Exploring Heterogeneity (Sales Taxes)

John Bonney

December 18, 2018

Distribution of Reforms

When are reforms happening?

As we can see in Figure 1, most of the increases are concentrated in 2010 Q3, and most of the decreases are concentrated in 2011 Q3. What is driving these large spikes?

```
# Which states are driving the spike in increases?
with(reforms[change == "Increase" & year_qtr == "2010 Q3"],
     table(state_name, change))
```

```
##                change
## state_name      Increase
##   Illinois             4
##   Kansas             105
##   Louisiana            1
##   Missouri             3
##   New Mexico          33
##   North Carolina       4
##   Ohio                 1
##   Washington           1
```

```
# Which states are driving the spike in decreases?
with(reforms[change == "Decrease" & year_qtr == "2011 Q3"],
     table(state_name, change))
```

```
##                change
## state_name      Decrease
##   California          58
##   North Carolina     100
##   Ohio                 1
```

These spikes in tax decreases and tax increases were specifically driven by four reforms:

- In July 2010, New Mexico **increased** its state sales tax by 0.125%, from 5% to 5.125% *(33 counties)*
- In July 2010, Kansas **increased** its state sales tax by 1.0%, from 5.3% to 6.3% *(105 counties)*
- In July 2011, North Carolina **decreased** its state sales tax by 1.0%, from 5.75% to 4.75% *(100 counties)*
- In July 2011, California **decreased** its state sales tax by 1.0%, from 8.25% to 7.25% *(58 counties)*

Where are reforms happening?

The only state-wide tax change that shows up in this map (Figure 2) but is not captured in the previous analysis is Arizona's June 2010 tax increase, which increased state sales tax by 1%, from 5.6% to 6.6%.

I show a breakdown of all county-level changes, state-by-state, in Table 1.

Predictive characteristics

We have some variables obtained from Zillow.com and IPUMS National Historical Geographic Information System (NHGIS) in addition to the Quarterly Census of Employment and Wages (QCEW) and the Local Area Unemployment Statistics, both from the Bureau of Labor Statistics (see Table 2). I use these data

