Impact of commercial venues, business activity and population on solid waste generation in the city of Lima, Peru

Glen Rodríguez
October 26th, 2019

2. Data acquisition and cleaning

2.1. Data sources

We will use 3 main types of data sources. The first one is FourSquare API, from which we will gather the number of venues around the center of each district. Later on, we will use that data and the area of each district to estimate the number of commercial venues per district.

The second type of data source is a set of reports from INEI, the Statistics Institute of the Peruvian government. These reports are in PDF format, and they contain the solid waste generated per district on 2015, the population of each district on 2015, the longitude and latitude of the administrative center of each district (accesible at the site of INEI: https://www.inei.gob.pe/media/MenuRecursivo/publicaciones_digitales/Est/Lib1477/libro.pdf), the number of business registered per district on 2015 (https://www.inei.gob.pe/media/MenuRecursivo/publicaciones_digitales/Est/Lib1382/cap06.pdf) and the residences per district on 2015 (on two contigous pages: https://www.inei.gob.pe/media/MenuRecursivo/publicaciones_digitales/Est/Lib1483/cap04/cap04.pdf). The PDF scraping will be done using the tabula package.
The last type of data source is Wikipedia. From the Wikipedia page about distrits of Lima: https://es.wikipedia.org/wiki/Anexo:Distritos_de_Lima, we will scrap the area of each district. The web scraping will be done using the BeautifulSoup package.

2.2. Data cleaning

We must correct some spelling mistakes into the district's names in some data sources, to merge them into a single dataframe. We will keep only the columns containing information for the year 2015, so will be drop everything else.

The main data cleaning task is the conversion of text into numbers. Some columns have a space separating thousand groups. We must delete the spaces and change the text into integers. The columns of latitude and longitude are in the format dd°mm'ss" (degrees, minutes and seconds); we will convert them into floats (dd + mm/60 + ss/3600).

We got all the information for all districts but one, Santa María del Mar. It lacks information about housing, so we used the population and housing information of its two neighbors districts (Pucusana and San Bartolo) to estimate its housing number.

Regarding the venues, some are commercial properties, but others are public facilities (that do not sell anything, do not generate waste directly and are not taxable anyway) and therefore should be excluded from our analysis. Categories that should not enter in the analysis are: Plaza, Beach, Scenic Lookout, Bridge, Park, Garden, Garden Center, Skate Park, Bus Line, Bus Station, Pool, Soccer Field, Water Park, Trail, City, Tourist Information Center, Mountain, EV Charging Station, Light Rail Station.

The number of venues per district can not be calculated directly. We count the number of venues around the administrative center of each district, with a radius of 600 m and area pi*0.6*0.6=1.13 km2. Then we extrapolate the number of venues multiplying the number of venues in the radius times the area of the district divide by 1.13 and use the result as an estimate.

2.3. Feature selection

After cleaning, we got 43 districts whit one dependent variable (waste) and 4 candidates for independent variable. Two are indicators of domestic generation of waste (population and housing units), another is related to commercial properties and the last one relates to business in general. We could not get any info on the number of industries, so we will use the number of businesses registered in the district as a proxy.

We calculate the matrix of covariance:

	waste2015	pop2015	houses2015	Venue	business2015
waste2015	1.000000	0.862409	0.879592	0.060521	0.897601
pop2015	0.862409	1.000000	0.988754	-0.004135	0.729946
houses2015	0.879592	0.988754	1.000000	-0.002199	0.749003
Venue	0.060521	-0.004135	-0.002199	1.000000	-0.015886
business2015	0.897601	0.729946	0.749003	-0.015886	1.000000

We see that all 4 variables have a positive correlation with solid waste, but population and housing units are also closely correlated. Therefore we will drop one of them, population, because it is redundant and housing units are directly related to the taxation structure (taxes are paid per property, not per person).