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. Motivation
2. Demand Analysis Pre-adjustment
3. Seasonal Adjustment
4. Demand Analysis Post-adjustment
5. Cyclical Sensitivity

Motivation

- Conventional data
  - Intentionally collected
    - Cleaned, adjusted, and checked
    - Compact and regular.
    - Mostly metrical (not qualitative) variables.
- New Data
  - Often by-product of economic and social activities.
  - Unconventional characteristics:
    - 3V: Volume, Variety, Velocity.
    - Can be non-metrical and not even numerical (eg. text).
    - Available as is: pre-processing is responsibility of user.

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.
  - use a second step that exploits cross section dependence to mop up residual nuisance variations.
  - automate using machine learning tools.
- Application: What does Nielsen scanner data say about consumer spending around the Great Recession

Disclaimer

- 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.

The Nielsen Scanner Data

The good:

- Volume (5TB):
- over 1000 products from $N_g \approx$110 product groups,
- Actual sales, not estimates from surveys.
- 2006:1:07 -2014:12:29. Financial Crisis
- Weekly frequency: (not monthly/quarterly)
- Local level economic data: major MSA (not 4 regions).

The not so good especially for macro analysis:

- Groceries, mass merchandise products: few durable goods.
- $T = 469$ weeks, $N_{\text{years}} = 9$, short span.
- Volume: 5TB, memory constraint.
- Variety (multilevel heterogeneity): product, spatial, time:
- Velocity: weekly, quasi-periodic.
  - Ng (2017): big data= big time cleaning data.
  - Few studies take advantage of weekly frequency. Why?
  - A Challenge: Strong Seasonality.

Setup

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

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

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

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

Demand Analysis of Product Groups

- Demand system: budget shares are functions of p and Y.

$$\text{share}_g = \sum_{k=1}^{r} \lambda_{kg}(\log \text{p}) F_k(\log Y/P).$$

  P is a theory based price index, F_1 constant for adding up.

- p and Y/P are common across groups g.

- The rank of a demand system r is the dimension of F, the functions that span Engel curve.

  - rank 1: homothetic demand independent of income.
  - rank 2: demands are generalized linear (Muellbauer 1980).

- Demand theory forms basis of consumer price indexes.

Traditional Demand Analysis

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

For AIDS (Deaton and Muellbauer 1980):

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

- Stone's price index: $\log P^* = \sum_j w_j \log p_j$
- Classical estimation: T large, N_g small.
- using seasonally adjusted data, typically $2 \leq r \leq 4$.

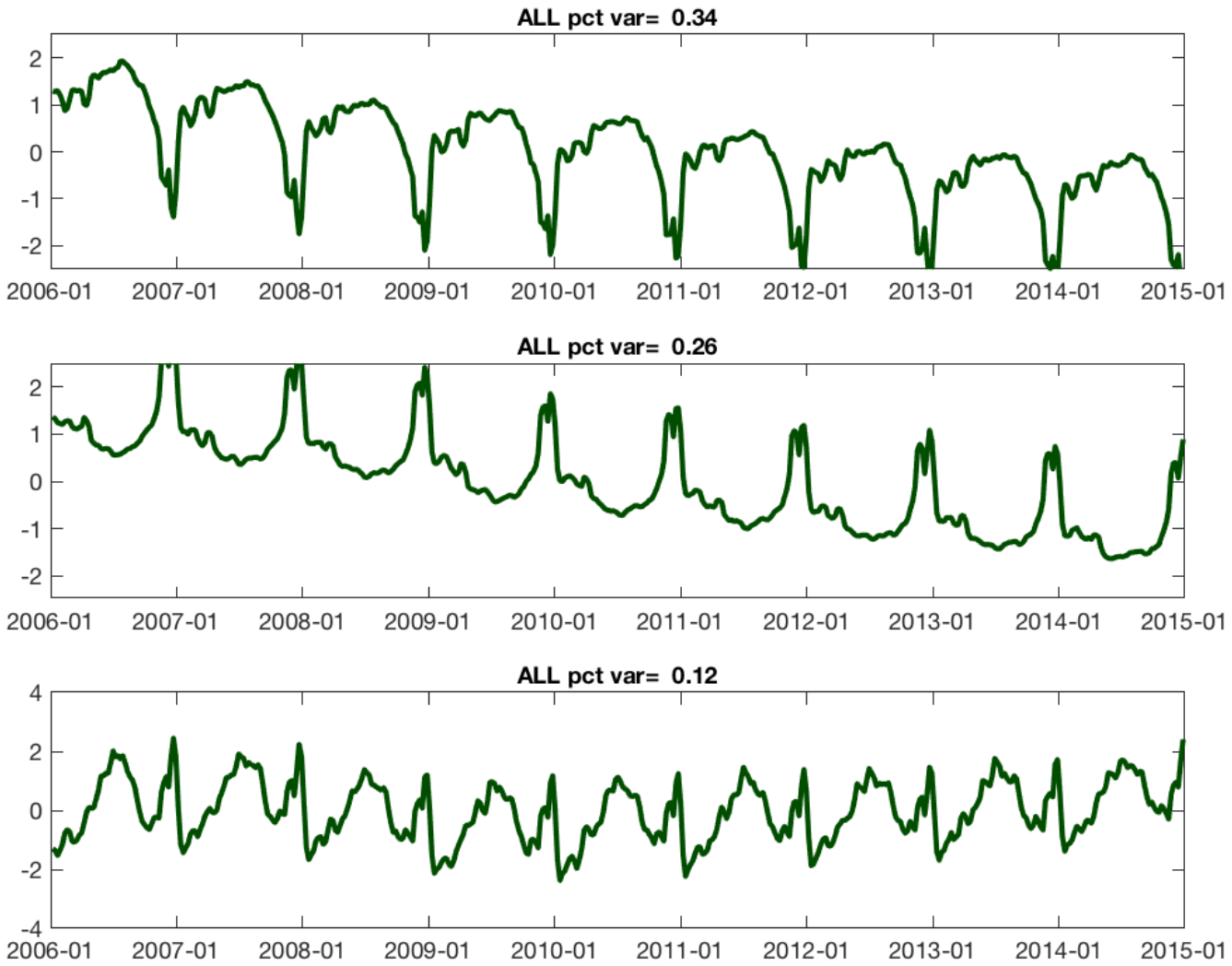
A Large N_g large T Approach

- Nielsen data: $T = 469$, $N_g = 108$ for each s.
  - budget shares have a factor structure.

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

  - Principle components estimates the space spanned by F
  - ie. can 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 official data, user has to deal with this.

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

- 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 fine tune parameters for every series.
- If $\text{sales}^{seas}_{gct}$ are correlated across c (counties), the residual will be cross-sectionally dependent.

Implication: Aggregate sales will have a seasonal component

$$\widehat{\text{sales}}_{gct}^{seas} - \text{sales}_{gct}^{seas} = \underbrace{\Phi_{gct}^{seas}}_{\text{correlated across c}} + \underbrace{e_{gct}^{seas}}_{\text{idiosyncratic}}$$

$$\widehat{\text{sales}}_{gt}^{nseas} = \sum_{c=1}^{N_c} \text{sales}_{gct}^{nseas} + \sum_{c=1}^{N_c} \Phi_{gct}^{seas} + e_{gct}^{seas}.$$

- Each individual series might appear de-seasonalized, but seasonality re-appears after aggregation.
- Need to remove the common seasonal variations when the span of data is short, and when the data have spikes.

A New Approach

- We use a year-by-year panel approach to complement the bottom up adjustments.
- An unconventional, periodicity-free, definition of seasonality.
- Motivation: sales of group g in neighboring counties have similar seasonal patterns regardless of county size.

