Data Fusion for Travel Analysis: Linking Travel Survey and Mobile Device Location Data

Guangchen Zhao
University of Maryland
Mohammad B. Al-Khasawneh
University of Maryland
Tiziana Tuoto
Italian National Institute of Statistics
Cinzia Cirillo (✉ ccirillo@umd.edu)
University of Maryland

Research Article

Keywords: Travel behavior, Mobile device location data, Travel survey, Probabilistic record linkage, Data panel

Posted Date: February 1st, 2024

DOI: https://doi.org/10.21203/rs.3.rs-3909560/v1

Additional Declarations: No competing interests reported.

Abstract

Travel surveys typically collect detailed information about demographics and travel behavior of households and persons; but their sample sizes are often limited, and trip information is usually limited to a single day. In contrast, Mobile Device Location Data (MDLD) provides extensive and accurate trip records spanning multiple days for each person from a much larger sample, while demographic information for the individuals are always lacking due to anonymization. This study constructs data panels combining high-precision, long-term trip records from MDLD with detailed demographic information from a regional travel survey (RTS). Two probabilistic record linkage algorithms are employed to identify individuals with similar travel behaviors between RTS and MDLD datasets. The data panels constructed by the linkage algorithm captured not only peak-hour commutes but also off-peak travel and non-home-related trips, shedding light on previously underreported travel behaviors and offering a more holistic view of individuals' travel patterns. This comprehensive dataset also exhibits comparable demographic characteristics to the original RTS, showing that such data panel is a reasonable representation of the entire population. The integration of diverse datasets holds promise for revolutionizing travel behavior analysis and shaping the future of transportation planning in the era of mobile technology and big data.

1. INTRODUCTION

Urban transportation systems play a critical role in shaping the quality of life in cities, and understanding travel behaviors is essential for effective transportation planning and policy development. Traditionally, travel surveys have served as the primary data source for studying travel behaviors, providing valuable insights into commuting patterns and travel preferences (Nakamya et al., 2007). Typically, travel surveys gather data regarding individuals’ socio-economic and demographic information, as well as a journey diary for a specific day, which includes information about the starting and ending locations, times, mode of travel, companions, and purpose of each trip (Hong et al., 2021). However, these surveys have several limitations that can impact the accuracy and comprehensiveness of the data collected, such as reporting bias, where people may inaccurately recall or report all of their travel activities, leading to underreporting or misrepresentation of certain trips (Clarke et al., 1981). Non-response bias is another concern, as certain groups of people may be less likely to participate in the survey, potentially skewing the representation of travel behaviors (Richardson et al., 1996). Moreover, travel surveys often have a limited timeframe, capturing data for only a specific day or short period (Stopher et al., 2008), and they may rely on relatively small sample sizes of participants (Stopher et al., 2011), limiting the representation of travel behaviors. Therefore, to overcome these limitations and obtain a more comprehensive and precise understanding of travel behaviors, there is a growing need to explore alternative data sources that can complement and enhance the insights derived from travel surveys.

In recent years, the landscape of travel data collection and analysis has been dramatically transformed by the rise of mobile devices and cutting-edge data collection technologies. With the widespread adoption of GPS-enabled smartphones and other mobile gadgets, an invaluable resource known as Mobile Device

Location Data (MDLD) has become available (Yang et al., 2023). This data source captures the movements and locations of individuals with unprecedented precision and continuity throughout their daily routines (Hu et al., 2023). In contrast to conventional travel surveys, which rely on sporadic and self-reported data, MDLD distinguishes itself by accurately documenting the precise locations and movement trajectories of individuals throughout their daily routines (Ratti et al., 2006). This continuous influx of high-resolution location data provides an intricate portrayal of travel behaviors, showcasing origin and destination points, travel routes, and durations spent at specific sites. Through extended data collection periods, MDLD unveils the temporal dynamics of travel patterns, revealing differences between weekdays and weekends, peak and off-peak hours, and even seasonal variations in travel habits (Bachir et al., 2019). Additionally, MDLD illuminates recurring travel behaviors, exposing habitual routes and preferred modes of transportation among individuals, along with their frequently visited destinations (Ashbrook & Starner, 2003).

Despite its promise, MDLD also has its own set of limitations. One of the main challenges lies in the lack of explicit demographic information, which limits its standalone applicability for conducting comprehensive travel behavior analysis (Rojas IV et al., 2016). Demographic factors play a crucial role in shaping travel patterns and preferences. For instance, age and employment status may influence the frequency and purpose of trips, with younger individuals and workers likely having different travel behaviors compared to retirees or students (Su et al., 2020). Similarly, household size and income level can impact choices of transportation modes and travel distances (Amoh-Gyimah & Aidoo, 2013). By combining MDLD with traditional travel survey datasets like the Regional Travel Survey (RTS), researchers can bridge the gap and create a more robust and comprehensive dataset. At present, the research and application of the integration of survey data and location data mainly focus on two directions. One is GPS-based travel survey, where participants are asked to use wearable GPS devices (Hawkins & Stopher, 2004) or smartphone Apps (Safi et al., 2015) to record their travel activities. A significant advantage of GPS-based travel survey is that the demographic information of participants are still self-reported, which makes it more reliable and detailed. However, a limitation of GPS-based surveys is the potential for sampling biases, as they depend on participants willing to respond to the GPS survey, similar to biases in traditional travel surveys (Stopher & Greaves, 2007). The other direction is applying population synthesis to location data, wherein socio-demographics in the location data are matched to the marginal control totals in aggregated census data (Janzen, 2017). However, this also requires a sub-sample of location data with known demographics, which is often scarce and challenging to obtain (Bwambale et al., 2021).

This study proposes a novel approach to enhance travel behavior analysis through the integration of MDLD and RTS datasets based on record linkage algorithms. Two distinct data linkage approaches are utilized to connect individuals with similar travel behavior across datasets. The resulting data panels offer comprehensive and accurate representations of travel behaviors over multiple days, capturing not only peak-hour commutes but also off-peak travel and non-home-related trips. Additionally, they include representative population demographics, enhancing the overall richness and reliability of the dataset.

The primary objectives of this study are as follows:

1. To integrate MDLD and RTS datasets using two data linkage approaches: one based on a probabilistic method and the other on similarity-based techniques.
2. To construct data panels that offer a longitudinal perspective on individuals' travel behaviors, overcoming the limitations of traditional survey methods.
3. To evaluate the effectiveness of each data linkage approach in capturing accurate travel patterns and compare the characteristics of the integrated data panels with the original RTS and MDLD datasets.
4. To explore the representativeness of the data panels and discuss the implications of the findings for transportation planning and policy development.

In the following sections, we present the data used for data linkage, the methodology employed for panel construction, the results of the linking process, and a comprehensive analysis of the data panels, discussing the implications and potential applications of our findings for transportation planning and policy development.

2. DATA

In this study, two primary datasets are utilized to identify individuals with similar travel behaviors, and construct the data panel: trips collected in the Regional Travel Survey (RTS), and trips imputed from MDLD.

2.1. Regional Travel Survey (RTS)

The Regional Travel Survey (RTS) conducted by the Metropolitan Washington Council of Government (MWCOG) from October 2017 through December 2018 collected demographic and travel information from a randomly selected sample of households in the metropolitan Washington region. Key components in the RTS Dataset include household, person, trip, and vehicle files. Trip details, including origin/destination, start/end times, mode of travel, and trip purpose are recorded. Demographic information of individuals, such as age, gender, household size, income level, and employment status are also included. While the RTS dataset provides valuable insights into the travel behavior and demographics of the surveyed individuals, it also has some limitations, for example, the RTS relies on a relatively small sample of participants, and is limited to information gathered on a single day, which may not fully capture the variations in travel behavior that can occur throughout different days of the week or across different seasons. Moreover, participants in the RTS might not recall or report all their trips, leading to incomplete trip records. Despite these limitations, the RTS remains a valuable source of information for understanding travel behavior in the region.

2.2. Mobile Device Location Data (MDLD)

The Mobile Device Location Data utilized in this study comes from a consistent national mobility data panel produced by the University of Maryland (UMD) team from multiple data vendors. Generated from various positioning technologies such as cell phones, Global Positioning System (GPS), and location-based services (LBS), the MDLD comprises anonymous and accurate trip records from a much larger sample of individuals. Like RTS Dataset, MDLD captures trip information including start and end time, coordinates of the origin and destination, the imputed travel mode, and the imputed trip purpose. Unlike the RTS, MDLD provides trip information for multiple days, offering a longitudinal perspective on individuals' travel behaviors. Also, MDLD collects information from a much larger sample of individuals due to the widespread use of mobile devices. By harnessing MDLD, researchers can gain a broader understanding of travel patterns over time for a broader sample of the population. However, the dataset lacks detailed demographic information due to anonymization processes that ensure privacy and data protection. While the MDLD provides valuable insights into travel patterns, the absence of demographic data limits the ability to analyze the relationship between travel behavior and specific demographic factors.

By combining the RTS's detailed demographic information with MDLD's high-precision, long-term trip records, the research aims to create a unique dataset that can provide insights into travel behavior trends and their connection to demographics at a broader scale. Anna Arundel County in the State of Maryland is chosen as the case study in this paper. To ensure data accuracy and relevance, both datasets were filtered to include only individuals and devices residing within Anna Arundel County. Additionally, the analysis was limited to trips occurring solely within the county, excluding any trips departing from or entering the county from different regions. It should be noted that the methodology is scalable to regional, state or national level.

3. METHODOLOGY

Figure 3 **– 1** presents the methodological framework of this study. There are three major components: Data Preprocessing of RTS and MDLD, Identification of individuals with similar travel behaviors by two distinct approaches, and Construction of the data panel.

