Ethan Arsht

24 February 2023

Research Strategy

I. Research Question

I will aim to answer the following research question: to what extent do CTA "ghost-buses" follow a geographically clustered pattern?[1]

II. Hypotheses:

In the context of this research question, I will consider two hypotheses. The first hypothesis holds that the clustering of ghost-buses is not randomly distributed, and therefore there are spatial clusters in which ghost-buses are more common. This will be tested in two contexts: first, the overall number of scheduled buses that failed to arrive, and second, the overall reliability of buses at any given stop. Rather than examining the ghost-buses associated with bus routes, each individual bus stop will be considered as a separate unit. By considering individual stops rather than routes, I can estimate the overall reliability likely to be experienced by users of CTA buses within a given geography.

The second hypothesis is contingent on the findings of the first hypothesis. It holds that the spatial clustering of ghost-buses, or bus unreliability, has a relationship with demographic, social, and economic indicators within Chicago. As far as specific variables, at a minimum I will examine the relationship between bus reliability/unreliability and race, crime, and income. To measure this data, I will aggregate bus reliability/unreliability to a neighborhood level, where it can be joined with the city of Chicago's open data.

III. Background

The primary data source for this project is ChiHackNight's Ghost Buses project. In addition to the data in their Github repository, they have provided me additional, more granular data at my request. This granularity of data will allow me to analyze reliability at both the route and stop data, although significant data cleaning will be necessary to format the data in terms of stops.

The Ghost Buses data includes the number of arrivals for each bus compared to its scheduled arrivals for each day since May 20th, 2022. The difference between scheduled arrivals and actual arrivals represents the number of ghost buses. CTA operates approximately 125 routes on weekdays and 85 routes on weekends and holidays. The overall dataset has 17370 observations (each observation represents one route's reliability on one day). After disaggregating this data to stops, the number of observations will likely increase.

IV. Methods

The first hypothesis, considering the spatial pattern of bus reliability, is well-suited to DBscan and HDBscan. In particular, HDBSCAN should effectively identify relative clusters of ghost-buses while reconciling the different baseline density of bus routes within Chicago. For example, The Loop and the surrounding areas will likely have a large number of ghost buses because of the large

[1] In the previous research question, I also included L service. I have now decided to focus exclusively on buses.

amount of overall bus traffic, while areas in the periphery and/or South Chicago have fewer bus routes.

To test the second hypothesis, machine learning methods such as Kmeans and/or Kmedioids may be most appropriate, as it can effectively cluster across multiple dimensions which will easily allow for the inclusion of non-spatial data in the analysis. While not the focus of the analysis, non-spatial methods such as linear regression may also be useful in verifying the hypothesis.