The Promise and Limitations of Locational App Data for Origin-Destination Analysis: A Case Study

APRIL 2018

U.S. Department of Transportation
Federal Highway Administration

TMIP
Better Methods. Better Outcomes.

Notice

This document is disseminated under the sponsorship of the U.S. Department of Transportation in the interest of information exchange. The U.S. Government assumes no liability for the use of the information contained in this document.

The U.S. Government does not endorse products or manufacturers. Trademarks or manufacturers' names appear in this report only because they are considered essential to the objective of the document.

Quality Assurance Statement

The Federal Highway Administration (FHWA) provides high-quality information to serve Government, industry, and the public in a manner that promotes public understanding. Standards and policies are used to ensure and maximize the quality, objectivity, utility, and integrity of its information. The FHWA periodically reviews quality issues and adjusts its programs and processes to ensure continuous quality improvement.

1. Report No. FHWA-HEP-20022	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle The Promise and Limitations of Locational App Data for Origin-Destination Analysis: A Case Study		5. Report Date October 2017
		6. Performing Organization Code
7. Authors Thomas Adler, Vince Bernardin, Jeff Dumont, Leah Flake, Hadi Sadrsadat		8. Performing Organization Report No.
9. Performing Organization Name and Address RSG 55 Railroad Row White River Junction, VT 05001		10. Work Unit No. (TRAIS)
		11. Contract or Grant No. DTFH61-12-D-00013
12. Sponsoring Agency Name and Address United States Department of Transportation Federal Highway Administration 1200 New Jersey Ave. SE Washington, DC 20590		13. Type of Report and Period Covered June 2017-October 2017
		14. Sponsoring Agency Code HEPP-30

15. Supplementary Notes

The project was managed by Task Manager for Federal Highway Administration, Sarah Sun, who provided detailed technical directions.

16. Abstract

Large, passively collected datasets from location-based services are a potential asset to transportation planning and modeling. These data have near-real-time availability and can capture travel behavior of a wide swath of the population. These data clearly hold promise for multiple transportation modeling methods, but practitioners should be aware of and account for various biases and other gaps inherent to the data. This volume seeks to further understand these gaps by comparing data from one large passively-collected data provider, Cuebiq, to data collected during a smartphone-based GPS household travel survey. As part of this comparison, the authors developed an algorithm to infer trips from the Cuebiq location data and identified smartphone users present in both datasets. Results from this comparison identify potential gaps in Cuebiq's representation of travel behavior, including demographic biases regarding traveler age and income and a bias toward trips longer than 9 miles (15 kilometers). The comparison also highlights the promising capability to capture detailed location behavior for a wide swath of the population, given the Cuebiq dataset's pervasive spatial coverage. This volume also summarizes persisting uncertainties regarding data from location-based services and describes potential future work to measure and account for inherent biases as these data are introduced to planning and modeling applications.

17. Key Words Origin-destination data; estimation, Big Data, GPS data, location-based services, rMove, travel surveys	18. Distribution Statement No restrictions.

19. Security Classif. (of this report) Unclassified	20. Security Classif. (of this page) Unclassified	21. No. of Pages 40	22. Price N/A

SI* (MODERN METRIC) CONVERSION FACTORS

APPROXIMATE CONVERSIONS TO SI UNITS

Symbol	When You Know	Multiply By	To Find	Symbol
		LENGTH		
in	inches	25.4	millimeters	mm
ft	feet	0.305	meters	m
yd	yards	0.914	meters	m
mi	miles	1.61	kilometers	km
		AREA		
in^2	square inches	645.2	square millimeters	mm^2
ft^2	square feet	0.093	square meters	m^2
yd^2	square yard	0.836	square meters	m^2
ac	acres	0.405	hectares	ha
mi^2	square miles	2.59	square kilometers	km^2
		VOLUME		
fl oz	fluid ounces	29.57	milliliters	mL
gal	gallons	3.785	liters	L
ft^3	cubic feet	0.028	cubic meters	m^3
yd^3	cubic yards	0.765	cubic meters	m^3
	NOTE: volumes greater than 1000 L shall be shown in m^3			
		MASS		
oz	ounces	28.35	grams	g
lb	pounds	0.454	kilograms	kg
T	short tons (2000 lb)	0.907	megagrams (or "metric ton")	Mg (or "t")
		TEMPERATURE (exact degrees)		
°F	Fahrenheit	5 (F-32)/9 or (F-32)/1.8	Celsius	°C
		ILLUMINATION		
fc	foot-candles	10.76	lux	lx
fl	foot-Lamberts	3.426	candela/m^2	cd/m^2
		FORCE and PRESSURE or STRESS		
lbf	poundforce	4.45	newtons	N
lbf/in^2	poundforce per square inch	6.89	kilopascals	kPa

APPROXIMATE CONVERSIONS FROM SI UNITS

Symbol	When You Know	Multiply By	To Find	Symbol
		LENGTH		
mm	millimeters	0.039	inches	in
m	meters	3.28	feet	ft
m	meters	1.09	yards	yd
km	kilometers	0.621	miles	mi
		AREA		
mm^2	square millimeters	0.0016	square inches	in^2
m^2	square meters	10.764	square feet	ft^2
m^2	square meters	1.195	square yards	yd^2
ha	hectares	2.47	acres	ac
km^2	square kilometers	0.386	square miles	mi^2
		VOLUME		
mL	milliliters	0.034	fluid ounces	fl oz
L	liters	0.264	gallons	gal
m^3	cubic meters	35.314	cubic feet	ft^3
m^3	cubic meters	1.307	cubic yards	yd^3
		MASS		
g	grams	0.035	ounces	oz
kg	kilograms	2.202	pounds	lb
Mg (or "t")	megagrams (or "metric ton")	1.103	short tons (2000 lb)	T
		TEMPERATURE (exact degrees)		
°C	Celsius	1.8C+32	Fahrenheit	°F
		ILLUMINATION		
lx	lux	0.0929	foot-candles	fc
cd/m^2	candela/m^2	0.2919	foot-Lamberts	fl
		FORCE and PRESSURE or STRESS		
N	newtons	0.225	poundforce	lbf
kPa	kilopascals	0.145	poundforce per square inch	lbf/in^2

*SI is the symbol for the International System of Units. Appropriate rounding should be made to comply with Section 4 of ASTM E380.
(Revised March 2003)

SI* (MODERN METRIC) CONVERSION FACTORS

APPROXIMATE CONVERSIONS TO SI UNITS

Symbol	When You Know	Multiply By	To Find	Symbol
		LENGTH		
in	inches	25.4	millimeters	mm
ft	feet	0.305	meters	m
yd	yards	0.914	meters	m
mi	miles	1.61	kilometers	km
		AREA		
in^2	square inches	645.2	square millimeters	mm^2
ft^2	square feet	0.093	square meters	m^2
yd^2	square yard	0.836	square meters	m^2
ac	acres	0.405	hectares	ha
mi^2	square miles	2.59	square kilometers	km^2
		VOLUME		
fl oz	fluid ounces	29.57	milliliters	mL
gal	gallons	3.785	liters	L
ft^3	cubic feet	0.028	cubic meters	m^3
yd^3	cubic yards	0.765	cubic meters	m^3
	NOTE: volumes greater than 1000 L shall be shown in m^3			
		MASS		
oz	ounces	28.35	grams	g
lb	pounds	0.454	kilograms	kg
T	short tons (2000 lb)	0.907	megagrams (or "metric ton")	Mg (or "t")
		TEMPERATURE (exact degrees)		
°F	Fahrenheit	5 (F-32)/9 or (F-32)/1.8	Celsius	°C
		ILLUMINATION		
fc	foot-candles	10.76	lux	lx
fl	foot-Lamberts	3.426	candela/m^2	cd/m^2
		FORCE and PRESSURE or STRESS		
lbf	poundforce	4.45	newtons	N
lbf/in^2	poundforce per square inch	6.89	kilopascals	kPa

APPROXIMATE CONVERSIONS FROM SI UNITS

Symbol	When You Know	Multiply By	To Find	Symbol
		LENGTH		
mm	millimeters	0.039	inches	in
m	meters	3.28	feet	ft
m	meters	1.09	yards	yd
km	kilometers	0.621	miles	mi
		AREA		
mm^2	square millimeters	0.0016	square inches	in^2
m^2	square meters	10.764	square feet	ft^2
m^2	square meters	1.195	square yards	yd^2
ha	hectares	2.47	acres	ac
km^2	square kilometers	0.386	square miles	mi^2
		VOLUME		
mL	milliliters	0.034	fluid ounces	fl oz
L	liters	0.264	gallons	gal
m^3	cubic meters	35.314	cubic feet	ft^3
m^3	cubic meters	1.307	cubic yards	yd^3
		MASS		
g	grams	0.035	ounces	oz
kg	kilograms	2.202	pounds	lb
Mg (or "t")	megagrams (or "metric ton")	1.103	short tons (2000 lb)	T
		TEMPERATURE (exact degrees)		
°C	Celsius	1.8C+32	Fahrenheit	°F
		ILLUMINATION		
lx	lux	0.0929	foot-candles	fc
cd/m^2	candela/m^2	0.2919	foot-Lamberts	fl
		FORCE and PRESSURE or STRESS		
N	newtons	0.225	poundforce	lbf
kPa	kilopascals	0.145	poundforce per square inch	lbf/in^2

*SI is the symbol for the International System of Units. Appropriate rounding should be made to comply with Section 4 of ASTM E380.
(Revised March 2003)

The Promise and Limitations of Locational App Data for Origin-Destination Analysis: A Case Study

Original: **October 2017**
Final: **April 2018**

Prepared for:

Federal Highway Administration

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviations

FHWA	Federal Highway Administration
IP	Internet Protocol
LBS	location-based service
OD	origin-destination
ODOT	Ohio Department of Transportation

Executive Summary

Passively collected data aggregated from cellular networks, navigation devices, and smartphone apps ("Big Data") are increasingly being used to inform transportation modeling. Given their new prevalence, it is important to recognize that the understanding of these datasets, particularly of their representativeness, is incomplete. Better understanding the limitations and biases of Big Data will enable the transportation planning and modeling community to build a knowledge base for improving transportation planning applications. This understanding will also help ensure that the transportation planning community accounts for known biases and limitations when working with passively collected data and avoids misinformation when applying these data toward transportation applications.