3.1. Data Preprocessing

The data record linkage requires that both datasets have the same structure and comparable attributes. In the case of this study, the RTS already includes person information such as the census tract of the home location, as well as trip details like census tract of origin/destination, start/end times, and more. However, the raw MDLD only contains location sightings with anonymized device identifiers (ID), latitude and longitude coordinates, time stamps, and positioning accuracies. To bridge this gap, a series of imputation algorithms are applied to derive the home location and trip information for each individual in the MDLD dataset (Zhang et al., 2021). A behavior-driven approach is utilized to identify devices’ home locations, which are imputed as the census tract where the device is observed with the highest count of nights, hours, and sightings during nighttime (Pan et al., 2023).

For extracting trip information from the sightings, a tour-based method is employed. The algorithm recognizes essential tours based on the daily life centers—such as a typical home-destination-home tour composed of multiple trips between the daily life centers and other stops—and further examines location observations within each tour to construct the complete chain of trips. Between the major activity locations, the point-to-point travel time, distance, and speed from the current location observation to the previous sighting and the next sighting are examined through a recursive algorithm to identify the stops and trips. Finally, a daily trip roster is generated, where for each trip, a unique trip identifier, start and end time, and coordinates of the origin and destination are included.

Additionally, the MDLD dataset may contain data from one or multiple recorded days for each device, while the RTS dataset records only one day for each person. To account for this difference, one day is carefully selected for each device in the MDLD data. Specifically, the focus is on including trips from the day with the highest number of recorded trips, as this day is deemed to best represent the person and allows for meaningful comparisons between the datasets.

3.2. Data linkage

In this research study, we present our methodology that builds upon the existing conventional approach for data record linkage. Figure 3 **– 2** visually represents the data record linkage framework:

The framework presented above encompasses the fundamental steps involved in any data record linkage model, whether it is exact matching or supervised/unsupervised probabilistic matching. These steps include indexing/blocking, similarity measurement, and linking decision. However, each step may vary in its implementation across different record linkage models. The selection of a specific model depends on various factors, such as the specific needs of the research, the characteristics of the data, the availability of previous knowledge, and whether partially true data is accessible for model training or not (Zhu, 2017).

After data processing step described in the previous section, these datasets are structured at the trip-unit level, where each device/person may have one or more records representing the total number of recorded trips. However, this data structure presents a unique challenge for the matching process, given our aim to link persons/devices. Most matching techniques are commonly applied to datasets involving entities with single records, such as health data records and other administrative data sources (Enamorado et al., 2019; Fleming et al., 2012; Sayers et al., 2016). To address the challenges posed by the trip-unit level data structure in the matching process, we have developed two distinct approaches, both based on probabilistic record linkage principles. In the following sections, before presenting the two approaches, we provide a more detailed explanation of each of the record linkage steps that were incorporated into both approaches.

3.2.1. Indexing/Blocking

The first step of data linkage is called indexing or blocking. This involves creating a list of potential candidate pairs between the two datasets for the comparison in the subsequent step. Initially, all possible pairs are formed into a matrix including all possible cases by exhaustively trying every entity from dataset A with every entity from dataset B. However, this exhaustive approach can result in significantly high computational time and reduce the algorithm's efficiency in finding matches. To address this issue, a more efficient approach is employed by eliminating many pairs through a process known as "blocking." Blocking involves grouping entities based on certain criteria, such as shared attributes or properties (Jaro, 1989). For example, candidate pairs may only be considered if they share the same zip code area. This way, the algorithm can reduce unnecessary comparisons and focus only on potentially relevant matches, thereby improving its efficiency. Blocking techniques are diverse and not constrained to merely grouping entities with identical attributes (Standard Blocking). In our study, we used Standard Blocking and we restricted our comparisons to individuals/devices residing in the same home tract area.

3.2.2. Field Comparison

Following the indexing/blocking process, potential candidate pairs are compared for similarity using selected attributes called "matching variables." These variables are common to both datasets, and the comparison is done individually for each one. The results range from 0 to 1, with 0 indicating no match, 1 indicating a perfect match, and any value in between representing the degree of similarity. The similarity depends on the type of variables compared, such as string values, numeric, date, time, or location. This study focuses on time and geographic variables. Distances between departure and arrival times are computed in minutes, while distances between departure and arrival locations are computed in kilometers for each candidate pair. These values represent distances and are not similarity scores yet. To convert distances into similarity scores ranging from 0 to 1, a linear decay function is used (de Bruin, 2022). The process involves two tuning parameters: "offset" and "scale." If a distance falls within the specified range, it gets a similarity score of 1. However, if the distance value exceeds the offset, the score starts to linearly decay from 1 to 0 at a rate of 1/scale (Fig. <link rid="fig2">3</link>–3).

The selection of tuning parameters in this study is driven by the desired accuracy level and the allowable error defined by the user. For this investigation, the tuning parameters "offset" and "scale" were set to 15 minutes and 25 minutes, respectively, for departure and arrival times. Additionally, for departure and arrival locations, the tuning parameters "offset" and "scale" were configured to 1 km and 3 km, respectively. We made these choices to the best of our ability. We believe that the choices meet the specific accuracy requirements and permissible errors outlined for the study.

3.2.3. Approach 1

The first approach uses a probabilistic method based on the Fellegi and Sunter theorem (Fellegi & Sunter, 1969) for record linkage. We adapt this method to the trip-unit level, which helps identify matching trips among the datasets. All the previous steps were applied at the trip-unit level to compute similarity scores across potential candidates. However, it is essential to note that these are similarity scores, not probabilities. This is why we have chosen to use the Fellegi and Sunter method to convert these similarity scores into probabilities of matching. To convert the similarity score into a match probability, Bayes' Theorem is applied. To do this, we require the following information:

- P-Probability (Overall): This is the prior probability, which represents the expected proportion of duplicates in a dataset before any comparisons are made between records.
- M-Probability (for each variable): P(Variable Matches | Record is a Match) is the probability that two records with the same entity agree on the linkage variable.
- U-Probability (for each variable): P(Variable does not match | Record is not a match) is the probability that two records with different entities agree on the linkage variable.

Unfortunately, none of this information is available in our case, as we lack ground truth data or training data. Therefore, in approach 1, we estimated these parameters for each matching variable using an unsupervised algorithm facilitated by the expectation/maximization algorithm (Bauman, 2006). In the decision step, at the trip-unit level, matched trips are accepted if their similarity score exceeds the specified threshold value of 0.5.

To find matches/links between devices and persons, we developed an algorithm that leverages unique matches to establish connections between individuals or devices. Since the matching at the trip level can result in multiple possible matches for each device/person, the algorithm searches for the most unique connections, assuming that the most unique links are the more reasonable matches.

3.2.4. Approach 2

The second approach is based on a similarity-based method. In this approach, we restructured the dataset, transforming it from the trip-unit level to the person/device level. This involved aggregating all trip information pertaining to each individual or device into a single trip itinerary. Each trip itinerary is represented as a vector of one or more trips, with each trip including selected matching variables as departure and arrival times, as well as departure and arrival locations in longitudinal and latitudinal format. By restructuring the data in this way, we create a more manageable dataset that facilitates the application of similarity measures. To measure the similarity between any candidate pairs, we have chosen to use the dynamic time warping algorithm (Muller, 2007). This algorithm calculates the distance between the two candidate pairs, taking into account variations in the time and location axes. The resulting distance is then converted into a similarity score using a linear decay function, similar to the tuning parameter used in approach 1. The dynamic time warping technique is particularly effective in capturing patterns and similarities in time-series data, making it well-suited for our trip-based datasets. This method has been widely used in various data trajectory similarity studies (Gong et al., 2020; Hung et al., 2015; Sun et al., 2017). In the decision step, we determine the acceptance of matched devices based on a similarity score. Devices are linked if their similarity score is equal to the maximum value and greater than or equal to a specified threshold value. For this study, the threshold value was set to 0.5, meaning that candidate pairs with a similarity score of 0.5 or higher are accepted as matches. By leveraging trip itineraries, we achieved direct linking of devices to individuals without the need for the uniqueness algorithm utilized in approach 1.

For indexing/blocking and estimating probabilities (P, M, and U) in both approaches, we utilized the 'recordlinkage' package in Python (de Bruin, 2022). By incorporating these two complementary approaches, the challenges posed by the original trip-unit level data structure can be effectively addressed.

3.3. Data panel construction

Data panels are constructed for travelers linked by approach 1 and approach 2, respectively. The trip records in the data panel are sourced from MDLD, which continuously captures trip information over multiple days without reporting bias, providing a longitudinal perspective on individuals' travel behaviors. The demographic information, including age, gender, household size, income level, and employment status, is obtained from RTS, where detailed and reliable demographic data is reported. By integrating these two data sources, we combine geospatial accuracy with detailed demographic information, addressing the limitations of both datasets and enabling a more nuanced analysis.

In the next section, we will present and compare the different characteristics of the four datasets: the original RTS, MDLD, and the panel data constructed with travelers linked by approach 1 and approach 2. We will also discuss and analyze the effectiveness of the two linking algorithms and the representativeness of the panel data.

4. RESULTS

4.1. Sample size, timespan, and trip rate of the data panel

Table 4 – 1 presents a summary of key statistics, including sample size, timespan, and trip rate, for different datasets: RTS, MDLD, travelers linked by approach 1, and travelers linked by approach 2. Note that the linkage process involved utilizing RTS and a representative day of MDLD to identify linked travelers. Subsequently, the travel patterns of these linked travelers were analyzed for the entire month to which the observed week belonged. This approach simulates the construction of such data panel, allowing for longer-term tracking of travel behaviors using MDLD data.

As shown in Table 4 – 1, MDLD originally covers a significantly larger sample size compared to RTS, with 23,224 travelers observed within just one week. This number is over 16 times greater than the total number of travelers in RTS, which includes 1,383 individuals. The substantial sample size in MDLD enhances the likelihood of finding travelers with similar travel behaviors in both datasets. Consequently, approach 1 successfully matched 617 travelers from MDLD with RTS, while approach 2 achieved 977 matches. These numbers represent approximately 44.6% and 70.6% of the travelers in RTS, respectively. This suggests that incorporating trip records from more days in MDLD could further increase the number of travelers matched with RTS.

