Journal of Big Data

CASE STUDY Open Access

Using electronic transaction data to add geographic granularity to official estimates of retail sales

Brian Dumbacher*, Darcy Steeg Morris and Carma Hogue

*Correspondence:
brian.dumbacher@census.gov
U.S. Census Bureau, 4600 Silver Hill Road, Washington, DC 20233, USA

Abstract

Introduction: Economists are interested in more granular, more frequent data to aid in their understanding of the U.S. economy. The most frequent economic data currently available from the U.S. Census Bureau come from monthly economic indicators such as the Monthly Retail Trade Survey, which produces national estimates of retail sales. On the other hand, the most granular data (in terms of geographic and industry detail) come from the Economic Census, which is conducted every five years. The Census Bureau is researching whether organic, third-party Big Data sources, in conjunction with survey data, allow for the production of retail sales estimates that are both monthly and subnational.

Case description: This case study explores the feasibility of using aggregated electronic transaction data from First Data (FD), a large payment processor, to calculate experimental regional and state-level monthly estimates of retail sales. Quality criteria are devised to understand this data source's representativeness of the target population and consistency with existing survey data. Five retail industries in the FD transaction data are identified as having acceptable quality for estimation. Estimation methodology is developed based on linear mixed models in a Bayesian framework. These models try to take advantage of the timeliness of the FD transaction data and smooth over artifacts of FD's business activity. Experimental estimates of retail sales are calculated for the period January 2015 through March 2018.

Discussion and evaluation: The experimental estimates are evaluated quantitatively via correlations between external estimates of the number of employees by industry and qualitatively with respect to additional information about the economy. Many features of the experimental estimates seem reasonable, but there are also caution flags such as anomalous trends related to identified FD quality issues.

Conclusions: The FD transaction data offer insight into economic activity at a more granular level. However, using this data source to enhance official estimates of retail sales is challenging; the FD aggregates have limitations in terms of suppression, coverage, and trends. Consequently, fewer industries than expected are identified as having acceptable quality for estimation. Future work involves calculating experimental estimates for more recent months and researching alternative methods for evaluating their accuracy.

Keywords: Bayesian methods, Electronic transactions, Model-based estimation, Monthly Retail Trade Survey, Official statistics, Organic data, Third-party data, U.S. Census Bureau

Introduction

The U.S. Census Bureau produces high-quality official economic statistics using traditional sample surveys and censuses. Data collection for each survey and census program is designed to produce reliable statistics of economic output for a specific time period and level of geography. For example, the Monthly Retail Trade Survey (MRTS) is designed to estimate national retail sales and inventories on a monthly basis from a representative sample of about 13,000 companies in the retail and food services sector [25]. The Economic Census (EC) is conducted every five years to produce subnational estimates for a wealth of economic variables, including retail sales, for years ending in "2" and "7" [26]. There is growing demand, however, for timelier and more geographically granular data products [20]. At the same time, respondent cooperation is declining, and the costs of conducting traditional sample surveys are increasing. This makes it increasingly challenging to meet the needs of the Census Bureau's economic data users in businesses, academic institutions, and other government agencies.

Third-party data—defined in this paper as data collected by a nongovernment entity in the course of providing a service or product—could help address these needs. The Census Bureau envisions leveraging third-party data sources in conjunction with existing survey data in various ways: providing timelier data products, improving efficiency and quality throughout the survey life cycle, and offering greater insight into the nation's economy through detailed geographic estimates [1, 15]. Research so far has focused on examining the potential use of third-party data sources to enhance retail programs, in particular MRTS [5]. Third-party electronic payment data from credit card companies and point-of-sale processors, for example, provide a measure of the retail sales economy at a more granular level. However, they suffer from classic Big Data complications [19]: they can be large in volume and lack veracity [two of the four characteristics often referred to as the four V's of Big Data [24]]. Understanding the uncertain and imprecise nature of these data—especially with respect to representativeness—is critically important when considering their use with official statistics. In fact, uncertain veracity is a primary characteristic of what is termed secondary, found, organic, or nonprobability sample data in survey research [2, 13, 14, 16, 17]. Third-party electronic payment data are collected for purposes not related to producing official statistics yet contain relevant information for measuring the retail trade economy.

Statistical modeling can integrate third-party electronic payment and records data with survey data to produce frequent and geographically granular estimates of economic activity. For example, there is a growing literature on using electronic data in nowcasting models to make economic forecasts for the recent past, present, or near future. Galbraith and Tkacz [11] use monthly aggregates of the value and number of debit and check transactions to nowcast change in the Canadian gross domestic product and in retail sales. In a similar vein, D'Amuri and Marcucci [3] use Google job-search activity to nowcast the monthly unemployment rate in the United States. Marchetti et al. [18] use Global Positioning System car journey data as a covariate in a Fay-Herriot small area

estimation model [7, 22] to estimate poverty rates for local areas in Tuscany, Italy. Using a similar small area estimation approach, Porter et al. [21] utilize Google search activity involving common Spanish words to estimate the relative change in rates of percent household Spanish-speaking for states in the eastern half of the United States. This literature presents a variety of statistical modeling techniques for using aggregates from large volume, high velocity third-party data to enhance measures of (socio-)economic output.

This case study involves using electronic transaction data from First Data (FD) to complement existing MRTS survey data in order to provide more geographically granular estimates of retail sales. FD is a large, global payment processor that processes about 72 billion credit, debit, gift, and prepaid card payment transactions per year in the United States and Canada [9, 10]. The Census Bureau receives monthly datasets containing aggregates of FD transaction value, or spend, and the corresponding number of FD merchants broken down by geography, industry, and month. The research question is whether these aggregates, together with statistical models and publicly available auxiliary data from the EC and other sources, can reasonably produce regional and state-level retail sales estimates. These estimates would offer finer geographic granularity than retail sales estimates currently produced from MRTS. The complexity in this case study is in assessing the quality of the FD transaction data. The volume of the aggregate data allows the use of standard statistical computing techniques. Indeed, the computing environment consists of a Linux server with SAS/STAT® and R software. No distributed framework is required. However, evaluating the veracity of the data requires careful exploratory analysis and thought.

The rest of this paper is organized as follows. The "First Data" section describes FD itself and the electronic transaction data used in this case study. The subsequent "Case description" section is divided into two parts: "Quality criteria" describes an industry-by-industry quality evaluation of the FD aggregates, and "Estimation methodology" details how experimental regional and state-level estimates of retail sales can be calculated. The fitted models and estimates are then presented in "Results". The "Discussion and evaluation" section describes an evaluation of the estimates, and "Conclusions" summarizes findings and outlines future work.

First Data

As an organic, third-party Big Data source, the FD transaction data may reflect business activity more than the true economic activity that the Census Bureau is trying to measure. Therefore, it is important to understand some features of FD's operations and payment processing in general. This section covers this information and describes the FD transaction data in more detail.

Payment processors such as FD link merchants to payment networks including Visa, MasterCard, and American Express. They serve as intermediaries between a customer swiping a card-form of payment and the card's financial institution [4]. FD captures electronic transactions from credit, debit, gift, and prepaid cards but not cash transactions such as checks and direct transfers. The FD transaction data represent a very large and rich data source. FD processes approximately 72 billion commercial transactions per year in the United States and Canada [9]. Considering just

the United States, FD processes approximately $1.9 trillion of card payments per year, or about ten percent of U.S. gross domestic product [9].

The FD transaction data reside in various business infrastructure units known as platforms, which may be skewed to particular geographies and industries. Some platforms exist as a result of FD having acquired other payment processors. Data from different platforms are incorporated into the FD aggregates at different times. Merchants join or leave FD often enough that there is appreciable birth and death activity, also known as merchant churn. To optimize costs, merchants switch to other payment processors temporarily or allocate their transactions across multiple payment processors. Merchants may also submit transactions to FD for processing in unusual ways to avoid paying inactivity or early termination fees. All of these FD business characteristics affect how well FD represents all retailers at different points in time and in different geographies.

FD classifies merchants by Merchant Category Code (MCC), which describes the merchant's type of business and is based on the main goods and services that it provides. Before the transaction data are aggregated and provided to the Census Bureau, the MCCs are mapped to North American Industry Classification System (NAICS) codes. NAICS is a similar classification system based on a hierarchical coding structure [27]. A NAICS code consists of six digits. The first two digits indicate industry sector (for example, retail), and subsequent nonzero digits add industry detail. The MCC-to-NAICS mapping is generally known to have issues such as the handling of e-commerce transactions, but the specifics of the problem have not been identified and therefore cannot be corrected.