One data provider that is aggregating location-based service (LBS) data from smartphone apps is Cuebiq. Cuebiq partners with over 100 smartphone apps to provide locational traces when those apps are activated. This volume describes a case study of Cuebiq data, comparing it to location and trip data collected via a smartphone app (rMove™) specifically built for travel diaries. The case study provides several insights:

1. Cuebiq location data have robust spatial coverage, even when limited to the most accurate location traces. The most obvious advantage of Cuebiq data is the level of user penetration and resulting geographic coverage and complete coverage of the OD solution space achieved given the myriad apps Cuebiq leverages for data collection.
2. Location accuracy varies significantly by device type in the Cuebiq dataset, with locations from devices using the iOS operating system having significantly greater locational accuracy in this case study.
3. Location-based data from the Cuebiq dataset shows a high variation in the median time between points. Frequency and sparsity of location data are likely different for each smartphone app leveraged by Cuebiq or impacted by individual device settings.
4. The Cuebiq data demonstrate strong potential for trip inference. Accounting for dwelled activity duration versus the duration of travel is critical for accurately inferring trips and trip travel time. Further research on more refined and robust trip-inference algorithms is recommended.
5. Conversely, short-duration trips are difficult to derive from the Cuebiq data. Trips less than 30 minutes in duration are likely underrepresented in the Cuebiq data, which in relative terms leads to an overrepresentation of trips longer than 30 minutes. Failing to account for this systematic bias will lead to skewed estimates of trip lengths and durations and OD patterns in general.
6. The comparative analysis indicates the presence of demographic biases in the Cuebiq data, related to the age *and* incomes of users. It is likely necessary to control for and expand LBS data to correct for these demographic biases to avoid skewing travel metrics.
7. The comparative analysis identified a small sample with both good confidence and completion from users generating GPS traces to both Cuebiq and to rMove. The sample size of these paired traces is too small to support many statistically significant conclusions,

but among device pairs that appear to represent the same user, the Cuebiq data often exclude some portion of the travel collected by rMove.

This comparative analysis identified the following key challenges:

- The absence of the list of apps used to derive the Cuebiq data. This can introduce bias among what is collected by device or by type of user.
- The absence of known demographics of Cuebiq users. This information can be imputed with some unknown level of certainty from the location data, but this introduces a level of error on top of the uncertainties inherent to the dataset. Matching devices in a smartphone-based GPS travel survey can produce more certainty (but only with a sufficiently large matched sample).
- Difficulty in measuring bias among users who remain in the dataset for multiple days vs. those who drop out after one day.

Uninformed use of Cuebiq data without correcting for systematic biases can result in faulty analyses and conclusions. Key biases identified in this comparative analysis relate to demographics, and to temporal sparsity at the individual level or the infrequency of observations. As this analysis shows, these biases can skew the observed duration of trips and other activities.

Finally, the spatial and temporal coverage of the Cuebiq data is far more robust than smartphone surveys like rMove, and even more so compared to traditional surveys. For this reason, reliance on survey data alone could limit the ability of models to present an accurate and complete picture of travel patterns.

1.0 Introduction

1.1 Disclaimer

The views expressed in this document do not represent the opinions of FHWA and do not constitute an endorsement, recommendation, or specification by FHWA. The document is based solely on the research conducted by RSG.

1.2 Acknowledgments

This volume is a collaboration between transportation professionals at FHWA, FTA, Cuebiq, and RSG.

1.3 Introduction and Overview

Traditional travel surveys are still the predominant source of origin-destination (OD) travel patterns in urban areas, but passively collected data from commercial sources offer complementary information. Passively collected data—aggregated from cellular networks, navigation devices, and smartphone apps—also deliver massive datasets for statistical analysis.

The travel forecasting community is using these datasets to support transportation planning. Some Big Data providers, for example, provide an already processed and aggregated OD matrix based on their processed location data. In addition, some agencies use the data to provide information on trips not covered by household surveys, such as external trips and visitor trips, and to independently validate travel demand models developed primarily from household survey data. The data are increasingly being applied in new ways—including demand model parameter estimation and data-driven modeling techniques (e.g., pivot point forecasting).

One data provider that is aggregating location-based app data is Cuebiq. Cuebiq has partnered with over 100 app providers to incorporate code in their apps that allows Cuebiq to access the location details polled by those applications. The result is a dataset that includes locational traces for each device on which one or more of such apps are installed.

2.0 Importance of Understanding Passive Data Sources

Use of these datasets is expanding, but an incomplete (yet developing) understanding of these datasets persists, particularly of their representativeness. The research documented in this volume seeks to further develop the knowledge base for these data and inform potential data users of the characteristics of select passively collected location data offerings. The research presented in this volume consists of a case study of Cuebiq data compared to location and trip data collected via a smartphone app, rMove, specifically built for travel diaries.

3.0 Case Study (April 2017)

3.1 Case Study Purpose

This case study seeks to understand Cuebiq data, its representativeness, and its potential OD trip and activity information applications. The project team compared Cuebiq data to actively collected location data from controlled random-sample household surveys conducted using rMove. This was done to understand the data, users, and gaps in the Cuebiq dataset. The Ohio Department of Transportation (ODOT) granted access to its location and trip data collected in Franklin County, Ohio, during the spring of 2017. The project team compared these data to Cuebiq data at the location level and trip level. An algorithm inferred trips from the Cuebiq dataset based on various location data characteristics. The process identified users whose location data were present in both datasets. This helped further compare and understand Cuebiq data beyond observing the general properties of each dataset. This step also helped fine-tune the trip-inference algorithm and identify gaps at the location and trip level within the Cuebiq data. Subsequent sections describe the trip-inference and user matching processes and results.

3.2 Cuebiq Data Size and Format Considerations

Cuebiq provided location data collected throughout the United States for the month of April 2017 (between April 3 and April 30). These data are stored in Amazon Web Service S3 Buckets, with one bucket per day of data. Each bucket consists of 1,500 zipped flat files, which were downloaded and converted to working data files to process internally. Cuebiq data are at the location level, with one row per collected location point. Dataset fields include the following:

- Device identification number (ID).
- Latitude and longitude (in decimal format).
- Accuracy radius (in meters).
- Internet Protocol (IP) address of the device at the time of capture.
- ID for the source of the location point.
- Timestamp at which the location was collected.
- Timestamp of the last capture for the device.
- Device type, model, and carrier.
- ZIP Code and state.

The April Cuebiq data files contain approximately 1,400,000,000 data points per day among 11,000,000 devices. Given the vast amount of data provided even for a single day, the scope of assessment is limited to a single day of data over a specific geography. The project team selected Franklin County, Ohio, for comparison.

3.3 The Comparison Dataset

The Ohio Moves Transportation Study, for which rMove data were collected, involved an address-based household sample recruited via mail. Eligible households participated for seven days using

rMove, which detects trips using a smartphone's GPS sensors and Wi-Fi capabilities and asks users to complete a survey about each trip. Users are prompted to review their trips each day and can add, delete, split, or merge their trips. Machine learning algorithms process the raw data from rMove, and analysts flag potentially inaccurate data during spatial review. This comparison used processed rMove data.

Cuebiq data were initially filtered to include location data in Ohio for one day in April, then filtered again to include only locations among devices that were present in Franklin County, Ohio, any time during that day. The selected day in April was based on the weekday with the highest number of trips in the Ohio household travel survey dataset collected via rMove. To achieve robust evaluation of Cuebiq data, two other weekdays with a high number of trips were also evaluated for some analyses, as described below.

3.4 Comparison Data Characteristics

Table 1 summarizes rMove and Cuebiq data for all Cuebiq devices observed in Franklin County and all rMove devices in the sample on the same day.

Table 1. Location data summary for one day in April.

	Cuebiq	rMove
Total devices	95,697	222
Points	12,128,310	124,681
Median time between points	147 seconds (2.5 minutes)	4 seconds (0.06 minutes)
Mean time between points	2.3 hours (138 minutes)	0.5 hours (32 minutes)
Standard deviation between points	2.2 hours (132 minutes)	0.72 hours (43 minutes)

Table 2. Composition of location data by device type for one day in April.

		Cuebiq	rMove
% of points	iOS	54%	74%
	Android	46%	26%
% of devices	iOS	57%	50%
	Android	43%	50%

The total number of devices in the Cuebiq dataset observed in Franklin County is approximately 8% of the total population of the county (1.25 million people). Because not everyone passing through the county necessarily lives there, imputing home locations from the dataset would provide a better estimate of Cuebiq penetration in the county, but this was not undertaken as part of this study.

Overall, frequency of Cuebiq point collection is relatively low, though frequency varies greatly among devices. The median time between points is generally the better measure of the frequency of observations; the mean is skewed because both Cuebiq LBS and rMove stop making as

frequent observations when they detect that the device has stopped moving (i.e., there are long periods without locational observations while a person remains in one location, such as at home overnight).

Based on the median time between points, the frequency of data points is sufficient for most purposes in both datasets. However, rMove provides a much more granular and frequent set of trace points with a median time of 4 seconds between observations compared to 147 seconds in Cuebiq. Because rMove is designed to collect locations frequently during trips, a relatively high number of locations have one second or less between points collected, resulting in the low median time between points. The Cuebiq data demonstrate that median time between points varies throughout the dataset, as frequency and sparsity of location data may be different for each smartphone app leveraged by Cuebiq—or impacted by individual device settings or operating systems. The difference in mean time between points among the two datasets results in a t-statistic of 16.3, indicating a highly significant difference; however, the means and standard deviations are skewed because of dwell time. Basic comparison of the medians reveals significant differences in frequency of observations. This lower frequency of observations in Cuebiq results in longer delay in recognizing that the traveler/device has begun moving again after a long dwell, which likely accounts for the failure to observe some short trips.

Accuracy varies significantly by device type in the Cuebiq dataset, as shown in Table 3, which provides the composition of location data by device type for each accuracy level (i.e., of all points with 164 feet [50 meters] of accuracy or less, 66% are from iOS devices). In terms of overall location makeup in the datasets (Table 2), the rMove data present a slightly more even split between Android and iOS devices, although the Cuebiq split is similar. Because Cuebiq relies on various smartphone apps for location data, the set of available or commonly used apps enabling Cuebiq may differ between Android and iOS, which would account for differences in accuracy levels. For example, a Cuebiq-enabled app that offers navigation services with high accuracy may only be available on iOS devices. Unfortunately, the full set of applications that include the Cuebiq-SDK (software development kit) is proprietary and not publicly disclosed.

The rMove data are not affected by app use, but these data also include more points from iOS devices than Android devices, likely due to differences in how each type of device leverages GPS and Wi-Fi sensors.

Table 3. Location composition by accuracy range, by device type in Cuebiq data.

		3.3 feet (1 m)	< 33 feet (10 m)	< 164 ft. (50 m)	< 328 ft. (100 m)	< 3,280 ft. (1,000 m)	All
Total	**Devices**	33,716	48,300	91,915	94,123	95,300	95,697
	Points	1,037,151	3,553,114	8,093,624	9,863,018	10,786,426	12,128,310
% of points	**iOS**	100%	98%	66%	61%	59%	54%
	Android	0%	2%	34%	39%	41%	46%

Figure 1 illustrates **the persistence of user IDs in the Cuebiq dataset**. Approximately 20% of devices “drop out” of the dataset after the first day. Weekends also produce a meaningful drop, with a small percentage of devices reappearing in the dataset when the week begins on Monday.

This indicates that some Cuebiq users may only use Cuebiq-enabled apps for one day, and that some Cuebiq-enabled apps are in use only during weekdays.

