

# A MACHINE LEARNING ANALYSIS OF SEASONAL AND CYCLICAL SALES IN WEEKLY SCANNER DATA

Rishab Guha\*  
Harvard University

Serena Ng†  
Columbia University and NBER

March 1, 2019

## Abstract

This paper analyzes weekly scanner data collected for 108 groups at the county level between 2006 and 2014. The data display multi-dimensional weekly seasonal effects that are not exactly periodic but are cross-sectionally dependent. Existing univariate procedures are imperfect and yield adjusted series that continue to display strong seasonality upon aggregation. We suggest augmenting the univariate adjustments with a panel data step that pools information across counties. Machine learning tools are then used to remove the within-year seasonal variations. A demand analysis of the adjusted budget shares finds three factors: one that is trending, and two cyclical ones that are well aligned with the level and change in consumer confidence. The effects of the Great Recession vary across locations and product groups, with consumers substituting towards home cooking away from non-essential goods. The data are thus informative about local and aggregate economic conditions once the seasonal effects are removed. The two-step methodology can be adapted to remove other types of nuisance variations provided that these variations are cross-sectionally dependent.

Keywords: Seasonal adjustment, Random forest, fixed effects, demand system.

---

\*Littauer Center, Harvard University, 1805 Cambridge Street, Cambridge, MA 02137. Email: rishab\_guha@g.harvard.edu

†Department of Economics, Columbia University, 420 W. 118 St. MC 3308, New York, NY 10027. Email: serena.ng@columbia.edu

This analysis is based in part on data from The Nielsen Company (US), LLC and marketing databases provided through the Nielsen Datasets at the Kilts Center for Marketing Data Center at The University of Chicago Booth School of Business. The conclusions drawn from the Nielsen data are our own and do not reflect the views of Nielsen. Nielsen is not responsible for, had no role in, and was not involved in analyzing and preparing the results reported herein.

We thank Gary Cornwall at the BEA, the discussant Ramon Huerta, and seminar participants at the NBER Summer Institute and Amazon for helpful comments. All errors are our own. Finally support from the National Science Foundation (Ng, SES-1558623) is gratefully acknowledged.

# 1 Introduction

A mind-boggling amount of data is now available for economists to analyze. This is made possible by improved technology in data collection and storage. The modern data differ from the conventional ones in at least two important ways. They tend not to be provided by government agencies, and they have what data scientists refer to as the “three V” characteristics: volume, variety, and velocity. Econometricians may think of them as short panels of big, often unbalanced, high-frequency, highly heterogeneous data. Such granular data can potentially allow learning about economic behavior. However, a full analysis of the data comes with unique challenges. A case in point is the weekly Nielsen Retail Scanner data collected since 2006. With roughly half a terabyte of data added each year, the database has grown in size to five terabytes in 2016.

The Nielsen dataset has three features of interest. First, it consists of real-time sales and unit price recorded at the store/upc-code level. Such transactions data are distinctly different from official price indices that are survey based. Second, the data are not subject to revisions once a transaction is completed; they are also less susceptible to measurement errors as the data are digitally recorded. Third, the data are available for the major metropolitan areas and thus provide spatial and temporal information distinct from the official monthly retail sales data. The weekly data also provide higher frequency information than that in quarterly and annual surveys. Fourth, many memorable events have occurred over the span of the sample: a big recession, destructive hurricanes, several elections, new tax initiatives, and a government shutdown. Though the weekly aspect of the data seems like it should appeal to researchers, work thus far has mostly aggregated the data to a monthly or a quarterly frequency without taking advantage of the weekly information. With just a peek of the data, one understands why: the data exhibit strong seasonal patterns that are highly heterogeneous in the product and spatial dimensions. As will be shown below, this is true not only in our base case analysis for the four most populous states, but also in an extended analyses that include more regions and states. The weekly data have limited use for business cycle analysis without a way to deconvolve the seasonal variations from the cyclical ones.<sup>1</sup>

The obvious solution is to seasonally adjust one series at a time. Unfortunately, there are few satisfactory methods for seasonally adjusting weekly data, let alone for a massive number of series. We argue below that the short span and the quasi-periodic nature of the Nielsen data make perfect adjustment of each series highly unlikely. This is problematic because in our data, counties within a state are likely to share common seasonal patterns.

---

<sup>1</sup>With some abuse of terminology, holiday effects will also be treated as seasonal variations.

Even if the residual seasonal effects are negligible at the individual series level, they can become non-trivial when aggregated across counties.

This paper develops a framework for seasonally adjusting a large panel of data in which common and idiosyncratic seasonal variations coexist. We suggest complementing univariate seasonal adjustments with a second step that pools counties within a state to remove the within year common seasonal variations, one year at a time. Our premise is that a good deal of the within year variations are highly predictable *ex-ante*. Hence we treat the within-year seasonal adjustment as a prediction problem. To find the prediction model of unknown functional form in the face of a large set of potential predictors, we use machine learning methods to perform estimation and variable selection. This by-passes the need to specify a single data generating process, which is a difficult task when the data are so highly heterogeneous. Though our approach is rather agnostic, the adjusted data are no longer dominated by seasonal effects so that insights about consumer behavior can be learned from analysis of demand systems.

In theory, Engel curves should be spanned by functions of prices and income that are common across product groups. Traditional demand analyses indirectly parameterize these latent processes by flexible functions. Consistent estimation of the underlying parameters is possible when  $T$  (the number of time periods) tend to infinity with  $N_g$  (the number of product groups) fixed. Given that  $N_g=108$  and  $T = 469$  are reasonably large in our data, we can take advantage of results developed in large dimensional factor analysis to estimate the latent functions of prices and income directly. Big data therefore provide a perspective of demand analysis that was not possible in the conventional small  $N_g$  large  $T$  setting.

Our demand analysis of the seasonally adjusted data leads to four conclusions. First, the demand systems are well described by three common factors relating to the trend, level, and curvature of Engel curves. Second, even though the data are primarily based on sales at grocery and mass merchandise stores, there is surprisingly clear evidence of cyclical spending patterns. The cyclical components move closely with measures of consumer sentiment and consumer confidence, indicating that the actions and 'feelings' of consumers are aligned. Third, an analysis of the loadings on the cyclical factor yields a 'distribution' of recession sensitivity across product groups. The budget share of a FOOD-IN basket tends to be strongly counter-cyclical while that of a LUXURY basket is pro-cyclical, consistent with evidence from the monthly CPI weights. Fourth, recession sensitivity has a spatial dimension; changes in spending on our FOOD-IN basket are larger in the metropolitan than the rural areas. We use heatmaps to illustrate the changes in FOOD-IN as the economy moves through the business cycle. Overall, the business cycle information in the scanner data seems roughly

consistent with the less granular official data. This is good news because this suggests that there is valuable higher frequency information about consumer spending at the aggregate and local levels once the seasonal variations are removed. Dealing with nuisance variations and some form of aggregation are almost inevitable in analysis of big data. The proposed two-step procedure can be adapted to other panels so long as the variations to be removed are sufficiently pervasive for pooling to be effective.

The rest of the paper proceeds as follows. We begin in Section 2 with a description of the data and highlight the presence of common seasonality. Section 3 discusses the challenges posed by cross-section dependence that seasonal adjustments have to overcome. Section 4 presents our two-step approach and elaborate how the second step is formulated as a prediction problem. Section 5 analyzes the properties of the seasonally adjusted data and documents how the different products and regions react to changing economic conditions. Section 6 concludes.

## 2 The Data

The Nielsen Retail Scanner data are collected by the Nielsen marketing group and managed by the Kilts Center for Marketing at the University of Chicago. The data have over 1000 products belonging to over 115 product groups (e.g., beer, wine, eggs) that can in turn be organized into 10 categories: dry groceries, frozen, dairy, deli meat, fresh food, non-food, alcoholic beverage, general merchandise, health and beauty. The data are heavily weighted towards groceries and mass-merchandise goods with limited coverage of consumer durables. Specifically, the products cover over 3 million universal product codes (UPCs) collected from over 35,000 participating stores in 55 (Metropolitan Statistical Areas) MSAs across the U.S. Each store reports weekly data for every UPC code that had any sales volume during the week. Nielsen uses a Saturday week-ending label to identify the week that the data are reported. We have information about the location of the retailer (but not the name), the units sold and the volume weighted average of the product for that week. Following Nielsen’s documentation<sup>2</sup>, the week’s total dollar sales is calculated as

$$\text{sales} = \frac{\text{price}}{\text{prmult}} \times \text{units}.$$

The *movement files* of the database provide data for UNITS (the number of units sold), PRICE (the volume weighted average price of the product for the week), PRMULT (a price multiplier to indicate deals such as 3 for one dollar).

---

<sup>2</sup>Kilts Nielsen Retail Scanner Dataset Manual 2014, Table 3, Part E.

We analyze the total sales of products within a product group (hereafter simple referred to as “groups”). The sales data are constructed as follows. For each state,  $s$ , we first compile a list of stores that report a sale in at least one of the 115 groups in each of the 469 weeks between Saturday, January 7, 2006 and Saturday, December 27, 2014. Restricting attention to groups with data in every week reduces the number of groups from 115 to  $N_g = 108$ . This gives a balanced panel of stores. The variables are indexed as follows:

	<b>county</b>	<b>group</b>	<b>week</b>	<b>year</b>
index	$c$	$g$	$t$	$\tau$
total	$N_c$	$N_g = 108$	$T = 469$	$N_{yr} = 9$

Group by group, we construct a measure of weekly total sales at the county level by aggregating over all stores located in each county within the compiled list. The variable of interest is  $\log(\text{SALES}_{gct})$ , the log sales of group  $g$  in county  $c(s)$  in week  $t$ <sup>3</sup>. Since a county is state-specific, the state index  $s$  will be suppressed when the context is clear. At each week  $t$ , the budget share of an arbitrary group  $g \in [1, N_g]$  is

$$\text{share}_{gt}^s = \frac{\sum_{c(s)} \text{SALES}_{gc(s)t}^s}{\sum_g \sum_{c(s)} \text{SALES}_{gc(s)t}^s} = \frac{\text{sales of group } g \text{ in state } s \text{ at week } t}{\text{total sales in state } s \text{ at week } t}$$

Our base case analysis uses data from the four most populated states in the U.S: California (CA), Florida (FL), New York (NY), and Texas (TX). We also construct a measure of total sales, labeled FOUR, that aggregates sales over the four states. Our extended case adds states from the Midwest (Illinois, Indiana, Iowa, Kansas, Michigan, Minnesota, Missouri, Nebraska, North Dakota, South Dakota, Ohio, Wisconsin), Mid-Atlantic (Delaware, Maryland, New Jersey, Pennsylvania, Virginia, and Washington D.C.), and Southwest (Arizona, Nevada and New Mexico). This sample encompasses a total of 24 states (plus D.C.) covering about 70% of the population in the U.S. with sales in 15,631 stores in 1,147 counties. Results that pool over all states in this extended data will be labeled SEVEN. The most comprehensive analysis groups the remaining states into MOUNTAIN, PACIFIC NORTHWEST, NEW ENGLAND, and SOUTH for a total of nine regions. The pooled results will be labeled ALL, and covers 24,280 stores in 2,095 counties.

The ten product groups with the largest sales in FOUR are listed in Table 1 below. The relative importance of the groups are reasonably similar across states, with bread, beer, juice, carbonated beverages, medication, and snacks making the list in each of the four states.<sup>4</sup>

<sup>3</sup>Throughout the text we use “county” to reference a specific geographic county, e.g., New York County, New York. We let  $\mathbf{s}(i)$  denote the state containing county  $i$

<sup>4</sup>We work with shares instead of sales which tend to have even stronger seasonal effects.

