Big Data Sources from GPS-enabled Smartphone Applications: An Exploratory Analysis of Transit App Data

Candace Brakewood · Niloofar Ghahramani · Coline Remy · Jonathan Peters

Abstract This paper considers a new automated transit data source, which is passively collected data from a multi-city transit and shared mobility smartphone application known as "Transit." A three-part analysis is conducted. First, this new data source is qualitatively compared to two commonly used transit data sources: travel surveys and automated fare collection (AFC) data. The results suggest that app data may provide unique insights about transit riders in some situations that are difficult to capture using traditional data sources. One example is studying transit users across multiple cities; therefore, the second part of this paper exploits the multi-city nature of this smartphone app to identify transit users who are intercity travelers. Another unique property of the app is that it continuously collects data, including during extreme and unusual events; in light of this, the third section of this paper conducts an exploratory analysis of app usage during a transit system shutdown due to an extreme weather event. These two quantitative analyses – identifying intercity travel and extreme event usage patterns – help to demonstrate the unique properties of the smartphone app dataset, which has many potential future uses in transit planning, management, and operations.

Keywords: Transit app · Smartphone application · Intercity travel · Extreme event

Candace Brakewood, PhD
Department of Civil and Environmental Engineering, University of Tennessee
Knoxville, Tennessee, USA
Email: cbrakewo@utk.edu

Niloofar Ghahramani
Department of Civil Engineering, The City College of New York
New York, New York, USA
Email: nghahra000@citymail.cuny.edu

Coline Remy
Ecole Nationale des Travaux Publics de l'Etat
Vaulx-en-Velin, Lyon, France
Email: coline.remy@entpe.fr

Jonathan Peters, PhD
Department of Business, The College of Staten Island
Staten Island, New York, USA
E-mail: jonathan.peters@csi.cuny.edu

1 Introduction

One of the fundamental components of transit planning is understanding passenger travel patterns. Rapid adoption of mobile phones has led to the creation of new data sources that can be used to study travel patterns, particularly smartphones that are GPS-enabled or location aware. The widespread usage of smartphone applications (or "apps") to access real-time transit information presents a unique opportunity to utilize the backend data generated by these apps to study transit travel behavior. In light of this, the objective of this paper is to conduct an exploratory analysis using an emerging "big data" source: backend data from a smartphone app that provides transit and shared mobility information.

This paper proceeds as follows. First, background information is provided about the specific app that is the focus of this paper, known simply as "Transit," and the contents of the dataset are discussed. Then, this new data source is qualitatively compared to two commonly used transit data sources – travel surveys and data from automated fare collection (AFC) systems – and potential advantages are identified. After this, two quantitative analyses are conducted, and both focus on the New York City region. The first quantitative analysis exploits the multi-city nature of the smartphone app, and transit users who are intercity travelers are identified. The second analysis explores app usage during a transit system shutdown due to an extreme weather event, which highlights the unique ability of the app to continuously collect data. The last section of this paper presents conclusions and areas for future research.

2 Background

2.1 Transit App

Transit is a company based in Montreal, Canada that has developed a freely available smartphone application (Transit, 2018). In 2012, the company released the first version of its application for iPhone, and in the initial version, the app provided transit schedule information for Montreal, Toronto, and Quebec City. Since then, Transit launched an Android version of its application and has expanded to over 125 cities in nine countries, including widespread coverage in the United States and Canada. Additionally, Transit has added many features, including real-time transit information, transit trip planning using schedule information, and multimodal support (including bikesharing, carsharing, and Uber). Fig. 1 shows the Transit app Android interface displaying real-time information for nearby transit lines and Uber in New York City (left); carshare, transit, and Uber information in Washington, DC (center); and Divvy bikesharing and transit information in Chicago (right).

