NEW YORK CITY DOT

CITYWIDE MOBILITY SURVEY

SURVEY USER GUIDE

RSG
the science of insight

PREPARED FOR THE NEW YORK CITY
DEPARTMENT OF TRANSPORTATION BY RSG

CONTENTS

1.0 SURVEY OVERVIEW

The New York City Department of Transportation (NYC DOT) conducts an annual travel survey called the Citywide Mobility Survey (CMS). The CMS assesses the travel behavior, preferences, and attitudes of New York City residents. The objectives of the survey are to:

- Identify the factors and experiences that drive transportation choices for New York City residents.

- Understand current views on the state of transportation within the City.

- Measure attitudes toward current transportation issues and topics in New York City.

- Determine the citywide mode share by collecting information on the trips made by a representative sample of residents over a set period of time.

- Track year over year changes in travel behavior.

The CMS began in 2017 and collects two primary types of data: 1) demographic information and 2) travel diary information where all trips made by a respondent during a set period of time are documented.

2.0 SURVEY DESIGN & METHODOLOGY

Under contract to NYC DOT, RSG managed the 2019 CMS effort collecting travel behavior data from May 2019 through June 2019. The 2019 CMS differs from previous iterations of the survey in two primary ways: 1) the sampling methodology changed from random digit dialing (RDD) to address-based sampling (ABS)[1] allowing all complete responses to be weighted and 2) survey participation through a smartphone-app was made available – in addition to the online and call center options previously offered. Online and call center participants completed a one-day retrospective travel diary and smartphone participants completed a real-time seven-day travel diary – increasing the number of complete travel days and the number of trips from complete travel days within the survey sample size.

The goal of the survey was to collect a statistically valid sample of 3,000 residents of New York City across ten geographic survey zones outlined by NYC DOT with approximately 300 participants per zone. The zones were created for the 2017 CMS as a way to examine travel patterns and behaviors at a sub-borough level that was still statistically valid. Each survey zone is composed of neighborhood tabulation areas (NTAs). For more information on which NTAs fall within each zone, refer to NYC Open Data. Overall, a total of 3,346 residents participated in the survey and are included in the final weighted dataset – 75% participated by smartphone, 20% participated online, and 5% participated through the survey call center. Despite the difference in length of the travel diary across participation methods, the same questionnaire was used across all three participation methods allowing for all data to be combined into a single weighted dataset.

---

[1] *Random digit dialing (RDD) is a survey sampling method where people are selected for survey participation by generating telephone numbers at random. Address-based sampling (ABS) is a survey sampling method where a random sample of addresses are selected from the U.S. Postal Service database of residential addresses. In recent years, an industry shift from RDD to ABS has occurred in survey research due to increased bias in RDD samples.*

FIGURE 1: CITYWIDE MOBILITY SURVEY REGION AND COMPLETE RECORDS BY ZONE
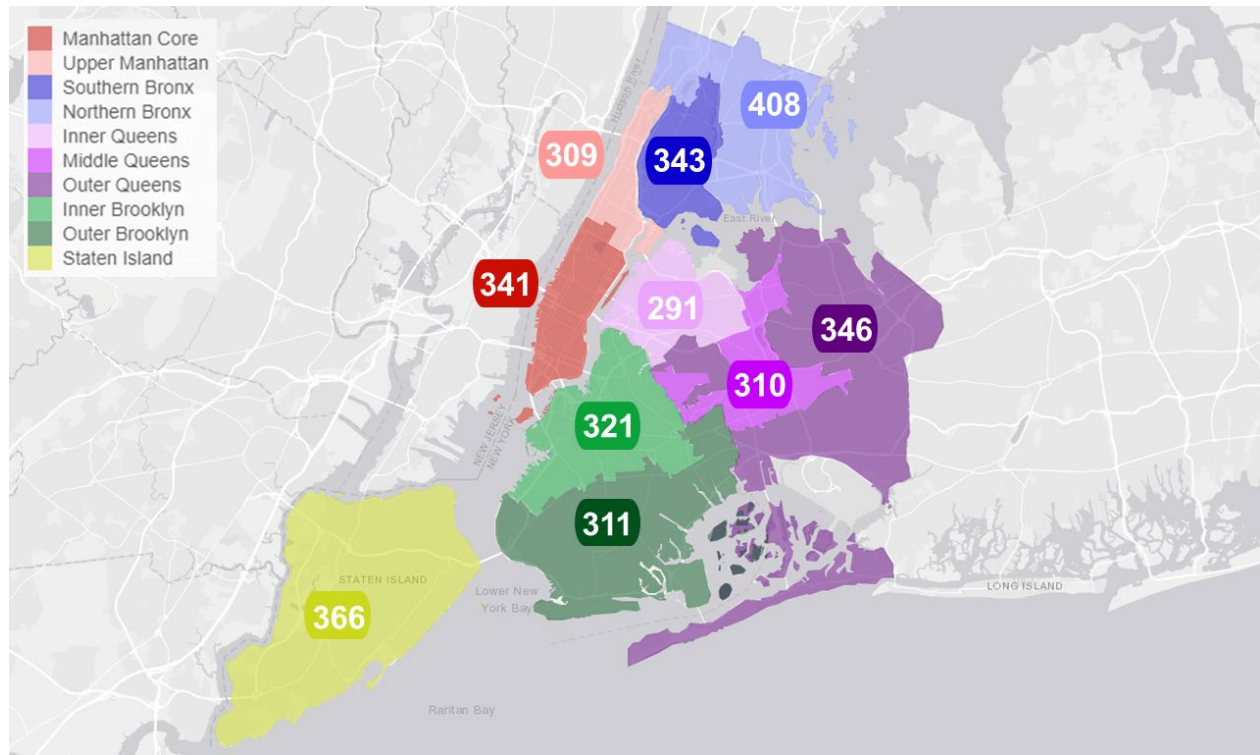
3.0 WEIGHTING METHODOLOGY SUMMARY

