**P18 GPS activity space and Twitter Data Analysis Plan**

1. **Datastets:**

**General Description of GPS Data:**

* 250 participants over 2 weeks with GPS coordinates every 10 seconds (protocol but actual data varies between individuals)
* Max size = 30,240,000 data points
* Missing/erroneous data via participant attrition, subway use (tracker does not function underground), tracker malfunction/interference
* Participants in two waves (different times of year)
* Cross streets of home
* Longitudinal survey data on behaviors, etc.
* Total N might vary from ~250-211 depending on limiting data

1. **General Description of Twitter Data:**

* (NYC-RacismNHomophobia\_till013118.csv) updated using the data till 01/31/2018. The details of the data including the description of each column can be found in DataDictionary.txt file

1. **Plan for Analysis:**

We are trying to find out the relationship between discrimination and sexual risk behavior. The plan are listed below:

a. Perform areal weighted mean of Twitter grid cell variables (“SSSOM\_Rac\_grid”, “SSSOM\_Hom\_grid” and “Rac\_tweets\_grid”) within GPS activity space variables as an indicator of individual exposure to neighborhoods and corresponding risk a single-number summary of risk

“areal weighted mean” will take a mean sum over all grid cells, *c*:

b. This will give us an average unit for further analysis of the AWM variables (“AWM\_SSSOM\_Rac”, “AWM\_Rac\_tweets”, “AWM\_SSSOM\_Hom” “AWM\_Zip\_Hom”, and “AWM\_Zip\_Rac))

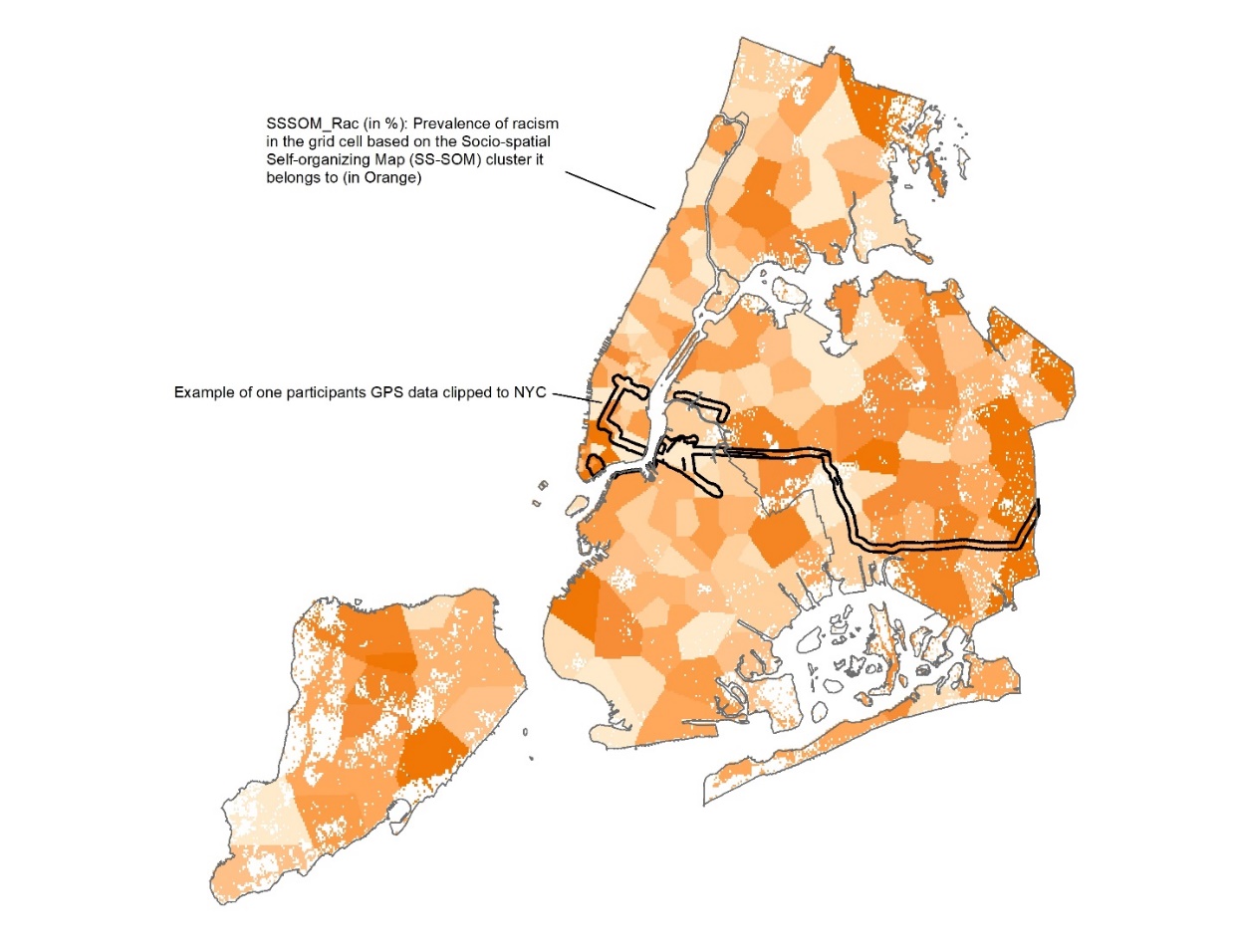
c. Right now we are not controlling for time, rather this is a cross sectional design.