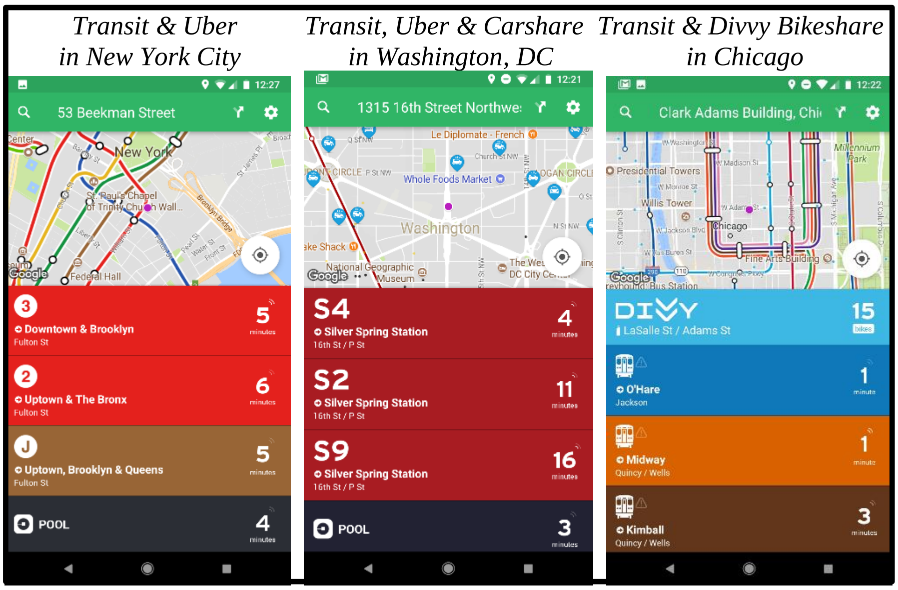

Fig.1 Transit app screenshots (Android version 4.4.4)

2.2 Data Tables

This section summarizes the structure of the Transit app dataset. Before continuing, it should be noted that personal information such as names or demographic information are not stored in this dataset to protect the anonymity of users.

Data are generated by the Transit app whenever a user opens the apps; their interactions with the app are stored in a backend database as a record of operations with the software. Each interaction is called a "session" and is identified by a unique identification number and timestamp. In order to provide relevant information for nearby transit service, the app needs to identify the location of the device. Therefore, each time the app is opened, a session is generated, the device location is sent to the server, and this record is then stored in the backend database (Brakewood et al., 2017).

The database generated by user interactions with the Transit app is divided into tables that capture data pertaining to the various functions within the app. Table 1 describes thirteen of these tables and the data that are captured in each. It should be noted the structure of the backend can change when there are application updates to include new features, and the tables discussed in the following paragraph represent a sample from the backend provided to the authors in 2016 (Brakewood et al., 2017).

As can be seen in Table 1, the first file is called the ***locations*** table; this is the primary file in the Transit app backend server that contains interactions every time a user opens the app (i.e., "sessions"). The second file is called ***session complete***, and it provides a concise, event-based view of each session. The third file, known as ***placemarks***, includes data about an optional feature to store frequently used places, such as home and work, which can be saved within the app to help users find relevant information quickly. The next three files shown in Table 1 pertain to shared mobility features in the Transit app. The fourth table is called ***sharing system actions***, and it provides information on the booking of carshare, bikeshare and other services. The fifth table is ***sharing system purchase***, which provides information on the successful purchase of shared mobility passes, primarily bikeshare. Sixth is known as ***Uber request***; in the 2016 version of the Transit app, the user could request an Uber via the Transit app, but that request would then be handed off to Uber's app for fulfilment. The remaining files shown in Table 1 are not used in the following analyses and are not discussed here; for additional details, the reader is referred to Brakewood et al. (2017).

Table 1 Summary of tables in the Transit app dataset

#	Table Name	Description of Contents
1	Locations	Includes the location (latitude/longitude) and a unique session ID for each time the app is opened.
2	Session Complete	Provides an event based view of each session, including the beginning and ending location.
3	Placemarks	Includes location data from an optional function that stores places users often go (e.g., home or work).
4	Sharing System Actions	Provides data on the booking of carshare, bikeshare and other services, including the location of shared vehicles.
5	Sharing System Purchase	Provides purchase records for shared vehicles, which are primarily bikeshare passes.
6	Uber Request	Lists requests for service from Uber, which are then handed off to Uber's app for fulfillment.
7	Trips	Contains information about usage of the trip planning feature, including start and end coordinates.
8	Nearby View	Contains information about the transit routes presented to a user in each session upon opening the app.
9	User Feed Session	Includes the number of times the app is opened and the different transit agency's data accessed by the user.
10	Installed App	Reports on other installed apps on the user's device that can impact functionality, such as the Uber app.
11	Feed Download	Provides a summary of activity on the Transit app by day.
12	Favorite	Provides information on user designated favorites in terms of transit routes.
13	Device	Contains a Transit app specific identification number (device ID), device type, model of device, operating system, version of Transit app, and last date of app use.

Adapted from Brakewood et al. (2017).