A Prediction-Based Approach to Seasonality

- For each (c, g), standardize year by year, remove size effect:

$$y_{gct} \equiv \frac{\log(\text{SALES}_{gct}) - \mu_{gc\tau}}{\sigma_{gc\tau}} \quad t \in \tau.$$

$$\begin{aligned} y_{gct} &= d_{gct} + q_{gct} + u_{gct} \\ &= \underbrace{\text{county specific seas.} + \text{correlated seas.}}_{\text{seasonal}} + \text{non-seasonal}. \end{aligned}$$

- 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: remove idio. seasonals: regress y_{gct} on d_{gct}, get resid.

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

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

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

Step 2: pool information across counties and years

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

Optional: Reweight and rescale:

$$
\begin{aligned}
y_{gct} &= \alpha_{p0} + \alpha_{g1} \cdot \widehat{d}_{gct} + \alpha_{g2} \cdot \widehat{q}_{gct} + u_{gct} \\
x_{gct} &= \widehat{\text{log adjusted sales}}_{gct} \equiv \widehat{u}_{gct} \cdot \sigma_{g\tau} + \mu_{g\tau}
\end{aligned}
$$

Step 2: Intuition

- For each (g, s, τ): predict q_{gct} based on $\widehat{q_{gct} + u_{gct}}$ $\forall c \in s$.
- 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.

$$
\left(\begin{array}{cccc}
y_{1,1} & y_{1,2} & \cdots & y_{1,N_g} \\
y_{2,1} & y_{2,2} & \cdots & y_{2,N_g} \\
\vdots & \vdots & \vdots & \\
y_{52,1} & y_{52,2} & \cdots & y_{52,N_g} \\
\hline
y_{53,1} & y_{53,2} & \cdots & y_{53,N_g} \\
\vdots & \vdots & \vdots & \\
y_{104,1} & y_{104,2} & \cdots & y_{104,N_g} \\
\hline
\vdots & \vdots & \vdots & \\
\vdots & \vdots & \vdots & \\
\hline
y_{417,1} & y_{417,2} & \cdots & y_{417,N_g} \\
\vdots & \vdots & \vdots & \\
y_{469,1} & y_{469,2} & \cdots & y_{469,N_g}
\end{array}\right)
$$

Step 2: Predictors $\mathbb{Z}_{gct}$

- Many (391) predictors

  i (base set) 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).

- Data $\mathcal{D}_{g\tau} = (Y_{g\tau}, Z_{g\tau})$=(response,predictors)
  $(\mathcal{D}_{1,g\tau}, \mathcal{D}_{2,g\tau})$=(training, validation)

- ncol$(\mathcal{D}_{g\tau}) =$ # of predictors+1.

- nrow $(\mathcal{D}_{1,g\tau}) = (469 - \text{weeks in year } \tau) \times N_c^s$.

step 2: Machine Learning

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

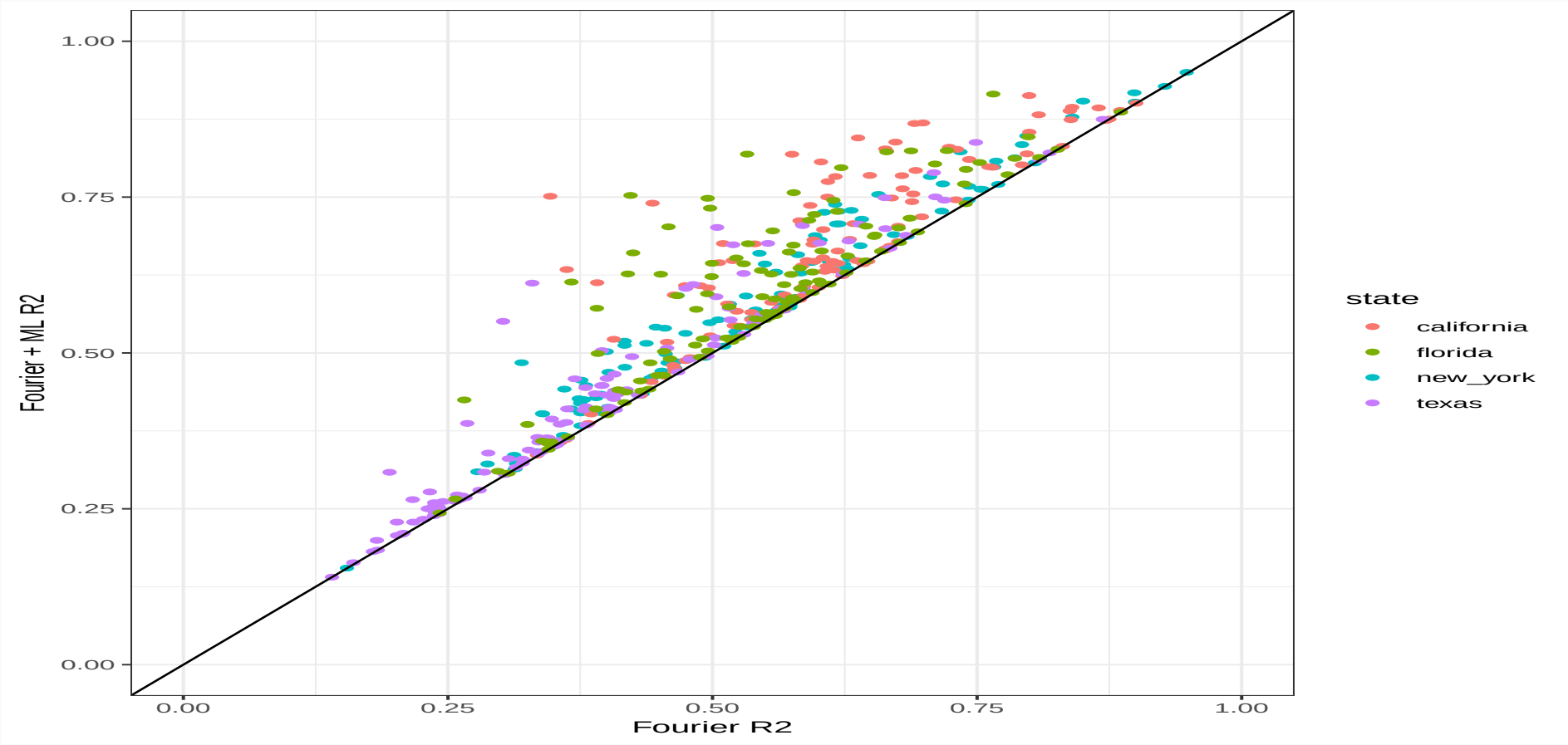
Week Ending	CA	FL	NY	TX	CA	FL	NY	TX
	Adjusted Data				Raw Data			
	The 2009 July 4th Effect on Beer Spending							
June 27	3.5	2.9	2.5	2.6	4.1	3.3	3.2	3.0
July 4	3.5	2.8	2.5	2.7	4.9	3.2	3.8	3.6
July 11	3.2	2.8	2.4	2.2	3.8	3.5	3.3	2.8
	The 2009 Superbowl Effect on Beer Purchases							
Jan 31	3.3	2.6	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