Table 1: Budget Shares (%): Most Purchased Product Groups

CA: $N_c = 53$		FL: $N_c = 58$		NY: $N_c = 58$		TX: $N_c = 161$	
share	description	share	description	share	description	share	description
3.4	bread	4.5	medications	4.2	medications	3.8	carbon. bev
3.4	beer	4.3	tobacco	3.3	fresh produce	3.7	medications
3.3	juice	3.1	carbon. bev.	3.2	bread	3.4	snacks
3.2	wine	2.9	liquor	3.1	candy	3.0	bread
3.1	fresh produce	2.8	beer	2.9	snacks	2.8	tobacco
3.1	carbon. bev	2.7	juice	2.8	juice	2.7	packaged meat
3.0	snacks	2.7	candy	2.6	tobacco	2.6	candy
2.8	packaged meat	2.5	snacks	2.6	beer	2.6	fresh produce
2.7	salad dressing	2.3	milk	2.4	carbon. bev	2.6	juice
2.7	medication	2.3	bread	2.3	milk	2.5	beer

A more systematic analysis of the data requires a framework. We appeal to demand theory which also forms the basis of price indexes and measures of cost of living. A (product group based) demand system expresses  $N_g$  budget shares in terms of  $r$  functions of prices  $p = (p_1, \dots, p_{N_g})'$ , and income  $Y$ . Hence we may write  $\text{share}_g^s = \sum_{k=1}^r \lambda_{gk}^s(\log p) F_k^s(\log p, \log Y)$ . Importantly, the functions  $F = (F_1, \dots, F_r)$  are common across groups. The adding up constraint requires that  $F_1^s$  is a constant.<sup>5</sup> The value  $r$  is the dimension of the space spanned by Engel curves and is known in the literature as the rank of a demand system. A rank one system occurs when budget shares are independent of the level of income, in which case all income elasticities equal one. Rank two demand systems are linear in log-prices but not in log income. Examples include the translog and the linear expenditure system. Many rank two systems belong to the PIGLOG class discussed in Muellbauer (1975). Quadratic Engel curves can be rank two or rank three. Gorman (1981) shows that exactly aggregable demand systems must have rank no larger than three.

Product-based demand systems were commonly estimated in the 1970s and 80s to obtain price elasticities and to understand substitutability across products until characteristic-based demand systems became popular. A demand analysis typically proceeds by using flexible functions to approximate the expenditure function. Imposing the axioms of demand theory then allows the shares or sales to be expressed as linear functions of prices, income, and a theoretical price index, say  $P$ . Under two stage budgeting, income can be replaced by total expenditure on the  $N_g$  groups. In empirical work, a proxy variable  $P^*$  that can be constructed

<sup>5</sup>For a discussion on the rank of demand systems, see Lewbel (1991).

prior to estimation is often used to bypass the cross-equation restrictions imposed by the ideal price index  $P$ . For example, the Almost Ideal Demand System (AIDS) of Deaton and Muellbauer (1980) uses Stone’s price index defined by  $\log P_t^* = \sum_{g=1}^{N_g} \text{share}_{gt} \log(p_{gt})$ . Given data for  $N_g$  shares and prices, the AIDS regression model is

$$\text{share}_{gt}^s = \lambda_{0g}^s + \sum_{j=1}^{N_g} \lambda_{jg}^s \log p_{gt}^s + \beta_g^s \log(Y_t^s / P_t^{s*}) + e_{gt}^s, \quad t = 1, \dots, T.$$

The term  $e_g^s$  can be due to measurement error or anything that shifts spending for reasons other than changes in prices and income, such as omitted time variation in preferences. In cross-section analysis,  $T$  would be the number of households whose spending patterns are recorded. In time series analysis,  $T$  would be the number of observations on aggregate spending over long periods of time. With panel data, the same household may be observed more than once. Using data with  $N_g$  small and  $T$  large, the rank of demand systems is typically estimated to be two or three, and at most four.<sup>6</sup>

We are interested in analyzing all product groups in the Nielsen data available, which is well over one hundred in number, not five or six. A large  $N_g$  may appear to hinder analysis at first glance. But because  $N_g$  and  $T$  are both large, we may deviate from traditional demand analysis to let the shares data identify the space spanned by the latent functions and its dimensionality without directly using data on prices or of  $P^*$ , or make approximations of the expenditure function. To do so, consider the factor representation of the budget shares:

$$\text{share}_{gt}^s = \lambda_g^{s'} F_t^s + e_{gt}^s$$

where  $F_t^s$  is a  $r \times 1$  vector of latent factors and  $\Lambda_g^s$  is the corresponding vector of factor loadings. Appealing to theoretical results in the literature for large dimensional factor analysis, we can consistently estimate the factors and the loadings by applying the method of principal components to the shares data alone, see Bai and Ng (2008).

In implementation, we take a three week rolling average of budget shares so as to smooth out the variations due to temporary promotional sales. Principal components are then estimated from the standardized data.<sup>7</sup> In  $s=(\text{NY}, \text{TX}, \text{and FOUR})$ , the largest factor  $\hat{F}_1^s$  explains over 0.9 of the variations of PACKAGED-MILK. As egg-nog is one of the products in the group, we may think of  $\hat{F}_1^s$  as a *Christmas* factor. Other product groups also exhibit recurring patterns towards the end of the year. In CA for example, the share of juice takes a big dip in week 51, while in FL, the share of hair-care products bottoms around week 51.

<sup>6</sup>See, for example, Lewbel (2003, 1997); Banks et al. (1997).

<sup>7</sup>Since our data are demeaned, the constant factor is controlled for prior to estimation.

Figure 1 plots the first four factors from the pooled data FOUR. The factors are only identified up to sign, so they are plotted to be pro-cyclical. Recalling that the factors are mutually orthogonal by construction, Figure 1 indicates the presence of a multitude of seasonal effects. Though all four factors have spikes around week 48, the exact week of the spike is different over years and across factors. Indeed, the spectrum of these factors peaks around but not exactly at the seasonal frequency of  $\frac{2\pi j}{52} = .1208j$  for  $j \geq 1$ . Though the first two factors are strongly periodic,  $\hat{F}_3$  and  $\hat{F}_4$  appear somewhat cyclical. Evidently, cyclical and seasonal common factors co-exist.

The criterion of Bai and Ng (2017)<sup>8</sup> finds five factors in CA, FL, NY, four factors in TX, and five factors in FOUR. In all cases, the first factor explains about one-third of the variations in the data, the first two factors together explain just under sixty percent, while four factors explain around seventy-five percent of the variations in budget shares. Taking into account that we demeaned the data before estimation by principal components, the actual rank is one larger than reported above, making the rank of the demand systems in the Nielsen data about twice as large as the estimates typically reported in the literature. In the next section, we attribute seasonality to be a reason for this finding.

### 3 Seasonality and Cross Section Dependence

Consumer theory suggests that it is desirable to smooth consumption over time. But in reality, spending is uneven over the course of a year. It tends to be concentrated around holidays, special events, and towards the last six weeks of the year. In addition to 'egg-nog' whose sales peak around Christmas, sales of stationary and school supplies peak around week 36. Sales of 'cough and cold' products are higher during the winter months while ice-cream sales are higher in the summer months. Flower sales are higher around Valentine's Day and Mother's Day than the rest of the year. Beer sales tends to be highest around July fourth while wine sales are higher around Thanksgiving and Christmas. The point to highlight is that such seasonal sales tend to recur every year though not necessarily on the same day or even the same week. Furthermore, for many of the product groups, the seasonal pattern is similar irrespective of location.

The challenge that seasonality poses for factor analysis is that principal components can identify pervasive variations but is blind to the source of pervasiveness. Hence in the presence

---

<sup>8</sup>The criterion is defined as  $\bar{r}^s = \min_{k=0, \dots, r_{\max}} \log(1 - \sum_{j=1}^k (\sigma_j^s - \gamma)_+^2) + k \cdot \text{PENALTY}(N, T)$ . where  $\text{PENALTY}(N, T) = \frac{(N+T)}{NT} \log(\frac{NT}{N+T})$  and  $\sigma_j$  is the  $j$ -largest eigenvalue in a  $T \times N$  panel  $Z = \frac{X}{\sqrt{NT}}$ , where  $X$  is the given panel of standardized budget shares. A regularization parameter of  $\gamma = 0.05$  is used to penalize common variations due to outliers. The the maximum number of factors is set to 10.



of strong common seasonality across product groups, the dominant principal components can be unrelated to cyclical economic conditions. This being the case in our data, a natural approach would be to seasonally adjust each series prior to demand estimation. A variety of univariate methods are available to de-seasonalize monthly and quarterly data, the most notable being X13ARIMA-SEATS and TRAMO-SEATS. However, these methods do not seem appropriate for data with the three-V features. For one thing, although we have 469 weekly observations, they only span nine years, meaning that we only have nine data points from a seasonal perspective. Time series seasonal adjustment methods typically assume that we have a large number of observations at the seasonal frequency of interest. Even if an ideal seasonal filter were available, finite sample bias is unavoidable since the span of our data is reasonably short.

Weekly data pose an additional challenge because weekly variations are not exactly periodic. For example, Thanksgiving and Christmas do not always fall on the same numbered week of the year, and July 4th is sometimes in week 26 and sometimes 27. This a consequence of the fact that we are on a Gregorian calendar. We cannot ‘difference away’ the seasonal effects like we could with monthly and quarterly data. Week-of-year and day-of-year differencing mitigate the problem to some extent, but it cannot capture events that occur on different day of the year, the most difficult to handle being Easter. In our sample, Easter was as late as April 24 in 2011, and as early as March 23 in 2008. Further complicating the problem is that these events have differential impact depending on the product in question and location of the sale. A ‘one size fit all’ seasonal filter is unlikely to ever exist.

The literature for adjusting weekly data is quite sparse. Notable exceptions are the fully parametric state-space analysis of Harvey et al. (1997) and the non-parametric approach of Pierce et al. (1984), Cleveland and Scott (2007), Cleveland et al. (2014). Structural time series modeling requires careful choice specification of the model for the series under investigation. The non-parametric approach is to approximate the seasonal component by basis functions such as trigonometric series. Cleveland et al. (2014) suggests to control for weekly effects, holiday effects, and outliers using locally-weighted regressions and applies the method to unemployment income claims and steel production data. But our data have several features that are distinct from these series.

First, the Nielsen sales data tend to be ‘spiky’. For many groups, the spikes only occur once per year, usually around Black Friday. For other series, the spikes can be observed a few times a year and can be attributed to temporary sales. Spikes are problematic as they tend not to be well approximated by non-parametric regressions that are smooth by design. As noted above, we use a three-week rolling average of the data in demand estimation, but

this may not be enough to annihilate the problem. Second, some variations do not repeat over the course of the year. Instead, they repeat in reference to a date  $t$ 's position within the *month*. As an example, sales increases around food stamp distributions, or end-of-month price changes. Strictly periodic functions based on fixed positions within the year may be too restrictive for these variations.

A third characteristic of our data is the volume. We have not one, but a large number heterogeneous and short time series that need to be adjusted. It would be quite impossible to write down a parametric model that would suit all series. It seems unrealistic to expect any statistical procedure to be able to completely de-seasonalize every series in the panel.

The possibility that a conventionally adjusted series will likely have some residual seasonality has implications for any analysis that involves aggregation of the individually adjusted series. Consider an arbitrary variable  $Z$  that has a seasonal and a non-seasonal component:

$$Z_{gct} = Z_{gct}^{nseas} + Z_{gct}^{seas}.$$

In our case,  $Z_{gct}$  is normalized sales of product group  $g$  in county  $c$  at  $t$ . Let  $\widehat{Z}_{gct}^{seas}$  be some  $\sqrt{T}$  consistent univariate estimate of the seasonal component of  $Z_{gct}$ . The seasonal adjustment error can be decomposed into a term  $\widehat{e}_{gct}^{seas}$  that is uncorrelated across counties  $c$ , and a term  $\widehat{\Phi}_{gct}$  that is correlated across  $c$ , viz

$$\widehat{Z}_{gct}^{seas} - Z_{gct}^{seas} = \widehat{\Phi}_{gct}^{seas} + \widehat{e}_{gct}^{seas}.$$

Aggregating the data over counties, we have

$$\widehat{Z}_{gt}^{nseas} = \sum_{c=1}^{N_c} Z_{gct}^{nseas} + \sum_{c=1}^{N_c} \widehat{\Phi}_{gct}^{seas} + \sum_{c=1}^{N_c} \widehat{e}_{gct}^{seas}$$

While  $\sum_{c=1}^{N_c} \widehat{e}_{gct}^{seas}$  tends to zero as  $N_c \rightarrow \infty$ , the sum of  $\widehat{\Phi}_{gct}^{seas}$  over  $c$  may not be mean zero. Chamberlain (1984) pointed out that Euler equation errors that are mean zero over time need not be mean zero in the cross-section dimension if the units face common shocks. For a similar reason, the seasonal variations left-over from an imperfect univariate adjustment can survive aggregation in the presence of common seasonality. This is relevant because we aggregate over counties to obtain total sales for the product group in the state. Since sales in neighboring counties will likely have similar seasonal patterns, aggregation will likely preserve the common seasonal component. Univariate seasonal adjustments yield group level sales data that may be better characterized by model with two distinct types of common factors, one that is seasonal, one that is not, and the seasonal ones dominate. This would be consistent with the finding in Ng (2017) that the principal components of budget shares constructed from data adjusted from bottom up continue to exhibit seasonal variations.

