### Author(s)

**Maintainer**: Ingrid Romero Pinilla <`ingridpinilla11@gmail.com`>

### See Also

Useful links:

- `https://github.com/roming20/lightsf`

## `mastigouche_poly` *Mastigouche Lake Network Data Set*

### Description

This dataset, mastigouche_poly, is a list containing spatial and network information for 42 lakes in the Mastigouche region. The dataset includes the XY geographical coordinates of the lakes and a site-by-edge matrix describing how the lakes influence each other. The network is defined by 66 directional edges of influence between the lakes.

### Usage

```
data(mastigouche_poly)
```

### Format

A list with 2 elements:

**xy** A data frame with 42 observations and 2 variables: X (numeric), Y (numeric) coordinates of the lakes

**siteEdge** An integer site-by-edge matrix describing 66 edges of influence among lakes

### Details

The dataset name has been kept as 'mastigouche_poly' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the lightsf package and assists users in identifying its specific characteristics. The original content has not been modified in any way.

### Source

Data taken from the **adespatial** package version 0.3-28

## nc_points *Mildly Clustered Points in North Carolina, United States*

### Description

This dataset, 'nc_points', is a data frame containing a set of spatial point coordinates representing mildly clustered points in North Carolina, United States. The dataset consists of 2,304 observations and 2 variables, corresponding to the X and Y coordinates of the points. The data can be used for examples of point pattern analysis, clustering, or spatial statistics.

### Usage

```
data(nc_points)
```

### Format

A data frame with 2,304 observations and 2 variables:

**X** X coordinate (numeric)

**Y** Y coordinate (numeric)

### Details

The dataset name has been kept as 'nc_points' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the 'lightsf' package and assists users in identifying its specific characteristics. The suffix does not include '_df' because the dataset primarily represents a spatial point pattern rather than general tabular survey data. The original content has not been modified in any way.

### Source

Data taken from the **chopin** package version 0.9.4

---

## `worldbank_poly` *World Bank Socioeconomic Indicators by Country*

---

### Description

This dataset, worldbank_poly, is a data frame containing selected socioeconomic indicators compiled from the World Bank. The dataset includes 177 observations (countries) and 7 variables such as Human Development Index (HDI), urban population percentage, unemployment rate, population growth, and literacy rate. Some values may be missing.

### Usage

```
data(worldbank_poly)
```

### Format

A data frame (tibble) with 177 observations and 7 variables:

**name** Country name (character)

**iso_a2** ISO 2-letter country code (character)

**HDI** Human Development Index (numeric)

**urban_pop** Urban population percentage (numeric)

**unemployment** Unemployment rate (numeric)

**pop_growth** Population growth rate (numeric)

**literacy** Literacy rate (numeric)

### Details

The dataset name has been kept as 'worldbank_poly' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the lightsf package and assists users in identifying its specific characteristics. The original content has not been modified in any way.

### Source

Data taken from the **spData** package version 2.3.4

# Index