Adjusted Data: Some Observations

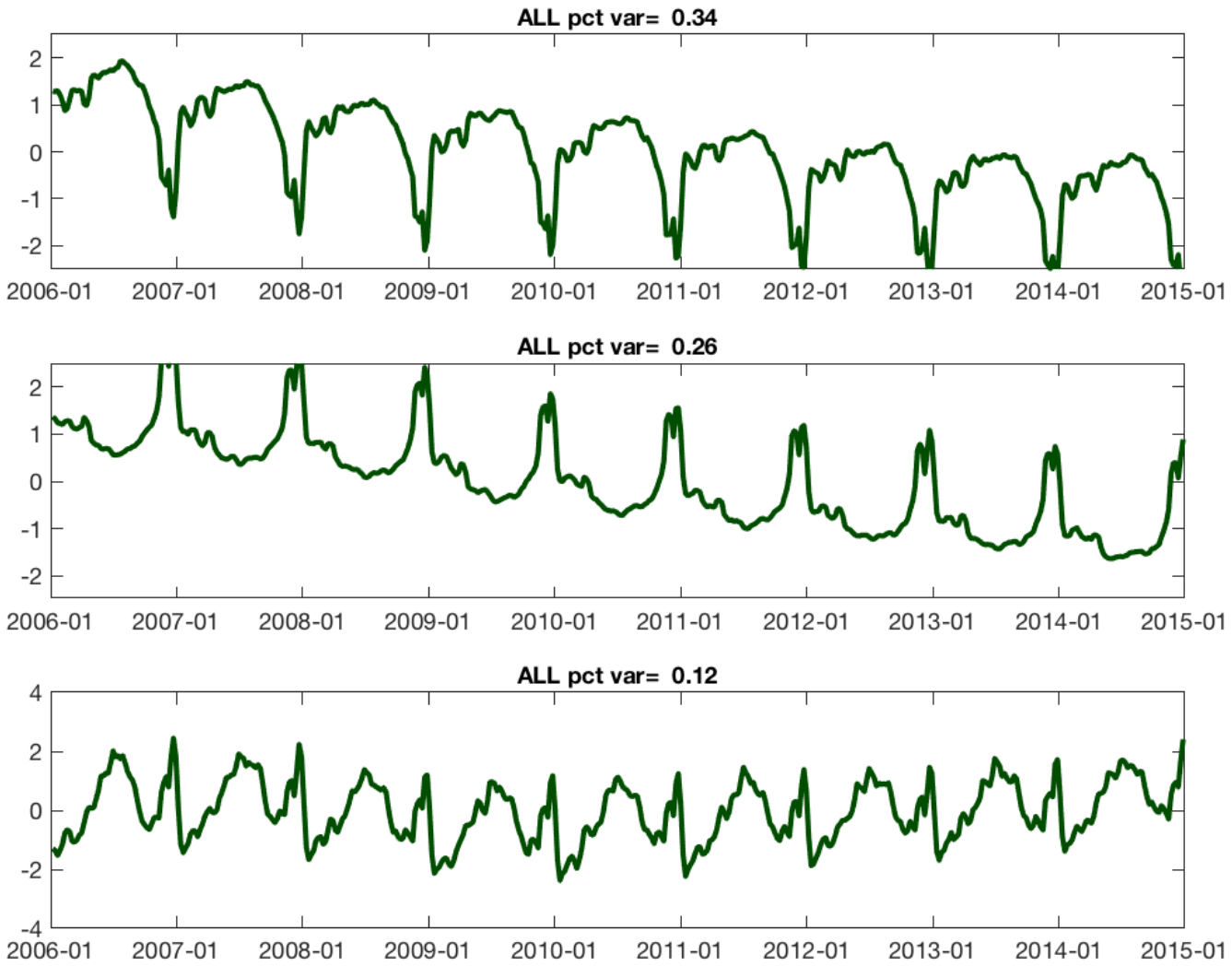
- Largest change: shares in 'insectisides', 'ice', 'stationary', 'candies', 'fragrance'.
- Value added of Step 2: Evaluate R^2 of
  $y_{gct} = \alpha_{p0} + \alpha_{g1} \cdot \widehat{d}_{gct} + \alpha_{g2} \cdot \widehat{q}_{gct} + u_{gct}$.

	Average of R^2					
method	mean	median	max	q75	q25	min
Fourier	0.49	0.49	0.91	0.60	0.37	0.14
OLS	0.50	0.50	0.91	0.61	0.37	0.14
Lasso	0.52	0.53	0.91	0.65	0.39	0.14
Random Forest	0.54	0.55	0.91	0.68	0.39	0.14