d. The idea is to include the AWM variable (“AWM\_SSSOM\_RAC”, etc.) as a covariate in a regression model (e.g. quasi-poisson regression, negative binomial regression). Other covariates and outcomes for the model are described below.

**Table 1: codebook for grid cell variables.** Spreadsheet: P18\_Final\_Data\_07162019.csv

|  |  |  |
| --- | --- | --- |
| **Variables Name** | **Description** | **Name in spreadsheet** |
| SSSOM\_Rac\_grid | Prevalence of racism in the grid cell based on the SS-SOM cluster it belongs to. | SSSOM\_Rac (in %) |
| SSSOM\_Hom\_grid | Prevalence of homophobia in the grid cell based on the SS-SOM cluster it belongs to. | SSSOM\_Hom (in %) |
| Rac\_tweets\_grid | Prevalence of racist tweets in the grid cell based on the SS-SOM cluster it belongs to. | NormalRacistTweets (in %) |
| Zip\_Rac\_grid |  | Zip\_Rac (in %) |
| Zip\_Hom\_grid |  | Zip\_Hom (in %) |

Twitter data and one participants GPS activity space (200m)



1. **Descriptive Statistics of GPS Activity Derived Variables:**

**Table 2: codebook for AWM variables.** Spreadsheet: P18\_GPS\_AWM\_Twitter\_data\_summarystats.csv

|  |  |  |
| --- | --- | --- |
| **Variables Name** | **Description** | **Name in spreadsheet** |
| AWM\_SSSOM\_Rac | Areal weighted mean of SSSOM\_Rac\_grid. | AWM\_SSSOM\_Rac |
| AWM\_SSSOM\_Hom | Areal weighted mean of SSSOM\_Hom\_grid. | AWM\_SSSOM\_Hom |
| AWM\_Rac\_tweets | Areal weighted mean of Rac\_tweets\_grid. | AWM\_Norm\_R\_tweets |
| AWM\_Zip\_Rac | Areal weighted mean of Zip\_Rac\_grid. | AWM\_Zip\_Rac |
| AWM\_Zip\_Hom | Areal weighted mean of Zip\_Hom\_grid. | AWM\_Zip\_Hom |
| Area\_km | Area in square kilometers of activity spaces. | Area\_km |