FD has a large client base in the United States, but it is important to emphasize that the FD aggregates represent a nonprobability sample of merchants. Because of FD's confidentiality procedures, the Census Bureau knows neither who FD's clients are nor which merchants comprise the aggregates. Furthermore, a substantial proportion of merchants have opted not to have their transactions be used to calculate the aggregates. Suppression rules are also applied to the aggregates before the Census Bureau receives them. These rules are based on the number of merchants and the merchants' share of spend for each combination of geography, industry, and month. Altogether, it is challenging to assess how representative the FD transaction data are of the entire payment processing industry and of the target population—all retailers are in scope to MRTS.

The Census Bureau receives monthly datasets containing aggregates of FD spend and the corresponding number of FD merchants. Spend is the key variable, but the number of merchants can inform the coverage of FD spend. For example, as is done in calculating the experimental estimates, the number of merchants can be compared to the Census Bureau's estimates of the number of business establishments. The aggregates are broken down by geography [the United States, Census Bureau-defined region, or state (including the District of Columbia)], industry (as classified by NAICS), and month. The aggregates represent 56 geographies and 378 industries, of which 102 are in the retail and food services sector. Therefore, there are 5712 data series, or combinations of geography and industry, that are in scope to MRTS. This case study is based on monthly FD aggregates of spend and number of merchants from January 2012 through March 2018.

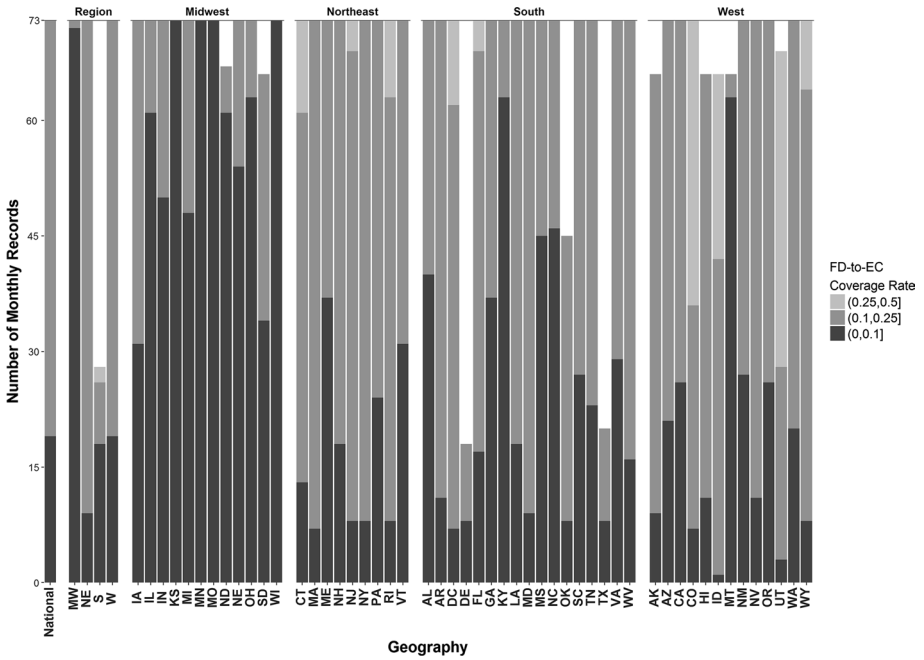

Fig. 1 First Data suppression rates and First Data-to-2012 Economic Census (FD-to-EC) coverage rates across months by geography for NAICS 447 (gasoline stations)

Case description

This section is divided into two parts. First, quality criteria are devised to evaluate the fitness of the FD aggregates for use in estimation and are applied by industry at the 3-digit NAICS level. The second part details the estimation methodology. This methodology is used to produce regional and state-level monthly estimates of retail sales for the industries identified as having acceptable quality.

Quality criteria

The FD aggregates have classic Big Data and nonprobability sample limitations such as uncertain representativeness of the target population. To address these limitations, modeled estimates are produced only for select industries that exhibit acceptable consistency in representativeness, reliability, and tracking to official statistics over geographies and time. A direct measure of representativeness and reliability is not available; thus industry selection relies on exploratory quality criteria metrics to inform a broad subjective quality profile. This screening step is crucial for understanding data quality at industry, geography, and time granularity based on historic data. It is important to recognize that industry quality profiles can change at any time as FD business activity is dynamic.

Consistency in data availability and representativeness

Data availability is assessed via suppression rates. The FD aggregates are received with missing values dictated by suppression rules applied to each geography and industry data series. A series with fewer suppressed values is naturally desirable to achieve better assessment of data quality and better model fit with a fuller set of data. As an example, Fig. 1 displays a visualization of suppression rates—and coverage rates—by region

and state from July 2012 through July 2018 (73 months) for NAICS 447 (Gasoline Stations). The height of the bar indicates the number of months with unsuppressed values; conversely, the white space at the top of the bar indicates the number of months with suppressed values. About 60% of FD aggregate spend values are suppressed for the South region for NAICS 447. At the state level, about 70% of FD aggregate sales values are suppressed for Texas and Delaware. As a result, for the majority of months during this period, there are estimation limitations for the South, Texas, and Delaware. Furthermore, information from these geographies is not available to inform modeled estimates for other geographies.

Data representativeness is assessed through coverage rates and trends comparing (1) FD and 2012 Economic Census (FD-to-EC), (2) FD and MRTS (FD-to-MRTS), and (3) MRTS certainty and noncertainty unit aggregate values. Two coverage rates are assessed by industry: (1) FD-to-EC compares coverage at the region and state level with the denominator fixed in time, and (2) FD-to-MRTS compares coverage at the national level over time. Coverage rates are defined as FD aggregate spend divided by the EC sales or MRTS estimate. A series with roughly consistent and substantial coverage rates is desirable to allay concerns about mistaking business activity for economic activity. Both metrics are indirect measures of representativeness based on relevant and available but imperfectly comparable census and survey data. For the FD-to-EC coverage rates, changes in FD representativeness are comingled with the time variation expected by comparing monthly FD data to time-constant 2012 EC data. In contrast, the national aggregation for the FD-to-MRTS coverage rates may mask changes in FD representativeness by naturally smoothing over finer level geographic variation. Both are useful coverage metrics, especially considered jointly, because they utilize relevant and publicly available official estimates at the appropriate geographic granularity (FD-to-EC) or frequency (FD-to-MRTS).

As an example, Fig. 1 provides a visualization of FD-to-EC coverage rates—and previously described suppression rates—by region and state for NAICS 447. The shading of the bar indicates the FD-to-EC coverage discretized into categories < 10%, 10–25%, and 25–50%. Figure 1 shows that for NAICS 447, the Midwest has consistently lower coverage (less than 10%) than the majority of the data points in the Northeast and West (10–25%). At the state level, the coverage is mostly between 10 and 25%, except for states in the Midwest where the coverage is mostly in the lower range of less than 10%. The mean and standard deviation of the national FD-to-MRTS coverage rate is 0.164 (i.e., 16.4%) and 0.041, respectively. Excluding the earlier period when new platforms were introduced in this industry (July 2012 through February 2014), the coverage consistently hovered around 18% (a mean of 0.184 and a standard deviation of 0.007).

As mentioned previously, information about the specific merchants that are included in the FD transaction data is not provided to the Census Bureau. Coverage rates assist in understanding the characteristics of these merchants, specifically with respect to size. However, even though low coverage may indicate exclusion of larger merchants, the effects of the exclusion depend on characteristics of the particular NAICS. Potential biases are explored through aggregated certainty versus noncertainty unit levels and trends of MRTS sales. Certainty units in MRTS are sampling units that are selected with probability one, represent only themselves, and are typically the largest units in terms

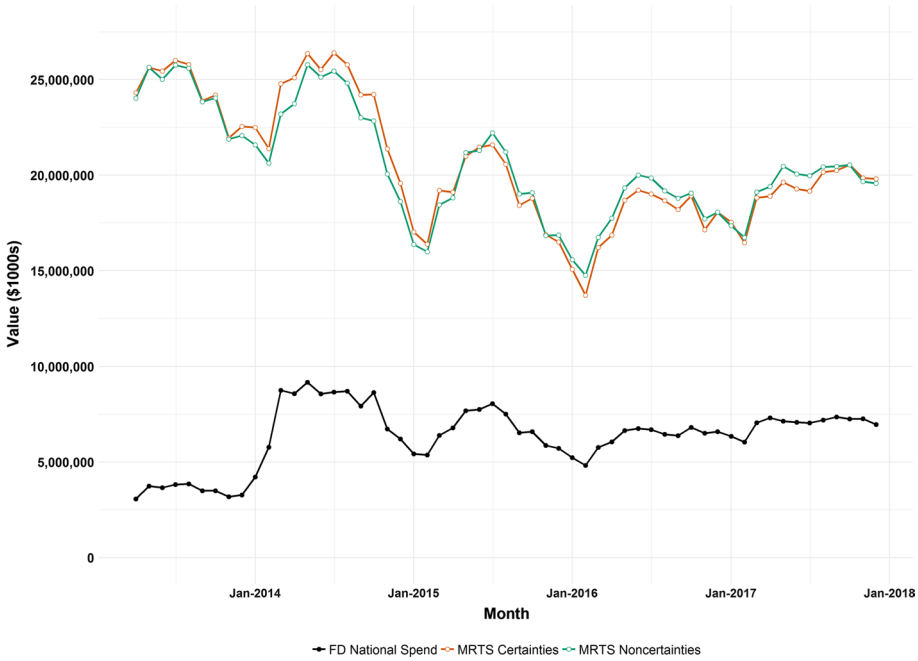

Fig. 2 Monthly Retail Trade Survey (MRTS) certainty and noncertainty sales estimates and First Data (FD) national spend over time for NAICS 447 (gasoline stations)