3 Qualitative Comparison of Data Sources

This section provides a brief comparison of commonly used transit travel behavior data sources – specifically, travel surveys and data from automated fare collection systems – with the new dataset from the Transit app (Brakewood et al., 2017). ***Travel surveys*** are one of the most widely used data sources to study transit travel, and they are often conducted as household travel surveys or in stations/onboard transit vehicles (Schaller, 2005; Pratt et al., 2018). ***Automated fare collection (AFC)*** systems, which have become commonplace in transit systems over the last two decades, enable passengers with contactless smart cards to pay fares by "tapping" their cards at faregates or on fareboxes; this provides a rich source of data about transit travel (Bagchi & White, 2005; Pelletier et al., 2011). These two traditional data sources are qualitatively compared with the ***Transit app data*** on four dimensions: [1] geographic area, [2] timescales, [3] modes, and [4] sample size, and the results are shown in Table 2.

Table 2 Comparison of Transit app data with traditional data sources

	Travel Survey Data	**Automated Fare Collection (AFC) Data**	**Transit App Data**
Geographic Area	Single region	Single region	Multi-region
Timescales	Cross-sectional	Continuous (when transit system open)	Continuous
Modes	Transit and/or others	Transit	Transit, Ride-hailing, Bike-sharing, Carsharing
Sample Size	Small	Large	Large
Adapted from Brakewood et al. (2017)			

As can be seen in Table 2, the first dimension is the ***geographic area*** for each dataset. Surveys are typically conducted within a single metropolitan by the local transit agency or planning organization, with some notable exceptions (e.g., the American Community Survey from the United States Census Bureau). Similarly, AFC systems are typically not compatible across different metropolitan areas (Acumen Building Enterprise, Inc. 2006), and subsequently, the data generated by these systems typically encompasses transit travel records for a single region. Because the Transit app includes over 125 metropolitan regions in nine countries, it collects data for many different metropolitan regions, and this data can then be used to understand intercity travel patterns, which will be discussed more later in this paper.

The second dimension, ***timescales***, refers to the period when data are generated. Most travel surveys are conducted at a single point in time and therefore provide a cross-sectional snapshot of transit travel behavior. Panel surveys conducted at multiple points in time are also possible, but in practice, this is infrequently done. AFC data are collected whenever the transit system is in operation, which means this data source is (nearly) continuous in time. The Transit app functions twenty-four hours a day, and therefore continuously collects user interaction data. The continuous nature of this dataset could advantageously allow for analysis of travel behavior during events that are difficult to capture in cross-sectional datasets or during transit system shutdowns (e.g., due to extreme weather events).

Third is the ***modes*** of transportation for which data are collected. Surveys conducted by the local transit agency typically focus on transit travel and occasionally include a limited number of questions about access and egress mode to transit stations/stops; surveys conducted by planning organizations are more likely to capture multiple modes, including trips made by automobile, transit, and non-motorized modes. AFC systems typically include only data pertaining to transit travel, since data are captured when transit fares are paid. The Transit app, on the other hand, integrates information about numerous shared mobility modes, including bikesharing, carsharing, and ridehailing (Uber), and allows users to purchase passes and utilize shared mobility vehicles in some cities, such as Divvy bicycles in Chicago.

Last, the ***sample size*** refers to the quantity of data collected. Travel surveys typically sample only a small portion of the population of interest. Despite the relatively small sample size, it is worth highlighting that surveys are usually conducted using methods that aim to be representative of the entire population of interest. AFC systems generate vast quantities of data, and depending on the level of AFC adoption by riders in a region, they can represent all or nearly all transit riders. The Transit app also generates vast quantities of data because many riders use the app on a daily basis (Ghahramani & Brakewood, 2016). However, it is important to note that the sample for which the Transit app data is generated depends on the app's adoption and utilization levels, which could be biased compared to the overall population; future research is needed to understand the potential biases of this new dataset.

In summary, the Transit app dataset has numerous potential advantageous properties, which include the ability to study [1] multi-city transit travel, [2] continuous records over time, such as during transit system shutdowns, [3] multi-modal trips, and [4] large sample sizes. These aspects have the potential to provide unique travel behavior information that is not easily captured in traditional transit datasets, such as travel surveys and AFC data.

4 Analysis of Intercity Travel

The multi-city nature of the Transit app provides a unique opportunity to understand how transit riders travel between cities. The objective of this section is to identify intercity travelers going to or from a single metropolitan region (in this case, the New York City region). The following sections provide a summary of the Transit app data sample used in the analysis, the methodology to identify intercity travelers, and the results. Additional details can be found in Ghahramani et al. (2017).

