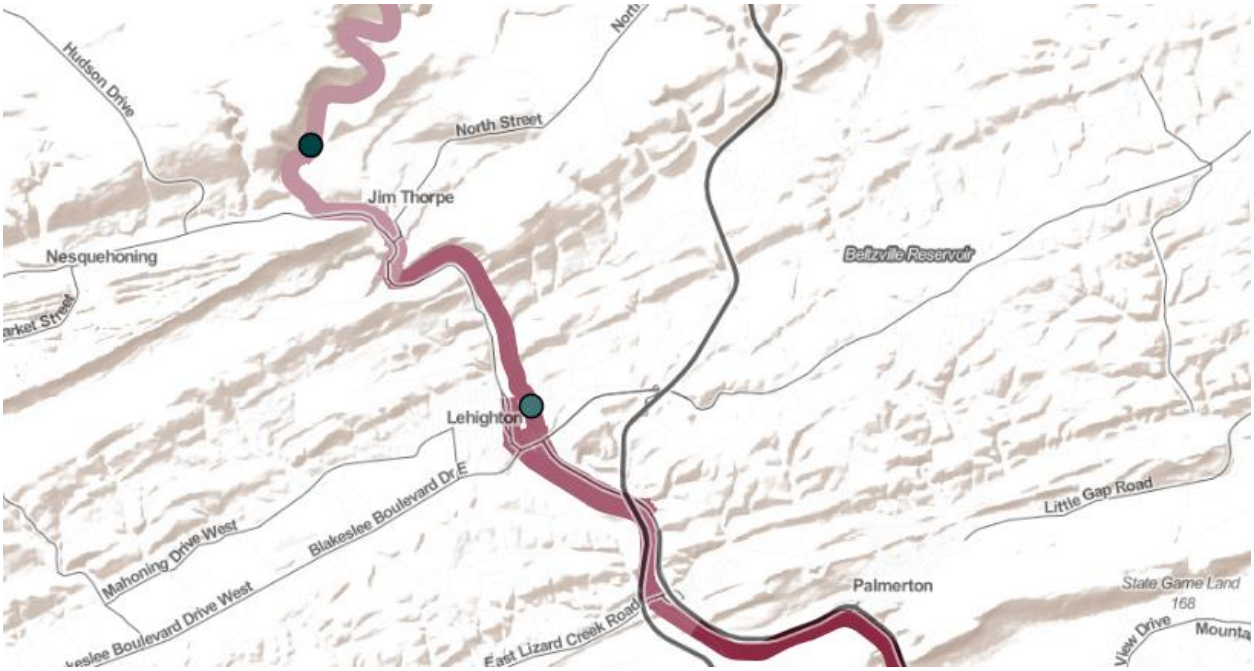
DELAWARE & LEHIGH
NATIONAL HERITAGE CORRIDOR

Trail Health Benefits Data Analysis



Prepared by

azavea

990 Spring Garden Street, 5th Floor
Philadelphia, PA 19123
(215) 925-2600
http://www.azavea.com

Executive Summary

The Delaware and Lehigh National Heritage Corridor is a national heritage area in 5 counties in Eastern Pennsylvania from Luzerne to Bucks County. Azavea performed data analysis to calculate health benefits statistical summaries and breakdowns by different categorizations of data from the organization's "Get Your Tail On the Trail" health incentive program. From 2017 to 2019 over 7,000 users reported over 370,000 activities where they recorded time, distance, mode of activity, and a text based location description.

First, Azavea cleaned and categorized the data by Location, Mode, Trail and Trail segment. Once the data was cleaned and categorized, we used a variety of statistical tools and packages in R to summarize the data by joining health survey information from users to calculate a caloric expenditure for the trail, its segments, user modes and locations.

We found that users of the GYTOT program burned **an estimated 213,594,448 calories** between January 1, 2017 and December 31, 2019. In addition, users of the GYTOT burned an **average of 572 calories per activity**. The most popular activities, by far, are walking and biking. Using the GYTOT data as a source to estimate calories burned on the D&L Trail we combined daily trail counts from twelve locations to estimate that **535,514 calories are burned a day on the D&L Trail for a total of 195 million calories a year.** The Bethlehem-Easton segment of the trail shows the highest amount of calories burned (36 million). However, on average users of the Wilkes-Barre - White Haven segment are burning more calories per trail activity (960 calories).

Methodology

Data Sources

- GYTOT Activity Log (January 1, 2017 to December 31, 2019)
- GYTOT Health Survey
- Trail counts

The Get Your Tail On The Trail Activity Log data contained over 370,000 records from users recording their physical activity. The users categorized their activity by Mode and could enter Location, Time and Distance. The GYTOT Activity Log data provided the fundamental data source for this analysis.

The GYTOT Health Survey data contained over 7,000 records submitted by users containing their personal information such as height, weight and age. This data can be linked to the Activity Logs based on username.

The Trail count data and corresponding shapefile contained data on trail counts at a dozen relevant locations along the D&L corridor. This data was recorded in 15 minute increments. Additional calculations will be described in the Caloric Expenditure section of the Methodology.

Data Cleaning

The GYTOT Activity Log data contained two fields which needed to be standardized and categorized for the analysis. The *Location* and *Modality* field can be filled by users with a variety of pre-populated options or users can enter free-form text. Since one of the goals of the project is to produce statistics on caloric expenditure by Mode and Location, these fields had to be categorized into a reasonable number of options for the analysis. Through consultation with D&L, we settled on the following categories. The total number of records after data cleaning for each Mode is listed.

Mode	Total Records
Biking	58518
Elliptical	3036
Hiking	10940
Other	15006
Paddling	1659
Running	56765
Stationary Bike	1359
Swimming	1178
Treadmill	26747
Walking	197880

We also wanted to categorize by location to compare statistics by trail segment and a boolean value for whether the activity was recorded on the D&L Trail.

Location	Total Records
Community	41817
Gym	19095
Home	27223
Neighborhood	23543
Other	159655
Trail	*92779*
Work	8976

We were able to identify 92,779 records that were recorded on the trail. We were further able to categorize the records based upon which segment of the D&L Trail the activity took place.

D&L Segment	Total Records
Allentown - Bethlehem	3004
Bethlehem - Easton	5516
Easton - Riegelsville	1064
Jim Thorpe - Lehigh Gap	5483
Lehigh Gap - Allentown	6366
Morrisville - Bristol	84
New Hope - Morrisville	94
Riegelsville - New Hope	343
White Haven - Jim Thorpe	2370
Wilkes-Barre - White Haven	668

We identified 24,992 records that were recorded on a trail segment.

Cleaning and Categorization Methods

In this section we describe the methods used for cleaning and categorizing the data. The goal of this process was to correct for spelling inconsistency and other text variations between entries with the same intended meaning.

