**Abstract**

The increasing availability of longitudinal Consumer Location Datasets (CLDs) presents substantial opportunities for the demographic analysis of residential mobility and migration. Despite their growing popularity, these datasets – produced by private companies for sale to commercial interests – are not explicitly designed for research and have not been assessed systematically in terms of data quality and representativeness. This paper carries out a comparative analysis of consumer location datasets produced by two different private companies – Data Axle and Infutor – in King County, Washington. Comparing these datasets with each other and with census data, we identify a number of strengths and weaknesses associated with each dataset. We conclude that despite notable limitations, CLDs have the potential to provide valuable and novel insights into heretofore unobserved patterns of residential mobility at a range of spatial scales.

**Introduction**

The era of “big data” has presented exciting opportunities for novel approaches to demographic research (Bohon, 2018). However, big data also presents a number of challenges: compared with traditional data sources such as the U.S. Census and the Panel Study of Income Dynamics, non-traditional data sources such as social media data are often created by private companies for commercial purposes (CITE). Research involving residential mobility and migration has increasingly employed a new form of big data: Consumer Location Datasets (CLDs). These datasets are produced by synthesizing an array of public and private datasets including tax assessment records, utility bills, and a variety of other sources, and are specifically designed for sale to commercial interests such as advertisers. Despite their increasing popularity as a research tool, the representativeness of these datasets has not been comparatively assessed. This paper seeks to address this gap in the literature through an analysis of two consumer location datasets that have been increasingly being employed in various academic research contexts: one created by Infutor, and the other by Data Axle (formerly Infogroup). In comparing the two datasets along with census data, we explore the following questions. First, to what extent do each of these datasets under- or over-represent particular social groups, as defined by race, socioeconomic status, tenure, or marital status? Second, are there any systematic biases in representation between these two datasets across different neighborhoods? Finally, how do residential mobility trends compare between the two datasets?

CLDs have several potential advantages over other possible data sources: on one hand, they provide more detailed and comprehensive household information than is generally available in traditional data sources; on the other hand, they are more deliberately curated than big “exhaust” data and therefore are more likely to relatively comprehensive in scope. Several recent studies have utilized consumer location datasets to analyze patterns of residential mobility. The Data Axle CLD has been used to measure neighborhood origin-destination flows (Greenlee, 2019) and to assess the impacts of nearby housing development on possible displacement (Chapple et al., 2022). Meanwhile, residential location histories provided by Infutor have been deployed to measure the effects of new housing construction on possible residential displacement (Asquith et al., 2019; Mast, 2019; Pennington, 2021) and the impacts of rent control policies on residential mobility (Diamond et al., 2019). Phillips (2020) offers, to date, the primary systematic validation of the Infutor dataset, showing that measures of residential instability derived from Infutor align with known instances of instability derived from other sources. However, no such validation has been conducted for the other primary CLD ­­– Data Axle – and the relative strengths and weaknesses of the two datasets have not been systematically compared. Furthermore, the validity of the household-level demographic characteristics contained within these datasets – including race, income, tenure, and marital status – requires further consideration.

* Provide examples of them being used – like UDP (EDR, Karen’s work, see if Crowder and team have done anything).
* Given the newness of these datasets you’ll have to try and provide as much detail about these companies, how they collect data, and what they claim is good or bad about them, maybe how they differ.

**Data Description**

Each of the datasets examined here take a distinct approach to storing information, with important implications for the longitudinal analysis of residential mobility. The Infutor dataset consists of two types of files: a “demographic” dataset that provide an array of household characteristics within a given year, and a “consumer history” dataset providing information on the last *n* moves that a household has experienced (up to 10) and the date (month and year) at which the household was first observed in each location. For this analysis, we merge the demographic data files from 2014 and 2021 with the consumer history file from 2021 to create a single comprehensive dataset. By contrast, Data Axle structures its consumer location data as a panel, with one record and one location for each household in each year; for this analysis, we examine annual Data Axle records for each year from 2006 through 2019. Households in the Data Axle records are not necessarily available for each year; a household with a given unique identifier may appear after 2006 or disappear before 2019, and households may appear and disappear sporadically.