Incremental Predictive Power of Random Forests

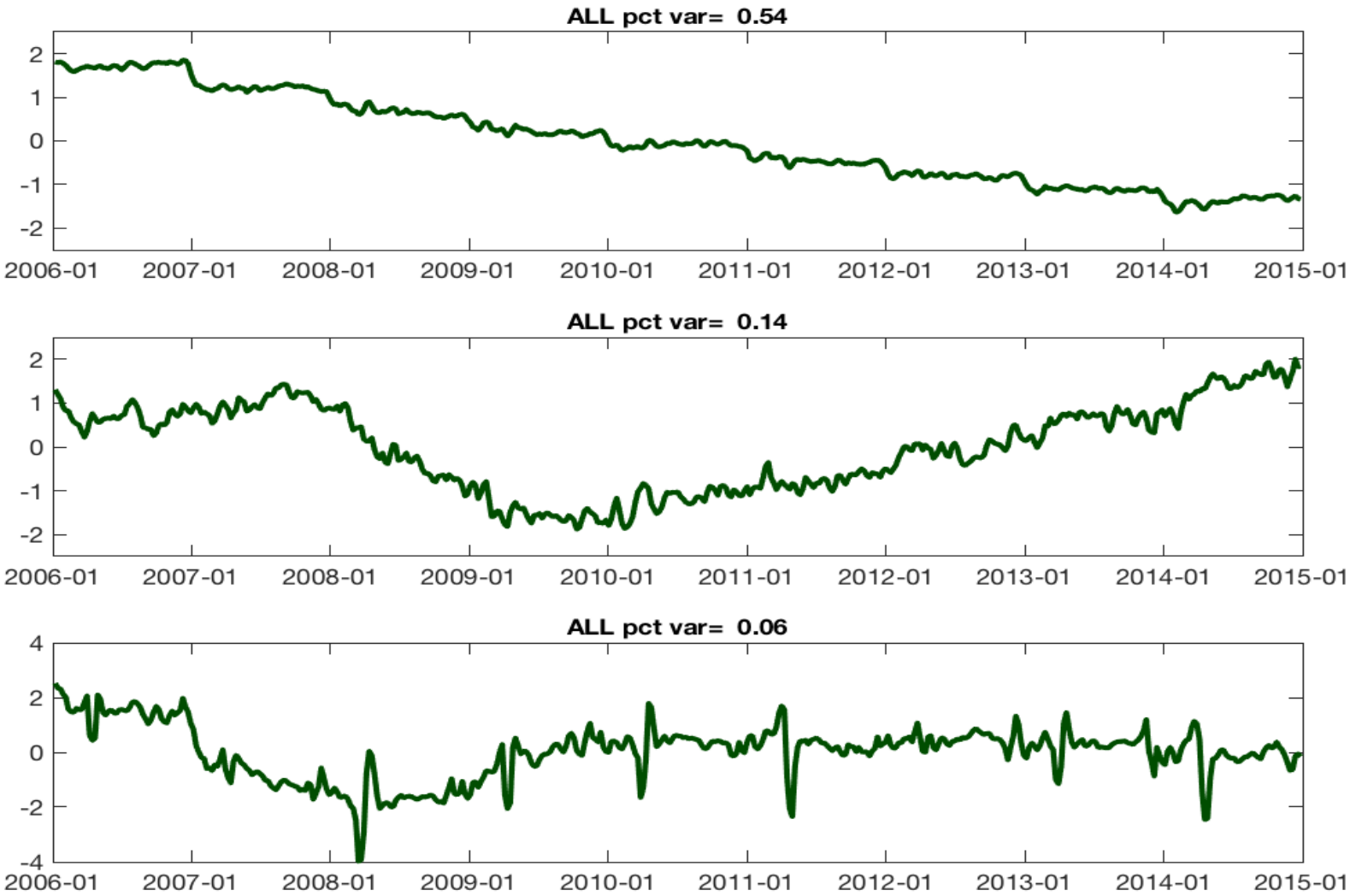


Factors Estimated from Raw Shares: All States



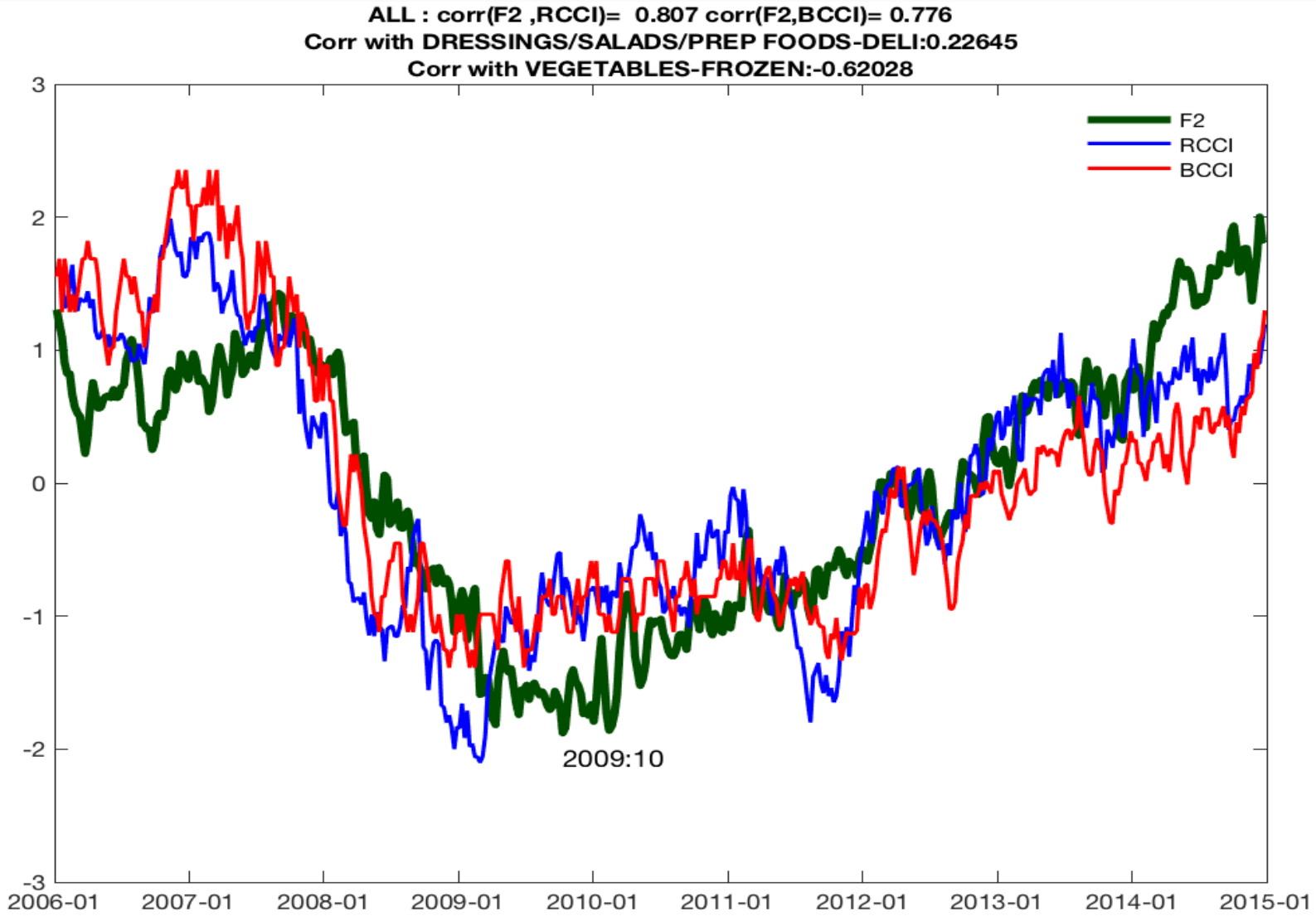
Strong and heterogeneous seasonal effects! Where is the cycle?