## 4 Seasonal Adjustment as a Prediction Problem

Aggregation of the seasonally adjusted data will not, in general, be the same as seasonal adjustment of the aggregate data. If we are only interested in the aggregate series, direct seasonal adjustment of the aggregate series might well be the simplest approach. But when the county and the group level seasonally adjusted information are both of interest, as is the case here, there is no choice but to perform seasonal adjustment from the bottom up, one (county, group) pair at a time. But the foregoing discussion suggests that existing filters will likely leave residual seasonal variations in the adjusted data. Our proposed approach is to complement the univariate adjustments with an additional step to 'mob-up' the residual seasonality prior to aggregate analysis.

To motivate our approach, note first that if there is commonality in seasonal patterns, it would seem inefficient to seasonally adjust each series in isolation. Seasonality is in fact a common feature in the sense of Engle and Kozicki (1993), but there is little work in this dimension. Geweke (1978) suggests that a multivariate adjustment might dominate a univariate adjustment in a mean-squared error sense, but the population analysis assumes that the model is correctly specified and abstracts from model and sampling uncertainty. McElroy (2017) considers a multivariate procedure in a large  $T$ , small  $N_g$  setting. Fok et al. (2007) considers a large  $T$ , large  $N_g$  panel of data and uses a hierarchical Bayes method to avoid the proliferation of dummy variables needed to control for seasonal fixed effects. Like Fok et al. (2007), we also pool information across clusters and over time. But instead of treating all dummy predictors as relevant, we train machine learning algorithms to determine which ones to use, and how they are to be used. In other words, we treat a large  $N_g$  as a big data blessing. Furthermore, we pool the data across counties and perform adjustment year by year while allowing the prediction model to differ every year.

### 4.1 A Two-Step Panel Approach

In time series analysis, seasonal variations are those that recur with seasonal periodicities. For example, monthly variations are those that recur every twelve months. However, as discussed above, weekly variations are not strictly periodic. This motivates us to use a definition of seasonality does not depend on periodicity.

Recall that for each state  $s$ , we have county level data over 469 weeks, and group level sales is the sum over sales in the counties. Our maintained assumption is that sales in the same group  $g$  collected in different counties  $c$  share common seasonal patterns over the course of a year. In other words, two neighboring counties share seasonal pattern even if one county

has ten times as many sales as the other. As some counties are much larger than others, we demean the data year-by-year to remove the size effect. We further standardize the data to ensure scale-independence across years and locations. Normalized sales within each year and each county is defined as:

$$y_{gct} = \frac{\log(Z_{gct}) - \mu_{pc\tau}}{\sigma_{gc\tau}}, \quad \tau = \text{yr}(t)$$

where  $\mu_{pc\tau}$  denotes the mean of log sales of group  $g$  in county  $c$  over the year  $\tau$  containing week  $t$ , and  $\sigma_{gc\tau}$  is the corresponding standard deviation. The within-year normalization isolates within-year seasonal patterns while preserving long-term trends in aggregate sales and volatility, which the econometrician can model separately. The normalization also allows us to pool observations across counties in subsequent estimation. Pooling county level data compensates for the relatively short-time span of data for each county.

Next, we posit that  $y_{gct}$  has three components: a group specific seasonal component, a common seasonal component, and a cyclical component:

$$\begin{aligned} y_{gct} &= (\text{county specific seasonal sales}) + (\text{common seasonal sales}) + (\text{non-seasonal sales}) \\ &= d_{gct} + q_{gct} + u_{gct}. \end{aligned}$$

In this decomposition, the seasonal component of sales is  $d_{gct} + q_{gct}$ . The goal is to extract  $u_{gct}$  when only  $y_{gct}$  is observed. An overview of the estimation methodology is as follows:

**Step 1: Estimation of  $d_{gct}$**  : For each  $(g, c)$  pair, perform time series estimation of

$$y_{gct} = \alpha_{gc}^0 + \text{Fourier}_{gct}(\beta_{gc}, \psi_{gc}) + \epsilon_{gct}$$

where with  $\delta_{tj} = 2\pi j \frac{\text{day of year}_t}{\text{days in year}}$  and  $m_{tj} = 2\pi j \frac{\text{day of month}_t}{\text{days in month}}$ ,

$$\begin{aligned} \text{Fourier}_{gct} &= \sum_{j=1}^{p_d} \beta_{1,gcj} \sin(\delta_{tj}) + \beta_{2,gcj} \cos(\delta_{tj}) \\ &\quad + \sum_{j=1}^{p_m} \psi_{1,gcj} \sin(m_{tj}) + \psi_{2,gcj} \cos(m_{tj}). \end{aligned}$$

The variables  $\delta_{tj}$  and  $m_{tj}$  are strictly periodic. The regression only includes an intercept and will preserve any trends in sales that might be in the data. This non-parametric time model is the deterministic seasonal component specified in Pierce et al. (1984) to seasonally adjust weekly data. Hereafter, we will refer to Step 1 as the Fourier regression.

Let  $\hat{\epsilon}_{gct} = \widehat{q_{gct} + u_{gct}}$  be the least squares residuals from Step 1. Because this step is based on estimation of a smooth regression, these residuals will have spikes. Furthermore, as pointed out above, the leftover seasonality in  $\hat{\epsilon}_{gct}$  may be negligible at the series level but can be non-trivial at the group level. To proceed, we assume that (i)  $d_{gct}$  and  $q_{gct}$  are partially predictable *over the course of a year*, and (ii)  $q_{gct}$  has variations that are common across counties. Hence we exploit cross-section dependence amongst counties to remove the within-year seasonal variation. This leads to the next step.

**Step 2: Estimation of  $q_{gct}$  from  $\hat{\epsilon}_{gct}$ :** We want to recover the common seasonal component  $q_{gct}$  by pooling information across counties. This is done by entertaining a large number predictors, which are mostly dummy variables. To alleviate the problem of overfitting, we use machine learning algorithms to pick out the most important predictors, leaving the functional form of the model unspecified. Details will be explained in the next subsection.

Let  $\hat{q}_{gct}$  be the prediction of the common seasonality in  $\hat{\epsilon}_{gct}$  obtained from Step 2. From these, we can obtain an  $\hat{u}_{gct}$ , an estimate of  $u_{gct}$ . A seasonally-adjusted value of log-sales is obtained by plugging the estimated residual  $\hat{u}_{gct}$  into

$$\widehat{y_{gct}^{sa}} \equiv \hat{u}_{gct} \cdot \sigma_{g\tau} + \mu_{g\tau} \quad (1)$$

The log seasonally-adjusted series has the intuitive interpretation of being the unpredictable part of the series's variation around its overall mean for the year. An estimate of seasonal adjusted sales is  $\exp(\widehat{y_{gct}^{sa}} + \text{adj}_{gc})$  where  $\text{adj}_{gc} = \sigma_{g\tau}/2$  is a Jensen's inequality adjustment for going from log levels to levels.

An optional step, which we use in the application, is to let the relative importance of  $\hat{q}_{gct}$  and  $\hat{d}_{gct}$  to vary across products. These weights can be estimated using the method of least squares to determine the weights of the two predictable component on  $y_{gct}$ :

$$y_{gct} = \alpha_{g0} + \alpha_{g1} \cdot \hat{d}_{gct} + \alpha_{g2} \cdot \hat{q}_{gct} + u_{gct} \quad (2)$$

Inserting  $\hat{u}_{gct}$  into (1) and inverting gives an alternative estimate of log adjusted sales.

We use a simple Fourier regression to fit the seasonal variations at the  $(g, c)$  level in Step 1 because least squares regression is simple to implement, and the predictors are the same across  $(g, c)$  pairs. Furthermore, there is a history of using Fourier regressions in seasonal adjustments. Cleveland and Devlin (1980) suggests to use the spectrum to detect pre-identified calendar and holiday effects in monthly data. Subsequent work such as Pierce et al. (1984) and Cleveland and Scott (2007) use Fourier regressions to model deterministic

seasonal component, and then employ ARMA models or local regressions to capture stochastic seasonality. As discussed in Cleveland et al. (2014), these regressions are still used in practice. However, these methods require careful choice of tuning of the parameters which is not practical when the number of series to fit is large. We use machine learning methods at this step and rely on algorithms to find the best model for each series. Regardless of the method used in Step 1, we are limited to nine seasonal observations for both training and validation. Instead of fine tuning Step 1, we augment Step 1 with a Step 2 that pools information across clusters year-by-year. This is now a panel-based prediction problem that involves a large number of potentially relevant predictors. Subsection 4.2 first discusses the predictors used. The prediction methodology will be discussed in Section 4.3.

## 4.2 Predictors $\mathbb{Z}_{gct}$

To properly remove the seasonal effects, we need to first understand the nature of the seasonal variations in the data. Consider the event Cinco de Mayo. It occurs on a fixed calendar date, and so is not strictly periodic with respect to the weeks within a year. The seasonal effects of Cinco de Mayo may be more important for counties with a higher Hispanic population. Another example is the event of Thanksgiving, which is always on the fourth Thursday in November and is celebrated across the country. The day in the year that Thanksgiving takes place shifts over time.

We need a flexible methodology to capture not just the week of the year and location effects, but also the day of the year effects. The last consideration may seem surprising since our data are weekly. But a major challenge is precisely that many of our seasonal events occur at different days of the year that cannot be parametrically modeled. With this in mind, we consider date and week specific dummies as well as demographic and spatial predictors collected into  $\mathbb{Z}_{gct}$ . These are defined as follows. Let  $\text{START}_t$  denote the date on which week  $t$  starts, and  $\text{END}_t$  denote the date on which week  $t$  ends

- A. Date-specific predictors: a dummy variable for each potential calendar date (MM-DD) which is 1 if that date is contained in  $[\text{START}_t, \text{END}_t]$  and 0 otherwise. As an example, if  $t = \text{Feb } 4, 2006$ , the date-specific predictors  $\mathbb{Z}_{gct}^A$  is as follows:

$$\begin{array}{cccccccc} 01/01 & \cdots & 01/29 & \cdots & 02/04 & \cdots & 12/31 \\ \hline 0 & \cdots & 1 & \cdots & 1 & \cdots & 0 \end{array}$$

- B. Week specific predictors.

- i.  $\text{START}_t$  and  $\text{END}_t$ 's positions within the year (out of 366)

- ii.  $\text{START}_t$  and  $\text{END}_t$ 's position within the months (out of 31)
- iii.  $\text{START}_t$ 's position within the month containing  $\text{END}_t$  (this will be a negative number, and differ from the previous column, if and only if the week ending on  $t$  crosses two different months)
- iv. A dummy variable which is 1 if Easter is in the week ending on  $t$ , and 0 otherwise
- v. A dummy variable encoding the month in which  $\text{END}_t$  falls

For example, for  $t = \text{Feb } 4, 2006$ , the week -based predictors  $\mathbb{Z}_{gct}^B$  will be:

(i)	(i)	(ii)	(ii)	(iii)	(iv)	(v)	Jan	(v)	Feb	...	(v)	Dec
28	34	29	4	-2	0	0			1	...		0

C. Demographic predictors depend only on county  $c$ . These variables are drawn from the 2013 American Community Survey, and held constant across time:

- i the percentage of the county that is Black, Hispanic, White, and Asian
- ii the percentage of the county on SNAP, in poverty, and median household income
- iii the percentage of the county  $c$  over 60 and under 18
- iv Centroid latitude, and centroid longitude for the county
- v NOAA's 30-year estimates of average rainfall and temperature for county  $c$  during the week of  $t$  (which depend on  $c$  and  $t$ )

The predictors in list A are day of the year dummies. As distinct from list A, the predictors in list B captures the Gregorian calendar effects at the week level. For example, some months have four Saturdays but other months may have five; a week may begin in one month and ends in the other. The interaction of the three sets of predictors generates as many as 400 potentially relevant predictors. Ex-post, the 366 date-based predictors are the most important. Results will be reported treating these predictors as the the base case.

