Monthly State Retail Sales Technical Documentation

1. Background

The Census Bureau is producing new monthly state retail sales as an *experimental product*.[1] This product includes state-level year-to-year percent changes of monthly retail sales both for Total Retail Sales excluding Nonstore Retailers and for 11 North American Industry Classification System (NAICS) retail subsectors. These measures are composite estimates combining independently-obtained synthetic estimates and hybrid estimates comprising third-party and directly-collected establishment (point of sale) sales data and modeled establishment data. Retail subsector NAICS 454 is not included in the experimental release. Consequently, the aggregated industry by state estimates will not equal the published MRTS total but will sum to the published retail sales total of this subset of the MRTS three-digit NAICS subsectors

The MRTS is a mail-out/mail-back survey of about 13,000 retail and food services businesses with paid employees, whose sampling unit is the firm. The one-stage stratified sampling design is designed to produce reliable industry-level estimates. A new sample is selected from the Business Register approximately every five years, and the sample is updated quarterly to reflect employer business "births" and "deaths" by adding new employer businesses identified in the Business and Professional Classification Survey and dropping firms and EINs when it is determined they are no longer active. Sampled firms report for *all* their retail establishments. For more details on the MRTS sampling design and data collection, see https://www.census.gov/retail/mrts/how_surveys_are_collected.html.

Because the MRTS sampling unit is the firm, there is no geographic component to design. Sampling weights represent a firm's contribution to the industry and do not reflect industry-state share in total sales. With MRTS, the geographic information is *only* available from the single-unit (SU)[2] firms and for the multi-unit (MU) firms that operate within a single state.

Throughout the remainder of the document, we use the following definitions:

Estimation Industry:	3-, 4-, or 5-digit NAICS code assigned to a study unit (establishment or MRTS sample unit), using the most specific disaggregation level supported by the frame and the third party data described in Section 2.
Tabulation Industry:	3-digit NAICS code (ALL = aggregated over all relevant MRTS industries)
Geography:	FIPS state code (ALL = aggregated over all states)

Appendix One provides the cross-walk between tabulation industries and estimation industries.

2. Data Sources

Monthly Sales:

- MRTS sample data at statistical period *t* from the sampling unit (firm)
  - Sampling unit data may be split into separate reporting parts (tabulation units) to represent the different industries in which the firm operates.

[1] 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-FY20-356)

[2] A single-unit (SU) firm owns or operates a business at a single physical location (establishment), whereas multi-unit (MU) firms comprise two or more establishments that are owned or operated by the same firm.

- Third-Party Data
    - Retailer point-of-sale data purchased from The NPD Group, Inc.
        - Monthly point-of-sales sales data for all establishments in twenty-two large companies identified by the Census Bureau.
        - Aggregated 3-digit NAICS by state point-of-sale sales data for a designated set of multi-unit companies.

<u>Frame (Annual) Data</u>

- Business register (BR): complete list of businesses in the U.S. Used to provide
    - Geography for all establishments in MRTS frame
    - Activity status of all establishments in MRTS frame
    - NAICS code for all establishments in the MRTS frame
    - Gross payroll (2018) for all establishments in MRTS frame obtained from tax returns
- MRTS sampling frame: union of the original sampling frame based on the Business Register as of December 2015 and subsequent birth sampling frames

3. State-Level Monthly Retail Sales Estimators

The composite estimate of monthly retail sales in tabulation industry *i* and state *g* at statistical period *t* is given by:

$$\hat{Y}_{igt}^{C} = f_{it}\left[\emptyset_{igt}\left(\hat{Y}_{igt}^{B}\right) + (1 - \emptyset_{igt})\left(\hat{Y}_{igt}^{T}\right)\right]$$

where

$\hat{Y}_{igt}^{T}$= "Top Down" synthetic estimate of monthly retail sales in tabulation industry *i* and state *g* at statistical period *t* (See Section 3.1)

$\hat{Y}_{igt}^{B}$= "Bottom Up" hybrid estimator of monthly sales in tabulation industry *i* and state *g* at statistical period *t* (See Section 3.2)

Due to the independent estimation procedure, $\hat{Y}_{igt}^{T} \perp \hat{Y}_{igt}^{B}$. This is a conservative assumption that could lead to overestimation of the variance of $\hat{Y}_{igt}^{C}$.