Benchmark function

This is essentially a score to tell us how many of the records have been cleaned and classified into the target categories. We calculated the benchmark after most intermediate steps to internally assess the quality of our decisions and the performance of algorithms. Before changing values, we created logs so that we could check that the cleaning algorithms were making only desired corrections.

String fuzzy matching

For each field that required cleaning, the program we used walked through low occurring entries and tried to match and change them to either the predetermined target values or, if none fit, other commonly occurring values that could be later mapped to a target. To accomplish this step, we used a fuzzy string matching library in the python programming language (https://pypi.org/project/fuzzywuzzy/) utilizing 4 algorithms to calculate string similarity scores.

- Base/Simple Ratio
- Partial Ratio
- Token Sort Ratio
- Token Set Ratio

For each word pair, consisting of a low occurring word and a target or commonly occurring word, we took the maximum of the 4 calculated scores to represent the overall similarity. The uncommon value was then changed to the word for which it had the highest score, if that score fell above the minimum similarity threshold. This change was also written to the logs to be verified and adjusted if needed.

Map results to predetermined categories

In a final step, we mapped the commonly-occurring values to the predetermined targets. These were often different words for expressing the same meaning as the target value, rather than simply a spelling or text variation.

Outlier Analysis and Handling

As with any database of user-entered information, it can be expected that users will make mistakes and enter incorrect information. These can be identified and handled to a reasonable degree by performing an outlier analysis to determine which records fall outside a certain degree or range. This can be an iterative process where summary statistics are generated to determine which values fall outside of a reasonable range.

To identify outliers, we generated box plots to first visualize the data and look at long tails on either end. Then, we generated the IQR (Interquartile Range). Since the definition of outliers is somewhat subjective, we used the interquartile range of the MPH value as a "rule" to determine which records are outliers. MPH was chosen because it is a combination of distance and time, therefore would be able to identify outliers in those values as well. The IQR was generated for MPH summarized by Mode. Outliers were primarily identified in the Biking and Walking categories (which also reflect the majority of the data), so only those outliers were actually replaced. An activity log record was considered an outlier if it is more than 1.5 * IQR above the third quartile. We did not eliminate or replace outliers on the low end, as it is reasonable to expect that some people may walk or cycle very slowly.

To handle the outliers, we removed them from the dataset then calculated summary statistics on the activity logs by Mode. The median value for time and distance for each walking and biking was then used to fill back the outlier data. MPH was recalculated using the outlier-handled time and distance in the activity logs. We plotted the MPH again to make sure it was reasonable for each Mode.

Outliers and user-entered errors were also identified in the Health Survey data. In that section below we will describe how they were handled.

MET Values

MET values are used to determine the caloric expenditure for each activity log. MET values are based on standard literature. Azavea used the values found in the Compendium of Physical Activities. Below, find the MET values we used for each data classification of Mode.

Bicycling

MPH	MET
< 10	4.0
10-11.9	6.8
12-13.9	8.0
14-15.9	10.0
16-19.9	12.0
20 or >	15.8

Walking

MPH	MET
< 2	2.0
2-2.4	2.8
2.5-2.7	3.0
2.8-3.2	3.5
3.3-3.4	3.9
3.5-3.9	4.3
4-4.4	5.0
4.5-4.9	7.0
5 or >	8.3

Hiking

MPH	MET
All	6.0

Running and Treadmill

MPH	MET
< 5	6.0
5-5.1	8.3
5.2-5.9	9.0
6-6.6	9.8
6.7-6.9	10.5
7-7.4	11.0
7.5-8.5	11.8
8.6-8.9	12.3
9-9.9	12.8
10-10.9	14.5
11-11.9	16.0
12-12.9	19.0
13-13.9	19.8
14 or >	23.0

Elliptical

MPH	MET
All	5.0

Paddling

MPH	MET
>3.9	2.8
4.0-5.9	5.8
>6	12.5

Stationary Bike

MPH	MET
> 10 MPH (Very light effort)	3.5
10-15MPH (light to moderate)	4.8
15-16MPH (Moderate to vigorous)	6.8
17-19 MPH (Vigorous)	8.8
20-21 (Vigorous)	11.0
>22 (Very Vigorous)	14.0

Swimming

MPH	MET
All	8.0

Other

MPH	MET
All	4.3

Azavea used the median value of all records, **4.3**, for any record with Other as the Mode.

Caloric Expenditure

In this section, we will describe the methods for generating a caloric expenditure calculation for each activity log.

Health Survey Data

The Health Survey data provided by D&L contained 7,623 records. This data tracks the weight for each user in a time series, from March 2017 to May 2019. This dataset is also prone to data error since it is user-entered information. We generated summary statistics on each field for weight to identify outliers and manually fix them. We also gave the data a cursory look to identify mistakes in each row, for example one user entered *160* for each month except one where they entered *610*. These easily identifiable mistakes were manually corrected.

In addition, users did not enter weight for every time period. In the case where they were missing data, we used the mean value for the user's weight for all other time periods to populate their missing values. For example, if someone entered their weight as *165*, *167*, *160*, *blank*, *164*, we filled the blank record

with the mean of 164. We used the same methodology to fill values of *0* where there were other values present for the user. Finally, in some cases users did not enter any weight at all. For those records, we used the population mean of *172*.

To calculate caloric expenditure, we needed the users' weight for each activity log record. The health survey data was converted from *wide* to *long* format to facilitate a tabular join. Since we do not have the user weight for every single time period, we had to approximate their weight at that time by joining the user health survey weight *closest* to the activity log date. For example, for an activity log recorded in June 2019, it would have user weight for May 2019. An activity log with a date of December 2018 would have the user weight for February 2019, since it is closer than May 2018.

Caloric Expenditure Calculation

Using standard literature, we can calculate caloric expenditure with the formula:

$$Calories = (MET * 3.5 * user\ weight\ (kg) / 200) * Minutes$$

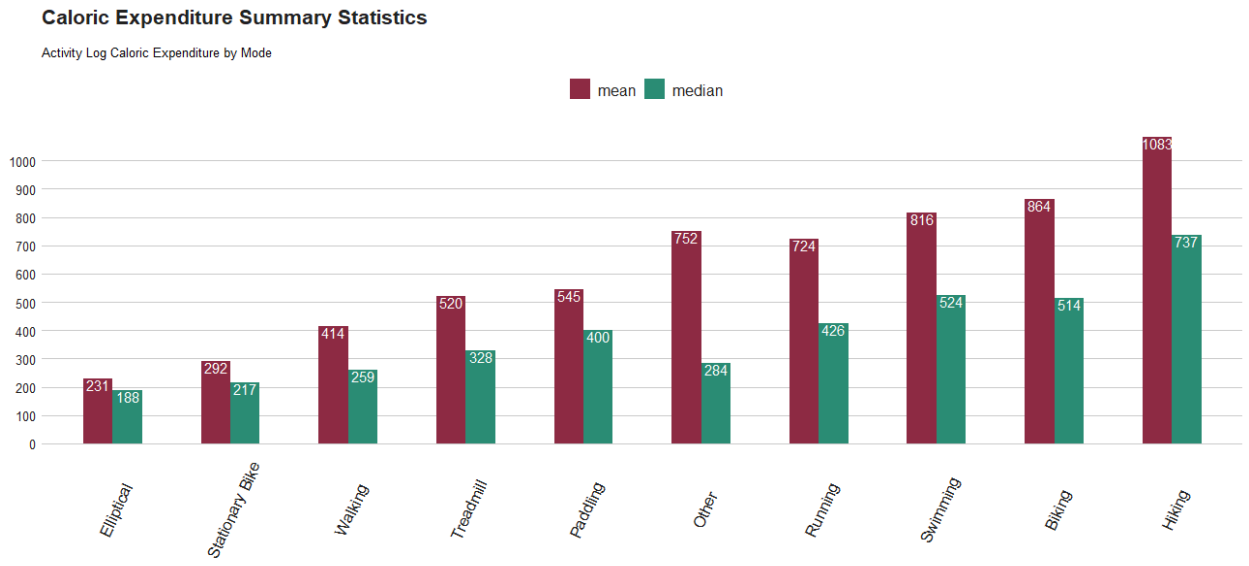
This formula was used to calculate a caloric expenditure for each Activity Log in the GYTOT data. Fields for MET, MET Minutes, and User Weight in Kilograms were added to the data to complete the calculation.

Findings

By using the data in each Activity Log record, we can summarize the caloric expenditure by categorizing the data into different buckets by Mode, Location, Segment and D&L Trail.

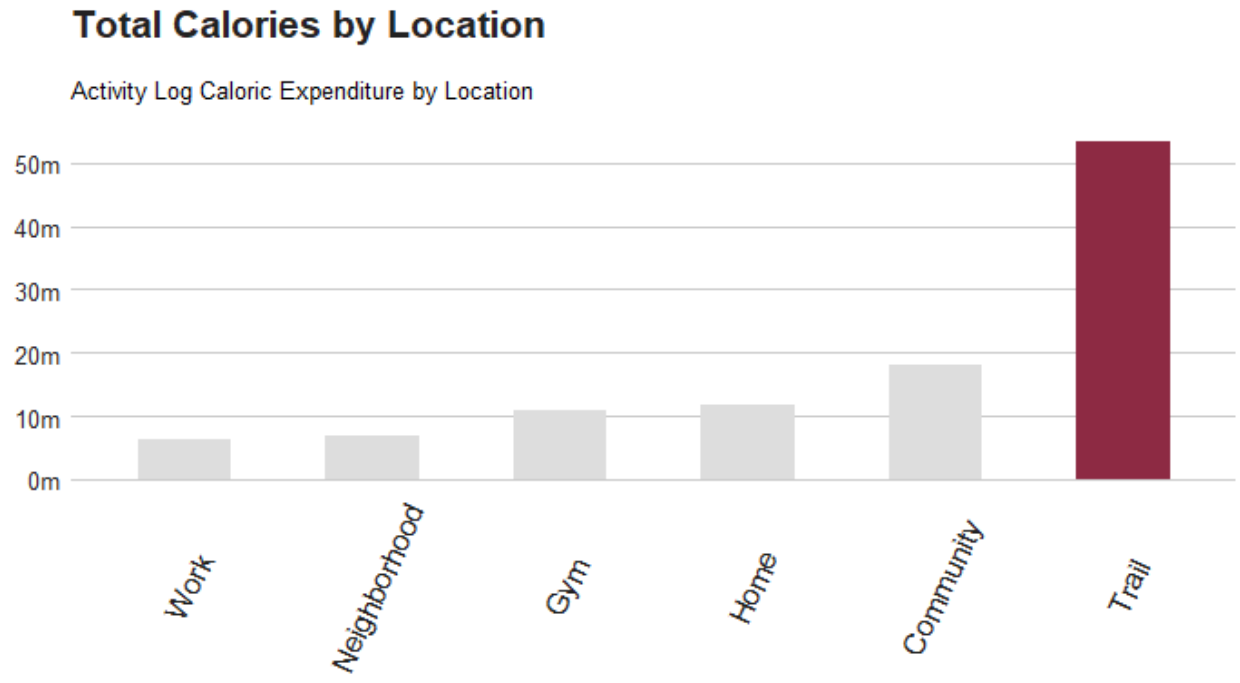
By Mode

We find that on average users are expending the most calories while hiking, followed by biking and swimming.

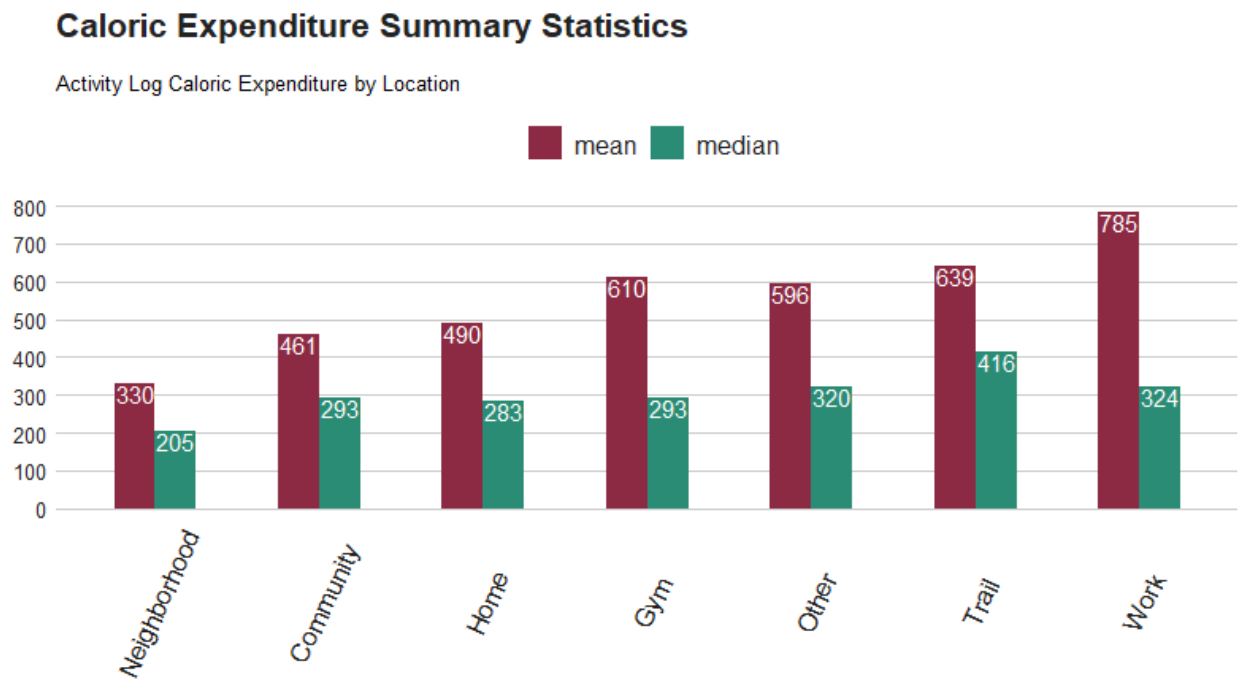


By Location

By far, the Activity Logs indicate that people are burning the most calories at locations indicated as a Trail. This can include the D&L trail or any other trail.