### 4.3 The Prediction Model

Generically denote data with  $N$  cases by  $\mathcal{D} = (\mathbb{Y}, \mathbb{Z})$  where  $\mathbb{Y}$  is the response variable and  $\mathbb{Z}$  is a set of observed predictors. To make predictions for all weeks in year  $\tau$ , we partition  $\mathcal{D}$  into  $\mathcal{D} = (\mathcal{D}_{1\tau}, \mathcal{D}_{2\tau})$  where  $\mathcal{D}_{1\tau}$  collects data for all weeks  $t \in \text{yr}(\tau)$ , and  $\mathcal{D}_{2\tau}$  collects all data not in year  $\tau$ . The  $N_{1\tau}$  cases in  $\mathcal{D}_{1\tau}$  will be used for training, and the  $N_{2\tau}$  cases in  $\mathcal{D}_{2\tau}$  will be used for validation, with  $N = N_{1\tau} + N_{2\tau}$ . The goal is prediction of points  $z^*$  in  $\mathcal{D}_{2\tau}$ .

Since we are interested in predicting the common seasonal variations in the composite error that emerges from the Fourier regression in Step 1, the mapping into  $\mathcal{D}$  notation is

$$\begin{aligned}\mathcal{D} &= (\{\hat{\epsilon}_{gct}\}, \{\mathbb{Z}_{gct}\}) = (\hat{\epsilon}_{g\tau}^s, \mathbb{Z}_{g\tau}^s), \quad \forall t : \text{yr}(t) = \tau \\ \mathcal{D}_{1\tau} &= (\{\hat{\epsilon}_{gct}\}, \{\mathbb{Z}_{gct}\}) = (\hat{\epsilon}_{g\tau}^s, \mathbb{Z}_{g\tau}^s), \quad \forall t : \text{yr}(t) \neq \tau\end{aligned}$$

where  $\hat{\epsilon}_{gt}^s$  is a stacked vector of  $\hat{\epsilon}_{gct}$  for all  $c$  in state  $s$ , and  $\mathbb{Z}_{g\tau}^s$  is similarly defined. In words, the training data  $\mathcal{D}_{1\tau}$  consists of observations for all counties in state  $s$  over all 469 weeks, less those weeks in year  $\tau$  (which is 52 except in a leap year). Thus the training data are indexed by the triplet  $(g, s, \tau)$ .

State by state, we train algorithms to fit a prediction model for each product group in each of the nine years. Thus for each state, the exercise involves training  $N_g \times N_{\text{yr}}$  models. For a given predictor set  $\mathbb{Z}$ , we use training data  $\mathcal{D}_1$  to estimate several models:

1. linear panel model using all predictors by POOLED OLS.
2. linear panel model using LARS-type methods to perform variable selection.
3. regression trees using RANDOM FOREST-type methods to determine the tree size.

We have close to 400 potential predictors, but we also have  $(469 - 52)$  weeks of data for each county. Hence even a pooled least squares regression that uses all predictors (method 1) will have enough degrees of freedom. However, using all predictors is unlikely to be efficient. Hence we consider two machine learning procedures.

Introduced in Efron et al. (2004), the *least angle regression* estimator LARS is a functional gradient descent method that repeatedly fits a model to the residuals of the previous step. LASSO, forward stagewise regressions, and boosting can be obtained as special cases of LARS. Under the boosting view, each model (also known as learner) is individually weak, but is 'boosted' to produce a strong learner via averaging. Averaging in this case reduces bias. Our implementation of LARS-type methods is actually based on LASSO because it requires fewer choices of tuning parameters. The base learner is thus a linear model rather than a regression tree. The LARS perspective helps understand the difference with random forests.

Random forest (RF), due to Breiman (2001), is an ensemble method that builds a prediction from a collection of regression trees. Each tree is fitted to a randomly selected subset of predictors in a bootstrapped sample. Like LARS type estimators, the final model is also an average over trees. But unlike LARS, these trees are built either separately or in parallel rather than sequentially. Regression trees can uncover complex relations and are strong learners, but they tend to have high variance. Averaging in the case of random forests



reduces the variance of models that have low bias. One advantage of regression tree over non-parametric regressions is that the smoothness condition on the regression function can be relaxed. Random forest is an extension of BAGGING which averages over trees grown on bootstrapped samples using all predictors.

The prediction provided by LARS or random forest is implicitly formed by averaging over the predictions of models that use only a subset of available predictors. Hence they are more resistant to overfitting. Though these methods have been widely applied to iid data, applications to time series data are more limited. Success of these algorithms in the present setting is very much an empirical matter. Of the three methods, random forest is the most flexible since it does not impose linearity or smoothness. We use it as benchmark in the discussion of results. We implement random forests using the R package RANGER with default parameter settings.<sup>9</sup> We find that the LARS-type methods do not achieve very much sparsity: our trained estimators have nonzero loadings on over 80% of the included variables, with worse performance than the random forest. By contrast, variable-importance tests for the random-forest show that a small number of predictors (mostly having to do with a week’s position within the year) are being used in highly nonlinear ways. This suggests that the underlying seasonal process is highly nonlinear, and a better fit for the random forest algorithm than the LARS algorithm.

## 5 The Seasonally Adjusted Data

The crux of our two-step procedure is to first remove deterministic seasonal effects using univariate Fourier regressions, and then exploit cross-section dependence to remove the residual common seasonal/holiday effects. Once this is accomplished, the seasonally adjusted budget shares can be computed as the ratio of seasonally adjusted sales for the group to total adjusted sales summed across groups. The largest difference between the unadjusted and adjusted shares are in groups like floral, insecticides, canning, ice, fragrances, toys, stationary, and candies. These results make sense since effects due to seasonal holiday events are precisely what we want to remove.

Table 2 uses two products to contrast the seasonal patterns in the raw and the adjusted data. Consider first beer sales, which tends to be higher in the summer and peaks around July 4th. In 2009, July 4th (week 183) fell on a Saturday when the Nielsen data were

---

<sup>9</sup>The default size of forest is `—ntree—=500` trees, and the default value of `—mtry—` (the number of independent variables considered for each split) is the square root of the total number of independent variables. The `—min node size—` parameter, which controls the depth of each tree grown, is set to 5 by default. It is possible that fine tuning the parameters can yield improved results.

collected. As July 4th is a common event, high beer sales likely occurred across counties. Our Step 2 should smooth out this holiday effect. As shown in the top of Table 2, the adjusted data is indeed smoother and exhibit a smaller spike than the raw data. Take the case of NY as an example. The share of beer computed from the raw data is 3.8 for the week ending July 4, but is 2.4 for the week ending February 7. The adjusted data exhibit smaller differences, being 2.5 and 2.7 for the two weeks in question. Beer sales nonetheless spike each winter around the first week of February because of Superbowl. This is illustrated for 2009 when Superbowl took place on Sunday, February 1. The adjusted shares are smoother within and between months.

Table 2: Effects of the Seasonal Adjustment on Selected Series' Share (%)

Week Ending	CA	FL	NY	TX	CA	FL	NY	TX
	Adjusted Data				Raw Data			
	The 2009 July 4th Effect on Beer Spending							
June 27	3.5	2.9	2.5	2.6	4.1	3.3	3.2	3.0
July 4	3.5	2.8	2.5	2.7	4.9	3.2	3.8	3.6
July 11	3.2	2.8	2.4	2.2	3.8	3.5	3.3	2.8
	The 2009 Superbowl Effect on Beer Spending							
Jan 31	3.3	2.6	2.6	2.5	3.3	2.4	2.2	2.1
Feb 7	3.7	2.7	2.7	2.6	3.3	2.7	2.4	2.3
Feb 14	3.0	2.5	2.3	2.3	2.5	2.2	1.9	1.9
	The April 1 2009 Cigarette Tax Hike							
April 4	1.2	4.4	2.7	3.2	1.2	4.8	2.6	3.2
April 11	1.1	4.1	2.4	2.7	1.0	4.1	2.3	2.7
April 18	1.3	4.4	2.8	3.3	1.3	4.3	2.8	3.3

It is also important that the second step adjustment does not remove spikes and variations that are non-seasonal. To check this, we consider the 62-cent Federal tax hike on cigarettes in April 1, 2009, which corresponds to week 171 in our data. Recall that the data for 2009 are adjusted using training data for all years except 2009. Since the tax hike is a one time event, nothing in the training data should predict the tax hike specific to 2009. The bottom panel of Table 2 reports the share of tobacco for the week before, during, and after the tax hike. According to the raw data, the tax hike led to a temporary decline in sales and hence in the budget share of tobacco. The seasonal adjustment preserves this feature. As in the raw data, the average share of tobacco is generally higher in the 170 weeks after the tax hike than the 170 weeks before the tax hike. For CA, it is 3.9% before and 4.6% after the tax

hike. Averaged over all states, the share is 2.1% (before) compared to 2.3% (after).

The premise of our analysis is that the residuals from the univariate Fourier regressions in Step 1 have co-movements that are predictable. To evaluate the incremental predictive power provided by different adjustment methods, we consider the  $R^2$  corresponding to (2), which is a regression of log sales  $y_{gct}$  on the two estimated seasonal components:  $\hat{d}_{gct}$  and  $\hat{q}_{gct}$ . Table 3 summarizes the distribution of  $R^2$  over all groups and states. A bit over 50% of the variations in log-sales are seasonal and predictable. The degree of predictability varies across groups, ranging from 14% to over 90%. Notably, Step 2 improves upon the univariate Fourier regressions implemented in Step 1 alone. The highest and lowest quantiles of the  $R^2$  do not depend on the procedure. This suggests that the improvements apply not to a few groups with extreme seasonality, but to a large number of groups.

Table 3: Importance of the Seasonal Component

sample	method	Average of $R^2$ in Equation (2)					
		mean	median	max	q75	q25	min
FOUR	Fourier	0.52	0.53	0.95	0.63	0.40	0.14
	RF	0.58	0.59	0.95	0.70	0.44	0.14
SEVEN	Fourier	0.52	0.51	0.95	0.62	0.40	0.14
	RF	0.57	0.57	0.95	0.70	0.44	0.14
ALL	Fourier	0.53	0.52	0.96	0.63	0.41	0.14
	RF	0.57	0.57	0.96	0.70	0.44	0.14

Figure 2 illustrates the difference between using Step 1 alone and the two step procedure by plotting the  $R^2$  of random forest against that based on the Fourier method. If the random forest estimator provides relatively little additional information, the optional step regression after Step 2 will push  $\lambda_g$  towards zero. In such cases, the  $R^2$  values will be bunched along the 45 degree line. Figure 2 indicates such groups do exist. However, many other groups have values in the scatterplot located above the 45 degree line. For some of these groups, the improvement in fit from adding the panel data step is quite significant. A quarter of groups see increases in  $R^2$  of 13% or greater.

At face value, it may seem that the improvement of a few percentage points in predictability over the univariate Fourier regression is trivial. However, the adjusted data that emerge from Step 3 have far fewer spikes than those adjusted using the Fourier regressions alone. This difference has direct implications for demand estimation.

## 5.1 The Factor Estimates

A main finding in the demand analysis of the raw data is that largest factors are exhibit strong seasonal patters. We now explore features of the common factors obtained from the first step alone, and from the two step procedure. We find four factors in the data adjusted by the Fourier Step alone. The first two factors explain over 68% of the variation in the data and consists of a trend and a cyclical component. However, factors three and four remain spiky and quasi-periodic, indicating that the Fourier regressions by themselves leave non-trivial seasonal variations unexplained. In contrast, we find either three or four factors depending on the state in the shares data adjusted by our two step procedure, whether it is based on LASSO or RANDOM FOREST. Compared to factors estimated from no adjustment and Step 1 alone, the most notable difference is absence of large spikes.

Figure 3 plots the three factors in FOUR using data adjusted by random forests. These factors, denoted  $\hat{F}_{RF}$ , are to be distinguish from the ones estimated from the unadjusted data, now denoted  $\hat{F}_{NSA}$ . Though not immediately evident,  $\hat{F}_{2,RF}$  is strongly correlated with  $\hat{F}_{4,NSA}$ . A regression of  $\hat{F}_{4,NSA}$  on  $\hat{F}_{2,RF}$  yields an  $R^2$  of 0.6. The largest residuals of that regression are precisely spikes between weeks 46 and 50, indicating that Step 2 is picking up the spikes not accounted for in Step 1.

The first three factors together explain about 80% of the variations of the adjusted shares, with  $\hat{F}_{1,RF}$  explaining 56%, and  $\hat{F}_{2,RF}$  explaining 15%. As can be seen from Figure 3,  $\hat{F}_{1,RF}$  is highly persistent and has a trend component. An investigation into the factor loadings finds that  $\hat{F}_{1,RF}$  always loads heavily on books and magazines, ethnic hair treatment, and photographic supplies. These product groups appear to have experienced secular trends in share during our sample period.