Table 4
– 1. Sample size, timespan, and trip rate comparison

	RTS	MDLD	Linked_approach1	Linked_approach2
Number of travelers	1,383	23,224	617	977
Median number of days in a month	1	10	10	13
Mean number of days in a month	1	11.4	10.89	13.14
Median number of longest consecutive days in a month	1	4	3	5
Mean number of longest consecutive days in a month	1	5.413	4.843	6.273
Median trips per day	2	2	2	2.5
Mean trips per day	2.837	2.772	2.215	3.140

One notable distinction is that RTS collects trip records for each responder on a single travel day, while MDLD provides trip records for an average of 11.4 days per traveler in a month. On average, the longest consecutive travel days observed for each MDLD traveler is 5.413 days. This highlights one of the advantages of MDLD, as it captures daily, weekly, and seasonal variations in travel behaviors due to its multiple-day coverage. For travelers linked by approach 1, the average number of days is 10.89, and the average longest consecutive days is 4.843, both of which are less than those in the entire MDLD dataset. Conversely, travelers linked by approach 2 show averages of 13.14 days and 6.273 longest consecutive days, both exceeding the entire MDLD dataset. This indicates that approach 1 tends to link travelers with shorter timespans in MDLD to RTS, while approach 2 is more capable of identifying travelers appearing over more days.

Regarding the trip rate, travelers in RTS are recorded to have an average of 2.837 trips per day, which closely aligns with the average of 2.772 trips per day made by travelers in MDLD. It should be noted that these numbers are slightly lower than the typical trip rate found in other datasets. This discrepancy can be attributed to our exclusion, as mentioned in Section 2, of any trips departing from or entering the county from different regions. This filtering step was applied to ensure the accuracy and relevance of the data. Upon examining the travelers linked by approach 1, an average of 2.215 trips per day was observed. Conversely, travelers linked by approach 2 demonstrated an average of 3.140 trips per day. This suggests that approach 1 is more inclined to identify individuals who undertake fewer trips, whereas approach 2 is better suited for detecting more active travelers. Considering that approach 2 can also identify travelers appearing in more days, it is reasonable to assume that the data panel constructed based on the linking results from approach 2 will include a higher proportion of active travelers.

4.2. Trip characteristics

4.2.1 Trip duration

Figure 4 **– 1** illustrates the distribution of trip durations across the four datasets, with dashed lines representing the respective means. The average trip duration in RTS is 14.8 minutes, closely followed by MDLD with an average of 14.9 minutes, making the red and green dashed lines nearly indistinguishable. However, variations are observed among linked travelers. Those linked by approach 1 display an average trip duration of 17.56 minutes, indicating that although they make fewer trips per day, their individual travel times are longer. Conversely, travelers linked by approach 2 exhibit an average trip duration of 13.00 minutes, indicating a tendency for more frequent short trips per day. This pattern is consistent with the density curves, where the share of short-duration trips (less than 15 minutes) for travelers linked by approach 1 (represented by the blue curve) is significantly lower than that for travelers linked by approach 2 (represented by the purple curve).

Furthermore, Fig. 4 – 1 reveals that trip durations in RTS cluster around multiples of five minutes (e.g., 5, 10, 15...), indicative of "reporting bias" or "rounding bias" commonly encountered in surveys (Rietveld, 2002). When respondents are asked to report specific trip details, such as start and end times, they tend to round or approximate their responses to convenient intervals, like 5-minute increments. In contrast, the density curves of trip duration in the other three MDLD-based datasets appear smoother. This is due to the passive collection of travel information in MDLD, which does not rely on self-reported data from survey respondents. Consequently, reporting bias is significantly reduced or eliminated in MDLD, making it a valuable and reliable resource for studying and understanding travel patterns and behaviors.

4.2.2 Time of day and day of week

The departure hour and arrival hour distributions of recorded trips are presented in Fig. 4 **– 2(a)** and Fig. 4 **– 2(b)**, respectively. Across all four datasets, a consistent trend is observed in the temporal distribution, with significantly higher traffic volume during peak periods (7:00–9:00 and 15:00–18:00) compared to other times of the day. However, a noticeable discrepancy exists between RTS and the other three MDLD-based datasets, with RTS reporting fewer trips during off-peak hours, particularly during the night (21:00–5:00). This reveals the presence of off-peak travel underreporting in travel surveys, wherein individuals tend to report fewer off-peak trips compared to what is actually revealed in MDLD. This phenomenon has also been reported in other studies (Chapleau et al., 2018).

The reasons for off-peak travel underreporting in RTS may be twofold: Firstly, respondents may have difficulty accurately recalling and reporting off-peak trips, especially if they occur infrequently and are not as memorable as peak-hour travel. Secondly, compared to peak-hour trips, respondents might perceive off-peak trips as less relevant or important, especially if they are part of continuous activities or involve short distances, leading to their omission from the reported trips.

This underreporting of off-peak travel in RTS can lead to an imbalanced representation of travel behavior, resulting in an underestimation of off-peak travel patterns. Consequently, RTS may not fully capture the true extent and nature of off-peak travel activities. The higher proportion of trips during midday and late night in MDLD-based datasets indicates that MDLD can be advantageous in addressing this bias. By providing continuous tracking of travel behavior, MDLD offers a more comprehensive and accurate representation of all travel activities, regardless of the time of day.

Regarding the comparison between travelers linked by the two approaches, it is evident that travelers linked by approach 2 exhibit a relatively higher share of off-peak trips and a lower share of peak trips compared to travelers linked by approach 1. This suggests that approach 1 tends to identify individuals following a more regular timetable throughout the day, while approach 2 captures a broader range of travel patterns, including more off-peak travel activities.

Figure 4 **– 3** presents the distribution of trips by day of the week. As RTS only includes weekdays as travel days, Fig. 4 **– 3(a)** displays the relative frequencies of trips during weekdays for the MDLD-based datasets to facilitate a meaningful comparison. Meanwhile, Fig. 4 **– 3(b)** illustrates the distribution of trips for all seven days of the week in the three MDLD-based datasets. This distinction highlights one of the major contributions of MDLD, as it overcomes the limitations of self-reported surveys by providing continuous tracking of travel activities.

From Fig. 4 **– 3(a)**, it is evident that RTS records a higher proportion of trips on Monday and Wednesday compared to MDLD, whereas MDLD records a higher proportion of trips on Thursday and Friday compared to RTS. Figure 4 **– 3(b)** reveals that travelers linked by approach 1 make slightly more trips during weekdays and fewer trips during weekends compared to travelers linked by approach 2. This finding suggests that approach 1 may tend to link more regular commuters, who primarily travel on weekdays, while approach 2 captures a broader range of travel patterns, including individuals who travel on both weekdays and weekends.

4.2.3 Home-based trips and heat map of trip destinations

The analysis highlights the variation in the proportion of home-based trips, as depicted in Fig. <link rid="fig6">4</link>–4, which underscores the disparities in travel patterns captured by the RTS and MDLD datasets. In RTS, the proportion of home-based trips is recorded at 81.67%, significantly higher than the three MDLD-based datasets, where home-based trips constitute 61%-62% of the total. This pattern corroborates the discussion in section 4.2.2, indicating that the design and limitations of the RTS data lead to a primary emphasis on trips originating from home and then returning home, which are common and essential travel patterns. As a result, other types of trips, especially those made during off-peak hours for non-home-related purposes, are not adequately represented in the RTS dataset. This omission leads to an incomplete picture of overall transit use and may result in an underestimation of traffic volume during off-peak periods. Therefore, by integrating MDLD into the analysis, researchers can gain valuable insights into the full spectrum of travel patterns, enabling more accurate and inclusive assessments of traffic volume during both peak and off-peak hours. The proportions of home-based trips for travelers linked by approach 1 and approach 2 are both similar to the proportion observed in the entire MDLD dataset. This indicates that the linkage algorithm employed in the study does not introduce bias in terms of home-based trips.

Figure 4–5 presents heat maps of trip destinations from the four datasets, with a focus on trips within Anna Arundel County as mentioned previously. The spatial distribution of trip destinations is found to be generally consistent across all four datasets. Annapolis, the county's largest city, emerges as a prominent hotspot, along with other densely populated towns like Glen Burnie and Crofton. This observation reinforces the reliability and validity of the MDLD data and the linkage approaches employed in the study, as the trip destinations exhibit a reasonable and coherent spatial pattern.

However, there is still a noticeable discrepancy between the heat maps of RTS and MDLD. In RTS, trip destinations appear to be more discrete, with certain rural areas showing no trip records (evident in the bottom left area of the map). Conversely, the MDLD-based datasets exhibit a more widely distributed pattern, covering nearly all areas within Anna Arundel County. This difference may be attributed to the fact that RTS has a smaller sample size compared to MDLD, with trip records collected for a single day for each participant. Consequently, trips made in some rural areas are likely underreported in RTS. MDLD, on the other hand, emerges as a promising data source to address this limitation. By providing continuous movement records over multiple days and encompassing a larger sample size, MDLD offers a more comprehensive representation of travel behavior. This comprehensive coverage ensures that trips made in various areas, including rural regions, are adequately captured, leading to a more accurate depiction of spatial travel patterns within the county.

4.3. Socio-Demographics

This section presents the results of socio-demographics for the individuals residing within Anna Arundel County in RTS, individuals linked by approach 1, and individuals by approach 2.