4.1 Data Sample

The specific data sample used in the following analysis contains Transit app data for any user that opened the app at least once during a single month in 2014 in the New York City region. The raw dataset used for this analysis was provided directly by software developers at Transit to the research team in Comma Separated Values (CSV) format. The main file used for the following analysis is the ***locations*** table, which contains records for each time an individual user opens the app. When the app is opened, the user's location is sent to Transit's server based on the coordinates from the location services in their smartphone. Approximately 13 million records were sent to the Transit app server by over 169,000 unique devices (i.e., individual users) during the one-month period. Last, it should be noted that the Transit app developers offset all geographic coordinates contained in this dataset by a random number for privacy purposes.

4.2 Method

A three-step method was used to identify intercity travelers in the Transit app dataset.

Step 1: Data Cleaning

First, data cleaning was conducted to address some minor issues pertaining to inaccurate timestamps and geographic coordinates. A small number of records had dates other than the month of study and were removed. Also, some records were "simulated" sessions, meaning that the user moved the GPS point on the map interface of the app to a location other than their actual location; these "simulated" records were deleted. After removing the irrelevant dates and simulated locations, the dataset was reduced to 10,8444,349 records made by 146,597 unique devices.

Step 2: Geographic Bounding Box

Second, each record in the cleaned dataset was classified as being inside or outside the New York metropolitan area based on the user's location (latitude/longitude). The New York City region was selected for this analysis for three reasons: [1] it is a popular tourist destination (McGeehan, 2016), suggesting that there should be many intercity travelers; [2] New York City has the highest concentration of transit trips in the United States (McKenzie & Rapino, 2011); and [3] it is a region with high utilization levels of the Transit smartphone application.

The New York City region was defined using a bounding box drawn at the maximum and minimum latitude and longitude specified by local metropolitan planning organization, known as the New York Metropolitan Transportation Council (NYMTC, 2016). This boundary roughly represents the dimensions of the regional commuter rail network. Transit app usage within the borders was considered local, and usage outside of the box was considered another region (Ghahramani et al., 2017).

Step 3: Identifying Days Outside the New York Region

Third, each record was classified as being inside or outside the New York metropolitan area bounding box based on the location (lat/long) of the user. Then, the devices that had records both inside and outside of the bounding box were identified, and these smartphone devices were assumed to belong to intercity travelers.

4.3 Results

The results reveal that 3,778 unique devices (2.58% of the devices in the cleaned dataset) used the Transit app both inside and outside of the bounding box and were deemed to be intercity travelers. These 3,778 intercity travelers had a total of 552,280 records inside and 64,715 records outside of the New York metropolitan area during the one-month study period.

Fig. 2 displays the locations of app records from intercity travelers outside of New York metropolitan area. This map shows some of the 64,715 records - specifically those in the United States but outside of the New York metropolitan area - made by the 3,778 intercity travelers. As can be seen in Fig. 2, intercity travelers used the Transit app in many cities in the United States and likely made many long-distance trips as well as numerous shorter trips during the month of study.

Last, one intercity traveler was selected to visualize his/her travel patterns, and the results are shown in Fig. 3. In this map, each dot represents the user's aggregated Transit app records in that area; larger sized dots represent higher usage in that specific area. The user shown in Fig. 3 searched for transit information in Florida on four days during the month of study in 2014. On the fourth day, this individual used the Transit app in Florida and New York, and the map in Fig. 3 shows records in both the Fort Lauderdale airport and New York's LaGuardia airport (Ghahramani et al., 2017).

