Modelling for Estimation: State-Level Expenditure Estimates

Organisation:	U.S. Bureau of Labor Statistics
Author(s):	Yezzi Lee & Clayton Knappenberger
Date:	27.09.2021
Version:	1.0

1. **Background** The U.S. Bureau of Labor Statistics (BLS) collects household expenditure data through the Consumer Expenditure Surveys as an input for our Consumer Price Index. These data are collected to be representative of the entire nation, but users understandably want sub-national spending estimates. One of the most frequent requests we get is for spending at the state-level, but so far our program has only been able to provide state-level spending estimates for 5 U.S. states: California, Florida, New York, New Jersey, and Texas. However, the existing approach is limited and BLS wants to explore other options for providing users with this data. BLS wants to use the ML 2021 project as an opportunity to explore a different method for estimating state-level spending using machine learning based model-assisted estimates that we believe will allow us expand the set of state-level estimates we can provide to our users.

2. Data

2.1 Input data - developing model-assisted state expenditure estimates requires us to merge several different datasets. At this stage we have explored two sources of auxiliary data, and brief descriptions of these datasets are provided below. Future progress reports may highlight different datasets if we find datasets that better meet our needs.

A) The Consumer Expenditure Survey contains expenditure, income, assets and liabilities, and demographics data for a nationally representative sample of U.S. households. We will be using expenditure data and linking them to various auxiliary datasets. The Public-Use Microdata is useful for getting an idea of what the data look like. For this study, we used 2017 – 2019 Consumer Expenditure Surveys' Quarterly Interview.

B) The Census Planning Database provides housing, demographic, and socioeconomic characteristics for every Census Tract in the U.S. The data from decennial censuses and from five-year averages of the largest household survey conducted in the U.S. The Response Outreach Area Mapper is also useful for getting an idea of the geographic level and variables captured in this dataset.

C) Quarterly Census of Employment and Wages (QCEW) offers subnational estimates of the number of establishments, employment, and payroll for a large number of industries organized by North American Industry Classification System (NAICS) code. The data is from Census Bureau and we used 2017 – 2020 QCEW data. So far we only use the QCEW annual summary files, however we anticipate being able to use the quarterly files to more closely link expenditures to employment and wages in the relevant reference period. We began using QCEW data in lieu of the County Business Patterns (CBP) data because it includes most government employees in their estimates, while CBP does not.[1]

2.2 Data preparation - since the auxiliary datasets we are using are themselves official statistics, some data are suppressed that might identify a respondent. Missing data in these datasets come in roughly two forms:

1) Variables that are suppressed for individual observations because the true value might enable someone to identify a respondent. These cases are most often counts (i.e. the number of people in a Census Tract under the age of 5) and are suppressed because the count is below some defined threshold. We impute these by replacing the missing value with the minimum value for the variable.
2) Observations that are suppressed because the value is based on a small sample size. For example if there are five households in an area, the entire row might be suppressed and not just the count variables. In these cases we have no reason to believe that variables like average household income are necessarily any different from the average household income of other areas so we imputed them using the matrix completion method described in "Matrix Completion via Iterative Soft-Thresholded Singular Value Decomposition"[2] and implemented in the R *softImpute* package.[3]

The QCEW dataset is organized according to hierarchical industry codes (NAICS) and lower-level industries are more likely to have missing values. To mitigate this, we limit ourselves only to higher level industry codes (2-digit NAICS) and drop all lower level industry classifications.

Some variables have formatting meant to make them easier for humans to read, but which make them more difficult to store and use programmatically. Dollar denominated variables may have a dollar-sign '$' prepended and may also have commas separating digits within the number. We strip these formats from the numbers during our data cleaning process.

2.3 Feature selection - prior to modelling, we only included a subset of the variables that we thought were most likely to be useful to the models. Many of the variables we dropped were conceptually equivalent to, calculated from, or strongly correlated with another variable in the data. Roughly half of the variables in The Census Planning Database represent the margin of error for a related variable and these were also dropped from our dataset. Additionally a few variables were dropped because they had a large share of missing values. After doing this, we were left with 179 variables. From there we performed Principle Components Analysis to keep only the top component corresponding to 95% of the total variance in the features. This typically reduces the number of variables by about 70% in each year.

[1] *U.S. Bureau of Economic Analysis*, "What is the difference between BEA employment and wages and BLS and Census employment and wages?" (January 2006), www.bea.gov/help/faq/104.

