Volume Title: Big Data for Twenty-First-Century Economic Statistics

Volume Authors/Editors: Katharine G. Abraham, Ron S. Jarmin, Brian Moyer, and Matthew D. Shapiro, editors

Volume Publisher: University of Chicago Press

Volume ISBNs: 978-0-226-80125-4 (cloth),
978-0-226-80139-1 (electronic)

Volume URL:
https://www.nber.org/books-and-chapters/big-data-twenty-first-century-economic-statistics

Conference Date: March 15-16, 2019

Publication Date: Februrary 2022

Chapter Title: A Machine Learning Analysis of Seasonal and Cyclical Sales in Weekly Scanner Data

Chapter Author(s): Rishab Guha, Serena Ng

Chapter URL:
https://www.nber.org/books-and-chapters/big-data-twenty-first-century-economic-statistics/machine-learning-analysis-seasonal-and-cyclical-sales-weekly-scanner-data

Chapter pages in book: p. 403 – 436

14

A Machine Learning Analysis of Seasonal and Cyclical Sales in Weekly Scanner Data

Rishab Guha and Serena Ng

14.1 Introduction

A mindboggling amount of data is now available for economists to analyze. This is made possible by improved technology in data collection and storage. Modern data differ from conventional data 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 new analyses of economic behavior. However, a full analysis of the data comes with unique challenges. A case in point is the weekly Nielsen Retail Scanner dataset, which has been collected since 2006 and has added roughly half a terabyte of data each year, reaching a size of about five terabytes in 2016.

Rishab Guha is a John M. Olin Fellow in Empirical Law and Finance, and a PhD candidate in business economics at Harvard University.

Serena Ng is the Edwin W. Rickert Professor of Economics at Columbia University, and a research associate at the National Bureau of Economic Research.

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. For acknowledgments, sources of research support, and disclosure of the authors' material financial relationships, if any, please see https://www.nber.org/books-and-chapters/big-data-21st-century-economic-statistics/machine-learning-analysis-seasonal-and-cyclical-sales-weekly-scanner-data.

The Nielsen dataset has three features of interest. First, it consists of real-time sales and unit prices 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 because the data are digitally recorded. Third, the data are available for the major metropolitan areas and thus provide spatial information distinct from the official monthly retail sales data. The weekly data also provide higher frequency information than 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 quarterly frequency without taking advantage of the weekly information. With just a peek at 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 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.[1]

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 quasiperiodic 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. Even if the residual seasonal effects are negligible at the individual series level, they can become nontrivial 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 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 bypasses 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 model-agnostic, the adjusted data are no lon-

[1] With some abuse of terminology, holiday effects will also be treated as seasonal variations.

ger 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, which collects goods related to home production of food, tends to be strongly countercyclical, while that of a LUXURY basket is procyclical, consistent with evidence from the monthly Consumer Price Index (CPI) weights. Fourth, recession sensitivity has a spatial dimension as cyclical changes in spending on the FOOD-IN basket are larger in metropolitan than rural areas. We use heatmaps to illustrate the changes in FOOD-IN as the economy moves through the business cycle. The data also reveal how consumers in the New York area adapted to changes in spending due to Hurricane Sandy. Overall, the business cycle information in the scanner data seems roughly consistent with the less granular official data. This is good news because it suggests that there is valuable higher-frequency information about consumer spending at the aggregate and local levels once the seasonal variations are removed. 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 14.2 with a description of the data and highlight the presence of common seasonality. Section 14.3 discusses the challenges posed by cross-section dependence that seasonal adjustments must overcome. Section 14.4 presents our two-step approach and elaborates how the second step is formulated as a prediction problem. Section 14.5 analyzes the properties of the seasonally adjusted data and documents how the different products and regions react to changing economic conditions. Section 14.6 concludes.

14.2 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 1,000 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, nonfood, alcoholic beverage, general merchandise, health, and beauty. The data are heavily weighted toward 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 MSAs (Metropolitan Statistical Areas) across the United States. 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 in which 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, a 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 three for $1).

We analyze the total sales of products within a product group (hereafter simply 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. Throughout, we will use "county" to reference a specific geographic county (e.g., New York County, New York). We let $s(i)$ denote the state containing county i.

The variables are indexed as shown in table 14.1.

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 $\log(sales_{gct})$, the log sales of group g

Table 14.1 **Index variables**

	County	Group	Week	Year
Index	cc	gg	tt	$\tau\tau$
Total	NN_{cc}	$NN_{gg} = 108$	$TT = 469$	$NN_{\text{yr}} = 9$

Table 14.2 **Budget shares (%): Most-purchased product groups**

CA: $NN_{cc} = 53$		FL: $NN_{cc} = 58$		NY: $NN_{cc} = 58$		TX: $NN_{cc} = 161$	
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

in county $c(s)$ in week t. 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

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

Our base case analysis uses data from the four most populated states in the US: 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 DC), and Southwest (Arizona, Nevada, and New Mexico). This sample encompasses a total of 24 states (plus DC) covering about 70 percent of the population in the US with sales in 15,631 stores in 1,147 counties. Results that pool over all states in this extended dataset 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 cover 24,280 stores in 2,095 counties.

The 10 product groups with the largest sales in FOUR are listed in table 14.2 below. The relative importance of the groups is reasonably similar across states, with bread, beer, juice, carbonated beverages, medication, and snacks making the list in each of the four states.[2]

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

[2] We work with shares instead of sales, which tend to have even stronger seasonal effects.

of cost of living. A (product group based) demand system expresses N_g budget shares in terms of r functions of prices $p = (p_1, \ldots, p_{N_g})'$, and income Y. Hence, we may write $\text{share}^s_g = \sum_{k=1}^{r} \lambda^s_{gk}(\log p) F^s_k(\log p, \log Y)$. Importantly, the functions $F = (F_1, \ldots, F_r)$ are common across groups. The adding-up constraint requires that F^s_1 is a constant.[3] 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 1980s 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 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

$$\text{share}^s_{gt} = \lambda^s_{0g} + \sum_{j=1}^{N_g} \lambda^s_{jg} \log p^s_{gt} + \beta^s_g \log(Y^s_t / P^{s*}_t) + e^s_{gt},\ t = 1, \ldots, T.$$

The term e^s_g 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.[4]

We are interested in analyzing all product groups in the Nielsen data available, which is well over 100 in number, not five or six. A large N_g may

[3] For a discussion on the rank of demand systems, see Lewbel (1991).

[4] See, for example, Lewbel (1997, 2003); Banks, Blundell, and Lewbel (1997).

appear to hinder analysis at first glance because the number of parameters in a demand system is quadratic in N_g. But because N_g and T are both large, we may deviate from traditional demand analysis and 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:

$$\text{share}^s_{gt} = \Lambda^{s\prime}_g F^s_t + e^s_{gt},$$

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

In implementation, we take a three-week rolling average of budget shares to smooth out the variations due to temporary promotional sales. Principal components are then estimated from the standardized data.[5] In s= (NY, TX, and FOUR), the largest factor $\hat{F}^s_1$ explains over 0.9 of the variations of PACKAGED-MILK. As eggnog is one of the products in the group, we may think of $\hat{F}^s_1$ as a *Christmas* factor. Other product groups also exhibit recurring patterns toward the end of the year. In California, for example, the share of juice takes a big dip in week 51, while in Florida, the share of haircare products bottoms around week 51.