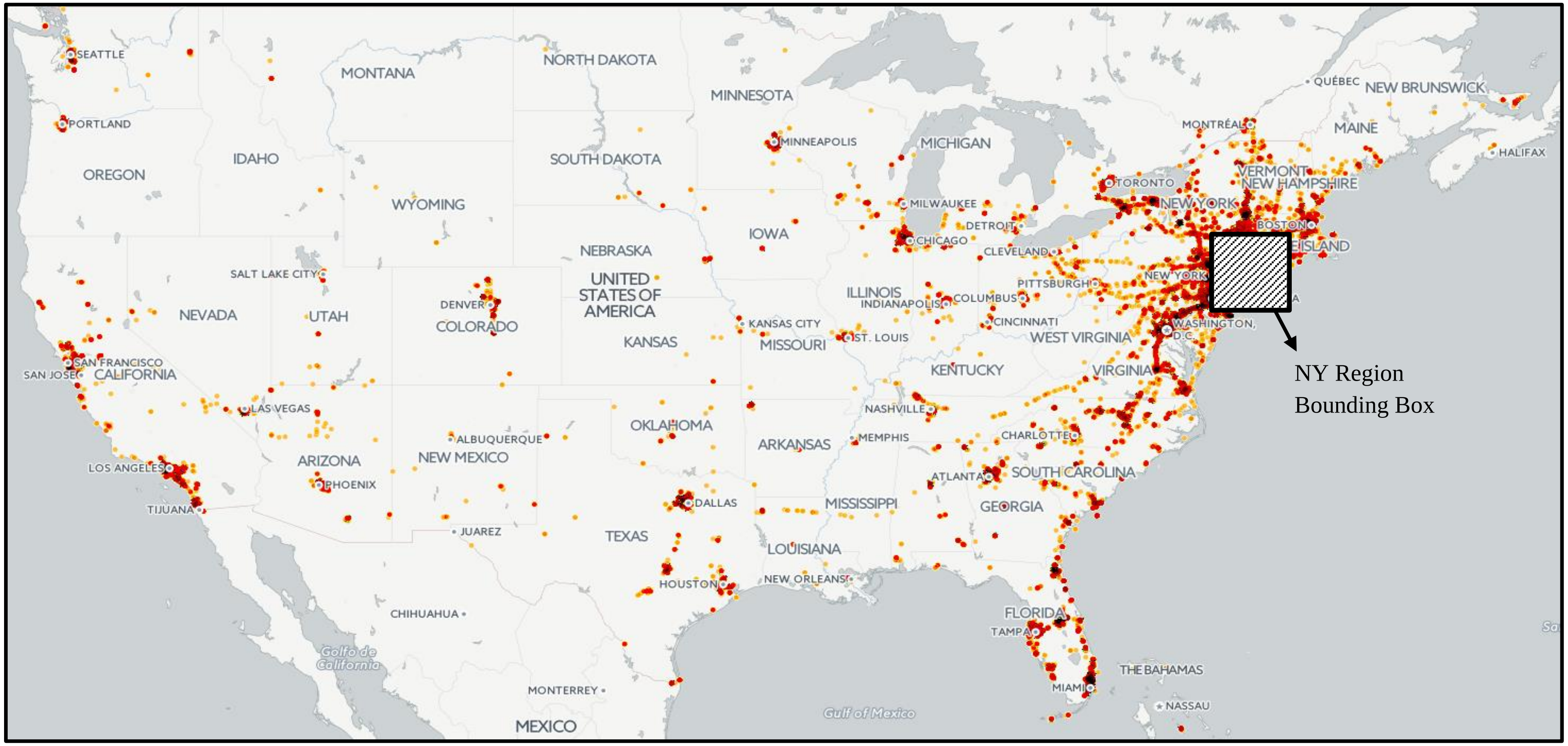

Fig. 2 Transit app records of intercity travelers outside of the New York region bounding box

Note: Records outside of the United States are not shown. Adapted from Ghahramani, Brakewood & Peters (2017)

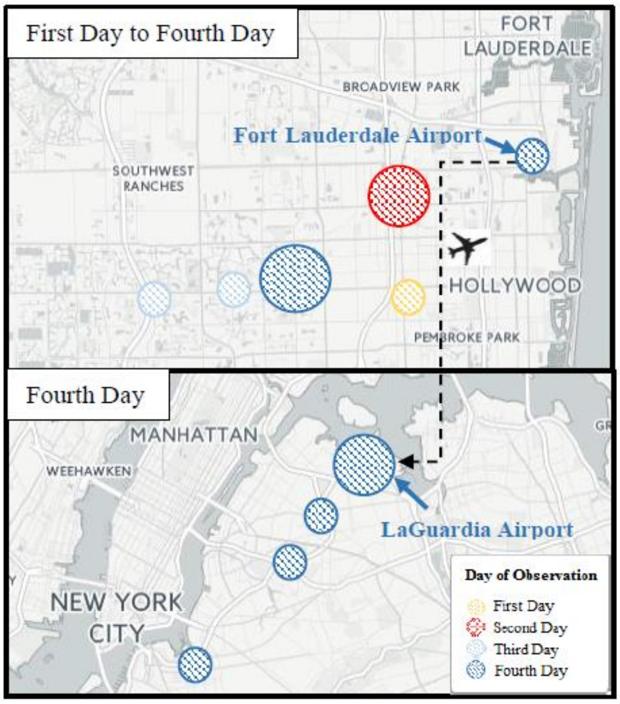

Fig. 3 Example of an intercity traveler who used the Transit app
Adapted from Ghahramani et al. (2017)

5 Analysis of an Extreme Weather Event

The continuous nature of data collection from the Transit app provides a unique opportunity to examine extreme and unusual events, which are typically difficult to capture in traditional transit datasets. The objective of analysis is to explore if and how the Transit app is used during a transit system shutdown due to an extreme event. The specific case is Winter Storm Jonas, which crippled the New York City region from Friday, January 22 until Sunday, January 24, 2016 (Solomon, 2016).

