NEW YORK CITY DOT CITYWIDE MOBILITY SURVEY

Survey User Guide

CONTENTS

1.0 Overview

The New York City Department of Transportation (NYC DOT) conducts a household travel survey called the Citywide Mobility Survey (CMS) to assess New York City residents' travel behavior, preferences, and attitudes. Launched in 2017, additional surveys were conducted in 2018, 2019, 2020, and 2022. The continuity of the survey allows us to track changes over time and analyze how factors such as COVID-19 have affected travel by New York City residents. The objectives of the survey are to better understand:

- Changes in travel behavior over time
- New Yorkers' basic transportation conditions
- New Yorkers' experience of and opinions of transportation
- The equity of access to transportation across the city

The survey also provides valuable data for building models of transportation behavior.

The CMS collects two primary types of data: 1) demographic information and 2) travel diaries that document all trips made by a respondent during a set period of time.

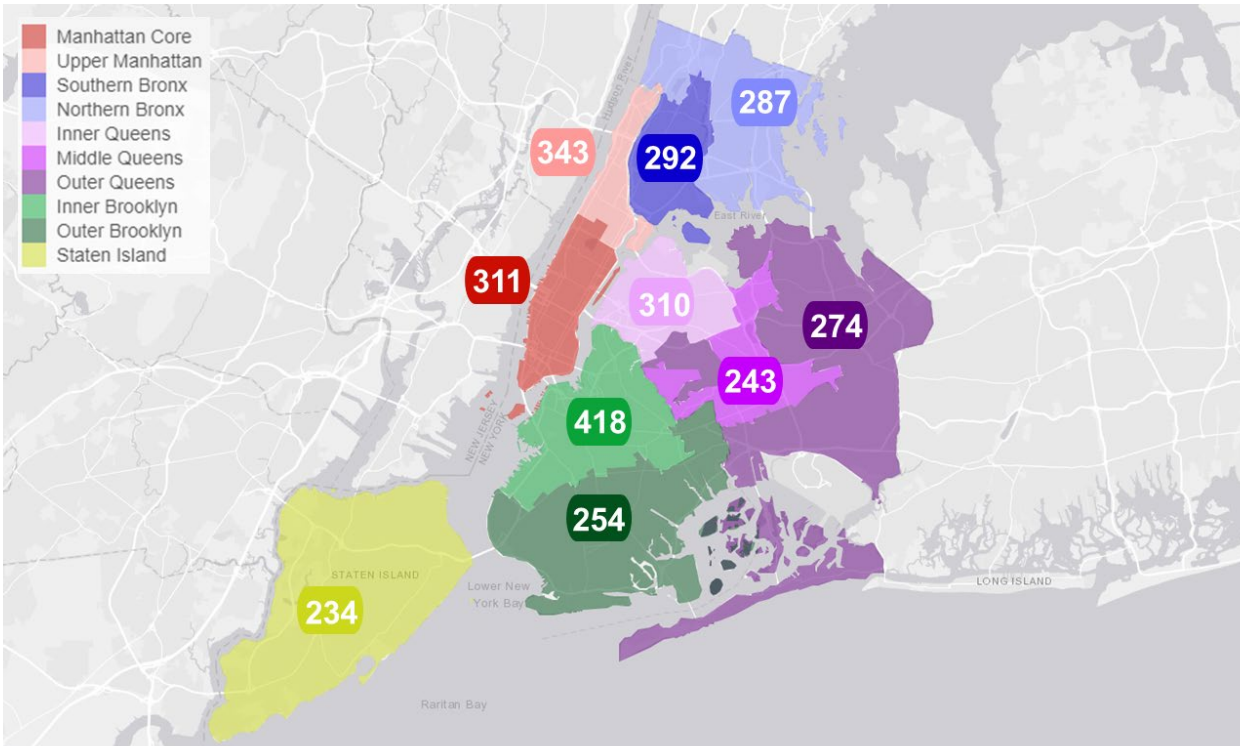
2.0 Survey Design & Methodology

Like the 2019 CMS, the 2022 CMS used address-based sampling (ABS) to recruit most participants; to compensate for an unusually low response rate, NYC DOT augmented the sample by re-inviting some participants from the 2019 CMS. To allow for meaningful analysis of travel patterns at a sub-borough level, the sample was constructed with the aim of enrolling 300 participants in each of ten geographic Citywide Mobility Survey zones into which NYC DOT has segmented the City. (Survey zones are aggregations of neighborhood tabulation areas (NTAs), as defined after the 2010 Census. NYC Open Data provides a shapefile of the CMS zones.)

Nearly three-fourths of respondents – 73% – chose to take the survey via smartphone app, while 23% used the online version and 4% participated via telephone (call center) interview. Online and call center participants completed a one-day travel diary and smartphone participants completed a real-time seven-day travel diary. Despite the difference in the 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.

The 2022 CMS effort collected complete travel behavior data from 2,966 resident adults from September 28 through November 17, 2022. The survey was conducted by RSG, which had previously conducted the 2019 and 2020 surveys. Figure 1 shows the actual number of respondents in each CMS Zone.

Figure 1: Citywide Mobility Survey region and complete records by zone



3.0 Weighting Methodology

This section describes the methodology used to expand[1] the survey data collected in the 2022 CMS to the 2021 American Community Survey Public Use Microdata Sample (ACS PUMS) 1-year data[2]. The weighting methodology is intended to allow the sample data to represent the entire survey population across several key dimensions related to travel behavior.

The weighting methodology also adjusted for biases inherent to data collection methods and user response. This includes survey non-response, participation mode (online, call center and smartphone app), and geographic bias due to oversampling and other factors.

The weighting process included four main steps summarized below and elaborated on in the following sections:

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 entropy maximization-based list balancing routine to several key household and person dimensions (described below) to reduce sampling biases and ensure the weighted data accurately represent the

[1] For the purposes of this memo, the terms expansion, expansion factors, and weights are used interchangeably and are synonymous. They all represent the concept of an expansion weight.
[2] https://www.census.gov/programs-surveys/acs/microdata/access/2021.html

entire survey region. This routine was performed using an open-source application, PopulationSim[3]. This method included imputation of missing data elements such as income, gender and race or ethnicity.

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 simple adjustments ensure that travel analyses accurately reflect the entire survey region and do not over-represent smartphone respondents with multiple travel days.
4. **Adjusting for non-response bias in trip rates**: Adjusting the trip-level weights by data collection method (smartphone, online, call center) to account for under-reporting of trips by online and call center users . These adjustments help make the trip-level data more consistent and increase the accuracy of trip rates across survey participation methods.

The following sections describe these four steps in more detail.

Overall, the goal of the weighting process is to make the survey sample representative of the entire survey region 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 the region as a whole.

3.1 Initial expansion

The purpose of the initial expansion is to expand each person that completed the survey to the population that was eligible to participate in the survey. As described in the sampling plan, the NYC CMS sample included data collected via supplemental non-probability[4] methods by reinviting 2019 CMS participants in addition to the traditional addressed-based sampling (ABS) approach. The inclusion of non-probability sample primarily impacts the calculation of the initial expansion factors aiming at ‘reversing’ the sampling plan. The primary point is how to allocate the population to the ABS and the supplemental sample and how that is implemented. The approach is described below.

3.1.1 Initial Weights for the Address-based (ABS) Sampling

The initial expansion weights for the ABS were calculated using:

H(***s***) = the actual number of households in each sampling stratum ***s***, based on ACS estimates.

R_{abs}(***s***) = the number of household responses obtained from each sampling strata ***s,*** via address-based sampling, including only those households/persons with at least one complete weekday from Monday to Thursday. The 2022 NYC CMS used the same strata as the 2019 CMS, i.e., 20 subzones of 10 NYC CMS zones x 2 (hard-to-survey census block groups vs. non-hard-to-survey census block groups). See the zones in figure 2 and Table 1.

[3] https://activitysim.github.io/populationsim/
[4] We use the terms “non-probability” and “convenience” interchangeably for the purposes of this memo.

The selection of “weekdays” essentially assumes that trip rates and behavior on those days are similar enough to consider them interchangeable, with an average weekday being the average of travel across those days.

C = percentage of population assigned to the convenience-based sample.

The initial expansion weight (IW) for the address-based sample is then equal to:

$IW_{abs}(\boldsymbol{s}) = H / R_{abs}(\boldsymbol{s}) * (1 – C)$

The initial expansion weights were calculated for each sampling segment and used as the starting weights for further reweighting to correct for non-response biases in the data, which is described in the following sections. The next sub-step is introduced by non-probabilistic sampling through reinviting 2019 CMS participants.

3.1.2 Initial Weights for the Convenience-based Sampling

Based on the above formulas, the initial expansion weights for the ABS for sample strata ***s*** will add up to a proportion of the total population (1 – C). The initial expansion weight for the convenience-based sample would then be defined as:

$IW_{cbs}(cbs) = H / R_{cbs}(cbs) * C$

We would set C to be small for strata where the ABS methods performed well in achieving a representative sample. For the 2022 NYC CMS, the 2019 CMS re-invitation sample is considered as the convenience-base sample. Given the statistical advantages of the address-based sample, RSG would recommend that C be set no larger than 20% and that a target of an average of no more than 10%. For the 2022 NYC CMS, RSG set C to 10%. See initial weights by each subzone including the convenience-based samples in Table 2.