[2] Hastie, Mazumder, Lee, and Zadeh, "Matrix Completion and Low-Rank SVD via Fast Alternating Least Squares," *Journal of Machine Learning Research.* 16 (2015): 3367 – 3402. https://www.jmlr.org/papers/volume16/hastie15a/hastie15a.pdf

[3] Hastie and Mazumder, "softImpute: Matrix Completion via Iterative Soft-Thresholded SVD," (May 2021), https://cran.r-project.org/package=softImpute

2.4 Output data – the models themselves output predicted expenditure amounts at the Census Tract level. We then aggregate these to state-level amounts by multiplying the predicted value by the number of households in the tract, and then adding the tracts together. At this stage, we have what are referred to as Model-Based Estimates and there is one for each state. We get to Model-Assisted Estimates by subtracting the actual value from the predicted value for each household we have collected expenditure data from, weight that residual by the number of households in the Census Tract and then add them up to get state-level values. Doing this allows us to use the survey data we collect to "assist" the estimates for areas where we do collect data. Adding the Model-Based Estimate to the survey correction gives us Model-Assisted Estimates for each state. Calculating mean expenditures for each category from the sums is as simple as dividing the sum by the number of housing units in the State.

3. Machine learning solution

3.1 Models tried So far, we have tried a Gradient-Boosting Machine (GBM), the Least Angle Shrinkage and Selection Operator (LASSO), and K-Nearest Neighbors (KNN) to estimate six expenditure categories: total expenditures, food, housing, transportation, healthcare, and entertainment. We are also using chained equations of these models to take advantage of correlations between each category when modelling.

3.2 Model(s) finally selected and quality criteria used (e.g. accuracy, time)

At this stage we have used two criteria for determining which model gives us the best results. First we perform 5-fold cross-validation to obtain Root-mean-square errors (RMSE). Second, since there are currently state-level estimates for five states, we also compared our results against these estimates and provided relative coverage rates. Based on results from prior quarters we have focused exclusively in this quarter on the GBM models.

3.3 Hardware used

Personal Computer: Intel Core i7-8665U, 1.90 GHz, 32.0 GB Installed RAM
Server: Intel Xeon Gold 6140 2.30GHz, 72 CPUs

4. Results

Since our previous progress report, we have been able to construct MAEs for all 50 States and the District of Columbia for the 2017 – 2020 time period. Figure 1 below shows a heat map of total spending for each state in each year. Unsurprisingly, California, Texas, Florida, and New York State show the highest levels of consumption spending in each year. These four states represent the largest share of the US population, have large metropolitan centers, and correspond to a large share of US GDP as well. These results are experimental, and furthermore we have reason to believe that year-to-year comparisons with these results are not very useful. Therefore, we will refrain from any providing any more detailed values about specific spending in a given year or from making any economic analysis on them.