5.1 Data Sample

The dataset used in the following analysis contains Transit app data for any user that opened the app at least once during 418 days from 2015 and 2016 in any region; this is a significantly larger dataset than the one used in the previous analysis of intercity travel. The raw data were provided directly by software developers at Transit to the research team in JSON format, which was then converted to Comma Separated Values (CSV) files. The file used for the following analysis is the ***session complete*** table, which provides an event-based view of how a user interacts with the app, including the beginning and ending location for each session.

5.2 Method

A three-step method was used to explore Transit app usage during an extreme event. Additional details can be found in Remy et al. (2018).

Step 1: Period of Analysis

First, data for the study period were pulled from the larger Transit app dataset. The time period considered for this analysis was Friday, January 8, 2016 to Friday, January 29, 2016; that is two weeks before the snowstorm and one week afterward.

Step 2: Geographic Bounding Box

Second, records in the geographic area of analysis, which was the New York City region, were retrieved. Similar to the previous analysis, the New York City region was defined using a bounding box drawn at the maximum and minimum latitude and longitude specified by local metropolitan planning organization, known as the New York Metropolitan Transportation Council (NYMTC, 2016). Transit app records with locations inside this bounding box were used for the following analysis.

Step 3: Visualization and Rate of Transit App Use

Overall Transit app usage patterns were visually explored to assess if and how the app was being used during the snowstorm. For the three-week period of analysis, the total Transit app usage per hour was calculated. This was then visually compared to the mean app usage per hour during the two first weeks before the snowstorm when there were normal weather conditions (January 8 to January 15). Then, the rate of Transit app use was calculated by comparing mean hourly usage prior to the snowstorm with hourly app usage during the snowstorm. For every hour t, the rate of use $R(t)$ was calculated by dividing the hourly app usage, $U(t)$, by the mean hourly usage, $Q(t)$, during the reference period, which was the two weeks before the snowstorm. Equation (1) shows the rate of use formula:

$$R(t) = U(t) / Q(t) \tag{1}$$

where

$R(t)$ = rate of use;
$U(t)$ = hourly app use; and
$Q(t)$ = mean hourly app use.

5.3 Results

The results of the visualization are shown in Fig. 4, which shows hourly Transit app usage (the black line) compared to the mean hourly usage (the red line). Utilization of the app has a strong periodic weekly pattern. There is a peak during the morning rush hour and another during the evening rush hour for weekdays, and app usage decreases during the weekend. However, this typical pattern was interrupted during the snowstorm period, as can be seen in Fig. 4. The vertical line in yellow shows the period when the snowstorm hit, and the vertical line in blue displays when the transit system was (mostly) closed. During this period, the amount of overall app usage per hour was less than the mean but non-zero.

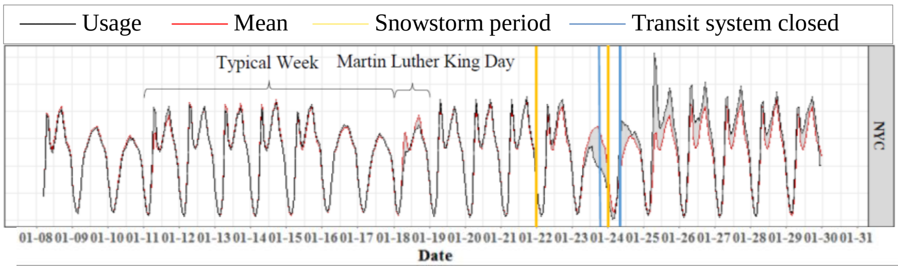

Fig. 4 Transit app usage and mean usage in New York City
Adapted from Remy et al. (2018)

The results of the rate of use calculation are shown in Fig. 5 as a black line. The red line in Fig. 5 shows 100%, which is the typical usage during the reference period. The rate of use is fairly consistent during the two weeks prior to the storm, but there is a dip in use on the Martin Luther King, Jr. holiday, which was Monday, January 18, 2016. During the snowstorm, the rate of app usage dropped below 100% on Saturday, January 23 before noon and remained low until the following day (Sunday, January 24). On Saturday, most transit service in New York City was shut down during the snowstorm (with the exception of some underground areas of the subway), and it was during this time that the rate of use reached its lowest point (approximately 65% of mean usage). After the snowstorm finished and transit service was largely restored on Monday, January 25, a peak appeared during the morning rush hour, and the rate of use reached about 165%. This may be because Transit app users wanted more information as they went back to work.