Factors Estimated from Adjusted Shares: ALL States

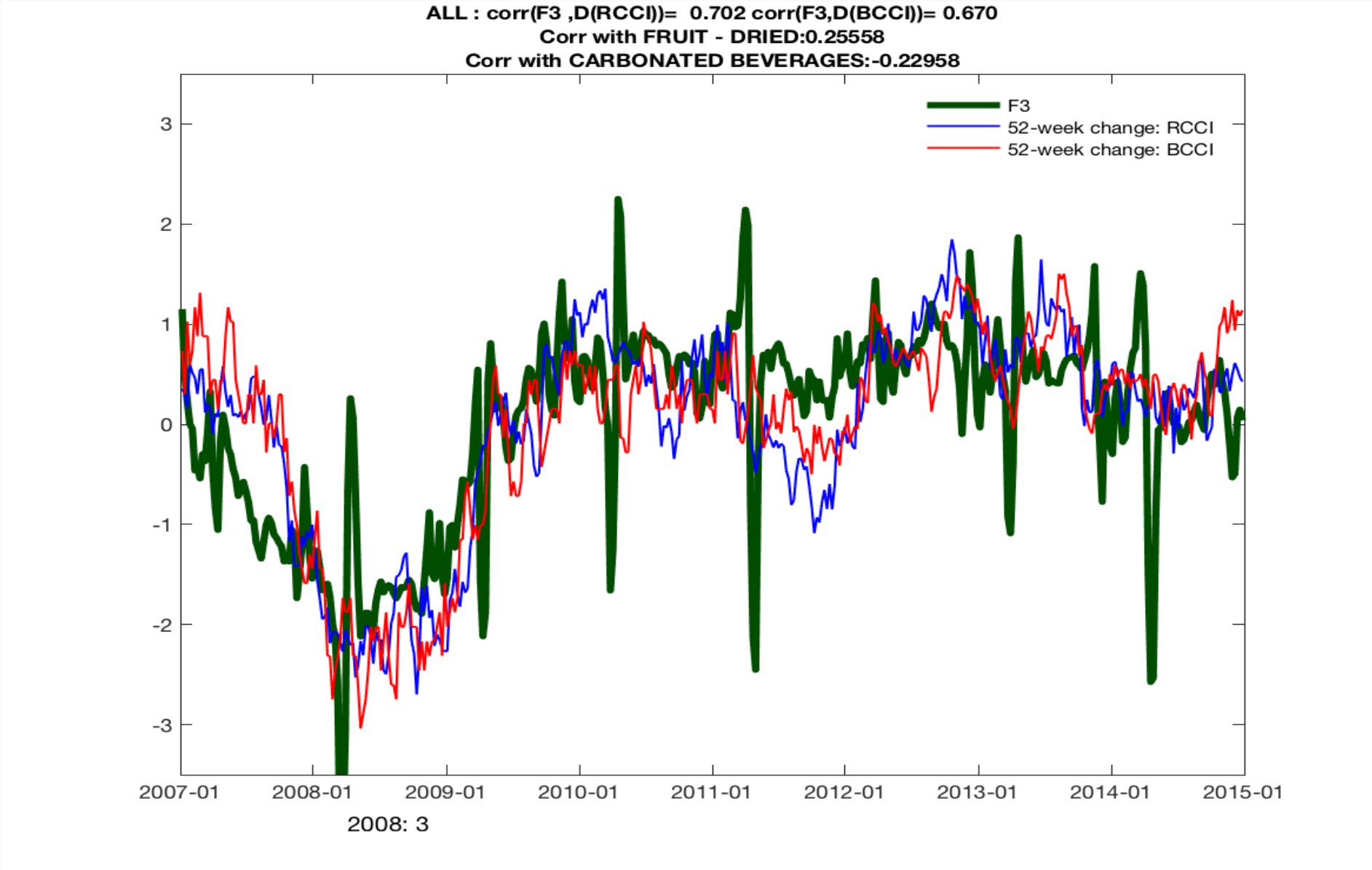


3 Factors: Trend, Level, Slope.

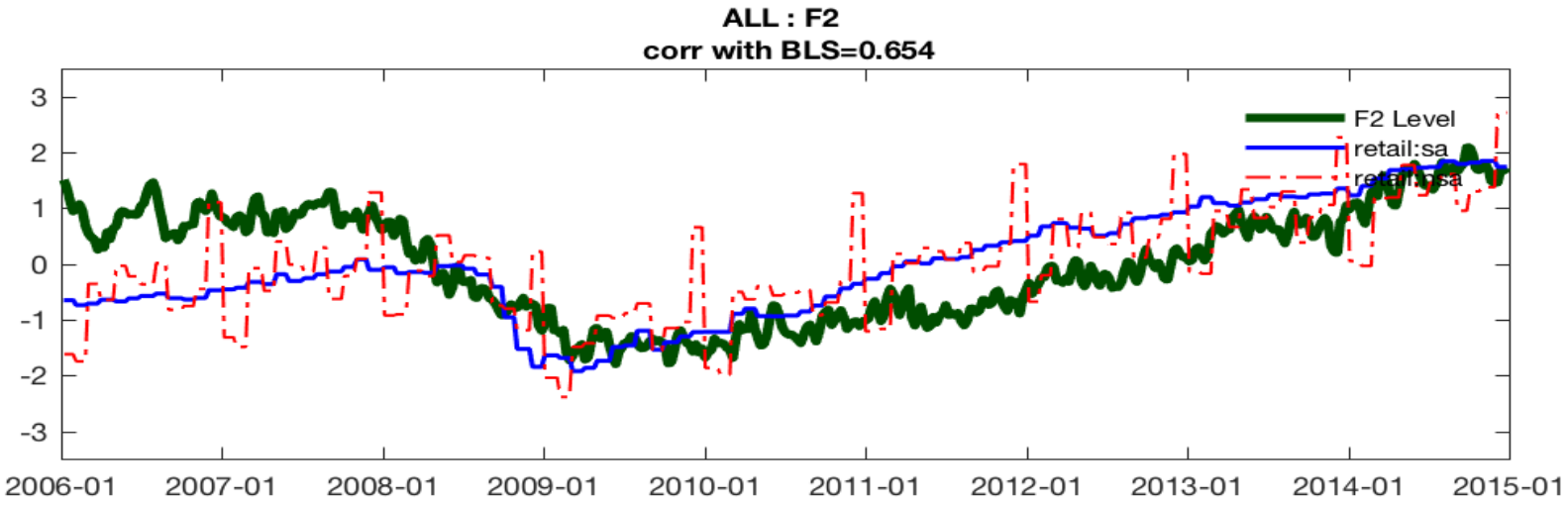
The Level Factor

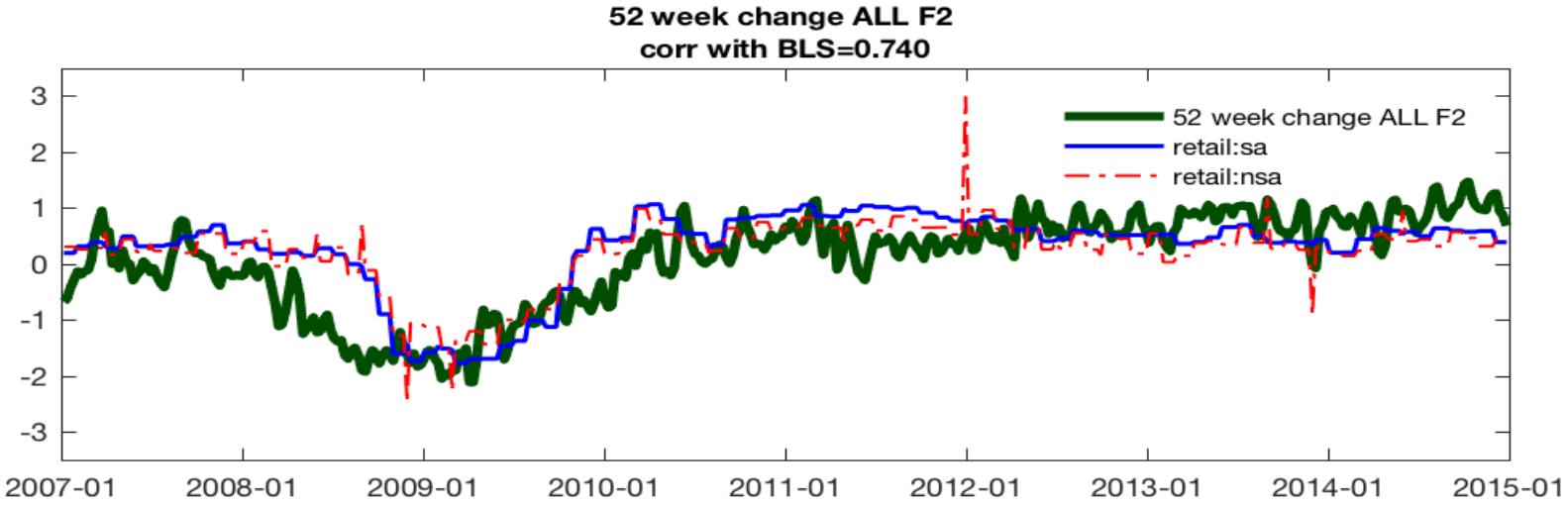


The Slope Factor



Comparison with the BLS Monthly data





