Reliability and Compatibility of Major Journey-to-Work Origin-Destination (OD) flows

investigating CTPP, LODES, Cellphone Signal, and Mobile Apps location data for selected zones in NYC.

Kyeongsu Kim
Senior Planner
Louis Berger Group
48 Wall Street, 16th floor, New York NY 10005

H. Joon Park, Ph.D., AICP
Office of Project Analysis/CEQR
NYCDOT - Division of Transportation Planning & Management
55 Water Street, 6 Floor New York, NY 10041

Louis Berger

Study Background

Trip Origin-Destination (OD) data, for smaller geographic levels such as traffic analysis zones (TAZ), are a fundamental input for transport planning and analysis (e.g., regional travel demand forecasting, traffic analysis, and environmental impact studies). A robust trip OD matrix helps better answer "when, where, why, how, and how far people travel"; ultimately, it supports policy-decision making for future transport infrastructure investments.

However, existing OD data sources (e.g., the Census Transportation Planning Package (CTPP) in the U.S., and household travel surveys) have limited reliability mainly due to their small sample size. Recently, the Longitudinal Employer-Household Dynamics Origin-Destination Employment Statistics (LODES), based on government administrative records of resident and workplace employment profiles, has become a new resource in the U.S. Yet the LODES also has data quality issues such as missing cash-paid jobs, counting telecommute employees, and failing to identify an actual workplace for employees at firms with multiple worksites; often, the headquarters or a regional accounting office is reported as a worksite.

A newly emerging discussion in the transport data community is how to use cellphone signal data and Global Positioning System (GPS) location data from mobile devices' apps as relevant analytic techniques become available. Currently, those data are owned by private firms and only tabulated matrices, from their proprietary data extraction methods (algorithm) are provided based upon clients' request. Hence, albeit the presumed strengths of big data with respect to trip OD counts and trajectory of all day trips, it is still uncertain whether the quality of private location data is reliable enough to employ in practice. This uncertainty also comes from the fact that the locational accuracy of cellphone signal data and mobile devices' GPS data has barely been validated when it comes to trip OD matrix.

This study is the first attempt to validate commercially available cellphone tower signal data and mobile app's location data with two public data sources: the CTPP and the LODES. We investigated journey-to-work trip ODs from each data source for three selected neighborhood zones in NYC and other broader zones.

Selected Zone Boundary

In this study, we selected three neighborhood zones located in Staten Island and Queens, NYC: Zone 1 in Staten Island, Zone 2 and Zone 3 in Queens (see figure 1 to figure 3; the scarlet polygons in figure 2 and 3 represents neighborhood zones). These neighborhoods are excellent places to investigate thanks to the dissimilar socio-economic and demographic characteristics of each neighborhood and different accessibility to major trip attraction sites (e.g., employment, shopping, and dining), which assume to affect travel behaviors in a typical transport model.

Zone 1 is a low-income neighborhood where a public housing complex is located. The public housing complex occupies almost two-thirds of the area of zone 1 (see the light blue polygon in figure 3-1; the yellow line represents TAZ boundary). Zone 2 is a residential neighborhood in Rego Park, Queens where a high proportion of the population are immigrants, especially from southwest Asia. Residents in Zone 2 enjoy good subway access to Midtown Manhattan. It is also important to note that residents in Zone 2 enjoy great access to shopping and employment opportunities in Rego Center, which is adjacent to the zone (see the light blue polygon in figure 3-2). The walking-distance urban style mall hosts commercial establishments such as Costco, Kohl's, Century 21 department store, Staples, and ToysRus. Lastly, Zone 3 is a residential neighborhood in Astoria, Queens. The neighborhood used to be a Greek community but for the last decade many millennials have moved in thanks to affordable rent and great subway access to Midtown Manhattan. Today, the neighborhood is one of many gentrifying neighborhoods in NYC. Unlike Zone 2, however, no major shopping or employment opportunities are available nearby; only a few retail stores and restaurants are located in this neighborhood.

Data Sources

This study examined cellphone tower signal data from AirSage and mobile apps location data from PlaceIQ (new to transport community) with two public data sources: the CTPP and the LODES. AirSage compiles and examines the cellphone signal data from two major cellphone carriers' while PlaceIQ compiles the foreground mobile apps location data from almost all mobile device (including smartphone and tablet) users. The foreground location data is only reported when the mobile apps are opened or running. Remember that when you download almost any mobile apps, you are asked to share your location information. This means that when you open an app or an app is running, your location is identified by the app's developers who then sell the location information to marketing companies.