The compositing factor in tabulation industry *i* and state *g* at statistical period *t* is given by

$$\emptyset_{igt} = \frac{v(\hat{Y}_{igt}^{T})}{v(\hat{Y}_{igt}^{T}) + v(\hat{Y}_{igt}^{B})}$$

This minimizes the variance of the composite estimate. Traditionally, composite estimates minimize the mean squared error of the estimate (variance plus squared bias). However, it is not possible to estimate the bias of these composite estimates or the component estimates presented in Sections 3.1 and 3.2 because "true" monthly state-level retail sales totals by 3-digit tabulation industry are not available and comparable state-level benchmarks are limited to a small number of states and industries.

A final ratio adjustment is applied to each composite estimate within tabulation industry to ensure additivity to the corresponding *published* MRTS retail sales total ($\hat{Y}_{it}^{MRTS,PUB}$), computed as $f_{it} = \hat{Y}_{it}^{MRTS,PUB} / \sum_{g=1}^{51} \hat{Y}_{igt}^{C}$.

The variance of the composite monthly estimate of sales in 3-digit tabulation industry *i* and state *g* at statistical period *t* is obtained as

$$v\left(\hat{Y}_{igt}^{C}\right) \approx f_i^2\ \emptyset_{igt} v\left(\hat{Y}_{igt}^{B}\right)$$

3.1 Top Down (Synthetic) Estimator

The "Top Down" (TD) monthly retail sales estimate for state *g* in tabulation industry *i* at statistical period *t* is given by

$$\hat{Y}_{igt}^{T} = \sum_{j \in i} \left(\frac{X^{jg}}{X^{j}}\right) \hat{Y}_{jt}^{MRTS}$$

where

X^{jg} is total BR gross annual payroll in state *g* in estimation industry *j* obtained from the frame;
X^{j} is total BR gross annual payroll in estimation industry *j;* and
$\hat{Y}_{jt}^{MRTS}$ is the MRTS Horvitz-Thompson retail sales estimate *before* benchmarking from estimation industry *j*, hereafter referred to as the *unpublished* MRTS retail sales estimate

This synthetic estimator provides computationally simple state level estimates of monthly retail sales within an industry that add exactly to the survey total. However, it has numerous disadvantages. Specifically, it

1. Requires very strong assumptions about the relationship between gross payroll (annual) and monthly sales for all states;
2. Ensures that month-to-month and year-to-year change estimates for each state are equivalent to the corresponding industry total estimates;
3. Fails to capture regional or state seasonal patterns; and
4. Cannot be "improved" as its reliability is a function of the MRTS sample design and response.

Furthermore, as mentioned in Section 2 above, it is impossible to estimate the bias of the synthetic estimates. Consequently, it is a useful fallback estimator when third party or MRTS state-level data are not available, but in not favored in the compositing procedure.

The TD estimator variance is given by

$$v\left(\hat{Y}_{igt}^{T}\right) \approx \left\{\sum_{j \in i} \left[\frac{X^{jg}}{X^{j}}\right]^2 \hat{v}\left(\hat{Y}_{jt}^{MRTS}\right) + \left[\frac{\hat{Y}_{jt}^{MRTS}}{X^{i}}\right]^2 \hat{v}\left(X^{jg}\right)\right\}$$

where

$\hat{v}(\hat{Y}_{j}^{MRTS})$ = variance estimator for MRTS estimation industry-level monthly retail sales that accounts for sampling error, nonresponse error, and imputation error using the methods outlined in Kim and Rao (2009); and
$\hat{v}(X^{jg})$ = variance of state-level gross annual payroll, assuming that $x_{jgk} \sim \left(\mu_{jg}, \sigma_{jg}^2\right)$ in the superpopulation from which the frame data are drawn, where *k* is the establishment within estimation industry *j* and state *g,* estimated as

$$\hat{v}\left(X^{gj}\right) \approx n_{jg} \frac{\sum_{k \in jg}\left(x_{jgk} - \bar{x}_{jg}\right)^2}{n_{jg} - 1} = \left(\frac{n_{jg}}{n_{jg} - 1}\right) \sum_{k \in jg} \left(x_{jgk} - \bar{x}_{jg}\right)^2$$