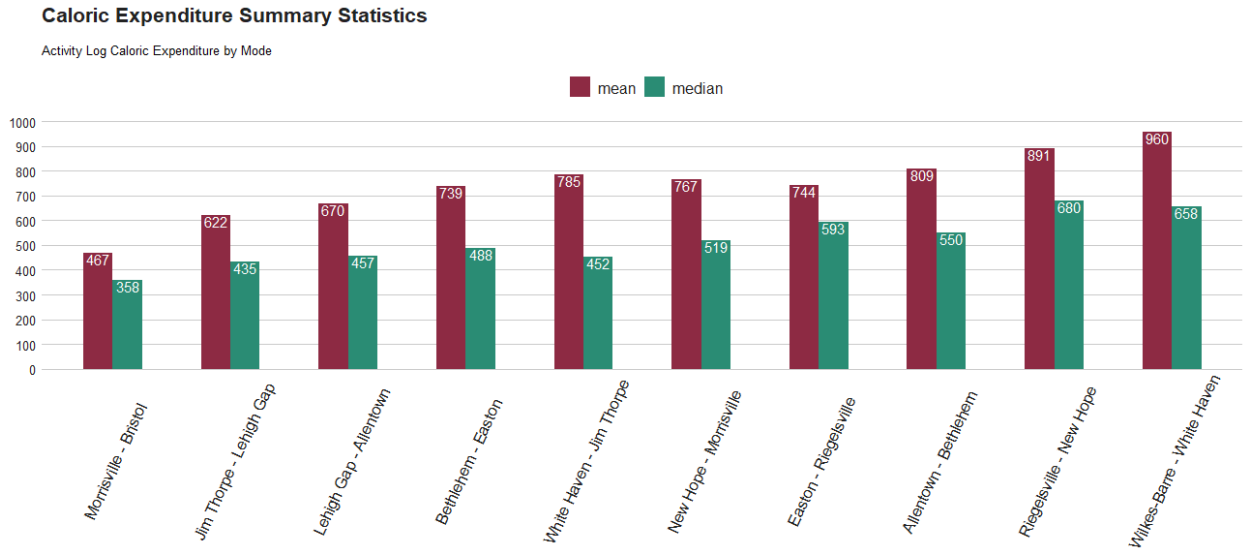
That would be logical considering most of the records were recorded with a Location of Trail. However, when normalizing the data, we find that the average caloric expenditure is also quite high - second only to the Location of Work. **The median value for caloric expenditures of 416 calories burned per Activity by Location is highest for Trail.**



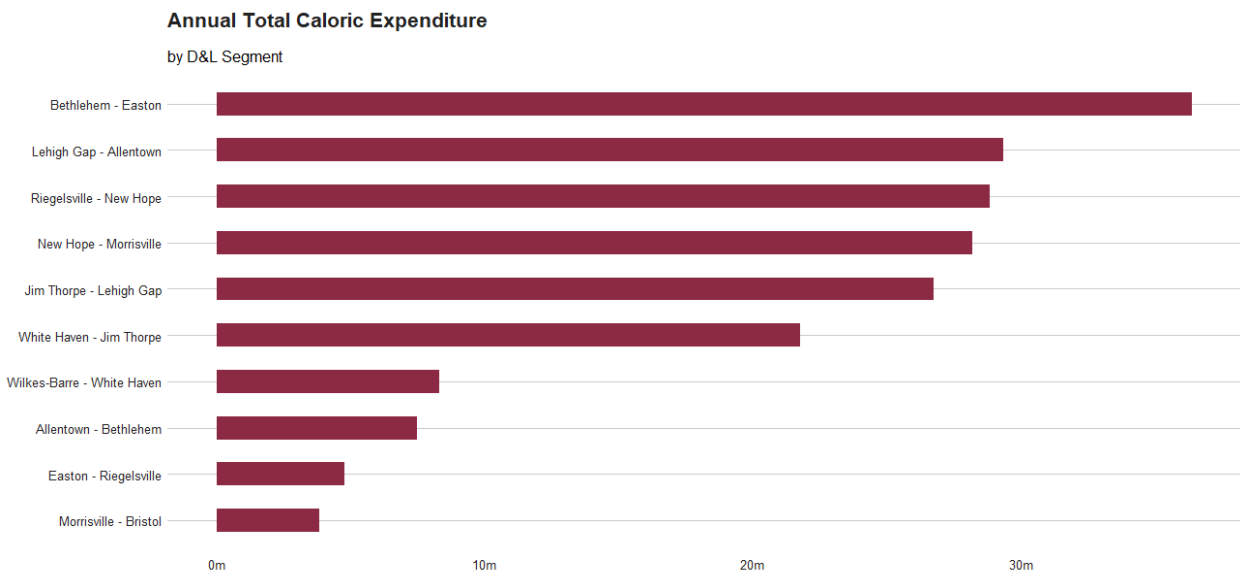
While the average caloric expenditure is higher for Work, the data is noisy and has a much lower sample size - only about 9,000 records or less than 3% of the total activity logs.

By D&L Trail Segment

We summarized the data by D&L trail segment in the Activity Log records. **On average, the highest caloric expenditure per activity was on the Wilkes-Barre - White Haven trail segment.**



However, in total, the most calories were expended on the Bethlehem - Easton trail segment with an estimate of over 36 million calories expended annually.



In the table below, trail counter data was used to estimate the daily and yearly total calories expended by trail segment. The average daily trail visits can be multiplied by the Mean caloric expenditure from the Activity Logs (AL) to determine the Estimated Average Daily Calories and Estimated Annual Total Calories expended for each trail segment.

