**General features**

* **Demographics**
  + In EMA: age, sex (dummy-coded male versus female), race/ethnicity (dummy-coded W/N-H versus other)
  + SES, baseline severity?
* **Weather**
  + *(Heller et al., 2020, p. 20): To extract weather information, weather data was collected using the API at http:// weatherunderground.com. API requests for mean temperature and total precipitation data were made using participants’ modal longitude/latitude for each day.*
* **Sociodemographic feature space – measures of psychosocial stress?**
  + *CSI (Kwan et al., 2019): Community SES was operationalized in this study as a composite socioeconomic index (CSI) based on the modified Darden–Kamel composite index developed by Darden et al. (2010). It uses nine variables from census data and assigns a higher score to communities with higher SES. The nine variables are the percentage of residents with university degrees, median household income, the percentage of managerial and professional positions, median value of dwelling, median gross rent of dwelling, the percentage of homeownership, the percentage of households with vehicle, the percentage of population below poverty (reverse scored) and unemployment rate (reverse scored). Kwan 2019 also used measures of crime.*
  + *(Heller et al., 2020, p. 20): The first step in this analysis was to map each GPS coordinate that appeared in our dataset to its corresponding Federal Information Processing Standard (FIPS) code. A FIPS code uniquely identifies locations at the spatial resolution of a “Block Group”, the smallest geographical unit for which the U.S. Census Bureau publishes data. The areas encompassed by Block Groups differ in size as they are determined by the approximate population size of roughly 600 to 3,000 people each. Block Groups are used by US government agencies and other geographic information systems (GIS) to link geographic and demographic information to specific physical locations. We downloaded 53 variables capturing demographic and socioeconomic features of a given block group (e.g., population density, median age, total businesses, educational attainment, unemployment population, race, gender - see Supplementary Appendix 1 for a complete list of data sources) from the U.S. Census Bureau.*
  + <https://data.census.gov/>
  + <https://www.census.gov/programs-surveys/geography/guidance/geo-identifiers.html>
* **Time of day (5pm – midnight versus any other time)**
* **Day of week**
  + Weekends versus weekdays?

**GPS features**

**Context-specific**

Perhaps all duration features would be better represented as percentages to better compare across individuals?

* **Type of place**
  + Do we need to reduce the dimensionality of these responses?
  + *Muller 2020: To analyse places visited at the daily level and at the two–week aggregate level, we referred to Oldenburg's (Oldenburg & Brissett,*[*1982*](https://journals.sagepub.com/doi/full/10.1002/per.2262#bibr43-per-2262)*) classification system of places and combined the place types into three overall categories with home as a stand–alone category, social places (bar/party, café /restaurant, fraternity/sorority house, and friend's house), and work–related places (campus, library, and work).*
  + Base question response, frequency (number of times in past X amount of period this type of place has been visited), duration (total amount of time spent at this type of place in past X amount of time)
* **Drank (Have you drank alcohol here before?)**
  + Dummy-coded yes/no
  + Base question response, frequency (number of times in past X amount of period participant has visited locations where previously drank), duration (total amount of time spent at places where has previously drank in past X amount of time)
* **Alcohol (Is alcohol available here?)**
  + Dummy-coded yes/no
  + Base question response, frequency (number of times in past X amount of period participant has visited locations where alcohol is available), duration (total amount of time spent at places where alcohol is available in past X amount of time)
* **Emotion (How would you describe your experiences here?)**
  + Pleasant, unpleasant, mixed, neutral
  + Base question response, frequency (number of times in past X amount of period participant has visited differently valenced locations – maybe a ratio?), duration (total amount of time spent at differently valenced locations – maybe also as a ratio?)
    - Ratio 🡪 time spent in pleasant locations relative to other valenced locations?
* **Risk (Does being at this location put you at any risk to begin drinking?)**
  + No risk, low risk, medium risk, high risk
  + Base question response, frequency (number of times in past X amount of period participant has visited different risk level locations – ratio?), duration (total amount of time spent at different risk level locations – maybe also a ratio?)
    - Ratio 🡪 time spent at high risk locations relative to other risk levels?
* **Avoid (Did the participant identify this place as a risky location they are trying to avoid now that they are sober?)**
  + Dummy-coded yes/no
* **Number of frequently visited locations defined in the interview**