Figure 4–6 illustrates the age distribution across the three datasets, with dashed lines representing the respective means. The average age in RTS is 43.34 years old, which matches the average age of individuals linked by approach 1. Individuals linked by approach 2 exhibit a slightly older average age of 44.13 years old. While there are some slight differences in age distribution among the three datasets, such as more individuals between 45–60 years old and fewer individuals between 25–45 years old in individuals linked by approach 1, the Kolmogorov-Smirnov Test reveals that the p-values for comparing the entire RTS with individuals linked by approach 1 and approach 2 are 0.7851 and 0.9155, respectively. Both p-values are significantly greater than 0.05, indicating insufficient evidence to reject the null hypothesis that these datasets do not come from the same age distribution. Figure 4–7 displays the gender distributions across the three datasets, showing similar proportions with approximately 52% of individuals being female and 48% being male in all datasets. Figure 4–8 shows the race distributions, with the highest percentage of individuals being white, accounting for nearly 80% across all datasets. The distribution of employment status is presented in Fig. 4–9, demonstrating comparable patterns among the three datasets. More than 50% of individuals in all datasets are workers. It is notable that the percentage of workers among individuals linked by approach 1 and approach 2 is higher than in the overall RTS. This finding may be attributed to the fact that workers generally make more trips than non-workers, increasing their likelihood of being linked with MDLD, given that the linkage algorithms are based on trip characteristics.

In conclusion, the distribution of various socio-demographic attributes across the three datasets is generally similar, suggesting that the data panel constructed by the linkage algorithm exhibits comparable demographic characteristics to the original RTS. Thus, it can be considered as representative of the population, similar to the RTS dataset itself. The consistency in socio-demographic profiles enhances the validity and applicability of the data panel for studying travel behavior and making informed decisions in transportation research and planning.

5. CONCLUSION AND DISCUSSION

In this study, we proposed and implemented a novel approach to enhance travel behavior analysis by integrating Mobile Device Location Data (MDLD) with the Regional Travel Survey (RTS) datasets. Our goal was to address the limitations of traditional travel surveys and leverage the advantages of MDLD to create more comprehensive and accurate representations of individuals' travel behaviors. Two distinct data linkage approaches were utilized to connect individuals across datasets with similar travel characteristics, resulting in the construction of data panels that provide valuable insights into travel patterns.

MDLD offered continuous and high-resolution information on individuals' movements over multiple days, providing a detailed and accurate picture of daily, weekly, and even seasonal travel activities. By combining this geospatially accurate data with the rich demographic information from RTS, our data panels provided comprehensive and reliable representations of travel behaviors. Our approach captured not only peak-hour commutes but also off-peak travel and non-home-related trips, shedding light on previously underreported travel behaviors and offering a more holistic view of individuals' travel patterns. This comprehensive dataset also allowed us to analyze travel behaviors over extended periods, providing a longitudinal perspective on travel habits that was challenging to achieve with traditional survey data.

While our approach showed significant promise, it is essential to acknowledge some limitations. While MDLD provides continuous location data, there may still be gaps in the dataset due to factors such as poor network coverage, battery depletion, or device malfunctions. The effectiveness and accuracy of the home location identification and trip imputation algorithm also needs further validation. The data linkage process also presents challenges, particularly in dealing with multiple possible matches at the trip level. Our algorithm addressed this by leveraging unique matches to establish connections between individuals or devices. While this approach was effective, it is crucial to further explore and refine data linkage methods to ensure accuracy and reliability in large-scale applications.

Despite these challenges, the findings of our study have significant implications for transportation planning and policy-making. By understanding travel behaviors across various time periods and demographic groups, policymakers can tailor transportation services to meet the specific needs of individuals and communities, and cities can develop targeted interventions to enhance mobility and reduce congestion, ultimately contributing to a more livable and sustainable urban environment. As data collection technologies continue to advance, the integration of diverse datasets holds promise for revolutionizing travel behavior analysis and shaping the future of transportation planning in the era of mobile technology and big data.

Declarations

AUTHOR CONTRIBUTION STATEMENT

The authors confirm contribution to the paper as follows: study conception and design: G.Z, M.B.A, T.T, C.C.; data collection: G.Z, M.B.A; analysis and interpretation of results: G.Z, M.B.A, T.T, C.C.; draft manuscript preparation: G.Z, M.B.A, C.C.. All authors reviewed the results and approved the final version of the manuscript.

Author Contribution

Study conception and design: G.Z, M.B.A, T.T, C.C.; data collection: G.Z, M.B.A; analysis and interpretation of results: G.Z, M.B.A, T.T, C.C.; draft manuscript preparation: G.Z, M.B.A, C.C.. All authors reviewed the results and approved the final version of the manuscript.

ACKNOWLEDGEMENTS

This study was conducted as part of a collaboration among the U.S. Department of Transportation's Federal Highway Administration, Center for Advanced Transportation Technology Laboratory (CATT), and Maryland Transportation Institute (MTI) at the University of Maryland College Park.

References

1. Amoh-Gyimah, R., & Aidoo, E. N. (2013). Mode of transport to work by government employees in the Kumasi metropolis, Ghana. *Journal of Transport Geography*, *31*, 35-43.
2. Ashbrook, D., & Starner, T. (2003). Using GPS to learn significant locations and predict movement across multiple users. *Personal and Ubiquitous computing*, *7*, 275-286.
3. Bachir, D., Khodabandelou, G., Gauthier, V., El Yacoubi, M., & Puchinger, J. (2019). Inferring dynamic origin-destination flows by transport mode using mobile phone data. *Transportation Research Part C: Emerging Technologies*, *101*, 254-275.
4. Bauman, G. J. (2006). *Computation of weights for probabilistic record linkage using the EM algorithm*. Brigham Young University.
5. Bwambale, A., Choudhury, C. F., Hess, S., & Iqbal, M. S. (2021). Getting the best of both worlds: a framework for combining disaggregate travel survey data and aggregate mobile phone data for trip generation modelling. *Transportation*, *48*, 2287-2314.

6. Chapleau, R., Gaudette, P., & Spurr, T. (2018). Strict and deep comparison of revealed transit trip structure between computer-assisted telephone interview household travel survey and smart cards. *Transportation research record*, *2672*(42), 13-22.
7. Clarke, M., Dix, M., & Jones, P. (1981). Error and uncertainty in travel surveys. *Transportation*, *10*(2), 105-126.
8. de Bruin, J. (2022). Record Linkage Toolkit Documentation.
9. Enamorado, T., Fifield, B., & Imai, K. (2019). Using a probabilistic model to assist merging of large-scale administrative records. *American Political Science Review*, *113*(2), 353-371.
10. Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, *64*(328), 1183-1210.
11. Fleming, M., Kirby, B., & Penny, K. I. (2012). Record linkage in Scotland and its applications to health research. *Journal of clinical nursing*, *21*(19pt20), 2711-2721.
12. Gong, X., Huang, Z., Wang, Y., Wu, L., & Liu, Y. (2020). High-performance spatiotemporal trajectory matching across heterogeneous data sources. *Future Generation Computer Systems*, *105*, 148-161.
13. Hawkins, R., & Stopher, P. R. (2004). Collecting data with GPS: Those who reject, and those who receive.
14. Hong, S., Zhao, F., Livshits, V., Gershenfeld, S., Santos, J., & Ben-Akiva, M. (2021). Insights on data quality from a large-scale application of smartphone-based travel survey technology in the Phoenix metropolitan area, Arizona, USA. *Transportation Research Part A: Policy and Practice*, *154*, 413-429.
15. Hu, S., Xiong, C., Chen, P., & Schonfeld, P. (2023). Examining nonlinearity in population inflow estimation using big data: An empirical comparison of explainable machine learning models. *Transportation Research Part A: Policy and Practice*, *174*, 103743.
16. Hung, C.-C., Peng, W.-C., & Lee, W.-C. (2015). Clustering and aggregating clues of trajectories for mining trajectory patterns and routes. *The VLDB Journal*, *24*, 169-192.
17. Janzen, M. (2017). Population synthesis for long-distance travel demand simulations. 6th symposium of the European association for research in transportation (hEART 2017),
18. Jaro, M. A. (1989). Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. *Journal of the American Statistical Association*, *84*(406), 414-420.
19. Muller, M. (2007). Dynamic time warping in information retrieval for music and motion. *Dynamic time warping Information retrieval for music and motion*, 69-84.
20. Nakamya, J., Moons, E., Koelet, S., & Wets, G. (2007). Impact of data integration on some important travel behavior indicators. *Transportation research record*, *1993*(1), 89-94.
21. Pan, Y., Sun, Q., Yang, M., Darzi, A., Zhao, G., Kabiri, A., . . . Zhang, L. (2023). Residency and worker status identification based on mobile device location data. *Transportation Research Part C: Emerging Technologies*, *146*, 103956.
22. Ratti, C., Frenchman, D., Pulselli, R. M., & Williams, S. (2006). Mobile landscapes: using location data from cell phones for urban analysis. *Environment and planning B: Planning and design*, *33*(5), 727-

748.

23. Richardson, A., Ampt, E., & Meyburg, A. (1996). Nonresponse issues in household travel surveys. Conference proceedings,
24. Rietveld, P. (2002). Rounding of arrival and departure times in travel surveys: an interpretation in terms of scheduled activities. *Journal of transportation and statistics*, *5*(1), 71-82.
25. Rojas IV, M. B., Sadeghvaziri, E., & Jin, X. (2016). Comprehensive review of travel behavior and mobility pattern studies that used mobile phone data. *Transportation Research Record*, *2563*(1), 71-79.
26. Safi, H., Assemi, B., Mesbah, M., Ferreira, L., & Hickman, M. (2015). Design and implementation of a smartphone-based travel survey. *Transportation Research Record*, *2526*(1), 99-107.
27. Sayers, A., Ben-Shlomo, Y., Blom, A. W., & Steele, F. (2016). Probabilistic record linkage. *International journal of epidemiology*, *45*(3), 954-964.
28. Stopher, P., Zhang, Y., Armoogum, J., & Madre, J.-L. (2011). National household travel surveys: The case for Australia. 34th Australasian Transport Research Forum (ATRF), Adelaide, South Australia,
29. Stopher, P. R., & Greaves, S. P. (2007). Household travel surveys: Where are we going? *Transportation Research Part A: Policy and Practice*, *41*(5), 367-381.
30. Stopher, P. R., Kockelman, K., Greaves, S. P., & Clifford, E. (2008). Reducing burden and sample sizes in multiday household travel surveys. *Transportation Research Record*, *2064*(1), 12-18.
31. Su, R., McBride, E. C., & Goulias, K. G. (2020). Pattern recognition of daily activity patterns using human mobility motifs and sequence analysis. *Transportation Research Part C: Emerging Technologies*, *120*, 102796.
32. Sun, L., Zhou, W., Jiang, B., & Guan, J. (2017). A real-time similarity measure model for multi-source trajectories. 2017 International Conference on Computing Intelligence and Information System (CIIS),
33. Yang, M., Luo, W., Ashoori, M., Mahmoudi, J., Xiong, C., Lu, J., . . . Kabiri, A. (2023). Big-Data Driven Framework to Estimate Vehicle Volume Based on Mobile Device Location Data. *Transportation Research Record*, 03611981231174240.
34. Zhang, L., Darzi, A., Pan, Y., Yang, M., Sun, Q., Kabiri, A., . . . Xiong, C. (2021). Next generation National Household Travel Survey National Origin Destination Data Passenger Origin-Destination Data Methodology Documentation. *Federal Highway Administration.[Google Scholar]*.
35. Zhu, S. (2017). *Integration of commercial vehicle GPS and roadside intercept survey data*. University of Toronto (Canada).