Following data collection, the survey dataset was weighted to match the population across key sociodemographic dimensions. Weighting corrects for bias that may be present in the sample of people who completed the survey and ensures the dataset is representative of the sample region. The data collected in the 2019 CMS was weighted to the 2017 American Community Survey Public Use Microdata Sample (ACS PUMS) data. The weighting methodology applied adjusts for survey non-response, survey participation mode, as well as geographic bias due to oversampling and other factors. In addition, trip rates were adjusted between the three participation methods offered for the survey: 1) online, 2) call center, and 3) smartphone-app.

The weighting process applied included four primary steps:

1. **Initial Expansion**: Calculating an 'initial weight' based on the probability of selection. This essentially 'reverses' the sample plan, providing higher initial weights to areas where less sampling occurred.
2. **Reweighting to account for non-response bias**: Performing an iterative proportional fit (IPF) routine to several key household and person dimensions to ensure the weighted data accurately represent the entire region (and reduce sampling biases).
   a. To do this step, a few missing data elements (income, gender, ethnicity and race) need to be imputed for those who did not provide that information.
3. **Creating day-level weights to account for multi-day survey data:** Adjusting the day-level and trip-level data to account for the fact that smartphone respondents provided multi-day travel

diaries, while online and call center respondents provided a single-day travel diary (this is the "multi-day adjustment"). These relatively simply adjustments ensure that travel analyses accurately reflect the entire region and do not over-represent smartphone respondents with multiple travel days.

4. **Adjusting for non-response bias in day pattern and trip rates**: Adjusting the trip-level weights by data collection method (smartphone, online, call center) to account for reporting biases that RSG has detected in this survey and prior travel surveys. These adjustments help make the day and trip-level data more consistent and increase the accuracy of trip rates across survey participation methods.

The weighting process applied to the CMS created the following three final weights that are provided with the dataset:

- **weight**: The resulting weights from expanding to the PUMS data. Should be used for household-level, person-level, and vehicle-level analyses.

- **day_weight**: The same as *weight* but divided by the number of complete days of data for each household. Should be used for household-day and person-day analyses – represents an average day.

- **trip_weight**: The resulting adjustment factors from the trip correction process. Should be used for trip-related analysis.

Overall, the goal of the weighting process is to make the survey sample representative of the entire city across a number of key dimensions related to travel behavior. Users should use the weighted survey data in any analysis wishing to draw conclusions about New York City as a whole or any of the survey zones.

4.0 DATA USER GUIDE

Transportation planners rely on detailed travel data to inform their planning. The 2019 CMS dataset includes highly detailed information about how, where, when, and why New Yorkers make trips. The CMS allows NYC DOT to obtain descriptive statistics about citywide travel and to analyze trends over time.