Field: AWM\_SSSOM\_Rac

Count: 249

Minimum: 0.399662

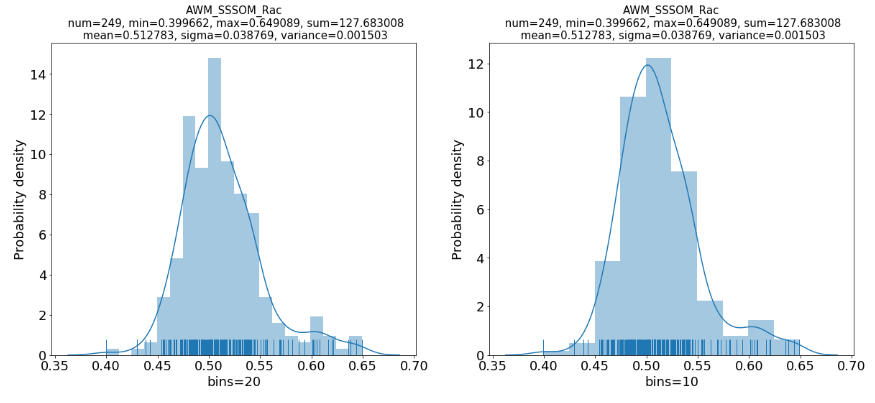
Maximum: 0.649089

Sum: 127.683008

Mean: 0.512783

Standard Deviation: 0.038769

Nulls: 0



Field: AWM\_SSSOM\_Hom

Count: 249

Minimum: 1.231941

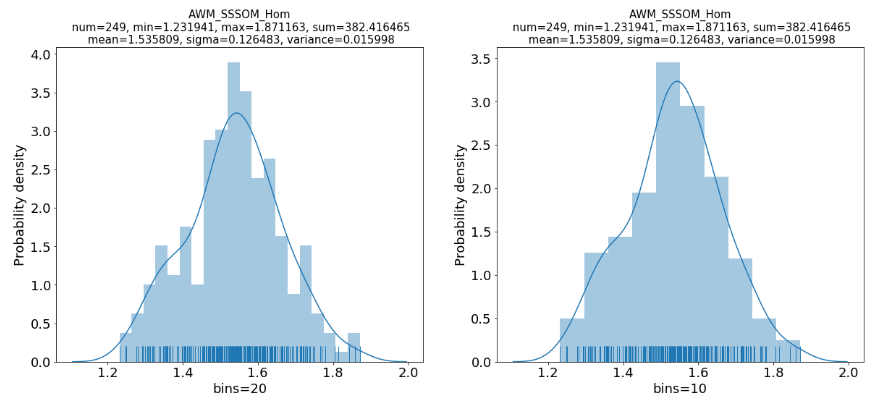
Maximum: 1.871163

Sum: 382.416465

Mean: 1.535809

Standard Deviation: 0.126483

Nulls: 0



Field: AWM\_Rac\_tweets

Count: 249

Minimum: 0

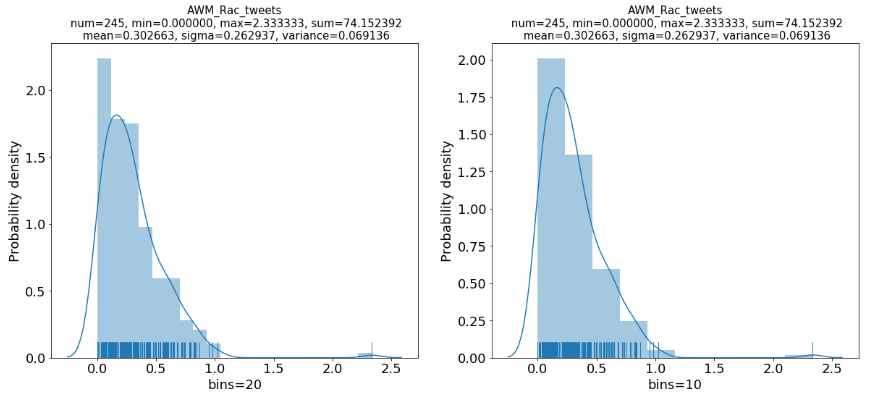
Maximum: 2.333333

Sum: 74.152392

Mean: 0.297801

Standard Deviation: 0.263578

Nulls: 0



Field: AWM\_Zip\_Rac

Count: 249

Minimum: 0.550821

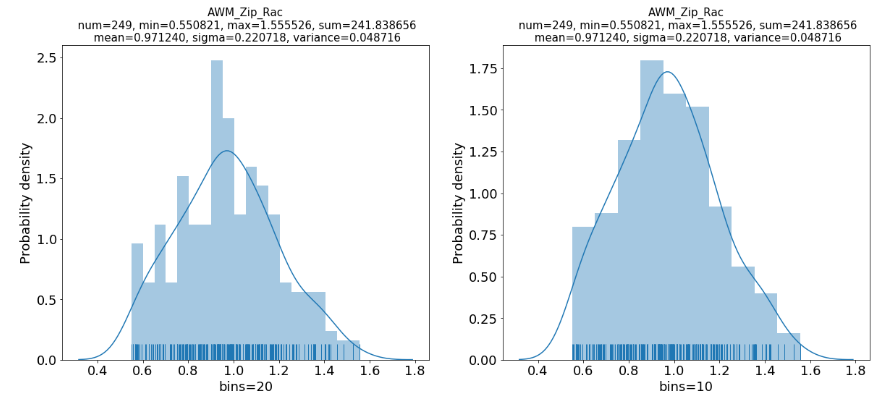
Maximum: 1.555526

Sum: 241.838656

Mean: 0.97124

Standard Deviation: 0.220718

Nulls: 0



Field: AWM\_Zip\_Hom

Count: 249

Minimum: 1.306398

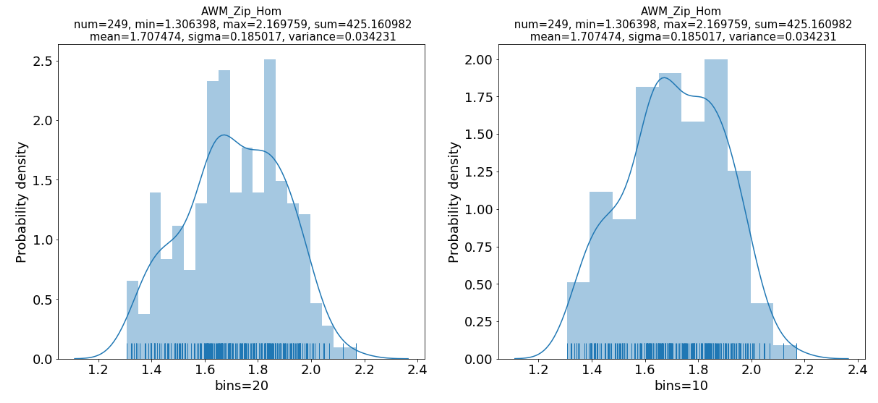
Maximum: 2.169759

Sum: 425.160982

Mean: 1.707474

Standard Deviation: 0.185017

Nulls: 0



Field: Area\_km

Count: 249

Minimum: 0.426594

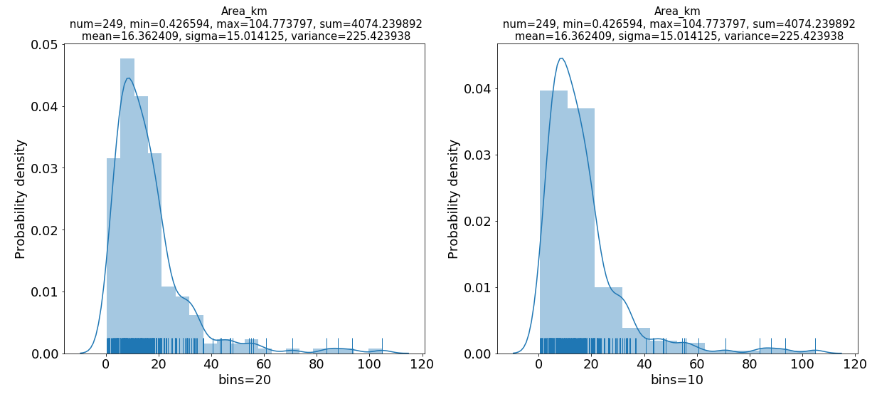
Maximum: 104.773797

Sum: 4074.239892

Mean: 16.362409

Standard Deviation: 15.014125

Nulls: 0



1. **Covariates:**

*Potential list of covariates pulled from our recent manuscript Associations of Spatial Mobility with Sexual Risk Behaviors among Young Men Who Have Sex with Men in New York City: The Project 18 Neighborhood Study. I think we still need to discuss what statistical model we want to run. I am less familiar with the survey level data but the covariates from the manuscript are listed below.*