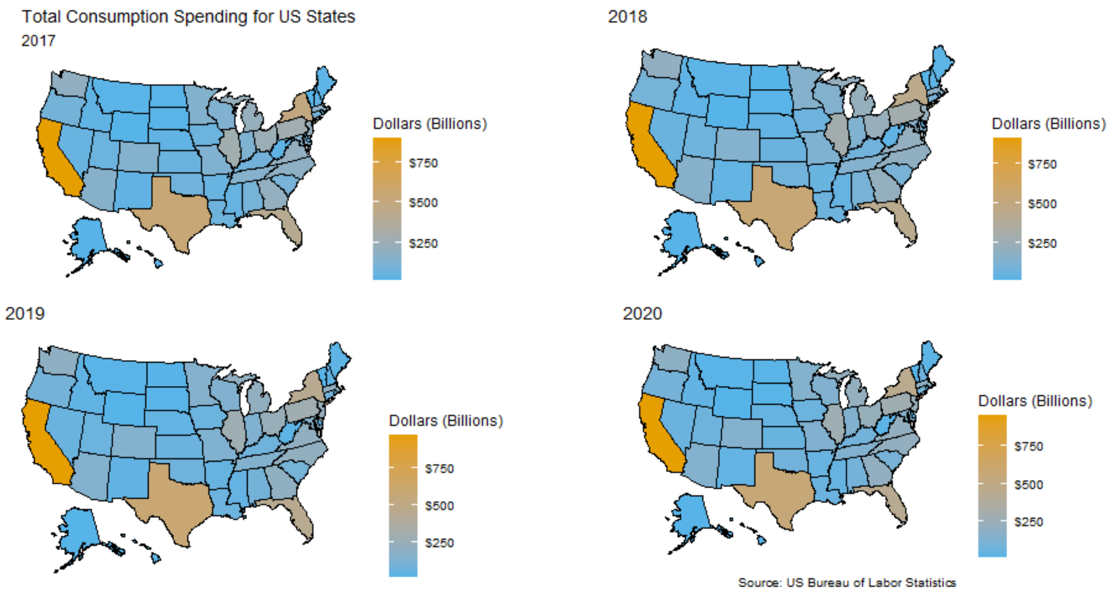

Figure 1: Total Consumption Spending for US States 2017 - 2020

Figure 2 below shows average spending on all categories by State from 2017 – 2020 and reveals a far more interesting story about the state of State consumption. In 2019, the top five states by average consumption spending were Massachusetts ($75,503), Maryland ($75,278), California ($74,073), Connecticut ($73,407), and New Jersey ($71,961). Additionally, the impacts of the COVID-19 pandemic on economic activity are also clearer when looking at averages. Three of the States which had the highest average consumption spending in 2019 were hit hard with declines of -$5,254 (Maryland), -$2,603 (Massachusetts), and -$1,098 (California). Other States from the 2019 top five saw increases in their average consumption sending of $2,543 (New Jersey) and $842 (Connecticut). This might be because these two States are close to New York City and saw COVID-19 restrictions ease up more quickly than New York City. Utah also jumped from 9th place in 2019 to 1st in 2020 with average consumption spending increasing by $5,024, and this might also be because Utah was impacted less by the COVID-19 pandemic and reopened more quickly than other States. The average consumption spending for New Hampshire ($131,569) is almost certainly an outlier.

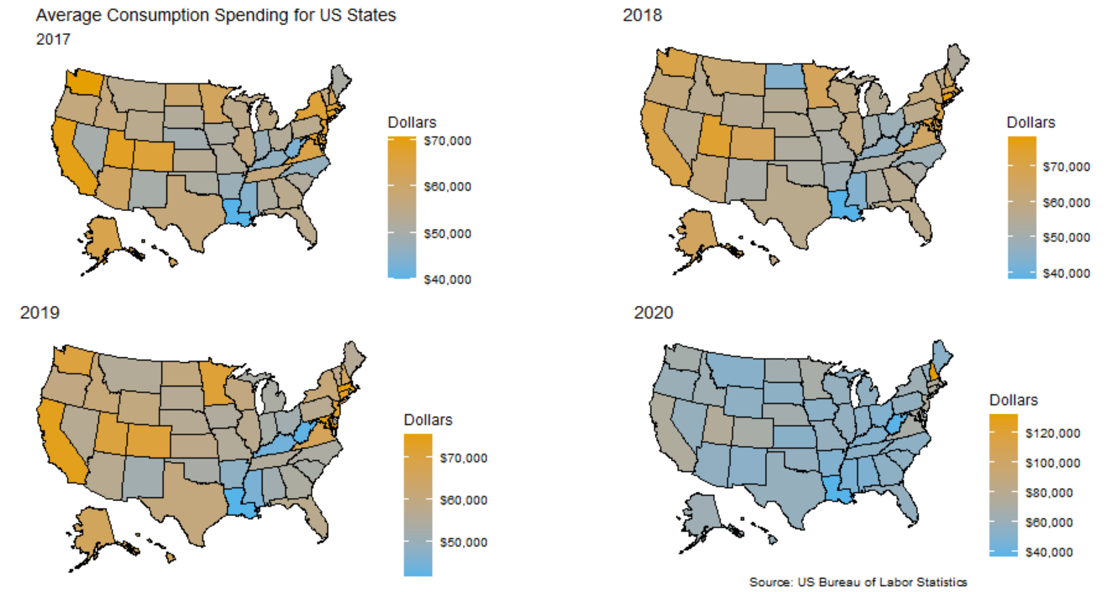

Figure 2: Average Consumption Spending for US States 2017 - 2020

Our comparison in Table 1 below takes the Model-Assisted Estimate for each model and divides it by state expenditure estimate that CE has already published. Values closer to 100% are preferred since that means the Model-Assisted Estimate is closer to the published estimate. Values over 100% indicate that the Model-Assisted Estimate is larger than the published estimate and values under 100% indicate the opposite.