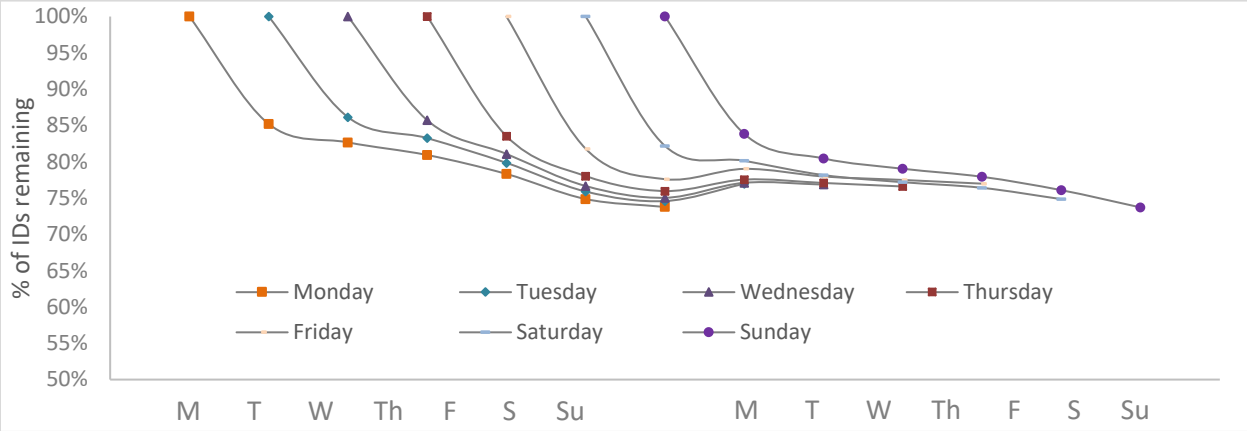

Figure 1. Persistence of user IDs in Cuebiq data over time0. (Sources: RSG, Cuebiq, 2017).

Cuebiq location data have robust spatial coverage, even when only considering the highest accuracy points. Figure 2 through Figure 5 show all points with 3.3 feet (1 meter) of accuracy radius among devices that were observed in Franklin County on the day of interest, mapped at various zoom levels. Franklin County is outlined in blue. rMove location data from the same day demonstrates sparser coverage of the region, which becomes more apparent as smaller geographic ranges are considered.

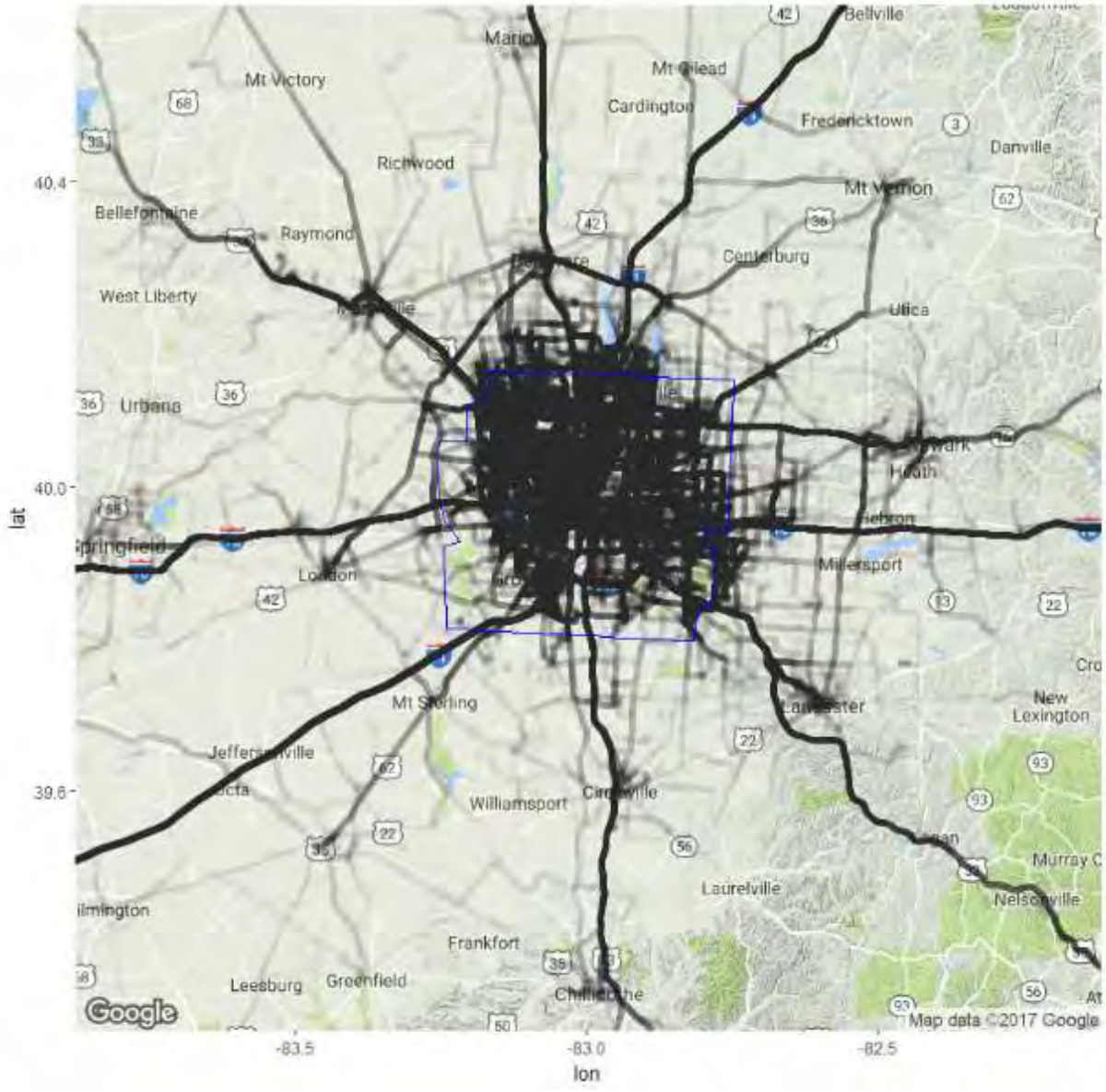

Figure 2. Cuebiq locations for one day (Sources: RSG, Cuebiq, Google Maps, 2017).

Figure 3. Cuebiq locations for one day where accuracy = 3.3 feet/1 meter; higher "zoom" magnitude (Sources: RSG, Cuebiq, Google Maps, 2017).

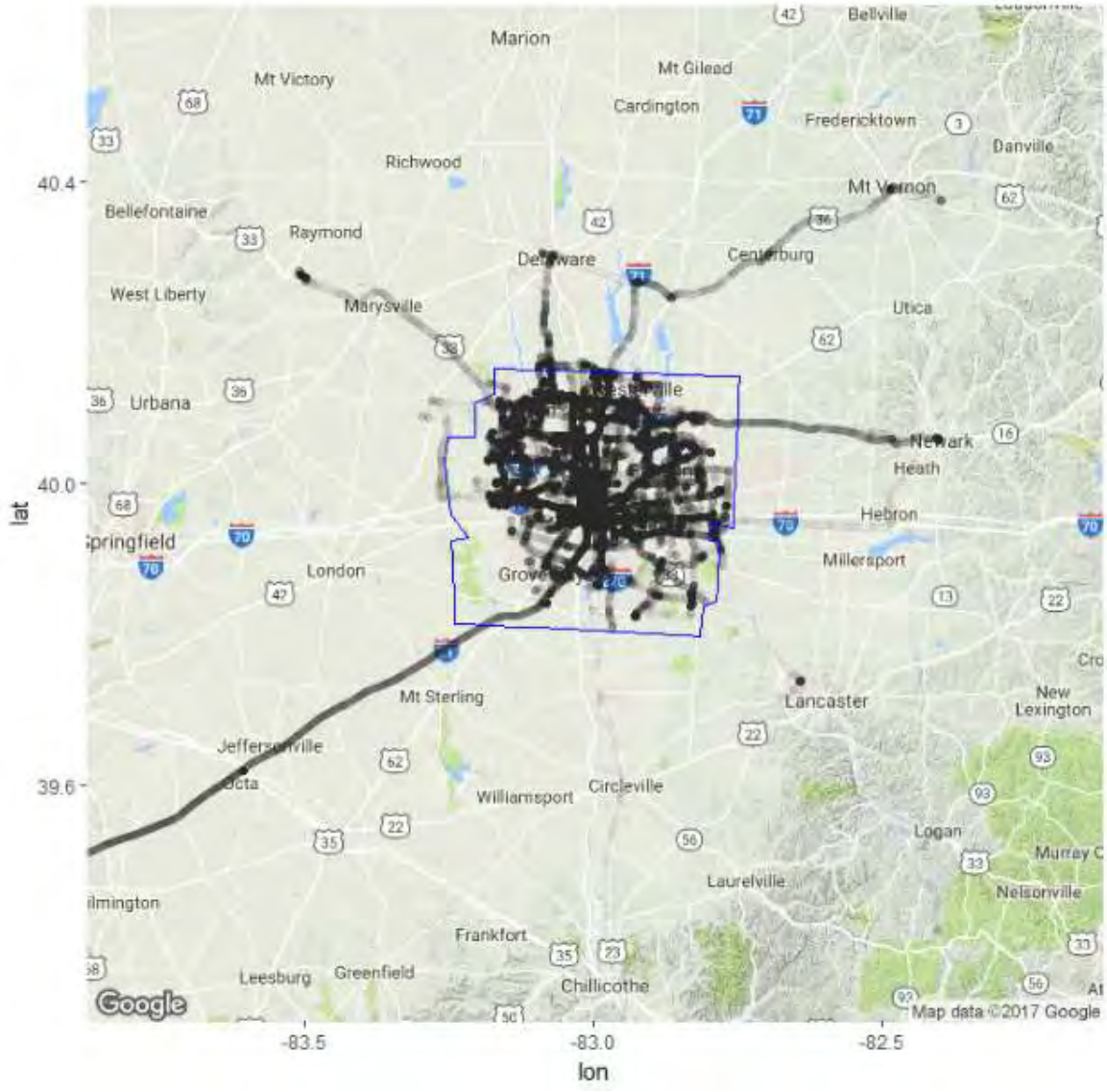

Figure 4. rMove locations for one day. (Sources: RSG, Google Maps, 2017).

