DATA FUSION
Advanced Survey Sampling & Market Research

With the explosion of information and digital technologies, now more than ever before, survey and market researchers stand to benefit from an unprecedented wealth of ancillary data that can be appended to residential addresses. The available sources that make such data fusions possible include a myriad of publicly available statistics from the Census and other government entities, as well as a growing number of commercial databases that are licensable from private data compilers.

Addition of such data not only provides more nuanced targeting and stratification options for survey sampling applications, but also furnish an assortment of covariates that can enrich inferential capabilities of surveys. In particular, with the declining rates of response, ancillary data at the frame level make it possible to identify and apply effective adjustments for differential nonresponse to improve the external validity of survey estimates.

On the other hand, fusion of ancillary data can offer huge possibilities for market research applications. Oftentimes, the limited information companies have about their customers are too scant to support important business decisions. Enhancing customer databases offers rich empirical foundations from which multivariate techniques can extract actionable intelligence to improve profitability – way beyond what basic descriptive statistics could otherwise produce from meager customer data.

Fusion of Publicly Available Data requires bridging the divide between the USPS and Census, as these two entities operate based on completely different geographic delineations. While the Delivery Sequence File (DSF) of the USPS is ZIP-based, the smallest geographic unit for which the latter reports are the so-called Census Block Groups (CBG). By geocoding every delivery point in the US to a unique coordinate, however, MSG makes it possible to marry the DSF with the plethora of estimates that the Census provides at various levels of aggregation. The following table provides a catalog of the nearly 150 million delivery points in the latest DSF.
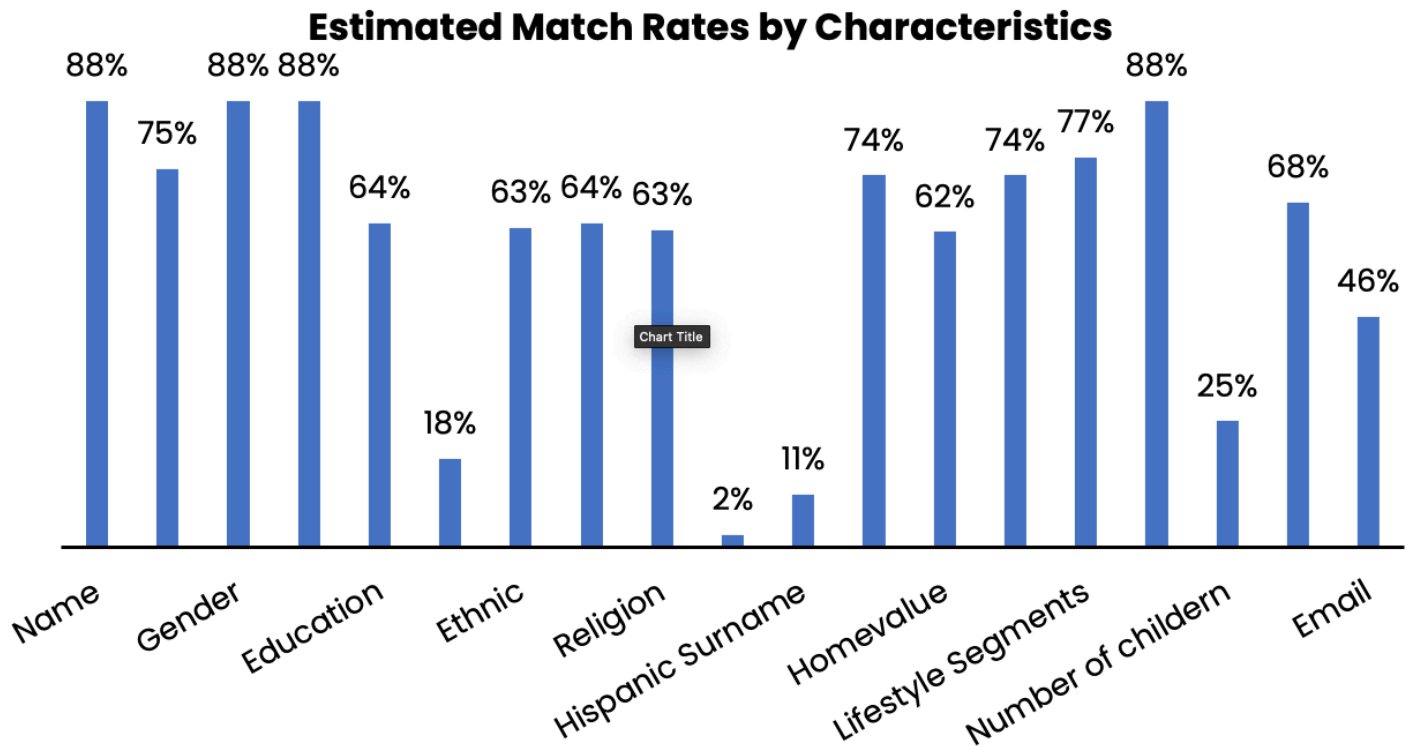
MSG-Enhanced ABS Frame