The variance estimator is a linearization estimator, since the frame totals are random variables, not constants. Business populations are not static, and the frame represents a “point in time” snapshot (sample). Furthermore, the assembled frame used to produce these estimates is subject to linking errors. Treating the frame totals of gross annual payroll as fixed appeared to underestimate the variance substantively in comparison to the corresponding “Bottom Up” estimator variances presented in Section 3.2. More important, the additional variance component increased the variability of the state-level monthly retail sales estimates within industry, with the variability increase approximately inversely proportional to the number of establishments in the industry and state. Finally, the linearization estimator does not include a covariance term because the unpublished MRTS estimation industry monthly retail sales estimates are independent of the frame estimation industry and estimation industry by state gross annual payroll estimates.

3.2. Bottom Up Estimator

3.2.1. Estimation

The “Bottom Up” (BU) estimate of monthly retail sales for state *g* in tabulation industry *i* at statistical period *t* is given by

$$\hat{Y}^{B}_{igt} = Z^{igt}_{MU} + Y^{igt}_{MU*} + Y^{igt}_{SU} + \left[\frac{\hat{Y}^{MRTS}_{it} - \sum_{g\in i} Z^{igt}_{MU} + Y^{igt}_{MU*} + Y^{igt}_{SU})}{\sum_{g} \breve{Y}^{igt}_{MU}}\right] \breve{Y}^{igt}_{MU}$$

where

Z^{igt}_{MU} is the third-party data aggregate retail sales in state *g* from the preselected MU companies in tabulation industry *i* at statistical period *t*;

Y^{igt}_{MU*} are the aggregated unweighted retail sales in state *g* in tabulation industry *i* at statistical period *t* for MRTS MU companies that operate entirely within a single state;

Y^{igt}_{SU} are the aggregated unweighted retail sales in state *g* in tabulation industry *i* at statistical period *t* from MRTS SU establishments; and

$\breve{Y}^{igt}_{MU}$ is the aggregate value of the *imputed* MU retail sales in state *g* in tabulation industry *i* at statistical period *t*

The bracketed term is a tabulation industry-level ratio adjustment that enforces consistency with the unpublished MRTS monthly retail sales estimates in each tabulation industry and accounts for units that are not eligible for imputation such as single unit establishments, MRTS sampling units that did not match to the frame, and new retail firms (births) without business register payroll data.

The BU estimator has several advantages in terms of statistical properties over the TD estimator presented in Section 3.1. Because it maximizes use of auxiliary and directly-collected data, it could yield theoretically unbiased estimates if all retail trade establishments in the industry and state were available. As additional auxiliary or directly collected data are available and are incorporated, the accuracy and precision will improve. Variance can also be reduced by improving the imputation models discussed in Section 3.2.2. This estimator has the following disadvantages:

1. It assumes there is no measurement error from auxiliary and directly-collected data.
2. It is a poor estimator for the single unit component, as single unit “imputation” is accomplished via the national industry level ratio adjustment. The overall effect of this poor imputation is minimized when third-party data total is close to corresponding MRTS industry total.
3. It can be extremely variable (see Section 3.2.2.)

The BU estimator variance is given by

$$v(\hat{Y}_{igt}^{B}) = v\left(\left[\frac{\hat{Y}_{it}^{MRTS} - \sum_{g \in i} Z_{MU}^{igt} + Y_{MU*}^{igt} + Y_{SU}^{igt})}{\sum_g \check{Y}_{MU}^{igt}}\right] \check{Y}_{MU}^{igt}\right) \approx \left[\frac{\hat{Y}_{it}^{MRTS} - \sum_{g \in i} Z_{MU}^{igt} + Y_{MU*}^{igt} + Y_{SU}^{igt})}{\sum_g \check{Y}_{MU}^{igt}}\right]^2 v(\check{Y}_{MU}^{igt})$$

where $v(\check{Y}_{MU}^{igt})$ is the multiple-imputation variance estimate for each industry and state. Note that since the BU estimator approximates a population instead of a sample, there is no within-imputation term in the multiple imputation variance estimate (Vink and van Buuren 2014).