Figures

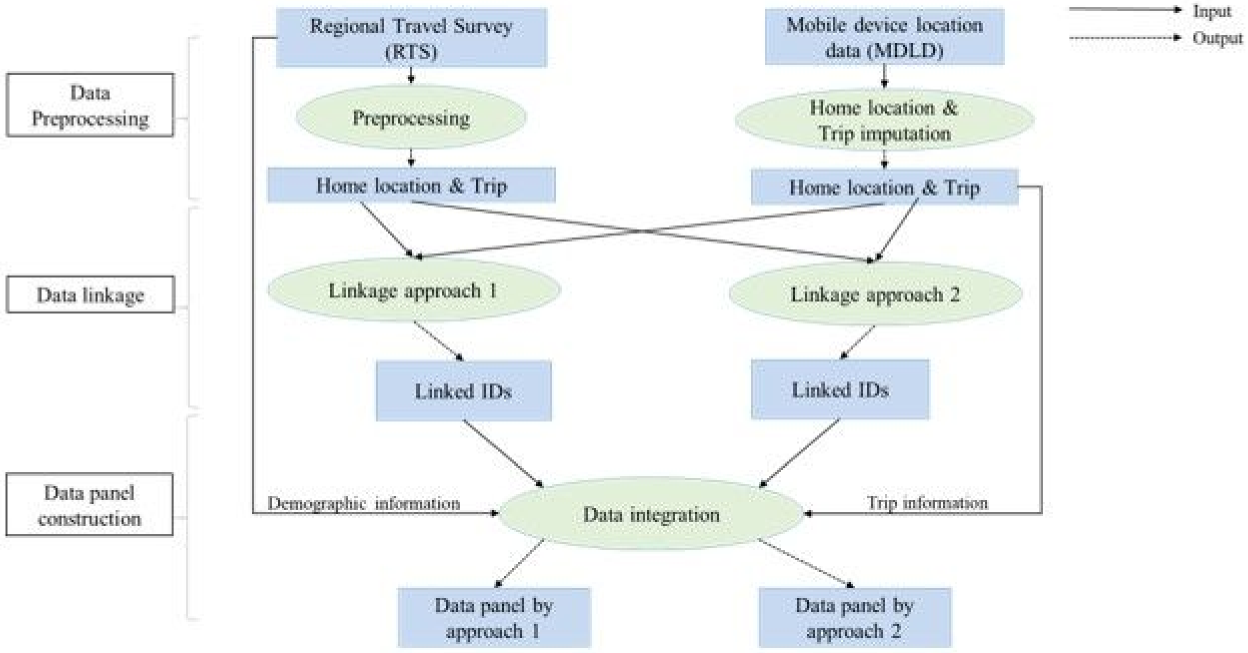


Figure 1

Figure 3-1. Methodological Framework

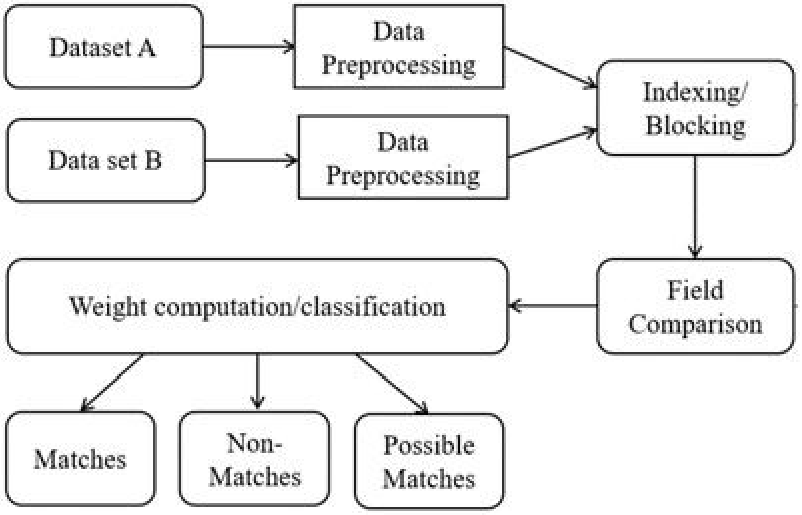


Figure 2

Figure 3-2. Data Record Linkage Framework

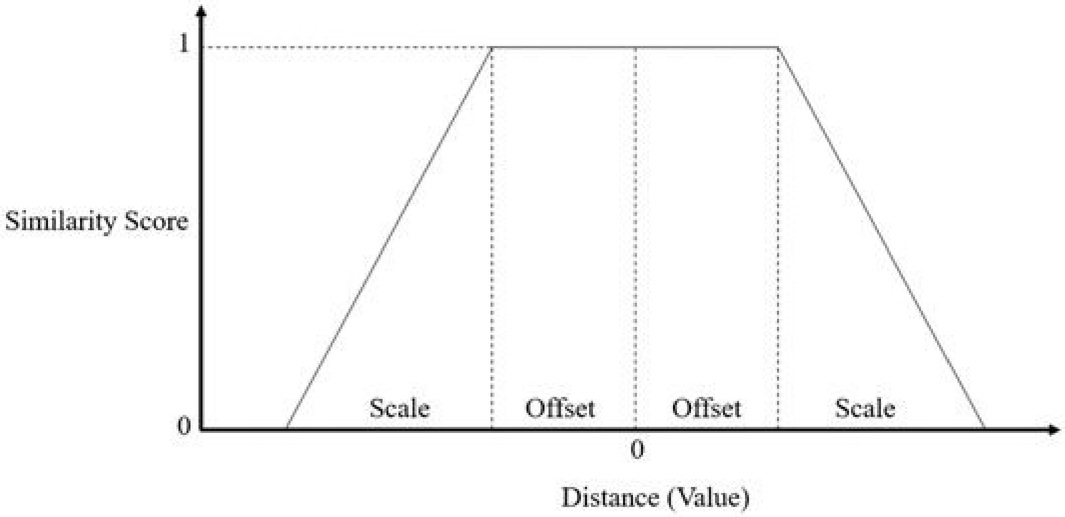


Figure 3

Figure 3-3. Conversion of Distances to Similarity Scores using Linear Decay Function