Figure 14.1 plots the first four factors from the pooled data FOUR. The factors are only identified up to sign, so they are plotted to be procyclical. Recalling that the factors are mutually orthogonal by construction, figure 14.1 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 $(2\pi j/52)\,208j$ 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 coexist.

The criterion of Bai and Ng (2019)[6] 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 60 percent, while four factors explain around 75 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

[5] Since our data are demeaned, the constant factor is controlled for prior to estimation.

[6] The criterion is defined as $\overline{r}^s = \min_{k=0,\ldots,\text{rmax}} \log(1 - \sum_{j=1}^{k}(\sigma^s_j - \gamma)^2_+) + k \cdot \textit{penalty}\,(N, T)$, where $\textit{penalty}\,(N, T) = (N + T)/NT \log[NT/(N + T)]$ and σ_j is the j-largest eigenvalue in a $T \times N$ panel $Z = X/\sqrt{NT}$, 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 maximum number of factors is set to 10.

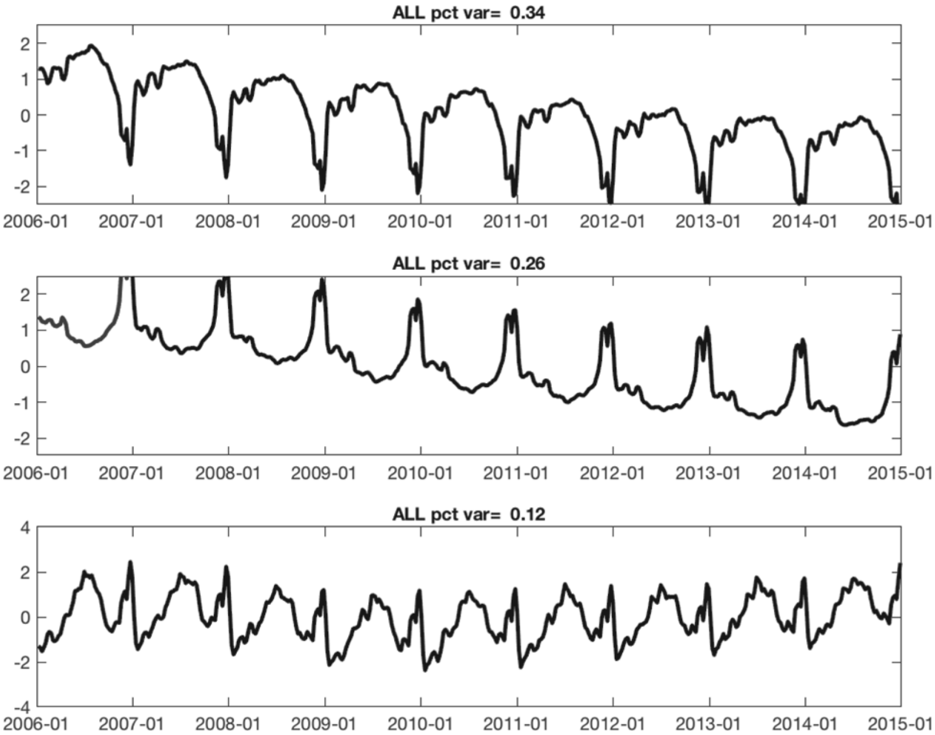


Fig. 14.1 Factors estimated from raw shares: FOUR states

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 show that this finding can be attributed to seasonality.

14.3 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 toward the last six weeks of the year. In addition to eggnog sales that peak around Christmas, sales of stationery 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 tend to be highest around July 4th, 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 are blind to the source of pervasiveness. Hence in the presence 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 because 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 is 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 days 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 fits 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, Koopman, and Riana (1997) and the nonparametric approach of Pierce, Grupe, and Cleveland (1984), Cleveland and Scott (2007), and Cleveland, Evans, and Scott (2014). Structural time series modeling requires careful specification of the model for the series under investigation. The nonparametric approach is to approximate the seasonal component by basis functions such as trigonometric series. Cleveland, Evans, and Scott (2014) suggest to control for weekly effects, holiday effects, and outliers using locally weighted regressions and apply 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 because they tend not to be well approximated by nonparametric 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, consider 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 of heterogeneous and short time series that need to be adjusted. 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 nonseasonal 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 $\hat{Z}_{gct}^{seas}$ be some $\sqrt{T}$ consistent univariate estimate of the seasonal component of Z_{gct}. The seasonal adjustment error can be decomposed into a term $\hat{e}_{gct}^{seas}$ that is uncorrelated across counties c, and a term $\hat{\Phi}_{gct}$ that is correlated across c; namely,

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

Aggregating the data over counties, we have

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

While $\sum_{c=1}^{N_c} \hat{e}_{gct}^{seas}$ tends to zero as $N_c \to \infty$, the sum of $\hat{\Phi}_{gct}^{seas}$ 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 a model with two distinct types of common factors, seasonal and nonseasonal. Figure 14.1 suggests that the seasonal factors dominate.

This is also consistent with the finding in Ng (2017) that the principal components of budget shares constructed from data adjusted from the bottom up continue to exhibit seasonal variations.

14.4 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 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, Franses, and Paap (2007) consider a large T, large N_g panel of data and use a hierarchical Bayes method to avoid the proliferation of dummy variables needed to control for seasonal fixed effects. Like Fok, Franses, and Paap (2007), we also pool information across counties 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.

14.4.1 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 that 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 patterns even if one county has 10 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 are defined as:

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

where $\mu_{pc\tau}$ denotes the mean of log sales of group g in county c over the year τ containing week t, and $\mu_{pc\tau}$ 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:

$$\begin{aligned} y_{gct} &= (\text{countyspecificseasonalsales}) + (\text{commonseasonalsales}) \\ &\quad + (\text{non} - \text{seasonalsales}) \\ &= d_{gct} + q_{gct} + u_{gct}. \end{aligned}$$

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

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

where using strictly periodic predictors $\delta_{tj} = 2\pi j(\text{dayofyear}_t / \text{daysinyear})$ and $m_{tj} = 2\pi j(\text{Dayofmonth}_t / \text{daysinmount})$,

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

The regression only includes an intercept and will preserve any trends in sales that might be in the data. Hereafter, we will refer to step 1 as the Fourier regression.

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, especially when the predictors are the same across (g, c) pairs. Furthermore, there is a history in using Fourier regressions to first remove deterministic weekly seasonality, followed by ARMA modeling or local regressions to remove the stochastic seasonality one series at a time. But this approach is not practical when the number of series to fit is large. We use machine learning methods in the second step, and we pool information in the spatial dimension.[7]

Step 2: Estimation of q_{gct} from $\hat{\varepsilon}_{gct}$: Let $\hat{\varepsilon}_{gct} = \widehat{q_{gct} + u_{gct}}$ be the least squares residuals from step 1. Because this step is based on estimation of a smooth regression, these residuals will have spikes and can be cross-sectionally correlated. 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. This allows us to exploit cross-section dependence among counties to remove the within-year seasonal variation. In a nutshell, we pool information across counties to predict the common seasonal component q_{gct} using a large number of 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.

