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

NBER/CRIW Conference

Big Data for 21st Century Economic Statistics

March 2019

Rishab Guha Harvard University
Serena Ng Columbia University and NBER

Outline

1. Goal
2. Demand Analysis Pre-adjustment
3. A Two-Step Approach
4. Demand Analysis Post-adjustment
5. Cyclical Sensitivity

Overview

- Generic problem: economic info hidden in VVV data, need to
  - remove some type of nuisance variations (here, seasonality)
  - aggregate data over some dimension (here, counties)
  - univariate procedures do not well.
- Proposed procedure
  - start with some simple univariate filter.
  - exploit cross section dependence to mop up residual nuisance variations. Automate using machine learning tools.
  - remove 'enough' so that economic insights can be obtained.
- Application: Nielsen Scanner Data

Disclaimer: Nielsen Scanner

- Calculated (or Derived) based on data from The Nielsen Company (US), LLC and marketing databases provided by the Kilts Center for Marketing Data Center at The University of Chicago Booth School of Business.
- The conclusions drawn from the Nielsen data are those of the researchers 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.

Nielsen Scanner Data

	county	group	week	year
index	$c(s)$	g	t	τ
total	$50 \leq N_c \leq 2000$	$N_g = 108$	$T = 469$	$N_{\text{yr}} = 9$

- find stores with at least one sale each week
- At each week t, the budget share of group $g \in [1, N_g]$ is

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

Budget Shares: Most Purchased Categories

CA: $N_c = 53$		FL: $N_c = 58$		NY: $N_c = 58$		TX: $N_c = 161$	
3.4	bread	4.4	medications	4.1	medications	3.7	carb. bev
3.3	beer	4.3	tobacco	3.2	fresh prod.	3.7	medications
3.3	juice	3.1	carb. bev.	3.1	bread	3.4	snacks
3.2	wine	2.9	liquor	3.0	candy	2.9	bread
3.0	fresh prod.	2.8	beer	2.8	snacks	2.8	tobacco
3.0	carb. bev	2.6	juice	2.8	juice	2.6	pkgd meat
3.0	snacks	2.6	candy	2.7	tobacco	2.6	candy
2.7	pkgd meat	2.4	snacks	2.5	beer	2.5	fresh prod.
2.7	salad dress.	2.3	milk	2.4	carb. bev	2.5	juice
2.6	medication	2.6	bread	2.3	milk	2.5	beer

Traditional Demand Analysis

- Approximate expenditure function eg. LES, translog.
- Impose restrictions of consumer theory.
- P imposes cross-equation restrictions. Proxy simplifies.

e.g. AIDS (Deaton and Muellbauer 1980):

$$\text{share}_g = \lambda_{0g} + \sum \lambda_{jg} \log p_g + \beta_g \log(\mathrm{Y}/P^*) + error_g .$$

- Stone's price index: $\log P^* = \sum_j w_j \log p_j$
- Classical estimation: T large, N_g small.
- Using seasonally adjusted data, rank of demand system typically estimated to be no larger than 4.

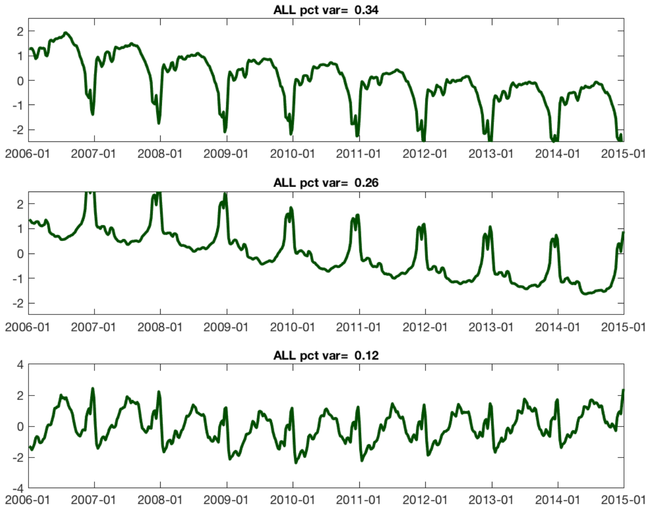
A Large N_g large T Approach

- Nielsen data: $T = 469$, $N_g = 108$ for each s.
  - A factor analytic approach to demand analysis:

$$\text{share}_{gt} = F_t' \Lambda_g + e_{gt}.$$

  - Principal components can consistently estimate F up to a rotation matrix without using price/income data
  - Non-parametric in economic and econometric sense.
  - Rank of demand system in Nielsen data $>> 5$. Why?

