HEAL Parks 2023 Analysis

Ronald Buie

Table of contents

1 Front Matter

This document outlines procedures, technical considerations, and analytic results for the 2023 analysis of data from the HEAL Parks Study. The primary purpose of this PDF is technical review by analyst and project managers to confirm the process and data quality.

For general information about the project please review the git or contact Seth Schromen-Wawrin.

1.1 Major inputs

Inputs to this script are contained at ./inputs/. and include

- Access to the ITHS REDCap project: "Public Health - Seattle & King County Park Observations"
- or data extracted from that project
- file of park meta data, including address, zip, city, neighborhood, official name, and REDCap name for each park
- (optional) file containing record ids to redact

1.2 Output categories

- data-metadata - tables of analyzable line item data from different stages of preparation leading up to analysis, generally in csv format
- tables - tabular outputs of analysis, generally in xlsx format
- charts - chart outputs of analysis, generally in pdf and/or png format

1.3 This is a quarto generated document

By rendering/knitting the qmd file, the analysis is re-executed, this document rebuilt, and new outputs are generated. To learn more about Quarto see https://quarto.org.

2 Setup & Environment

This script was last executed using R version 4.3.1 (2023-06-16 ucrt).

2.1 Secrets and tokens

In order to pull data directly from REDCap, API information must be provided. You should create a file called "secrets.txt" in a subdirectory "./local_only/" This file should include two lines of R code:

```
api_token       <- 'yourapikeyhere'
api_url         <- 'https://redcap.iths.org/api/'
```

Note, that the .gitignore for this project is configured to exclude your secrets.txt and anything else in the local_only direcotry by default, it will not upload to github, and you will not be able to see other users' secrets.txt. They are only stored on your machine.

3 Data Preperation

3.1 Extraction

Data are extracted directly from REDCap via API.

3.2 Transformation

Data types are assigned. And character dates configured to POSIX dates.

POSIX dates are used to generate individual variables for the day, month, and a weekend indicator variable.

3.3 Cleaning

Description	Details
Drop pre-study data	Study start date is 7/1/23 with the first park being Garfield Playfield. Observations prior to this date or the first observation of Garfield are dropped
Drop incomplete entries	observations where the REDCap status is not "Complete", that are missing a timestamp, or missing a park name are dropped
Drop duplicates	duplicate entries (exclusive of redcapID) are dropped. Note that where this creates incomplete days, this is corrected in later QA
Drop inacurate subareas	Some parks were identified as having having sub area data input without having multiple sub areas (and so shouldn't be processed as sub areas). The subarea entry for these parks is removed.
Drop observations identified by human review	Some observations were identified through review of base data by program managers. These are identified in the file "Records-to-Remove.xlsx"

3.4 Create table of activities including sub areas

Park activities are extracted from other park data for more accurate analysis.

3.5 Collapse sub areas

For this analysis, we are not using sub areas. These can be collapsed into single areas. For each observation, sub areas will be collapsed using the following rules:

- numerical observations will be added together -categorical observations become affirmative/existing if any of the subareas are affirmative/existing -timestamp of the earliest observation in the set will be used

Sub areas of the same target area will be assumed to be of the same observation period based on the following logic:

-for a sequence of sub areas observed in the same 50 hours period apparently missing sub-areas will be ignored (assumption: not all sub areas are necessary)

If non unique sub area labels are identified:

- check if the expected list of sub areas are in teh data set if too many, attempt to identify redundancies and remove these if too few, interpolate missing information

In this step we also append changed record_ids to our activity table as "record_id_aggregated" and save the updated activity table.

3.6 Quality Checks

The QA process attempts to identify and correct errors. The process initially performs a series of checks on all park data and reports results. For each park that fails QA, various strategies are executed to attempt to cure that park's data. The final results, including which strategies were executed, and the final QA status are saved in a csv file for review.