Table 1: MAE to State Weight Comparisons

State	Total	Food	Housing	Transport	Health	Entertain
California						
2017	107.62%	104.64%	107.75%	113.14%	110.17%	110.17%
2018	104.88%	101.37%	107.19%	105.78%	103.94%	114.34%
2019	107.85%	103.83%	111.99%	112.95%	104.68%	113.00%
Florida						
2017	107.62%	100.54%	107.29%	116.21%	106.49%	143.19%
2018	105.50%	100.52%	107.57%	116.70%	111.17%	103.90%
2019	101.66%	98.70%	103.12%	115.58%	105.75%	98.85%
New Jersey						
2017	90.29%	93.52%	91.38%	87.95%	92.40%	89.57%
2018	93.57%	95.06%	93.38%	104.06%	101.43%	106.38%
2019	97.31%	100.16%	96.54%	101.20%	97.91%	102.36%
New York						
2017	111.40%	101.23%	103.51%	115.79%	99.98%	113.55%
2018	97.81%	96.33%	98.59%	108.40%	94.23%	95.10%
2019	98.89%	102.11%	100.06%	103.91%	103.33%	94.65%
Texas						
2017	103.48%	100.94%	104.16%	105.21%	106.34%	99.99%
2018	99.76%	100.80%	99.81%	102.52%	99.79%	100.85%
2019	99.05%	101.93%	101.78%	98.43%	97.91%	108.19%

The Q2 comparisons for sums of spending in each category ranged 26.25 percentage points from 83.89% for Transportation Spending in Florida in 2017 to 110.14% for Total Spending in California in 2017. On the other hand, the above comparisons of the mean spending in each category ranged 55.24 percentage points from 87.95% for Transportation Spending in New Jersey in 2017 to 143.19% for Entertainment Spending in Florida in 2017. This wider range obscures an important difference between the two sets of estimates. Table 2 below shows that prior estimates had a tendency to underestimate consumption spending relative to the State Weight estimate while the current estimates tend to be closer to the State Weight estimate on average estimates.

Table 2: Summary Statistics of State Weight Comparisons

Progress Memo	Minimum	1^{st} Quartile	Median	Average	3^{rd} Quartile	Maximum
Quarter 2	81.54%	89.46%	93.54%	93.44%	96.55%	110.14%
Quarter 3	87.95%	98.93%	102.44%	103.33%	107.27%	143.19%

A sign of a good estimate would be one that on average is close to the State Weights numbers, but does not consistently over or underestimate them. If you consider the 10 percentage point range between 95% and 105%, only about 34% of the Q2 estimates fall in that range. On the other hand, over half of the Q3 estimates fall in that same 10 percentage point band. See Figure 3 below for the full empirical cumulative distributions for these two quarters of estimates.