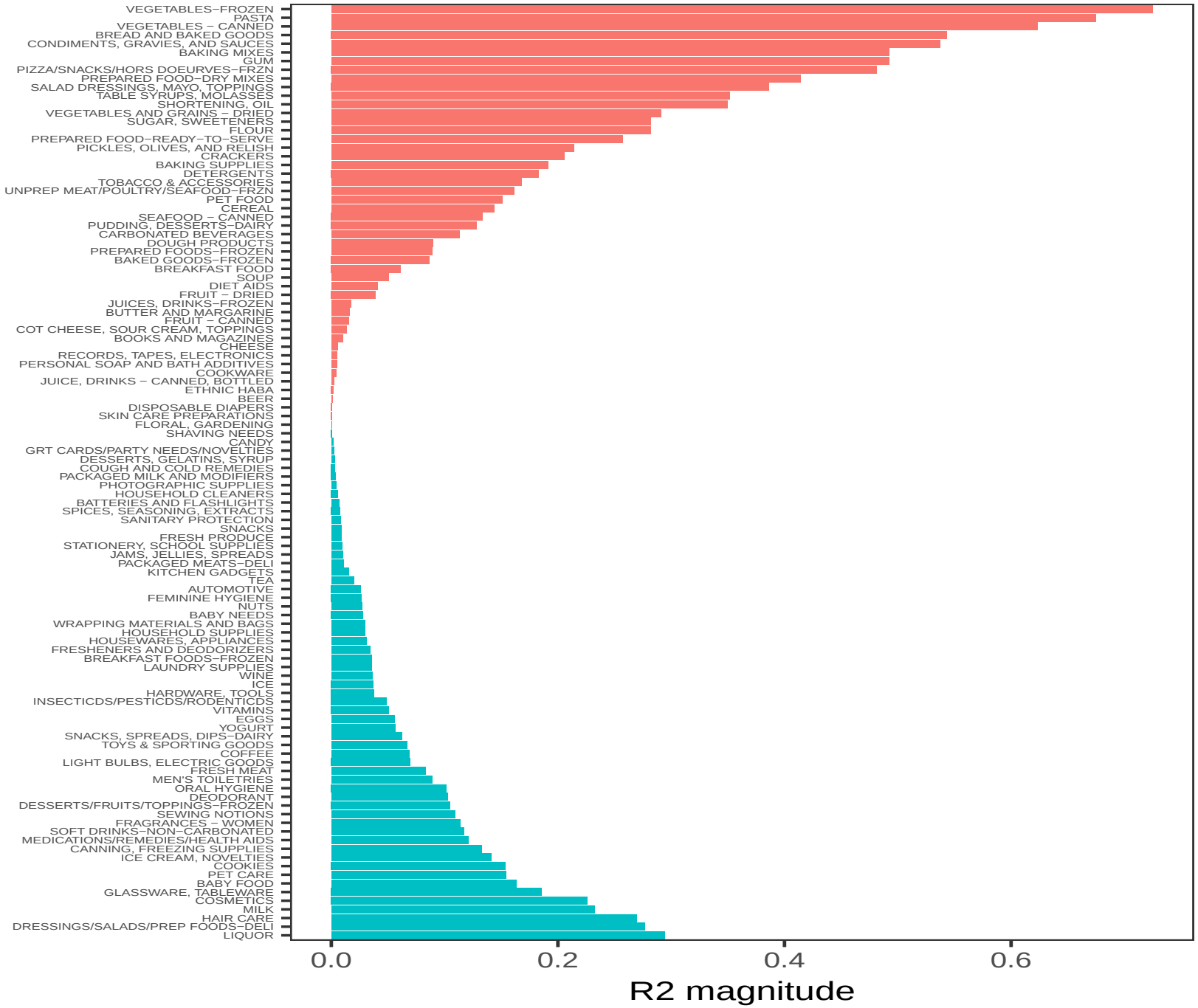
Sensitivity to Cyclical Factor: group level

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

  (group,week) panel regression:

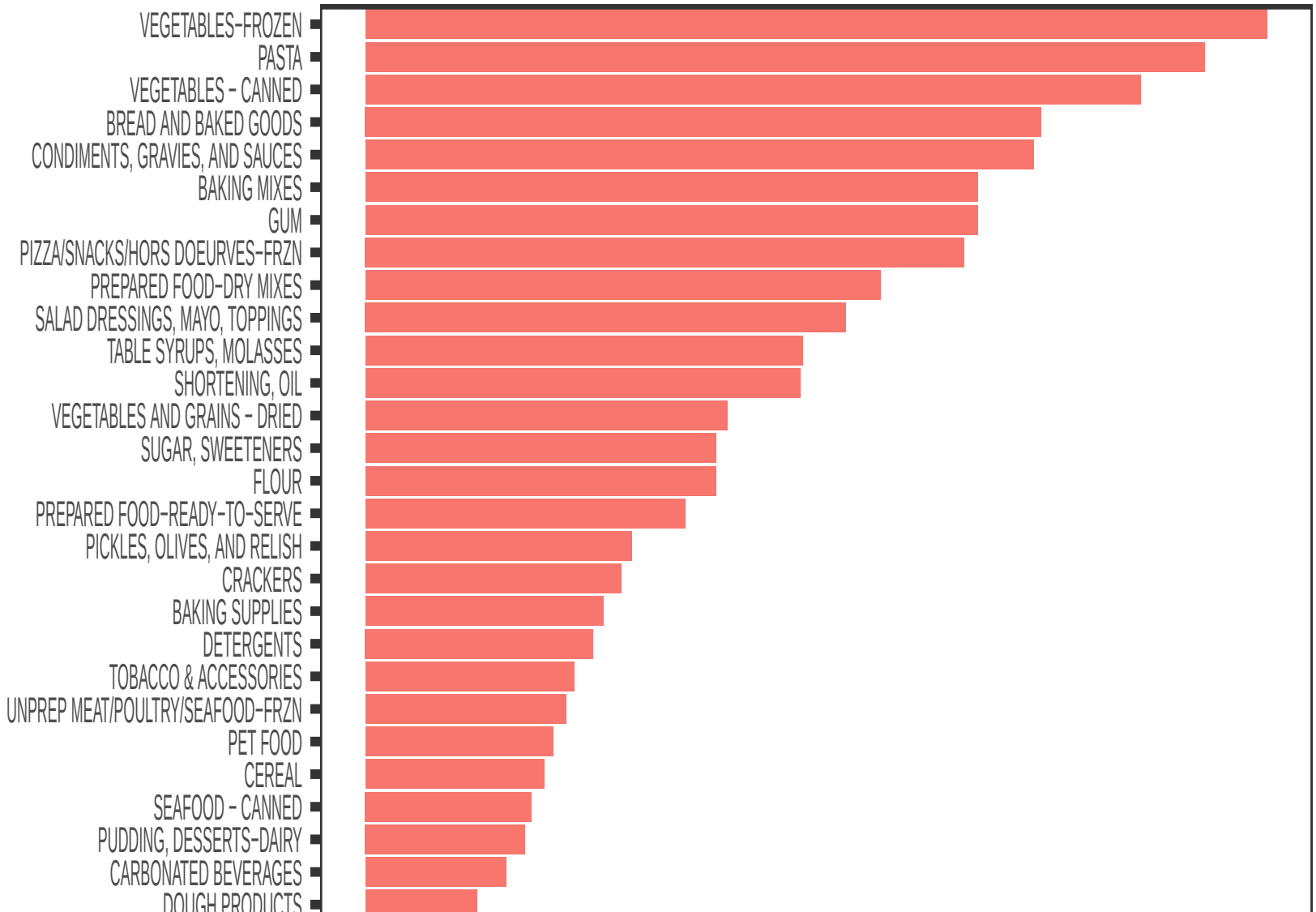
$$share_{gt} = a_0 + a_1 \widehat{F}_{2,RF,t} + error_{gt}$$