**Context-agnostic**

*What should the window be for these?*

**Theoretically easy to calculate**

* **Home stay**
  + Percentage of time spent at home relative to other locations
  + (Saeb, Zhang, Karr, et al., 2015): 75.9% accuracy / 80.5% sensitivity / 71.7% specificity
  + (Saeb et al., 2016): significant correlation with PHQ-9 scores at end of study
  + (Saeb, Zhang, Kwasny, et al., 2015): significant correlation with PHQ-9 scores
  + (Palmius et al., 2016)
* **Location variance**
  + Logarithm of the sum of statistical variances of latitude and longitude per subject using only location data of stationary states
    - Log compensates for skewness in distribution of variance across participants
  + (Saeb, Zhang, Karr, et al., 2015): 75.7% accuracy / 80.2% sensitivity / 71.5% specificity; greater depressive symptoms = lower variance
  + (Saeb et al., 2016): significant correlation with PHQ-9 at end of study
  + (Saeb, Zhang, Kwasny, et al., 2015): significant correlation with PHQ-9 scores
  + (Palmius et al., 2016)
* **Number of clusters**
  + *K*-means clustering on data classified as stationary
    - Start with one cluster, increase number of clusters until distance of farthest point falls below predetermined threshold (Saeb uses 500m, Palmius uses 400m)
  + Maybe instead of deriving these clusters we can do number of frequently visited locations that we have from interviews?
  + (Saeb, Zhang, Karr, et al., 2015): 41.5% accuracy / 47.4% sensitivity / 35.5% specificity
  + (Saeb et al., 2016): significant correlation with PHQ-9 scores at beginning and at end of study
  + (Saeb, Zhang, Kwasny, et al., 2015)
  + (Gruenerbl et al., 2014)
  + (Palmius et al., 2016)
* **Total distance**
  + (Saeb et al., 2016; Saeb, Zhang, Karr, et al., 2015; Saeb, Zhang, Kwasny, et al., 2015)
  + A mathematical equation with numbers and symbols

    Description automatically generated(Heller et al., 2020)
  + Think we can just calculate this by summing our distance values over a given period?
  + (Saeb, Zhang, Karr, et al., 2015): 56.4% accuracy / 69.6% sensitivity / 43.4% specificity
  + (Saeb et al., 2016)
  + (Saeb, Zhang, Kwasny, et al., 2015)
  + (Heller et al., 2020) 🡪 included as a covariate
  + (Gruenerbl et al., 2014)
  + (Palmius et al., 2016)
* **Transition time**
  + Amount of time spent in non-stationary state
  + Number of GPS location samples in transition states / total number of samples
  + (Saeb, Zhang, Karr, et al., 2015): 41.1% accuracy / 43.4% sensitivity / 38.7% specificity
  + (Saeb et al., 2016)
  + (Saeb, Zhang, Kwasny, et al., 2015)
  + (Palmius et al., 2016)

**Difficult but want to try**

* **Circadian movement**
  + To what extent do participants’ sequence of locations follow a 24-hour (circadian) rhythm?
    - This number will be high if someone leaves for work and returns home from work at the same time every day, for example
  + Steps:
    - First, obtain GPS location data spectrum (least-squares spectral analysis/Lomb Scargle Periodogram)
    - Then calculate amount of energy that falls into frequency bins within a 24+/-0.5 hour period:
    - This is calculated separately for latitude and longitude, then summed and log-transformed:
  + (Saeb, Zhang, Karr, et al., 2015): 78.6% accuracy / 80.1% sensitivity / 77.5% specificity; more disrupted (lower) values in individuals with high PHQ-9 scores
  + (Saeb et al., 2016): significant correlation with PHQ-9 scores at end of study
  + (Saeb, Zhang, Kwasny, et al., 2015)
  + (Palmius et al., 2016) 🡪 referred to as “diurnal movement”
    - Diurnal movement on normalized coordinates: calculated the same way but latitude and longitude are scaled to have zero mean and unit variance within calculation period
    - Diurnal movement on the distance from home: calculated using Euclidean distance from home instead of latitude and longitude, normalized to have zero mean and unit variance within calculation period
* **Entropy**
  + Variability of the time participant spent at location clusters
    - High entropy: time distributed uniformly across locations
    - Low entropy: time unevenly spent across locations
  + (Saeb, Zhang, Karr, et al., 2015): 69.7% accuracy / 66.8% sensitivity / 72.7% specificity
  + (Saeb et al., 2016): significant correlation with PHQ-9 scores at end of study
  + (Saeb, Zhang, Kwasny, et al., 2015): weak correlation with PHQ-9 scores
  + (Palmius et al., 2016)
* **Roaming (normalized) entropy**
  + (Saeb et al., 2016; Saeb, Zhang, Karr, et al., 2015; Saeb, Zhang, Kwasny, et al., 2015)
  + A close-up of a black text