Even though the Nielsen data are concentrated on grocery store sales with few consumer durables that are traditionally known to be cyclical,  $\hat{F}_{2,RF}$  is visually cyclical and warrants further investigation. We use two measures of consumer confidence index as benchmark of cyclicity: the Rasmussen RCCI index and the Bloomberg index of consumer confidence. The former is a daily national survey collected by the Rasmussen group that tracks 1,500 consumers concerning their confidence, expectations, and sentiment toward the U.S. economy. The latter started as the ABC News consumer comfort index and has been under the control of the Bloomberg Corporation since 2011. Figure 4 plots  $\hat{F}_{2,RF}$  (in dark green), RCCI (in blue), along with BLOOMBERG (in red). It is evident that spending moves positively with consumer sentiment. These confidence measures have absolute correlation with  $\hat{F}_{2,RF}$  of about 0.8. In this regard, consumers' actions are aligned with how they feel. Because

our data covers a very large sample of stores, which is distinct from the much smaller set of consumers surveyed by Bloomberg and Rasmussen, we are able to correlate beliefs with purchasing actions without worrying about the confounding influence of “mere-measurement” effects studied in Morwitz and Fitzsimons (2004) by which asking consumers about their beliefs might affect their ensuing purchasing decisions.

Turning now to  $\hat{F}_{3,RF}$ , it takes a big dip in the week ending March 22, of 2008. As point of reference, JP Morgan purchased Bear Stearns on March 17 of 2008. Furthermore, oil price spiked up to nearly 110 dollars per barrel a few days earlier. Upon examination, the factor is actually highly correlated with the 52 week change in consumer confidence. Figure 5 plots  $\hat{F}_{3,RF}$  estimated using data for four states along with the 52 week change in RCCI and BLOOMBERG. Their correlation with  $\hat{F}_{3,RF}$  are 0.74 and 0.68, respectively. If  $\hat{F}_{2,RF}$  indicates the level of economic activity,  $\hat{F}_{3,RF}$  indicates direction of change. We may think of the three factors in the seasonally adjusted demand system as characterizing the trend, level, and curvature of Engel curves. These estimates of the latent functions are interesting in their own right because classical estimation of demand system cannot consistently estimate the latent functions of prices and income.

It remains to check how our aggregate weekly adjusted sales data compare to the official monthly retail sales. The U.S. Census Bureau releases both the raw and the seasonally adjusted data for retail sales each month.<sup>10</sup> To compare with our weekly series, we interpolate values for the weeks in a month to the officially released sales for the month. Figure 6 plots both series along with  $\hat{F}_{2,RF}$ . The top panel shows that our  $\hat{F}_{2,RF}$  has a correlation of 0.65 with the officially adjusted series. The bottom panel plots 52 week change in the series. The correlation of the adjusted series is 0.74. The most notable difference is around the 2008 Financial Crisis during which the  $\hat{F}_{2,RF}$  shows a steeper decline than the official data. But the weekly series generally tracks the monthly series reasonably well. Some discrepancy is to be expected because our weekly data does not cleanly line up with the monthly calendar.

The results so far have focused on four states: CA, FL, NY, TX. However, similar results are obtained in an extended analysis that groups additional states into three regions: the Midwest (IL, IN, MI, OH, WI), the Mid-Atlantic (DC, DE, MD, VA), and the South-West (AZ, NM, NV). In each of the three regions,  $\hat{F}_1$  is a trend,  $\hat{F}_2$  is correlated with the level of consumer confidence, while  $\hat{F}_3$  is correlated with the 52 week change in consumer confidence. Not surprisingly, pooling data for the four states and three regions also give three factors with very similar properties. Hereafter, we use the extended data when appropriate. These results will be labeled SEVEN and ALL.

<sup>10</sup>The series are RETAILSMNSA and RETAILMSA in FRED.

## 6 Cyclical Sensitivity

A unique feature of the Nielsen scanner data is the availability of weekly information at the spatial and product group levels. This presents an opportunity to study the timed response of spending to economic conditions at a disaggregated level. Subsection 1 considers cyclical sensitivity of product groups while subsection 2 considers cyclical sensitivity of counties.

### 6.1 Variation Across Product Groups

We first turn to the sensitivity of the product groups to business cycle conditions. Since  $\hat{F}_{2,RF}^s$  is positively correlated with RCCI, a positive loading indicates that the share of product  $j$  is pro-cyclical, while a negative value means that the share of product  $j$  is high when  $\hat{F}_{2,RF}^s$  is low. The dispersion of sensitivity to aggregate conditions across product-groups is best seen from the distribution of  $SGNR_{sj}^2$ . This is defined as the signed-fraction of variance of  $SHARE_j^s$  explained by  $\hat{F}_{2,RF}^s$ , where  $SGN$  is the sign of the loading of  $\hat{F}_{2,RF}^s$  on  $SHARE_j^s$ . Though there are some minor differences across states and regions, the pattern across states is broadly similar. Figure 7 presents results for SEVEN. The distribution is noticeably asymmetric because there are more counter-cyclical product groups and the magnitude of the absolute loadings are larger (top) than pro-cyclical ones (bottom). Goods little affected by  $\hat{F}_{2,RF}^s$ , plotted in the middle of Figure 7, are disposable diapers, shaving products, cold and cough remedies, and somewhat surprisingly, beer.

The effect of the cyclical factors on the shares is highly heterogeneous. According to the factor loadings, a decrease in  $\hat{F}_{2,RF}^s$  has the largest marginal impact on the share of frozen, canned vegetables, and pasta. The impact of an increase in  $\hat{F}_{3,RF}^s$  is most adverse (i.e. most negative) on eggs and most positive on dried-fruit, which is often marketed as a snack. These results suggest less eating out during downturns in favor of preparing meals at home. There is increasing evidence for adaptive changes in the pattern of food consumption during the Great Recession. The USDA finds not only that total food spending fell during the Great Recession, but also that recovery was slow.<sup>11</sup> Cha et al. (2015) aggregates the weekly Homescan data to annual level and finds that food consumed at home is countercyclical. Grittith et al. (2015) finds that households also adjusted food spending in the U.K. Our results reinforce these findings using a completely different approach.

To further explore this phenomenon at a more granular level, state by state we aggregate spending on the five product-groups with large negative loadings. These are frozen vegetables, canned vegetables, pasta, bread, and condiments/sauces. Because these products all

---

<sup>11</sup>See <https://ageconsearch.umn.edu/bitstream/120969/2/10FoodSpending.pdf>.

seem related to home-cooking, we designate them the FOOD-IN group. We also identify the five products with large positive loadings on  $\hat{F}_{2,RF}^s$ : liquor, prepared food, milk, hair care, and cosmetics. These five products are then aggregated to form a LUXURY good basket, one for each state. Note that because our data are restricted to grocery-store goods, our LUXURY goods are relatively less “luxurious” than goods included in most standard luxury-good baskets.

Next, we use a five variable VAR to evaluate the dynamic response of FOOD-IN and LUXURY to an unanticipated increase in the two cyclical factors  $\hat{F}_{2,RF}$ ,  $\hat{F}_{3,RF}$ , and to RCCI. We report results for FOUR, but results for SEVEN and ALL are similar. The dynamic responses to one standard deviation shocks are shown in Figure 8. A positive  $F_{2,RF}$  shock, which is an increase in economic activity, has a negative effect on FOOD-IN that peaks after two weeks and nearly recovers after five weeks. This negative effect on FOOD-IN is mirrored by an opposite effect on LUXURY. The absolute impact on LUXURY is actually larger than that on FOOD-IN. The effect of a  $\hat{F}_{3,RF}$  shock is mainly on LUXURY; the impact on FOOD-IN is negligible. In terms of decomposition of variance, about 55% of the variations in food-in are explained by its own lag, 35% explained by  $\hat{F}_{2,RF}$ , 7% by  $\hat{F}_{3,RF}$  with little attributed to RCCI. About 37% of the variations in luxury are explained by its own lag, 28% by  $\hat{F}_{2,RF}$  and 32% by  $\hat{F}_{3,RF}$ . It thus appears that FOOD-IN is primarily affected by the level factor, while LUXURY is affected by both the level and the curvature factors, ie. where the economy is, and where it is going. The results are robust to whether RCCI is ordered second or last. Interestingly, even though the correlation between RCCI and food-in is well over 0.75, shocks to RCCI account for little of the variations in FOOD-IN and LUXURY once conditioned on  $\hat{F}_{2,RF}$  and  $\hat{F}_{3,RF}$ .

## 6.2 Variation Across Regions

According to the NBER’s Business Cycle chronology, the downturn in economic activity leading to the Great Recession began in December 2007 when the last business cycle peaked, and continued to decline until it reached a trough in June 2009. This subsection looks at the spatial aspect of the change in food spending before, during, and after the Great Recession.

The consumer price index (CPI ) is based on a comprehensive consumer expenditure survey conducted by the Bureau of Labor Statistics (BLS) every two years. The CPI weights reflect the relative importance of the particular good in consumption basket. The top panel of Table 4 reports the CPI weights for *food consumed at home* and *luxury* as defined by the BLS. In their own study on how consumer spending change during boom recession and

Table 4: Spending Over the Business Cycle

	Dec 2007	Dec 2009	Dec 2011	Dec 2013
	CPI Weights (%)			
FOOD-IN	7.6	-	8.6	8.1
FOOD-OUT	6.1	-	5.6	5.7
	Seasonally-adjusted Nielsen Shares (%)			
FOOD-IN:FOUR	5.6	6.0	5.8	5.6
FOOD-IN:SEVEN	6.0	6.5	6.2	6.1
FOOD-IN:ALL	6.3	6.7	6.6	6.4
FOOD-IN:FLORIDA	4.3	5.0	4.7	4.6
FOOD-IN:MIDATL	6.9	7.4	7.2	7.0
LUXURY:FOUR	8.3	8.6	8.8	9.0
LUXURY:SEVEN	7.9	8.2	8.3	8.5
LUXURY:ALL	7.8	8.2	8.1	8.3
LUXURY:MIDATL	6.6	6.5	6.5	6.6
LUXURY:NY	7.3	7.3	8.1	8.1

recovery, 2007 was used as a boom year, 2011 as recession, and 2013 a year of recovery.<sup>12</sup> The CPI weights indicate an increased importance of FOOD-IN and a reduced importance of LUXURY items during recessions.

How well do our adjusted shares corroborate with the CPI weights? The bottom panel of Table 4 reports the shares of FOOD-IN and LUXURY averaged over the weeks in December for four years that represent different stages of the business cycle. Notably, FOOD-IN is much higher in 2009 and 2011 than in 2007 and 2013, while LUXURY is lower in 2009 than in 2013. Even though our definitions of FOOD-IN and LUXURY are data-driven, factor based, and restricted to grocery-store nondurables, the Nielsen data also indicate an increased importance of FOOD-IN and reduced importance of LUXURY items during recessions, similar to the more comprehensive CPI weights.

An appeal of the Nielsen data is that it provides granular information in both the time series and cross-section dimensions. The share of FOOD-IN ranges between 5% in Florida to 7.8% in the Mid-Atlantic regions with an average of 6.6% over the entire sample. The series is most persistent in CA and least persistent in the Mid-West, with first order autocorrelation coefficient of 0.83 and 0.50 respectively. The share of LUXURY ranges between 6.6% in the Mid-Atlantic regions to 12% Florida with an average of 8.9% over the full sample. The

<sup>12</sup>See <https://www.bls.gov/opub/btn/volume-3/how-does-consumer-spending-change-during-boom-recession-and-recovery.htm>.



series is most persistent for the Mid-West and least persistent in the South-West, with autocorrelation coefficient of 0.86 and 0.5, respectively. The contemporaneous correlation between FOOD-IN and LUXURY is strongly negative in CA, NY and the Mid-West, with cross-correlations in excess of 0.6 in absolute value. The correlation is much weaker in the South-West and even positive in FL. The heterogeneity across states in spending behavior can be important in the design of economic policies.