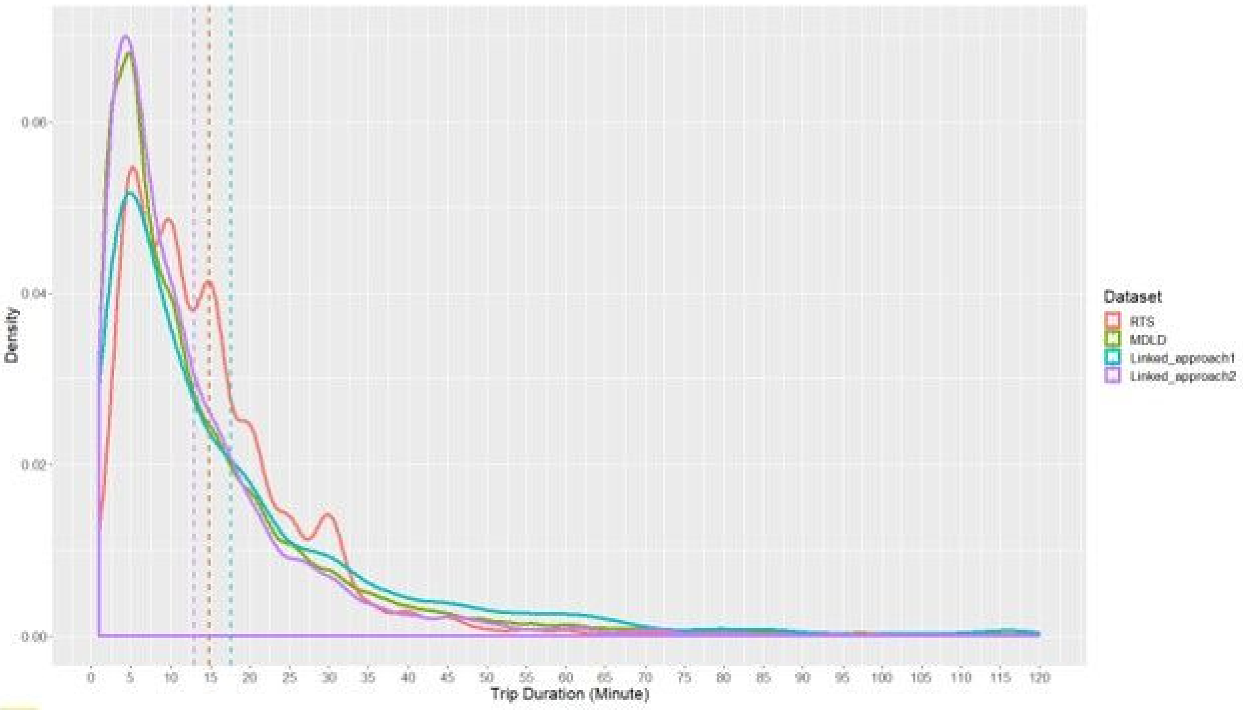
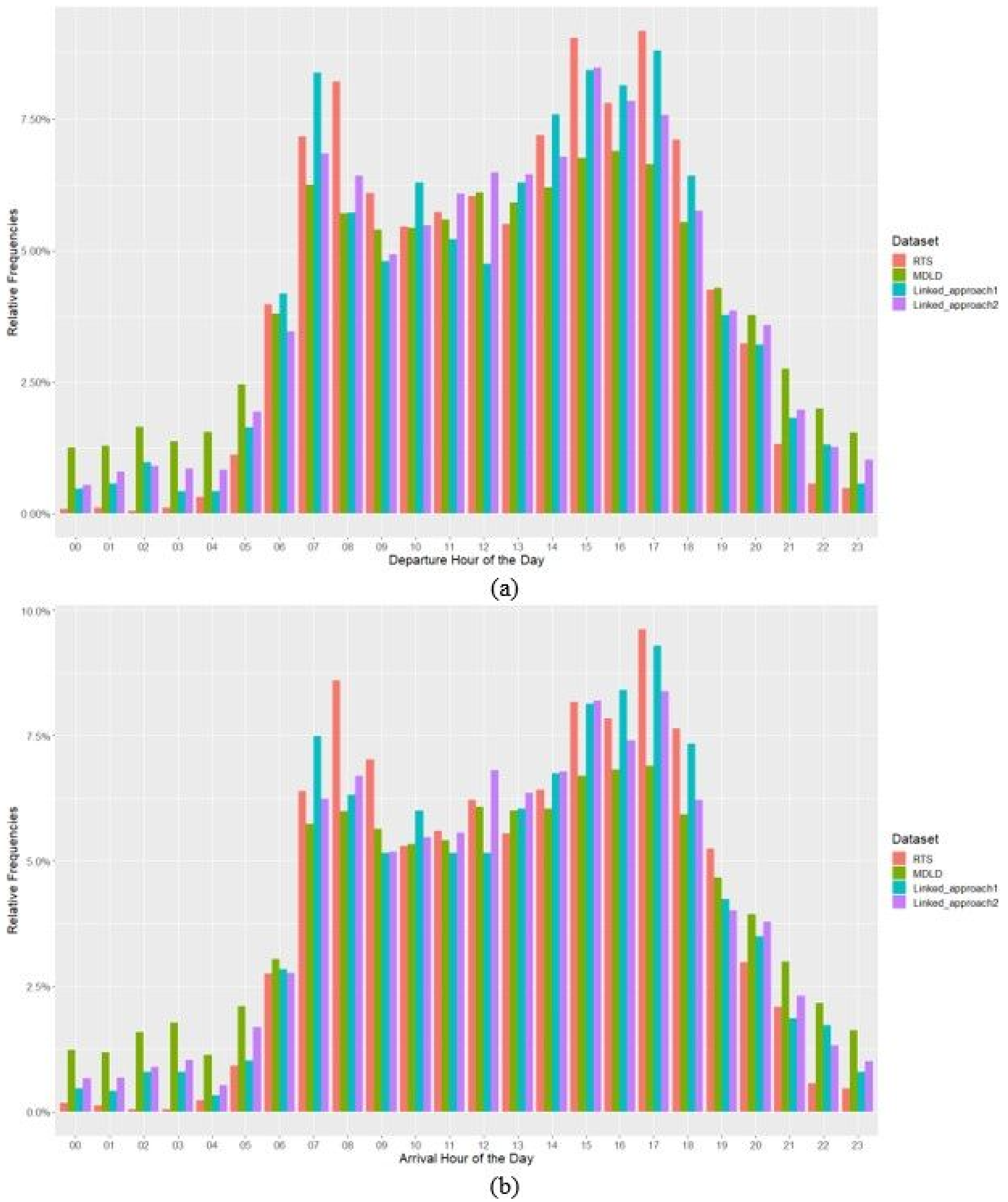


Figure 4

Figure 4-1. Density of trip duration in minute



(a)

(b)

Figure 5

Figure 4-2. Distribution of trips by the hour of day for (a) the departure time, and (b) the arrival time

