Applied Geography 97 (2018) 152–160

Contents lists available at ScienceDirect

Applied Geography

journal homepage: www.elsevier.com/locate/apgeog

Evaluation of commercial demographic data accuracy: Testing for input bias in the location analysis process

William Graves[a,*], Brian Gerney[b]

[a] Department of Geography and Earth Sciences, UNC Charlotte, USA
[b] Gerney Research Group, USA

ABSTRACT

Business geographers devote considerable effort to develop models of space to answer complex spatial questions. Unfortunately, little thought is given to the accuracy of the input data in these models. This paper examines the accuracy of tract-level population and income data reported for the second quarter of 2015 from five spatial-demographic data vendors commonly used in the corporate location-decision process (Experian, Synergos/PopStats, ScanUS, ESRI and EASI). Data are compiled for 80 Census tracts in the 40 fastest growing US metropolitan areas. Little agreement between data vendors was found – mean absolute percent errors (MAPE) for 2015 population and income estimates were 8.97% and 11.98% respectively. Error rates increase substantially in urban tracts, low income tracts, and high-density tracts. These accuracy biases are likely to contribute to the under provision of retail in urban settings. Strategies for minimizing the impact of this error and for improving commercial demographic data quality are also discussed.

The question that is fundamental to any applied geographer's job is: "how reliable are my results?" This question becomes particularly difficult to answer for those working with demographic data in the latter half of each decade as the enumerations from the previous decennial Census grow stale. Regardless of the ability of the Census Bureau's ability to update its spatial demographic products, the business community's demands for "real time" small-area demographic data is growing. This need for currency in commercial spatial analysis forces the retail, insurance and home-building industries to rely on estimates of key demographic metrics produced by an assortment of private data vendors (such as Experian, ESRI or Synergos) for their site selection process. Unfortunately, little is known about how these vendors derive their proprietary estimates and evaluation of data accuracy by consumers is difficult due to the cost, a lack of methodological transparency and the reluctance of vendors to provide historical data. Thus, site selection professionals are forced to take the accuracy of their spatial demographic data inputs for granted – a situation that prevents the site selection process from being scientific thus eroding the quality and credibility of applied spatial modeling (Berry, 2002; Pickles, 1995; Weinberger, Dock, Cohen, Rogers, & Henson, 2015). The significance of this issue is not limited to the corporate world. Data from these vendors has been used for scholarly research into telecommunications access (Rappoport, Alleman, & Taylor, 2007), measuring childhood poverty (Vash, 2015), measuring community health disparities (Kantor & Constantin, 2017; Mehrotra, Story, Guest, & Fedunyszyn, 2012), public health (Duncan et al., 2014) and foreclosures and community health (Katz, Wallace, & Hedberg, 2013) to name a few.

This paper expands the single-city case study of demographic data biases by Graves and Gerney (2016) by examining variations in 2015 population and median household income estimates from five commercial data vendors across 40 US metropolitan areas. Data from the following vendors were examined:

- ESRI Business Analytics
- Synergos Technologies (formerly Popstats)
- EASI Analytics
- Scan/US Professional Market Mapping
- Experian Demographic

These data are updated either quarterly or annually and are derived by modeling demographic change - typically the application of a proprietary algorithm to base-year figures from the Census. Consumers typically receive the data as a package with analytical software (e.g. ESRI Business Analyst, Trade Area Systems or Alteryx) or accessed via a subscription model (e.g. Synergos, Scan/US or EASI Analytic). These 'real time' data are costly-subscriptions typically range between $2000 and $60,000 per year depending on specific user needs.

Our data comparison is conducted for 2015 population and median

* Corresponding author.
E-mail address: bgraves@uncc.edu (W. Graves).

https://doi.org/10.1016/j.apgeog.2018.06.009
Received 15 June 2017; Received in revised form 4 June 2018; Accepted 22 June 2018

household income statistics for 80 census tracts selected from the 40 largest metro areas. Our analysis is conducted from the perspective of a retail industry data user; our sample is selected from the set of tracts consider to be the most attractive to contemporary retail development. Half of our 80 tracts were selected from predominantly single-family suburban areas while the other half were revitalizing center city tracts. Our decision not to include low-income or non-urban tracts was driven largely by the location preferences of the multi-establishment retail industry. The American Community Survey (ACS) 5-year estimates were used as a baseline for evaluating the consistency of the vendor data. A Mean Absolute Percent Error (MAPE) figure for each tract, data vendor, and sub-category of tract (urban and suburban) was calculated to measure the degree of deviation from the ACS[i] with the goal of measuring the degree of variation between each vendor's estimates for each tract. We also seek to identify any spatial biases (relative to the ACS) which may exist in the vendors' estimation techniques. Our analysis determined that the vendor data in our sample deviates from ACS estimates by as much as 40%. In addition, deviations from ACS data were smaller for suburban tracts and larger for urban ones. We believe our finding of this suburban accuracy[ii] bias may have important implications for the adequate provision of retail in urban communities as well as present challenges for effective and efficient retail location decision making.