3.2.2. Imputation

The imputation model is a Bayesian formulation of a linear mixed model that uses regression and random effects parameters to predict monthly retail sales given gross payroll, state, and NAICS code. Multiple imputations from the predictive posterior distribution estimate the missing MU establishment data. Model parameters are estimated using MU establishment data with non-missing sales data; MU establishments with missing monthly retail sales data are imputed from the estimated model using frame data. SU establishments' sales data are not imputed but are accounted for in the final ratio adjustment to the bottom up estimates (see Section 3.2.1). Models are estimated and imputations are produced independently for each month and imputation industry, which are generally defined at the 3-digit NAICS level (see Appendix Two for a cross-walk between tabulation and imputation industries). All model parameters are estimated using R (R Core Team 2017, https://www.R-project.org).

Establishment level sales ($z_{i(dgk)}$) are either provided by third-party data or estimated from reported MRTS firm data[3], where *d* indexes the most disaggregated NAICS code available for establishment *k;* the parenthesis in the subscripts indicates nesting within tabulation industry. The regression parameters for this model capture the national level relationship between an establishment's logged monthly sales and logged annual gross payroll within 3-digit tabulation industry for a given month. For the remainder of this section, the *t* subscript (indexing statistical period) is omitted, gross annual payroll is referred to as payroll, monthly retail sales are referred to as sales, and the disaggregated NAICS codes indexed by *d* are referred to as detailed NAICS.

Geography variations are modeled with state-level random effects, which capture deviation from the national industry trend. An Intrinsic Conditional Auto-Regressive (ICAR) prior is used for the random state effects, which smooths estimates by modeling correlation between adjacent states (Morris et al., 2019). Hawaii and Alaska are treated as "islands" and are modeled independently from other states, but with the same variance. An additional detailed NAICS random effect is included in the imputation model when sales data is observed in a majority of detailed industries included in the imputation industry.

The general form of the imputation model is given by

$$z_{i(dgk)} = \alpha_i + \beta_i x_{i(dgk)} + \tau_{i(g)} + \delta_{i(d)} + \varepsilon_{i(dgk)} \tag{1}$$

where

$z_{i(dgk)}$ log(sales + 1) for establishment *k* in state *g* for detailed NAICS *d* within imputation industry *i*
x_{idgk} log(payroll + 1) for establishment *k* in state *g* for detailed NAICS *d* within imputation industry *i*
α_i National level intercept for imputation industry *i*
β_i National level industry slope for imputation industry *i*
$\tau_{i(g)}$ State random effect for state *g* within imputation industry *i*
$\delta_{i(d)}$ Detailed NAICS random effect for NAICS *d* within imputation industry *i*

[3] Establishment-level estimates of monthly sales are obtained by pro-rating the reported value from the MRTS firm to firm's establishments by each establishment's proportion of gross payroll (to the total firm). This procedure essentially mimics the Top down estimation procedure at the firm level. Firms with imputed values of monthly sales are excluded.

$\varepsilon_{i(dgk)}$ Residual error term for establishment *k* in state *g* for detailed NAICS *d* within imputation industry *i*

Error terms are modeled as

$$\varepsilon_{i(dgk)} \sim normal\left(0, \sigma_{\varepsilon_i}^2\right)$$

$$\tau_{i(g)} \sim normal\left(0, \sigma_{\delta_i}^2\right)$$

$$\delta_{i(d)} \sim normal\left(0\,, \sigma_{\tau_i}^2 [D - A]^{-1}\right)$$

where A is an G x G matrix that defines neighbors and D is a diagonal G x G matrix that defines the number of neighbors.