Figure 2: NYC CMS Zones and Hard-To-Reach Census tracts

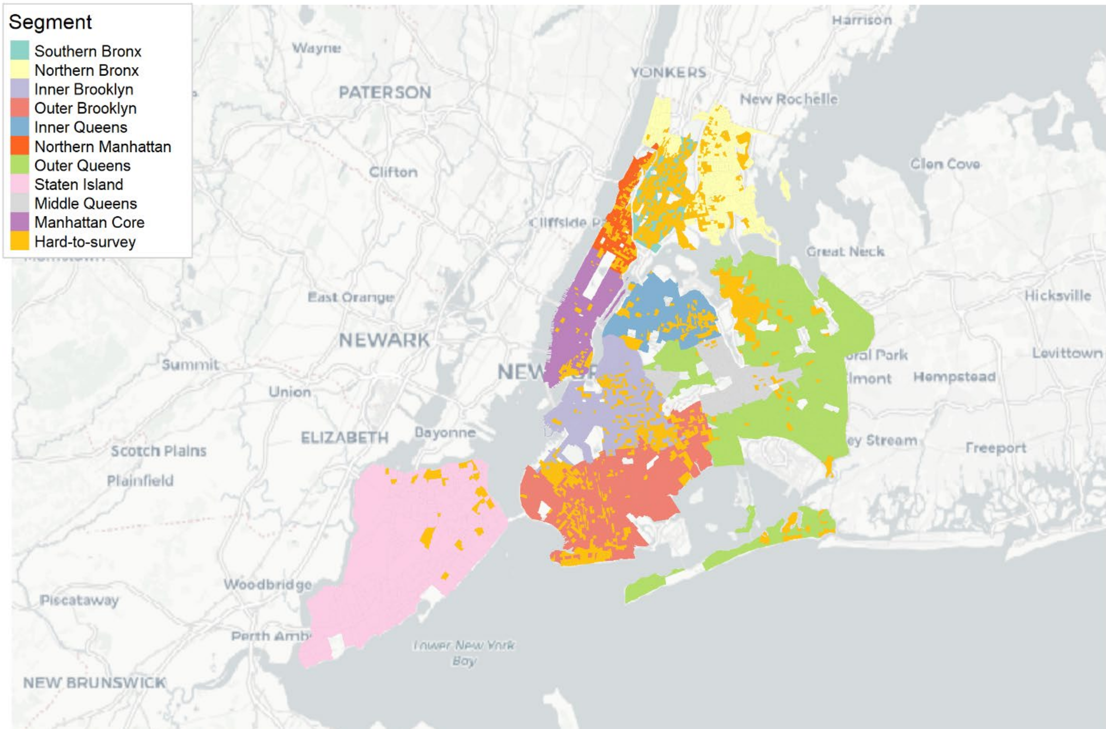


Table 1: 2022 NYC CMS Zones

CMS ZONE	LOCATION
1	Southern Bronx
2	Northern Bronx
3	Inner Brooklyn
4	Outer Brooklyn
5	Inner Queens
6	Northern Manhattan
7	Outer Queens
8	Staten Island
9	Middle Queens
10	Manhattan Core

Table 2: Initial Expansion factor by zone, Hard-to-Reach subzone & 2019 Reinvites

Sample Segment	Respondents	ACS Persons at least 18 years old	Initial Expansion Factor
Southern Bronx	56	113,051	2,018.76
Southern Bronx-Hard-to-survey	201	353,213	1,757.28
Northern Bronx	176	349,771	1,987.34
Northern Bronx-Hard-to-survey	69	124,578	1,805.47
Inner Brooklyn	294	612,191	2,082.28
Inner Brooklyn-Hard-to-survey	79	237,267	3,003.37
Outer Brooklyn	160	725,228	4,532.68
Outer Brooklyn-Hard-to-survey	51	243,525	4,774.99
Inner Queens	219	334,171	1,525.90
Inner Queens-Hard-to-survey	52	123,334	2,371.81
Northern Manhattan	221	271,385	1,227.99
Northern Manhattan-Hard-to-survey	87	138,494	1,591.89
Outer Queens	198	746,876	3,772.10
Outer Queens-Hard-to-survey	32	101,105	3,159.52
Staten Island	187	313,787	1,678.00
Staten Island-Hard-to-survey	11	26,992	2,453.85
Middle Queens	169	311,528	1,843.36
Middle Queens-Hard-to-survey	27	62,101	2,300.05
Manhattan Core	246	679,605	2,762.62
Manhattan Core-Hard-to-survey	24	73,101	3,045.89
Southern Bronx 2019 reinvited	8	12,561	1,570.15
Southern Bronx-Hard-to-survey 2019 reinvited	26	39,246	1,509.46
Northern Bronx 2019 reinvited	34	38,863	1,143.04
Northern Bronx-Hard-to-survey 2019 reinvited	9	13,842	1,538.00
Inner Brooklyn 2019 reinvited	33	68,021	2,061.25
Inner Brooklyn-Hard-to-survey 2019 reinvited	12	26,363	2,196.91
Outer Brooklyn 2019 reinvited	30	80,581	2,686.03
Outer Brooklyn-Hard-to-survey 2019 reinvited	13	27,058	2,081.41
Inner Queens 2019 reinvited	26	37,130	1,428.08
Inner Queens-Hard-to-survey 2019 reinvited	13	13,704	1,054.14
Northern Manhattan 2019 reinvited	30	30,154	1,005.13
Northern Manhattan-Hard-to-survey 2019 reinvited	5	15,388	3,077.65
Outer Queens 2019 reinvited	40	82,986	2,074.66
Outer Queens-Hard-to-survey 2019 reinvited	4	11,234	2,808.46
Staten Island 2019 reinvited	35	34,865	996.15
Staten Island-Hard-to-survey 2019 reinvited	1	2,999	2,999.15
Middle Queens 2019 reinvited	39	34,614	887.55
Middle Queens-Hard-to-survey 2019 reinvited	8	6,900	862.52
Manhattan Core 2019 reinvited	38	75,512	1,987.15
Manhattan Core-Hard-to-survey 2019 reinvited	3	8,122	2,707.46
Total	**2,966**	**6,714,576**	

3.2 Reweighting to account for non-response bias

The 2021 American Community Survey Public Use Microdata Sample (ACS PUMS) 1-year data served as the target data for weighting the 2022 CMS dataset. We used an entropy-maximization (EM) algorithm, as implemented in PopulationSim, to adjust the weights to match these targets. This routine was seeded with the initial expansion weights described in the previous step. The benefit of this approach relative to IPF is that the resulting weights tend to have reduced variance resulting in reduced margins of error when working with weighted statistical analysis. For more details on the mathematical steps employed, see Multi-level Population Synthesis Using Entropy Maximization-Based Simultaneous List Balancing by Paul et al[5].

3.2.1 Weighting geography

Ten NYC CMS zones were used for the target geographies. Then, a set of weighting controls were generated for the target geographies in the study areas to adjust for non-response bias. The PUMS weighting controls were allocated to the CMS zones using a crosswalk from PUMAs provided by NYC DOT for the 2019 CMS.

3.2.2 Household and Person Weighting Targets

Different household compositions and personal attributes impact response (for example a person with multiple children who travel a lot each day may be less likely to provide a response due to the length of the corresponding survey). Therefore, a variety of person-level and household-level target categories were developed to account for these potential biases. The person-level targets were designed to identify the person types that are typically used in activity-based modeling software. The weighting targets were derived from PUMS data using the person-level weights. This step also included model-based imputation of missing values (described below) for respondents who might choose to respond with ‘prefer not to answer’ to target variables.

The household and person-level variables used in the non-response adjustment step are included below in Table 3 and Table 4.

Table 3: Household-Level Target Variables

VARIABLE	CATEGORIES
Income *(Imputed if non-response)*	Under $25,000 $25,000 - $49,999 $50,000 - $74,999 $75,000 - $99,999 $100,000 - $149,999 $150,000 - $199,999 $200,000 or more
Household Size	1-person 2-person 3-person 4-person 5-person or more

5 https://trid.trb.org/view/1496005

Number of children	0-child 1-child 2-children 3-children or more
Number of adults	1-adult 2-adults 3-adults 4-adults or more

Table 4: Person-Level Target Variables

VARIABLE	CATEGORIES
Gender ***(Imputed if non-response)***	Male Female
Age	18 – 24 years 25 – 44 years 45 – 64 years 65 years or older
University Student Status	University student Non-university student
Educational Attainment	Some college education No college education
Race ***(Imputed if non-response)***	White Non-White
Ethnicity ***(Imputed if non-response)***	Non-Hispanic Mexican, Puerto Rican Other Hispanic Origin (Cuban and Dominican were initially assigned to a separate group but eventually combined to this group due to low samples)
Total Persons	Not applicable

Unlike the previous 2019 CMS, the 2022 CMS did not control for three transportation-related variables for non-response adjustment: vehicle ownership, worker status and primary commute mode. NYCDOT and RSG believe that not controlling for these variables with the 2021 ACS PUMS data would better reflect reality given the rapid change in these variables each year since COVID, i.e., the 2022 CMS would better reflect the new reality.