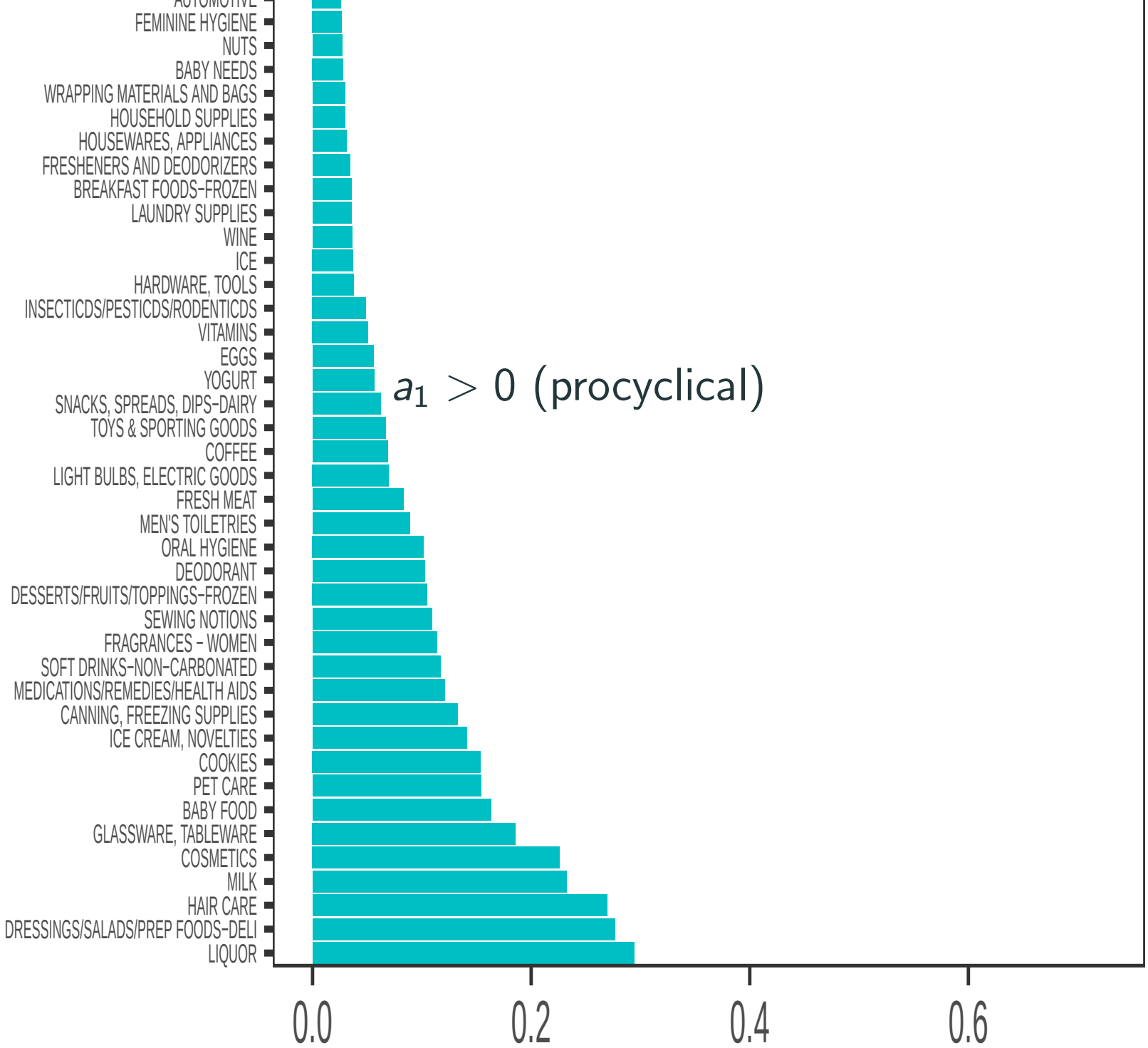
$R^2 \approx 0$: beer, cough/cold remedies, disp. diapers, batteries



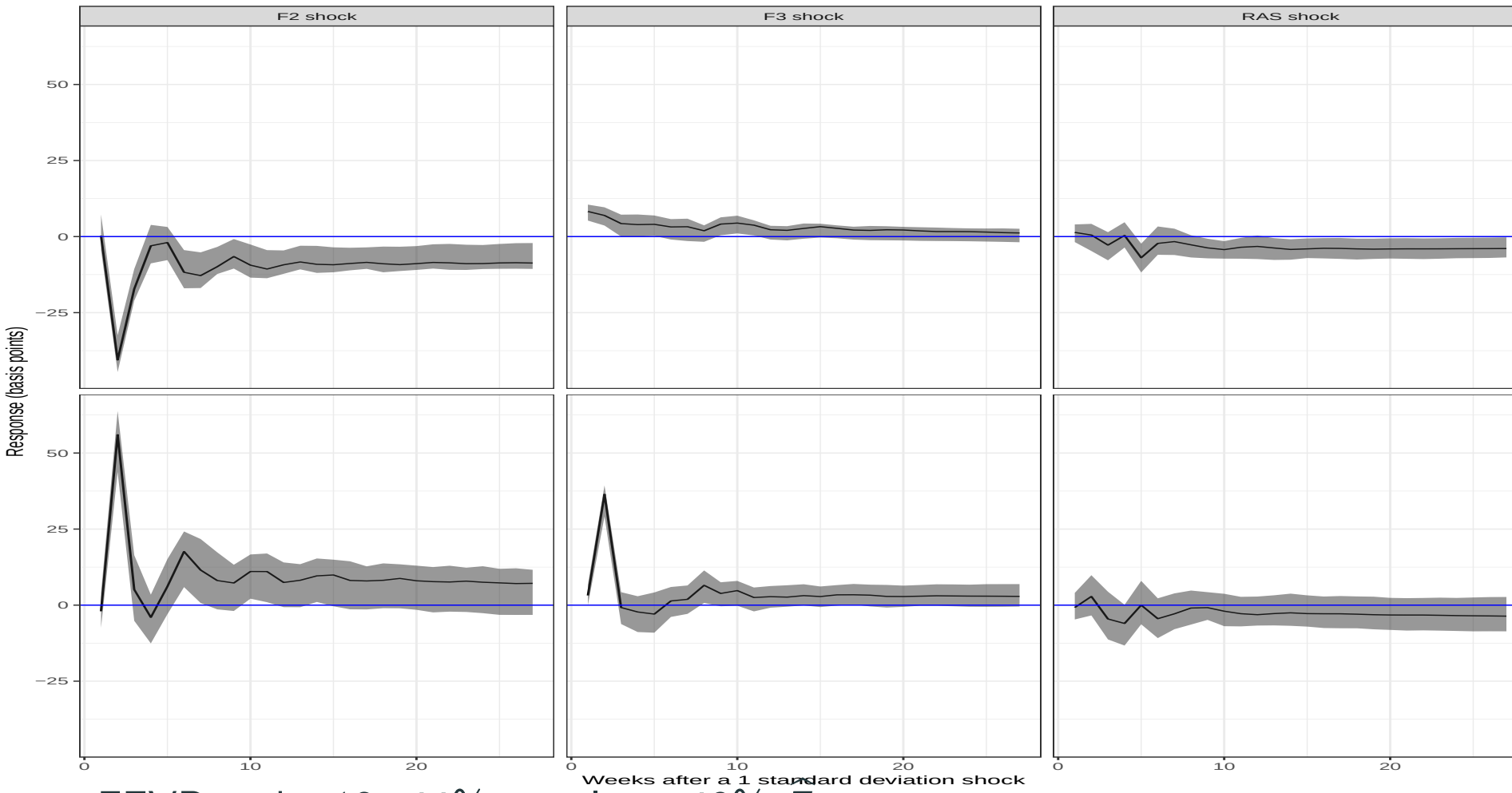
Distribution of R^2

$a_1 < 0$ (countercyclical)




AUTOMOTIVE
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

Response of food-in to $\widehat{F}_2$ in VAR



- FEVD at h=12: 44% own lags, 48% $F_{2,RF}$.
- RCCI accounts for little of variations in food-in even in bivariate VAR.