3.6.1 strategy 1: correct missingness where observations are only missing but correct number of days

This strategy looks at the number of days of data observed, and, if only 3 days are observed, then checks to confirm that, for each target area, 8 or fewer observations are made. If both of these conditions are met, this strategy attempts to insert the missing observations as blank entries in the part of the day that appeaer to be missing them. It identifies part of day by looking at each data collection period and inserting observations into the underweighted period(s) until 8 exists for each target area.

When executing this script, the user may choose to use all data, or only data that have passed QA. It is generally suggested to only use data that have passed QA.

3.7 Parks metadata

Metadata are provided for each park by Seth. The formal name, address, city, zip, neighborhood, tract, equity score, image status, planned park change notes, and general notes for each park will be appended to the analysis table.

3.8 Assign population within half-mile radius

For each park we estimate the number of people within a half-mile. This is based on the generally accepted threshold of a 10-minute walk to a park as a metric goal.

This relies on access to APDE population data and use of the RADS library and on park longitudes and latitudes provided in the meta data.

3.9 Study tracker

For the 2023 analysis we advised creating a study tracker. For each time a park is studied, a new entry is created. Each entry should includ the park name, the date the study staretd, and a brief description of the study. We generate and save a study tracker that covers 2022 and 2023 studies.

In a later step we will use this table to append a study count to our parks

4 Parameterizing Data For Analysis

4.1 Time of day

For each observation we need to assign a time period to that observation. Because time stamps and the actual time of data collection may not align, the assignment of the time period requires some assumptions and may effect results.

4.1.1 assigning time periods based on sequence

Our primary approach to assigning time periods relies on the assumption that timestamps are in order for each target area (e.g. the first observation of park A area 1 is always “morning1”, the second, “morning2” and so on.)

In turn, with this approach, we assign the sequence of the day such that the first morning1 is day 1, the second is day2, and so on.

We assume that there are the correct number of observation periods such that 1 can be assigned to each period, and 8 to each day.

For each park, there should be 3 days of observation. We will detect this by extracting the date per park and putting observations in order of day "1" "2" "3".

Finally, we cross check each day and period assignment to confirm that:

- All all parks have the expected number of period observations (e.g. if a park has 4 target areas, it should have 12 morning1 observations).
- All observations assigned to a particular day-count (e.g. "1") actually are on the same date (this assumption may need to be relaxed or data corrected before all data pass QA).
- That each day has the expected number of observations.

4.2 Aggregation of periods for analysis

For all analyses, we do not want to distinguish between the first and second of a period (e.g. morning1 and morning2). We aggregate these according to the following:

positive indicaters	aggregate to
accessible	Yes
usable	Yes
lit	Yes
occupied	Yes
supervised	Yes
organized	Yes
equipped	Yes

counts of	are aggregated as
num_child_prim	ceiling of mean
num_child_snd	ceiling of mean
num_child_spec	ceiling of mean
num_teen_prim	ceiling of mean
num_teen_snd	ceiling of mean
num_teen_spec	ceiling of mean
num_adult_prim	ceiling of mean
num_adult_snd	ceiling of mean
num_adult_spec	ceiling of mean
num_senior_prim	ceiling of mean
num_senior_snd	ceiling of mean
num_senior_spec	ceiling of mean

Activities are independently captured in the activity table and not aggregated here. The non aggregated observations table may contain multiple activities separated by a ";" where they were grouped from sub-areas. It is recommended to use the activity table for analysis, and join other park observation data to it using the record_id and record_id_aggregated where necessary.

4.3 Integrate metadata and analysis sets

We join our metadata and park observations for analysis and for use in other software such as excel.