3.2.3 Imputation of missing values

Income

RSG imputed income using an approach based on an ordered logit model, where missing income was predicted based on a set of independent variables including:

- Income distribution of the respondent's block group
- Number of non-working adults in the household
- Educational attainment of the household
- Age of the primary survey respondent

This model has been tested across many travel survey projects and adequately matches the income values that were reported, indicating it is reliable to predict the missing income values. Model specification and coefficients are shown in Table 5

Table 5: Income imputation model summary

PARAMETER	DESCRIPTION	ESTIMATE	STD ERROR	T-STAT
finc_0k_25k	Fraction of people in block group with incomes under 25k	-1.49	0.50	-2.95
finc_25k_50k	Fraction of people in block group with incomes 25k-50k	-1.44	0.61	-2.35
finc_50k_75k	Fraction of people in block group with incomes 50k-75k	-1.01	0.65	-1.57
finc_100k_150k	Fraction of people in block group with incomes 100k-150k	0.81	0.66	1.23
finc_150k_200k	Fraction of people in block group with incomes 150k-200k	1.67	0.74	2.25
finc_200k_plus	Fraction of people in block group with incomes more than 200k	2.53	0.58	4.38
nonworking_adult_n	Number of non-working adults in household	-0.70	0.26	-2.69
full_time_graduate_degree_n	Number of full–time workers with graduate degrees in household	1.41	0.26	5.36
part_time_graduate_degree_n	Number of part–time workers with graduate degrees in household	-0.26	0.38	-0.68
full_time_bachelor_degree_n	Number of full–time workers with bachelor's degrees in household	1.17	0.26	4.47
part_time_bachelor_degree_n	Number of part–time workers with bachelor's degrees in household	-0.93	0.37	-2.56
full_time_low_education_n	Number of full-time workers with no advanced degrees in household	0.16	0.27	0.6
part_time_low_education_n	Number of part-time workers with no advanced degrees in household	-0.62	0.31	-1.98
head_under_35_n	Head of household under 35 years	-0.24	0.08	-2.97
head_over_65_n	Head of household over 65 years	-0.18	0.12	-1.48
own_home	Owns home (doesn’t rent)	1.18	0.09	13.23

PARAMETER	DESCRIPTION	ESTIMATE	STD ERROR	T-STAT
single_family_home	Lives in single family housing	-0.24	0.10	-2.44

McFadden's rho-squared: 0.13

Gender

Missing gender was probabilistically assigned using a Monte Carlo procedure based on the sample data's gender distribution within the respondent's age category.

Race and Ethnicity

Missing race and ethnicity were probabilistically assigned using a Monte Carlo procedure based on the ACS data's race/ethnicity distribution within the respondent's home block group.

3.3 Expansion of Household and Person Data

Table 6 provides the distribution of the final weights that were calculated for each weighting geography in the sample and Table 7 summarizes the ratio of the final weight against the initial expansion factor (the weight derived based on the probability of being sampled). In the weighting process, the ratio of the final weight to the initial weight was constrained to be in the range of 0.01 to 5 for each household/person, with a maximum absolute weight of 25,000. In general, allowing the weights to be outside this range would enable the process to match the ACS PUMS targets more exactly, but at the cost of having extremely high or extremely low weights and the introduction of more variance. Considering that the PUMS targets are themselves estimates based on Census survey data, it is not good practice to try to match the targets too precisely by allowing the survey weights to vary widely. In contrast however, relaxing the weight ratio limits can at times result in less extreme weights and variance due to unique combinatorial population covariances that would otherwise be difficult to match. The target ratio range of 0.01 to 5 was arrived at after testing alternative limits and judging the best trade-off between accuracy and variability.

Table 6: Summary statistics of the final weights[6]

GEOGRAPHY	SAMPLE SIZE	MIN	MEAN	MEDIAN	MAX
Southern Bronx	291	152.01	1,780.15	587.94	10,086.66
Northern Bronx	288	114.40	1,830.02	799.64	9,935.85
Inner Brooklyn	418	206.39	2,257.89	1,067.03	15,016.08
Outer Brooklyn	254	279.60	4,237.51	2,454.54	22,661.99
Inner Queens	310	105.96	1,638.93	214.94	11,857.27
Northern Manhattan	343	100.83	1,327.77	391.18	7,957.95
Outer Queens	274	208.42	3,438.72	1,468.96	18,858.98
Staten Island	234	101.39	1,616.12	700.14	8,384.87
Middle Queens	243	87.09	1,705.55	186.89	11,500.00
Manhattan Core	311	334.32	2,689.05	1,994.15	15,198.76

[6] The maximum weights in the 2019 were around 15,000.

Table 7: Summary statistics for the ratio of final to initial Weights

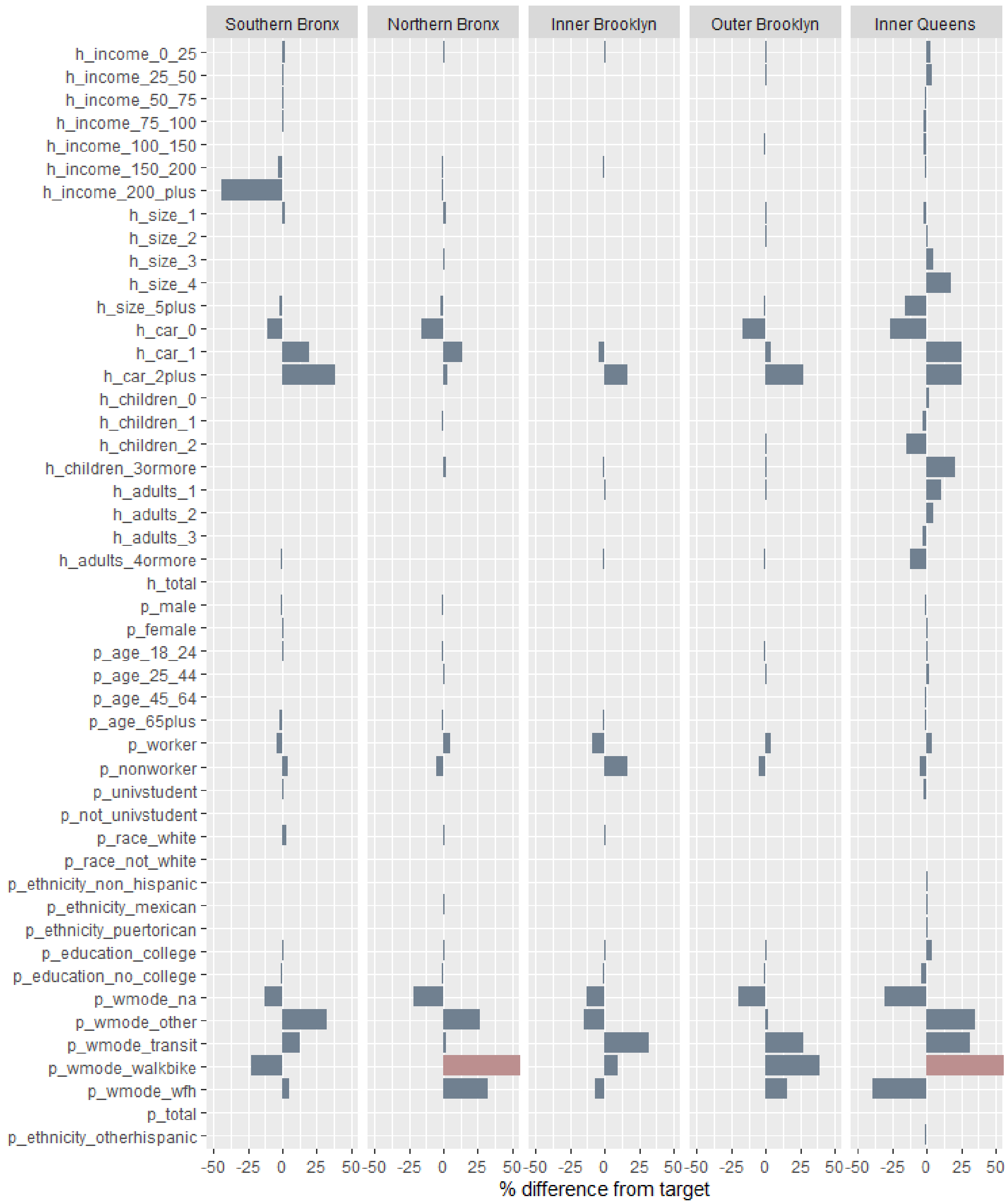
GEOGRAPHY	SAMPLE SIZE	MIN	MEAN	MEDIAN	MAX
Southern Bronx	291	0.10	1.01	0.33	5.00
Northern Bronx	288	0.10	1.00	0.44	5.00
Inner Brooklyn	418	0.10	0.96	0.49	5.00
Outer Brooklyn	254	0.10	1.00	0.56	5.00
Inner Queens	310	0.10	0.99	0.11	5.00
Northern Manhattan	343	0.10	0.99	0.30	5.00
Outer Queens	274	0.10	1.00	0.42	5.00
Staten Island	234	0.10	0.99	0.48	5.00
Middle Queens	243	0.10	0.98	0.10	5.00
Manhattan Core	311	0.12	1.00	0.78	5.00

Final Household and Person Weights

The final weights are effective in facilitating close matches to the regional totals for people, households, persons-in-households, and vehicles-in-households when using this dataset. The overall expanded and weighted survey values match the targets well (Figure 3 through Figure 5), with a mean absolute percent error (MAPE) of 2.1% and 1.6% to ACS values for household and person level variables, respectively. (Note that three target groups (car ownership: *h_car*; worker status: *p_worker*; and typical commute mode: *p_wmode*) are included in these figures for reference but were not matched in the weighting routine as described above; see section 3.2. Hence the relatively poor apparent match to the ACS targets is intentional for these three cases.)