Factors Estimated from Raw Data: four



Strong and heterogeneous seasonal effects!

Dealing with Seasonality

- The general Q4 effect
  - Spending is concentrated in the last 6 weeks of year.
  - Entry-exit is seasonal: more goods introduced in Q4.
- 3 challenges specific to Nielsen data
  - i Weekly data: not exactly periodic, (Gregorian calendar).
    - Earliest Easter: March 23, 2008, latest Easter, April 24, 2011.
  - ii Volume and heterogeneity: one model will not fit all.
  - iii Data are spiky. Promotional sales.
- 52 week differencing does not work well.

Unlike with official data, user has to deal with all these problems.

$\textbf{sales}_{gct} = \textbf{sales}^{nseas}_{gct} + \textbf{sales}^{seas}_{gct}$

- Univariate (parametric) procedures: X13, SEATS/TRAMO
- Perfect seasonal adjustment unlikely
    - i Spikes from holiday sales move around over the years.
    - ii Smooth functions are not good at picking up spikes.
    - iii Span of data is short. Finite sample bias.
    - iv Hard to tune $N_c = 2000 \times 108$ models
- If $\text{sales}^{seas}_{gct}$ are correlated across c (counties), the residual will be cross-sectionally dependent.
- Each individual series might appear de-seasonalized, but seasonality re-appears after aggregation.

Proposed Approach

- Key observation: sales of group g in neighboring counties have similar seasonal patterns regardless of county size.

$$y_{gct} = \overbrace{\underbrace{d_{gct}}_{\text{county specific seasonal}} + \underbrace{q_{gct}}_{\text{common across counties seasonal}}}^{\text{seasonal}} + \overbrace{u_{gct}}^{\text{non-seasonal}}$$

- Key Assumption: d_{gct} and q_{gct} are predictable.
- Treat seasonal adjustment as a prediction problem.

Overview: $y_{gct} = d_{gct} + q_{gct} + u_{gct}$

Step 1: For each (g, c) pair: Fourier regression

$$y_{gct} = \underbrace{\alpha^0_{gc} + \text{Fourier}_{gct}(\beta_{gc}, \psi_{gc})}_{d_{gct}} + \underbrace{\epsilon_{gct}}_{q_{gct}+u_{gct}}$$

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

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

Step 2: pool information across counties and years

- train algorithms to predict q_{gct} from $\widehat{q_{gct} + u_{gct}}$.

$$
\begin{pmatrix}
y_{1,1} & y_{1,2} & \cdots & y_{1,N_g} \\
y_{2,1} & y_{2,2} & \cdots & y_{2,N_g} \\
\vdots & \vdots & \vdots & \\
y_{52,1} & y_{52,2} & \cdots & y_{52,N_g} \\
\hline
y_{53,1} & y_{53,2} & \cdots & y_{53,N_g} \\
\vdots & \vdots & \vdots & \\
y_{104,1} & y_{104,2} & \cdots & y_{104,N_g} \\
\hline
\vdots & \vdots & \vdots & \\
\vdots & \vdots & \vdots & \\
\hline
y_{417,1} & y_{417,2} & \cdots & y_{417,N_g} \\
\vdots & \vdots & \vdots & \\
y_{469,1} & y_{469,2} & \cdots & y_{469,N_g}
\end{pmatrix}
$$

columns= 2000+ counties $\times$ 108 product groups

$$
\begin{pmatrix}
y_{1,1} & y_{1,2} & \cdots & y_{1,N_g} \\
y_{2,1} & y_{2,2} & \cdots & y_{2,N_g} \\
\vdots & \vdots & \vdots & \\
y_{52,1} & y_{52,2} & \cdots & y_{52,N_g} \\
\hline
y_{53,1} & y_{53,2} & \cdots & y_{53,N_g} \\
\vdots & \vdots & \vdots & \\
y_{104,1} & y_{104,2} & \cdots & y_{104,N_g} \\
\hline
\vdots & \vdots & \vdots & \\
\vdots & \vdots & \vdots & \\
\hline
y_{417,1} & y_{417,2} & \cdots & y_{417,N_g} \\
\vdots & \vdots & \vdots & \\
y_{469,1} & y_{469,2} & \cdots & y_{469,N_g}
\end{pmatrix}
$$

Train one model for each of 108 products for each year:

Step 2: Seasonality Adjustment as a Prediction Problem

- Control for multidimensional seasonal heterogeneity using lots of dummy predictors using a flexible seasonality function.
  - Intuition from LSDV: incidental parameter if T is short.
  - Fok, Franses, Paap (2007): hierarchical structure, Bayesian.
  - We use algorithms choose predictors and functional form.
- Many (391) predictors
  - i (base set) all date-specific dummies: holidays, sports events.
  - ii social-economic indicators at county level.
  - iii weather and location from NOAA.
  - iv interaction of (i) and (ii).

step 2: Machine Learning/Regularization

a. Pooled OLS, non-regularized, no averaging.

b. LARS type algorithms.