1. Corporate use of vendor data

An informal email survey of corporate spatial demographic data consumers was conducted in May and June 2016. Survey subjects were either employees of the location decision offices of large retailers or active consultants in the industry. Survey participants were asked what data vendor they currently used as well as their opinion of the accuracy of the data. Frequency of use results are shown in Table 1. Based on this survey, no single data provider dominates the industry, although the three most commonly used providers account for nearly 75% of the market. Users most frequently attributed their data preference to the vendor's ability to integrate their demographic updates within the corporate IT structure as well as a history with the specific product – sufficient data accuracy was assumed but was only evaluated by 25% of users. When specified, the evaluation consisted of either comparing new vendor data with current data or baselining new vendor demographics against store performance data. Notably, none of the firms reported relying on any type of Census data due to its perceived staleness as well as the difficulty of integrating it into their spatial decision support software.

Promises of anonymity prevent revealing the specific preferences of individual users; however, several industry preferences were identified. Grocers and discount retailers most commonly used Synergos data while quick service restaurants most commonly used ESRI data. While we did not survey users in the insurance or actuarial industries, some trade publications suggest those firms frequently use EASI data.

Approximately half of respondents noted problems with bias in their demographic data. Problems with inaccuracy were noted in areas that had undergone significant disruption (e.g. South Florida following the sub-prime crises or New Orleans following Hurricane Katrina), areas of high growth (14 mentions), areas of low growth (6 mentions), high-density areas (5 mentions), low income areas (2 mentions) and low-density (e.g. rural) areas (1 mention). There was a general sense among respondents that the vendor data are most accurate in stable, middle-class, single-family neighborhoods. Despite concerns, users were generally comfortable with their data providers and believed that accuracy was sufficient for their purposes. Only three of 72 respondents mentioned a past change of vendors for accuracy reasons. Our primary takeaway from the survey results was that corporate data users understood that their data contained spatial biases, but there were few efforts to measure or correct for these inconsistencies. In general, users felt that vendor data contained a tolerable level of error and bias but accepted their vendor's demographics as the best available.

Table 1
Demographic data packages used by major corporation for location analysis. Survey conducted by author in May and June 2016.

Data Vendor	Industrial Users	% of Total
Synergos/Popstats	24	33.33%
ESRI	16	22.22%
Neilson/Claritas	13	18.06%
Experian	7	9.72%
EASI	6	8.33%
Scan/US	4	5.56%
AGS	2	2.78%
Total	72	100.00%

2. Vendor data estimation models – the ACS as a foundational input

Corporate data users know little about the accuracy of the demographic data they purchase because the estimation models are proprietary, information about the methodologies used to produce the estimates is scarce, and the cost of conducting large scale comparisons of vendor data is prohibitive. We examined the methodological material that is publicly available from each data vendor and identified some common estimation techniques. The most significant methodological tool of the vendors is to derive their estimations from the most recent ACS five year estimates (Experian, 2012; ESRI, 2014; Synergos, 2016; Scan/US, 2016). While the ACS is considered by many scholars to be the highest quality demographic information available (Meyer, Mok, & Sullivan, 2015; Spielman, Folch, & Nagle, 2014), it is important to keep in mind that the ACS is created through survey samples and, as such, is significantly less accurate than the enumerations of the decennial Census. Sample sizes are quite small for individual tracts. Spielman and Folch (2015) estimate that each tract has an average of 32 respondents. These sample size limitations can produce substantial errors at the sub-county scale. The Census strategy for minimizing this error is to pool samples over multiple years (thus the ACS five-year estimates). While this strategy is somewhat effective at reducing statistical uncertainty, it substantially reduces its temporal resolution – five year estimates which are released a year after the most recent survey are not contemporary enough to satisfy most corporate data users. Despite the sample pooling, ACS error rates are so large that Spielman et al. (2014) explicitly question the usability of ACS data for small area (e.g. below the county level) analysis. Unfortunately, there is a documented tendency for users to dismiss ACS reliability problems (Francis, Tontisirin, Anantsuksomsri, Vink, & Zhong, 2015).

The accuracy of ACS data has been widely explored by spatial scientists. Using a case study of Census tracts in Nashville, Bazuin and Fraser (2013) compared ACS five year estimates to their on-the-ground experience with the area. They detected substantial errors in 5 year ACS estimates compared to their personal enumerations of the high-poverty tracts – their enumerations were as much as 60% higher than ACS five year estimates in their case study area. They attributed the error rates to an insufficient sample size as well as an overreliance on telephone

[i] The ACS was used as an arbitrary baseline for comparison due to its survey-based (rather than model-based) methodology. This methodology makes it a comparison which is independent of the estimation techniques used by the data vendors. While the five-year average of survey results used to derive ACS values deviates from the single-year perspective of data vendors we believed the independence of the ACS process made it the most useful baseline option. We are not suggesting that the ACS data represent an ideal of accuracy or reliability.

[ii] We use the term accuracy to refer to low MAPE values. Low MAPE values indicate that the vendor estimates are in close agreement with the ACS estimate. Since scholars are unwilling to assume that ACS figures represent a paragon of accuracy it may be best to think of our MAPE values as a measure of estimation consistency between vendors.