The regression parameter priors are estimated using restricted maximum likelihood via the "lme4" R package. All other priors are from a uniform distribution.

$$\alpha_i \sim normal\left(\alpha_i^*\,, \sigma_{\alpha_i^*}^2\right)$$

$$\beta_i \sim normal\left(\beta_i^*\,, \sigma_{\beta_i^*}^2\right)$$

If the relationship between log sales and log payroll appears to be nonlinear, a piecewise regression -- with at most two breaking points (c_{i1}, c_{i2}) -- may be used to model curvature.

$$z_{i(dgk)} = \begin{cases} \alpha_{i2} + \beta_{i2} x_{i(dgk)} + \tau_{i(g)} + \delta_{i(d)} + \varepsilon_{i(dgk)} & x_{i(dgk)} < c_{i1} \\ \alpha_{i2} + \beta_{i2}\left(x_{i(dgk)} - c_{i1}\right) + \tau_{i(g)} + \delta_{i(d)} + \varepsilon_{i(dgk)} & c_{i1} \leq x_{i(dgk)} < c_{i2} \\ \alpha_{i3} + \beta_{i3}\left(x_{i(dgk)} - c_{i2}\right) + \tau_{i(g)} + \delta_{i(d)} + \varepsilon_{i(dgk)} & c_{i2} \leq x_{i(dgk)} \end{cases} \tag{2}$$

where $\alpha_{i2} = \alpha_{i1} + \beta_{i1} c_{i1}$ and $\alpha_{i3} = \alpha_{i2} + \beta_{i2}(c_{i2} - c_{i1})$. Priors are added to the breaking points and are estimated using the bootstrap restarting algorithm described in Wood (2001) and implemented in the "segmented" R package: $c_{i1} \sim normal\left(c_{i1}^*\,, \sigma_{c_{i1}^*}^2\right)$ and $c_{i2} \sim normal\left(c_{i2}^*\,, \sigma_{c_{i2}^*}^2\right)$.

The industry level imputation models are evaluated each month. Each month, the first step of the model development process is to examine the percentage of reported zero sales from the establishments within the imputation industry[4]. In most statistical periods, zero sales values are treated as outliers and are therefore excluded from parameter estimation to prevent overrepresentation of "closed" businesses in a state and tabulation industry. However, if the frequency of establishments that reported zero sales is greater than one percent, then a two-step imputation procedure is adopted for the imputation industry. First, a logistic regression is used to model the probability of observing an establishment with zero sales, with geography as the sole predictor:

$$log\left(\frac{p(z_{i(dgk)}=0)}{1-p(z_{i(dgk)}=0)}\right) = \alpha_{i0} + \gamma_{i(g)}$$

$$\gamma_{i(g)} \sim normal\left(0\,, \sigma_{\gamma_i}^2 [D - A]^{-1}\right)$$

Then, the predicted nonzero establishment values are modeled using (1) or (2), depending on model fit diagnostics for the tabulation industry given the observed nonzero establishment data.

[4] This step is crucial for the April 2020 and May 2020 estimates to address the differing state-level responses to the COVID-19 pandemic.

$$z_{i(dgk)} = \begin{cases} \alpha_i + \beta_i x_{i(dgk)} + \tau_{i(g)} + \delta_{i(d)} + \varepsilon_{i(dgk)} & z_{i(dgk)} > 0 \\ 0 & z_{i(dgk)} \leq 0 \end{cases} \quad (3)$$

Appendix Two provides the imputation models used for each imputation industry. Unless otherwise specified, the general model is used.

Model parameters are estimated using Bayesian inference with the open-source probabilistic programming language "Stan" in R. Stan (https://mc-stan.org) uses a No-U-Turn sampler (NUTS), which is a variation of Hamiltonian Monte Carlo (HMC). Imputations and variances are estimated from 1,000 multiple imputations drawn from the posterior distribution. Point estimates for each establishment are obtained as the mean of the 1,000 independent draws. The imputation variances are estimated from 1,000 totals calculated from each posterior draw.