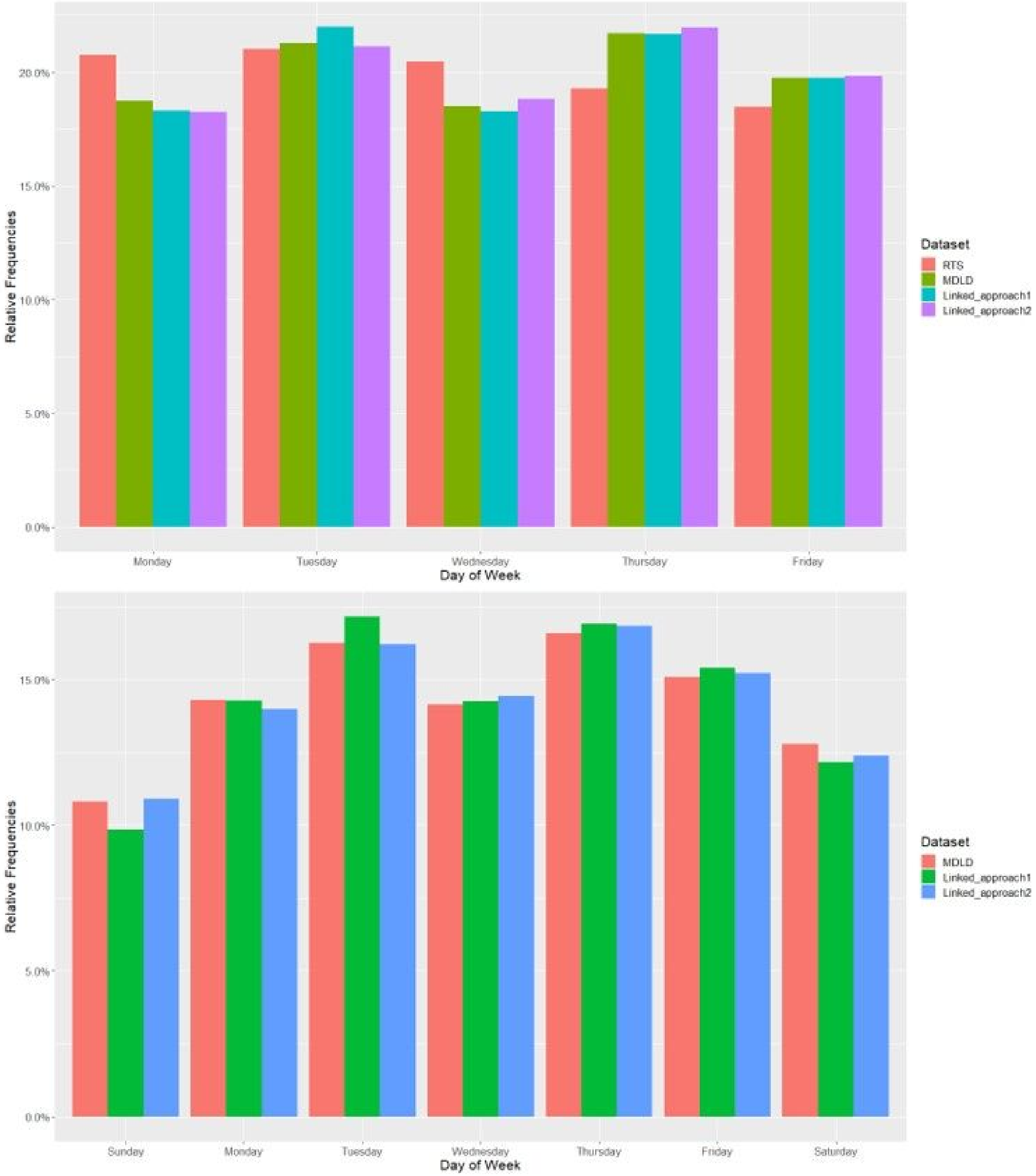


Figure 6

Figure 4-3. Distribution of trips by (a) weekdays, and (b) weekdays and weekends

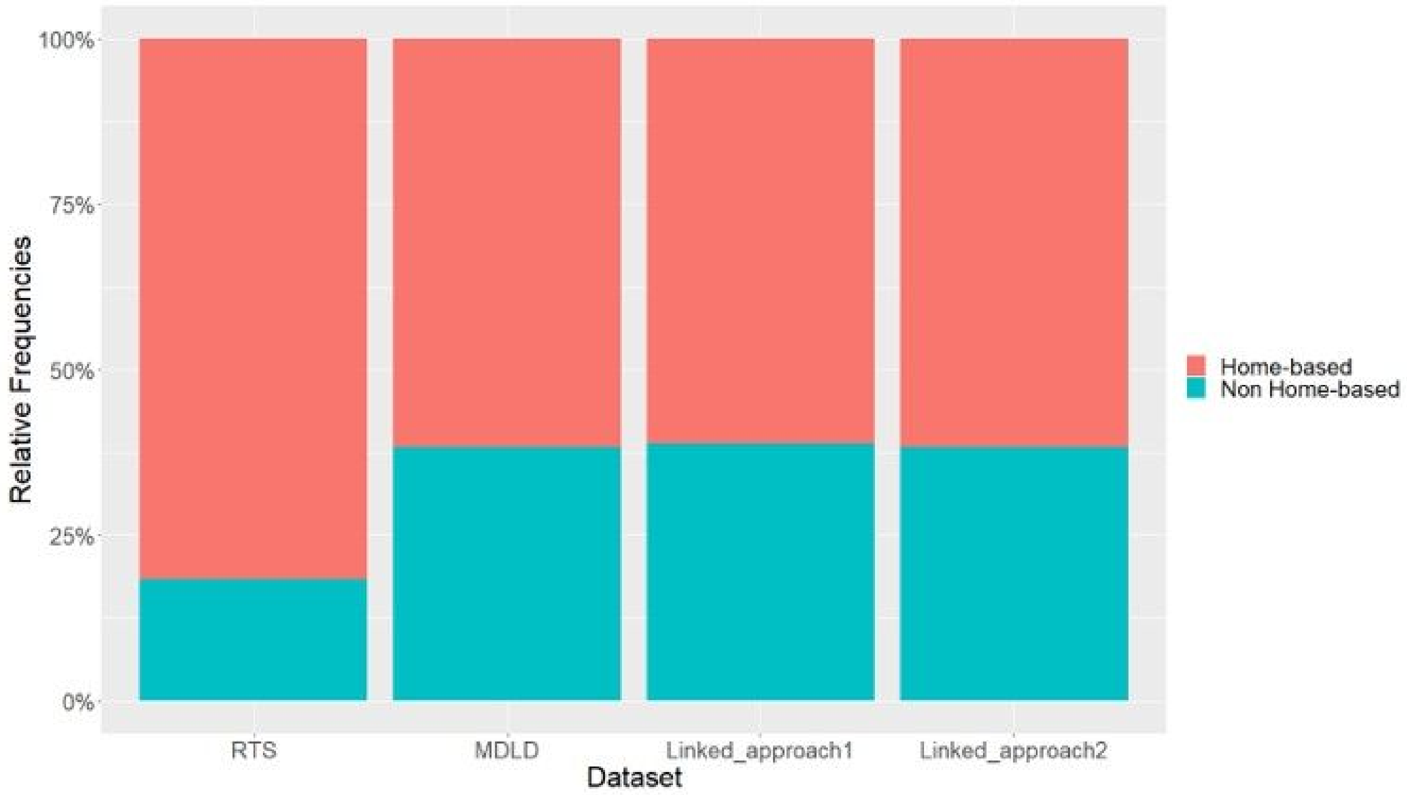
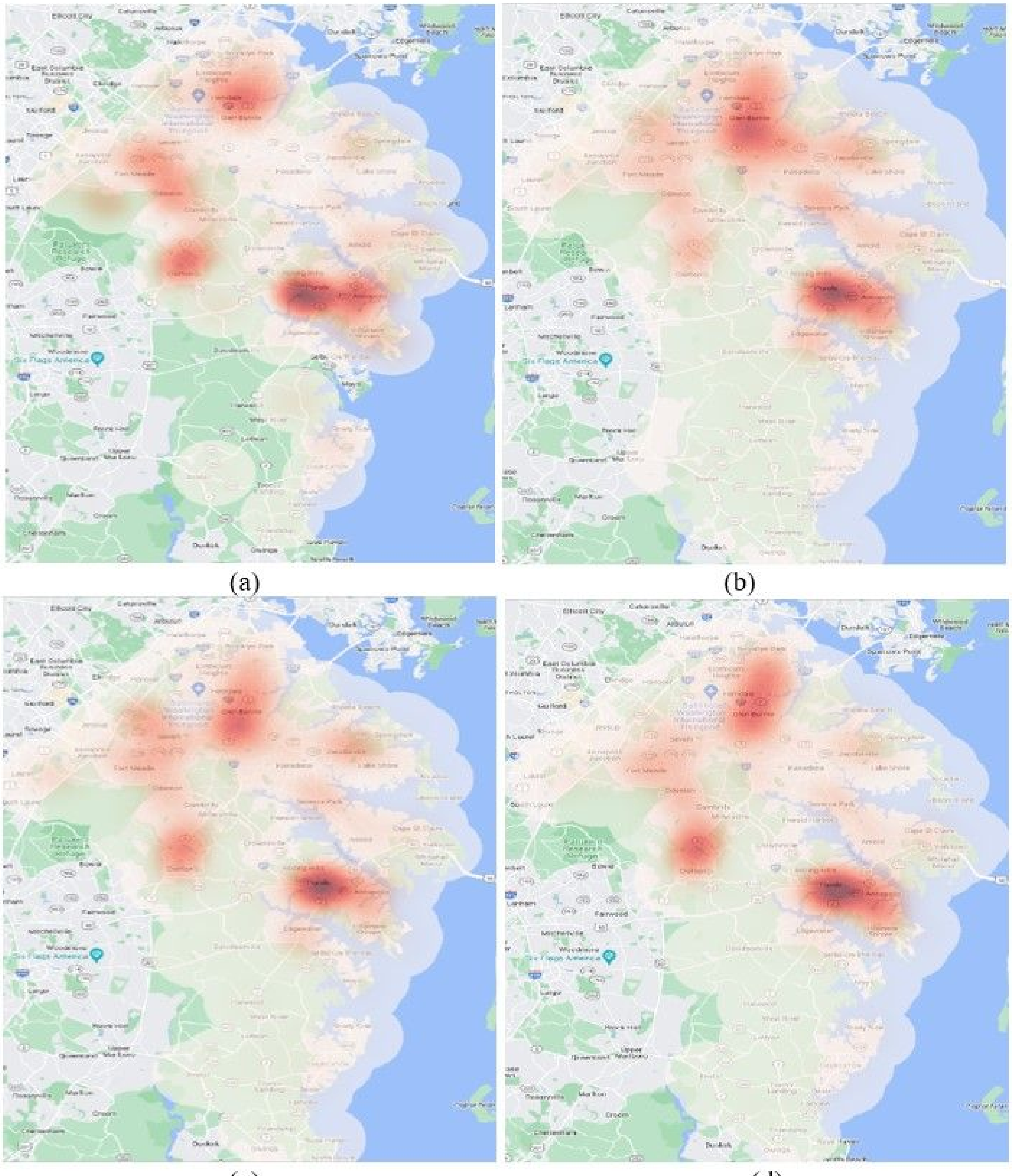


Figure 7

Figure 4-4. Proportion of home-based trips



(a) (b) (c) (d)

Figure 8

Figure 4-5. Heat maps of destinations of trips in (a) RTS, (b) MDLD, (c) travelers linked by approach 1, and (d) travelers linked by approach 2