In order to make these datasets comparable, we convert the Infutor dataset into a panel, assuming that an individual remains at the same location from the date at which that household was first observed until the date at which the household is first observed at a new location. We further convert the Data Axle dataset from a household-level panel to an person-level panel, attributing household-level characteristics to each individual within an identified household while retaining person-level name and race/ethnicity attributes. Finally, we subset to observations collected within a single geography. With a total population over 2 million, King County, Washington, is one of the largest counties in the United States and features a variety of urban, suburban, and rural geographies that make it an effective case study for the assessment of data quality.

**Table 1: Original Structure of Example Household Record in Infutor CLD**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Person ID | Location 1 | Location 1 Start Date | Location 2 | Location 2 Start Date | Location 3 | Location 3 Start Date |
| 001 | 111 A St | 01/2015 | 123 Main St | 06/2017 | 101 1st Ave | 09/2019 |
| 002 | 123 Main St | 06/2017 | 101 1st Ave | 09/2019 |  |  |
| 003 | 123 Main St | 06/2017 |  |  |  |  |

**Table 2: Original Structure of Example Household Record in Data Axle CLD**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Household ID | Year | Location | Individual 1 ID | Individual 2 ID | Individual 3 ID |
| 001 | 2015 | 111 A St | 001 | 002 |  |
| 001 | 2017 | 123 Main St | 001 | 002 | 003 |
| 001 | 2018 | 123 Main St | 001 | 002 | 003 |
| 001 | 2019 | 101 1st Ave | 001 | 002 |  |

**Analysis**

* Which types of households are over-/under-represented in the data? [Descriptive statistics based on race, income, tenure, marital status]

In order to compare the strengths and weaknesses of these datasets, [Infutor comparison \_\_\_\_\_.] Meanwhile, we find that while Data Axle undercounts the total population while significantly *over*counting the total number of households, which implies that Data Axle may relegate individuals into separate households.

**Table 3:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | 2019 1-year ACS | Infutor | Data Axle |
| **Population** | 2,252,782 |  | 1,953,613 |
| Non-Hispanic White | 1,302,544  (57.8%) |  | 1,401,964  (81.3%) |
| Black | 147,822  (6.6%) |  | 35,320  (1.8%) |
| Hispanic/Latinx | 222,642  (9.9%) |  | 116,626  (6%) |
| Asian | 424,590  (18.8%) |  | 247,748  (2.3%) |
| **Households** | 907,761 |  | 1,219,753 |
| Non-Hispanic White | 588,849 |  | 860,119 |
| Black | 55,233 |  | 24,018 |
| Hispanic/Latinx | 63,419 |  | 78,302 |
| Asian | 156,438 |  | 159,716 |
| Owner-Occupied | 510,957 |  | 702,671 |
| Renter-Occupied | 396,804 |  | 366,061 |
| Married Couples | 430,014 |  | 489,270 |

* Which types of neighborhoods are over-/under-represented in the data? [Linear model with neighborhood characteristics relative to population counts?]

Chart, scatter chart

Description automatically generated

Map

Description automatically generated

* How do the mobility patterns compare between the two datasets? [Comparison of in-out ratios and destination characteristics]

|  |  |  |  |
| --- | --- | --- | --- |
| Previous Location | ACS 1-Year 2019 PUMS | Data Axle | Infutor |
| No Move | 1,832,669  (82.3%) | 988,561  (93.8%) |  |
| Within King County, WA | 245,858  (11.1%) | 46,363  (4.4%) |  |
| Snohomish County, WA | 13,835  (0.6%) | 2,468  (0.2%) |  |
| Pierce County, WA | 11,913  (0.5%) | 1,870  (0.2%) |  |
| Los Angeles County, CA | 5,640  (0.2%) | 635  (0.05%) |  |
| No Previous Observations | -- | 165,842 |  |