As mentioned previously, matching the survey data to the target data even more closely can be achieved by relaxing the constraints on the ratio of the final to initial weights. However, this introduces more variance in the final weights and thereby increases the statistical error in any estimates. Allowing for more extreme weights also increases the likelihood of travel behavior analyses being impacted by extreme or outlier weights, which could unknowingly bias an estimate.

Figure 3: Comparison of weighted counts to targets for CMS Zone 1 – 5[7]



[7] Note that three target groups (car ownership: *h_car*; worker status: *p_worker*; and typical commute mode: *p_wmode*) are included in these figures for reference but were not matched in the weighting routine as described above (see section 3.2)

Figure 4: Comparison of weighted counts to targets for CMS Zone 6 - 10

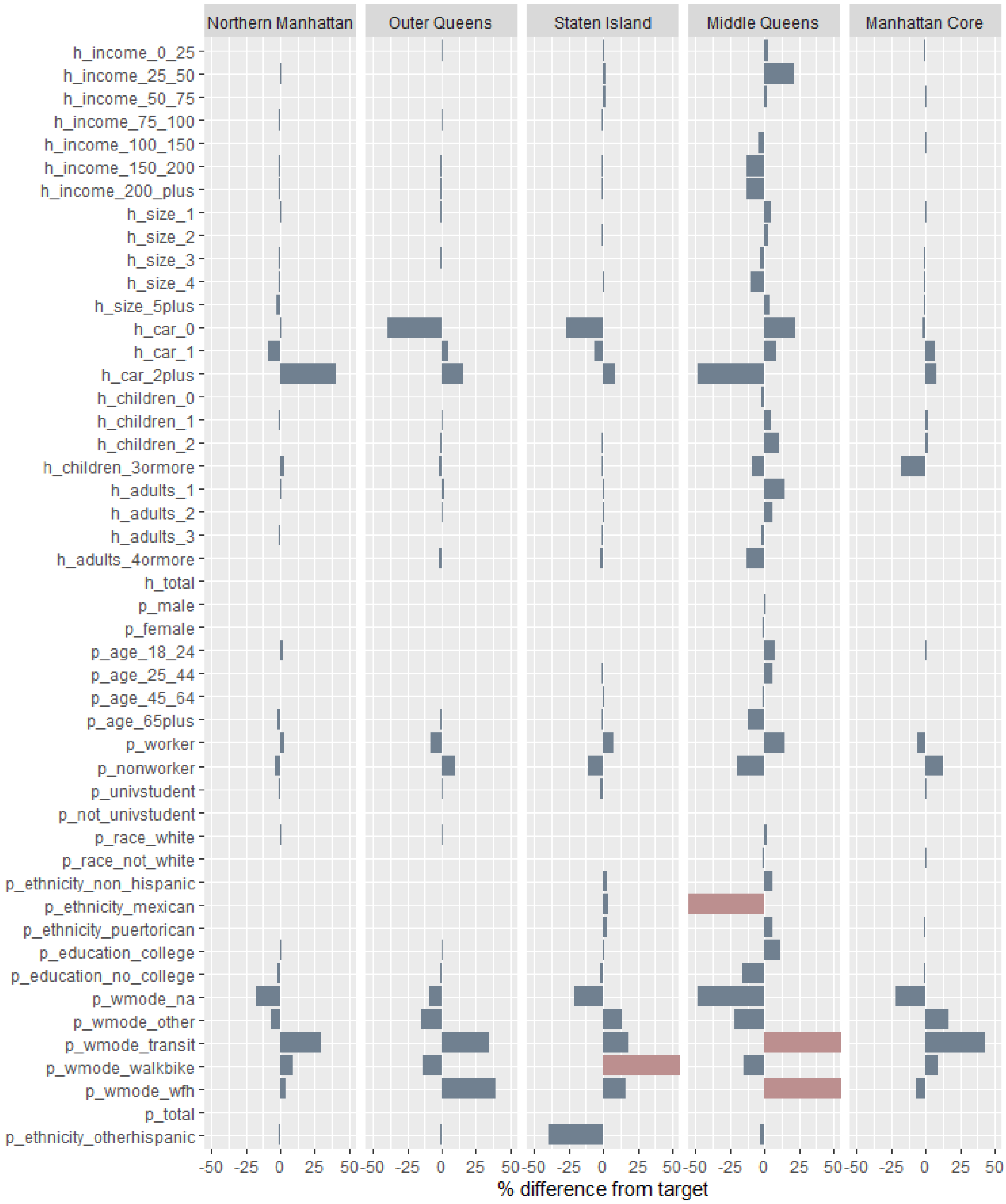
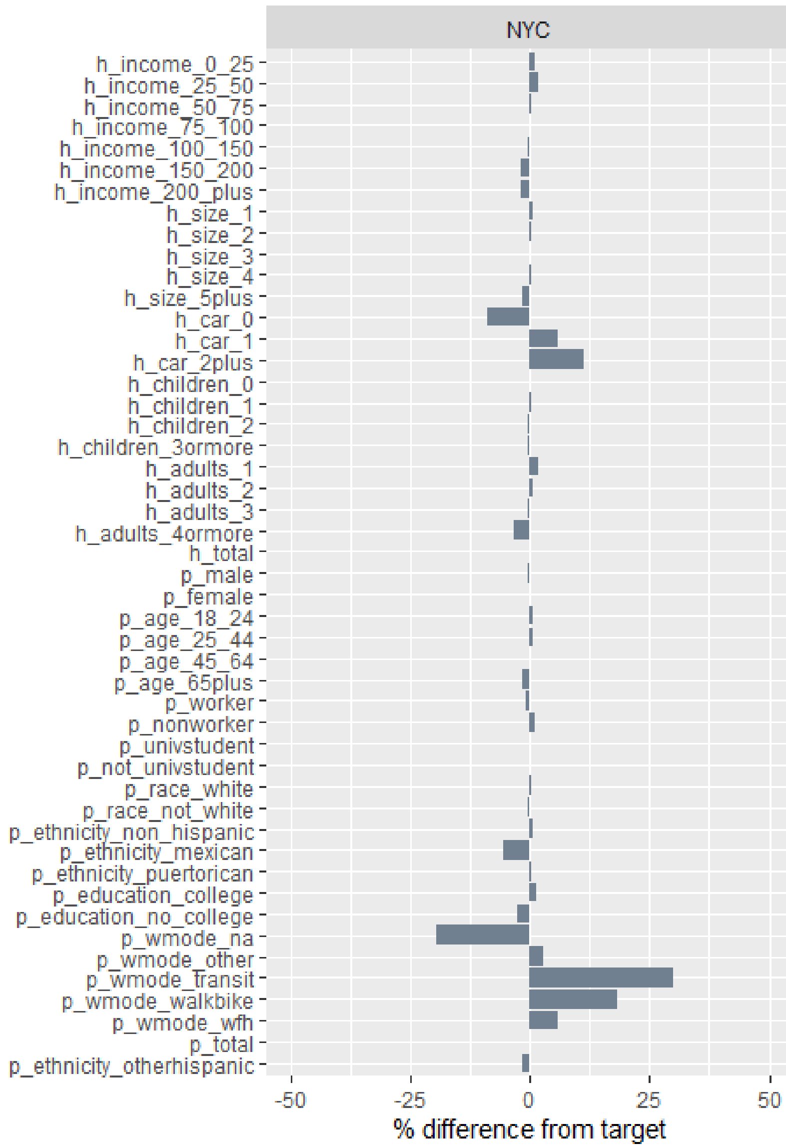


Figure 5: Comparison of weighted counts to targets for NYC



3.4 Creating Day-Level Weights

Because smartphone participants record up to seven days of travel, while online and call center participants only record one day, it is important to weight the trip data such that smartphone participants are not over-represented in data aggregations. RSG applied the following approach to create an "average weekday" day-level weight (wkday_weight) as follows:

- Define weekdays as Monday through Thursday as discussed previously.
- For each respondent, count the number of weekdays (N) for which the respondent provided complete and valid data. Set the person-day level weight equal to the person-level weight divided by N. In this way, when the data is weighted and aggregated, the sum of the person-day weights across days for each person is equal to the person weight, and the weighted results will reflect an average day for each respondent.