4. Year-to-Year Change Estimates

The year-to-year change (trend ratio) estimate is given by

$$\hat{\theta}_{igt}^{TR} = \frac{\hat{Y}_{igt}^{C}}{\hat{Y}_{ig,t-12}^{C}}$$

with the corresponding percentage change estimate given by

$$\hat{\theta}^{P} = \frac{\hat{Y}_{igt}^{C} - \hat{Y}_{ig,t-12}^{C}}{\hat{Y}_{ig,t-12}^{C}}$$

The variance estimate for is equivalent for the trend ratio and percentage change estimate and is estimated by the Taylor linearization variance estimator

$$\hat{v}\left(\hat{\theta}_{igt}^{TR}\right) \approx \left[\frac{\hat{Y}_{igt}^{C}}{\hat{Y}_{ig,t-12}^{C}}\right]^2 \left[\frac{\hat{v}\left(\hat{Y}_{igt}^{C}\right)}{\left(\hat{Y}_{igt}^{C}\right)^2} + \frac{\hat{v}\left(\hat{Y}_{ig,t-12}^{C}\right)}{\left(\hat{Y}_{ig,t-12}^{C}\right)^2} - 2\,\frac{Cov\left(\hat{Y}_{igt}^{C}, \hat{Y}_{ig,t-12}^{C}\right)}{\hat{Y}_{igt}^{C}\hat{Y}_{ig,t-12}^{C}}\right]$$

$$Cov\left(\hat{Y}_{igt}^{C}, \hat{Y}_{ig,t-12}^{C}\right) \approx \frac{\left[(\emptyset_{igt}\emptyset_{ig,t-12})\sqrt{\boldsymbol{v}(\hat{Y}_{igt,B})\boldsymbol{v}(\hat{Y}_{ig,t-12,B})}\boldsymbol{\gamma_{12,ig,B}}\right] + \left[(1-\emptyset_{igt})(1-\emptyset_{ig,t-12})\sqrt{\boldsymbol{v}(\hat{Y}_{igt,T})\boldsymbol{v}(\hat{Y}_{ig,t-12,T})}\boldsymbol{\gamma_{12,i,T}}\right]}{\hat{Y}_{igt}^{C}\hat{Y}_{ig,t-12}^{C}}$$

where

$\boldsymbol{\gamma_{12,ig,B}}$ is the lag-12 autocorrelation of the Bottom up estimator in industry *i* and state *g*, estimated from the imputed MU establishment data from the current statistical period; and

$\boldsymbol{\gamma_{12,i,T}}$ is the lag-12 autocorrelation of the Top down estimator in industry *i*, estimated from MRTS sales data.

Lag-12 autocorrelations (γ_{12}) are defined as

$$\gamma_{12} = \frac{\sum_{t=1}^{t=T-12}(Y_t - \bar{Y}_t)(Y_{t+12} - \bar{Y}_{t+12})}{\sum_{t=1}^{T}(Y_t - \bar{Y}_t)^2} = \frac{\sum_{t=1}^{T-12}\widehat{Cov}\,(Y_t, Y_{t+12})}{\sum_{t=1}^{T}\hat{v}\,(Y_t)}$$

given measurements $Y_1, Y_2, \ldots, Y_T$ from times *t* = 1, 2, …, *T* on the same units.

Both the TD and BU autocorrelation estimates include the same units in all periods. The two sets of monthly autocorrelations use *T* = 24 (12 pairs of covariances, 24 variance estimates) and are calculated monthly from January 2020 onward. Previous statistical periods use the averaged values from January 2020 through March 2020; the component estimates are extremely stable over this period (as expected) and did not appear to be subject to the pandemic effects of April 2020 and May 2020.

4.1. Top Down Lag 12 Autocorrelations

The covariance term in the lag-12 TD autocorrelation estimate is obtained for tabulation industry *i* at statistical period *t* as

$$\widehat{cov}(\hat{Y}_{it}^{m}, \hat{Y}_{i,t+12}^{m}) = \frac{1}{2}\left(v\left(\hat{Y}_{it}^{m}\right) + v\left(\hat{Y}_{i,t+12}^{m}\right) - v(\hat{Y}_{i,t+12}^{m} + \hat{Y}_{it}^{m}\right),$$

where $\hat{Y}_{it}^{m}$ is the Horvitz-Thompson estimate of monthly retail sales at time *t.* All stratified simple random sample variances are obtained using PROC SURVEYMEANS (SAS Institute Inc. 2016), where *m =* the set of all active eligible MRTS tabulation units in all *t*=1, 2, …,24 statistical periods (i.e. the intersection, not the union). Respondents and nonrespondents are included in all calculations with nonrespondent data set to missing. Tabulation industry is treated as a domain estimate to reflect the variability due to random sample sizes in respondents caused by the matching process and by the random response status. The variance and covariance estimates include a pseudo "finite population correction factor": the numerator is computed from the sampled units that <u>responded</u> in all statistical periods and the denominator is computed as the sum of the sampling weights assigned to the units in the current statistical period.