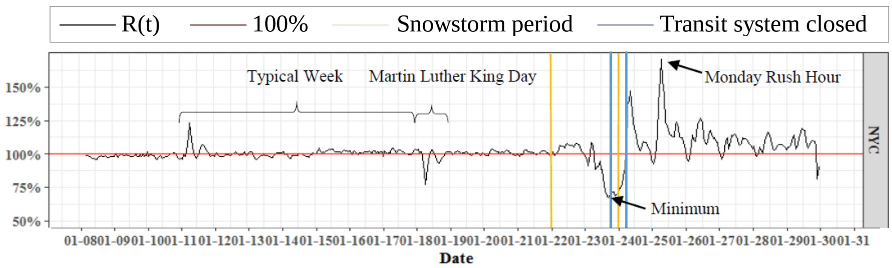

Fig. 5 Rate of Transit app use in New York City
Adapted from Remy et al. (2018)

Last, one app user was selected to visualize his/her travel patterns during the snowstorm. This user's Transit app records are represented by the dots in Fig. 6. The map on the left shows this app user's records on a normal weekday, which includes numerous late evening records in the Bronx that are presumably near his or her home because the device is frequently observed there. During the snowstorm, this user was observed checking the Transit app in numerous locations in the Bronx. suggesting that this person travelled short distances despite the snow.

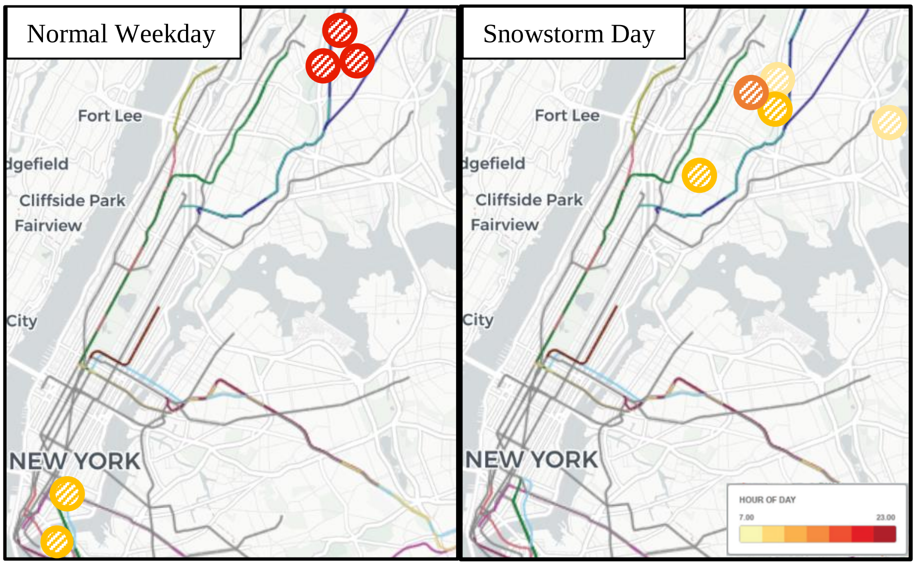

Fig. 6 Example of a Transit app user's records during Winter Storm Jonas

6 Conclusions and Areas for Future Research

This paper considered a new automated transit data source, which is passively collected data from a multi-city transit and shared mobility smartphone application known as "Transit." A three-part analysis was conducted. First, this new data source was qualitatively compared to two commonly used transit data sources: travel surveys and automated fare collection (AFC) data. The results suggest that app data may provide unique insights about transit riders in some situations that are difficult to capture using traditional data sources.