To complete the seasonal adjustment, let $\hat{q}_{gct}$ be the prediction of the common seasonality in $\hat{\varepsilon}_{gct}$ obtained from step 2. From these, we can obtain $\hat{u}_{gct}$, an estimate of u_{gct}. A seasonally adjusted value of log sales is obtained by plugging the estimated residual $\hat{u}_{gct}$ into

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

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

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

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

Inserting $\hat{u}_{gct}$ into (1) and inverting gives an alternative estimate of log-adjusted sales. We now elaborate on step 2.

14.4.1.1 Predictors $\mathbb{Z}_{gct}$

Regardless of the method used in step 1, we are limited to nine seasonal observations for both training and validation, so the seasonal adjustment is likely imperfect. To more thoroughly remove the seasonal effects, we need to first understand the nature of the seasonal variations in the data. Con-

[7] Cleveland and Devlin (1980) suggest using the spectrum to detect preidentified calendar and holiday effects in monthly data. For use of Fourier regressions in seasonal adjustment of weekly data, see Pierce, Grupe, and Cleveland (1984), Cleveland and Scott (2007), and Cleveland, Evans, and Scott (2014).

sider 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 on which Thanksgiving falls 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 because 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 $start_t$ denote the date on which week t starts, and 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 [$start_t$, end_t] and 0 otherwise. As an example, if t = Feb 4, 2006, the date-specific predictors $\mathbb{Z}^A_{gct}$ is as follows:

01/01	···	01/29	···	02/04	···	12/31
0	···	1	···	1	···	0

B. Week specific predictors.
   (i) $start_t$ and end_t's positions within the year (out of 366)
   (ii) $start_t$ and end_t's position within the months (out of 31)
   (iii) $start_t$'s position within the month containing 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 that is 1 if Easter is in the week ending on t, and 0 otherwise
   (v) A dummy variable encoding the month in which end_t falls

   For example, for t = Feb 4, 2006, the week -based predictors $\mathbb{Z}^B_{gct}$ 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 capture 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 end 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 base case.

14.4.1.2 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 τ, 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 τ. 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

$$\mathcal{D} = (\{\hat{\varepsilon}_{gct}\}, \{\mathbb{Z}_{gct}\}) = (\hat{\varepsilon}^{s}_{g\tau}, \mathbb{Z}^{s}_{g\tau}),\ \forall t : \text{yr}(t) = \tau$$

$$\mathcal{D}_{1\tau} = (\{\hat{\varepsilon}_{gct}\}, \{\mathbb{Z}_{gct}\}) = (\hat{\varepsilon}^{s}_{g\tau}, \mathbb{Z}^{s}_{g\tau}),\ \forall t : \text{yr}(t) \neq \tau,$$

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

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 $N_g \times N_{yr}$ 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. Though a pooled least-squares regression that uses all predictors (method 1) is possible, it will unlikely 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.

The random forest (RF), attributable 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 trees over nonparametric 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 i.i.d 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, the random forest is the most flexible because it does not impose linearity or smoothness. We use it as a benchmark in the discussion of results. We implement random forests using the R package RANGER with default parameter settings.[8] We find that the LARS-type methods do not uncover sparse models as our trained estimators have nonzero loadings on over 80 percent 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.

[8] 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.

14.5 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 differences between the unadjusted and adjusted shares are in groups like floral, insecticides, canning, ice, fragrances, toys, stationery, and candies. These results make sense because effects due to seasonal holiday events are precisely what we want to remove.

Table 14.3 uses two products to contrast the seasonal patterns in the raw and adjusted data. Consider first beer sales, which tend to be higher in the summer and peak around July 4th. In 2009, July 4th (week 183) fell on a Saturday when the Nielsen data were 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 at the top of table 14.3, the adjusted data are indeed smoother and exhibit a smaller spike than the raw data. Take the case of New York 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 the Superbowl. This is illustrated for 2009, when the Superbowl took place on Sunday, February 1. The adjusted shares are smoother within and between months.

It is also important that the second step adjustment does not remove

Table 14.3 **Effects of seasonal adjustment on selected series' share (%)**

	Adjusted data				Raw data			
Week Ending	CA	FL	NY	TX	CA	FL	NY	TX
	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

Table 14.4 **Importance of the seasonal component**

		Average of R^2 in equation (2)					
Sample	Method	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

spikes and variations that are nonseasonal. To check this, we consider the 62-cent federal tax hike on cigarettes on 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 14.3 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. In results not reported, we find that 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, suggesting that the tax did little to discourage cigarette consumption.

The premise of our analysis is that the residuals from the univariate Fourier regressions in step 1 have comovements 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 14.4 summarizes the distribution of R^2 over all groups and states. A little over 50 percent of the variations in log sales are seasonal and predictable. The degree of predictability varies across groups, ranging from 14 percent to over 90 percent. 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.

Figure 14.2 illustrates the difference between using step 1 alone and the two-step procedure by plotting the R^2 of random forest results against those based on the Fourier method. If the random forest estimator provides relatively little additional information, the optional step regression after step 2 will push λ_g toward zero. In such cases, the R^2 values will be bunched along the 45-degree line. Figure 14.2 indicates such groups do exist. However, many other groups have values in the scatterplot located above the 45-degree

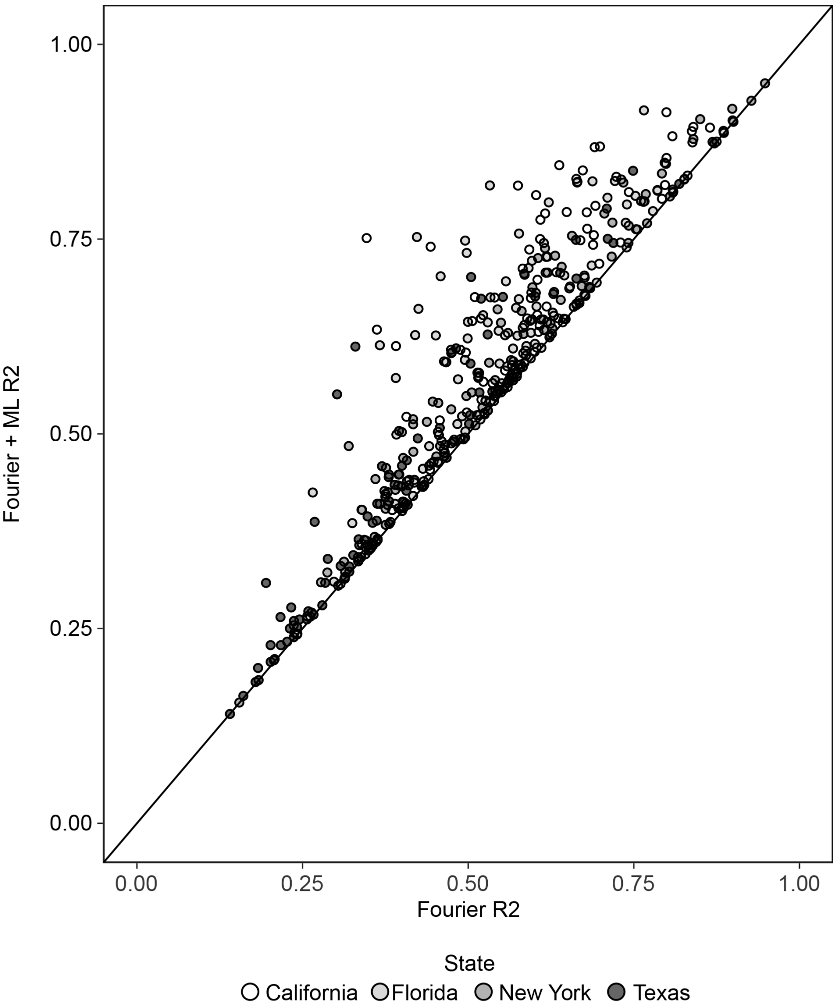


Fig. 14.2 Incremental predictive power of random forests

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 percent 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 have far fewer spikes than those adjusted using the Fourier regressions alone. This difference has direct implications for demand estimation.