We compare AirSage's HBW trips and PlaceIQ's home dwellers' day time location data (between 10 am to 12 pm and 1 pm to 3 pm) with CTPP and LODES. Our interest is to compare different sources' OD estimates between selected small zones (roughly, 300 square meters) and broader NYC community district boundaries (aggregate of multiple census tracts). Unfortunately, the neighborhood boundaries, initially delineated by NYCDOT and AirSage, align with block boundaries but are not necessarily conterminous with census tracts or TAZ boundaries. Since CTPP estimates are only available down to census tract- and TAZ- levels (in NYC, a TAZ boundary is equivalent to or smaller than a census tract) cannot be perfectly compared to LODES, AirSage, and PlaceIQ. In this study, we aggregated TAZ estimates that correspond with the selected zone boundary as a whole or in part for CTPP (see figure 2); that is, the CTPP OD estimates for this comparison study are based on the larger boundary than the actual neighborhood zone. It implies higher CTPP estimates than the real ODs, which cannot be extracted given the limited data availability of the CTPP. There are no boundary issues in comparing LODES, AirSage and PlaceIQ data because AirSage and PlaceIQ data was extracted based on pre-defined zones that were created by aggregating multiple census blocks.

Findings

The study found that the proportional distribution of worksite location looks similar, for each of the data sets analyzed, but the total amount of commute trips is dissimilar. . In general, the cellphone signal data appears to provide lower work trip estimates and undercount or miss short distance trips. In contrast, the mobile apps location data provides higher estimates; this can be explained by the fact that mobile apps data is based on a typical daytime location with includes locations such as schools, voluntary work, and cash-paid jobs, which are not captured in the LODES and unlikely in the CTPP (although reporting a cash-paid job is still possible). In summary, the result reveals the usefulness of cellphone signal and mobile apps location data but with careful interpretation. OD estimates can be undercounted in lower income neighborhoods where residents use non-major cellphone carrier service and overcounted in relatively affluent neighborhoods where multiple mobile device ownership prevails. This study calls for the necessity of more efforts to utilize different data resources by the transport planning community.

Comparison of each data source: from a Neighborhood zone to the rest (broader) of NYC zones