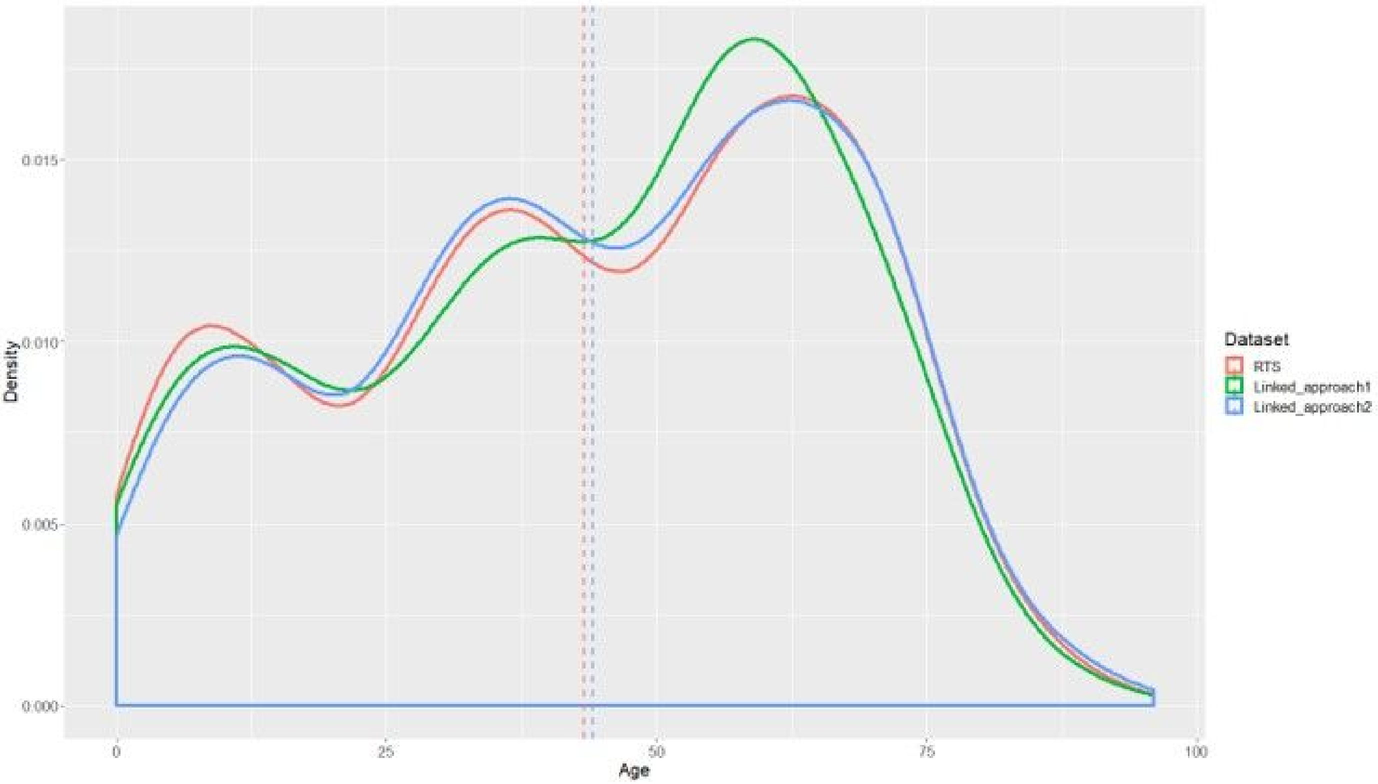


Figure 9

Figure 4-6. Age distribution

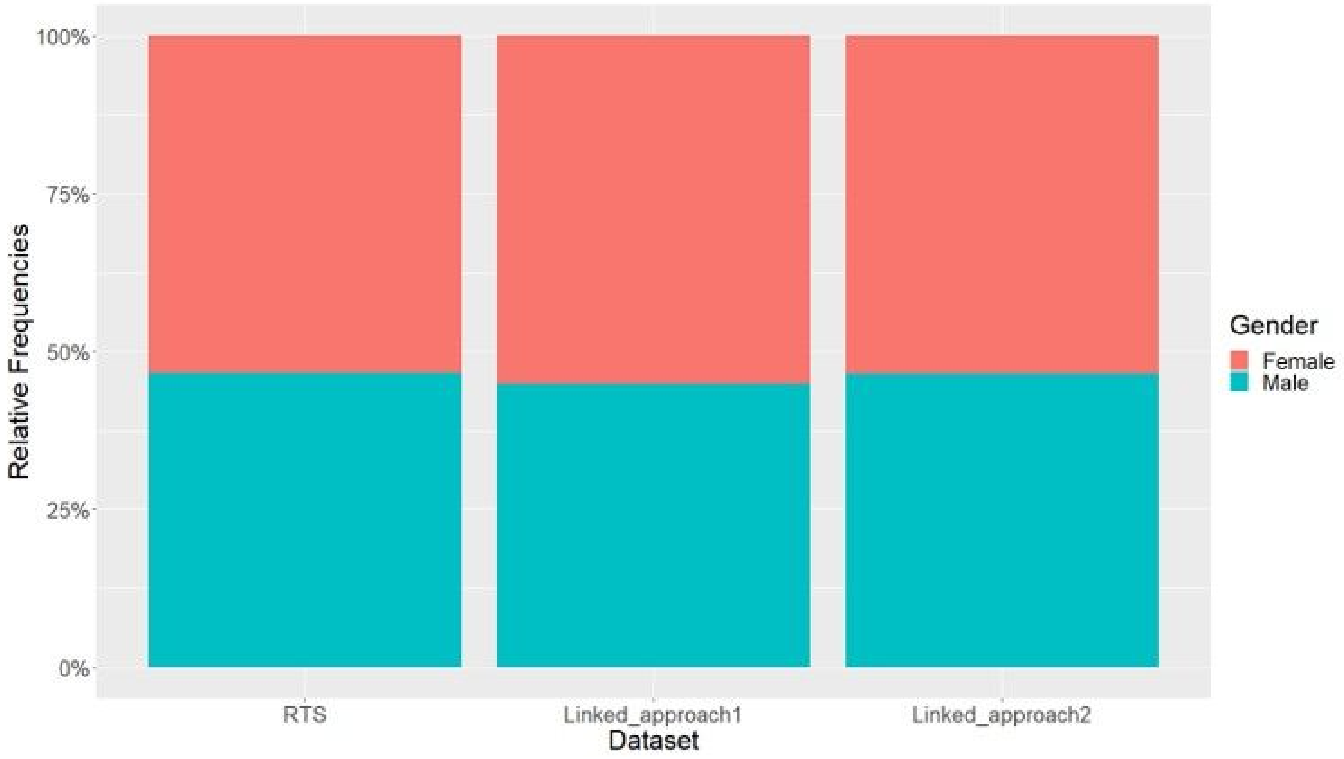


Figure 10

Figure 4-7. Gender distribution

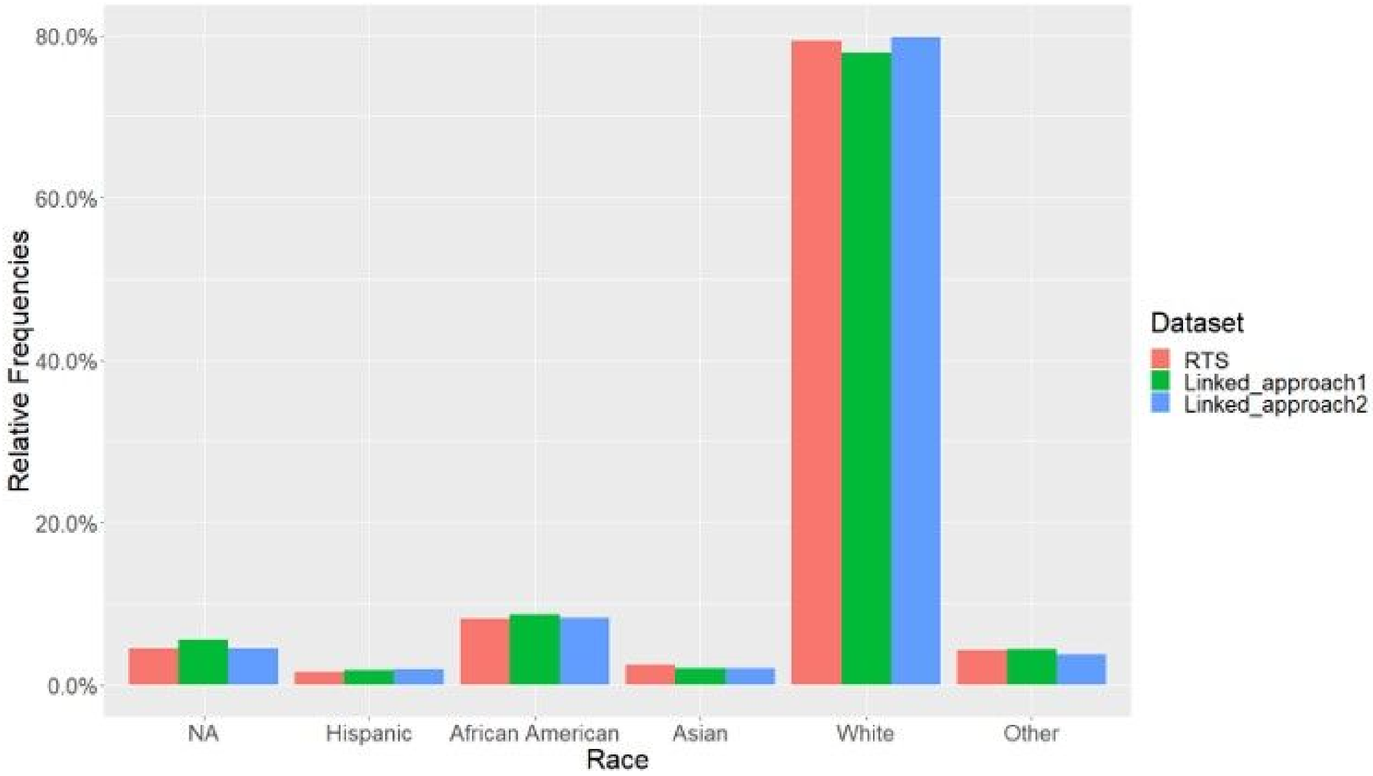


Figure 11

Figure 4-8. Race distribution

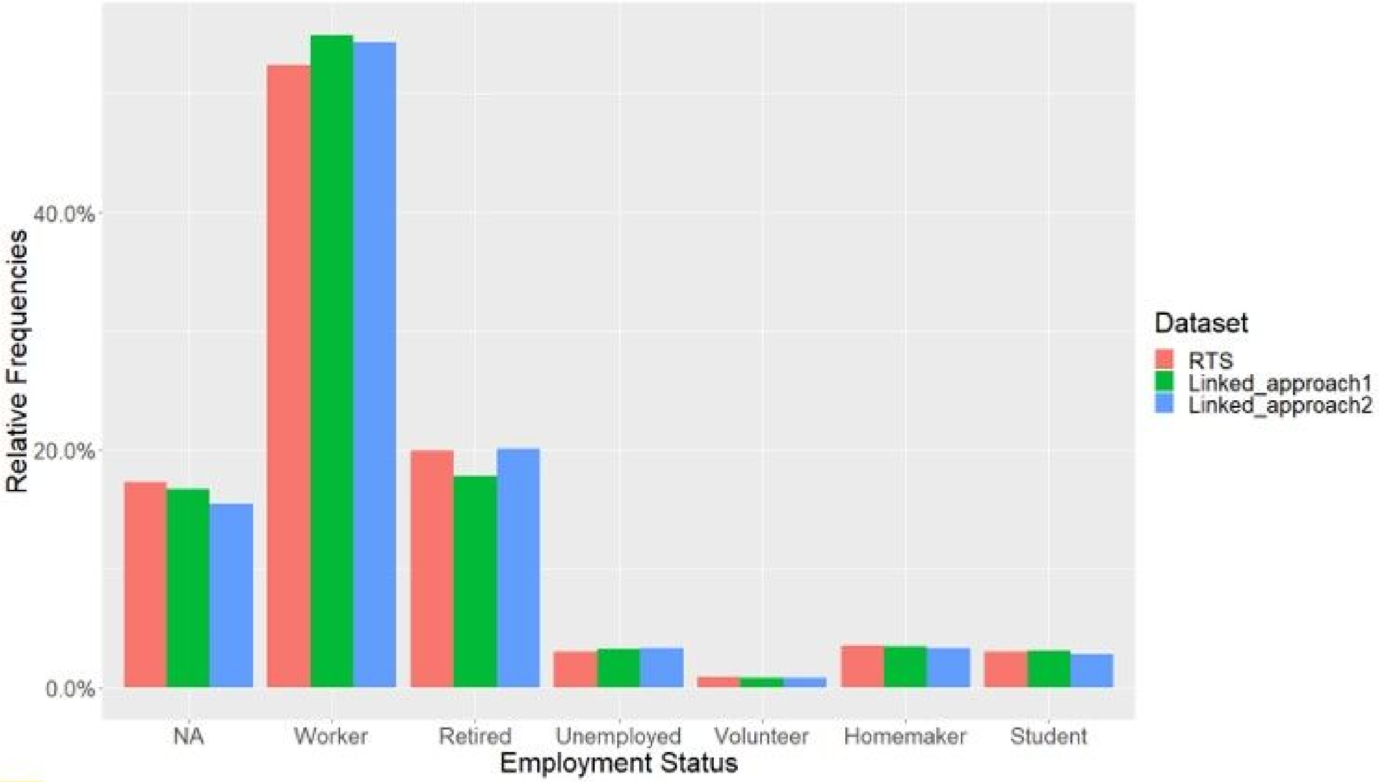


Figure 12

Figure 4-9. Employment status distribution