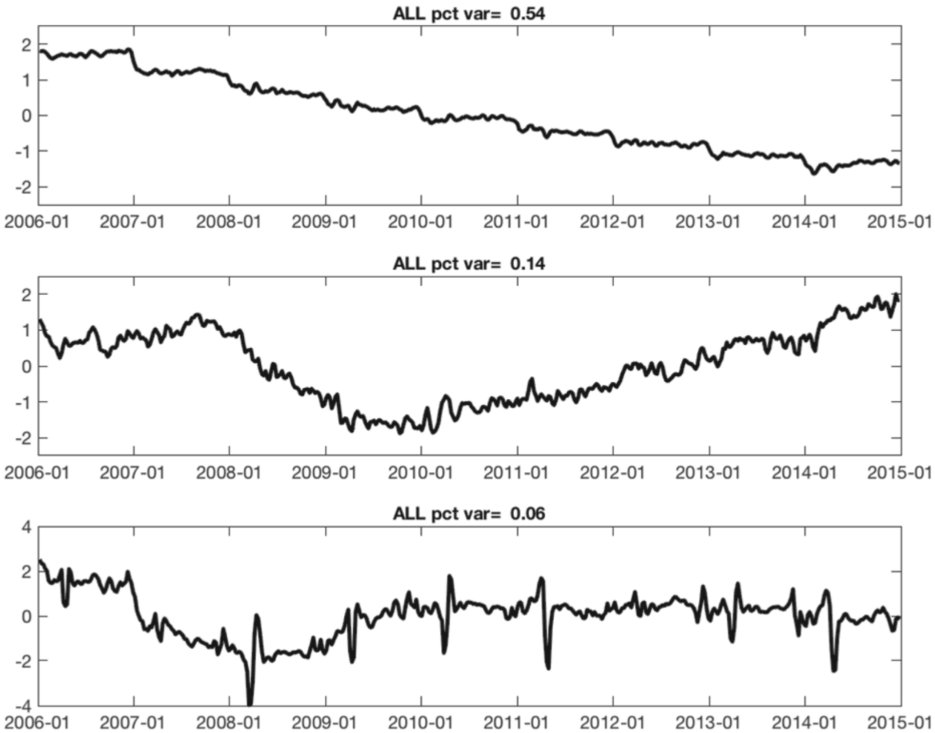


Fig. 14.3 Factors estimated from seasonally adjusted shares: FOUR states

14.5.1 The Factor Estimates

A main finding in the demand analysis of the raw data is that the first few factors exhibit strong seasonal patterns. 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 percent of the variation in the data and consist of a trend and a cyclical component. However, factors three and four remain spiky and quasiperiodic, indicating that the Fourier regressions by themselves leave nontrivial 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 the absence of large spikes.

Figure 14.3 plots the three factors in FOUR using data adjusted by random forests. These factors, denoted $\hat{F}_{\mathrm{RF}}$, are to be distinguished from the ones estimated from the unadjusted data, now denoted $\hat{F}_{\mathrm{NSA}}$. Though not immediately evident, $\hat{F}_{2,\mathrm{RF}}$ is strongly correlated with $\hat{F}_{4,\mathrm{NSA}}$. A regression of

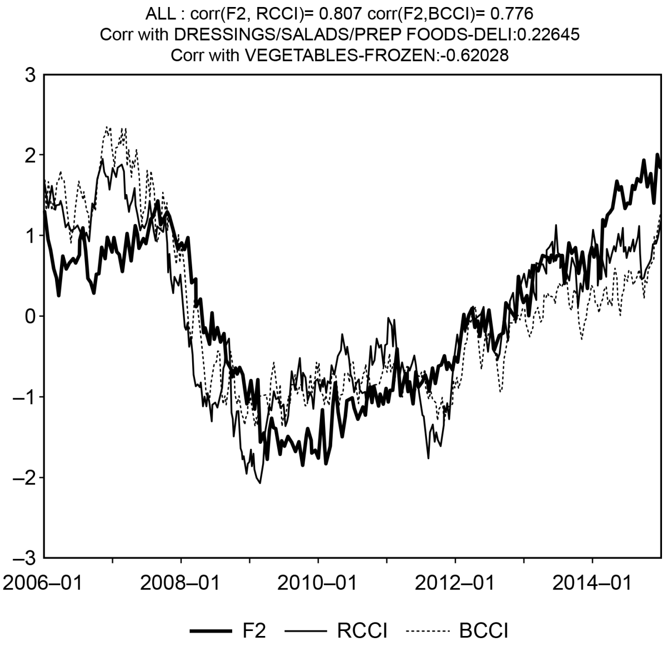


Fig. 14.4 The level factor: $\hat{F}_{2,RF}$

$\hat{F}_{4,\text{NSA}}$ 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 percent of the variations of the adjusted shares, with $\hat{F}_{1,RF}$ explaining 56 percent, and $\hat{F}_{2,RF}$ explaining 15 percent. As can be seen from figure 14.3, $\hat{F}_{1,RF}$ 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 during our sample.

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 as benchmarks of cyclicality: 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 about the US economy. The latter started as the ABC News consumer comfort index and has been under the control of the Bloomberg Corporation since 2011. Figure 14.4 plots $\hat{F}_{2\text{RF}}$ (thick solid line), RCCI (thin solid line), along with BLOOMBERG (dotted line). It is evident that spending moves posi-

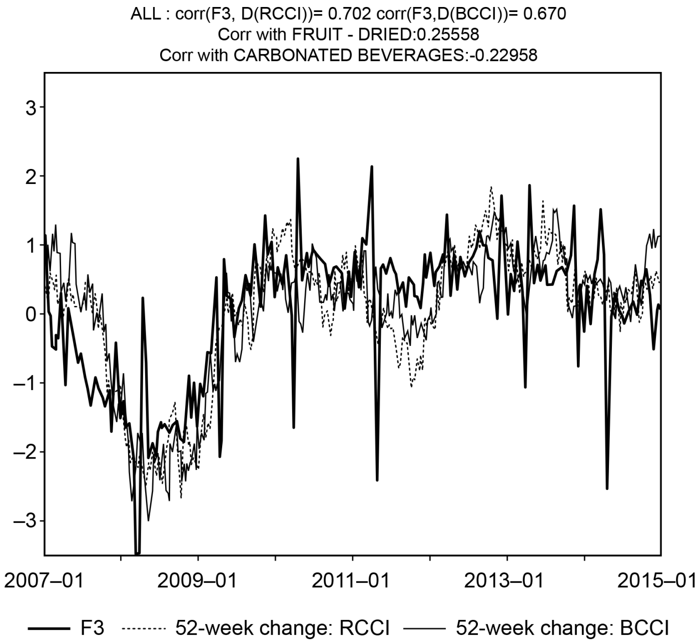


Fig. 14.5 The curvature factor: $\hat{F}_{3,RF}$

tively with consumer sentiment. These confidence measures have absolute correlation with $\hat{F}_{2,\mathrm{RF}}$ of about 0.8. In this regard, consumers' actions are aligned with how they feel. Because our data cover 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, 2008. As a point of reference, JP Morgan purchased Bear Stearns on March 17, 2008. Furthermore, oil prices spiked up to nearly \$110 per barrel a few days earlier. Upon examination, the factor is actually highly correlated with the 52-week change in consumer confidence. Figure 14.5 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