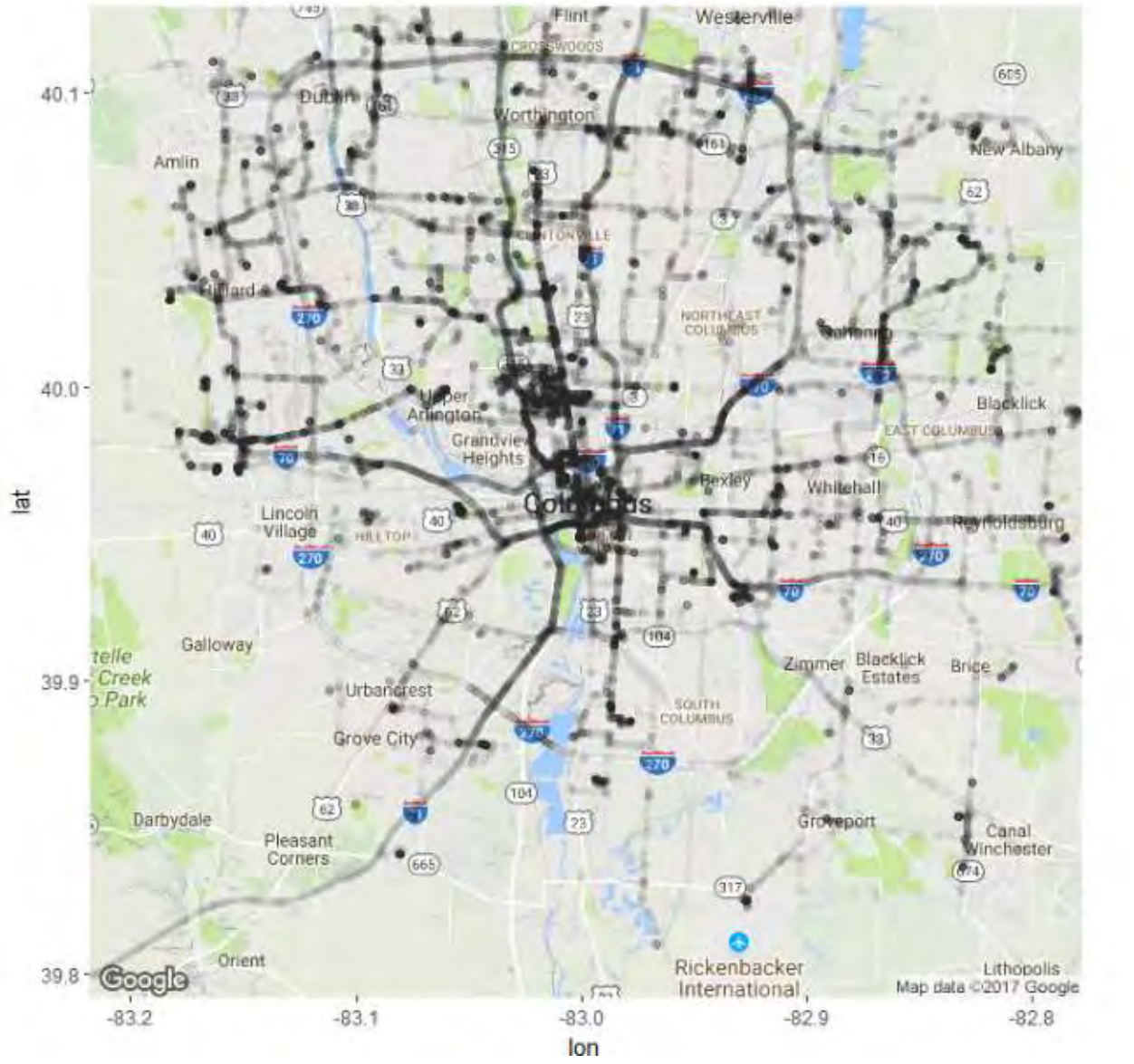

Figure 5. rMove locations for one day; higher "zoom" magnitude. (Sources: RSG, Google Maps, 2017).

3.5 Trip-Inference and Aggregate Comparisons

As part of the analysis, the project team created an algorithm that imputed trips from the Cuebiq location points. The algorithm used implied speed, cumulative distance traveled, dwell time, and distance between points to build a dataset of trips. The Appendix provides more detail of this algorithm in pseudocode. Matches between the Cuebiq and rMove datasets calibrated the algorithm. The project team applied the algorithm to three days of Cuebiq data, including the primary day of interest.

Analysis of the data revealed potential improvements to the algorithm. For example, some trips had to be filtered out of the trip duration analysis because the algorithm did not correctly identify

some trip ends. Although the location of the destination was *believed* to be correct, large portions of the activity duration for a small portion of trips were erroneously included in the travel duration, resulting in unreasonable implied travel speeds. The algorithm also falsely identified some brief trips. These trips were filtered out of the subsequent analysis. Further research on more refined and robust trip-inference algorithms is recommended.

Table 4 lists the characteristics of the Cuebiq trips imputed on each of the days. For comparison, rMove trip characteristics (Table 5) are shown below for the same three days. Cuebiq median beeline distance (straight line distance between origin and destination) is generally lower than rMove median beeline distance, while median travel time is much higher. One potential reason for this is the issue of short or zero-beeline-distance trips, as suggested above. Deriving path distance in the trip-inference algorithm using a map-matching process could provide a better understanding of inferred trip feasibility in terms of distance in duration. Median path distance for rMove is shown in Table 5 as "total distance" as the rMove data already included this information.

Table 4. Trip characteristics of Cuebiq inferred trips over three days.

	Day 1	Day 2	Day 3
Number of Trips	378,953	348,778	371,732
Median Distance (Bee Line)	1.56 mi (2.50 km)	1.67 mi (2.69 km)	1.53 mi (2.46 km)
Median Travel Time	34 minutes	36 minutes	35 minutes
Number of Trips per Device	4.73	4.40	4.73

Table 5. Characteristics of rMove recorded trips over three days.

	Day 1	Day 2	Day 3
Number of Trips	1,187	1,025	1,051
Median Distance (Bee Line)	1.64 mi (2.65 km)	2.29 mi (3.68 km)	2.48 mi (4.00 km)
Median Total Distance	2.6 mi (4.21 km)	3.15 mi (5.06 km)	3.82 mi (6.16 km)
Median Travel Time	12 minutes	12 minutes	14 minutes
Number of Trips per Device	5.34	5.63	5.13

While generally similar, the number of trips per device is more than 10% higher from rMove than from Cuebiq devices, suggesting that the Cuebiq data may be missing approximately 1 in 10 (or more) trips. rMove trip characteristics also present a reasonable median travel speed (about 15 miles per hour). Median beeline distance, travel time, and trips per device are all statistically significant ($p < 0.05$) between the two datasets. Because the Cuebiq trip inference resulted in a high number of trips with beeline distances under 0.15 miles (0.25 kilometers), for the reasons described previously, analyses include only trips greater than 0.25 kilometers in length in rMove and Cuebiq.

Figure 6 shows the distribution of trip distances between rMove and Cuebiq. While Cuebiq contains more inferred trips under 0.3 miles (0.5 kilometers), Cuebiq also has a higher percentage of trips greater than 9 miles (15 kilometers). Moreover, the trips less than 500 meters in the Cuebiq data almost certainly contain some false trips related to locational imprecision in the less precise

subset of Cuebiq data. A more refined version of the trip-inference algorithm could likely address this.

Overall, rMove tends to capture more short-distance trips (under 9 miles [15 kilometers]). One potential explanation for this is that rMove is an app designed for detecting *all* trips, and Cuebiq only collects location data when certain apps are in use. rMove may collect relatively more short-distance discretionary trips, while Cuebiq may collect trips that elapse a certain distance and time threshold.

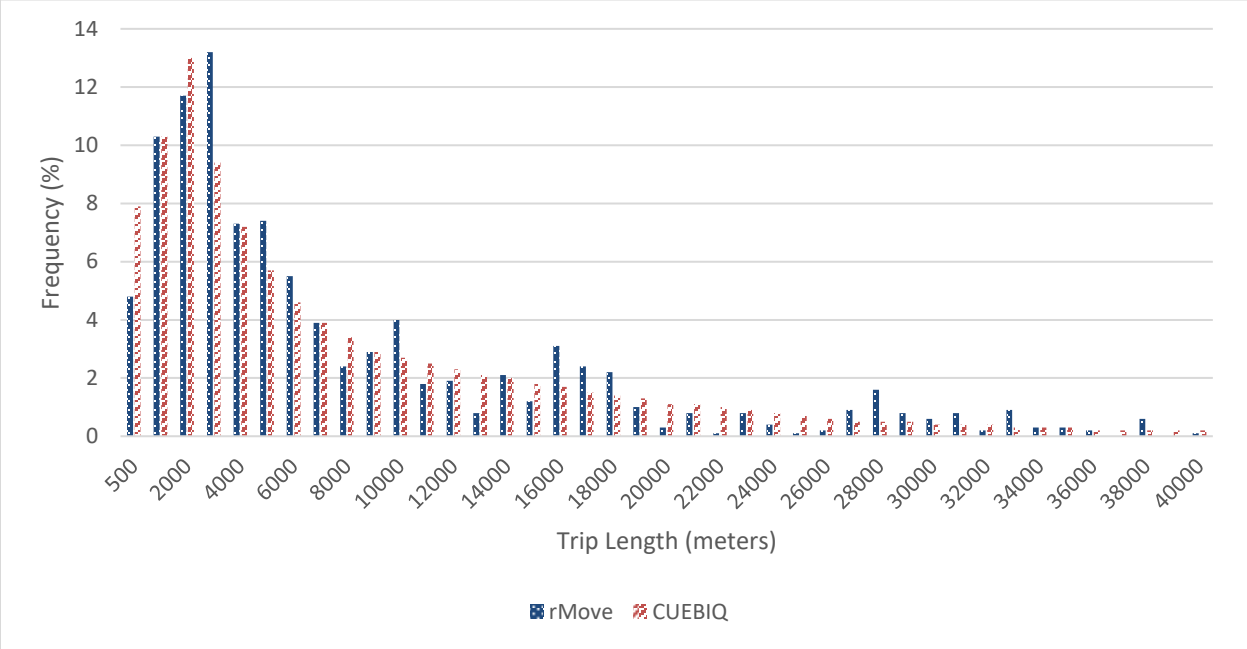

Figure 6. Trip distance distribution (meters) among rMove and Cuebiq over the same day in April. (Sources: RSG, Cuebiq, 2017).

Filtering out short trips smaller than the locational precision in the Cuebiq dataset produces the trip frequency by trip duration between rMove and Cuebiq, shown in Figure 7. This demonstrates similar and clearer patterns among shorter duration vs. longer duration trips, with higher-duration trips included at a higher rate in the Cuebiq dataset compared to rMove.

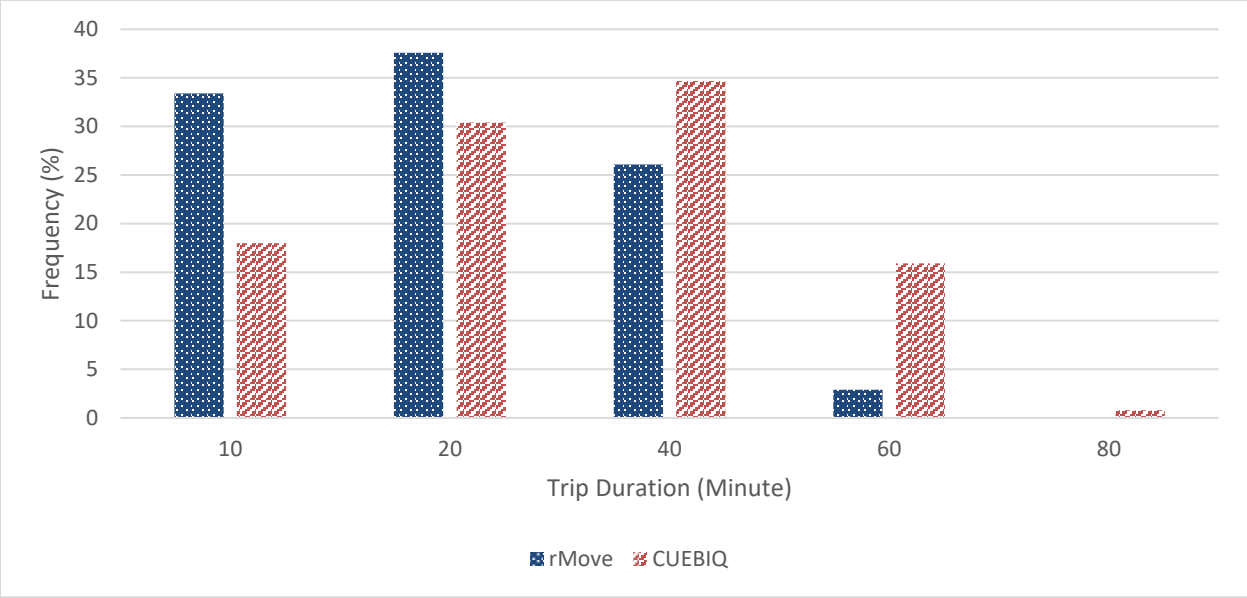

Figure 7. Trip duration distribution (minutes) among rMove and Cuebiq over the same day in April. (Sources: RSG, Cuebiq, 2017).