of sales. The total sales for certainty units in MRTS is utilized as a proxy for the sales of large merchants in the FD data. Noncertainty units in MRTS, on the other hand, are sampling units that are selected with probability less than one and represent other units in addition to themselves through sampling weights. It is reasonable to assume that industries with (1) similarly behaving certainty and noncertainty units and/or (2) a smaller share of certainty unit aggregate spend, are likely to have a smaller bias with respect to the presence/absence of larger merchants in the FD data. As an example, Fig. 2 displays total sales estimates in MRTS for certainty and noncertainty units (with sampling weights applied) compared to FD aggregate values in FD for NAICS 447. The MRTS values are not adjusted for seasonal variation, holiday and trading day differences, or for price changes. MRTS total sales for certainty and noncertainty units exhibit strikingly similar trends (correlation of 0.96) and levels over time, suggesting that exclusions of large merchants from the FD data (i.e., a lower coverage rate) may not be of concern for this industry.

Consistency in spend patterns

Consistency in spend patterns is assessed via comparisons of FD spend across geographies (region-to-region) and FD spend to national MRTS estimates (region-to-MRTS). In addition to graphical assessments, Pearson correlation coefficients are used to summarize association between geographies and data sources restricted to more recent months—January 2015 through March 2018. This restricted date range avoids periods when most of FD's platforms were incorporated into the data. An FD data series that exhibits strong correlation and visibly consistent tracking both

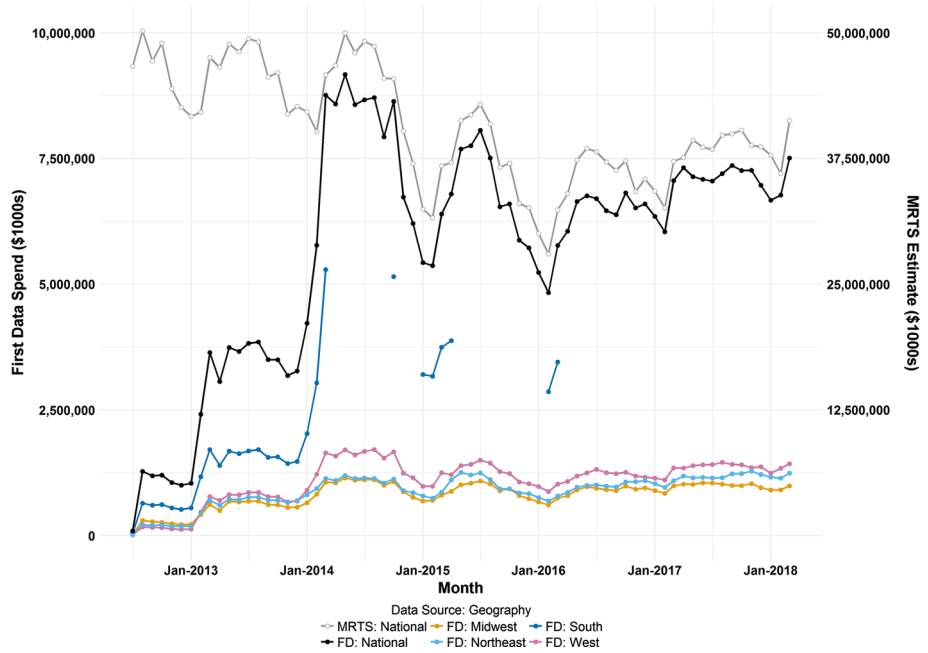

Fig. 3 National Monthly Retail Trade Survey (MRTS) estimates with regional and national First Data (FD) aggregate spend over time for NAICS 447 (gasoline stations)

internally across geographies and externally to MRTS is regarded as desirable. Such consistency allays concerns about mistaking business activity for economic activity, yet geographic differences and departures from MRTS trends are precisely the value of the geographic granularity of the FD transaction data. However, with only indirect measures of representativeness and reliability, a conservative judgment on the trade-off between geographic granularity and capturing true economic activity is deemed appropriate.

As an example, Fig. 3 presents national MRTS estimates as well as national and regional FD aggregate spend for NAICS 447 over time. The difference in the spend levels across regions and data sources is evident, but the inconsistency and suppression in the South series are most striking. The Midwest, Northeast, and West FD series are all strongly positively correlated with MRTS (0.91, 0.85, 0.95, respectively) and with each other (> 0.87 for all pairwise correlations). NAICS 447 illustrates good tracking to official estimates in three regions, but not in the South. As a result, caution is recommended in interpreting estimates relying on regional data that include the South.

Taken together, these indirect quality measures yield a strong quality profile for NAICS 447. This industry exhibits low suppression rates (except in the South), good coverage, similar certainty/noncertainty spend levels, strong positive certainty/noncertainty trend correlation, and strong spend tracking (except in the South). Therefore, experimental regional and state-level estimates are produced for NAICS 447, but with caution due to value suppression in the South.

Table 1 Summary of quality criteria by industry

Industry		Quality criteria				
		Suppression	Coverage (FD-to-EC and FD-to-MRTS)	Trend correlations		
				Certainty-to-noncertainty	FD region-to-region	FD region-to-MRTS
441	Motor vehicle and parts dealers	X		X		
442	Furniture and home furnishings stores	X*				
443	Electronics and appliance stores				X*	X*
444	*Building material and garden equipment and supplies dealers*	*X*	*X*	*X*	*X*	*X*
445	*Food and beverage stores*	*X*	*X*	*X*	*X**	*X**
446	Health and personal care stores	X*				
447	*Gasoline stations*	*X**	*X*	*X*	*X**	*X**
448	*Clothing and clothing accessories stores*	*X*	*X*	*X*	*X**	*X**
451	Sporting goods, hobby, musical instrument, and book stores					
452	General merchandise stores	X*		X		
453	Miscellaneous store retailers	X*				
454	Nonstore retailers			X		
722	*Food services and drinking places*	*X*	*X*	*X*	*X*	*X*

X indicates acceptable quality, and X* indicates acceptable quality with some notable geographic and/or time exceptions. Industries selected for estimation are in italics

Quality evaluation

Quality profiles are assessed for the thirteen 3-digit NAICS-level industries in scope to MRTS. Table 1 provides a summary of the quality assessment for each of the thirteen industries including the findings from NAICS 447 (Gasoline Stations) described in detail previously. As a contrary example, NAICS 454 (Nonstore Retailers) has notably high suppression rates and low coverage (for example, the average FD-to-MRTS coverage rate is 2%) indicating insufficient data availability and representativeness. This finding is likely an artifact of the nature of NAICS 454 containing e-commerce merchants for which there is no matching MCC in the FD data. Evaluation of each of the five objective criteria resulted in sufficient confidence in the FD data for five industries: Building Material and Garden Equipment and Supplies Dealers (NAICS 444), Food and Beverage Stores (NAICS 445), Gasoline Stations (NAICS 447), Clothing and Clothing Accessories Stores (NAICS 448), and Food Services and Drinking Places (NAICS 722).

Estimation methodology

State-level models are developed to calculate estimates of retail sales by state, Census Bureau-defined region, and month for the five industries identified in the previous subsection. The models are cross-sectional linear mixed models [6] that are fit separately

Table 2 Data sources, variables, and their uses in estimation

Data source	Variables	Uses
First Data (FD)	Transaction value, or spend, by geography, industry, and month Number of merchants by geography, industry, and month	Adjusting FD spend for FD merchant coverage Calculating dependent variables
U.S. Census Bureau Monthly Retail Trade Survey (MRTS)	Retail sales for the United States by industry and month	Calculating dependent variables Benchmarking
U.S. Census Bureau 2012 Economic Census (EC)	Retail sales by state and industry for 2012	Accounting for suppressed FD spend Model covariates Validating experimental estimates
U.S. Census Bureau County Business Patterns (CBP)	Number of employer establishments by state and industry for 2015 and 2016	Adjusting FD spend for FD merchant coverage
U.S. Census Bureau Nonemployer Statistics (NES)	Number of nonemployer establishments by state and industry for 2015 and 2016	Adjusting FD spend for FD merchant coverage
U.S. Census Bureau Population Estimates Program (PEP)	Resident population by state and year	Model covariates
Bureau of Economic Analysis State Personal Income (SPI)	Accommodation and food earnings by state and quarter Construction earnings by state and quarter Retail trade earnings by state and quarter	Model covariates
Bureau of Labor Statistics Quarterly Census of Earnings and Wages (QCEW)	Number of employees by state, industry, and month	Validating experimental estimates

by industry and month. There is no time series component. The models attempt to take advantage of the timeliness of the FD transaction data but also utilize covariates from various economic and demographic data sources to smooth the trends in FD spend.