This method results in an "average weekday" for each respondent regardless of the number of days of data provided making the multi-day smartphone-based data compatible with the single-day online and call center-based data.

RSG also calculated a seven-day weight (svnday_weight) that captures behavior across the entire week. The calculation is similar to that above, except that N is the total number of days for which the respondent provided complete and valid data. This provides a way to compare the 2022 survey with the 2019 survey, which collected responses from all participants across all seven days of the week. Note that in the 2022 survey, online and call-center respondents were only asked to provide data for Tuesdays, Wednesdays, and Thursdays, so the seven-day weights are somewhat biased toward weekday behavior.

3.5 Adjusting for Trip Rates Non-response

Trip-Rate Adjustments

In reviewing all rMove™ smartphone-based survey data, RSG has found that the trip rates from the smartphone-based survey data are frequently 15–20% higher than those from online survey data. There are three main reasons for this:

- Smartphone-owning households have different socio-demographic characteristics than non-smartphone households and tend to make more trips.
- There are about twice as many "stay at home" days with no reported trips in the online and call center-based data in comparison to the smartphone-based data.
- Even on days with one or more reported trips, there are more trips per day reported on average in the smartphone-based data than in the online and call center-based data.

In this study, the trip rates for the online and call-center surveys were adjusted to better match those of the smartphone-based surveys.

The starting point for the trip-rate bias correction was the person-day weights (wkday_weight and svnday_weight). The following steps were then taken to adjust trip rates:

1. Trips were segmented into the following four trip types that have different levels of underreporting. Then for each person-day in the sample, the number of trips were counted by type.
    a. Home-based work/school trips
    b. Home-based other trips
    c. Non-home-based work/school trips
    d. Non-home-based other trips
2. For each trip type, a Poisson regression model was estimated where the dependent variable was the number of trips of that type for the person-day. The independent variables were the set of household and person variables, including age, income, employment, student status, education, telework frequency, and dummy variables for online and call center-based person-days (see for example Table 8).

For each person-day and for each trip type, the estimated regression model was applied with and without the bias coefficients. The ratio of the two estimates resulted in a factor to apply to the trip weight for that person-day. For example, if the model predicted 1.10 trips with the estimated model and 1.32 trips with the bias parameters set to 0 for an online or call center-based person-day, then a factor of 1.32/1.10 = 1.2 was used to multiply the person-day weight to get an adjusted trip weight. For smartphone respondents, the bias coefficients do not apply, so the factor was always 1.0 and the trip weight equaled the person-day weight. A lower bound of 1.0 and an upper bound of 2.0 was placed on ratios to avoid extreme adjustment to the weights. The specifications for each of the four regression models are shown in Table 8, Table 9, Table 10, and Table 11. The resulting trip adjustment factors by diary method and trip type are shown in Table 12. The final trip weights are in the trip_wkday_weight and trip_svnday_weight variables.

Table 8: Home-Based Work trip model

PARAMETER	DESCRIPTION	ESTIMATE	STD ERROR	T-STATISTIC
(Intercept)		-1.3885	0.114	-12.17
online_data	Online diary	-0.6445	0.064	-10.15
call_center	Call center data	-1.5086	0.304	-4.95
age_under_25	Under age 25	0.4123	0.103	4.00
age_25_45	Age 25 to 45	-0.1057	0.082	-1.29
age_45_65	Age 45 to 65	-0.0020	0.081	-0.02
income_under_50k	Income less than 50k	-0.1775	0.064	-2.76
income_50k_to_100k	Income 50k – 100k	-0.0607	0.055	-1.09
income_100k_to_150k	Income 100k – 150k	0.2222	0.061	3.63
employed_ft	Employed full-time	1.2958	0.073	17.75
employed_pt	Employed part-time	1.0419	0.099	10.53
employed_self	Self-employed	0.6529	0.120	5.43

PARAMETER	DESCRIPTION	ESTIMATE	STD ERROR	T-STATISTIC
bachelors	Has bachelor's degree	-0.3974	0.057	-7.03
graduate_degree	Has masters/PhD	-0.0721	0.055	-1.32
is_student	Is student	-0.5527	0.074	-7.48
two_plus_jobs	Works 2+ Jobs	0.2119	0.063	3.36
sf_home	Lives in single family home	0.3158	0.054	5.80
telework_freq %in% c(4:5)	Telework 1-3 days per week	0.0669	0.094	0.71
telework_freq %in% 6	Telework 1-3 days per month	-0.4110	0.238	-1.73
telework_freq %in% c(7, 996)	Telework less than monthly or never	-0.0050	0.090	-0.06

McFadden’s rho-squared: 0.089

Table 9: Home-Based Other trip model

PARAMETER	DESCRIPTION	ESTIMATE	STD ERROR	T-STATISTIC
(Intercept)		0.5918	0.051	11.49
online_data	Online diary	-0.9295	0.029	-31.81
call_center	Call center data	-0.6048	0.057	-10.69
age_under_25	Under age 25	0.0325	0.049	0.67
age_25_45	Age 25 to 45	0.2611	0.031	8.34
age_45_65	Age 45 to 65	0.1482	0.031	4.83
income_under_50k	Income less than 50k	-0.3157	0.027	-11.86
income_50k_to_100k	Income 50k – 100k	-0.1624	0.025	-6.56
income_100k_to_150k	Income 100k – 150k	0.0560	0.028	2.03
employed_ft	Employed full-time	-0.4821	0.025	-19.57
employed_pt	Employed part-time	-0.2305	0.042	-5.54
employed_self	Self-employed	-0.2464	0.041	-5.99
bachelors	Has bachelor's degree	0.0965	0.023	4.19
graduate_degree	Has masters/PhD	0.0193	0.025	0.77
is_student	Is student	0.3555	0.041	8.70
two_plus_jobs	Works 2+ Jobs	0.0006	0.034	0.02
sf_home	Lives in single family home	0.0403	0.025	1.61

PARAMETER	DESCRIPTION	ESTIMATE	STD ERROR	T-STATISTIC
telework_freq %in% c(4:5)	Telework 1-3 days per week	-0.1975	0.053	-3.73
telework_freq %in% 6	Telework 1-3 days per month	0.2065	0.104	1.99
telework_freq %in% c(7, 996)	Telework less than monthly or never	-0.0706	0.045	-1.58

McFadden's rho-squared: 0.74

Table 10: non-Home-Based Work trip model

PARAMETER	DESCRIPTION	ESTIMATE	STD ERROR	T-STATISTIC
(Intercept)		-0.8783	0.076	-11.54
online_data	Online diary	-0.5092	0.037	-13.70
call_center	Call center data	-2.1073	0.287	-7.35
age_under_25	Under age 25	0.5487	0.067	8.17
age_25_45	Age 25 to 45	0.0922	0.055	1.67
age_45_65	Age 45 to 65	0.2032	0.055	3.71
income_under_50k	Income less than 50k	0.0058	0.039	0.15
income_50k_to_100k	Income 50k – 100k	0.1104	0.033	3.33
income_100k_to_150k	Income 100k – 150k	-0.0277	0.043	-0.65
employed_ft	Employed full-time	1.5886	0.050	31.77
employed_pt	Employed part-time	1.4137	0.064	22.24
employed_self	Self-employed	1.6139	0.065	24.91
bachelors	Has bachelor's degree	-0.3247	0.035	-9.36
graduate_degree	Has masters/PhD	0.0106	0.034	0.31
is_student	Is student	-0.6490	0.043	-14.92
two_plus_jobs	Works 2+ Jobs	0.1116	0.039	2.86
sf_home	Lives in single family home	0.0337	0.037	0.90
telework_freq %in% c(4:5)	Telework 1-3 days per week	0.3918	0.048	8.13
telework_freq %in% 6	Telework 1-3 days per month	-0.9801	0.192	-5.12
telework_freq %in% c(7, 996)	Telework less than monthly or never	-0.1882	0.057	-3.29

McFadden's rho-squared:0.132

Table 11: non-Home-Based Other trip model

PARAMETER	DESCRIPTION	ESTIMATE	STD ERROR	T-STATISTIC
(Intercept)		0.8080	0.042	19.21
online_data	Online diary	-0.1938	0.020	-9.77
call_center	Call center data	0.0479	0.040	1.20
age_under_25	Under age 25	-0.0488	0.040	-1.23
age_25_45	Age 25 to 45	0.1097	0.026	4.19
age_45_65	Age 45 to 65	-0.0373	0.026	-1.45
income_under_50k	Income less than 50k	0.0485	0.021	2.27
income_50k_to_100k	Income 50k – 100k	-0.0389	0.021	-1.88
income_100k_to_150k	Income 100k – 150k	-0.1386	0.026	-5.28
employed_ft	Employed full-time	-0.0985	0.021	-4.62
employed_pt	Employed part-time	0.1619	0.033	4.93
employed_self	Self-employed	-0.0420	0.035	-1.20
bachelors	Has bachelor's degree	0.0336	0.019	1.73
graduate_degree	Has masters/PhD	-0.0806	0.021	-3.78
is_student	Is student	0.1950	0.032	6.05
work_loc_varies	Work location varies	-0.0654	0.028	-2.29
sf_home	Lives in single family home	-0.1115	0.023	-4.92
telework_freq %in% c(4:5)	Telework 1-3 days per week	0.0043	0.039	0.11
telework_freq %in% 6	Telework 1-3 days per month	0.2721	0.081	3.37
telework_freq %in% c(7, 996)	Telework less than monthly or never	-0.1334	0.037	-3.57

McFadden's rho-squared: 0.013

Table 12: Trip adjustment factors

TRIP TYPE	RMOVE	CALL CENTER	ONLINE
Home-based work	1	2.00	1.91
Home-based other	1	1.83	2.00
Non-home-based work	1	2.00	1.66
Non-home-based other	1	1.00	1.21

4.0 Data User Guide

Transportation planners rely on detailed travel data to inform their planning. The 2022 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 frequency, with four complete travel surveys since 2017. This allows for trend analysis over time, which is often not feasible when collection occurs less often.