4.2. Bottom Up Lag 12 Autocorrelations

The variance and covariance terms the bottom up estimator lag 12 autocorrelation estimate at time *t* in tabulation industry *i* and state *g* are obtained using PROC CORR (SAS Institute Inc. 2016) applied imputed monthly retail sales values for MU establishments for the 24 consecutive statistical periods.

5. Quality Metrics

5.1. Statistical Quality Metrics of Composite estimator (Total and Trend)

Four metrics are produced for each state-level estimate of monthly retail sales within tabulation industry *i:*

Phi ($\emptyset_{igt}$): compositing factor, representing the percentage of composite estimator obtained from the Bottom Up estimator (see Section 3)

Coefficient of variation (c.v.), also known as the relative standard error: The c.v. of an estimator $\hat{\theta}$ is given as

$$cv(\hat{\theta}) = \frac{\sqrt{\hat{v}(\hat{\theta})}}{\hat{\theta}}$$

At the 10% significance level (the U.S. Census Bureau standard), there is no evidence that monthly retail sales totals and percentage change estimates whose c.v.'s are greater than 1/1.645 (≈ 0.67) are statistically different from zero.

90% Confidence Limits: The lower and upper (normal theory) confidence limits of an estimator $\hat{\theta}$ are given as

$$(\hat{\theta} - 1.645\sqrt{\hat{v}(\hat{\theta})}, \hat{\theta} + 1.645\sqrt{\hat{v}(\hat{\theta})})$$

At the 90% confidence level (= 10% significance level), if the confidence interval of a monthly retail sales total or of a percentage change estimate includes zero, then there is no evidence that these estimates are statistically different from zero. If the confidence interval of a trend ratio includes one, then there is no statistical evidence of a change in the year-to-year trend.

Margin of Error (MOE): $1.645\sqrt{\hat{v}(\hat{\theta})}$

The MOE (also known as the confidence interval half width) provides a measure of the variability of the point estimate.

5.2. Data Composition: Proportion of Estimate from Directly Collected Data

The proportion of the BU estimator obtained from directly collected data is computed as

$$P_{igt,B} = ZY_{igt}/\hat{Y}_{igt,B}$$

where $ZY_{igt} = [Z_{igt} + Y_{igt,\ MU} + Y_{igt,\ SU}]$ as defined earlier in Section 3.2.1. This metric is computed for industry by state estimates of monthly retail sales and is useful for providing information on the percentage of computed data as well as for evaluating the value of the compositing factor.

The proportion of the composite estimator obtained from directly collected data is for all monthly composite estimates, at the individual industry by state level and at aggregated levels as

Tabulation Industry by State: $P_{igt,C} = \emptyset_{igt}\left[ZY_{igt}/\hat{Y}_{igt,B}\right] = \phi_{igt}P_{igt,B}$
Tabulation Industry (Aggregated Over States): $P_{it,C} = \sum_g \emptyset_{igt}\, ZY_{igt}/\sum_g \hat{Y}_{igt}^{C}$
State (Aggregated Over Tabulation Industries): $P_{gt,C} = \sum_i \emptyset_{igt}\, ZY_{igt}/\sum_i \hat{Y}_{igt}^{C}$
Grand Total**:** $P_{t,C} = \sum_i \sum_g \emptyset_{igt}\, ZY_{igt}/\sum_i \sum_g \hat{Y}_{igt}^{C}$

The contribution of directly collected data is down-weighted by $\emptyset_{igt}$; the top down estimator component does not include any directly-collected industry by state level monthly sales data.

References

Kim, J.K. and Rao, J.N.K. (2009). A unified approach to linearization variance estimation from survey data after imputation for nonresponse. *Biometrika*, **96(4)**, pp. 917-932.