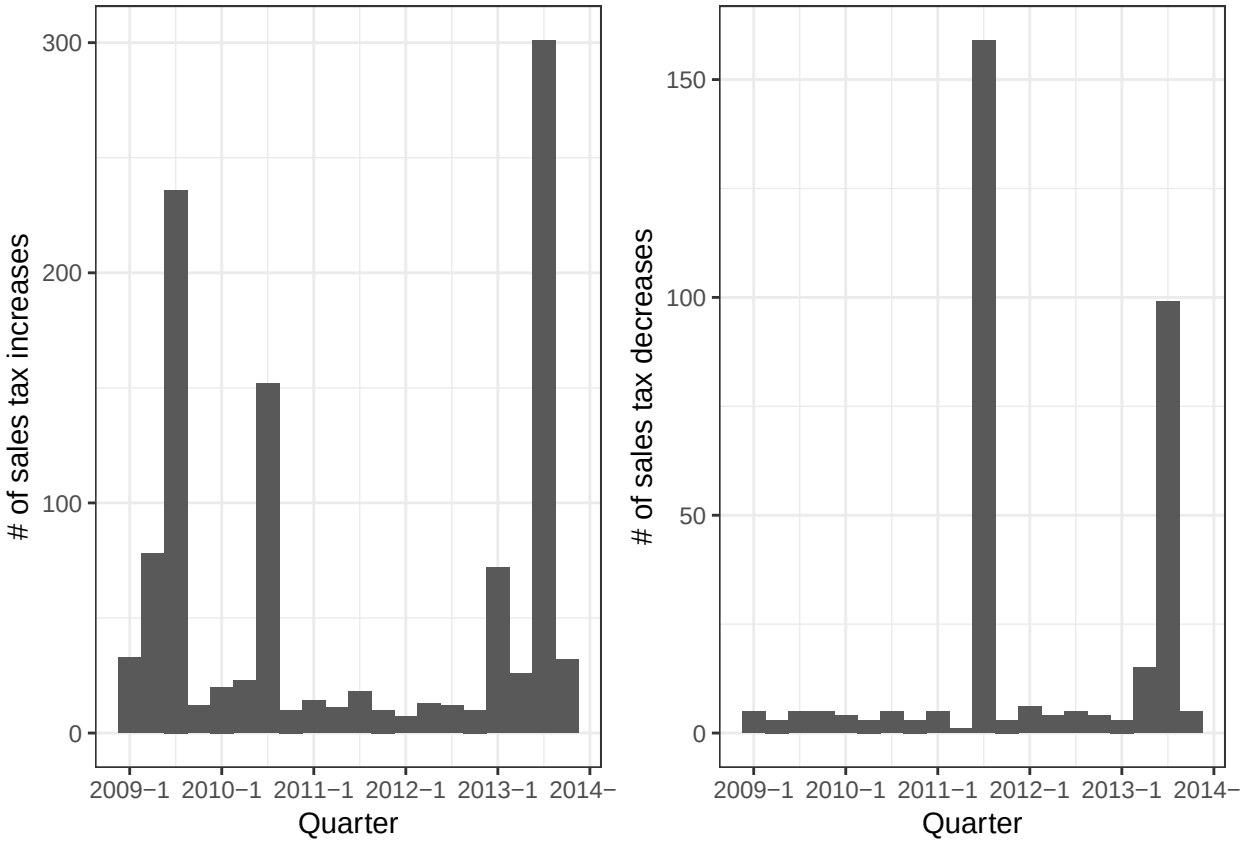

Figure 1: Distribution of Sales Tax Changes over Time

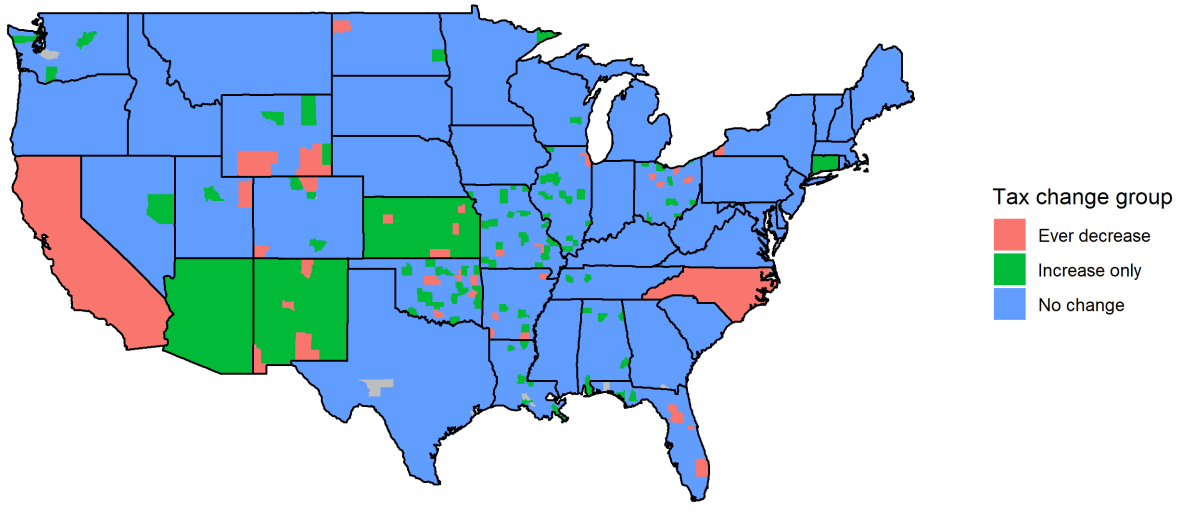

Figure 2: Geographical Distribution of Sales Tax Changes

Table 1: Number of counties that increased/decreased sales taxes between 2010 and 2012 (by state)

State	# ever increase	# ever decrease	Total
Alabama	11	1	67
Alaska	0	0	29
Arizona	15	15	15
Arkansas	75	8	75
California	58	58	58
Colorado	6	3	64
Connecticut	8	0	8
Delaware	0	0	3
Florida	6	5	67
Georgia	0	0	159
Hawaii	0	0	5
Idaho	0	0	44
Illinois	20	2	102
Indiana	0	0	92
Iowa	0	0	99
Kansas	105	100	105
Kentucky	0	0	120
Louisiana	13	0	64
Maine	16	0	16
Maryland	0	0	24
Massachusetts	14	0	14
Michigan	0	0	83
Minnesota	87	0	87
Mississippi	0	0	82
Missouri	36	5	115
Montana	0	0	56
Nebraska	0	0	93
Nevada	17	0	17
New Hampshire	0	0	10
New Jersey	0	0	21
New Mexico	33	5	33
New York	5	2	62
North Carolina	100	100	100
North Dakota	2	1	53
Ohio	88	6	88
Oklahoma	30	8	77
Oregon	0	0	36
Pennsylvania	1	0	67
Rhode Island	0	0	5
South Carolina	2	0	46
South Dakota	0	0	66
Tennessee	6	0	95
Texas	0	0	254
Utah	1	1	29
Vermont	0	0	14
Virginia	131	0	133
Washington	6	1	39
West Virginia	0	0	55
Wisconsin	3 2	0	72
Wyoming	11	9	23