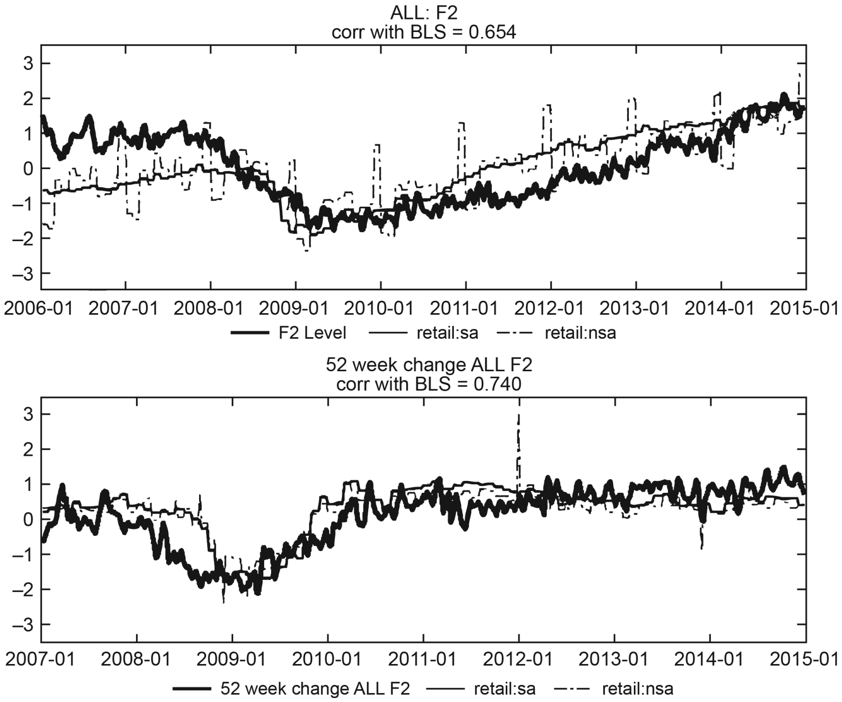


Fig. 14.6 Comparison with monthly retail series

right because the 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 US Census Bureau releases both the raw and seasonally adjusted data for retail sales each month.[9] To compare with our weekly series, we interpolate values for the weeks in a month to the officially released sales for the month. Figure 14.6 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 the 52-week change in the series. The correlation of the adjusted series is 0.74. The most notable difference is seen 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 do not cleanly line up with the monthly calendar.

The results so far have focused on four states: CA, FL, NY, TX. How-

[9] The series are RETAILSMNSA and RETAILSMSA in FRED.

ever, 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 Southwest (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 gives three factors with very similar properties. Hereafter, we use the extended data when appropriate. These results will be labeled SEVEN and ALL.

14.6 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 timing of the response of spending to economic conditions at a disaggregated level. Subsection 14.6.1 considers cyclical sensitivity of product groups, while subsection 14.6.2 considers spatial variations in spending.

14.6.1 Variation across Product Groups

We first turn to the sensitivity of the product groups to business cycle conditions. Since $\hat{F}^s_{2,RF}$ is positively correlated with RCCI, a positive loading indicates that the share of product j is procyclical, while a negative value means that the share of product j is high when $\hat{F}^s_{2,\mathrm{RF}}$ is low. The dispersion of sensitivity to aggregate conditions across product groups is best seen from the distribution of $SGNR^2_{sj}$. This is defined as the signed fraction of variance of $SHARE^s_j$ explained by $\hat{F}^s_{2,\mathrm{RF}}$, where SGN is the sign of the loading of $\hat{F}^s_{2,RF}$ on $SHARE^s_j$. Though there are some minor differences across states and regions, the pattern across states is broadly similar. Figure 14.7 presents results for SEVEN. The distribution is noticeably asymmetric because there are more countercyclical product groups and the magnitude of the absolute loadings are larger (top) than procyclical ones (bottom). Product groups little affected by $\hat{F}_{2,RF}$, plotted in the middle of figure 14.7, 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 $\hat{F}_{2,RF}$ has the largest marginal impact on the share of frozen vegetables, canned vegetables, and pasta. The impact of an increase in $\hat{F}_{3,RF}$ 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