Figure 1 Comparison of workplace distribution from each data source

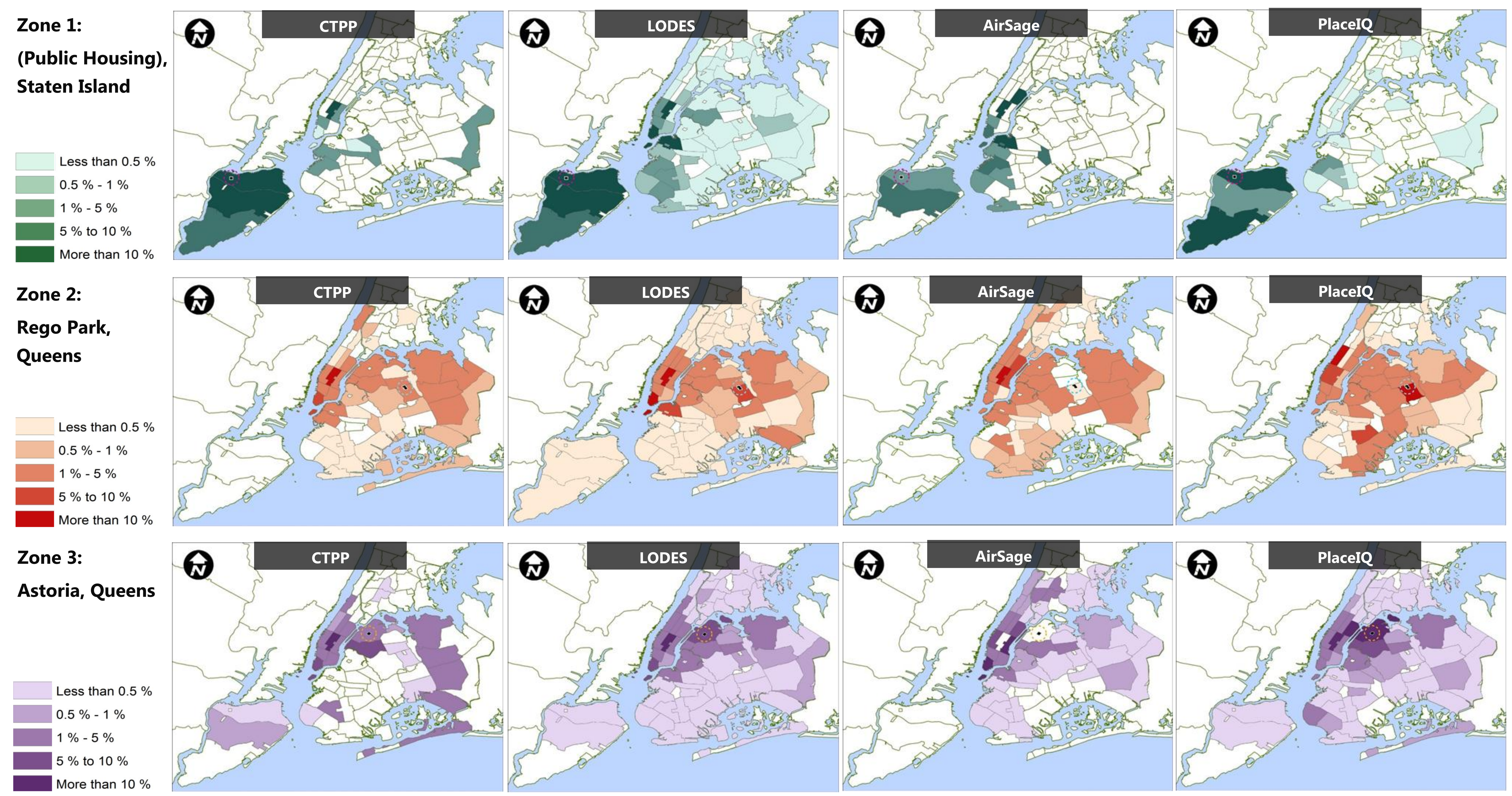

Figure 2 Selected neighborhood zones

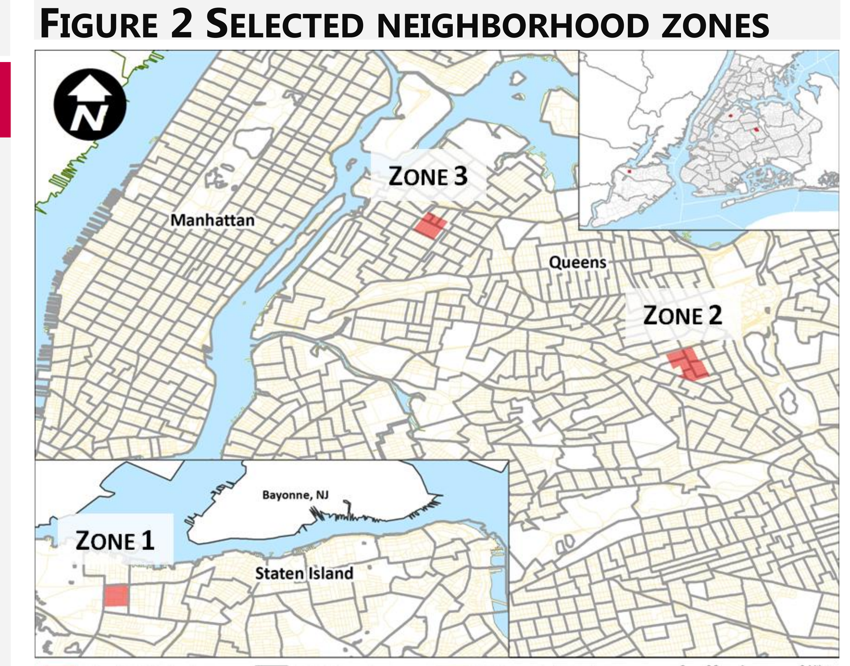

Figure 3 Google Earth view on the selected neighborhood zones with TAZ boundary

Table 1 Comparison of Each Data Estimates

	Zone 1				Zone 2				Zone 3			
Zone 1	Estimates	% by LEHD	% by age group (18 to 65yrs)	% by age group (over 18yrs)	Estimates	% by LEHD	% by age group (18 to 65yrs)	% by age group (over 18yrs)	Estimates	% by LEHD	% by age group (18 to 65yrs)	% by age group (over 18yrs)
ACS: Total POP	3,023	--	--	--	21,000	--	--	--	9,450	--	--	--
ACS: 18 to 65yrs	1,606	--	--	--	13,522	--	--	--	7,241	--	--	--
ACS: over18yrs	1,924	--	--	--	17,707	--	--	--	8,088	--	--	--
LEHD	853	100%	53%	44%	4,724	100%	35%	27%	2,674	100%	37%	33%
CTPP	979	115%	61%	51%	6,324	134%	47%	36%	2,983	112%	41%	37%
ASageHW	191	22%	12%	10%	2,189	46%	16%	12%	4,216	158%	58%	52%
PIQ_Day	1,900	223%	118%	99%	3,333	71%	25%	19%	4,956	185%	68%	61%

Data Sources:

1. *CTPP 2006-2010 Part 3: TAZ level flow (A302100 - Total Workers (1) (Workers 16 years and over)*
2. *LODES 2012 data for all jobs (NY_OD_MAIN_JT00_2012)*
3. *Airsage: based on the average of one month (Feburary, 2015) data*
4. *PlaceIQ: based on the average of five months (April to August, 2015) data*

Solutions for a better world