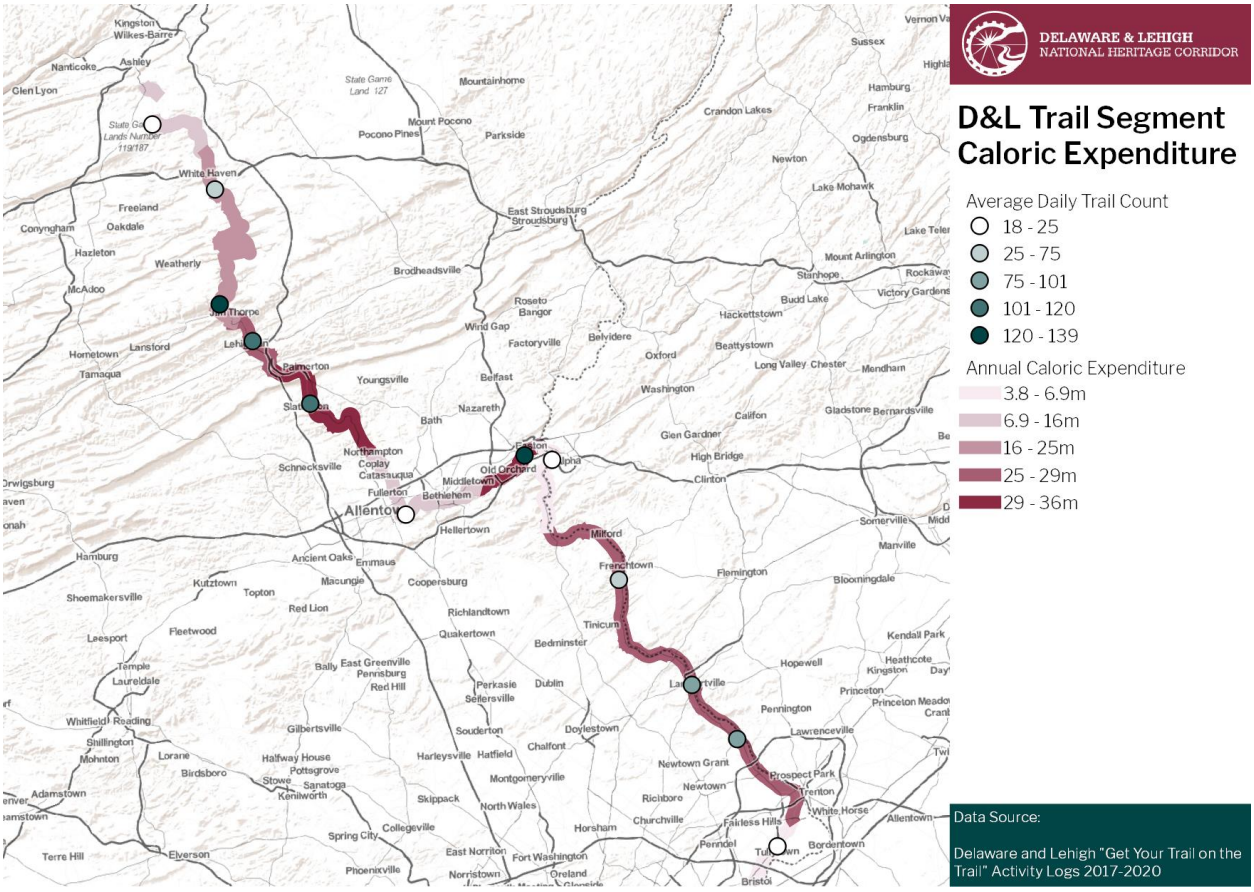
D&L Segment	Mean (from AL)	Median (from AL)	Average Daily Trail Visits	Estimated Total Daily Calories	Estimated Total Annual Calories
Allentown - Bethlehem	809	550	25.3	20,477.04	7,474,118.95
Bethlehem - Easton	739	488	134.75	99,574.24	36,344,597.65
Easton - Riegelsville	744	593	17.6	13,091.18	4,778,279.12
Jim Thorpe - Lehigh Gap	622	435	117.7	73,174.65	26,708,749.06
Lehigh Gap - Allentown	670	457	119.9	80,349.28	29,327,488.08
Morrisville - Bristol	467	358	22.55	10,521.30	3,840,275.35
New Hope - Morrisville	767	519	100.65	77,153.98	28,161,203.50
Riegelsville - New Hope	891	680	88.55	78,920.47	28,805,971.86
White Haven - Jim Thorpe	785	452	75.9	59,560.07	21,739,426.80
Wilkes-Barre - White Haven	960	658	23.65	22,692.34	8,282,704.30

By D&L Trail

Summarizing on the Activity Log data by whether the activity was on the D&L trail, we found that there was a substantially higher caloric expenditure on the D&L trail versus activities elsewhere, when one

takes into account trail counter data. This isn't an apples-to-apples comparison, since we have additional data for trail counts.

Below, we designed a map that displays the Trail Segment by Caloric Expenditure overlaid with Average Daily Trail Counts.



Map of Annual Caloric Expenditure by Trail Segment and Average Daily Trail Count

Using the trail count data, we estimate that there is an average of 914 daily trail visits. Using that information, we can estimate that **on an average day 535,000 calories are burned on the D&L trail**. Annually, we can estimate that **195 million calories** are burned on the D&L trail (using the trail count data). Note that this is slightly less than the 190 million calories burned through the GYTOT Activity Log, showing that there may be an undercount of trail visits due to the spacing of trail counters.

Accounting for Missing Trail Visits

Acknowledging that there is an undercount of trail activity due to spacing of the counters and trailheads, Azavea set out to estimate what that undercount might be. Below, we describe the two methodologies used. We refer to D&L trial segments as "sections" here to alleviate confusion between sections that contain multiple segments of trail.