The dependent variable equals the published national MRTS estimate allocated among the states according to FD spend adjusted for merchant coverage. This FD-adjusted MRTS value is modeled as a function of covariates from publicly available official statistics, as well as a geography-level random effect. The covariates include measures of quarterly earnings, resident population, and retail sales produced by federal statistical agencies such as the Census Bureau, the Bureau of Labor Statistics, and the Bureau of Economic Analysis. All values are transformed to the log scale to help satisfy the assumption of homoscedasticity. Bayesian methods are used to fit the models. Lastly, a benchmarking, or raking, procedure ensures the state-level estimates sum to regional estimates and the published national MRTS estimate.

Data sources

In addition to the FD transaction data, publicly available data from multiple Census Bureau programs and other federal statistical agencies are used in various ways to calculate estimates. Table 2 summarizes all of the data sources, variables, and their uses in estimation. Data from MRTS are obtained from the Census Bureau's monthly retail trade website [25]. Data from the 2012 EC, County Business Patterns (CBP) program, Nonemployer Statistics (NES) program, and Population Estimates Program (PEP) are

Table 3 Geography definitions

Region	Division	States
Midwest	East North Central	*IL*, IN, *MI*, *OH*, WI
	West North Central	IA, KS, MN, MO, ND, NE, SD
Northeast	Middle Atlantic	*NJ*, *NY*, *PA*
	New England	CT, *MA*, ME, NH, RI, VT
South	Potomac	DC, DE, MD, *VA*, WV
	South Atlantic	*FL*, *GA*, *NC*, SC
	East South Central	AL, KY, MS, TN
	West South Central	AR, LA, OK, *TX*
West	Mountain	*AZ*, CO, ID, MT, NM, NV, UT, WY
	Pacific	AK, *CA*, HI, OR, *WA*

Fifteen of the largest states in terms of population and retail sales are in italics

Table 4 Subscript and set notation

Notation	Definition
i	Region
j	Division
k	State
s	Set of states with unsuppressed FD spend
U	Set of all states

obtained from the Census Bureau's American FactFinder tool [28]. Lastly, data from the Bureau of Economic Analysis State Personal Income (SPI) program and the Bureau of Labor Statistics Quarterly Census of Earnings and Wages (QCEW) come from the Federal Reserve Economic Data database [8]. The uses of the variables in Table 2 consist of the following: adjusting FD spend for FD merchant coverage; calculating the dependent variables in the models; accounting for suppressed FD spend in these calculations; serving as model covariates; benchmarking estimates to published national MRTS estimates of retail sales; and validating the final experimental estimates.

Geography

Table 3 defines the geography used in estimation. The regions and divisions agree with those defined by the Census Bureau [29] except in one case; the Census Bureau's South Atlantic division, which consists of nine states, is split into a Potomac division and a new South Atlantic division. Fifteen of the largest states in terms of population and retail sales are presented in italics in Table 3: Arizona, California, Florida, Georgia, Illinois, Massachusetts, Michigan, New Jersey, New York, North Carolina, Ohio, Pennsylvania, Texas, Virginia, and Washington. Also, the term "state" includes the District of Columbia.

Notation

The models are fit separately by industry and month. Consider these values to be fixed in the following notation. Table 4 defines subscripts and sets.

Table 5 defines notation for basic input into estimation such as CBP and NES establishment counts that are used to adjust the FD aggregates of spend for FD merchant coverage. An assumption underlying this adjustment is that FD's merchants are comparable

Table 5 Basic input notation

Notation	Definition
EC_{ijk}	Industry-specific retail sales from the 2012 Economic Census (EC)
m_{ijk}	Number of merchants (FD)
n_{ijk}	Number of establishments (CBP and NES)
POP_{ijk}	Resident population (PEP)
QE_ACC_{ijk}	Quarterly earnings for those working in the accommodation and food sector (SPI)
QE_CON_{ijk}	Quarterly earnings for those working in the construction sector (SPI)
QE_RET_{ijk}	Quarterly earnings for those working in the retail trade sector (SPI)
t_{ijk}^{raw}	Aggregate of spend (FD)
t_{ijk}^{adj}	Adjusted aggregate of FD spend to account for FD merchant coverage $= \left(\frac{n_{ijk}}{m_{ijk}}\right) t_{ijk}^{raw}$
$\widehat{Y}^{MRTS}$	National-level retail sales estimate (MRTS)

Table 6 Dependent variables notation

Notation	Definition
$K_{ijk\|s}^{adj}$	Share of unsuppressed adjusted FD spend $= \frac{t_{ijk}^{adj}}{\sum_{k \in s} t_{ijk}^{adj}}$
L_s	Share of 2012 EC sales associated with unsuppressed states $= \frac{\sum_{k \in s} EC_{ijk}}{\sum_{k \in U} EC_{ijk}}$
y_{ijk}^{adj}	Dependent variable (an estimate of sales) $= K_{ijk\|s}^{adj} L_s \widehat{Y}^{MRTS}$

Table 7 Model notation

Notation	Definition
θ_{ijk}	Retail sales (estimand of interest)
$\boldsymbol{x_{ijk}}$	Vector of covariates from among EC_{ijk}, POP_{ijk}, QE_ACC_{ijk}, QE_CON_{ijk}, and QE_RET_{ijk} (all on the log scale) and also an intercept
$\boldsymbol{\beta}$	Vector of coefficients
q	Length of vectors $\boldsymbol{x_{ijk}}$ and $\boldsymbol{\beta}$
u_{ij}	Random effect to account for geographical division
τ^2	Variance of u_{ij}
ε_{ijk}	Residual error at the state level
σ^2	Variance of ε_{ijk}
B	Posterior sample size
b	Index for the posterior draws

to establishments, which are defined by the Census Bureau as physical locations where business is conducted or services are performed.

Table 6 defines notation for calculating the dependent variables in the models. The idea is to allocate the published national MRTS estimate among the states according to adjusted FD spend. A scaling factor based on retail sales from the 2012 EC accounts for suppressed FD spend.

Lastly, Table 7 defines model notation.

State-level models

The state-level models are cross-sectional linear mixed models [6] fit separately by industry and month in a Bayesian framework. The models attempt to take advantage of the timeliness of the FD transaction data but also utilize quarterly, annual, and quinquennial covariates to smooth the trends in the dependent variable. The dependent variable is itself a state-level estimate of retail sales based on MRTS and FD data. It is modeled as state-level retail sales (the estimand of interest) plus state-level residual error. The estimand of interest is further assumed to equal a linear combination of covariates plus a random effect to account for geographical division. Random effects are used instead of fixed effects in order to reduce the number of model parameters. Also, all values are transformed to the log scale. For states $k \in s$, the following data model is assumed:

$$\begin{aligned} ln\left(y_{ijk}^{adj}\right) &= ln\left(\theta_{ijk}\right) + \varepsilon_{ijk} \\ &= \boldsymbol{x}_{\boldsymbol{ijk}}^{T}\boldsymbol{\beta} + u_{ij} + \varepsilon_{ijk}, \end{aligned}$$

where

$$\begin{aligned} u_{ij} &\sim N\left(0, \tau^2\right) iid \\ \varepsilon_{ijk} &\sim N\left(0, \sigma^2\right) iid. \end{aligned}$$

The u_{ij} are division-level random effects, and the ε_{ijk} are residual state-level errors independent of the u_{ij}.

Commonly used prior distributions are assumed for the model parameters. For the vector of regression coefficients, $\boldsymbol{\beta}$, the noninformative multivariate normal prior $N_q\left(0, 100\boldsymbol{I_q}\right)$ is used. For the standard deviation parameters, τ and σ, the uniform prior distribution $U(0,1)$ is used. This prior distribution has a narrow domain over the unit interval and is thus more informative, but it helps prevent large posterior draws observed in preliminary models. Because the dependent variables and covariates are on the log scale, it is expected that the values of τ and σ are less than 1. Markov chain Monte Carlo (MCMC) is used to simulate draws from the posterior distributions. This method is implemented using SAS/STAT PROC MCMC [23]. For every quantity of interest, the posterior sample size B equals 1000.