Address Type	Delivery Units
Total Address	**149,249,675**
City Style	133,491,398
Drop Units (From 689,532 Drop Points)	1,903,690
Drop Unit Augments (MSG)	39,842
Seasonal	782,477
Educational	99,962
Vacant	3,044,307
Throwback	186,522
Augments (MSG)	71,078
Rural Route/Highway Contract	38,813
PO Box	15,719,464
Traditional	14,299,960
Only Way of Getting Mail (OWGM)	1,418,552

It should be noted that data available from the Census pertain to households that can be appended only at the block level and beyond, since by law the Census cannot provide any data at the person or household levels. The main sources from the Census include those from the public-release data from the American Community Survey (ACS) and the various monthly supplements of the Current Population Survey (CPS). On custom basis, it is possible to append estimates from what is available from the CDC and other government surveys as well.

Fusion of Commercially Available Data makes it possible to go beyond what is publicly available and append more granular data to individual households. With the rapidly evolving data streams from many commercial data compilers, it is now possible to license an assortment of databases at both household and individual levels. Among

others, key commercial data sources include Acxiom, Experian, Neustar, Data Axel, Aristotle, US Data, and Claritas. Depending on the source and type of data, as summarized in the following chart, appended data can be subject to different match rates.

Estimated Match Rates by Characteristics

Characteristic	Match Rate
Name	88%
Gender	75%
Education	88%
(unlabeled)	88%
Ethnic	64%
(unlabeled)	18%
Religion	63%
(unlabeled)	64%
Hispanic Surname	63%
(unlabeled)	2%
Homevalue	11%
Lifestyle Segments	74%
(unlabeled)	62%
Number of childern	74%
(unlabeled)	77%
Email	88%
(unlabeled)	25%
(unlabeled)	68%
(unlabeled)	46%

It is worth noting that appended data to addresses can come from multiple sources, since different sources compile different data types at different levels of aggregation. Some examples include:

1. **Commercial Data at the CBG Level:**

-   Total Population
-   Households, Housing Units, Families, and Group Quarters
-   Percent Change from 2010 to Current Year
-   Population by Gender, Age, and Race–Ethnicity
-   Population by Ancestry
-   Hispanic Population by Country of Origin
-   Population by Language Spoken at Home
-   Group Quarters Population by Type
-   Occupied Housing Units by Tenure and Age
-   Households by Ethnicity and Race
-   Homeownership Status by Unit Value

- Households by Number of People in Household
- Households by Presence of People Under 18
- Housing Units by Year Structure Built
- Housing Units by Units in Structure
- Occupied Housing Units by Year Moved In
- Occupied Housing Units by Vehicles Available
- Household Income
- Householder Age by Income
- Households by Race and Income
- Households by Ethnicity and Income
- Population by Ethnicity, Gender, and Education
- Population by Employment Status
- Workers by Means of Transportation to Work
- Workers by Travel Time to Work
- Employed Civilian Population by Class of Worker
- Employed Civilian Population by Occupation

2. **Claritas Lifestyle Segment**s are derived from a long list of geodemographic and behavioral attributes for each household. As such, all US households are modeled to belong to one of the 68 PRIZM segments that can range from Upper Crust to Movers & Shakers to Young & Influential. These market segments allow targeting of consumers based on a battery of attitudinal and behavioral characteristics, including their shopping habits, financial metrics, and media usage.

3. **Neustar Market Segments** are comprised of 172 groupings of the US households based on many available and modeled data for each household. These segments are partitioned into specific subsets that focus on households expected to be affluent, high income, middle income, moderate income, or low income.

4. **Miscellaneous Ancillary Data** include what could be appended from a variety of sources. Such data can span a wide range, including political affiliation, voter registration, specific product purchases and service subscriptions, as well as various contact information. For example, MSG can identify most of the "pay-as-you-go" cellular numbers with listed addresses. These numbers are purported to have a higher chance of reaching individuals who are more transient, lower income, and of minority groups with modest levels of acculturation.