- average over sequentially constructed predictions.
- solution path similar to Lasso.
- learner = linear model. Averaging reduces bias.

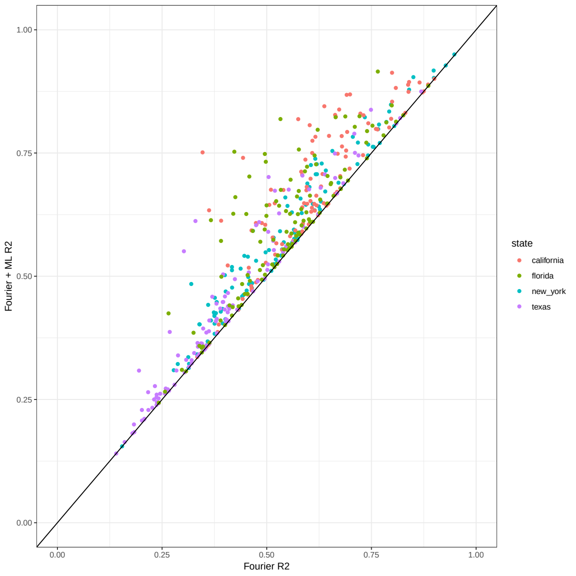
c. Random forest/bagging type algorithms.

- average over predictions from randomly chosen predictors.
- learner = regression tree. Averaging reduces variance.

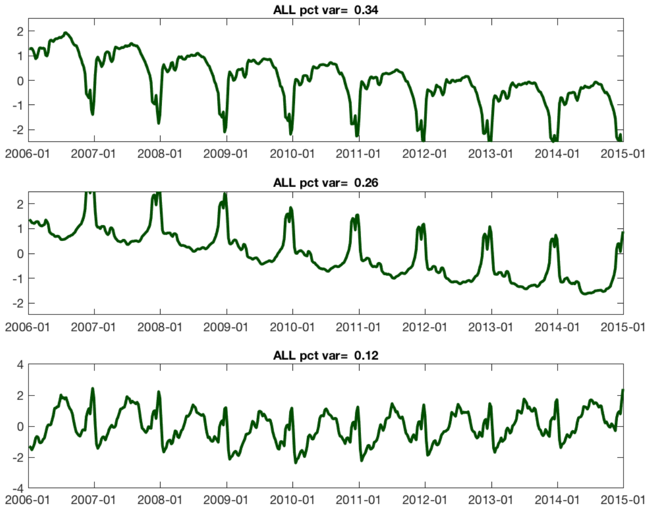
Effects of the Adjustment on Selected Shares

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 Purchases							
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