4.1 Dataset Overview

The data can be seen as composed of two parts:

- Part one, also called the “recruit survey,” collected information about household composition, demographics, and typical travel behavior.
- Part two, also called the “travel diary,” required participants to record their travel during an assigned travel period.

“Complete” households met the following conditions:

- The participant completed the recruit survey in full.
- The participant completed a travel diary for at least one weekday. Only one person in the household (identified as person 1) completed a travel diary.

The 2022 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**: contains data about the characteristics of the participant’s household, including household size, income, type of residence, number of vehicles, bicycles, and micromobility devices, package delivery location, and other demographic and transportation-related information.
- **Person**: contains characteristics of individual members of the participant’s household, including age, race, gender, disability status, education, employment, remote work, typical commute mode, biking frequency, shared services, and other demographic and transportation-related information.
- **Vehicle**: contains characteristics of each vehicle in the participant’s household.
- **Day**: contains information on each day of the participant’s assigned travel period, including deliveries, remote work duration, online shopping, reasons for no travel if applicable, and start/end location of the day, with between one and seven travel days recorded for each participant.
- **Trip**: contains characteristics of each trip made by the participants, including origins and destinations, mode, purpose, transportation related fees such as transit, parking, or taxi pay, parking location for the bikes or micromobility devices, number of travelers, trip duration, and trip distance.

- **Location**: Two or more records for each trip collected via smartphone (if any).

To protect the privacy of survey participants, the Location table is not posted on Open Data; all other location information, such as home and work location, is only provided at the CMS zone level.

4.2 Data Preparation

This section summarizes the methods used to prepare the data. Given that all data were collected in a "controlled" environment (e.g., survey answers are validated in real-time), data preparation was primarily focused on coding variables and deriving new fields to facilitate analysis.

Initial Data Review

Before reviewing the data for completion, RSG removed households from the dataset that met the following exclusion criteria:

- Household reported a home location outside the study region. Most households dropped during initial review were excluded for this reason.
- Household reported contact information that matches other households (indicating duplicates). In these cases, RSG kept the first "household" to report their travel diary and removed the subsequent records.

Geographic Variables

Census PUMA, tract, block group, and block shapefiles were downloaded via the R Tigris package and spatially joined to the reported home coordinates. The final reported block group may not always match the block group ascribed to the household's sample address (which is used to determine the sample segment) for a few reasons: Sample addresses are geocoded differently than survey addresses, sample addresses sometimes are coded to a mailbox location rather than a home location, and home addresses in the survey are not always geocoded to a person's exact home (e.g., a cross street nearby). Because a person's reported home address is considered to be more recent and typically more accurate than the sample address, the geographic variables are derived using this address.

Households retain their initial sample segment assignment, as this is what determines their probability of being invited according to the information in the sample address file.

4.3 Data Coding and Labeling

Time and location standards:

Unless otherwise noted, all timestamps are set to the local time at the time they were collected. All location latitude and longitude information are presented in WGS84 format.

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 survey data table cell may be missing data for one of four reasons:

1) Value or response is missing due to survey logic, participant non-response, or error.

   Example: Participants who traveled by bus were not asked if they were the driver or passenger on the trip.

   Coded as: 995 for categorical variables, blank/NA for continuous variables

2) A respondent indicated that the question was not applicable and skipped that question.

   Example: Some participants did not share how they pay to park at work because they do not park at work (e.g., carpool).

   Coded as: 996 (often labeled as “Not applicable”)

3) A respondent indicated that they didn’t know the answer and skipped that question.

   Example: Some participants who made a vehicle trip and paid to park the vehicle may not remember the amount they paid.

   Coded as: 998 (Don’t know)

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

   Example: Some participants chose not to provide their household income.

   Coded as: 999 (Prefer not to answer)

Other notes about missing survey data:

- For a survey to be complete, all survey questions asked of the participant must have been answered.
- Continuous variables (e.g., trip distance, trip duration) are not coded with missing value codes and are instead left empty when missing to avoid interfering with statistical calculations.
- Due to the large size of the location table, missing values were left exactly as they were collected. Speed, heading, and accuracy can all potentially contain missing values that are either stored as “-1”, NA, or 0. Analysis on those fields should filter to where the values are greater than zero.
- The ev_typical_charge set of variables are not included in the household data due to a technical survey error that overwrote the data during data collection.

Outliers

Continuous variables (e.g., trip distance, trip duration, parking cost) in the dataset may contain outliers. Data users should be aware of these outliers when calculating summary statistics (e.g., mean) for these variables.

4.4 Derived and Recoded Variables

This dataset includes a combination of variables that were actively collected via survey questions, passively collected via rMove or other metadata, implicitly assigned (e.g., administrative variables such as

ID numbers), and derived or recoded (calculated from some combination of other variables). Key derived or recoded variables in this dataset are summarized below.

Household-level Derived Variables

- Completion status
- Home geographies (block group, zone)
- Aggregate income (based on the initial and follow-up income questions)

Person-level Derived Variables

- Completion status
- Number of complete days
- Work/school geographies (state, county, block group, zone)

Day-level Variables

- Completion status
- Number of trips per day
- Day completion status

Trip-level Variables

- Trip speed
- Trip path distance (based on the GPS location data)
- Trip origin and destination geographies (state, county, block group, zone)
- Departure time (imputed in some cases)
- Trip purpose (imputed in some cases)
- Mode type and purpose categories

4.5 Imputation

Departure Time

In some cases, the rMove app may have detected the start of a trip after its true start time, which can yield invalid or extreme values for trip duration and speed. In these cases, the fields depart_date, depart_hour, and depart_minute were adjusted for “late pickup” conditions using the following approach:

- Departure time was imputed using the median speed between all locations along the trip, excluding the origin point, and the distance between the origin and the next point on the trip. For trips with fewer than three recorded locations, imputed departure time is set three minutes earlier than the original departure time to compensate for rMove’s 3-5-minute ping interval. Note that some trips that are the result of split loop trips may only have three or fewer points but will use the imputed depart time from before the loop trip was split and thus may not be included in this rule.

- If the imputed departure time overlaps with the previous trip's arrival time, the previous trip's arrival time was instead used as the departure time. Regardless of the number of locations along a trip, if the imputed departure time was later than the initially reported departure time, the imputed departure time is set to the original departure time. User-added trips as well as long distance passenger mode trips are also set to the original departure time, as user-added trips are not subject to "late pickup" conditions, and long-distance passenger modes are often plane trips where all collected traces contain speed information from other modes and thus are less reliable (as rMove cannot collect locations when a phone is in "airplane mode").

Duration and speed are calculated based on the imputed departure time.

Purpose

Respondents report the purpose of the trip destination in each trip survey. The origin purpose is derived from the destination purpose of the previous trip, except for the first trip in the travel period or where an rMove trip occurs after a trip with item non-response. For the first trip in the travel period, the origin purpose can be inferred from "begin_day" in the day table.

When purpose was not asked because an analyst split a user-reported trip during data cleaning (creating a new destination along a trip), purpose values are derived where possible based on proximity (within 150 meters) to estimated home, work, or school locations. If the location is not proximate to home, work, or school locations, the purpose is set to "other."

The purpose category variables (o_purpose_category, d_purpose_category) contain aggregated purpose values based on the type of purpose at the origin/destination of each trip. Dataset users are welcome to perform their own recoding of the purpose categories as well.

Trip purposes have been imputed in cases where a purpose reported by the user is assumed to be inaccurate based on information about that person's reported habitual locations and other trips (primarily to home, work, and school locations). The trip purpose imputation approach was applied to all rMove trips in person-days with at least 1 complete trip and no more than 10 incomplete trips. ("Incomplete" trips are trips for which the respondent did not answer the trip-specific survey questions about purpose, mode, etc. for the given trip.)