Methodology 1

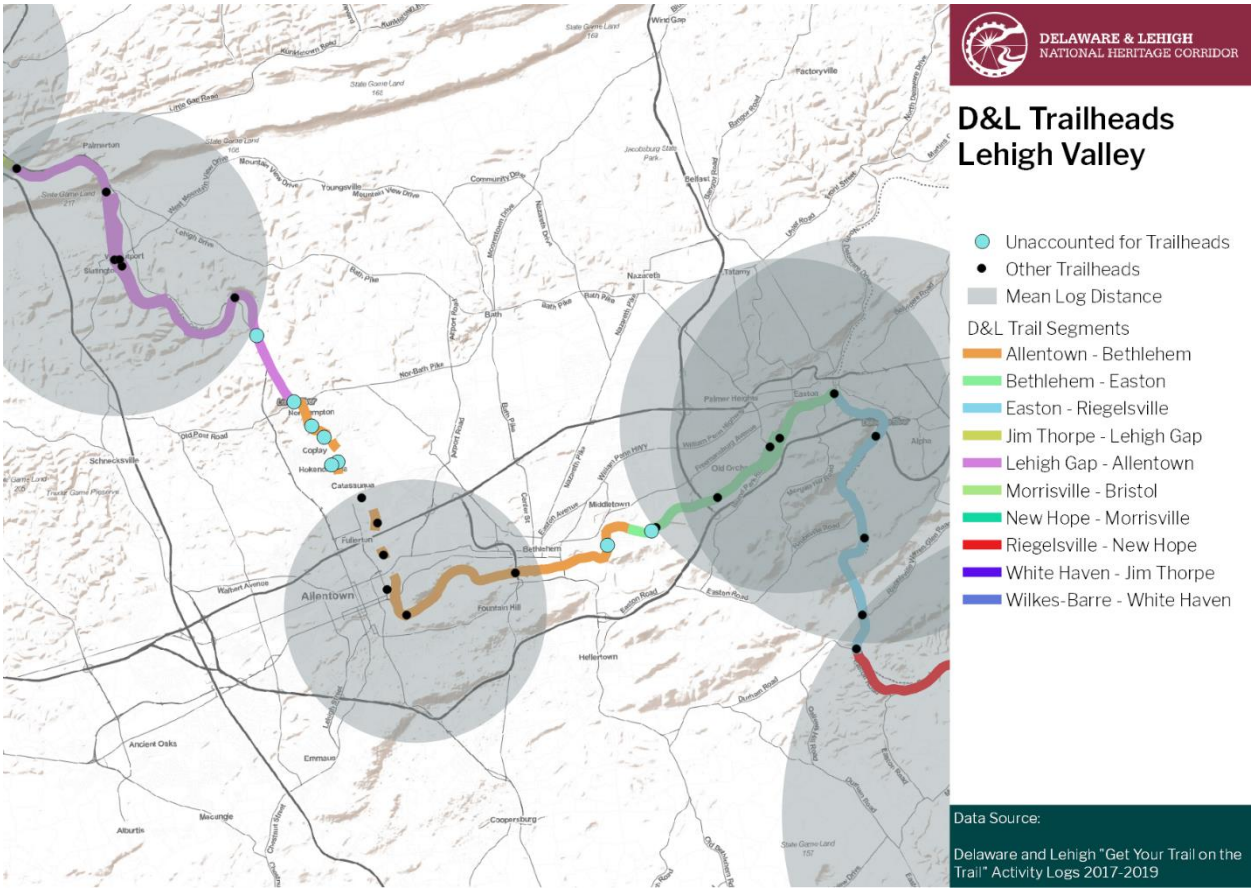
This methodology involves a spatial analysis of trailhead locations and their proximity to trail counters. We identified the trailheads closest approximation to a counted and identified the trail length half of the mean log distance[1] for that section was marked. Trailheads located outside that range were given new counter values equal to the mean of counter values (if more than one counter per section) or counter value for that section. In the cases of sections with multiple segments, we assumed that they had unique counters equal to the counter value for that segment. Most notably, with this methodology the estimated caloric expenditure is drastically higher for the Allentown-Bethlehem section. This section contains separate segments as well as a portion of the main segment well outside the mean log distance. While we suspect the original caloric expenditure of the Allentown-Bethlehem section may be an undercount, it is possible that this number is too high yet close to the real value, considering that it is in such a highly populated area.

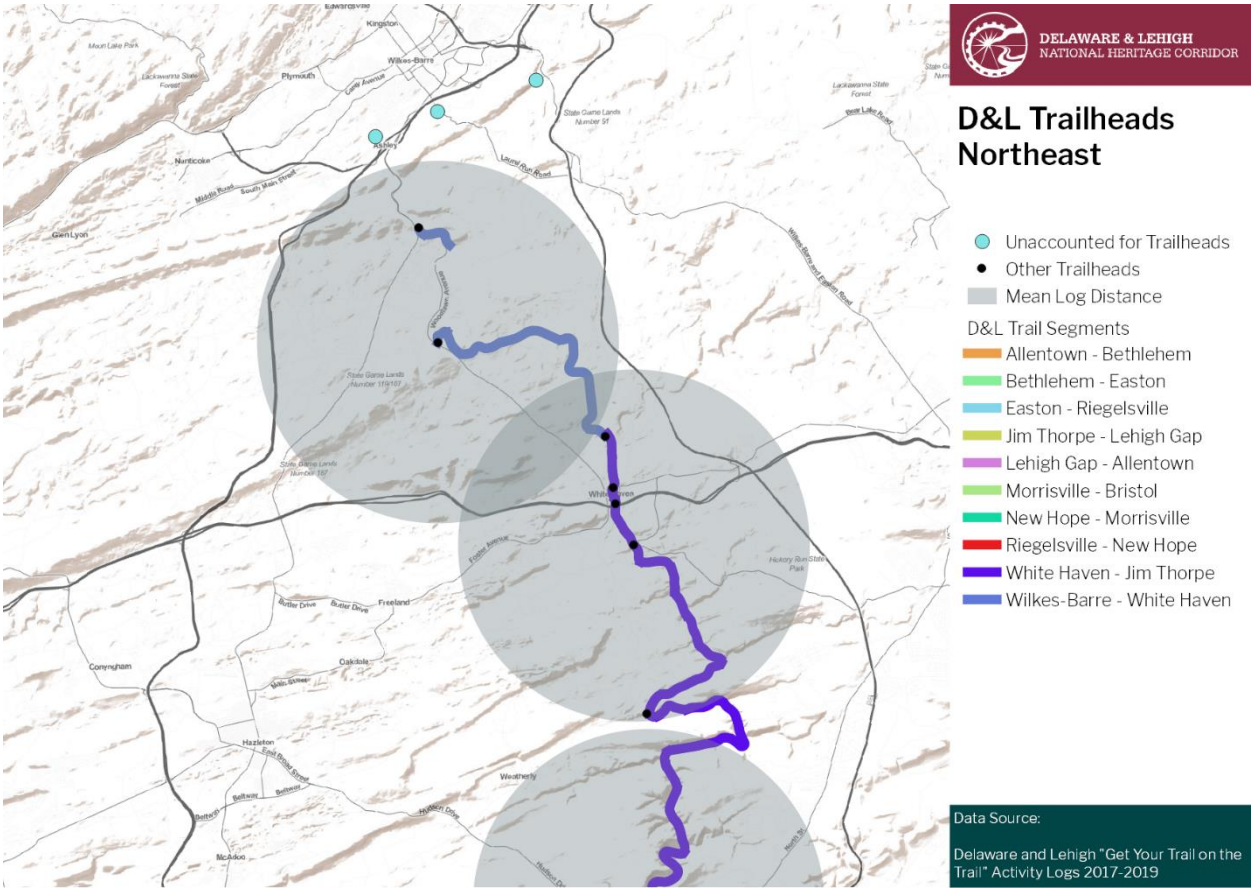
Section	New Average Daily Trail Visits	Average Daily Calories	Estimated Yearly Total Calories
Allentown - Bethlehem	220	178,061	64,992,339
Bethlehem - Easton	135	99,759	36,412,027
Easton - Riegelsville	18	13,389	4,886,876
Jim Thorpe - Lehigh Gap	118	73,361	26,776,826
Lehigh Gap - Allentown	240	160,833	58,703,896
Morrisville - Bristol	46	21,463	7,833,821
New Hope - Morrisville	202	154,845	56,518,262
Riegelsville - New Hope	174	155,078	56,603,491
White Haven - Jim Thorpe	176	138,110	50,410,265
Wilkes-Barre - White Haven	24	23,028	8,405,281