Data users can derive many key figures from the CMS dataset, including person-trip rates, travel mode shares, vehicle occupancy factors, geographic travel patterns, and more – with rich demographic detail. Beyond traditional travel survey uses, the CMS program is also somewhat unique in its recurrent nature, collecting data every year. This allows for trend analysis over time, which is often not feasible when collection occurs less frequently.

4.1 DATASET OVERVIEW

This dataset includes data collected for the Citywide Mobility Survey from May through June 2019. The study included two sections:

1. **Section one**, also called the "**recruit survey**," collected information about household composition, person demographics, and typical travel behavior.

2. **Section two**, also called the "**travel diary**," required participants to record their travel during an assigned travel period. RSG's travel survey smartphone app, rMove™, collected data for 97% of trips in the dataset, and RSG's travel survey instrument, rSurvey™, collected the remaining trip data either online or through the call center.

This dataset includes full data for the 3,346 records that met the survey completion criteria. A record is considered complete and was included in the final dataset when it met the following two conditions:

- **Online/Call Center**: a person completed the survey by answering all required questions including the 24-hour travel diary.

- **Smartphone**: a person completed the recruit survey and at least one 24-hour travel diary (for smartphone participants that completed at least one, but not all seven days – data from partially complete days has also been delivered and has been flagged as incomplete).

The 2019 CMS dataset includes six data tables, often referred to as their own "level" of data. These tables include all user-input survey variables, passively collected GPS and location data, survey metadata, and derived variables to support data analysis. The tables included in the dataset include:

- **Household**: One record for each valid, complete participant's household.

  - *Note: While the CMS is a person-level survey, a household table is still included to provide details about the overall household and residence.*

- **Vehicle**: One record for each vehicle (if any) in the participant's household.

- **Person**: One record for each member of the participant's household.

  - *Note: Other members of the primary survey respondent's household are included in this table, but only basic demographic information has been included. These other household members did not complete a trip diary, so they do not have corresponding records in the day and trip tables.*

- **Day**: One record for each day of the participant's assigned travel period (one to seven travel days per person).

- **Trip**: One record for each trip (if any) reported by the participant.

- **Location**: Two or more records for each trip collected via smartphone (if any). *Note: this table is not included in the public dataset.*

4.2 DATA CODING AND LABELING

A comprehensive list of data variables and values is provided in the dataset codebook. This section expands on time and location standards as well as missing value codes that have been used across multiple variables.

Time and location standards:

All timestamps reflect the local time zone for the study region (Eastern time), regardless of where the trip took place geographically (e.g., if a trip took place in another time zone, the timestamps for that trip are still in Eastern time).

Missing values and gaps in the data

A cell may not have data for the following reasons:

1. A value or response was not required under the circumstances.

    -   Example: Respondents who were not employed were not asked what industry they work in.

    -   Coded as: 995

    -   Continuous variables are not coded with missing value codes and are instead left empty when missing to avoid interfering with any statistical calculations. Text values are left blank when there is no response.

2. A respondent indicated that they didn't know the answer and skipped that question.

    -   Example: Some respondents didn't know how much they personally pay to park at work.

    -   Coded as: 998 (Don't know)

3. A respondent indicated that they preferred not to answer a question and skipped that question.

    -   Example: Some respondents preferred not to provide household income.

    -   Coded as: 999 (Prefer not to answer)

4. A respondent did not answer part of a survey that was required.

    -   Example: A smartphone respondent did not complete a trip survey and trip mode is missing.

    -   Coded as: -9998 (Non-response)

    -   *Note: This code only applies to smartphone participant data. Smartphone participants completed the recruit survey and at least one 24-hour travel diary. A survey day is considered complete when all required questions were answered at the trip and day level. For smartphone participants that completed at least one, but not all seven days – data from partially complete days has also been delivered and has been flagged as incomplete. For this reason, some variables are coded as -9998 for smartphone participants.*

5. A technological error occurred.

    -   Example: A logic or data collection error occurred resulting in missing responses for otherwise required data/questions.

    -   Coded as: -9999 (Technical error).

4.3 RECOMMENDATIONS FOR DATA USERS

Although travel survey data provides many opportunities for interesting analysis, data users should consider the context and best applications of the data. Data users should **keep the "universe" of data collection in mind** to ensure the analysis is logical for the data source.