The approach was to apply various "tests" in logical sequence to trips for which the stated purpose is not consistent with the location type based on the reported habitual locations. In general terms, the tests were designed to:

- Check the respondent's reported destination purpose when it conflicts with the destination location type. (The details of the tests depend on the trip purpose, with different criteria used for change-mode trips, escort trips, linked transit trips, trips with home destinations but other reported purposes, etc.)
- Identify cases where respondents swapped the order of two or more trips when reporting their details.
- Identify cases where respondents may have omitted a trip and shifted remaining reported trip details by one trip when reporting the rest of their trips.

- Fill in missing data by sampling destination purposes from other trips made to the same locations, either by the same respondent or by other respondents.

4.6 Reminders 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.

4.7 Practical Tips for Data Analysis

Many ways exist to view, join, summarize, and map 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 name	Variable(s) to join to other survey data tables
Household	hh_id
Person	hh_id, person_id
Vehicle	hh_id
Day	hh_id, person_id, day_num
Trip	hh_id, person_id, day_num, trip_id
Location	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 personal citywide travel for the time period studied (September – November 2022). This means it does not include commercial vehicle travel (including by delivery bicyclists), travel for persons residing in group quarters outside of the address-based sampling frame (e.g., college dorms, institutional housing), travel from nonresidents (e.g., commuters from outside the city, and visitors to the city), 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 day (svnday_weight) or one weekday (wkday_weight) for all persons residing in the survey region. This is equal to the number of adult persons in the region.
- **Trip**: Represents the total number of trips all adults 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.**
    - When working with trip and day data, it is important to consider that participants using smartphones app recorded more days and trips compared to one-day online and call center participants. To ensure equal representation, the trip rates (not attributes) for the online and call-center surveys participants were adjusted to better match those of the smartphone-based surveys. For more details, refer to the detailed methodology in section 3.
3. **Make sure to use the correct weight for the analysis you intend to perform.**
    - CMS is a weighted sample, and it is necessary to use the weighting variables to generate estimates and standard errors that accurately represent the population of the city. In general, use household weight for household-level analyses; use person weight for person-level analyses. Apply day weight for household-day and person-day analyses, and trip weight for trip-level analyses. When working with a merged file that includes household and person records, use person weights to estimate person characteristics. Remember to exercise caution when analyzing person and household characteristics from merged files.

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://www.nyc.gov/html/dot/html/about/citywide-mobility-survey.shtml

Appendix A

Table 13: Household-level unweighted Sample counts (Number of respondents)

VARIABLE	TOTAL	SOUTHERN BRONX	NORTHERN BRONX	INNER BROOKLYN	OUTER BROOKLYN	INNER QUEENS	NORTHERN MANHATTAN	OUTER QUEENS	STATEN ISLAND	MIDDLE QUEENS	MANHATTAN CORE
h_income_0_25	422	107	53	51	38	22	54	35	17	28	17
h_income_25_50	478	74	62	47	44	48	64	39	31	38	31
h_income_50_75	518	59	57	62	51	61	57	44	45	44	38
h_income_75_100	358	19	40	48	33	49	41	37	24	35	32
h_income_100_150	524	25	42	64	42	53	62	61	57	52	66
h_income_150_200	258	6	19	43	21	33	26	32	27	23	28
h_income_200_plus	408	1	15	103	25	44	39	26	33	23	99
h_size_1	953	99	106	111	68	92	148	60	44	76	149
h_size_2	989	72	74	173	76	135	103	87	88	77	104
h_size_3	474	53	54	70	44	37	46	52	37	46	35
h_size_4	339	34	34	41	42	32	31	46	37	25	17
h_size_5plus	211	33	20	23	24	14	15	29	28	19	6
h_children_0	2,236	193	213	320	174	257	288	186	162	175	268
h_children_1	379	52	36	53	33	30	27	52	34	37	25
h_children_2	256	27	27	36	34	16	20	26	27	26	17
h_children_3ormore	95	19	12	9	13	7	8	10	11	5	1
h_adults_1	1,049	123	118	124	78	98	156	66	48	83	155
h_adults_2	1,383	101	116	228	122	168	133	137	127	119	132
h_adults_3	345	47	35	49	34	25	34	41	38	25	17
h_adults_4ormore	189	20	19	17	20	19	20	30	21	16	7
Total	2,966	291	288	418	254	310	343	274	234	243	311

Table 14: Person-level unweighted Sample counts (Number of respondents)

VARIABLE	TOTAL	SOUTHERN BRONX	NORTHERN BRONX	INNER BROOKLYN	OUTER BROOKLYN	INNER QUEENS	NORTHERN MANHATTAN	OUTER QUEENS	STATEN ISLAND	MIDDLE QUEENS	MANHATTAN CORE
p_male	1,276	88	100	174	113	140	150	131	112	114	154
p_female	1,690	203	188	244	141	170	193	143	122	129	157
p_age18_24	169	31	14	17	12	13	25	17	13	11	16
p_age25_44	1,433	151	120	274	110	202	165	100	67	102	142
p_age45_64	895	89	95	91	81	62	109	91	93	94	90
p_age_65plus	469	20	59	36	51	33	44	66	61	36	63
p_univ_student	213	32	20	24	19	16	34	30	8	15	15
p_not_univstudent	2,753	259	268	394	235	294	309	244	226	228	296
p_education_college	2,519	203	234	374	212	276	294	224	190	220	292
p_education_no_college	447	88	54	44	42	34	49	50	44	23	19
p_race_white	1,364	38	106	258	112	146	159	97	142	107	199
p_race_non_white	1,602	253	182	160	142	164	184	177	92	136	112
p_ethnicity_non_hispanic	2,387	132	178	376	221	251	261	246	215	214	293
p_ethnicity_mexican	90	18	9	12	8	13	14	2	9	1	4
p_ethnicity_puertorican	201	64	51	11	8	11	23	9	8	12	4
p_ethnicity_otherhispanicorigin	288	77	50	19	17	35	45	17	2	16	10
Total	2,966	291	288	418	254	310	343	274	234	243	311

Note: Variables with decimals include variables with imputed values for respondents who chose not to provide that information and variables affected by the day pattern adjustments. Imputation was done in a probabilistic fashion.

Table 15: Household-level Target counts (Persons at least 18 years of age)

VARIABLE	TOTAL	SOUTHERN BRONX	NORTHERN BRONX	INNER BROOKLYN	OUTER BROOKLYN	INNER QUEENS	NORTHERN MANHATTAN	OUTER QUEENS	STATEN ISLAND	MIDDLE QUEENS	MANHATTAN CORE
h_income_0_25	1,164,843	190,856	113,884	170,430	202,047	52,560	108,338	142,057	36,179	51,035	97,457
h_income_25_50	1,150,698	167,620	119,783	138,353	217,002	74,656	80,863	156,165	56,448	60,183	79,625
h_income_50_75	943,571	80,814	89,351	120,511	185,491	77,592	56,409	149,632	51,204	68,170	64,397
h_income_75_100	793,512	60,119	64,544	106,959	142,811	64,876	48,004	133,401	51,656	55,506	65,636
h_income_100_150	1,035,624	54,326	77,320	136,555	197,161	74,690	45,525	197,325	83,170	75,575	93,977
h_income_150_200	637,212	23,666	45,871	99,418	93,618	40,989	29,101	121,369	58,262	43,841	81,077
h_income_200_plus	989,116	16,676	42,903	179,800	117,060	44,007	45,694	133,383	66,034	57,722	285,837
h_size_1	936,877	72,828	72,861	135,679	117,869	59,061	79,367	99,466	35,778	41,926	222,042
h_size_2	1,638,135	118,831	120,913	288,340	253,782	110,207	115,659	196,222	83,339	87,437	263,405
h_size_3	1,335,271	124,428	119,699	206,069	209,734	84,097	88,456	217,415	73,442	84,945	126,986
h_size_4	1,215,400	112,399	97,695	144,802	225,837	73,463	68,247	209,840	87,257	81,414	114,446
h_size_5plus	1,588,893	165,591	142,488	177,136	347,968	102,542	62,205	310,389	123,137	116,310	41,127
h_children_0	3,768,103	262,721	281,262	563,182	585,576	262,171	253,396	564,726	196,484	232,226	566,359
h_children_1	1,157,279	126,117	118,178	162,737	197,329	69,199	71,267	188,522	70,590	73,253	80,087
h_children_2	1,066,882	111,385	90,327	128,274	190,505	72,313	57,522	175,285	87,530	66,629	87,112
h_children_3ormore	722,312	93,854	63,889	97,833	181,780	25,687	31,749	104,799	48,349	39,924	34,448
h_adults_1	1,257,407	139,916	112,629	180,574	170,296	67,900	111,602	126,873	50,265	51,237	246,115
h_adults_2	2,904,993	223,545	215,497	490,418	498,789	172,835	178,548	386,051	172,146	161,297	405,867
h_adults_3	1,329,969	126,022	119,774	166,829	243,168	93,141	78,436	237,965	93,677	90,112	80,845
h_adults_4ormore	1,222,207	104,594	105,756	114,205	242,937	95,494	45,348	282,443	86,865	109,386	35,179
Total	6,714,576	594,077	553,656	952,026	1,155,190	429,370	413,934	1,033,332	402,953	412,032	768,006