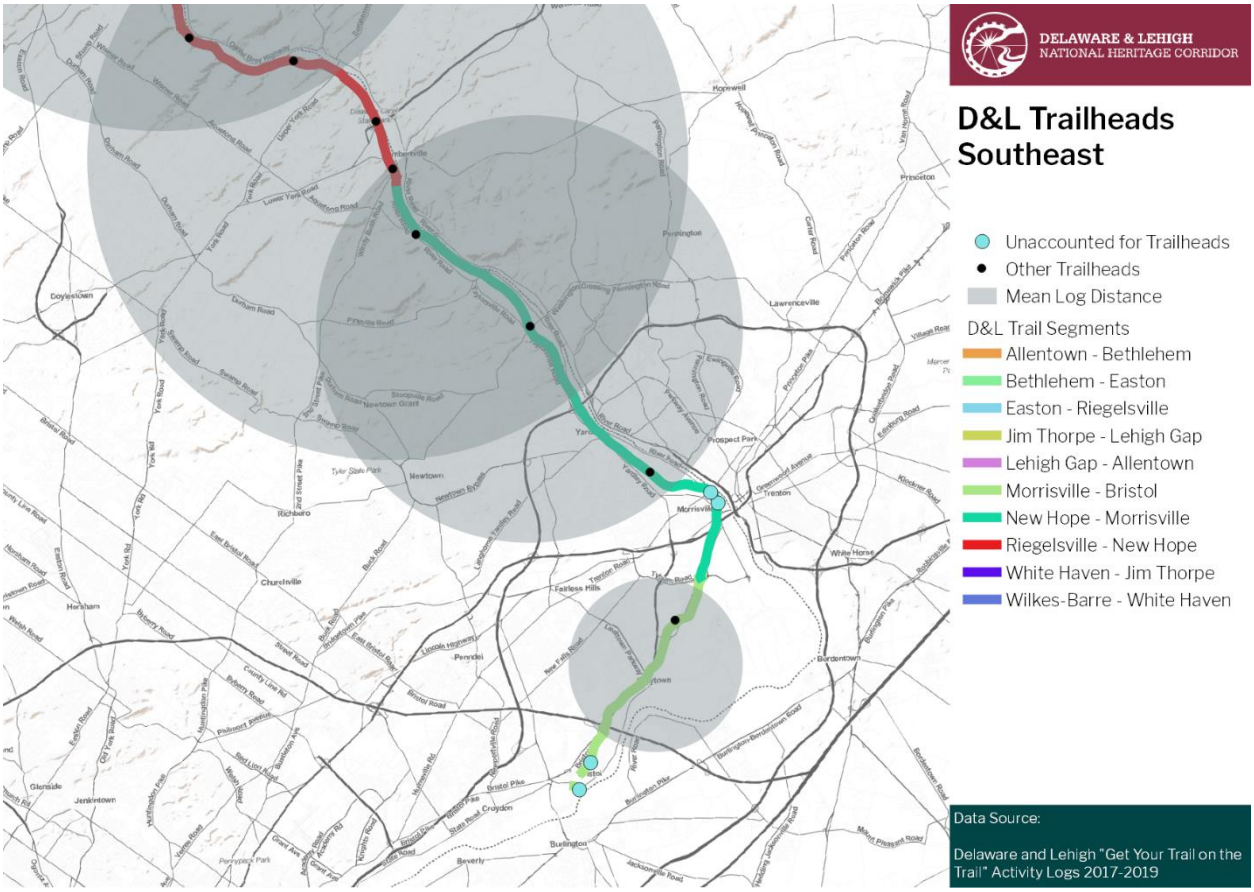
The estimated total annual calories for the entire trail using this methodology is **371,543,085**.

Below, we mapped the trailheads that are outside the mean log distance for each segment from each counter's nearest trailhead. We used an "as-the-crow-flies" buffer an easy visualization of the mean log distance.

[1] Mean Log Distance is the average distance a user traveled on the trail section and was calculated from the GYTOT Activity Log.


DELAWARE & LEHIGH
NATIONAL HERITAGE CORRIDOR
D&L Trailheads
Lehigh Valley
Unaccounted for Trailheads
Other Trailheads
Mean Log Distance
D&L Trail Segments
Allentown - Bethlehem
Bethlehem - Easton
Easton - Riegelsville
Jim Thorpe - Lehigh Gap
Lehigh Gap - Allentown
Morrisville - Bristol
New Hope - Morrisville
Riegelsville - New Hope
White Haven - Jim Thorpe
Wilkes-Barre - White Haven
Data Source:
Delaware and Lehigh "Get Your Trail on the Trail" Activity Logs 2017-2019
Allentown
Bethlehem
Easton
Palmerton
Nazareth
Catasauqua
Coplay
Hellertown
Emmaus
Coopersburg
Macungie
Ancient Oaks
Schnecksville
Factoryville
Tatamy


DELAWARE & LEHIGH
NATIONAL HERITAGE CORRIDOR
D&L Trailheads
Northeast
Unaccounted for Trailheads
Other Trailheads
Mean Log Distance
D&L Trail Segments
Allentown - Bethlehem
Bethlehem - Easton
Easton - Riegelsville
Jim Thorpe - Lehigh Gap
Lehigh Gap - Allentown
Morrisville - Bristol
New Hope - Morrisville
Riegelsville - New Hope
White Haven - Jim Thorpe
Wilkes-Barre - White Haven
Data Source:
Delaware and Lehigh "Get Your Trail on the Trail" Activity Logs 2017-2019