Incremental Predictive Power of Random Forests

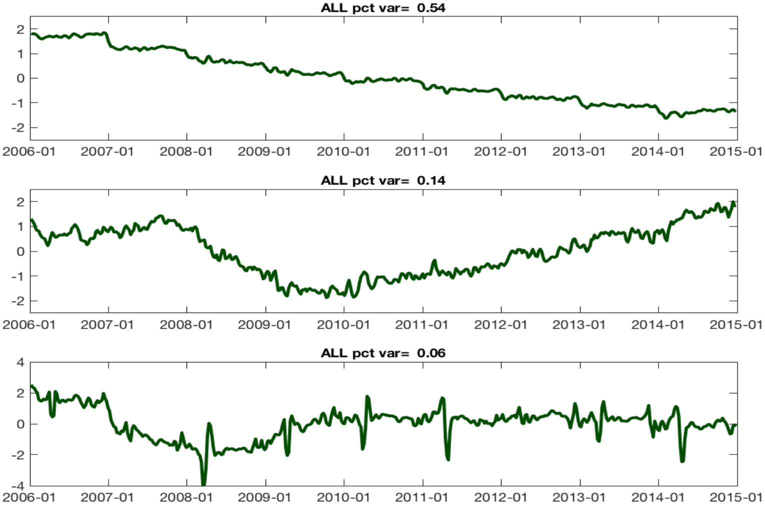


Factors Estimated from Raw Shares: All States



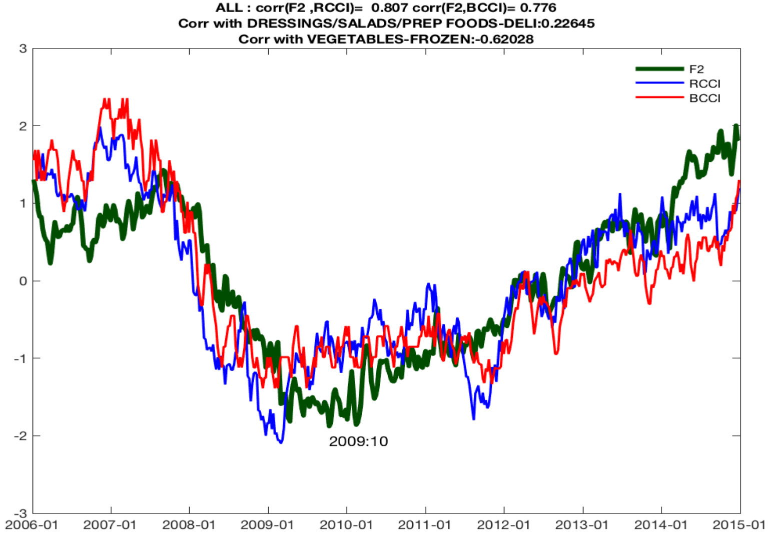
Strong and heterogeneous seasonal effects! Where is the cycle?

Factors Estimated from Adjusted Shares: ALL States

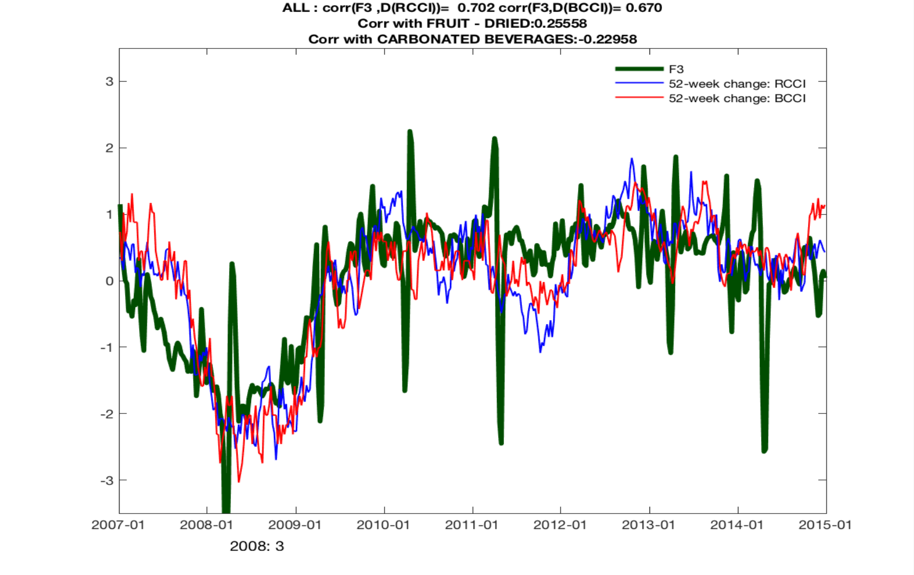


3 Factors: Trend, Level, Slope.

The Level Factor



The Slope Factor



Comparison with the BLS Monthly data

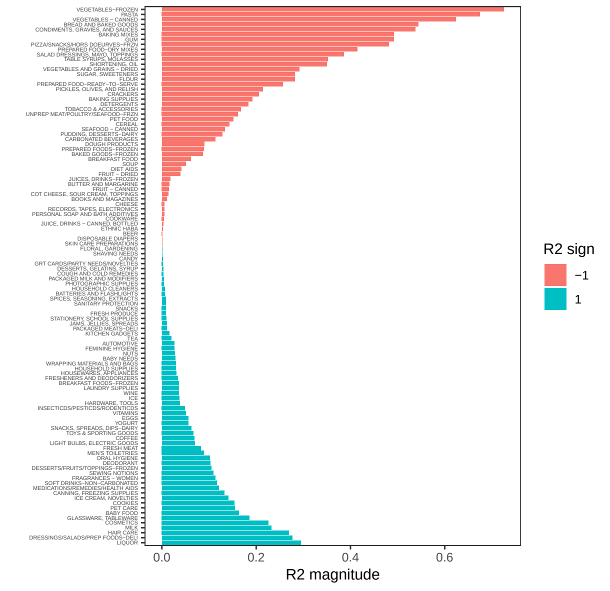
Sensitivity to Cyclical Factor: group level

- Product and regional level information at weekly frequency make the data unique.
- Which product groups are recession-proof?