To analyze local sensitivity to (aggregate) business cycle fluctuations, we also estimate for each county in each state, the regression

$$\text{food-in}_{ct} = a_{c1} + a_2 \hat{F}_{2,RF,t} + a_{3c} \hat{F}_{3,RF,t} + \text{error}_{ct} \quad (3)$$

The  $R^2$  provides a measure exposure of county  $c$  to the two common factors. Upon ranking the  $R^2$ s, the urban and densely populated counties are found to be more exposed to aggregate shocks. Take the state of NY as an example. The counties Rockland, Nassau, and Kings county have a combined population of over 4 million according to the 2010 census. Each of these counties has an  $R^2$  above 0.45. In contrast, the counties Seneca, Lewis, and Broome with a combined population of under 300,000 each have an  $R^2$ s of at most 0.01.

Heatmaps provide a more compact way to see how different regions are affected by economic conditions. The top panel of Figure 9 plots the change in FOOD-IN between 2006 and 2007. Regions shaded with darker red indicate larger reductions in FOOD-IN. With the exception of isolated regions in Michigan, this boom episode was associated with reductions in FOOD-IN. The reductions were largest in Nevada and Arizona, one possible explanation being the housing boom in those regions. The bottom panel presents the change in FOOD-IN from 2008 to 2009, an episode of economic downturn. Darker blue indicates larger increases in FOOD-IN share. Now there are more regions shaded blue than brown, with Arizona and Florida witnessing the largest increase in FOOD-IN. Thus while the effects of the Great Recession were pervasive, the regions were differentially affected.

The regressions based on equation (3) help understand the impact of aggregate economic conditions on weekly spending. It is also of interest to learn about the impact of local rather than aggregate economic conditions. But as mentioned above, there is limited data at the state and county levels. We use the county level unemployment rate which is only available monthly. Using data over the sample 2006-2014 with a total of 88,559 observations, we consider fixed effects estimation of

$$\overline{\text{BASKET}}_{c,m} = \alpha_{1c} + \gamma_{1m} + \beta_1 \text{UNEMP}_{c,m-1} + \text{error} \quad (4)$$

where  $\alpha_c$  is a county-level fixed effect,  $\gamma_m$  is a set of month dummies,  $\text{unemp}_{c,m}$  is unemployment rate in county  $c$  at month  $m$ , and  $\overline{\text{BASKET}}_{c,m}$  is the average share of either

FOOD-IN or LUXURY in county  $c$  across all weeks in month  $m$ <sup>13</sup>. Both UNEMP and BASKET are standardized to have standard deviation one.

Table 5: Effect of Unemployment on Spending: Fixed-Effects Estimates of  $\beta_1$

basket	FOOD-IN		LUXURY	
	sa	nsa	sa	nsa
FOUR	0.315 (0.020)	0.278 (0.017)	0.040 (0.024)	-0.013 (0.021)
SEVEN	0.338 (0.011)	0.265 (0.009)	-0.087 (0.013)	-0.103 (0.011)
ALL	0.363 (0.008)	0.296 (0.006)	-0.108 (0.009)	-0.128 (0.008)

The parameter  $\beta_1$  measures the marginal effect of local economic conditions on spending. The estimates are shown in Table 5. Estimates using seasonally adjusted data are in the column SA while those using raw data are labeled NSA. Standard errors are clustered at the county level. The estimates of  $\beta_1$  indicate that lagged local unemployment is a statistically significant predictor for the one-month ahead FOOD-IN in the cross-section of counties. A one standard deviation increase in the local unemployment rate predicts  $\frac{1}{3}$  of a standard deviation increase in seasonally adjusted FOOD-IN share. The estimates of  $\beta_1$  using seasonally unadjusted data are generally smaller, suggesting that the presence of seasonal effects weakens predictability. The results are similar with or without controlling for month of the year effect, and robust to adding a polynomial time control (not shown). The marginal effect of unemployment on LUXURY is, however, not as well determined.

Mian and Sufi (2010) reports a drop in consumer durables a full year (or  $j = 12$  months) before any significant change in unemployment. In the case of ALL, we do find a marginally significant coefficient in a fixed-effect regression of local unemployment on  $LUXURY_{c,m-j}$  for all  $j \in [1, 10]$ , suggesting the possibility that consumers reduce non-essential spending ahead of bad times, while cutbacks in FOOD-IN are deferred till an economic downturn appears imminent. However, this result is not present in FOUR and only marginally significant in SEVEN. One reason why the Mian-Sufi finding is not as clear in this data is that our LUXURY basket does not contain durables.

<sup>13</sup>Note that because the data are weekly we cannot cleanly isolate all sales from a given month.

## 7 Conclusion

Large volumes of highly heterogeneous data are increasingly available, but they are often not immediately useful for economic analysis without removing some nuisance variations and performing some form of aggregation. In this paper, the nuisance variations in question are the seasonal and holiday effects. As they cannot be adequately removed by conventional procedures, the adjusted data continue to exhibit seasonal patterns when aggregated over counties. We propose to augment univariate seasonal adjustments with a machine-learning step that pools information across counties. Validity of this second step relies on the presence of common seasonal patterns across counties.

Our two step procedure can be applied to other data. For example, output of firms in a given sector will likely be correlated. Unless we can perfectly remove seasonality at the firm level, the sectoral data obtained by aggregating over firms will likely exhibit seasonality. Staff researchers at the Bureau of Economic Analysis appear to encounter similar experiences. Though our focus has been on handling seasonal effects, the procedure can be adapted to remove other nuisance variations. In a big data setting, it is possible to improve upon the conventional way of removing nuisance variations one series at a time by taking advantage of cross-sectional dependence.

## References

- Bai, J. and Ng, S. 2008, Large Dimensional Factor Analysis, *Foundations and Trends in Econometrics* **3:2**, 89–163.
- Bai, J. and Ng, S. 2017, Regularized Estimation of Approximate Factor Models, mimeo, Columbia University.
- Banks, J., Blundell, R. and Lewbel, A. 1997, Quadratic Engel Curves and Consumer Demand, *Review of Economics and Statistics* **79**, 527–539.
- Breiman, L. 2001, Random Forests, *Machine Learning* **45**(1), 5–32.
- Cha, W., Chintagunta, P. and Dhar, S. 2015, Food Purchases During the Great Recession, Kilts Booth Marketing Series, Paper 1-008.
- Chamberlain, G. 1984, Panel Data, *Handbook of Econometrics*, Vol. 2, North Holland, pp. 1247–1318.
- Cleveland, W. and Devlin, S. 1980, Calendar Effects in Monthly Time Series: Detection by Spectrum Analysis and Graphical Models, *Journal of the American Statistical Association* **75**(371), 487–496.
- Cleveland, W. and Scott, S. 2007, Seasonal Adjustment of Weekly Time Series with Application to Unemployment Insurance Claims and Steel Production, *Journal of Official Statistics* **23**(2), 209–221.
- Cleveland, W., Evans, T. and Scott, S. 2014, Weekly Seasonal Adjustment- A Locally weighted Regression Approach, U.S. Bureau of Labor Statistics Working Paper 473.
- Deaton, A. S. and Muellbauer, J. 1980, An Almost Ideal Demand System, *American Economic Review* **70**, 312–326.
- Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. 2004, Least Angle Regression, *Annals of Statistics* **32:2**, 407–499.
- Engle, R. F. and Kozicki, S. 1993, Testing for Common Features, *Journal of Business and Economic Statistics* **11**, 369–379.
- Fok, D., Franses, P. and Paap, R. 2007, Seasonality and Non-Linear Price Effects in Scanner-Data-Based Market Response Models, *Journal of Econometrics* **138**, 231–251.
- Geweke, J. 1978, The temporal and sectoral aggregation of seasonally adjusted time series, in A. Zellner (ed.), *Seasonal Analysis of Economic Time Series*, NBER, pp. 411–432.
- Gorman, W. M. 1981, Some Engel Curves, in A. Deaton (ed.), *Essays in the Theory and Measurement of Consumer Behavior in Honor of Sir Richard Stone*, Cambridge University Press.
- Grittith, R., O’Connell, M. and Smith, K. 2015, Shopping Around: How Households Adjusted Food Spending over the Great REcession, IFS Working Paper W15/29.
- Harvey, A., Koopman, S. and Riana, M. 1997, The Modeling and Seasonal Adjustment of Weekly Observations, *Journal of Business and Economic Statistics* **15**, 354–368.

- Lewbel, A. 1991, The Rank of Demand Systems: Theory and Nonparametric Estimation, *Econometrica* **59:1**, 711–730.
- Lewbel, A. 1997, Consumer Demand Systems and Household Equivalence Scales, in M. H. Pesaran and P. Schmidt (eds), *Handbook of Applied Econometrics*, Vol. 2:4, Blackwell, Oxford, pp. 167–201.
- Lewbel, A. 2003, A Rational Rank Four Demand System, *Journal of Applied Econometrics* **18**, 127–135.
- McElroy, T. 2017, Multivariate Seasonal Adjustment, Economic Identities, and Seasonal Taxonomy, *Journal of Business and Economic Statistics* **35**, 511–525.
- Mian, A. and Sufi, A. 2010, Household Leverage and the Recession of 2007-09, *IMF Economic Review* **58(1)**, 74–117.
- Morwitz, V. and Fitzsimons, G. 2004, The Mere-Measurement Effect: Why Does Measuring Intentions Change Actual Behavior, *Journal of Consumer Psychology* **14:1-2**, 64–71.
- Muellbauer, J. 1975, Aggregation, Income Distribution, and Consumer Demand, *Review of Economic Studies* **62**, 269–283.
- Ng, S. 2017, Opportunities and Challenges: Lessons from Analyzing Terabytes of Scanner Data, in B. Honore, A. Pkes, M. Piazzesi and L. Samuelson (eds), *Advances in Economics and Econometrics, Eleventh World Congress of the Econometric Society*, Vol. II, Cambridge University Press, pp. 1–34.
- Pierce, D., Grupe, M. and Cleveland, W. 1984, Seasonal Adjustment of the Weekly Monetary Aggregate: A Model Based Approach, *Journal of Business and Economic Statistics* **2**, 260–270.