Methodology 2

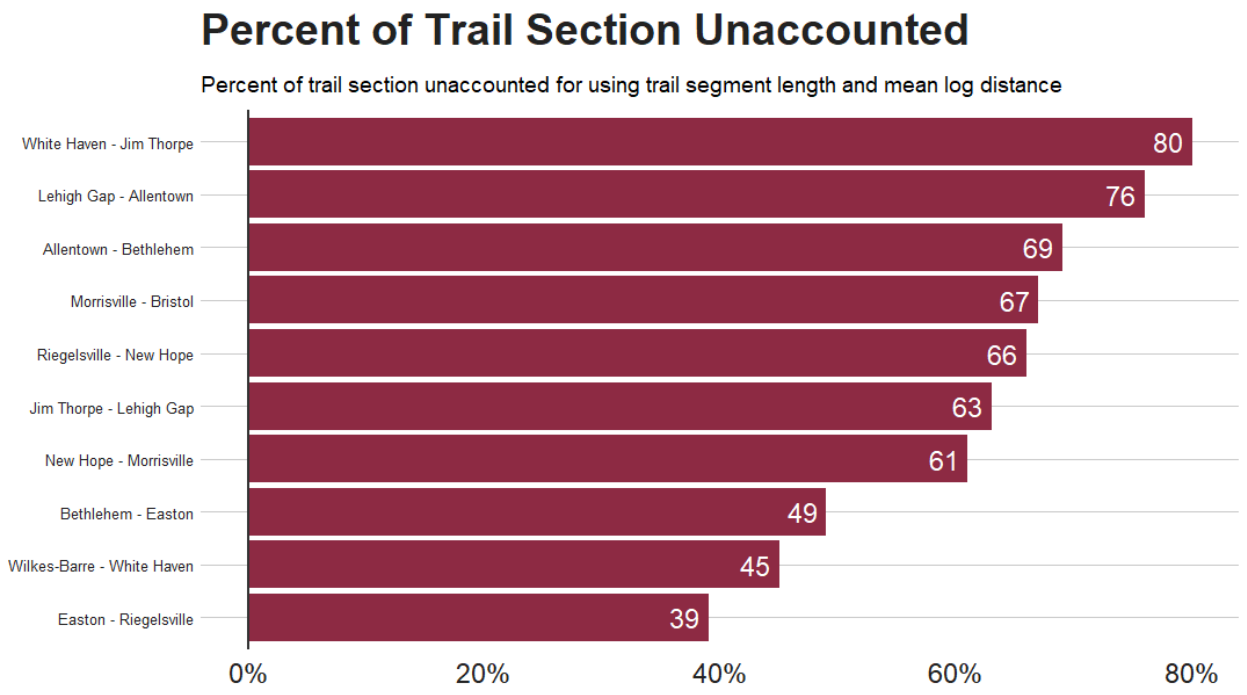
In the table below, we estimate the unaccounted-for trail visits using the difference between the total section miles and half[2] the mean log distance to calculate the "unaccounted" percentage of each segment. Then, we multiplied the average daily trail visits by that unaccounted-for percentage to get a new ADTV and estimated calories per section.

[2] Assumes out-and-back trips

Section	Segment Miles	Mean Log Distance (from AL[3])	% Unaccounted	Orig. ADTV	New ADTV	Estimated Total Daily Calories	Estimated Total Annual Calories
Allentown - Bethlehem	12.3	7.7	69%	25.3	42.7	34,520	12,599,733
Bethlehem - Easton	9.5	9.8	49%	134.75	200.3	148,049	54,037,723
Easton - Riegelsville	9.7	11.9	39%	17.6	24.5	18,189	6,639,068
Jim Thorpe - Lehigh Gap	10.5	7.8	63%	117.7	191.7	119,170	43,497,153
Lehigh Gap - Allentown	18.7	8.9	76%	119.9	211.3	141,592	51,681,237
Morrisville - Bristol	7.7	5.1	67%	22.55	37.6	17,537	6,400,880
New Hope - Morrisville	15.9	12.5	61%	100.65	161.7	123,989	45,256,141
Riegelsville - New Hope	26	17.6	66%	88.55	147.1	131,103	47,852,745
White Haven - Jim Thorpe	26.4	10.3	80%	75.9	137.0	107,499	39,236,955
Wilkes-Barre - White Haven	9.7	10.6	45%	23.65	34.3	32,949	12,026,551

The estimated total annual calories for the entire trail using this methodology is **319,228,186**.

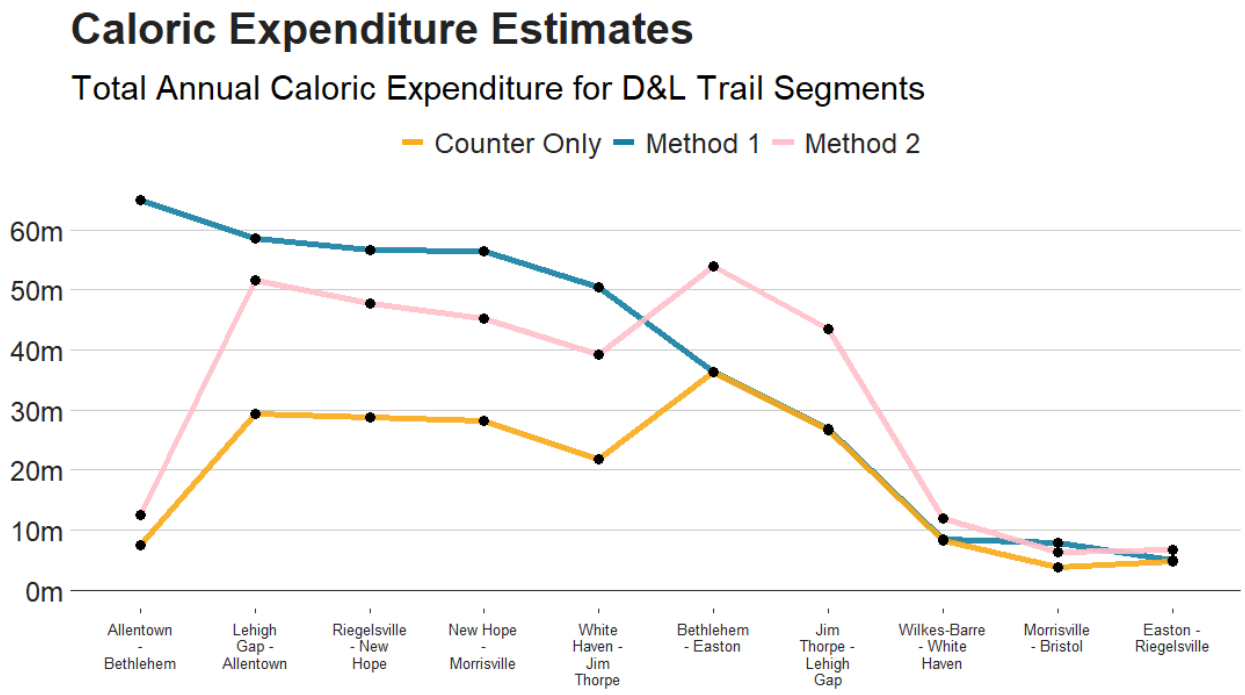
This methodology also exposes which trail segments may be the most at need for additional counters to track trail visits.



[3] Activity Log

Comparing Methodologies

Below, we plotted each of the caloric expenditure estimates. The *Counter Only* methodology, described in the "By D&L Trail Segment" section, can be read as a baseline but clearly an undercount for most trail sections. *Methodology 1* shows much higher estimates of caloric expenditure on trail sections that are broken, with several separate sections. *Methodology 2* is perhaps a good compromise between the two. Overall, these three methods show the level of uncertainty in estimating trail counts and caloric expenditure. We believe they do indicate a strong need for more data, primarily in the Lehigh Valley area.



Resources

Below, a list of technical resources that were used in this project.

- Python library for data cleaning: https://pypi.org/project/fuzzywuzzy/
- R packages for data analysis:
  - Dplyr
  - Tidyr
  - Ggplot
  - lubridate
  - Bbc_plot
- QGIS for mapping
- We will provide a private Github repository containing the scripts and interim datasets used in the analysis.