**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 2022: 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 from Kwan 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 2022: 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

**GPS features**

**Context-specific**

Set relevance threshold for distance from context 🡪 500 meters? 50 meters? (dist\_context default calculated in meters: <https://cran.r-project.org/web/packages/geosphere/geosphere.pdf>) – or maybe this could be a binary dummy-coded yes/no feature (within 50 meters versus not?)

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

**Context-agnostic**

*What should the window be for these?*

* Home stay
  + Percentage of time spent at home
* Location variance
  + Logarithm of the sum of statistical variances of latitude and longitude per subject
* Total distance
  + Saeb 2016:
  + Think we can just calculate this by summing our distance values over a given period?
* Number of clusters
  + Saeb 2015a/2016: *K*-means clustering
  + Maybe instead of deriving these clusters we can do number of frequently visited locations that we have from interviews?
* Roaming (normalized) entropy
  + Normalizes entropy by number of location clusters
  + This calculation used in Freund et al. 2013 and Sargosa-Harris et al., 2022, Heller et al., 2020 uses a variation of this
* Circadian movement
  + To what extent do participants’ sequence of locations follow a 24-hour (circadian) rhythm?
  + Steps:
    - First, obtain GPS location data spectrum (least-squares spectral analysis)
    - 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:
* Raw entropy
  + Entropy of discretized distribution of GPS coordinates
* Entropy
  + Variability of the time participant spent at location clusters
  + We could calculate this as variability of time participant spent at frequently visited locations defined in the interview?
* Proximity to risky locations (risk terrain mapping)