-age (years), ethnicity (Hispanic or non-Hispanic), race (Black, Asian, White, and others), education attainment (high school or less, some college/ technical school, college degree or more), current school enrollment status (yes/no), and foreign-born status (yes/no), Homelessness, total individual annual income was categorized as less than $15,000, between $15,000 and $35,000, more than $35,000 per year, which approximates national poverty level (<$15,000)

May I have the codebook for the variable names in file “P18\_Final\_Data\_07162019.csv”? I may need to know which variable is age, ethnicity, etc.

1. **Outcomes:**

Potential outcomes: sexual behaviors associated with risk of HIV infection among MSM. The Project 18 Cohort Study collected data on sexual behaviors including number of male sexual partners and number of condomless sexual encounters in past six months. The number of male sexual partners was assessed from two questions: (1) “In the past 6 months, how many male steady partners have you have anal or oral sex with?,” (2) “In the past 6 months, how many casual male partners have you had anal or oral sex with casual or non-steady partners?”.The total number from those two questions was used as one outcome and was considered a count type variable in the analyses [35-39]. In addition, numbers and types of sexual encounters were assessed to create three variables: total numbers of condomless anal intercourse acts, (2) total numbers of condomless insertive anal intercourse acts, and (3) total number of condomless receptive anal intercourse acts in past six months [35-40].

To be more specific, the outcome variables can be:

Anal Intercourse variables: "AI\_Total", "AI\_Condom", "AI\_Condomless"

Insertive Anal Intercourse variables: "IAI\_Total", "IAI\_Condom", "IAI\_Condomless"

Receptive Anal Intercourse variables: "RAI\_Total", "RAI\_Condom", "RAI\_Condomless"

Others: "drug\_use"