Second, data users should **use the weighted survey data** in any analysis wishing to draw conclusions about the city as a whole (as opposed to the survey takers). Applying weights ensures that the final analysis is regionally representative.

Finally, data users should **ensure a sufficient sample size (and acknowledge margins of error)** in any analysis. The smaller the sample size, the larger the margin of error. For example, travel survey data users can generally draw reasonable conclusions about trip rates by mode on an average day but should consider the sample size for modes with small shares of overall citywide travel.

4.4 PRACTICAL TIPS FOR DATA ANALYSIS

Many ways exist to view, join, summarize, and map the CMS data. To achieve the full data benefits, data users may need multiple tools. Data users can apply three key data analysis mechanics to make the most of their analyses.

Joining Tables on Unique IDs

All data tables can be joined into a single database as needed. Some unique IDs are a combination of two variables. In these cases, joining on only one of the variables will create duplicate records.

TABLE 1: UNIQUE IDS FOR JOINING DATA TABLES

	Household	Person	Vehicle	Day	Trip
Household	--	hh_id	hh_id	hh_id	hh_id
Person	hh_id	--	--	person_id	person_id
Vehicle	hh_id	--	--	--	--
Day	hh_id	person_id	--	--	person_id + day_num
Trip	hh_id	person_id	--	person_id + day_num	--
Location	hh_id	person_id	--	--	trip_id

Applying and Interpreting Weights

Analyses designed to draw conclusions about travel behavior in the city (as opposed to just the survey respondents) should use weighted data. When applied, the weights make the dataset representative of citywide travel for the time period studied (May - June 2019). This means it does not include commercial vehicle travel, travel for persons residing in group quarters outside of the address-based sampling frame (e.g., college dorms, institutional housing), travel from nonresidents (i.e., visitors to the region), or seasonal/holiday travel outside of the survey fielding period.

Using weighted data generally involves summing the weights for the groups of interest. The sum of weights in each table represents the following groups:

- **Household**: Represents the total number of households within the survey region.

- **Person**: Represents the total number of persons within the survey region.

- **Vehicle**: Represents the total number of personal vehicles of households in the survey region.

- **Day**: Represents one weekday for all persons residing in the survey region. For the CMS, this is equal to the number of persons in the region.

- **Trip**: Represents the total number of trips all persons residing in the survey region make on a typical day.
    - This differs from the number of trips made *in the survey region* on a typical day, given that some residents make trips outside the region.

To calculate weighted crosstabs or descriptive statistics, sum the weights for that table. Grouping variables or filters will provide weighted totals for specific subgroups, like two-person households, or by age group. Keep in mind the following when creating weighted statistics and summaries from travel survey data:

1. **Filter to the data relevant to your analysis (e.g., complete travel days).**

    - For example, focusing on travel days with complete survey information (i.e., no unanswered survey questions) is best. Also note that not all people are asked every question, so understanding the 'missing value' codes can be important.

2. **Remember the survey design when using and interpreting weighted values.**

    - For example, the survey design for the 2019 CMS included both one-day online and call center participation and seven-day smartphone app participation.

Generating Trip Rates

Trip rates – or the number of trips per day among groups – are useful for comparing several travel behaviors (e.g., travel by mode, travel by age group). Trip rates can be weighted or unweighted, but this section focuses on the former.

To calculate a weighted trip rate, data users must divide the number of weighted trips by the number of weighted travel days. For example, if there are 300,000 weighted person-trips across 75,000 person-days, then the average person-trip rate is 4.0 per day. If there are 225,000 person-trips by car across 75,000 person-days, then the person-trip rate for car trips is 3.0. This is different than calculating vehicle trip rates, which would require calculating the weighted vehicle trips taking place (in this example, if the average vehicle occupancy is two people, then the vehicle trip rate would be 1.5).

Data users should always calculate the number of weighted travel days using the day table rather than the trip table given that persons with zero-trip travel days do not have any records in the trip tables for those days.

5.0 FURTHER INFORMATION

For further information about the Citywide Mobility Survey program, please visit the CMS information page on the NYC DOT website.

CMS information page: https://www1.nyc.gov/html/dot/html/about/citywide-mobility-survey.shtml

RSG
the science of insight

55 Railroad Row
White River Junction, VT 05001
802.295.4999
www.rsginc.com

White River Junction & Burlington, VT

Arlington, VA

Chicago, IL

Evansville, IN

Portland, OR

Salt Lake City, UT

San Diego, CA