The second part of the paper exploited the multi-city nature of this smartphone app to identify Transit app users who are intercity travelers. App users were identified as intercity travelers if they were observed inside and outside of a bounding box drawn around the New York metropolitan area over the course of one month, and a total of 3,778 intercity travelers were identified (or 2.58% of devices in the sample). For

future research, a longer timeframe (i.e., at least one year) should be used to better capture the Transit app users' activities, particularly infrequent intercity travel.
The third section of this paper considered the continuous nature of data collection by the Transit app, which provides a unique opportunity to examine extreme and unusual events that are often difficult to capture in traditional transit datasets. An exploratory analysis of app usage during a transit system shutdown due to a large winter storm in the New York City region was conducted. The results reveal that, despite travel bans and a partial transit system shutdown, app users continued to search for transit information. For future research, additional analysis should be conducted to explore the locations of app users searching for information (e.g., searching for information from home while waiting for the storm to pass).

These two quantitative analyses – identifying intercity travel and extreme event usage patterns – help to demonstrate the unique properties of the smartphone app dataset, which has many potential future uses in transit planning, management, and operations.

Acknowledgements: The authors are very grateful to the Transit app for sharing data with the research team. They would also like to acknowledge City University of New York (CUNY) Graduate Center student Adam Davidson for his contributions to this research project. This research was funded in part by a 2015 CUNY Collaborative Incentive Research Grant (CIRG) grant and a University Transportation Research Center (UTRC) faculty-initiated research grant.

References

Acumen Building Enterprise, Inc. (2006). TCRP Report 115: Smartcard Interoperability Issues for the Transit Industry. Transit Cooperative Research Program of the Transportation Research Board, Washington DC.

Bagchi, M., and White, P. R. (2005) The Potential of Public Transport Smart Card Data. *Transport Policy*, 12(5), pp. 464–474.

Brakewood, C., Ghahramani, N., Peters, J., Kwak E., and Sion, J. (2017). Real-Time Riders: A First Look at User Interaction Data from the Backend of a Transit and Shared Mobility Smartphone App. *Transportation Research Record: Journal of the Transportation Research Board*, Volume 2658, pp. 56-63.

Ghahramani, N. and Brakewood, C. (2016). Trends in Mobile Transit Information Utilization: An Exploratory Analysis of Transit App in New York City. *Journal of Public Transportation*, Volume 19, Issue 3, pp. 139-160.

Ghahramani, N., Brakewood, C. and Peters, J. (2017). An Exploratory Analysis of Intercity Travel Patterns Using Backend Data from a Transit Smartphone Application. Proceedings of the 96th Annual Meeting of the Transportation Research Board, Washington, DC.

McGeehan, P. (2016). Record Number of Tourists Visited New York City in 2015, and More Are Expected This Year. New York Times, March 8, 2016. http://www.nytimes.com/2016/03/09/nyregion/record-number-of-tourists-visited-new-york-city-in-2015-and-more-are-expected-this-year.html?_r=0. Accessed May 25, 2018.

McKenzie, B. and Rapino, M. (2011) *Commuting the United States: 2009*. American Community Survey Reports, United States Census Bureau, http://www.census.gov/prod/2011pubs/acs-15.pdf.

New York Metropolitan Transportation Council (NYMTC) Website, https://www.nymtc.org/. Accessed August 1, 2016.

Pelletier, M.P., Trépanier, M., & Morency, C. (2011) Smart Card Data Use in Public Transit: A Literature Review. *Transportation Research Part C: Emerging Technologies*, 19(4), pp. 557–568.

Pratt, J., Lee-Gosselin M., and Burbridge, S. (2018) Chapter 3: Options for Travel Surveys. The Online Travel Survey Manual: A Dynamic Document for Transportation Professionals. http://www.travelsurveymanual.org/Chapter-3-1.html, Accessed May 24, 2018.

Remy, C., Brakewood, C., Ghahramani, N., Kwak, E., and Peters, J. (2018). Transit Information Utilization during an Extreme Weather Event: An Analysis of Smartphone App Data. Proceedings of the 97th Annual Meeting of the Transportation Research Board, Washington, DC.

Schaller, B. (2005) TCRP Synthesis 63: On-Board and Intercept Transit Survey Techniques: A Synthesis of Transit Practice. Transit Cooperative Research Program of the Transportation Research Board, Washington DC.

Solomon, B. (2016). Winter Storm Jonas Knocks Out Uber In New York, New Jersey. *Forbes.* January 23, 2016. Accessed May 25, 2018. https://www.forbes.com/sites/briansolomon/2016/01/23/winter-storm-jonas-knocks-out-uber-in-new-york-new-jersey/#12b164934cb2

Transit (2018), http://transitapp.com/, Accessed May 24, 2018.