As shown in Figure 7 and Figure 8, trips under 30 minutes duration are likely underrepresented in the inferred trips from Cuebiq data, which in relative terms makes trips longer than 30 minutes overrepresented in this inferred trip set. Failing to account for this systematic bias would lead to skewed estimates of trip lengths and durations and OD patterns in general. Similar bias in other types of passive OD datasets (e.g., GPS, cellular) has also been inferred based on comparisons both with aggregate traditional survey data and with traffic counts; however, underreporting of short trips in traditional household surveys without location tracing/verification may obscure this bias when compared to smartphone survey data.

The rMove data result from an algorithm designed specifically to collect trips, and these trips are reviewed by users and undergo rigorous data cleaning. As such, these data afford a reliable comparison to understand trip-inference results from the Cuebiq dataset. However, the evidence for similar (and possibly even greater) biases in other passive datasets suggests this may be a general issue with all passive data rather than an issue specific to LBS or Cuebiq, as can be seen in the discussion provided in Volume 1.

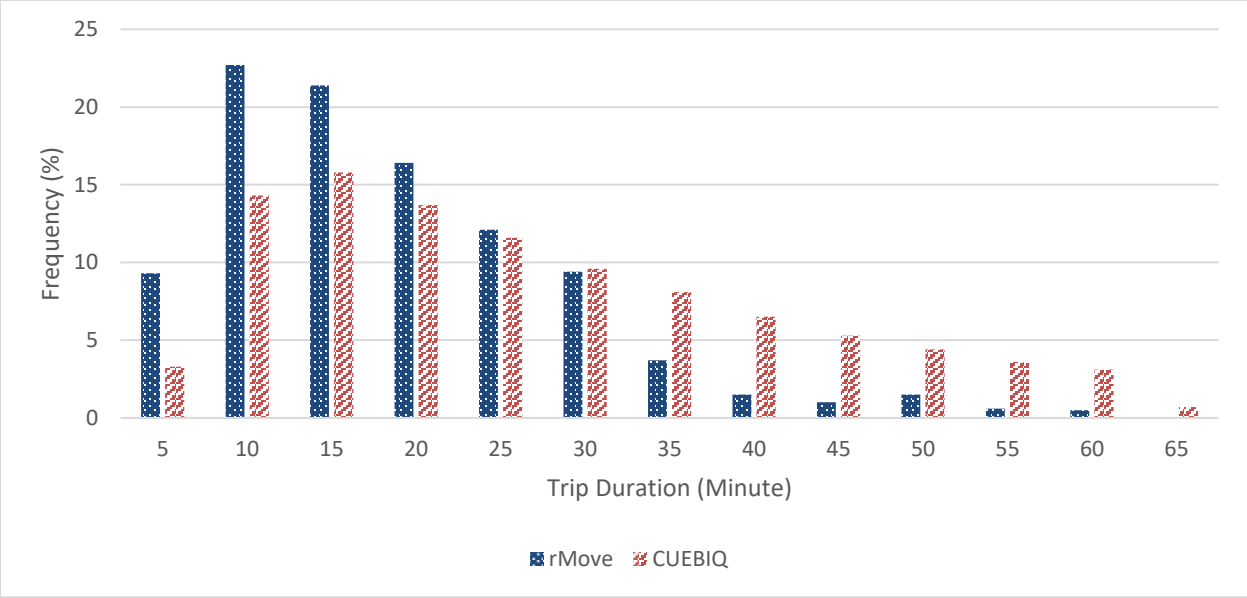

Figure 8. Detailed trip duration distribution (minutes) among rMove and Cuebiq over the same day in April. (Sources: RSG, Cuebiq, 2017).

The Cuebiq dataset misses a certain portion of trips from individual devices, but it offers much more complete geographic coverage over a single day. The project team compared geographic coverage of Cuebiq and rMove trips at the census tract level. Because rMove data are typically collected from different households over several months during a study, a single day of data represents only a small portion of the overall coverage provided by rMove studies. This also applies to passive data, especially considering availability of observations over prolonged periods.

Although the true/total coverage is understated for both Cuebiq and rMove datasets, the comparison of a single day of data from each dataset presents their relative geographic coverage for that day. Weighted rMove data were used for the geographic coverage comparison. Trip ODs at the census tract level (Figure 9) show that the Cuebiq data includes trips to/from every tract in the study area, while a significant portion of the tracts have no observations in the rMove data over a single day.

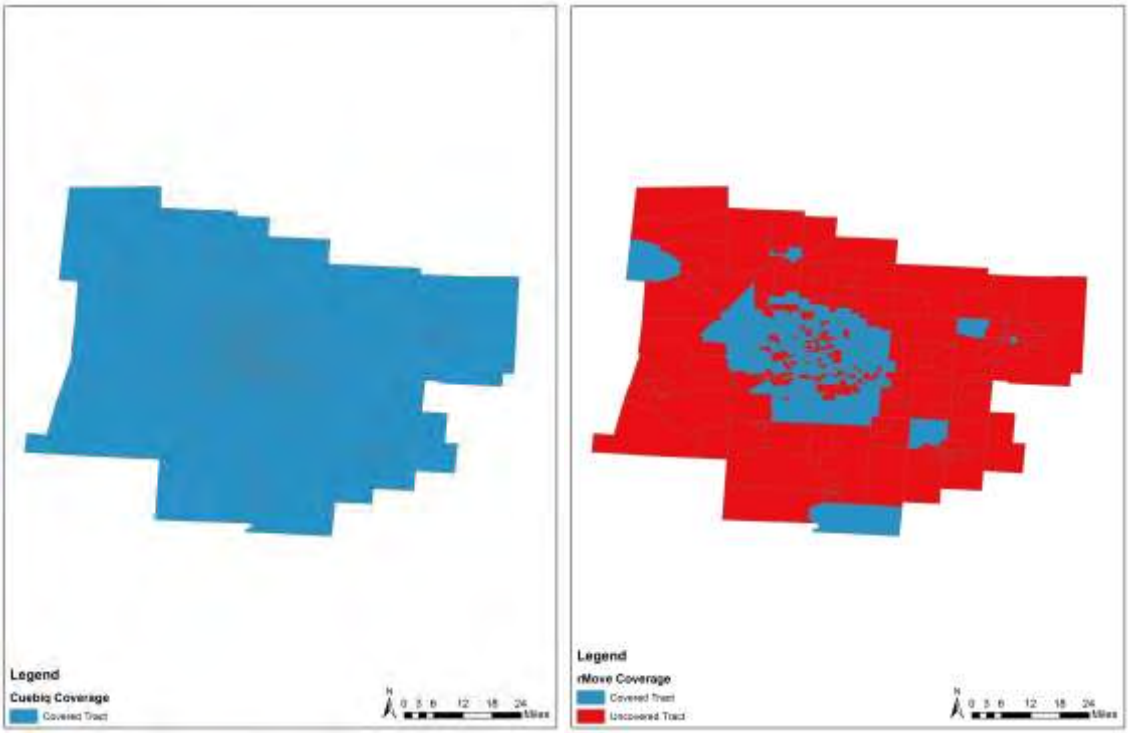

Figure 9. Geographic coverage of Cuebiq and rMove, by tract. (Sources: RSG, Cuebiq, Caliper Corp., 2017).

The project team calculated a trip rate comparison metric as the ratio of total trips to the sum of households, retail employment, and total employment (as rMove and Cuebiq capture both commercial and personal trips and nonhome trip ends). This was done to compare trip generation data captured by rMove and Cuebiq for the selected travel day. The mean and standard deviation of this coverage metric were separately calculated for rMove and Cuebiq trips over all tracts. "Low-coverage" areas are defined as tracts with coverage lower than mean coverage minus one standard deviation. Figure 10 shows tracts with low coverage for Cuebiq and rMove trips, respectively—revealing that Cuebiq trips are more evenly distributed than rMove over tracts for the given day of data.

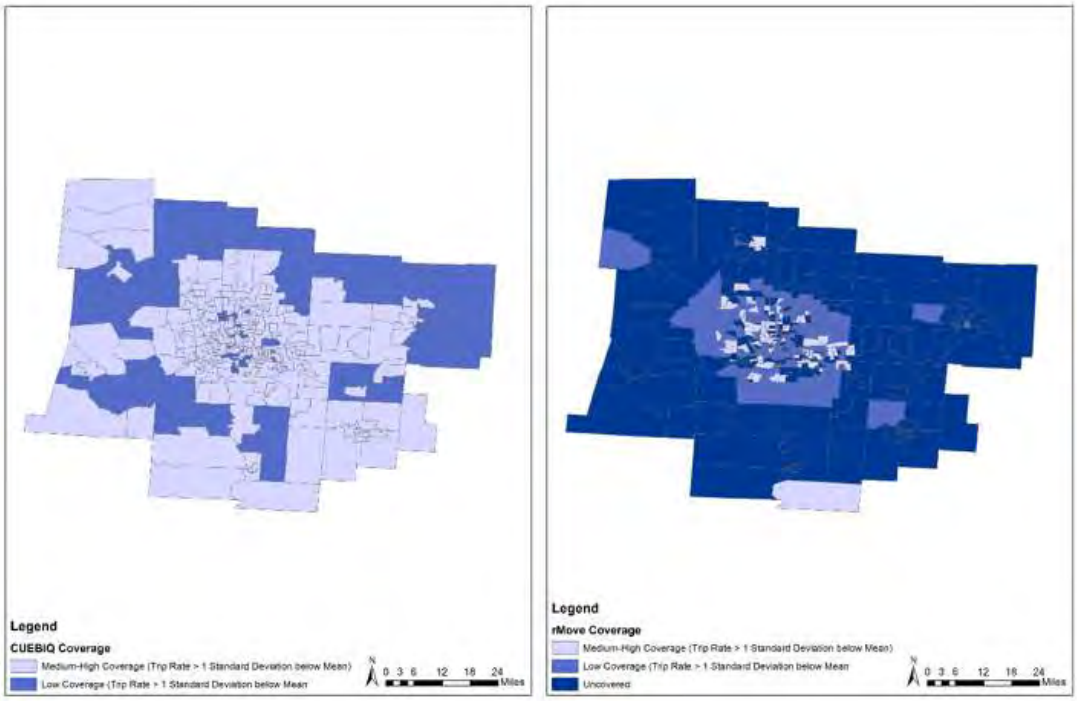

Figure 10. Low coverage tracts in rMove vs. Cuebiq. (Sources: RSG, Cuebiq, Caliper Corp., 2017).