Morris, M., Wheeler-Martin, K., Simpson, D., Mooney, S.J., Gelman, A., and DiMaggio, C. (2019). Bayesian hierarchical spatial models: Implementing the Besag York Mollié model in stan. *Spatial and Spatio-temporal Epidemiology* 31, ISSN 1877-5845, https://doi.org/10.1016/j.sste.2019.100301.

R Core Team (2017). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

SAS Institute Inc. (2016). SAS/STAT® 14.2 User's Guide. Cary, NC: SAS Institute Inc.

Vink, G., & van Buuren, S. (2014). Pooling Multiple Imputations when the Sample Happens to be the Population. *arXiv preprint arXiv:1409.8542*.

Wood, S. N. (2001). Minimizing Model Fitting Objectives That Contain Spurious Local Minima By Bootstrap Restarting. Biometrics 57, 240–244.

TABULATION INDUSTRY	ESTIMATION INDUSTRY	DESCRIPTION
441	4411	Automotive Dealers
	4412	Other Motor Vehicle Dealers
	4413	Automotive Parts, Accessories, and Tire Stores
442	4421	Furniture Stores
	4422	Home furniture stores
443	4431	Electronics and Equipment Stores
444	44411	Home Centers
	44412	Paint and Wallpaper Stores
	44413	Hardware Stores
	44419	Other Building Supply Dealers
	44421	Outdoor Power Equipment Stores
	44422	Nursery, Garden Center, and Farm Supply Stores
445	4451	Grocery Stores
	4452	Specialty Food Stores
	4453	Beer and Liquor Stores
446	44611	Pharmacies and Drug Stores
	44612	Cosmetics, Beauty Supplies, and Perfume Stores
	44613	Optical Goods Stores
	44619	Other Health and Personal Care Stores
447	4471	Gasoline Stations
448	4481	Clothing Store
	4482	Shoe Store
	4483	Jewelry, Luggage, Leather Goods Stores
451	45111	Sporting Goods Stores
	45112	Hobby Stores
	45113	Sewing, Needlework and Piece Goods stores
	45114	Musical Instruments and Supplies
	45121	Book Stores
	45122	News Dealers and Newsstands
452	452111	Department Stores(Except Discount Department Stores)
	452112	Discount Department Stores
	4529	Other General Merchandise Stores (incl. Warehouse Clubs and Supercenters)
453	4531	Florists
	45321	Office supply and stationary stores
	45322	Novelty and gift stores
	4533	Used Merchandise stores
	45391	Pet and Pet Supplies Stores
	45392	Art Dealers
	45393	Manufactured (Mobile) Home Dealers
	45399	Other Miscellaneous Store Retailers

NAICS3	Imputation NAICS	Third Party Aggregate Data	Third Party Establishment Data	General Model	April Model	May Model	June Model
441		Yes	No	reg0 state naics			
442		No	No	reg0 state naics	zero reg0 state naics	zero reg0 state naics	
443		No	No	reg0 state naics	zero reg0 state naics	zero reg0 state naics	
444		No	No	reg0 state naics			
445		No	No	reg0 state naics	zero reg0 state naics	zero reg0 state naics	
446	44611	Yes	No	reg0 state			
	No 44611	Yes	No	reg0 state naics	zero reg0 state naics		
447		No	No	reg0 state			
448		Yes	Yes	reg1 state	zero reg0 state	zero reg0 state	zero reg1 state
451		Yes	Yes	reg2 state	zero reg0 state	zero reg0 state	zero reg2 state
452		Yes	Yes	reg1 state	zero state	zero reg0 state	
453		Yes	No	reg0 state naics	zero reg0 state naics	zero reg0 state naics	

Unless specified in the April, May, or June column, then the General Model is used for all months.

- "Reg<breaking points>" indicates the number of breaking points included in the linear regression imputation component $\alpha_i + \beta_i x_{i(dgk)} + \varepsilon_{i(dgk)}$ from in equation (10).
- "State" and "naics" indicate the addition of respective random effects in equation (9), $\tau_{i(g)}$ and $\delta_{i(d)}$ respectively
- "Zero" indicates usage of the two-stage imputation model described in equations (3) in Section 3.2.2.