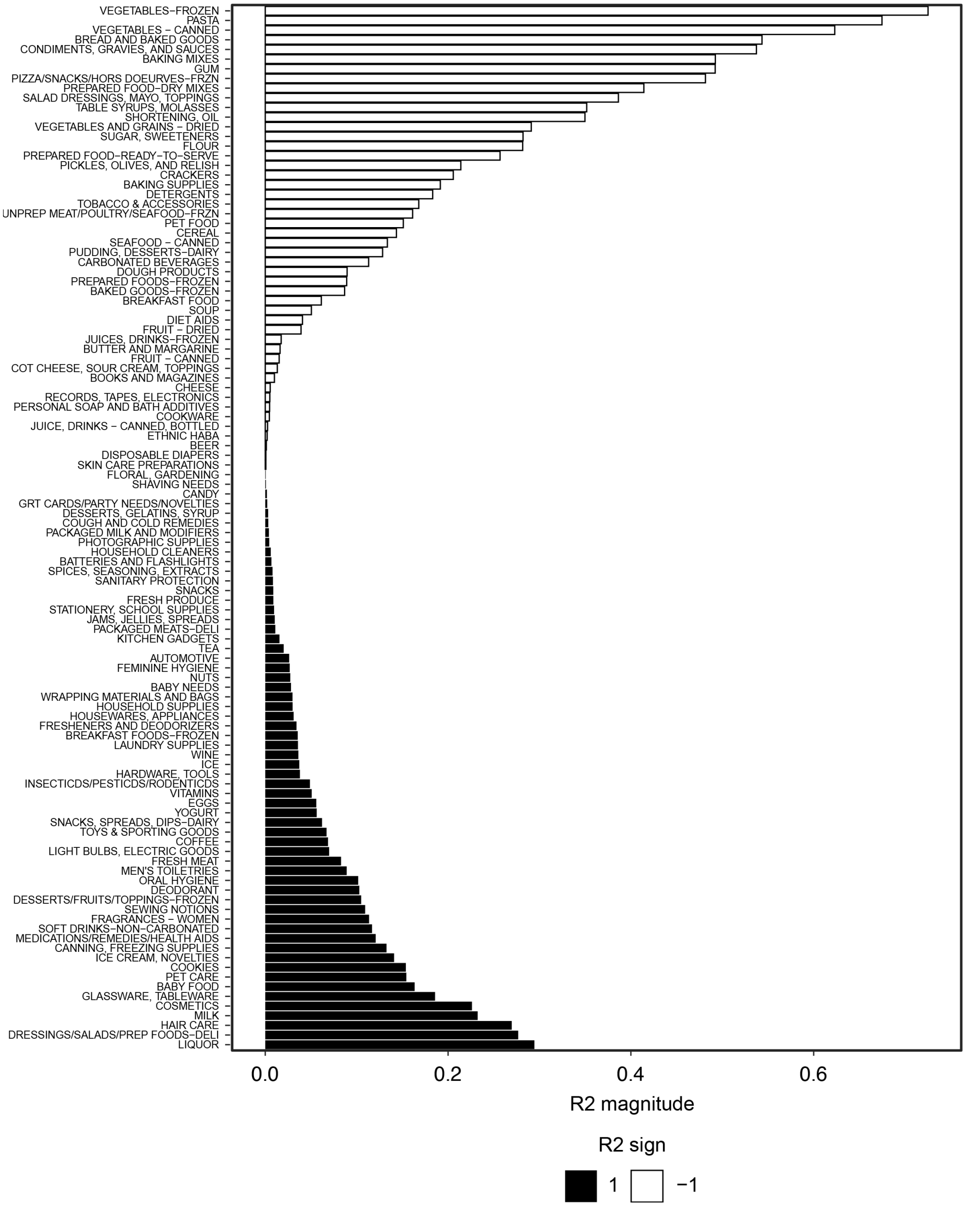


Fig. 14.7 R^2 **from regression of adjusted shares on** $\hat{F}_2$**: FOUR states**

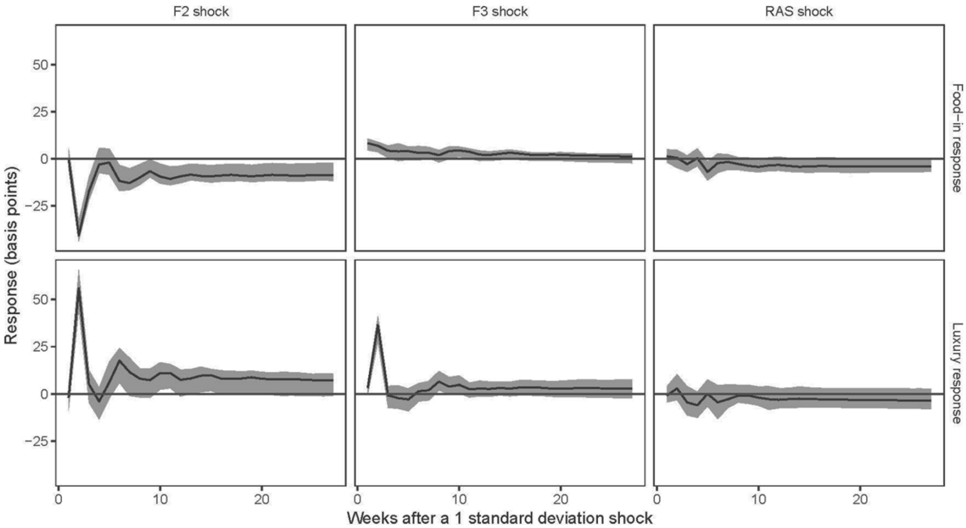


Fig. 14.8 Response of FOOD-IN to shock in $\hat{F}_2$

recovery was slow.[10] Cha, Chintagunta, and Dar (2015) aggregate the weekly Homescan data to annual level and find that food consumed at home is countercyclical. Grittith, O'Connell, and Smith (2015) find that households also adjusted food spending in the UK. 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 seem related to home cooking, we designate them the FOOD-IN group. We also identify the five products with large positive loadings on $\hat{F}^s_{2,\mathrm{RF}}$: 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 conventionally defined.

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 14.8. 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

[10] See https://ageconsearch.umn.edu/bitstream/120969/2/10FoodSpending.pdf.

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 percent of the variations in FOOD-IN are explained by its own lag, 35 percent explained by $\hat{F}_{2,RF}$, 7 percent by $\hat{F}_{3,RF}$, with little attributed to RCCI. About 37 percent of the variations in LUXURY are explained by its own lag, 28 percent by $\hat{F}_{2,RF}$ and 32 percent 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 (i.e., 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 $\hat{F}_{3,RF}$.

14.6.2 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 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 the consumption basket. The top panel of table 14.5 reports the CPI weights for *food consumed at home* and *luxury* as defined by the BLS. In their own study

Table 14.5 **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

on how consumer spending changes during boom, recession, and recovery, 2007 was used as a boom year, 2011 as recession, and 2013 a year of recovery.[11] 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 14.5 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.

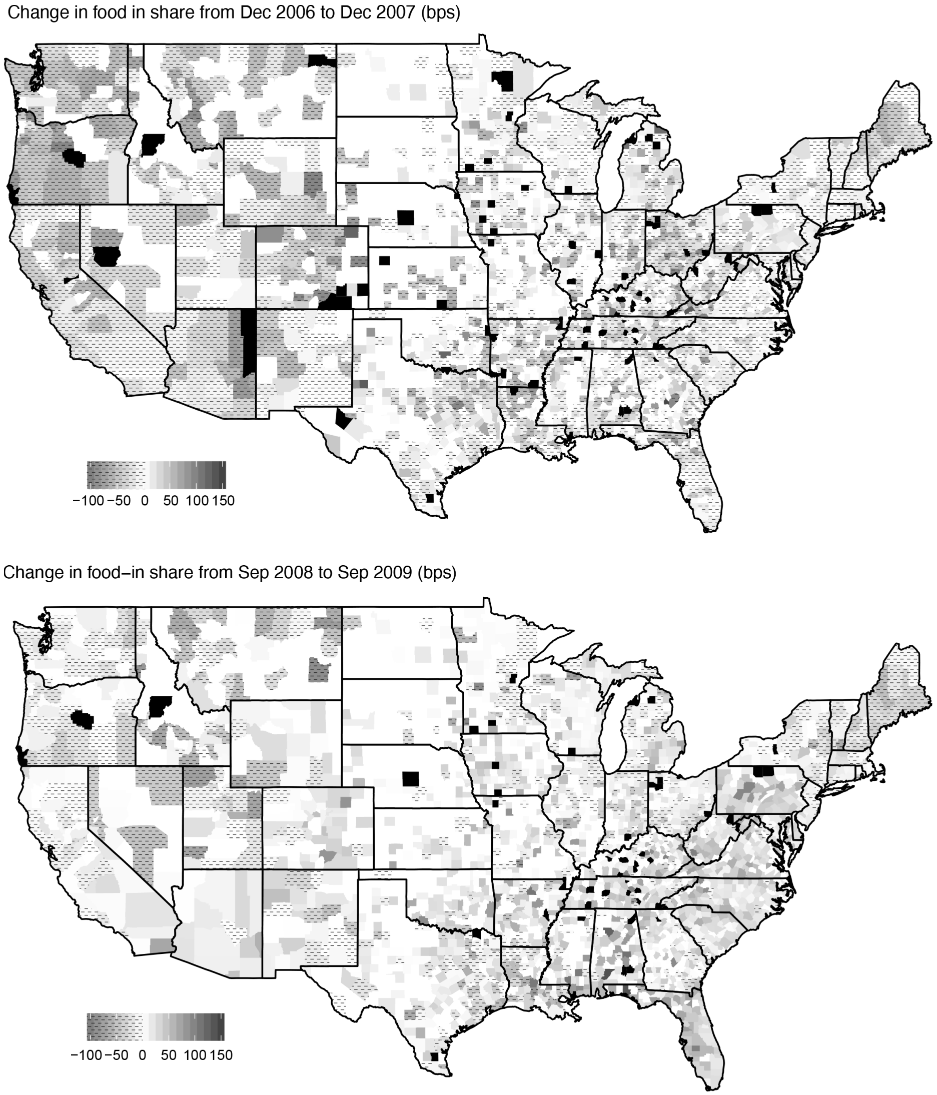
An appeal of the Nielsen data is that they provide granular information in both the time series and cross-section dimensions. The share of FOOD-IN ranges between 5 percent in Florida to 7.8 percent in the Mid-Atlantic regions, with an average of 6.6 percent over the entire sample. The series is most persistent in California and least persistent in the Midwest, with first order autocorrelation coefficients of 0.83 and 0.50, respectively. The share of LUXURY ranges between 6.6 percent in the Mid-Atlantic regions to 12 percent in Florida, with an average of 8.9 percent over the full sample. The series is most persistent for the Midwest and least persistent in the Southwest, with autocorrelation coefficients of 0.86 and 0.5, respectively. The contemporaneous correlation between FOOD-IN and LUXURY is strongly negative in California, New York, and the Midwest, with cross-correlations in excess of 0.6 in absolute value. The correlation is much weaker in the Southwest and even positive in Florida. The heterogeneity across states in spending behavior underscores the difficulty in designing policies that would satisfy all consumers.

