Vote-Time Export Codebook

Center for New Data
First Draft September 7 2021
Updated September 14 2021

Table of Contents

Introduction

The associated data were produced by way of the *Center for New Data*'s replication of the *Chen et al* methodology for using geolocation data to estimate arrival times and voting durations at polling locations on election day for the 2020 General Election.[1]
Notes on Data Quality and Data Cleaning

- Devices with UPPER_MINUTES_WAITING > 180 minutes have been removed.
- Polling locations (and associated devices) with < 25 devices have been omitted.
- There are devices in the dataset that were detected at the same polling location multiple times in the same day. The *Chen et al* method that was replicated for this work did not establish a filter for this type of aberrant data, and our replication of the data provided here do not offer a correction. Therefore, for these devices, correct interpretation is not currently clear. We advise multiple ways to deal with these data, including potentially:
    - Collapsing repeated visits into a single visit if they are under *X* minutes together (on the logic that perhaps they represent a single visit that was interrupted by an aberrant ping or itinerant movement activity in and outside the detected threshold); and/or
    - Excluding devices with more than one visit to the same polling location in the same day.

[1] The Chen et al paper was originally published at NBER, and was later covered in *Scientific American* and published as "Racial Disparities in Voting Wait Times: Evidence from Smartphone Data," Review of Economics and Statistics (Dec. 11, 2020): 1–27.

Codebook: Columns in the data

Device Characteristics

- **DEVICE_ID_HASH**. A hashed version of deviceid, aka advertiser id, which distinguishes each device in our dataset.

Polling Location Characteristics

- **SPATIALLY_DISTINCT_GEOHASH_KEY**. The seven-digit geohash corresponding to the latitude/longitude coordinates associated with the polling location. (A 7-digit geohash represents an area of 23,000 square meters, or about 150 meters by 150 meters.) This is very close to a unique id for polling locations, and can be used to identify incidences of repeat observations of polling locations due to slight variations in spelling or address. However in other cases there will be two or more polling locations with the same SPATIALLY_DISTINCT_GEOHASH_KEY due to the polling locations being very close in proximity such that they fall into the same 150m x 150m grid.
- **CND_POLL_UUID**. This is a unique identifier assigned to each polling location.
- **PRECINCT_ID**. Precinct of the polling location.
- **PRECINCT_NAME**. Name of the polling location.
- **POLLING_LOCATION_ADDRESS.** Address of the polling location
- **POLLING_LOCATION_NAME**. The name of the location provided associated with the address.
- **POLLING_LOCATION_COUNTY_FIPS**. The county of the polling location.
- **POLLING_LOCATION_CENSUS_TRACT**. Census tract associated with the polling location
- **POLLING_LOCATION_SOURCE**. The reporting source of the polling location data, either: cpi (center for public integrity), google_api (Google Civics API), DemocracyWorks, or correction.

Estimated Time at Polls

- **LOWER_MINUTES_WAITING**. A lower-bound estimate of time on-site, calculated as time of the first ping within the threshold radius around the polling location (i.e., **LATEST_TIME_ARRIVED_POLLS_LOCAL_TIME**) to the last ping detected within the radius i.e., **EARLIEST_TIME_LEFT_POLLS_LOCAL_TIME**).
- **UPPER_MINUTES_WAITING**. An upper-bound estimate of time on-site, calculated as time of the last ping *before* the first ping within the threshold radius around the polling location (i.e., **EARLIEST_TIME_ARRIVED_POLLS_LOCAL_TIME**) to the first ping *after* the last last ping detected within the radius (i.e., **LATEST_TIME_LEFT_POLLS_LOCAL_TIME**).
- **EXPECTED_MINUTES_WAITING**. The average of LOWER_MINUTES_WAITING and UPPER_MINUTES_WAITING.

- **RADIUS_FROM_POLL_USED_FOR_CALCULATION**. The radius around the centroid of the polling location used to detect potential voters. The value used for the primary analysis in *Chen et al* was 60 meters.
- **HAS_PING_IN_BUILDING**. An indicator for any of the pings detected from a device while visiting a polling location were detected within the shapefile associated with the building at the address of the polling location.
- **FIRST_PING_IN_CLUSTER_LOCAL_TIME** and **LAST_PING_IN_CLUSTER_LOCAL_TIME**. Respectively, the first and last pings detected in the "stationary cluster" corresponding to the time when the device was detected at the polling location. This does not directly bear on the *Chen et al* methodology but can be helpful for additional validation.
- **LAST_PING_IN_PREVIOUS_CLUSTER_LOCAL_TIME**. The last ping in the stationary cluster detected immediately before the cluster associated with the interval of time when the device was detected at the polling location.
- **FIRST_PING_IN_NEXT_CLUSTER_LOCAL_TIME**. The first ping in the stationary cluster detected immediately after the cluster associated with the interval of time when the device was detected at the polling location.