The Bayes estimate of the state-level total of retail sales θ_{ijk} under squared error loss equals the posterior mean

$$\begin{aligned} \widehat{\theta}_{ijk} &= \frac{1}{B}\sum_{b=1}^{B}\theta_{ijk}^{(b)} \\ &= \frac{1}{B}\sum_{b=1}^{B} exp\left(\boldsymbol{x}_{\boldsymbol{ijk}}^{T}\boldsymbol{\beta}^{(b)} + u_{ij}^{(b)}\right). \end{aligned} \tag{1}$$

Because the covariates $\boldsymbol{x_{ijk}}$ are available for every state, a Bayes estimate $\widehat{\theta}_{ijk}$ can be obtained for every state as well, even out-of-sample states ($k \notin s$). The corresponding

coefficient of variation (CV) equals the square root of the posterior variance of θ_{ijk} divided by the Bayes estimate $\widehat{\theta}_{ijk}$ given by (1):

$$CV\left(\widehat{\theta}_{ijk}\right) = \frac{\sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\theta_{ijk}^{(b)} - \widehat{\theta}_{ijk}\right)^2}}{\widehat{\theta}_{ijk}}. \tag{2}$$

Regional estimates and corresponding CVs are calculated in a similar fashion. The Bayes estimate of the regional total of retail sales for region i, $\theta_i = \sum_{k\in U_i}\theta_{ijk}$, equals

$$\begin{aligned}\widehat{\theta}_i &= \frac{1}{B}\sum_{b=1}^{B}\theta_i^{(b)} \\ &= \frac{1}{B}\sum_{b=1}^{B}\sum_{k\in U_i}\theta_{ijk}^{(b)} \\ &= \frac{1}{B}\sum_{b=1}^{B}\sum_{k\in U_i} exp\left(\boldsymbol{x}_{\boldsymbol{ijk}}^{T}\boldsymbol{\beta}^{(b)} + u_{ij}^{(b)}\right),\end{aligned} \tag{3}$$

and the corresponding CV equals

$$CV\left(\widehat{\theta}_i\right) = \frac{\sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\theta_i^{(b)} - \widehat{\theta}_i\right)^2}}{\widehat{\theta}_i}. \tag{4}$$

Benchmarking

The state-level estimates $\hat{\theta}_{ijk}$ are not guaranteed to sum to the published national MRTS estimate. This consistency is desirable especially considering estimation begins with allocating the MRTS estimate among the states. To achieve this consistency, a procedure is applied whereby the state-level posterior values $\theta_{ijk}^{(b)}$ are benchmarked, or raked, to the MRTS estimate. The manner in which this is done takes advantage of the availability of regional aggregates of adjusted FD spend.

First, regional estimates are calculated by allocating the published national MRTS estimate to the four regions according to regional adjusted FD spend. Based on the quality evaluation, these regional estimates are believed to be reasonable for the industries being studied. Next, for state $k \in U_i$ and posterior draw b, the benchmarked posterior state-level value is calculated as

$$\theta_{ijk}^{bench(b)} = z_i^{(b)}\theta_{ijk}^{(b)},$$

where

$$z_i^{(b)} = \frac{\left(\frac{t_i^{adj}}{\sum_{i'=1}^{4} t_{i'}^{adj}}\right)\widehat{Y}^{MRTS}}{\theta_i^{(b)}}$$

is the regional benchmarking ratio for region i and draw b. The benchmarked Bayes estimate and corresponding CV, respectively, equal

$$\widehat{\theta}_{ijk}^{bench} = \frac{1}{B}\sum_{b=1}^{B}\theta_{ijk}^{bench(b)} \tag{5}$$

and

$$CV\left(\widehat{\theta}_{ijk}^{bench}\right) = \frac{\sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\theta_{ijk}^{bench(b)} - \widehat{\theta}_{ijk}^{bench}\right)^2}}{\widehat{\theta}_{ijk}^{bench}}. \tag{6}$$

The benchmarked state-level estimates $\widehat{\theta}_{i}^{bench}$ sum to the regional estimates $\left(\frac{t_i^{adj}}{\sum_{i'=1}^{4} t_{i'}^{adj}}\right)\widehat{Y}^{MRTS}$:

$$\begin{aligned}
\widehat{\theta}_{ijk}^{bench} &= \frac{1}{B}\sum_{b=1}^{B}\theta_{i}^{bench(b)} \\
&= \frac{1}{B}\sum_{b=1}^{B}\sum_{k\in U_i}\theta_{ijk}^{bench(b)} \\
&= \frac{1}{B}\sum_{b=1}^{B}\sum_{k\in U_i} z_i^{(b)}\theta_{ijk}^{(b)} \\
&= \frac{1}{B}\sum_{b=1}^{B} z_i^{(b)}\sum_{k\in U_i}\theta_{ijk}^{(b)} \\
&= \frac{1}{B}\sum_{b=1}^{B} z_i^{(b)}\theta_{i}^{(b)} \\
&= \frac{1}{B}\sum_{b=1}^{B}\left(\frac{t_i^{adj}}{\sum_{i'=1}^{4} t_{i'}^{adj}}\right)\widehat{Y}^{MRTS} \\
&= \left(\frac{t_i^{adj}}{\sum_{i'=1}^{4} t_{i'}^{adj}}\right)\widehat{Y}^{MRTS}.
\end{aligned} \tag{7}$$

The last equality in (7) derives from the fact that the summand does not depend on b. It follows readily from (7) that the benchmarked state-level estimates also sum to the published national MRTS estimate $\widehat{Y}^{MRTS}$. It is important to note that there is no variability in the $\theta_i^{bench(b)}$ values from posterior draw to draw. Therefore, the CV calculated using (4), which captures the variability in the $\theta_i^{(b)}$ values, is reported for $\widehat{\theta}_i^{bench}$.

Table 8 Summary of model covariates

Model covariate		Industry				
		444	445	447	448	722
EC_{ijk}	Industry-specific retail sales from the 2012 Economic Census (EC)	X	X	X	X	X
POP_{ijk}	Resident population (PEP)	X		X	X	X
QE_ACC_{ijk}	Quarterly earnings for those working in the accommodation and food sector (SPI)					X
QE_CON_{ijk}	Quarterly earnings for those working in the construction sector (SPI)	X				
QE_RET_{ijk}	Quarterly earnings for those working in the retail trade sector (SPI)		X	X	X	

Limitations

Variance estimation does not take into account uncertainty associated with the MRTS estimates, so the CVs produced by the models are understated. The design-based CVs for the MRTS estimates have the following ranges over the period January 2015 through March 2018: NAICS 444, 1.1–2.5%; NAICS 445, 0.7–1.6%; NAICS 447, 1.5–2.0%; NAICS 448, 1.7–2.6%; NAICS 722, 1.5–2.5% [25]. Future research could involve determining how to incorporate this additional source of uncertainty.

If the regional FD aggregates are suppressed, then the previously described benchmarking procedure cannot be applied. Instead, the posterior state-level values $\theta_{ijk}^{(b)}$ can be raked directly to the MRTS estimate without taking region into account. The estimates and CVs can be calculated similarly as in (5) and (6).

Results

The estimation methodology is applied over the period January 2015 through March 2018 to the five 3-digit NAICS-level industries identified as having acceptable quality. This section summarizes the fitted models and experimental estimates.

Table 8 summarizes the model covariates, which differ by industry. All covariates are transformed to the log scale. The covariate EC_{ijk}, industry-specific retail sales from the 2012 EC, is included in every model because it helps yield estimates whose ordering from large to small seems reasonable. For NAICS 445, the covariate POP_{ijk}, resident population from PEP, is excluded from the model because it has an undesirable effect on this ordering. The three quarterly earnings covariates from SPI, QE_ACC_{ijk}, QE_CON_{ijk}, and QE_RET_{ijk}, are chosen for inclusion based on relevance to the industry.

In assessing model convergence in a Bayesian framework, good practice is followed such as checking trace plots, autocorrelations, posterior densities, and effective sample sizes [12]. These diagnostics show no cause for concern. The trace plots indicate good mixing of the Markov chains, the autocorrelations decrease quickly to 0, and the posterior densities are approximately normal for the regression coefficients and right-skewed for the variance parameters. Figures 4 and 5 display representative diagnostic plots for two parameters in the NAICS 444 March 2018 model.