5 Results

Results are saved in multiple locations:

- an excel workbook that contains a separate page for each table of analysis results
- folders with charts and tables of results designed to be integrated into documents
- outputs of various steps of the metadata, QA, and final analysis ready sets

5.1 Included parks

Table 4: Parks Included In Analysis

Park Name
Annex Park
Arbor Lake Park
Bicentennial Park
Cascade View Community Park
Cecil Memorial Park
Chelsea Park
Crestview Park
Crystal Springs Park
Dick Thurnau Memorial Park
Dotty Harper Park
Duwamish Gardens
Duwamish Hill Preserve
Duwamish Park
Fort Dent Park (North)
Fort Dent Park (South)

Table 4: Parks Included In Analysis *(continued)*

Park Name
Garfield Playfield
Hazel Valley Park
Hazelnut Park
Hilltop Park
Jacob Ambaum Park
Joseph Foster Memorial Park
Lake Burien School Memorial Park
Lakeview Park
Linde Hill Park
Manhattan Park
Mathison Park
Moshier Memorial Park
North Shorewood Park
Puget Sound Park
Riverton Park
Salmon Creek
Seahurst Park
Skyway Park
Southern Heights Park
Town Square Park
Tukwila Community Center
Tukwila Park
Tukwila Pond
White Center Heights Park

5.2 Utilization

The following metrics rely on counts of people observed. The underlying data are of people observed within an observation period (e.g. "morning1") and target area. It is possible that the same people may be observed across multiple blocks of time and multiple target areas.

Because of this user counts are more accurately understood as "person time of use per target area" and represents a target area being used by a person within the observation period. This explanation accounts for people being counted multiple times by crossing target areas during the observation period.

5.2.1 number of park users

The total number of users observed in the park across all observation periods.

Notes:

- Due to the study design, user counts are subject to both over and under counting
- User counts are more accurately understood as "person time of use per target area"

5.2.2 average number of daily park users

The daily average number of users observed in the park.

5.2.3 average number of daily park users by time period

The daily average number of users within each time period.

For each park:

$$(\text{Average Park Users By Period}) = \frac{\sum_{d=1}^{3} (\text{People}_p)_d}{3}$$

where $_d$ is the day of study and (People_p) are the sum of number of people observed in a time period of a given day.

The people within a given time period and day is defined as the average (rounded up) of people observed in a time period of a given day and target area $_t$, summed across across all target areas.

$$\text{People}_p = \sum_{t=1}^{n} \left(\lceil \frac{(\text{people observed first half of period}) + (\text{people observed second half of period})}{2} \rceil \right)_t$$

Notes:

- User counts are more accurately understood as "person time of use per target area"
  - "Person using target area during the observation"
- If taken strictly as "people using a park" then this may be an over or under count
  - If in the first morning observation 2 people are observed, and the second 3, this may be 3 unique people, or as many as 5, but we calculate this as 2.5 and round up to 3.
  - If two people walk across all target areas during an observation period, they would be counted each time. With 10 target areas, this would be 20 people observed.

5.2.4 rate of park use by time period

The proportion of the total users observed within each time period.

For each park:

$$(\text{Rate of Park Use By Period}) = \frac{\sum_{d=1}^{3} (\text{People}_p)_d}{\sum_{d=1}^{3} \text{People}_d}$$

where $_d$ is the day of study and (People_p) are the sum of number of people observed in a time period of a given day.

The people within a given time period and day is defined as the average (rounded up) of people observed in a time period of a given day and target area $_t$, summed across across all target areas.

$$\text{People}_p = \sum_{t=1}^{n} \left(\lceil \frac{(\text{people observed first half of period}) + (\text{people observed second half of period})}{2} \rceil \right)_t$$

The total number of people in a day, People_d, is defined as the sum of all People_p within a day.

Notes:

- This is accurately understood as the rate of park use during the period
- There is likely still some error from the under and overcounting of the underlying counts, but the more true that over and under counting is randomly distributed across all time periods, the less true this is.
- Rate is within-park (each park totals to 100%)

5.2.5 average number of daily park users by age

The daily average number of users observed in each age group.

Computationally this measure is similar to the average users by time period above.

Notes:

- User counts are more accurately understood as "person time of use per target area"
- If taken strictly as "people using a park" then this may be an over or under count
  - If in the first morning observation 2 people are observed, and the second 3, this may be 3 unique people, or as many as 5, but we calculate this as 2.5 and round up to 3.

  - If two people walk across all target areas during an observation period, they would be counted each time. With 10 target areas, this would be 20 people observed.

5.2.6 rate of park use by age

The proportion of the total users observed within each age group.