    Description automatically generated (Heller et al., 2020)
  + Normalizes entropy by number of location clusters (therefore invariant to number of clusters)
    - Ranges from 0 to 1
      * 0: all location data points belong to same location
      * 1: location data points are uniformly distributed
  + This calculation used in Freund et al. 2013 and Sargosa-Harris et al., 2022, Heller et al., 2020 uses a variation of this
  + (Saeb, Zhang, Karr, et al., 2015): 86.5% accuracy / 88.4% sensitivity / 84.9% specificity; high negative correlation with PHQ-9 scores
  + (Saeb et al., 2016): significant correlation with PHQ-9 scores at end of study
  + (Saeb, Zhang, Kwasny, et al., 2015): correlation with PHQ-9 scores
  + (Heller et al., 2020): greater positive affect associated with greater roaming entropy; negative affect not related to roaming entropy
  + (Palmius et al., 2016)
* **Proximity to risky locations (risk terrain mapping)**
  + Not for now, for future consideration w/ Jamie and Kendra

**Not interested in:**

* **Number of hours outdoors**
  + (Gruenerbl et al., 2014)
* **Average time outdoors per hour**
  + (Gruenerbl et al., 2014)
* **Times of day spent outdoors**
  + (Gruenerbl et al., 2014)
* **Variance of times spent outdoors**
  + (Gruenerbl et al., 2014)
* **Number of stays outdoors**
  + (Gruenerbl et al., 2014)
* **Percentage of time outside in 24 hours**
  + (Gruenerbl et al., 2014)
* **Number of cell towers “pinged”**
  + (*Behavioral Activities Collected through Smartphones and the Association with Illness Activity in Bipolar Disorder - Faurholt‐Jepsen - 2016 - International Journal of Methods in Psychiatric Research - Wiley Online Library*, n.d.; Faurholt-Jepsen et al., 2014)
* **Number of novel locations visited**
  + (Heller et al., 2020): more novel locations visited associated with greater positive affect
* **Speed mean**
  + A math equations with numbers and symbols

    Description automatically generated
  + Mean of instantaneous speed at each GPS data point (degrees per second)
  + (Saeb et al., 2016)
* **Speed variance**
  + Variance of instantaneous speed
  + (Saeb et al., 2016)
* **Raw entropy**
  + Entropy of discretized distribution of GPS coordinates
  + Uses the same formula as entropy but uses data points prior to location clustering
  + Latitude and longitude raw entropies are calculated separately and then summed together
  + (Saeb et al., 2016)

**Top/most predictive features by paper:**

(Saeb, Zhang, Karr, et al., 2015): circadian rhythm, normalized entropy, location variance

(Saeb et al., 2016): location variance, entropy, circadian rhythm, normalized entropy, home stay, number of clusters

(Saeb, Zhang, Kwasny, et al., 2015): circadian movement, location variance, normalized entropy, home stay

(Heller et al., 2020): roaming (normalized) entropy (only associated with positive affect and not negative affect)

(*Behavioral Activities Collected through Smartphones and the Association with Illness Activity in Bipolar Disorder - Faurholt‐Jepsen - 2016 - International Journal of Methods in Psychiatric Research - Wiley Online Library*, n.d.; Faurholt-Jepsen et al., 2014): cell tower ID changes (associated w manic symptoms)

(Gruenerbl et al., 2014): “fusion” metrics created, contributions of single features not examined

(Palmius et al., 2016): a five-feature model performed best, but what five features were most predictive is not listed; entropy and normalized entropy appear to have the lowest MAE in standard linear models

**Github:**

Saeb: <https://github.com/sosata/MobileDepression/tree/master>

Heller: <https://github.com/manateelab/GPS_Study/blob/master/filter_entropy_calculator.R>