Additional file 1 contains the experimental monthly estimates of retail sales in dollars for the four Census Bureau-defined regions, the 15 large states identified in Table 3, and the five select industries over the period January 2015 through March 2018. The state-level estimates and corresponding CVs are calculated using (5) and (6), respectively. The regional estimates are the regional benchmarking values given by (7), and the

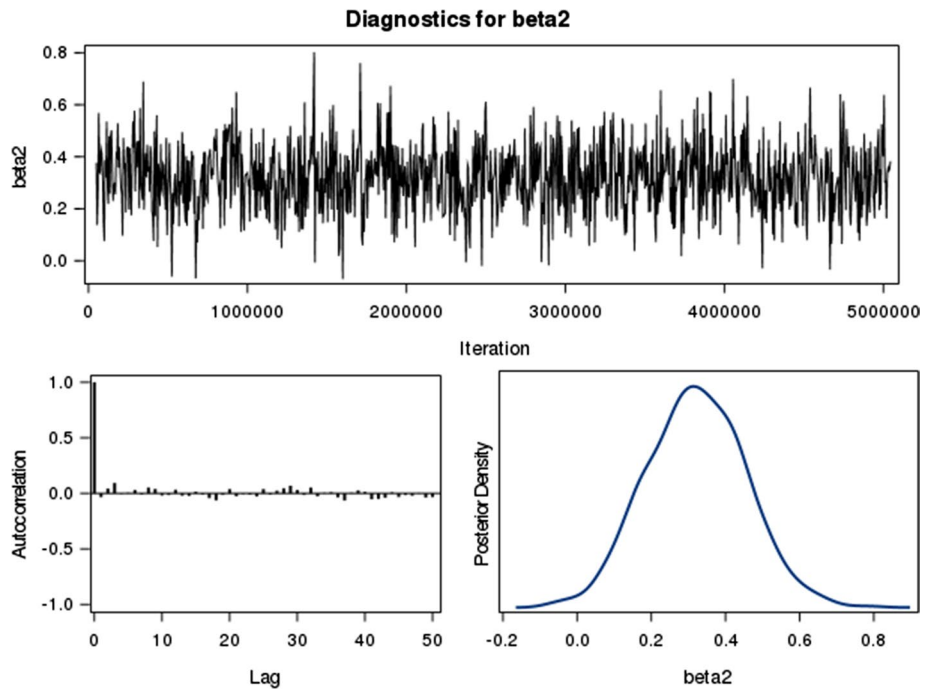

Fig. 4 MCMC diagnostics for β_2 (coefficient of QE_CON_{ijk}) in the NAICS 444 (building material and garden equipment and supplies dealers) March 2018 model

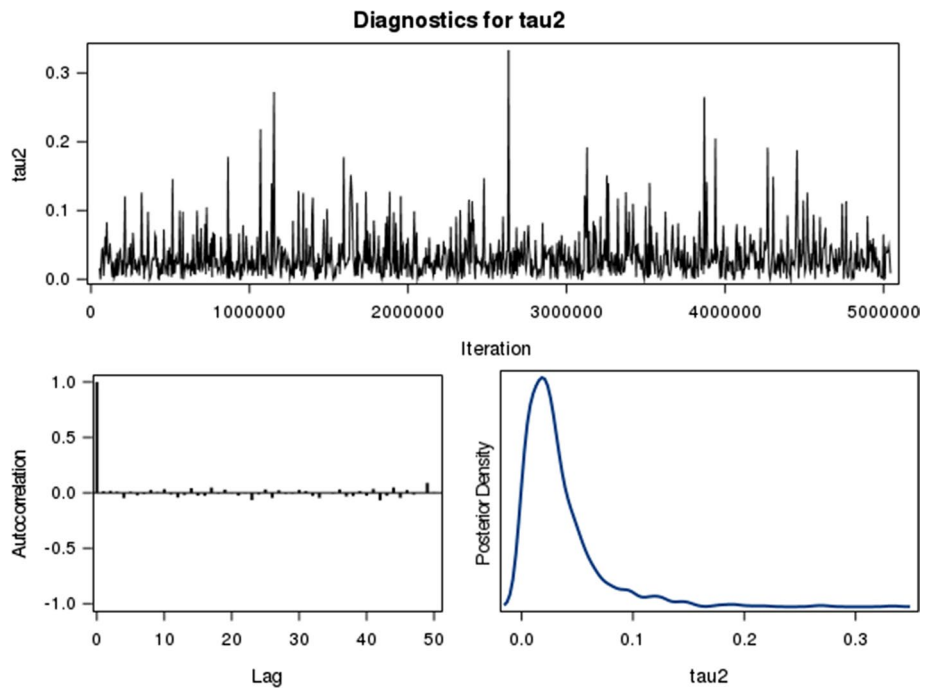

Fig. 5 MCMC diagnostics for τ^2 (variance of the random effect) in the NAICS 444 (building material and garden equipment and supplies dealers) March 2018 model

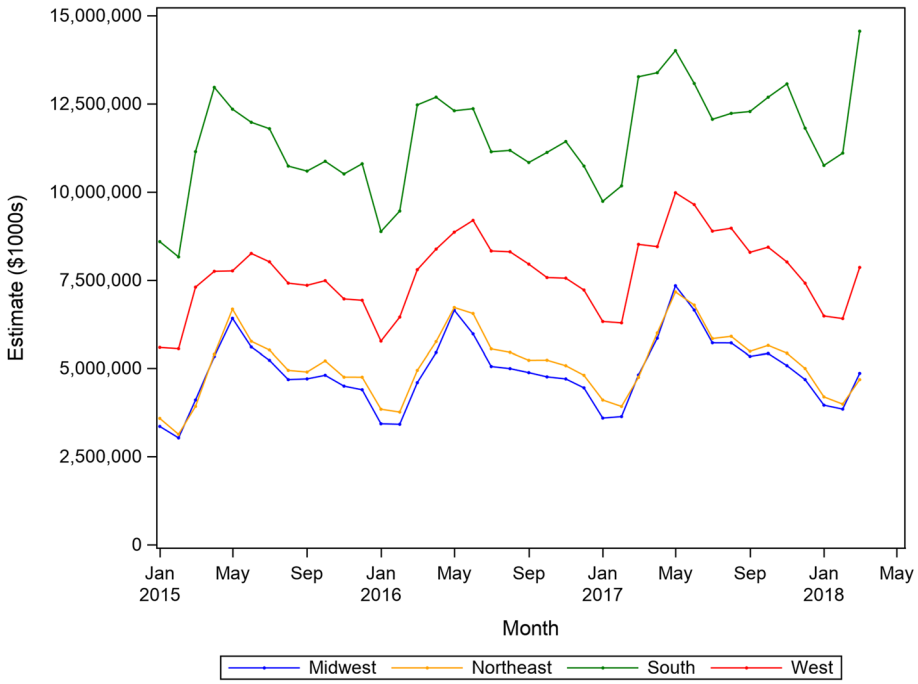

Fig. 6 Regional estimates of retail sales over time for NAICS 444 (building material and garden equipment and supplies dealers)

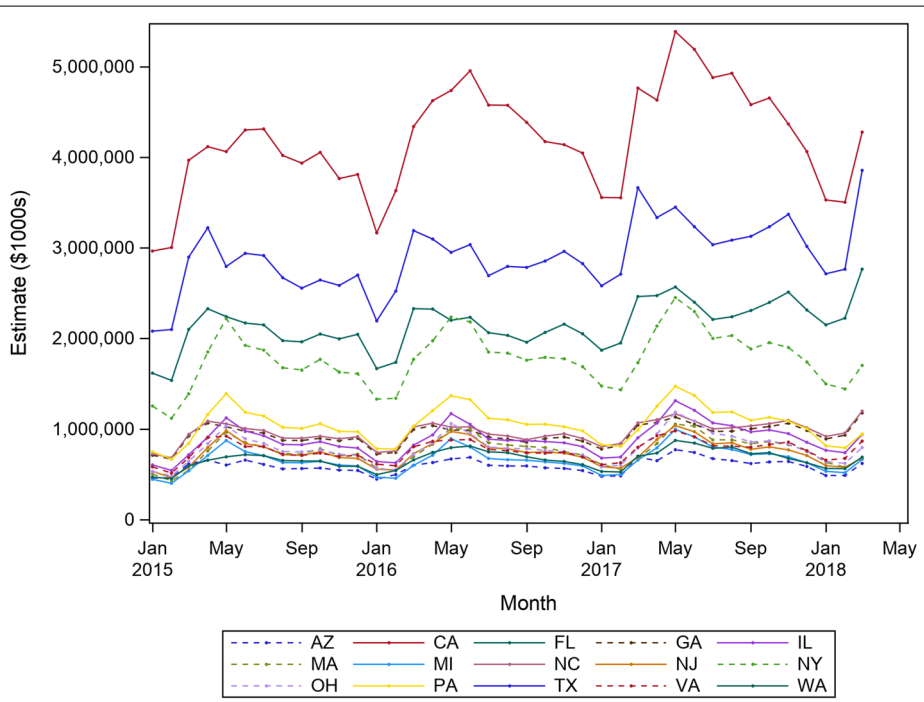

Fig. 7 State-level estimates of retail sales over time for NAICS 444 (building material and garden equipment and supplies dealers)

