

Write-up

Ernesto Truqui

3/16/2022

The central question the following project set out to answer was, is there a mismatch between police patrol allocation and crime in the city of Medellin, COL? Given a rich dataset, including geocoded and timestamped crime data, as well as demographic and socioeconomic indicators at the block and quadrant level, we were well-positioned to attempt to answer this question. It is an important issue to consider since police are a scarce resource and one of the only tools available to the state to combat and reduce crime. Therefore, it is of the utmost importance for this resource, police patrols, to be distributed efficiently across the city in order to prevent and combat crime.

We will discuss our approach in the following order. First, we will lay out what the raw data we received looked like and outline the steps we took to “clean” it to make it usable for our purposes. Then, we will describe how we used the clean data to visualize crime and police patrol distribution across Medellin. Next, we will provide details regarding our text processing And finally, we will fit a model to try to predict

1. Data wrangling

The raw data we received was directly from the National Police of Colombia. Rather than have everything well-organized within one dataset, they sent us several datasets, with overlapping information, but each containing a useful set of variables. One of them contained geocoded crime data at the street block level. Another contained geocoded crime data at the quadrant level (akin to police beats in the US). Another dataset contained the labels for the strangely coded variable names, to facilitate interpretation. We proceeded to clean each of these datasets and prepare them for merging, so that we could have all the information in a single place.

The cleaning process involved several steps, which we will briefly glance over. Having chosen the quadrant as the unit of analysis of interest, we created a variable in the block level data that identified which quadrant each street block belonged to given its geocoded location. We also averaged 2018-2021 data to get the yearly mean and facilitate plotting. Furthermore, using crimes’ timestamps, we identified during which police shifts crimes were committed, to analyze the mismatch between police distribution and crimes across all three daily police shifts. Finally, having made the data tidy, using the crime variable and knowing that officers were deployed uniformly across quadrants, we calculated a crimes per officer variable that would be central to measuring the success of our re-distribution proposal.

To wrap up the data wrangling, we put forth a police patrol re-distribution recommendation that would assign officers to quadrants proportionally to the crime levels, rather than uniformly so. The purpose of this was to visualize in the next step whether crime would be better combatted if there was a better strategy for police patrol distribution than the status quo.

2. Plotting

With clean data, the plotting was merely about reproducing reality in maps of Medellin. We started by showing how crime and police patrols were currently distributed across the city