Figure 11 illustrates the normalized rMove and Cuebiq trip densities, by census tract. rMove data show a higher rate of trips in Ohio State University, downtown, and nearby areas while Cuebiq trips show somewhat uniform distribution over downtown and suburban areas as a higher percentage of trips are expected in the central business district.

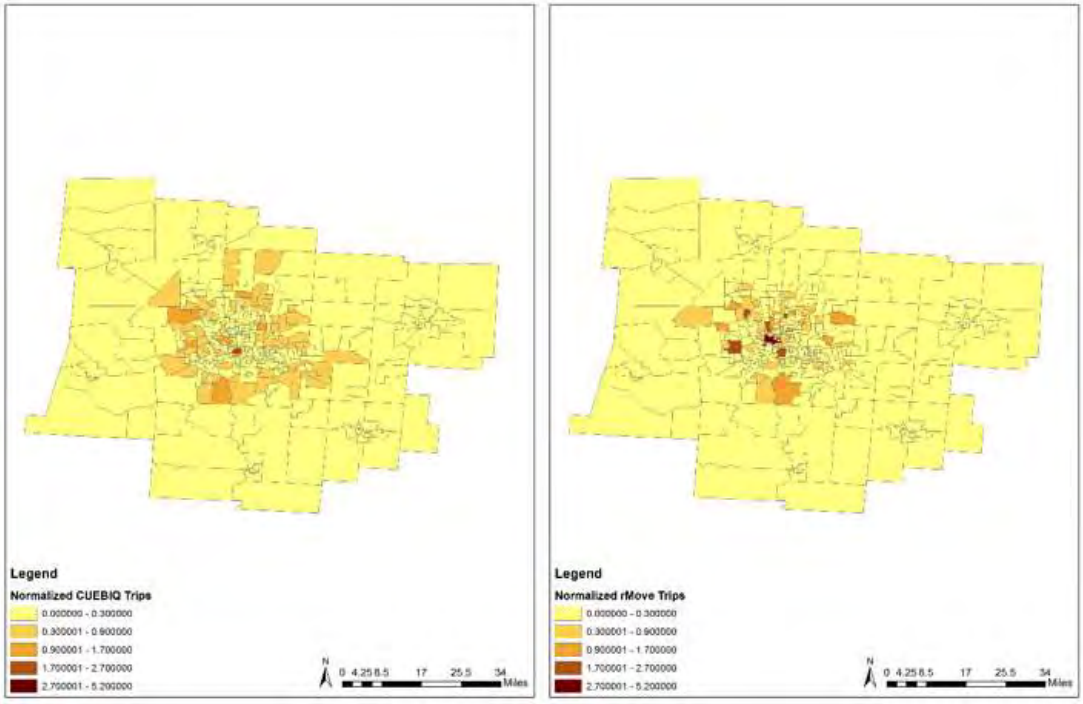

Figure 11. Normalized Cuebiq and rMove trip coverage, by tract. (Sources: RSG, Cuebiq, Caliper Corp., 2017).

3.6 Cuebiq and rMove: Equivalent Data Comparison

3.6.1 Matching Process: Cuebiq and rMove

Two matching process waves identified rMove users who appeared in the Cuebiq dataset. The first wave matched devices based on locations: cumulative "dwell" time was calculated for each location (with a 328-feet [100-meter] radius of buffer) per device in each dataset (i.e., the amount of time spent at a location by that device), and users were matched based on their highest "dwell" locations. Data used for the matching process were limited to two days in April 2017. The project team plotted locations among device pairs sharing more than one high-dwell location and evaluated these pairings to identify equivalent devices and possible device pairs. Figure 12 and Figure 13 show all locations for a given day from two sets of equivalent rMove and Cuebiq devices, while Figure 14 shows locations from one set of potentially equivalent devices.

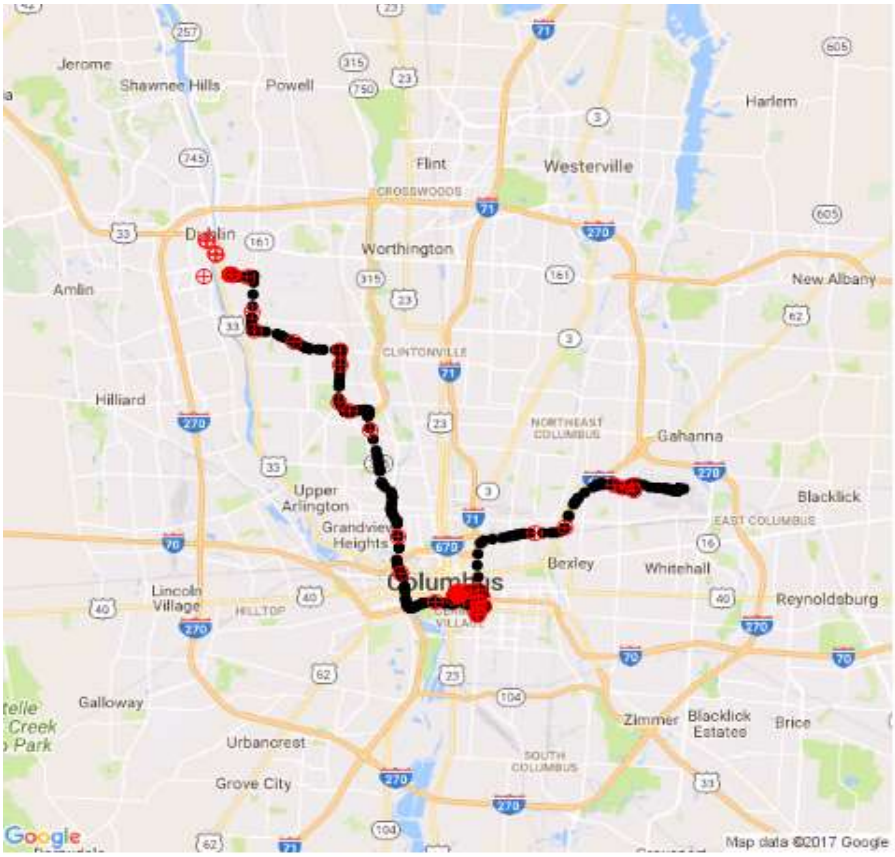

Figure 12. Locations of equivalent rMove devices. (Sources: RSG, Google Maps, 2017).

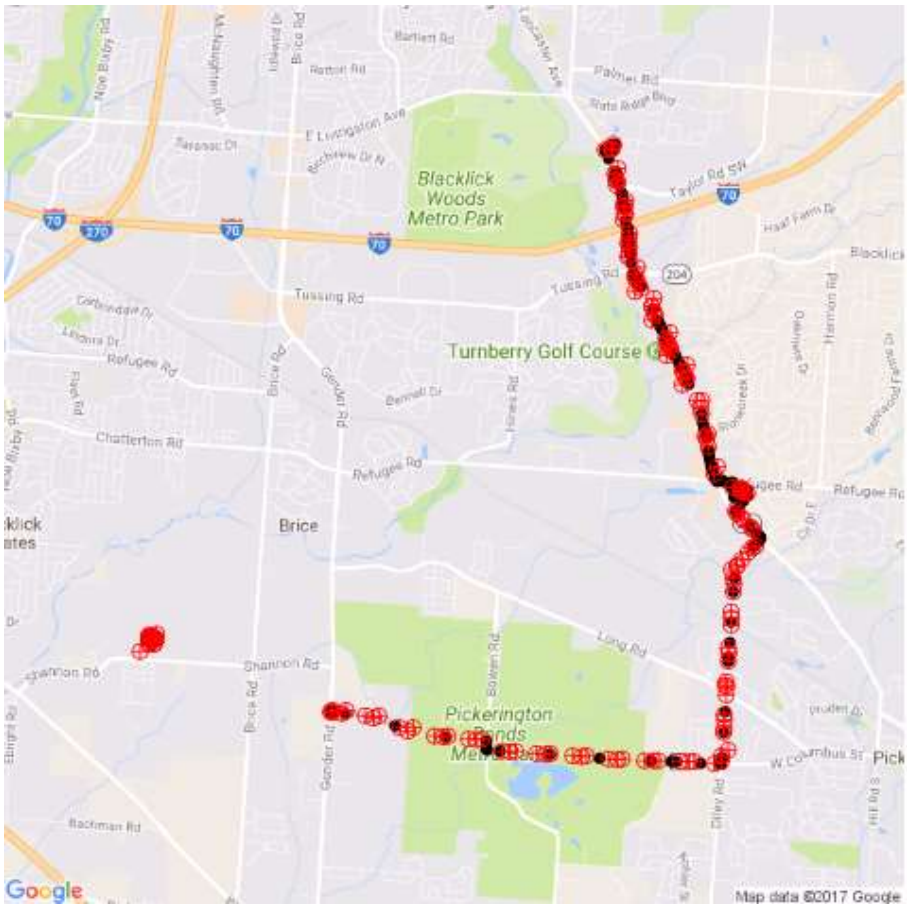

Figure 13. Locations of equivalent Cuebiq devices. (Sources: RSG, Cuebiq, Google Maps, 2017).

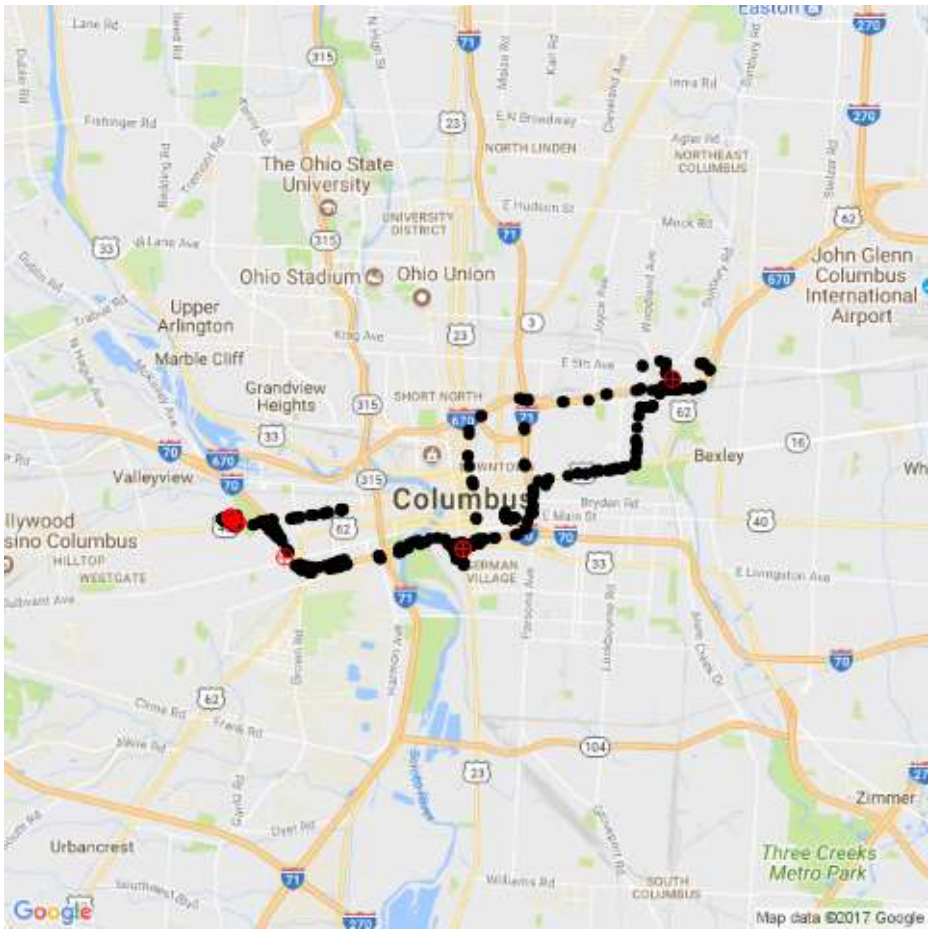

Figure 14. Locations among one set of potentially equivalent rMove and Cuebiq devices. (Sources: RSG, Cuebiq, Google Maps, 2017).