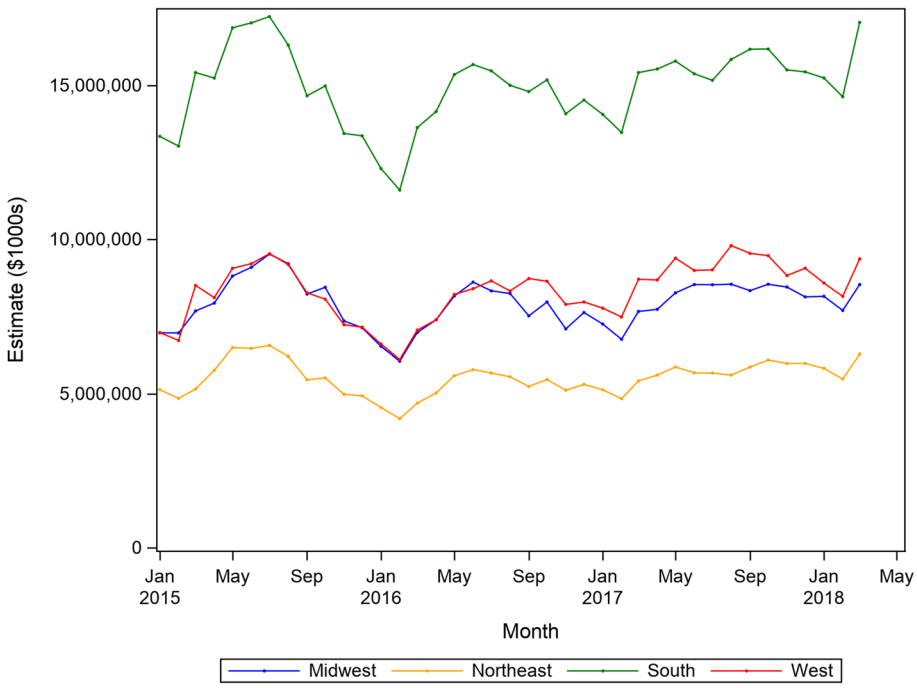

Fig. 8 Regional estimates of retail sales over time for NAICS 447 (gasoline stations)

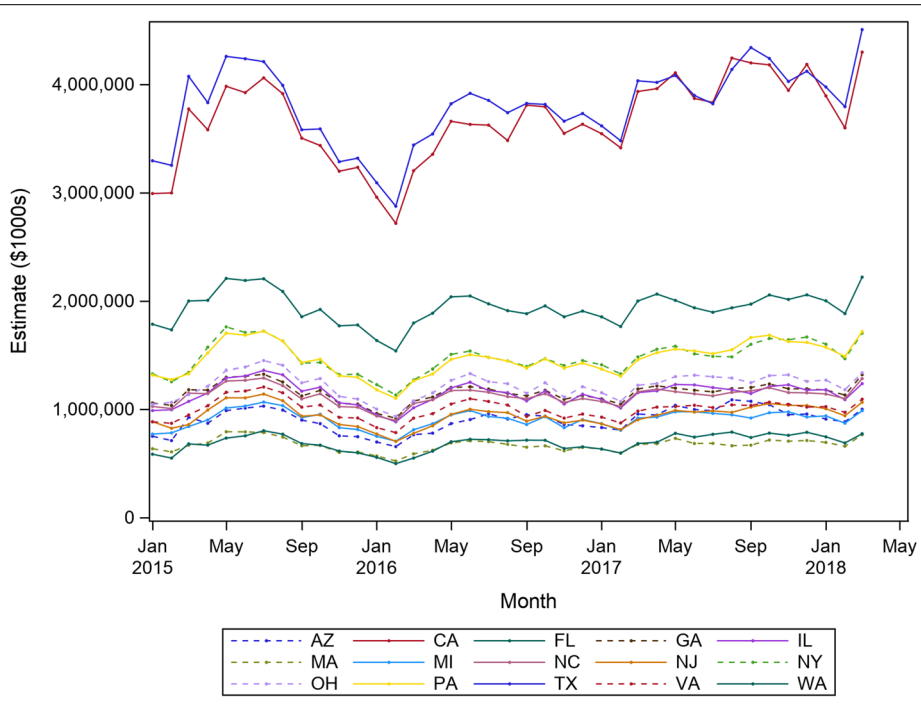

Fig. 9 State-level estimates of retail sales over time for NAICS 447 (gasoline stations)

Table 9 Correlations between state-level retail sales estimates and QCEW estimates

Industry	Number of large states with a QCEW series	Median state correlation	QCEW and MRTS national correlation
444	13	0.89	0.93
445	14	0.32	0.82
447	9	0.43	−0.58
448	13	0.67	0.67
722	15	0.75	0.93

corresponding CVs are calculated using (4). For NAICS 448, the estimates are missing for November 2017 because a suppressed aggregate of FD spend for the Midwest precludes applying the regional benchmarking procedure. For NAICS 447, the regional FD aggregates for the South are suppressed for many months. In this case, instead of applying the preferred regional benchmarking procedure when possible, a national benchmarking procedure is applied each month. To illustrate results, Figs. 6, 7, 8, and 9 plot regional and state-level estimates over time for NAICS 444 and 447.

Discussion and evaluation

A key challenge of this case study involves evaluating the experimental estimates. There are no readily available regional and state-level sample survey estimates based on MRTS that can be used for comparison. Instead, the reasonableness of the estimates is evaluated in two ways: quantitatively with respect to QCEW estimates of the number of employees by industry and qualitatively with respect to additional information about the economy.

For some external validation of the experimental estimates, correlations are calculated over the period January 2015 through March 2018 between the experimental state-level retail sales estimates and state-level QCEW estimates of the number of employees by industry. The QCEW data series are not used in calculating the experimental estimates and are thought to be correlated with retail sales. Unfortunately, a QCEW data series is not available for every state. Table 9 presents correlation summary statistics based on the 15 large states identified in Table 3. For comparison, also given is the correlation between the published national MRTS estimates and the national QCEW estimates over the same period.

The lowest median state correlation is 0.32 for NAICS 445, which is much less than the national correlation of 0.82. This, together with a visual inspection of estimates over time, suggests the model for NAICS 445 may not be picking up on state-level differences in trends very well. Another observation is that the models for NAICS 444 and 722 have the highest median correlations. The models for these two industries take advantage of the availability of an SPI quarterly earnings covariate that is much more industry-specific, whereas the models for NAICS 445, 447, and 448 rely on the SPI quarterly earnings for the entire retail sector. The negative national correlation for NAICS 447 suggests the QCEW estimates are not a good source of validation for this industry; retail sales for NAICS 447 are highly influenced by the price of gasoline.

The experimental estimates are also evaluated qualitatively based on additional information about the economy and other factors:

- *The agreement between the ordering of states in terms of estimated retail sales and the ordering in terms of retail sales totals from the 2012 EC*. In general, there is close agreement between the two orderings. In particular, the largest states California, Texas, Florida, and New York have the highest estimates. The covariate EC_{ijk}, industry-specific retail sales from the 2012 EC, is included in every model because it greatly improves this agreement. The covariate POP_{ijk}, resident population from PEP, is excluded from the model for NAICS 445 because it results in an unreasonable ordering of states.
- *Effects of regional climate and hurricanes*. For NAICS 444, it is sensible that the seasonal peak in sales for the South tends to occur before the onset of higher summer temperatures and before sales in the Midwest and Northeast begin to peak. The estimates should also show signs of the effects of hurricanes such as Hurricanes Harvey and Irma, which hit the South in August and September 2017, respectively. The estimates for Florida for NAICS 448 and 722, for example, do show more of a pronounced dip in sales in September 2017.
- *Presence of anomalous estimates and trends related to identified FD quality issues*. For example, for NAICS 448, the estimates for California and the West appear to be high during 2016, with an especially high peak in December 2016. These suspicious estimates are related to the same trend observed in the FD transaction data.
- *Additional knowledge about the economy and industries*. Evaluating the quality of estimates for NAICS 447 involved consulting the U.S. Energy Information Administration [30] for the average price of a gallon of gasoline over time. According to this source, the peak around May 2015 and the dip in February 2016 observed in the retail sales estimates are explainable. Regarding NAICS 445 and 722, these two industries deal with food and beverages, which are necessities. It seems reasonable that states would have similar sales patterns, although the QCEW correlations suggest the patterns in estimates for 445 may be too similar from state to state.