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

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

The R^2 provides a measure exposure of county c to the two common factors. Upon ranking the R^2s, the urban and densely populated counties are found to be more exposed to aggregate shocks. Take the state of New York as an example. The counties of Rockland, Nassau, and Kings 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^2s of at most 0.01.

[11] See https://www.bls.gov/opub/btn/volume-3/how-does-consumer-spending-change-during-boom-recession-and-recovery.htm.


Change in food in share from Dec 2006 to Dec 2007 (bps)

Change in food–in share from Sep 2008 to Sep 2009 (bps)

Fig. 14.9 Regional changes in FOOD-IN

Heatmaps provide a more compact way to see how different regions are affected by economic conditions. The top panel of figure 14.9 plots the change in FOOD-IN between 2006 and 2007. Regions with dotted gray shading 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 solid gray indicates larger increases in FOOD-IN share. Now there are more regions shaded solid gray than dotted gray, with Arizona and Florida witnessing the largest increase in FOOD-IN. This shows that the Great Recession differentially affected regional purchasing behavior of FOOD-IN goods.

14.6.3 Sandy Regression

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. To illustrate, we take advantage of the weekly and spatial information in the Nielsen data to examine purchasing behavior in New York around landfall of Hurricane Sandy on Monday, October 29, 2012.

In hindsight, Sandy was a much bigger storm than expected and consumers were caught somewhat unprepared. Figure 14.10 shows little evidence of stocking up during the week prior to Sandy, but that there was a distinct increase in FOOD-IN share during the week containing the storm. One might be concerned that the increase in the raw data shown in the top panel is an artifact of seasonality as the week ending November 3rd was close to the beginning of the Thanksgiving and Christmas shopping season. But the bottom panel shows that when using the seasonally adjusted data, there is a clear post-Sandy spike in 2012, which brings the seasonally adjusted FOOD-IN share to its highest value for the year.

To quantify the impact of Sandy, we estimate a simple panel data model. Let $y_{i,t}$ be the share of FOOD-IN in county i and week t, normalized to have standard-deviation 1 within each county. Let $sandy - county_i$ be a dummy variable that indicates if i is a coastal county that was hit by Hurricane Sandy. Let $landfall_t$ be a dummy variable that indicates if t is the week containing the landfall of Hurricane Sandy, which is the week ending November 3, 2012. We estimate the regression

$$y_{it} = \alpha_i + \lambda_t + \sum_{j=0}^{5} \beta_j \cdot sandy - county_i \times landfall_{t-j} + \text{error}.$$

Our results show that FOOD-IN consumption increases by about 2.5 standard deviations during the week that Sandy made landfall. The effects of Sandy on FOOD-IN purchases persisted for about one month. Figure 14.11 shows that the effects of Sandy were localized to the counties near New York City and Long Island, which were most exposed to the hurricane. Other counties in the state of New York were nearly unaffected by the storm.

14.7 Conclusion

Large volumes of highly heterogeneous data are increasingly available, but they are often not immediately useful for economic analysis without remov-

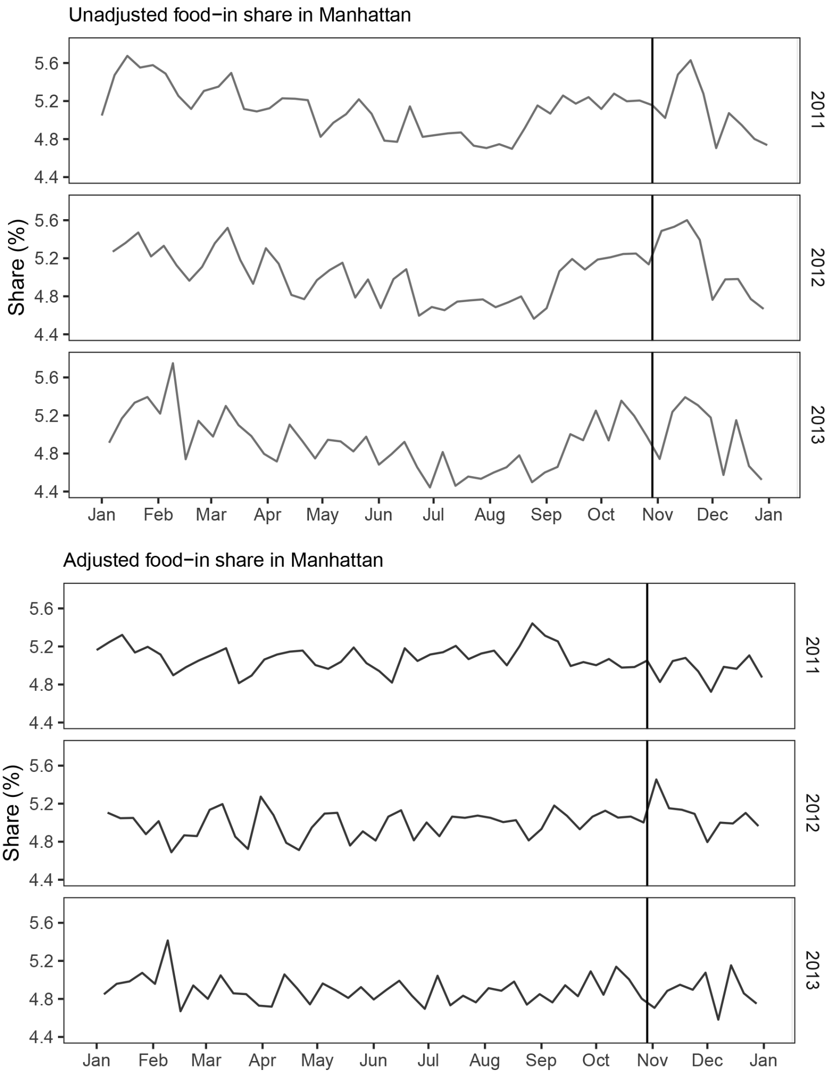


Fig. 14.10 Unadjusted and adjusted FOOD-IN share in Manhattan for 2011–2013
Note: Vertical line denotes October 29 (the date of Hurricane Sandy's landfall).

