Group Assignment - Data Descriptions

International DAAD Summer School on Geospatial Data Science
Jena, August 2019

1. Home Sales in Seattle

Description: The data for these sales comes from the official public records of home sales in the Seattle area, Washington State, U.S. Each of the 2000 rows represents a home sold from May through December 2014.

Objective: To predict home price in dollars per square foot as the response variable based on characteristics of the home (and its neighborhood) as predictors.

Number of cases: 6000

Variable names:

Lat: Latitude in degrees (positive = northern hemisphere) – not to be used as a predictor.

Long: Longitude in degrees (negative = west of Greenwich) – not to be used as a predictor.

Price_per_sqft: Selling price in dollars per square foot of each home sold – this is the response variable

Waterfront: Indicator variable for whether the home was overlooking the waterfront (=1) or not (=0)

Lot_to_living: Ratio of square footage of the land space to the square footage of the interior living space (i.e. sqft_living)

Basement_percent: Fraction (0-1) of the interior living space (sqft_living) that is below ground level

Bedrooms: Number of bedrooms

Bathrooms: Number of bathrooms, where .5 accounts for a room with a toilet but no shower

Floors - Number of floors

View: An index from 0 to 4 of how good the view of the property was

Condition: An index from 1 to 5 on the condition of the apartment

Grade: An index from 1 to 13, where 1-3 falls short of building construction and design, 7 has an average level of construction and design, and 11-13 have a high quality level of construction and design.

Yr_built: The year the house was initially built

Yr_renovated: The year of the house’s last renovation

Sqft_living: Square footage of the apartment’s interior living space

Sqft_lot: Square footage of the land space

Sqft_above: The square footage of the interior housing space that is above ground level

Sqft_basement: The square footage of the interior housing space that is below ground level

Sqft_living15: The square footage of interior housing living space for the nearest 15 neighbors

Sqft_lot15: The square footage of the land lots of the nearest 15 neighbors

2. Aconcagua fruit-tree remote sensing

Description: This data set contains inventoried fruit-tree crop types and Landsat satellite data from the Aconcagua river basin in central Chile. There are between 4 and 40 pixels within each of the 400 inventoried fields, and Landsat data is available for 9 image dates from one growing season. This is a subset of the data used by Peña, Liao & Brenning

(2017) in *ISPRS Journal of Photogrammetry and Remote Sensing*, who present additional background information and details on the data set.

Objective: To predict crop type as the response variable based on the available optical remote sensing data.

Number of cases: 300 fields (100 per crop type) with 3210 grid cells in total

Variable names:

croptype: Crop type, the response variable – crop1 = table grape; crop2 and crop3: walnut and peach (not sure which one is which)

field: Field identifier – not to be used as a predictor variable

utmx, utmy: UTM x/y coordinates – not to be used as predictors

b*ij*: value of Landsat band j in image i; images are numbered from early season = 1 to late season = 9; see Peña et al. (2017) for details, e.g. image dates

ndvi0*i*: NDVI value from image date *i*

ndwi0*i*: NDWI value from image date *j*

File contents:

The .Rdata file contains a data.frame `d` with the data set, a formula object called `formula`, and a character vector `predictors` representing the names of the predictor variables.

3. Rock glaciers in the Andes of Santiago, Chile

Description: This data set represent information on the presence/absence of flow structures related to the deformation of rock glaciers in the Andes of Santiago, and corresponding remotely-sensed texture attributes and terrain attributes as predictors. Texture attributes derived from high-resolution panchromatic IKONOS imagery is the main feature set in this study, and terrain attributes are used as additional predictors. A ‘filter bank’ of Gabor filters is used since Gabor features are capable of detecting ‘zebra stripe’ type patterns that relate to the troughs and ridges typically found on ‘ice-debris landforms,’ i.e. rock glaciers and debris-covered glaciers. This data set is a subset of the data used by Brenning, Long & Fieguth (2012) in *Remote Sensing of Environment*, specifically a subset of the Laguna Negra area. Note that areas that can “obviously” not present rock glaciers have been masked out (i.e. removed from the data set), e.g. steep slopes, in order to allow the classifier to focus on the “difficult” areas; see Brenning et al. (2012) for details.

Objective: To identify rock-glacier flow patterns based on the available texture and terrain attribute data.

Number of cases: 3403 grid cells (617 from flow patterns and 2786 from other terrain outside of rock glaciers). (The flow patterns occur within approximately 50 individual rock glaciers of different size.)

Variable names:

class: Factor variable (levels: “TRUE”, “FALSE”) representing the presence (“TRUE”) and absence (“FALSE”) of rock glacier flow patterns

dem: Elevation in metres above sea level (m a.s.l.)

slope: (Local) slope angle in degrees

cslope: Slope angle of the upslope contributing area in degrees

log.carea: Logarithm (to the base 10) of the size upslope contributing area in m²

log.cheight: Logarithm (to the base 10) of the height of the upslope contributing area in m

pisr: Annual potential incoming solar radiation

m30e*i*g*jx*: Gabor feature with the following settings (see Brenning et al. 2012 for details): i = axis ratio (1 or 2) of Gabor filter; j = wavelength of Gabor filter (5, 10, 20, 30, or 50 m); x = aggregation scheme (“min” = minimum; “max” = maximum; “rg” = range; “med” = median

File contents:

The .Rdata file contains a data.frame `d` with the data set, a formula object called `formula`, and a character vector `predictors` representing the names of the predictor variables