  (group,week) panel regression:

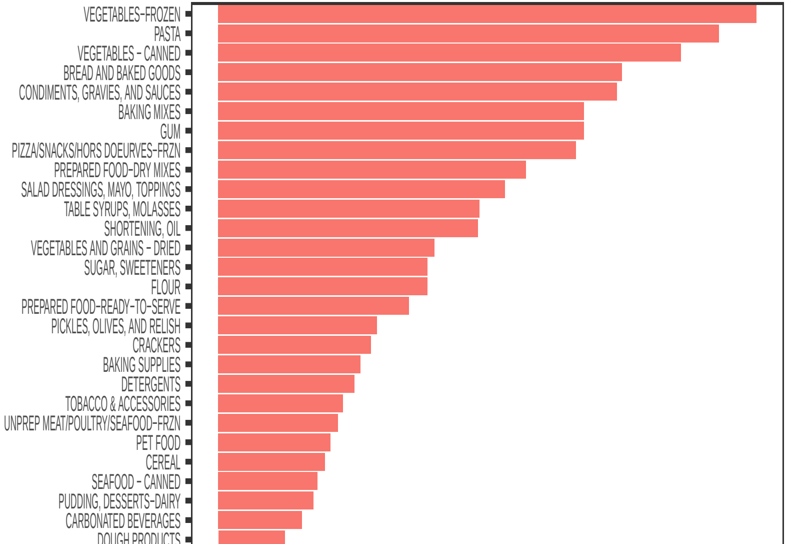
$$share_{gt} = a_0 + a_1 \widehat{F}_{2,RF,t} + error_{gt}$$

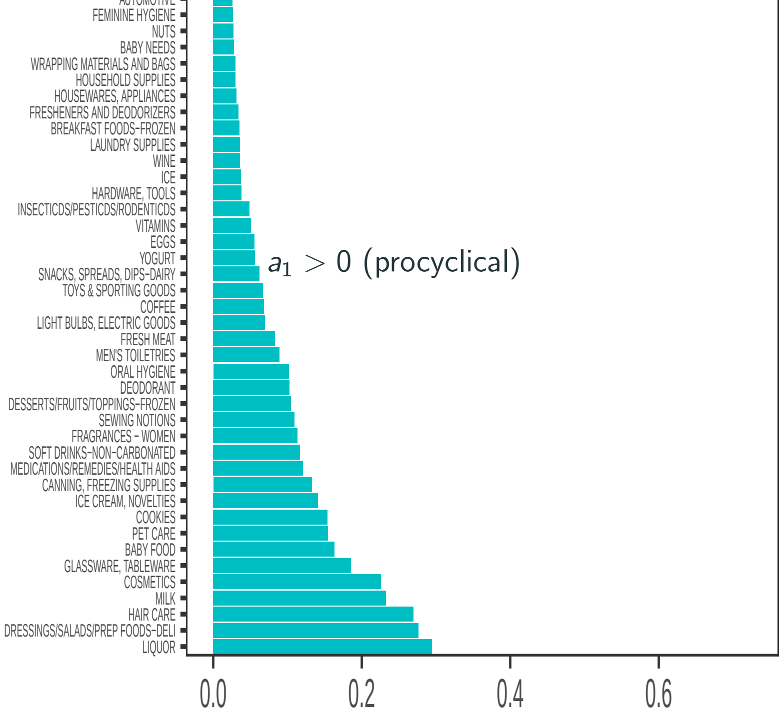
$R^2 \approx 0$: beer, cough/cold remedies, disp. diapers, batteries