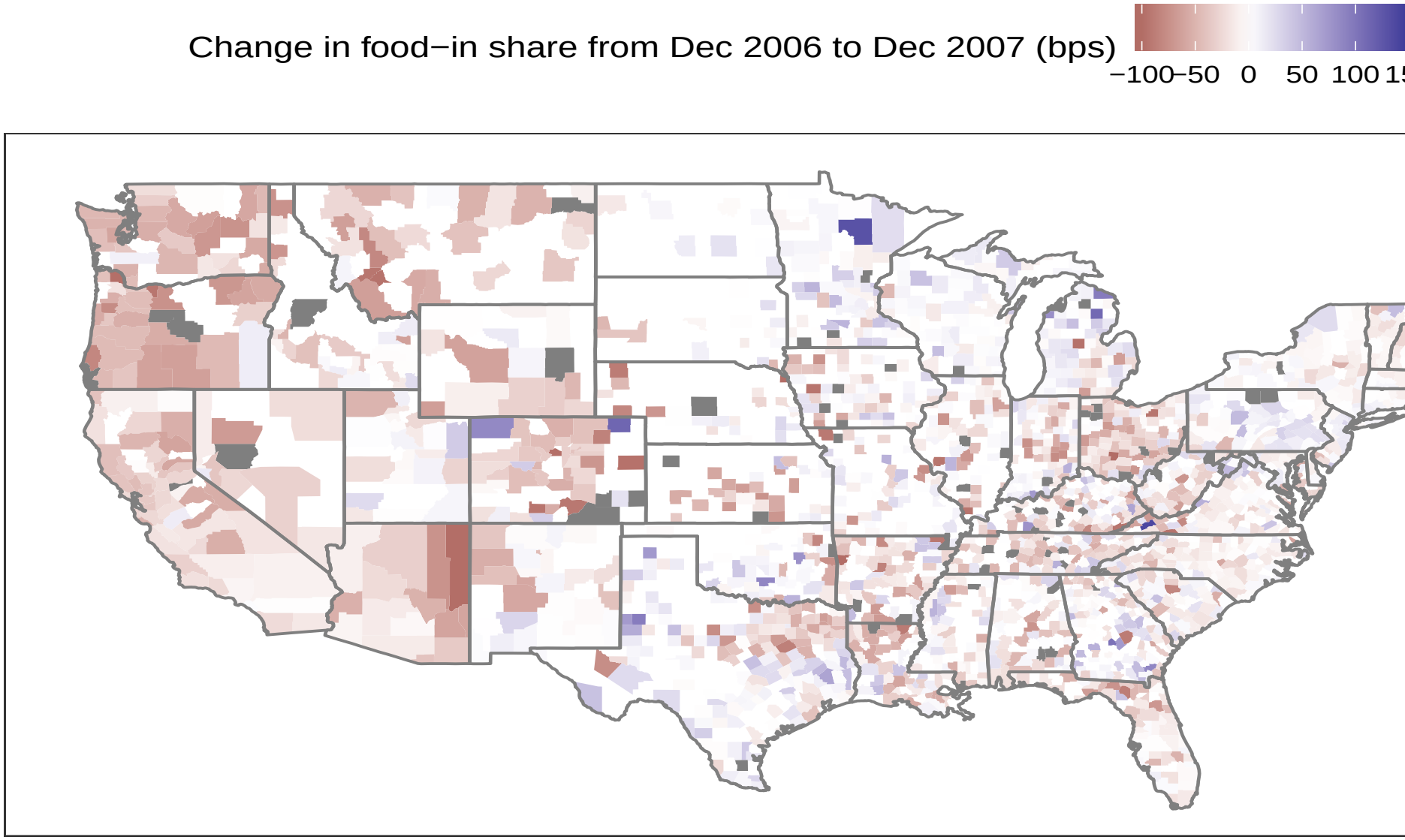
County-level heterogeneity

- Which counties are more exposed to aggregate risk?
- food-in: frozen and canned vegetables, pasta, bread, condiments and sauces.
- luxury: liquor, prepared food, milk, hair care, cosmetics.
- (week, county) panel regressions:

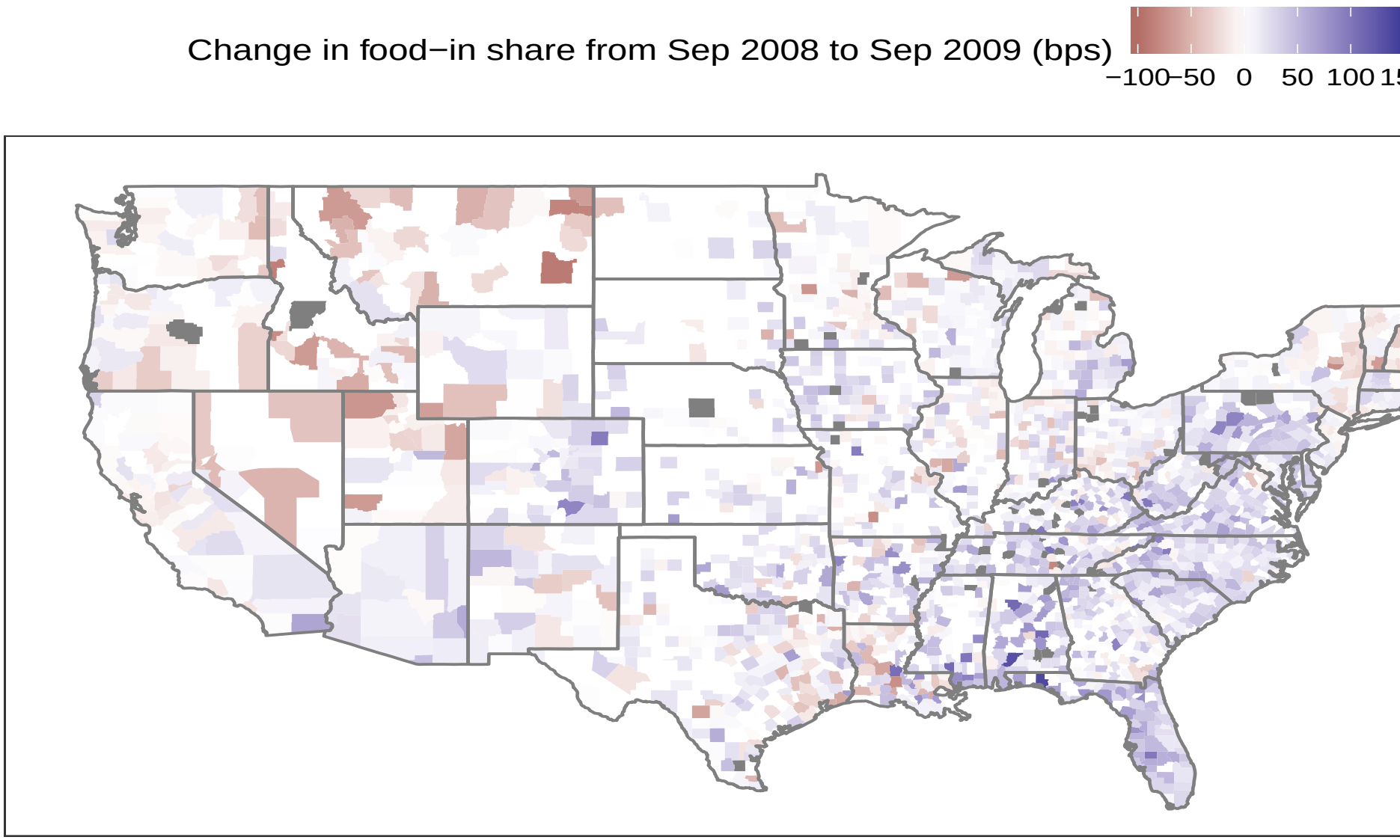
$$\text{food-in}_{ct} = a_1 + a_2 \widehat{F}_{2,RF,t} + a_3 \widehat{F}_{3,RF,t} + error_{ct}$$
$$\text{luxury}_{ct} = a_1 + a_2 \widehat{F}_{2,RF,t} + a_3 \widehat{F}_{3,RF,t} + error_{ct}$$

use R^2 as measure of exposure to common shocks.

Regional Changes in Food-in: Boom



Regional Changes in Food-in: Recession



Sensitivity to Economic Conditions: county level

county	CA	FL	NY	TX
	Most sensitive			
1	santa barbara	miami-dade	rockland	montgomery
2	los angeles	broward	nassau	dallas
3	orange	organge	kings	travis
	Least sensitive			
1	sutter	sarasota	seneca	callahan
2	kings	hamilton	lewis	willacy
3	humbodt	lafayette	broome	bexar

urban and densely populated counties are more exposed to aggregate shocks.

Concluding Remarks

- User has 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.