Conclusions

The use of Big Data sources opens many doors for official statistics and offers insight into economic trends that exist at a more granular level. However, data quality must be analyzed carefully. In this case study, the quality of the FD aggregates is evaluated with regard to multiple criteria such as suppression rates, coverage rates, and trends. Only five of the thirteen 3-digit NAICS-level industries in scope to MRTS are identified as having acceptable quality for estimation. Estimation methodology based on linear mixed models in a Bayesian framework is developed to produce regional and state-level monthly estimates of retail sales. These models try to take advantage of the timeliness of the FD transaction data. By incorporating variables from other demographic and economic data sources as covariates, the models also attempt to smooth over undesirable features of FD's business activity and the quirks

of payment processing. Many features of the resulting estimates seem reasonable, but this research does raise some caution flags. These include the difficulty of finding an external data source for validation and the presence of anomalous trends in the estimates related to identified FD data quality issues. Future work involves calculating estimates for more recent months and researching alternative methods for evaluating their accuracy.

Based on the Census Bureau's overall experience with Big Data sources, literature reviews should be performed to see whether others have successfully used the data. Also, transparency issues should be addressed in the contracts with third-party data vendors. On a final note, Big Data for official statistics involves using data for purposes other than the ones for which the data were originally created. As an example of such an organic data source, the FD aggregates may have limitations when it comes to adding geographic granularity to MRTS, but they may be extremely useful for other purposes such as understanding payment card usage in the United States.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s40537-019-0242-z.

Additional file 1. Experimental estimates.

Abbreviations

CBP: County Business Patterns; CV: coefficient of variation; EC: Economic Census; FD: First Data; MCC: Merchant Category Code; MCMC: Markov chain Monte Carlo; MRTS: Monthly Retail Trade Survey; NAICS: North American Industry Classification System; NES: Nonemployer Statistics; PEP: Population Estimates Program; QCEW: Quarterly Census of Earnings and Wages; SPI: State Personal Income.

Acknowledgements

The authors would like to acknowledge and thank colleagues at the U.S. Census Bureau for reviewing drafts of this paper and providing helpful comments.

Disclaimer

Any views expressed are those of the authors and not necessarily those of the U.S. Census Bureau. The Census Bureau has reviewed this data product for unauthorized disclosure of confidential information and has approved the disclosure avoidance practices applied (Approval ID: CBDRB-FY19-070).

Authors' contributions

BD is the primary author and led work on the estimation methodology. BD and DM performed exploratory data analysis for quality criteria and assessed findings with CH. CH also served as the champion for this project. All authors read and approved the final manuscript.

Funding

Not applicable.

Availability of data and materials

First Data's electronic transaction data are proprietary and not publicly available. The auxiliary data from the U.S. Census Bureau, the Bureau of Economic Analysis, and the Bureau of Labor Statistics are publicly available online. The experimental regional and state-level monthly estimates of retail sales and corresponding coefficients of variation can be found in Additional file 1.

Competing interests

The authors declare that they have no competing interests.

Received: 24 April 2019 Accepted: 13 August 2019
Published online: 29 August 2019

References

1. Bostic WG Jr, Jarmin RS, Moyer B. Modernizing federal economic statistics. Am Econ Rev. 2016;106(5):161–4. https://doi.org/10.1257/aer.p20161061.
2. Capps C, Wright T. Toward a vision: official statistics and big data. In: Amstat News (Aug). American Statistical Association, Alexandria, 2013, p. 9–13. https://magazine.amstat.org/blog/2013/08/01/official-statistics/. Accessed 22 Apr 2019.

3. D'Amuri F, Marcucci J. The predictive power of Google searches in forecasting US unemployment. Int J Forecast. 2017;33(4):801–16. https://doi.org/10.1016/j.ijforecast.2017.03.004.
4. DeGennaro RP. Merchant acquirers and payment card processors: a look inside the black box. Econ Rev. 2006;91(1):27–42.
5. Dumbacher B, Hutchinson R. Enhancing the foundation of official economic statistics with big data. Paper presented at the 2016 European Conference on Quality in Official Statistics, Círculo de Bellas Artes, Madrid, 1–3 June 2016.
6. Faraway JJ. Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models. 2nd ed. Boca Raton: CRC Press; 2016.
7. Fay RE, Herriot RA. Estimates of income for small places: an application of James-Stein procedures to census data. J Am Stat Assoc. 1979;74(366):269–77. https://doi.org/10.1080/01621459.1979.10482505.
8. Federal Reserve Bank of St. Louis. Federal Reserve Economic Data. 2018. https://fred.stlouisfed.org. Accessed 6 Dec 2018.
9. First Data Corporation. First Data 2016 annual report. In: First Data investor relations. 2017. https://investor.firstdata.com/~/media/Files/F/FirstData-IR/documents/ar-2016.pdf. Accessed 20 July 2018.
10. First Data Corporation. Merchant services, credit card processing & payment solutions. 2019. https://www.firstdata.com/en_us/home.html. Accessed 3 Apr 2019.
11. Galbraith JW, Tkacz G. Nowcasting with payments system data. Int J Forecast. 2018;34(2):366–76. https://doi.org/10.1016/j.ijforecast.2016.10.002.
12. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. Bayesian data analysis. 3rd ed. Boca Raton: CRC Press; 2013.
13. Groves RM. Three eras of survey research. Public Opin Q. 2011;75(5):861–71. https://doi.org/10.1093/poq/nfr057.
14. Japec L, Kreuter F, Berg M, Biemer P, Decker P, Lampe C, Lane J, O'Neil C, Usher A. Big data in survey research: AAPOR task force report. Public Opin Q. 2015;79(4):839–80. https://doi.org/10.1093/poq/nfv039.
15. Jarmin RS. Evolving measurement for an evolving economy: thoughts on 21st century US economic statistics. J Econ Persp. 2019;33(1):165–84. https://doi.org/10.1257/jep.33.1.165.
16. Kreuter F, Peng RD. Extracting information from big data: issues of measurement, inference and linkage. In: Lane J, Stodden V, Bender S, Nissenbaum H, editors. Privacy, big data, and the public good: frameworks for engagement. Cambridge: Cambridge University Press; 2014. p. 257–75. https://doi.org/10.1017/cbo9781107590205.
17. Landefeld S. Uses of big data for official statistics: privacy, incentives, statistical challenges, and other issues. Paper presented at the International Conference on Big Data for Official Statistics, Beijing, 28–30 Oct 2014.
18. Marchetti S, Giusti C, Pratesi M, Salvati N, Giannotti F, Pedreschi D, Rinzivillo S, Pappalardo L, Gabrielli L. Small area model-based estimators using big data sources. J Off Stat. 2015;31(2):263–81. https://doi.org/10.1515/jos-2015-0017.
19. Mayer-Schönberger V, Cukier K. Big data: a revolution that will transform how we live, work, and think. Boston: Houghton Mifflin Harcourt; 2013.
20. National Academies of Sciences, Engineering, and Medicine. Reengineering the Census Bureau's annual economic surveys. Abraham KG, Citro CF, White Jr. GD, Kirkendall NK, eds. Washington: The National Academies Press; 2018. https://doi.org/10.17226/25098.
21. Porter AT, Holan SH, Wikle CK, Cressie N. Spatial Fay-Herriot models for small area estimation with functional covariates. Spat Stat. 2014;10:27–42. https://doi.org/10.1016/j.spasta.2014.07.001.
22. Rao JNK, Molina I. Small area estimation. 2nd ed. Hoboken: Wiley; 2015. https://doi.org/10.1002/9781118735855.
23. SAS Institute Inc. The MCMC procedure. In: SAS/STAT 14.3 user's guide. 2018. http://documentation.sas.com/?docsetId=statug&docsetTarget=statug_mcmc_syntax01.htm&docsetVersion=14.3&locale=en. Accessed 6 Aug 2018.
24. Schroeck M, Shockley R, Smart J, Romero-Morales D, Tufano P. Analytics: the real-world use of big data: how innovative enterprises extract value from uncertain data, Executive Report. IBM Institute for Business Value and Saïd Business School at the University of Oxford. 2012. https://www.bdvc.nl/images/Rapporten/GBE03519USEN.PDF. Accessed 3 June 2019.
25. U.S. Census Bureau. Monthly retail trade. 2018. https://www.census.gov/retail/index.html. Accessed 15 Nov 2018.
26. U.S. Census Bureau. Economic Census. 2018. https://www.census.gov/programs-surveys/economic-census.html. Accessed 15 Nov 2018.
27. U.S. Census Bureau. North American Industry Classification System. 2018. https://www.census.gov/eos/www/naics/. Accessed 7 Dec 2018.
28. U.S. Census Bureau. American FactFinder. 2018. https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml. Accessed 6 Dec 2018.
29. U.S. Census Bureau. Census regions and divisions of the United States. In: Geography general reference maps. 2018. https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf. Accessed 23 July 2018.
30. U.S. Energy Information Administration. Petroleum & other liquids: weekly retail gasoline and diesel prices. 2018. https://www.eia.gov/dnav/pet/pet_pri_gnd_a_epm0_pte_dpgal_m.htm. Accessed 15 Nov 2018.

Publisher's Note