Distribution of R^2

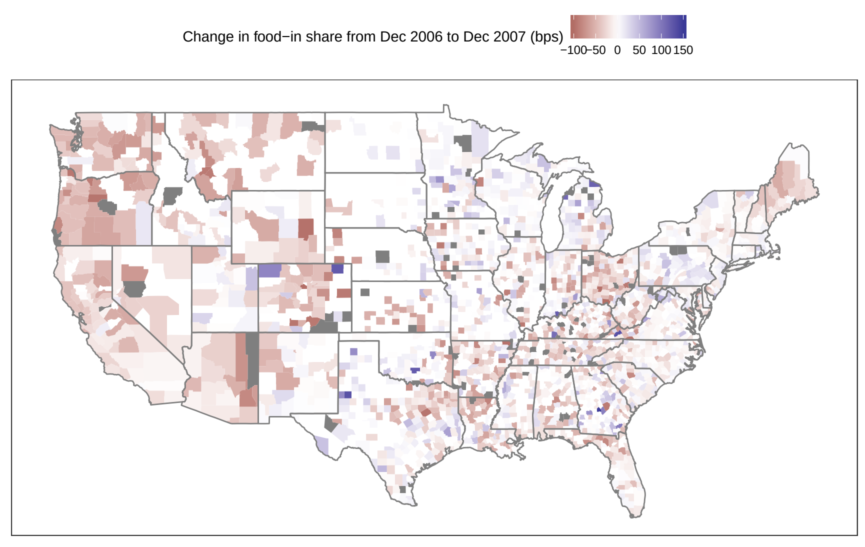
$a_1 < 0$ (countercyclical)



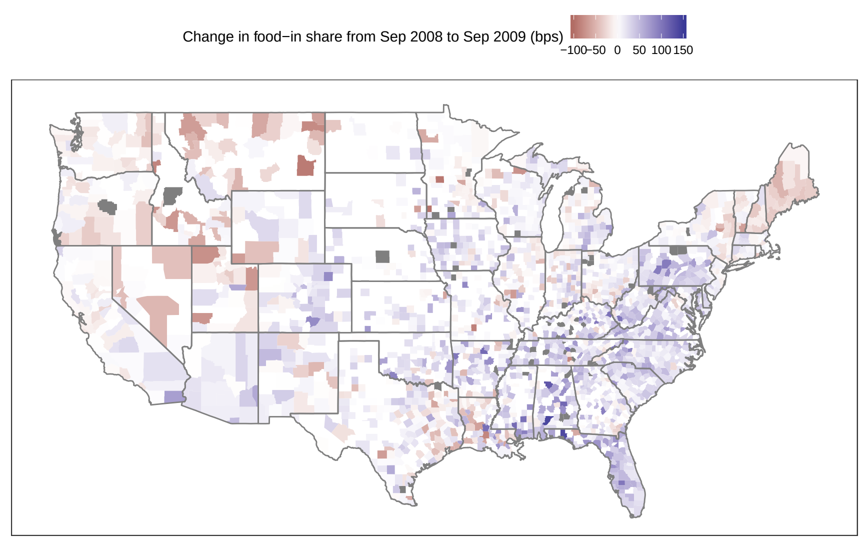


FEMININE HYGIENE
NUTS
BABY NEEDS
WRAPPING MATERIALS AND BAGS
HOUSEHOLD SUPPLIES
HOUSEWARES, APPLIANCES
FRESHENERS AND DEODORIZERS
BREAKFAST FOODS-FROZEN
LAUNDRY SUPPLIES
WINE
ICE
HARDWARE, TOOLS
INSECTICDS/PESTICDS/RODENTICDS
VITAMINS
EGGS
YOGURT
SNACKS, SPREADS, DIPS-DAIRY
TOYS & SPORTING GOODS
COFFEE
LIGHT BULBS, ELECTRIC GOODS
FRESH MEAT
MEN'S TOILETRIES
ORAL HYGIENE
DEODORANT
DESSERTS/FRUITS/TOPPINGS-FROZEN
SEWING NOTIONS
FRAGRANCES - WOMEN
SOFT DRINKS-NON-CARBONATED
MEDICATIONS/REMEDIES/HEALTH AIDS
CANNING, FREEZING SUPPLIES
ICE CREAM, NOVELTIES
COOKIES
PET CARE
BABY FOOD
GLASSWARE, TABLEWARE
COSMETICS
MILK
HAIR CARE
DRESSINGS/SALADS/PREP FOODS-DELI
LIQUOR
$a_1 > 0$ (procyclical)
0.0
0.2
0.4
0.6

Regional Changes in Food-in: Boom



Regional Changes in Food-in: Recession



Concluding Remarks

- Users have more data preprocessing responsibilities
- Methodology

   i start with some possibly imperfect method.

   ii exploit information across counties (Tweedie formula)

   iii automate using machine learning methods.

- Other applications:
  - Firm level industrial production, different sectors
  - For each sector, pool across firms.