The project team used the known set of equivalent devices and possible equivalencies to validate the trip algorithm and to inform the second wave of device matching. This process produced a more holistic set of device matches to statistically evaluate. The second wave joined rMove and Cuebiq location data based on location and time, and compared average normalized spatial and temporal distance between device locations to determine acceptable matches. Pseudocode for this algorithm is included in the Appendix.

The algorithm calculated spatial and temporal distances between relevant points in each possible device pairing. It assigned a confidence level to each pairing based on spatial/temporal difference values for known equivalent pairs. Pairs were also evaluated for "completion" relative to the rMove dataset (i.e., the percentage of rMove location data that corresponded to an equivalent location in the Cuebiq dataset). Table 6 and Table 7 show completion and confidence ranges used for evaluation. Because spatial difference and temporal difference are normalized, they no longer are associated with valid units of distance/time. Distance and time were initially measured in meters and seconds before normalization.

Table 6. Thresholds for confidence categories among rMove/Cuebiq user pairings.

	Normalized spatial distance	Normalized temporal difference
Good	<= 0.3	<= 0.4
Moderate	0.3 to 0.5	0.4 to 0.65
Low	> 0.5	> 0.65
Incomplete	NA	NA

Table 7. Thresholds for completion categories among rMove/Cuebiq user pairings.

	Pct. relative to rMove dataset
Good	>= 40%
Moderate	25% to 40%
Low	15% to 25%
Incomplete	< 15%

3.6.2 Matching Results: Cuebiq and rMove

The second matching process initially resulted in 29,731 possible equivalent pairs among 222 rMove users and 95,697 Cuebiq users. Because the first part of the matching algorithm ignores definite nonmatches, the set of possible pairs is fewer than 222 multiplied by 29,731. Table 8 includes the results of pair categorization, by completion and confidence. **Because of the small number of users within the rMove sample who could be paired to Cuebiq data, robust statistical analysis of confidence among pairs is not feasible**, though future work may facilitate such analysis. While few pairings result in high confidence and completion relative to possible pairings (given each rMove user presumably corresponds to a maximum of one Cuebiq user), the maximum number of truly equivalent pairs is 222. While the table below is not mutually exclusive (**one rMove user may have a possible pairing with several Cuebiq users**), unique pairs categorized with either “good” or “moderate” confidence and completion sum to 59, which is 29% of total rMove devices. This finding aligns with Cuebiq’s own estimates of reaching 25% of the adult population in the United States.

Table 8. Number of user pairings in each confidence/completion category.

	Good Completion	Moderate Completion	Low Completion	Incomplete
Good Confidence	8	16	44	2,680
Moderate Confidence	41	156	331	25,244
Low Confidence	3	9	2	2,798

Demographics among the pairs with "good" or "moderate" confidence and completion, shown in Figure 15 and Figure 16, reveal the demographic makeup of Cuebiq users and their general representativeness of the population. These charts show demographic details of the rMove users who were identified in the Cuebiq dataset compared to the rMove sample at large, and the weighted sample, which is weighted based on the census and American Community Survey. Demographic details were collected for rMove users as part of the travel study. The set of rMove users with Cuebiq equivalents are more likely to be younger than both the unweighted and weighted overall rMove sample, but these users are less likely to have midlevel incomes between $50,000 and $100,000.

In many cases, demographics are similar between the paired users and the unweighted rMove sample, which is unsurprising as both samples represent smartphone users, who are more likely to be younger and have relatively higher incomes. The paired users indicate a closer-to-representative percentage of the 18–24 age group in the Cuebiq dataset compared to the rMove unweighted sample. The age bias appears to be the more significant issue in terms of demographic representativeness, although income bias is also an issue. It may be necessary to control for and expand LBS data to correct for these demographic biases to avoid skewing travel metrics.

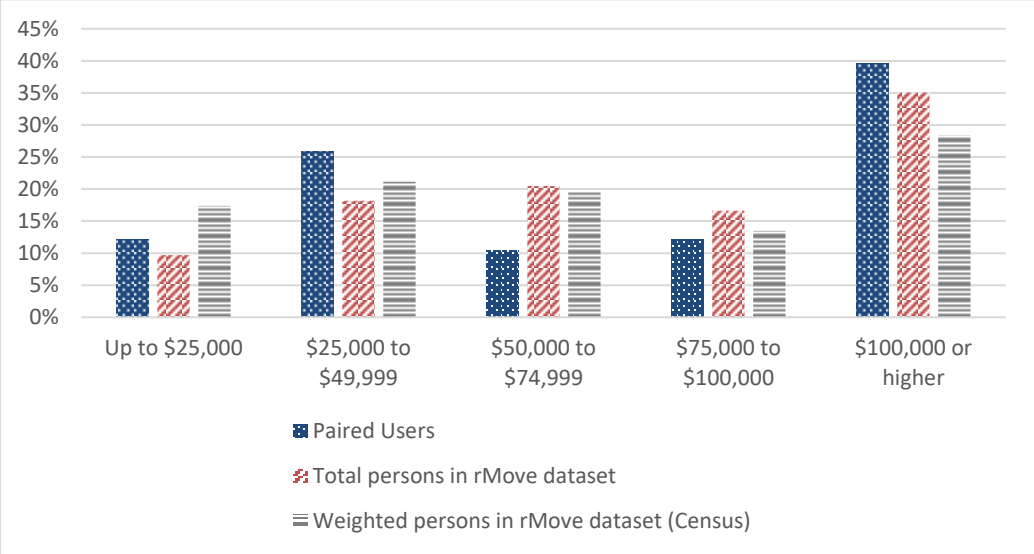

Figure 15. Income of paired users vs. persons in rMove dataset (weighted and unweighted). (Sources: RSG, Cuebiq, 2017).

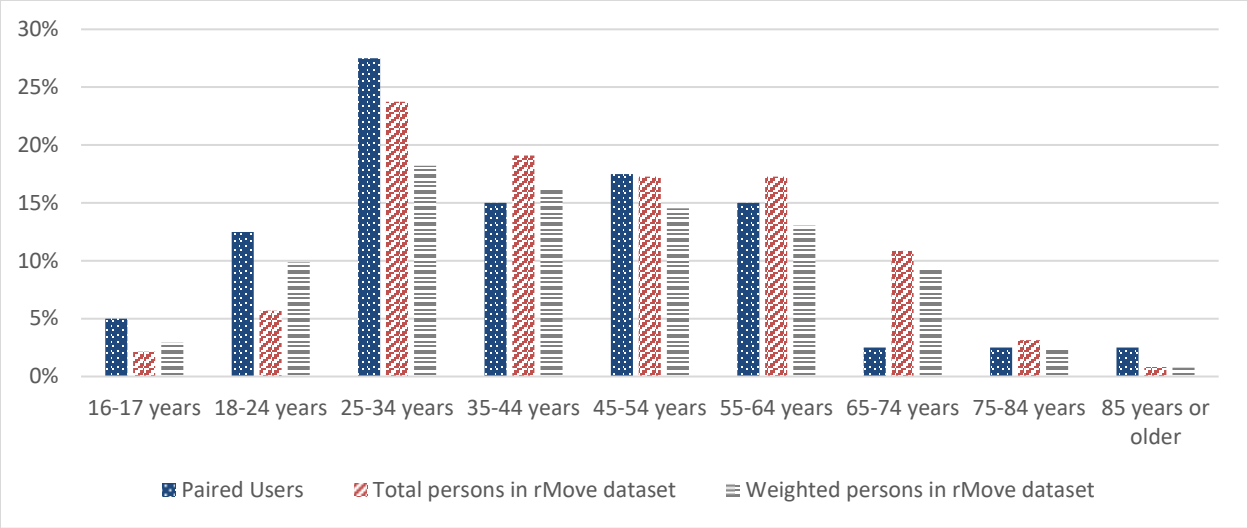

Figure 16. Age of paired users vs. persons in rMove dataset (weighted and unweighted). (Sources: RSG, Cuebiq, 2017).

While the sample size of paired traces is too small to achieve statistical significance among pairings with both good confidence and completion, rMove collected 5.1 trips per device and Cuebiq trip inference resulted in 4.1 trips per device for these "good" pairings. The maps in Figure 17 and Figure 18 illustrate trips inferred from Cuebiq data and trips recorded by rMove for one user pairing categorized as "good" confidence and completion. Each plot shows the same day of trips for the user in each dataset. Red points denote trip ends, while different shades of blue mark points along distinct trips. Because round trips covering the same route exist in both datasets, distinct trips are somewhat difficult to identify; however, a difference in travel collected by Cuebiq vs. rMove for this device is apparent. **Assuming data from Cuebiq and rMove represent the same user**, this implies that Cuebiq data exclude some portion of this user's travel. Moreover, the missing trips are shorter, as observed in the aggregate comparisons.