Notes:

- This is accurately understood as the rate of park use by each age group
- There is likely still some error from the under and over counting of the underlying counts, but the more true that over and under counting is randomly distributed across all time periods, the less true this is.
- Rate is within-park (each park totals to 100%)

5.3 Occupancy

5.3.1 occupancy rate

The percentage of observations where at least one user was observed in the park.

For each park, the number of observation periods with any target area in occupied status is divided by the total number of observation periods (24)

Notes:

5.4 Activities

5.4.1 number of users per activity

The number of users observed doing the activity in the park.

Notes:

- user counts are subject to over and under counting due to

5.4.2 rate of user activity

The percentage users observed doing the activity in the park.

For each park, the number of users engaged in an activity is divided by the total number of users observed in the park throughout the study duration.

Notes:

- All activities listed have at least 1 participant, but some may show 0 in the prepared tables due to being less than 0.01%
- The total rate of all user activities listed will equal 100%, the total amount of activity observed.
- This is calculated on the non aggregated population counts.

5.4.3 rate of activity observed

The percentage of observations where at least one user was observed doing the activity in the park.

For each park, the total the number of periods the activity was observed is divided by 24, the total number of observations periods possible for any particular activity.

Notes:

- Unlike many of the other measures provided, these rates are mostly independent of each other and do not have an additive meaning. This is because multiple activities may be observed in a single observation period. E..g "walking" may be observed in all 24 periods, and so have an observation rate of 100%, and "sitting" may be observed in 6 periods and so have a rate of 25%.

5.5 Spatial analysees

5.5.1 ratio of use to half-mile catchment area

The ratio of how many users were observed per day on average relative to how many people live within 0.5 miles of the park.

This is provided per 1,000 residents to improve readability.

Notes:

- The ratios are calculated on non-rounded data, and then rounded. The average number of users and populations provided in the formatted word document are rounded. This causes a rounding error where you wouldn't get the exact ratio if you were to divide these users and populations. For accurate results use the numbers provided in the pivot ready tables.

6 Closure and Next Steps

6.1 REDCap maintanence

Project has been in "development" status throughout the study period. It has now been moved into "analysis/cleanup". It is recommended to keep it in this stage until finished with any future studies and analyses in this line of work, e.g. if planing to use the data from this study in the future, it is good to keep this REDCap project in analysis mode for review. There is an additional "completed" status, which should generally only be used when completely finished with the body of work. This status makes the project largely inaccessible. Notably, changing the project to "complete" status would also break this script (or require it to be fed a different data source rather than pulling from the REDCap API.)

As currently envisioned, these data are part of a longitudinal study, and so this and future REDCap projects would ideally remain accessible.

I recommend conducting future data collection efforts in their own REDCap projects, naming them similarly, such as "Public Health - Seattle & King County Park Observations:

"

The project used for this analysis has been renamed to: Public Health - Seattle & King County Park Observations: 2023 Annual Study

6.1.1 recommended changes to future surveys instruments

- variable for capturing other primary activities is inconsistent to secondary and special ones
- modify instrument to have observers indicate the time of their data collection
- add a study name to data (hidden value)
- include additional testing and possibly consultation with ITHS REDCap staff for issues with data upload from cell phones. This cause problems for observers. It is unclear if this is a problem with the cell phone application or due to the study being executed with the project in "development" status.

6.2 Data management

Files in the provided data-metadata directory should be retained. These include QA information, updated metadata, and observational data at different stages of data preperation.

Most critical is the maintenance of the version of the data used for the above analysis. Over the years, these would be the ones expected for use in further analytic work and for comparing year-over-year results.

The analysis-ready files include:

- SOPARCAnalysisSet.csv (fully cleaned park observation)
- SOPARCAnalysisSetAggregatedPeriods.csv (analysis set with observations aggregated by period, the level of analysis most commonly used)
- SOPARCActivities.csv & SOPARCActivitiesExpanded.csv (all activity data at the individual record level. The expanded version includes some park information already attached for human readability)