Table 16: Person-level Target counts (Persons at least 18 years of age)

VARIABLE	TOTAL	SOUTHERN BRONX	NORTHERN BRONX	INNER BROOKLYN	OUTER BROOKLYN	INNER QUEENS	NORTHERN MANHATTAN	OUTER QUEENS	STATEN ISLAND	MIDDLE QUEENS	MANHATTAN CORE
p_male	3,162,598	243,809	240,417	440,208	505,537	222,350	217,958	497,015	188,305	201,314	405,685
p_female	3,551,977	288,998	285,881	507,448	581,076	220,598	247,987	555,890	202,520	213,959	447,620
p_age18_24	634,643	71,512	57,267	85,170	110,382	33,569	39,275	99,441	40,048	37,703	60,276
p_age25_44	2,595,996	207,238	186,545	466,868	369,777	200,877	190,866	341,606	132,039	143,459	356,721
p_age45_64	2,117,015	173,001	169,460	257,379	353,288	132,538	143,763	368,970	135,580	144,357	238,679
p_age_65plus	1,366,921	81,056	113,026	138,239	253,166	75,964	92,041	242,888	83,158	89,754	197,629
p_univ_student	540,474	52,129	41,700	73,637	90,039	33,624	40,094	86,794	33,072	35,200	54,185
p_not_univstudent	6,174,101	480,678	484,598	874,019	996,574	409,324	425,851	966,111	357,753	380,073	799,120
p_education_college	4,144,339	229,252	299,437	634,608	614,644	257,240	289,728	606,767	237,978	237,866	736,819
p_education_no_college	2,570,236	303,555	226,861	313,048	471,969	185,708	176,217	446,138	152,847	177,407	116,486
p_race_white	2,312,990	32,706	99,131	354,962	409,417	144,433	108,257	251,113	244,923	121,793	546,255
p_race_non_white	4,401,585	500,101	427,167	592,694	677,196	298,515	357,688	801,792	145,902	293,480	307,050
p_ethnicity_non_hispanic	4,844,043	176,804	288,566	744,982	916,990	267,849	244,170	851,526	324,436	282,820	745,900
p_ethnicity_mexican	232,359	38,188	15,227	37,165	35,115	34,972	13,906	13,419	13,115	17,956	13,296
p_ethnicity_puertorican	465,311	108,294	87,043	62,281	46,235	10,949	33,927	38,441	24,444	16,746	36,951
p_ethnicity_otherhispanicorigin	1,172,862	209,521	135,462	103,228	88,273	129,178	173,942	149,519	28,830	97,751	57,158
Total	6,714,575	532,807	526,298	947,656	1,086,613	442,948	465,945	1,052,905	390,825	415,273	853,305

Table 17: Household-LEVEL DIFFERENCES BETWEEN WEIGHTED SAMPLE AND TARGET PUMS DATA

DEMOGRAPHIC	TOTAL	% DIFFERENCE FROM TARGET	SOUTHERN BRONX	NORTHERN BRONX	INNER BROOKLYN	OUTER BROOKLYN	INNER QUEENS	NORTHERN MANHATTAN	OUTER QUEENS	STATEN ISLAND	MIDDLE QUEENS	MANHATTAN CORE
h_income_0_25	12,536	1.1%	1,513	-113	1,194	-1,336	2,262	3,637	2,172	-120	1,344	1,983
h_income_25_50	19,935	1.7%	-3,524	-1,288	190	-2,170	4,771	3,027	998	603	15,381	1,947
h_income_50_75	-1,954	-0.2%	-1,631	-1,021	212	-2,544	-378	1,910	-123	668	-1,252	2,205
h_income_75_100	-2,726	-0.3%	-165	-884	-477	-2,099	-1,608	692	2,161	-824	-1,264	1,742
h_income_100_150	-7,840	-0.8%	-1,475	-784	-412	-3,817	-2,140	1,629	935	-1,108	-3,882	3,214
h_income_150_200	-10,885	-1.7%	-2,710	-1,006	-672	-1,563	-535	596	-358	-1,392	-5,076	1,831
h_income_200_plus	-9,065	-0.9%	-6,859	-1,047	-106	-2,128	497	799	-57	-1,334	-6,244	7,414
h_size_1	26,757	2.9%	1,455	840	126	475	-477	6,322	-599	546	3,050	15,019
h_size_2	13,091	0.8%	-3,614	-1,260	-46	-1,696	-860	3,658	1,704	-1,044	6,213	10,036
h_size_3	-3,259	-0.2%	-690	55	-1,091	-2,924	4,010	2,831	-1,419	-1,001	-3,689	659
h_size_4	3,672	0.3%	-2,986	-940	328	-3,417	18,125	786	2,232	-722	-6,564	-3,170
h_size_5plus	-40,261	-2.5%	-9,017	-4,839	612	-8,095	-17,929	-1,306	3,810	-1,286	-2	-2,209
h_children_0	61,125	1.6%	227	-1,214	2,985	-6,984	15,677	8,143	9,034	370	5,815	27,072
h_children_1	534	0.0%	-3,798	-2,539	347	-2,521	-2,477	1,298	2,554	-302	1,704	6,268
h_children_2	-18,068	-1.7%	-7,025	-1,793	-501	-2,568	-15,790	1,318	-2,706	-2,147	3,777	9,367
h_children_3ormore	-43,590	-6.0%	-4,256	-598	-2,902	-3,584	5,459	1,532	-3,153	-1,428	-12,289	-22,371
h_adults_1	11,923	0.9%	-5,154	-1,322	1,156	-2,737	7,887	1,437	4,388	-1,261	5,933	1,596
h_adults_2	33,584	1.2%	-4,580	-1,575	303	-6,082	13,545	6,588	6,837	1,090	6,414	11,044
h_adults_3	-5,491	-0.4%	-895	-1,854	249	-2,669	-3,335	1,951	772	-1,009	-3,240	4,539
h_adults_4ormore	-40,019	-3.3%	-4,223	-1,392	-1,779	-4,170	-15,229	2,315	-6,269	-2,329	-10,100	3,157
Total	-1	0.0%	-14,852	-6,143	-71	-15,657	2,868	12,291	5,728	-3,508	-993	20,336

Table 18: Person-level differences between weighted sample and Target PUMS data

DEMOGRAPHIC	TOTAL	% DIFFERENCE FROM TARGET	SOUTHERN BRONX	NORTHERN BRONX	INNER BROOKLYN	OUTER BROOKLYN	INNER QUEENS	NORTHERN MANHATTAN	OUTER QUEENS	STATEN ISLAND	MIDDLE QUEENS	MANHATTAN CORE
p_male	-15,099	-0.5%	14,282	7,994	210	24,170	-8,175	-19,852	-6,325	4,499	-1,154	-30,748
p_female	15,101	0.4%	32,136	13,221	4,089	28,750	-2,535	-19,868	-7,520	4,122	-3,080	-34,214
p_age18_24	5,433	0.9%	5,587	1,737	632	4,485	-397	-2,406	-963	585	837	-4,664
p_age25_44	25,826	1.0%	23,612	9,124	4,115	21,279	-67	-14,976	-2,339	2,994	8,730	-26,646
p_age45_64	5,858	0.3%	14,559	7,596	1,019	17,049	-6,586	-12,051	-2,837	4,158	801	-17,850
p_age_65plus	-37,117	-2.7%	2,660	2,758	-1,468	10,107	-3,660	-10,288	-7,706	884	-14,601	-15,803
p_univ_student	1,536	0.3%	6,963	1,546	-267	4,186	-1,992	-3,708	-337	-141	-622	-4,092
p_not_univstudent	-1,532	0.0%	39,455	19,669	4,566	48,734	-8,717	-36,012	-13,507	8,762	-3,612	-60,870
p_education_college	46,710	1.1%	27,444	15,406	7,244	33,605	5,587	-20,581	-1,623	8,237	26,192	-54,801
p_education_no_college	-46,707	-1.8%	18,974	5,809	-2,945	19,315	-16,296	-19,139	-12,222	384	-30,426	-10,161
p_race_white	-14,549	-0.6%	5,164	5,827	4,461	20,488	-3,166	-8,001	-1,103	4,617	-363	-42,473
p_race_non_white	14,550	0.3%	41,254	15,388	-162	32,432	-7,544	-31,719	-12,741	4,003	-3,871	-22,490
p_ethnicity_non_hispanic	16,162	0.3%	16,155	11,739	4,144	44,753	-5,333	-19,614	-9,829	16,954	13,481	-56,288
p_ethnicity_mexican	-9,196	-4.0%	3,687	789	195	1,604	-526	-1,079	-17	830	-13,609	-1,070
p_ethnicity_puertorican	10,403	2.2%	9,428	3,725	-37	2,256	-136	-2,799	-534	1,166	344	-3,010
p_ethnicity_otherhispanic origin	-17,363	-1.5%	17,149	4,962	-3	4,307	-4,713	-16,228	-3,464	-10,329	-4,449	-4,595
Total	1	0.0%	46,419	21,215	4,299	52,920	-10,708	-39,720	-13,844	8,621	-4,233	-64,963