Figure 1: Factors Estimated from Raw Shares: FOUR States

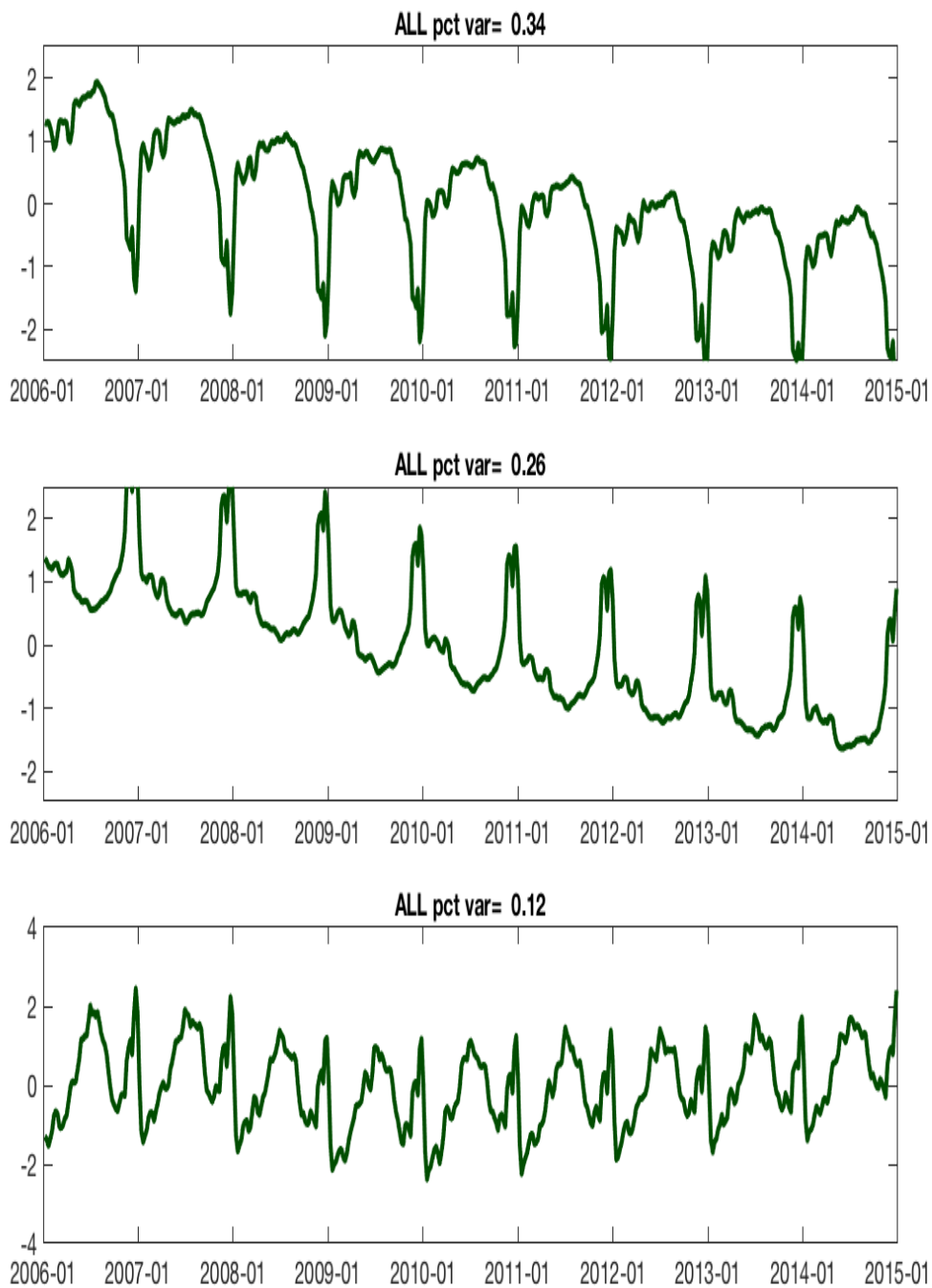


Figure 2: Incremental Predictive Power of Random Forests

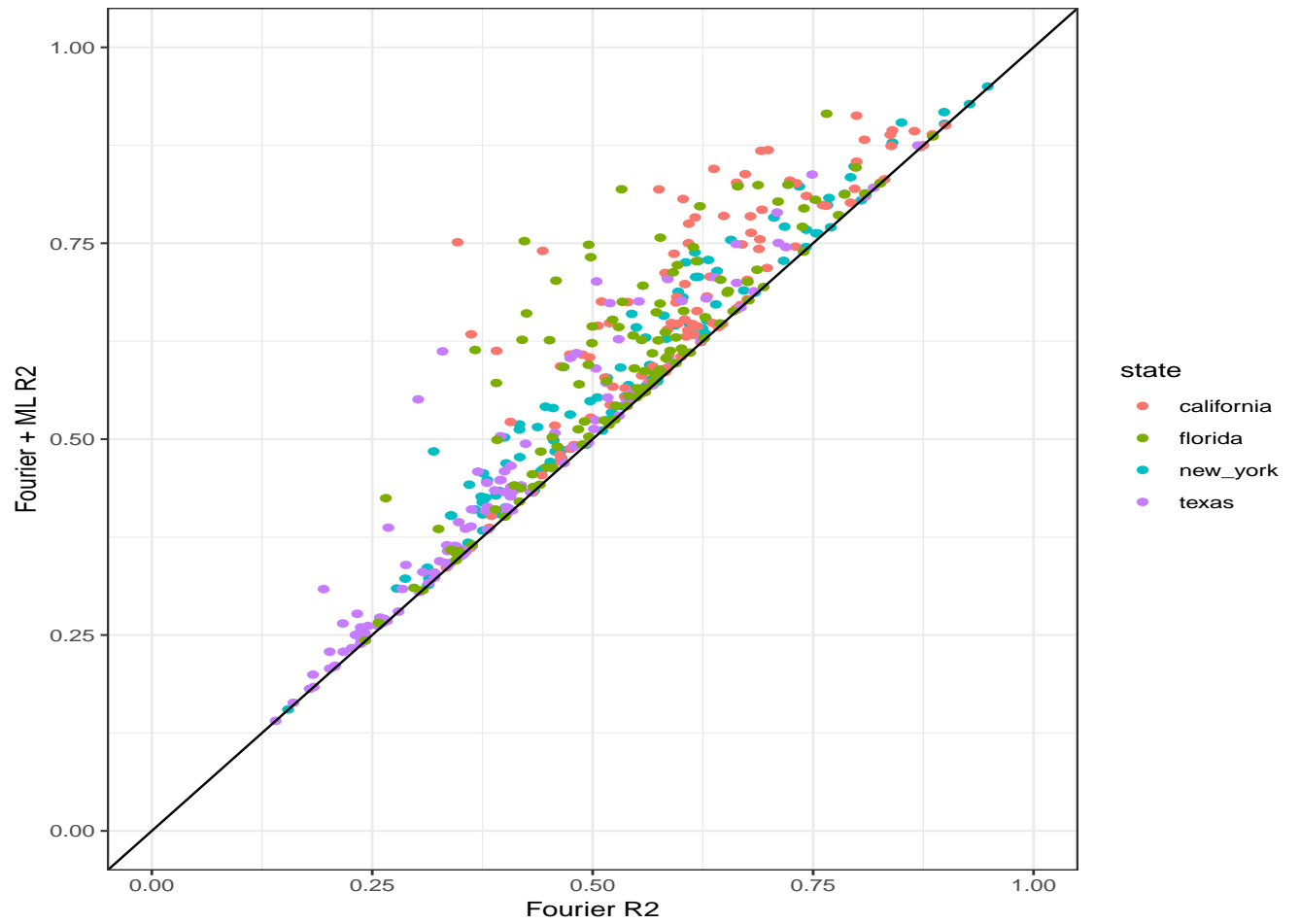


Figure 3: Factors Estimated from Seasonally Adjusted Shares: FOUR States

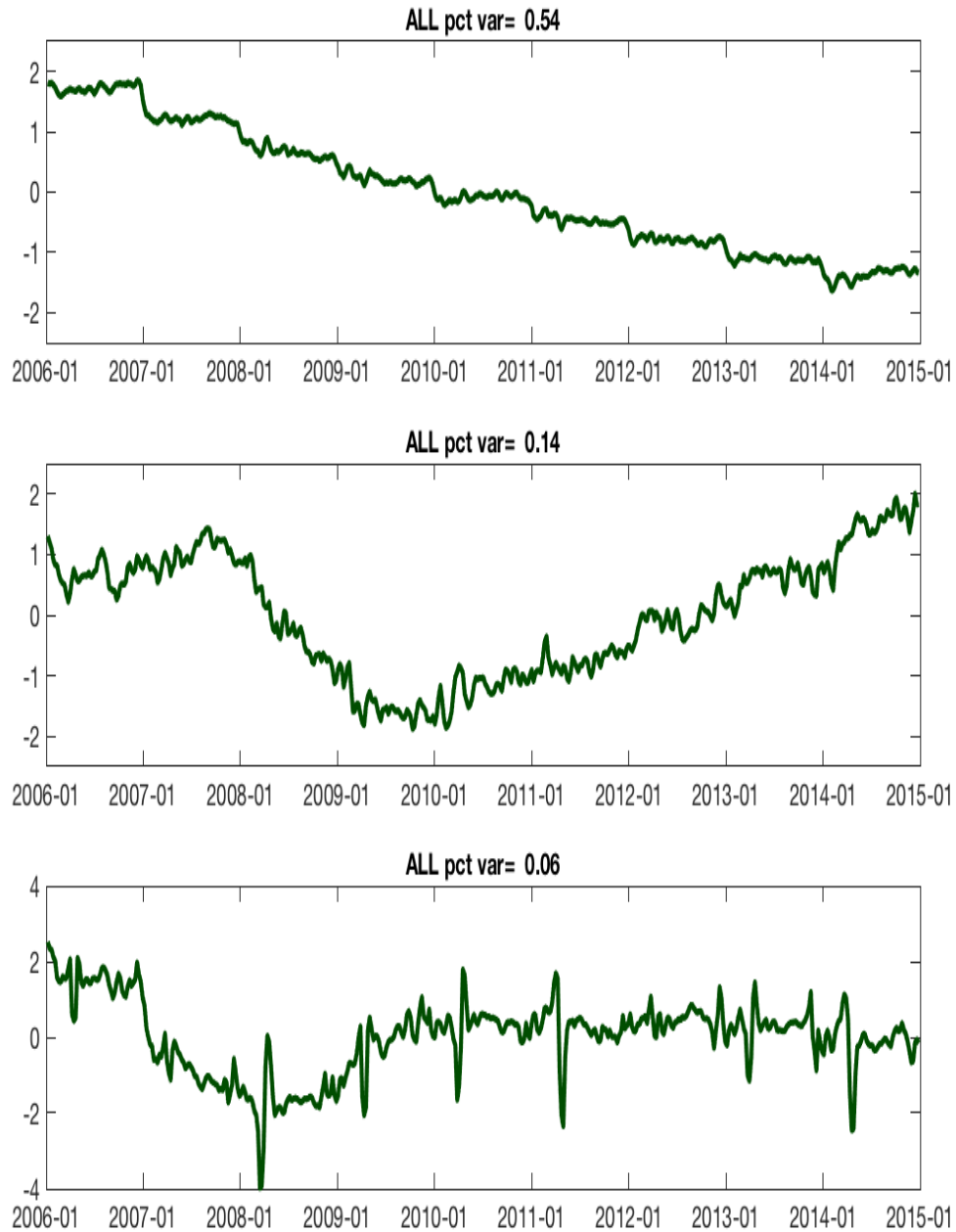




Figure 4: The Level Factor:  $\hat{F}_{2,RF}$

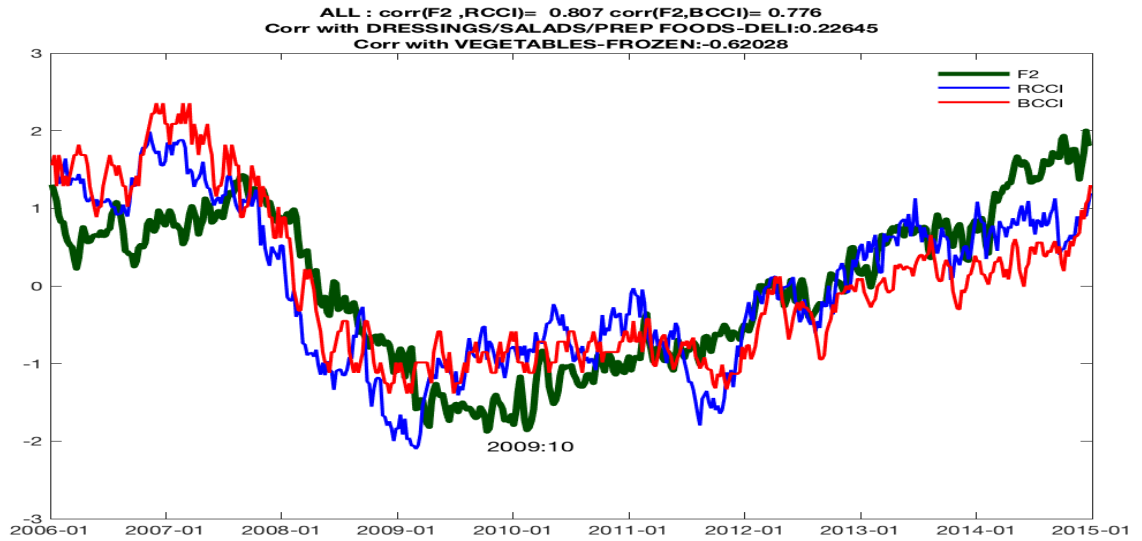


Figure 5: The Curvature Factor:  $\hat{F}_{3,RF}$