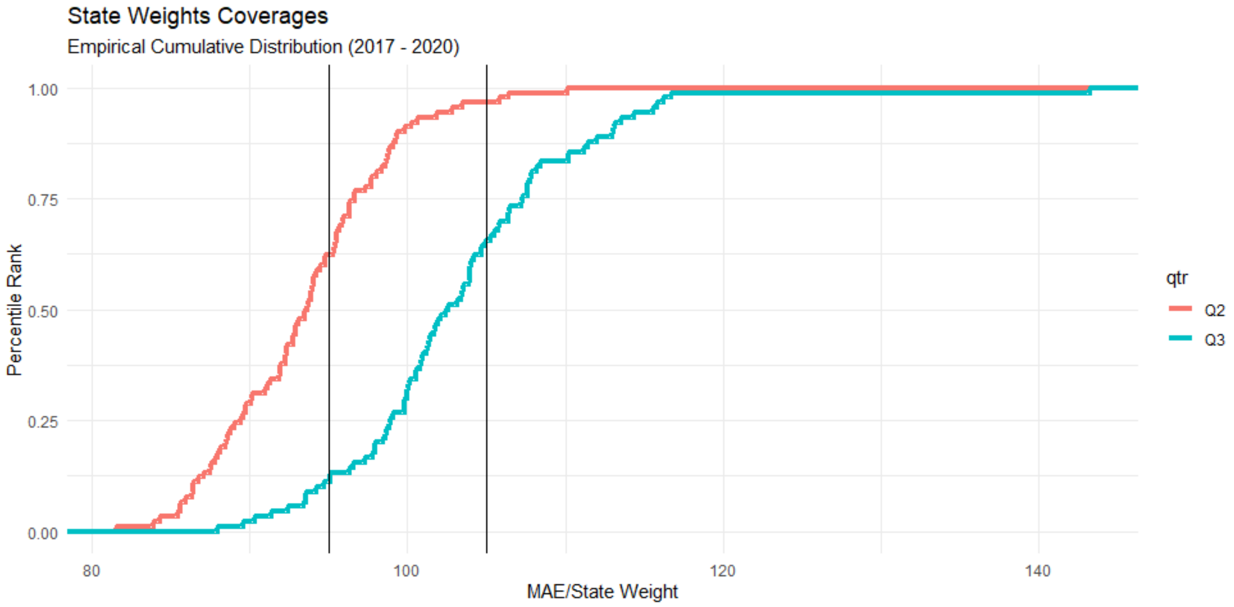

Figure 3: State Weight Coverages Empirical Cumulative Distribution

If we consider the previous estimates of state-level expenditures published by CE to be the “gold standard” estimate, then we can say that within a given year, the MAE we’ve produced is a reasonable estimate of spending in each state for each category of expenditures. However, once we start attempting to compare spending across time, the utility of our MAEs breaks down. We examined the pattern of consumption growth for the same five States over the 4 year period we have estimates for and provided those results in Figure 4 below.

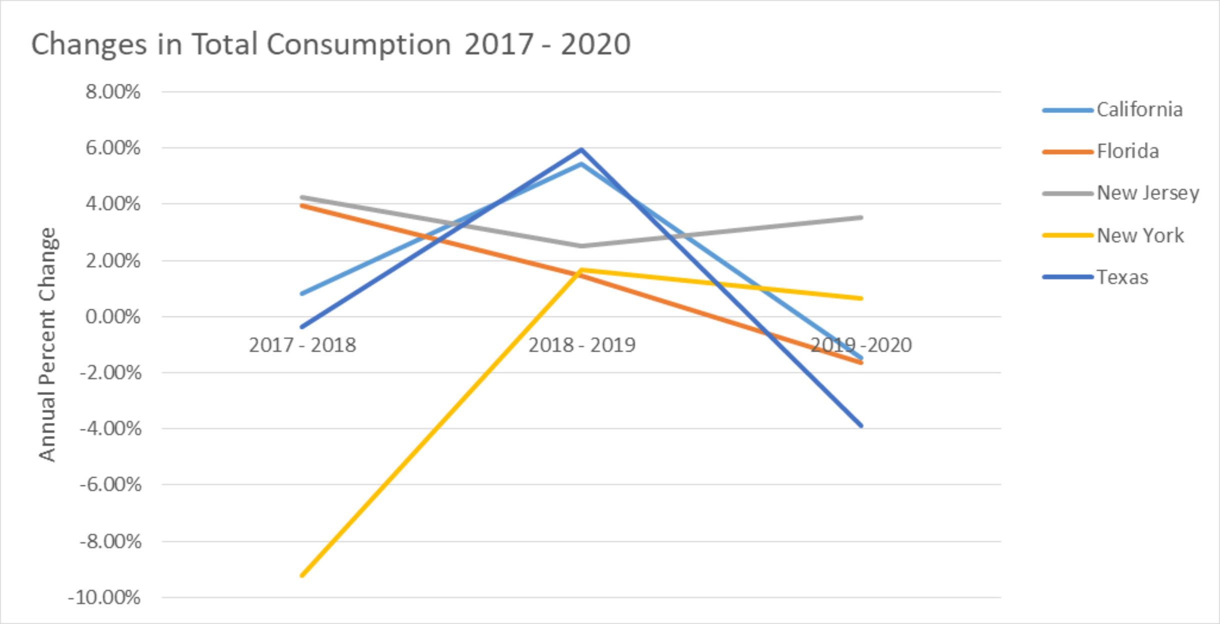

Figure 4: Changes in Total Consumption 2017 - 2020

Oddly, even as the estimates themselves compare favorably to the State Weights estimates, our MAEs show a great deal of volatility year-over-year. Based on other available economic data, we are fairly certain that average consumption spending did not decline by 9.24% New York from 2017 to 2018 and that it did not grow by 5.94% in Texas from 2018 to 2019. The results for the other States not shown here are even more dramatic than these

5. Who are the stakeholders?

Our main stakeholders are external data users and the BLS Consumer Price Index.

6. Fall Back

There are other groups working on other approaches to solve this problem of providing state-level expenditure estimates to our users. Our office has already produced estimates for five states, and we are using those to help validate our approach. Even if our approach does not end up working, CE may be able to use this approach to produce state-level of expenditure information for other, but not every state.

7. Robustness

We are trying to validate our results by comparing the estimates to several different sources. Additionally, we plan to perform these comparisons using multiple years of estimates to compare our Model-Assisted Estimates across time. Examples of different sources we can try to compare our estimates to are:

1. Existing state-level CE estimates
2. ACS state estimates for some categories of spending
3. PCE state level estimates (which are not fully comparable)

8. Next steps

Our next steps include rounding out our comparisons of our MAEs against ACS state estimates and against PCE state-level data where available.