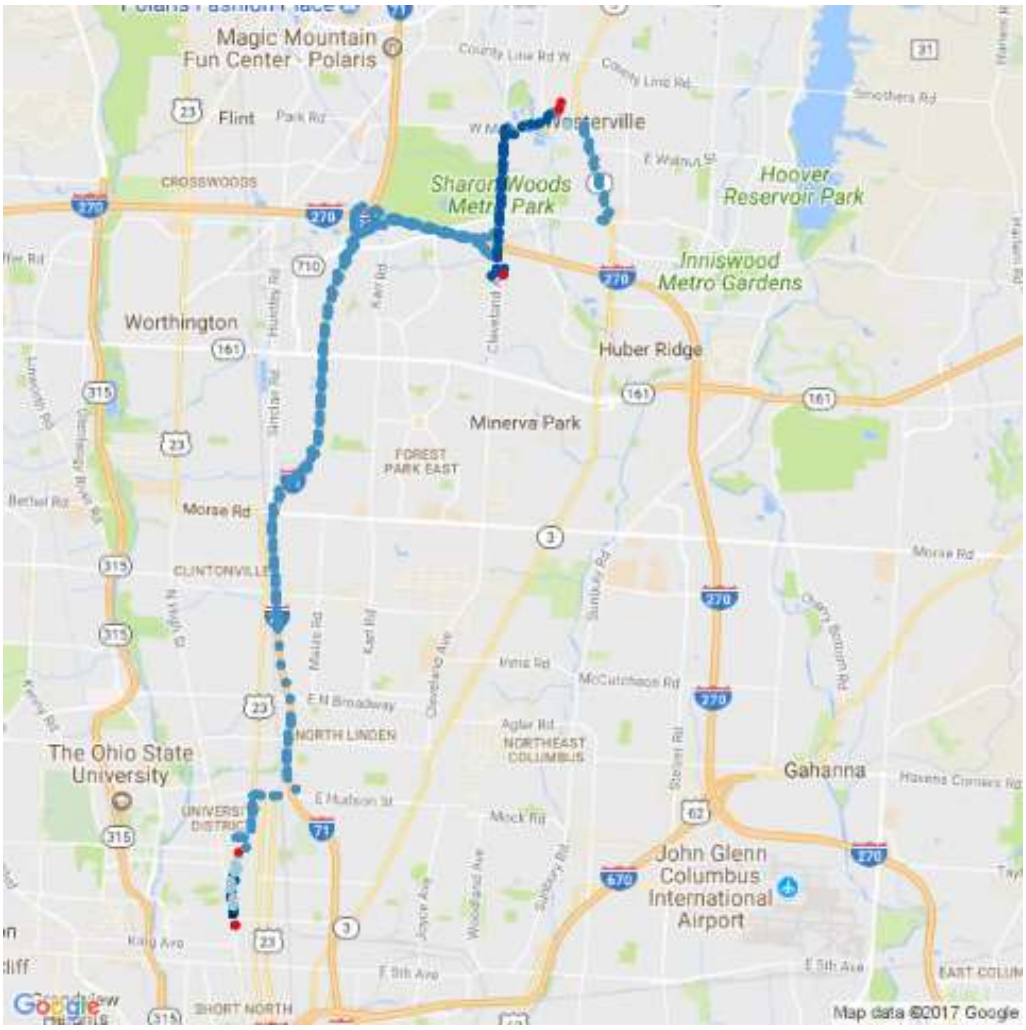

Figure 17. Trips inferred from Cuebiq for one user match over one day. (Sources: RSG, Cuebiq, Google Maps, 2017).

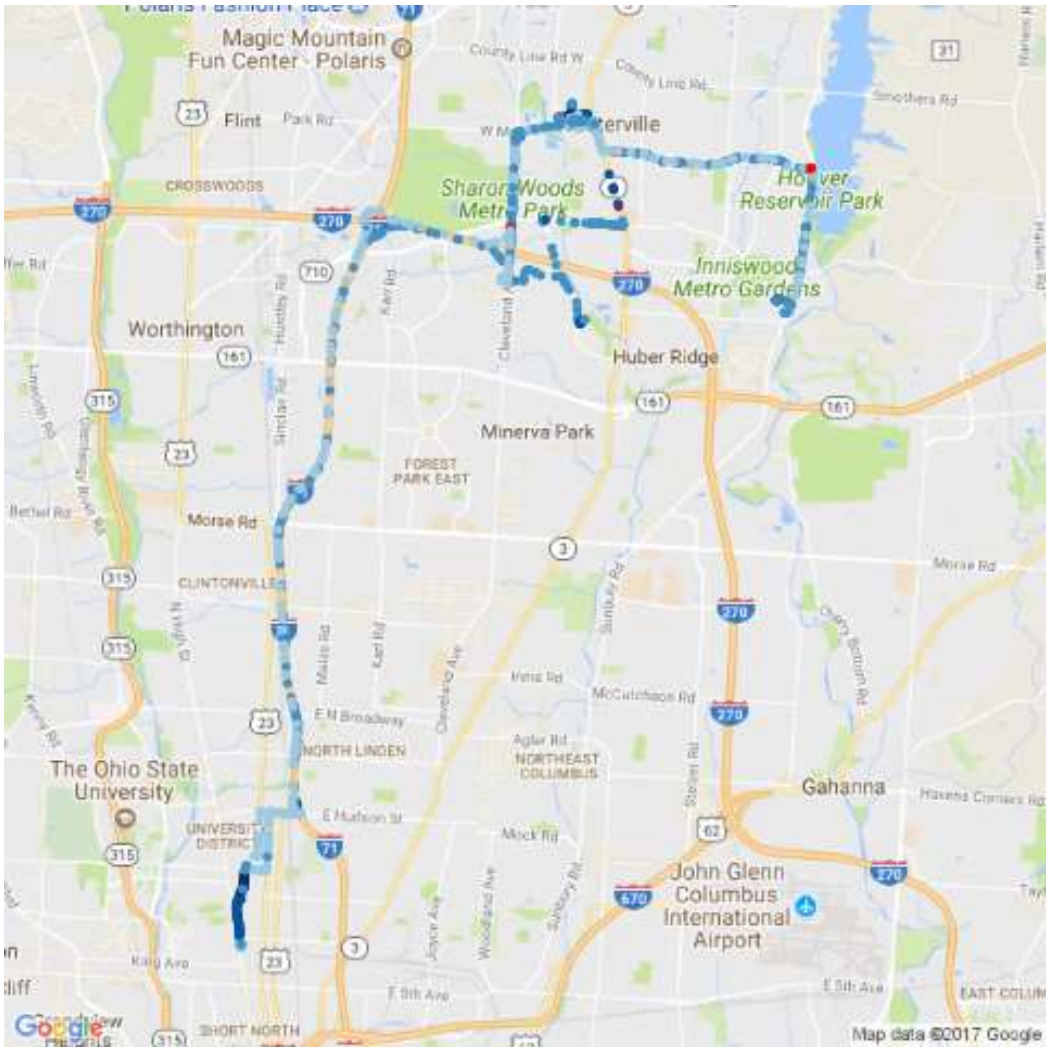

Figure 18. Trips recorded by rMove for one user match over one day. (Sources: RSG, Cuebiq, Google Maps, 2017).

3.7 Gaps in Knowledge and Understanding

The evaluation of Cuebiq location data and inferred trips illustrates the strengths of Cuebiq data as a large, passively collected location dataset. The evaluation also reveals several areas in which Cuebiq data may be incomplete or biased. The most obvious advantage of Cuebiq data is the level of user penetration and geographic coverage achieved given the hundreds of apps Cuebiq leverages for data collection. Coverage maps at the location and trip level, and the large number of devices represented in just one day of data, demonstrate this strength. However, based on the evaluation, data collected do not appear to holistically represent each user's complete travel record. In particular, short-duration trips are difficult to derive from the Cuebiq data.

Despite these findings, knowledge and understanding gaps persist regarding Cuebiq data. First, data collected come from an unknown (proprietary) list of applications, potentially introducing bias to data collected on each device and from each user (e.g., if apps are targeted to specific age groups, transit riders vs. drivers). Second, demographics of Cuebiq users are unknown. This

information can be imputed with some level of certainty from the location data provided, but this introduces a level of error on top of the uncertainties inherent to the dataset. Third, bias among users who remain in the dataset for multiple days vs. those who drop out after one day is difficult to measure and is likely tied to certain types of users or use of certain apps.

While these gaps in knowledge are based on unknown aspects of Cuebiq data, several additional questions were unaddressed in the scope of this research but could be analyzed using similar methodologies, including the following:

- Identify mode and purpose of trips present vs. trips missing in the Cuebiq dataset.
- Identify variation in trip characteristics (other than duration), by weekday.
- Understand uncertainty/error associated with matching users in separate databases.

3.8 Challenges Based on Gaps in Knowledge and Understanding

This analysis points to two important and practical conclusions regarding Cuebiq data and survey data:

- Cuebiq data provides far greater geographic and temporal coverage than smartphone surveys (even more when compared to traditional surveys), and reliance on survey data alone could limit the ability of models or analyses to present an accurate and complete picture of travel patterns. Biases in the survey sample itself must also be accounted for when leveraging the coverage advantages of Cuebiq data.
- Uninformed use of Cuebiq data without correcting for systematic bias related both to demographics and to temporal sparsity at the individual level or the infrequency of observations, which skews the observed duration of trips and other activities, would result in faulty analyses and conclusions.

These two conclusions point to the promise and value of data fusion to leverage the strengths of both passive LBS data and survey data together. Combined, both types of data could be used by practitioners to develop a more complete and accurate picture of travel patterns than either data type could provide alone. Because data provided from Cuebiq only encompassed one month of data—and the feasible scope of analysis was limited to a specific region—transferability of this analysis to other geographies and times of year is unknown. The use of the set of smartphone apps from which Cuebiq data are drawn could vary by geography and other time frames. While many important details may differ from region to region, it seems unlikely that regional or temporal variations in the data would affect general conclusions from analyzing these data. As the analysis presented in this volume indicates, Cuebiq data are a promising resource for transportation applications, and these data can be valuable for model estimation and other uses if analysts are aware of the challenges of using these data and prepared to account for the inherent biases.

4.0 Appendix

4.1 Trip-Inference Algorithm

1. Filter points by 328 feet (100 meter) accuracy.
2. Within ID, sort by timestamp.
3. Loop-through points:
    a. Calculate implied speed between points t and t-1.
    b. If speed exceeds 1.5 mph (2.4 km/h) (half walking speed) and origin not defined, then define point t-1 as an origin.
    c. If speed exceeds 1.5 mph (2.4 km/h), then define point t as waypoint.
    d. If speed is 2.4 km/h or below and distance from point t to origin is greater than 164 feet (50 meters), then define point t as a destination.

4.2 Cuebiq-rMove Device Pairing Algorithm

1. Merge Cuebiq and rMove location datasets where locations are within 0.62 miles (1 kilometer) and +/- 5 minutes of one another (for the same day).
2. Calculate haversine distance and duration between merged locations.
3. Normalize distance and duration between merged locations on a scale of zero to one, where the smallest value is zero and the largest value is one.
4. Calculate the average normalized distance and duration between locations for each device pairing.
5. Calculate the percentage of paired locations/total locations collected for the day by rMove for each device pairing ("completion" percentage).
6. Derive a confidence level for each device pairing given the following "confidence" thresholds (based on normalized distance/duration values for known equivalent pairs between rMove and Cuebiq devices):
    a. Good confidence: Average normalized time difference <= 0.4; average normalized distance <=0.3.
    b. Moderate confidence: Average normalized time difference between 0.4 and 0.65; average normalized distance between 0.3 and 0.5.
    c. Low: Average normalized time difference >0.65; average normalized distance >0.5.
7. Derive a completion level for each device pairing given the following "completion" thresholds (based on completion percentages of known equivalent pairs):
    a. Good completion: 40% or greater.
    b. Moderate completion: 25–40%.
    c. Low completion: Less than 25%.

NOTICE

This document is disseminated under the sponsorship of the U.S. Department of Transportation in the interest of information exchange. The United State Government assumes no liability for its contents or use thereof.

The United States Government does not endorse manufacturers or products. Trade names appear in the document only because they are essential to the content of the report.

The opinions expressed in this report belong to the authors and do not constitute an endorsement or recommendation by FHWA.

This report is being distributed through the Travel Model Improvement Program (TMIP).

U.S. Department of Transportation
Federal Highway Administration
Office of Planning, Environment, and Realty
1200 New Jersey Avenue, SE
Washington, DC 20590

April 2018

FHWA-HEP-20-022

U.S. Department of Transportation
Federal Highway Administration