Table 14.6 Consumption increases

j	0	1	2	3	4	5
$\hat{\beta}_j$	2.541*** (0.242)	0.318 (0.203)	0.152 (0.202)	0.323** (0.153)	–0.606*** (0.166)	–0.123 (0.197)

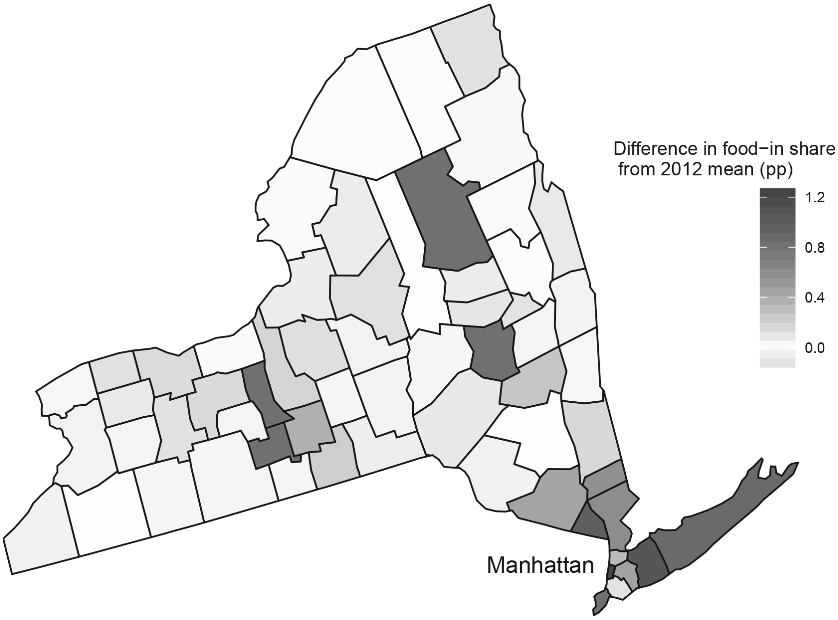


Fig. 14.11 Difference between $y_{i,landfall}$ the FOOD-IN share for the week containing Hurricane Sandy's landfall, and $\bar{y}_{i,2012}$, the average FOOD-IN share for 2012, by county in New York state

ing 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. The validity of this second step relies on the presence of common seasonal patterns across counties.

There is no shortage of examples in which common seasonality would be a feature of the raw data. For example, employment and 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. Informal discussions with staff researchers at the Bureau of Economic Analysis confirm such experiences. Our analysis provides an explanation for why a bottom-up approach to seasonality might be inadequate. 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. Though our focus has been on handling seasonal effects, the procedure can be adapted to remove other nuisance variations. A limitation of our analysis is the lack of a way to assess sampling uncertainty of the two-step procedure. This is left for future research.

References

Bai, J., and S. Ng. 2008. "Large Dimensional Factor Analysis." *Foundations and Trends in Econometrics* 3 (2): 89–163.

———. 2019. "Rank Regularized Estimation of Approximate Factor Models." *Journal of Econometrics* 212 (1): 78–96.

Banks, J., R. Blundell, and A. Lewbel. 1997. "Quadratic Engel Curves and Consumer Demand." *Review of Economics and Statistics* 79:527–39.

Breiman, L. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32.

Cha, W., P. Chintagunta, and S. Dhar. 2015. "Food Purchases during the Great Recession." Kilts Booth Marketing Series, Paper 1-008. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2548758.

Chamberlain, G. 1984. "Panel Data." In *Handbook of Econometrics*, vol. 2, 1247–1318. Amsterdam: North Holland.

Cleveland, W., and S. Devlin. 1980. "Calendar Effects in Monthly Time Series: Detection by Spectrum Analysis and Graphical Models." *Journal of the American Statistical Association* 75 (371): 487–96.

Cleveland, W., T. Evans, and S. Scott. 2014. "Weekly Seasonal Adjustment—A Locally Weighted Regression Approach." Working Paper No. 473, US Bureau of Labor Statistics, Washington, DC.

Cleveland, W., and S. Scott. 2007. "Seasonal Adjustment of Weekly Time Series with Application to Unemployment Insurance Claims and Steel Production." *Journal of Official Statistics* 23 (2): 209–21.

Deaton, A. S., and J. Muellbauer. 1980. "An Almost Ideal Demand System." *American Economic Review* 70:312–26.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani. 2004. "Least Angle Regression." *Annals of Statistics* 32 (2): 407–99.

Engle, R. F., and S. Kozicki. 1993. "Testing for Common Features." *Journal of Business and Economic Statistics* 11:369–79.

Fok, D., P. Franses, and R. Paap. 2007. "Seasonality and Non-Linear Price Effects in Scanner Data-Based Market Response Models." *Journal of Econometrics* 138:231–51.

Geweke, J. 1978. "The Temporal and Sectoral Aggregation of Seasonally Adjusted Time Series." In *Seasonal Analysis of Economic Time Series*, edited by A. Zellner, 411–32. Washington, DC: National Bureau of Economic Research.

Gorman, W. M. 1981. "Some Engel Curves." In *Essays in the Theory and Measurement of Consumer Behavior in Honor of Sir Richard Stone*, edited by A. Deaton. Cambridge: Cambridge University Press.

Grittith, R., M. O'Connell, and K. Smith. 2015. "Shopping Around: How Households Adjusted Food Spending over the Great Recession." *Economica* 83 (330): 247–280. https://onlinelibrary.wiley.com/doi/abs/10.1111/ecca.12166.

Harvey, A., S. Koopman, and M. Riana. 1997. "The Modeling and Seasonal Adjustment of Weekly Observations." *Journal of Business and Economic Statistics* 15:354–68.

Lewbel, A. 1991. "The Rank of Demand Systems: Theory and Nonparametric Estimation." *Econometrica* 59 (1): 711–30.

———. 1997. "Consumer Demand Systems and Household Equivalence Scales." In *Handbook of Applied Econometrics*, vol. 2, edited by M. H. Pesaran and P. Schmidt, 167–201. Oxford: Blackwell.

———. 2003. "A Rational Rank Four Demand System." *Journal of Applied Econometrics* 18:127–35.

McElroy, T. 2017. "Multivariate Seasonal Adjustment, Economic Identities, and Seasonal Taxonomy." *Journal of Business and Economic Statistics* 35:511–25.

Morwitz, V., and G. Fitzsimons. 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* 62:269–83.

Ng, S. 2017. "Opportunities and Challenges: Lessons from Analyzing Terabytes of Scanner Data." In *Advances in Economics and Econometrics*, Eleventh World Congress of the Econometric Society, vol. 2, edited by B. Honoré, A. Pakes, M. Piazzesi, and L. Samuelson, 1–34. Cambridge: Cambridge University Press.

Pierce, D., M. Grupe, and W. Cleveland. 1984. "Seasonal Adjustment of the Weekly Monetary Aggregate: A Model Based Approach." *Journal of Business and Economic Statistics* 2:260–70.