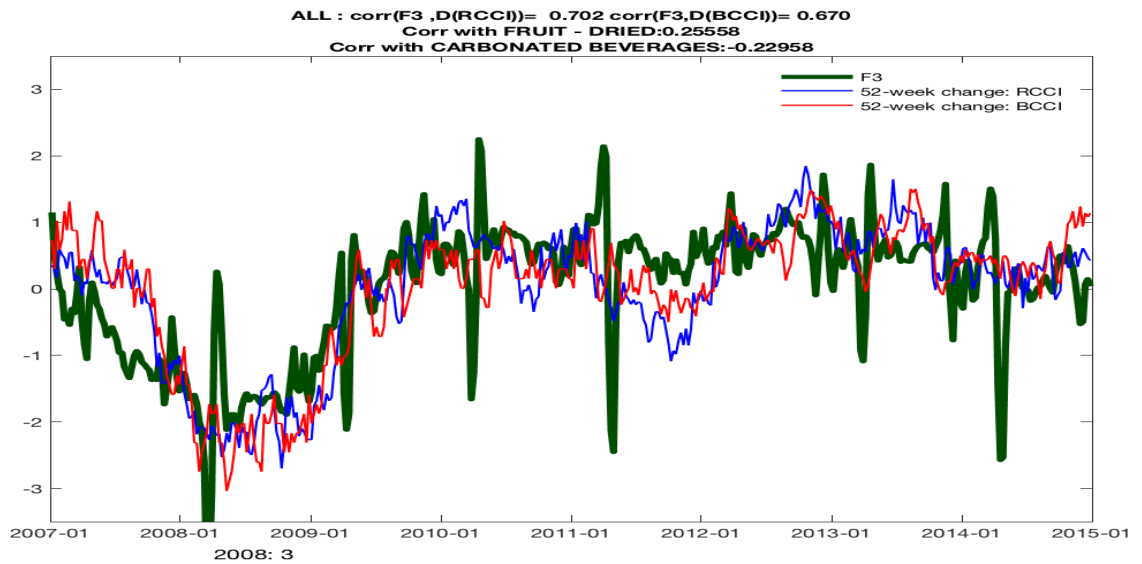


Figure 6: Comparison with Monthly Retail Series

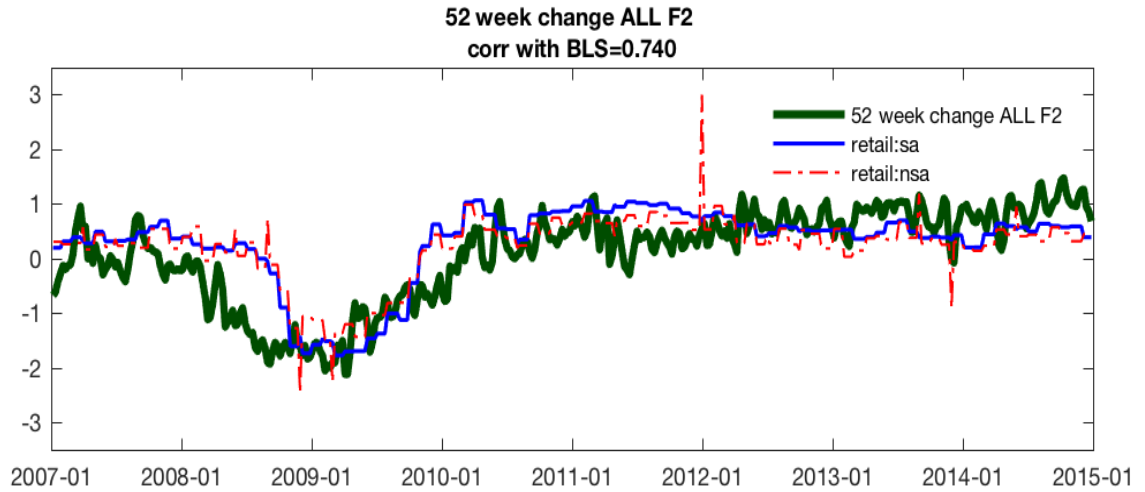
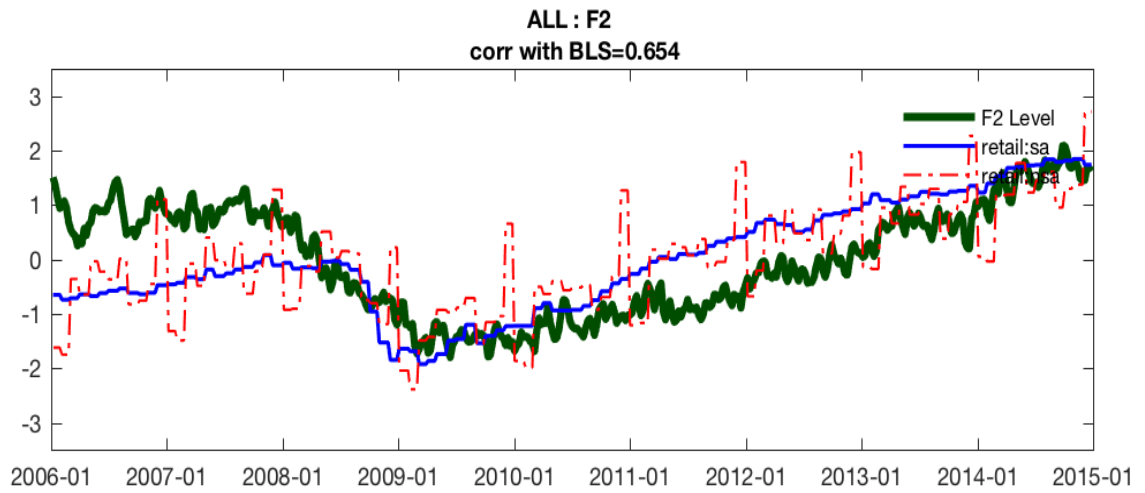


Figure 7:  $R^2$  from Regression of adjusted shares on  $\hat{F}_2$ : FOUR States

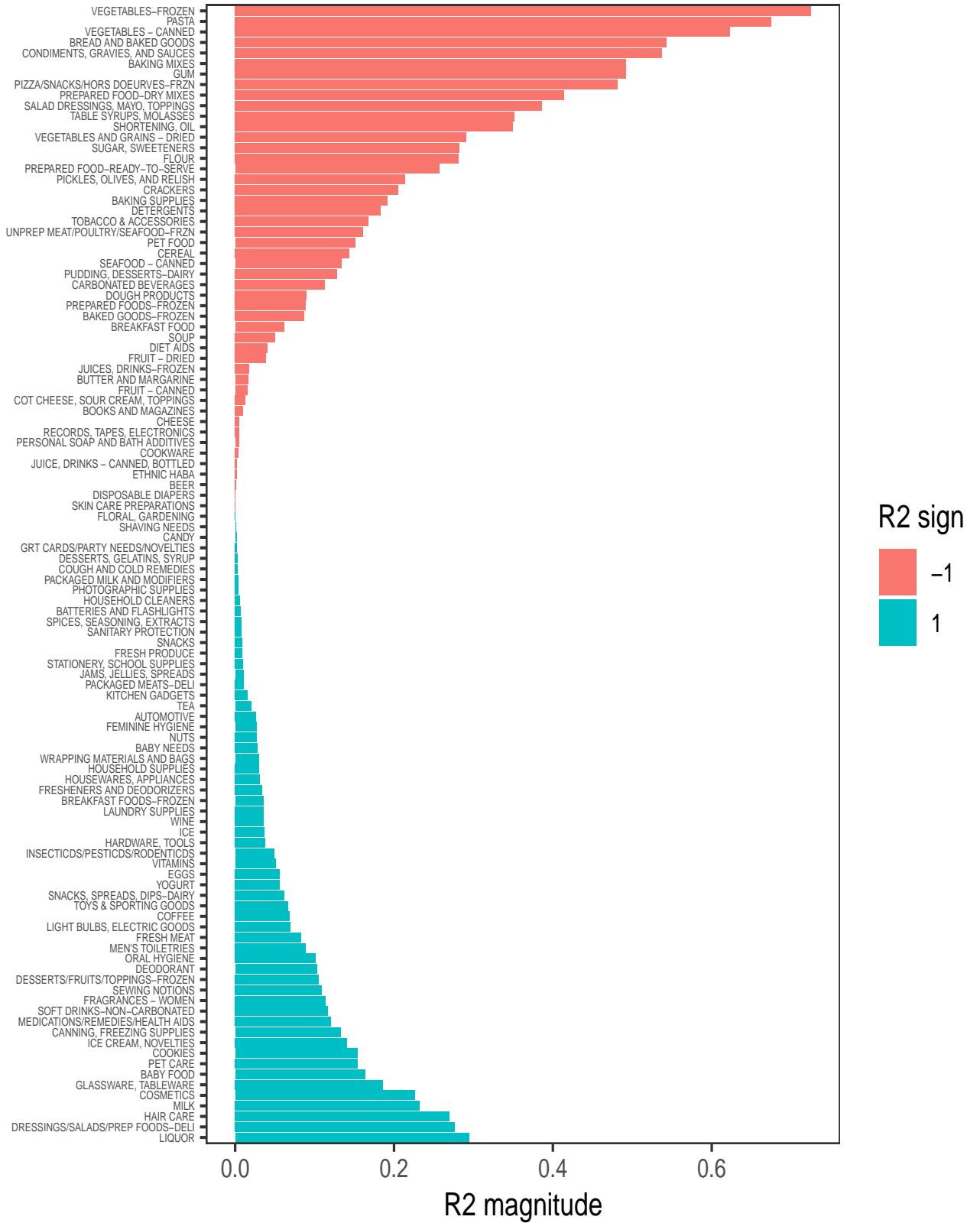


Figure 8: Response of FOOD-IN to Shock in  $\hat{F}_2$

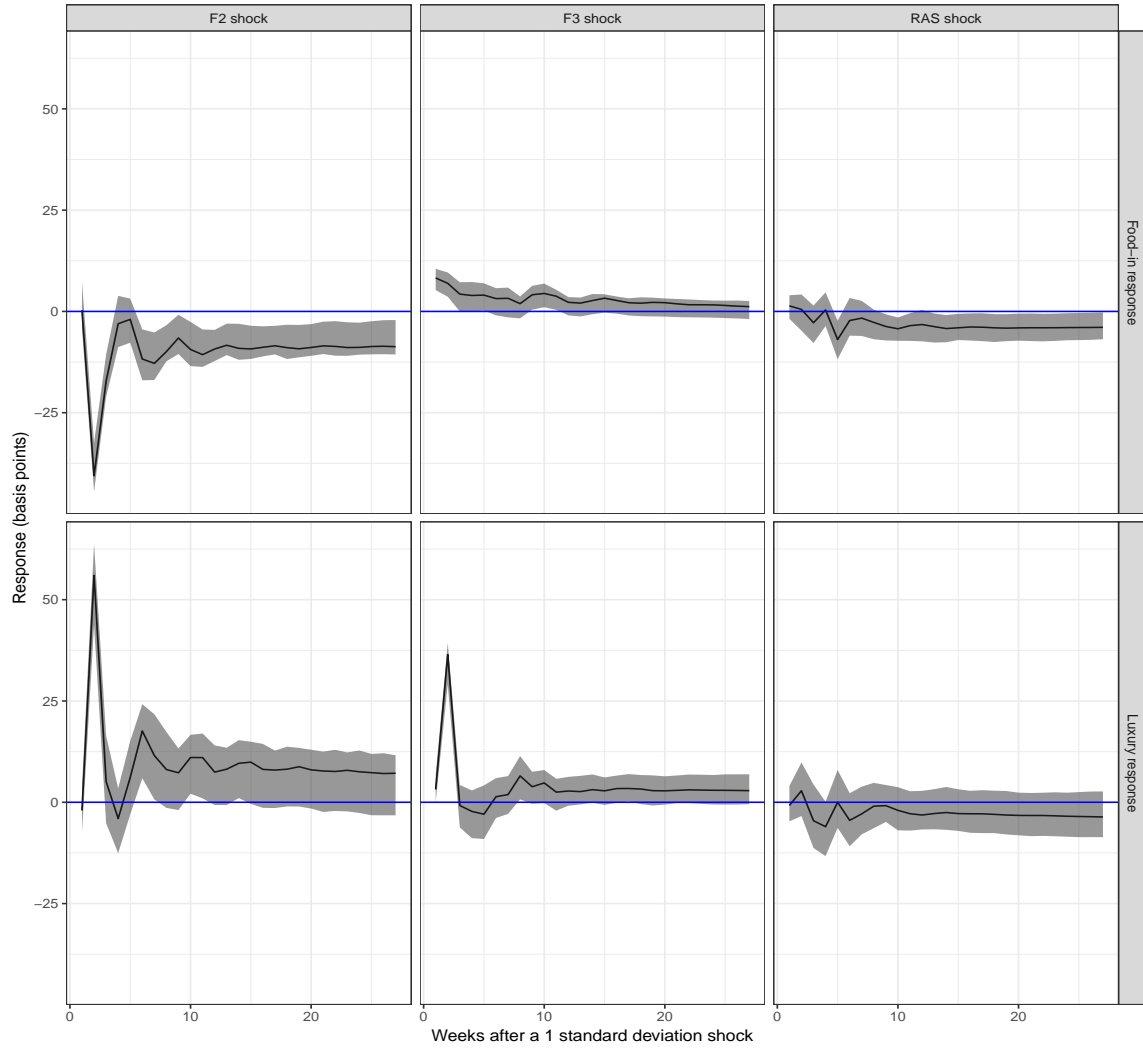


Figure 9: Regional changes in FOOD-IN