Table 2: Summary of variables

Variable	Source	Description
SizeRank	Zillow	Rank of county size (by pop.) (Jan. 2006)
median_home_price	Zillow	Median home price (Jan. 2006)
pop	NHGIS	County pop. (2000)
pop_urban	NHGIS	Pop. living in urban areas (2000)
pop_black	NHGIS	Pop. identifying as black (2000)
pop_over_65	NHGIS	Pop. over age 65 (2000)
pop_under_25	NHGIS	Pop. under age 25 (2000)
pct_pop_no_college	NHGIS	Pct. pop. without a college degree (2000)
pct_pop_bachelors	NHGIS	Pct. pop. with bachelor's degree or higher (2000)
median_income	NHGIS	Median income (2000)
total_housing	NHGIS	Total housing units (2000)
owner_occupied_housing	NHGIS	Total owner-occupied housing units (2000)
annual_avg_emplvl*	QCEW	Establishment employment level (2006)
retail_employment	QCEW	Retail industry employment (2006)
total_establishments	QCEW	Total number of establishments (2006)
retail_establishments	QCEW	Total number of retail establishments (2006)
food_and_drugstores_empshare	QCEW	Fraction of workers in food and drugstores (2006)
total_mean_wage	QCEW	Mean wage (2006)
retail_mean_wage	QCEW	Mean wage in retail industry (2006)
laborforce	LAUS	Labor force size (2006)
employed*	LAUS	Total employment level (2006)
unemployed	LAUS	Unemployed level (2006)
unemp_rate	LAUS	Unemployment rate (2006)

* Note that there are two measures of employment, one from the QCEW and one from the LAUS. This is because the QCEW estimates employment levels using establishment data, but the BLS adjusts these measures from a place-of-work basis to a place-of-residence basis using commutation factors calculated from the ACS. For more information, see the BLS estimation methodology.

(restricted to information determined prior to the tax change) to see which county characteristics "predict" sales tax changes.

I adapted many of the variables in Table 2 to be in percentage or log terms for comparability across counties. In addition to these variables, we have industry employment shares for a number of different industries from the QCEW.

To determine what features are indicative of a future sales tax change, I run predictive regressions similar to those run by Deshpande and Li (2018). I estimate three different models for each type of sales tax reform (increase and decrease). The first model includes population-specific covariates relating to housing, employment, income, race, and education. The second model includes industry employment shares for the county for six specific industries. The third model combines the covariates from the first two models, and is as follows:

$$y_c = \beta PopChar_{c2000} + \gamma Housing_{c2000,2006} + \delta LaborForce_{c2006} + \kappa EmpShare_{c2006} + \epsilon_c,$$

where $PopChar_{c2000}$ is a vector of population characteristics; $Housing_{c2000,2006}$ is a vector of housing characteristics, including the local median home price in 2006 and the urbanization and home ownership rates in 2000; $LaborForce_{c2006}$ is a vector of local labor force characteristics in 2006; $EmpShare_{c2006}$ is a vector of

Table 3: Predictive Characteristics

	Ever decrease			Ever increase		
Log(pop.)	0.304***		0.509***	0.022		0.191**
	(0.117)		(0.139)	(0.076)		(0.092)
Unemp. rate	0.179***		0.147**	0.029		0.002
	(0.056)		(0.068)	(0.044)		(0.051)
Log(median home price)	1.509***		1.471***	1.263***		1.081***
	(0.222)		(0.282)	(0.164)		(0.198)
Urbanization rate	-0.018***		-0.014*	-0.007*		-0.011**
	(0.006)		(0.007)	(0.003)		(0.004)
Pct. black	0.007		0.006	0.004		-0.001
	(0.007)		(0.008)	(0.005)		(0.006)
Pct. over 65	0.029		-0.044	0.016		-0.015
	(0.036)		(0.044)	(0.025)		(0.03)
Pct. under 25	0.009		-0.074**	-0.019		-0.048*
	(0.029)		(0.037)	(0.021)		(0.025)
Pct. no college	-0.05***		-0.056***	-0.016**		-0.023**
	(0.012)		(0.015)	(0.008)		(0.01)
Log(median income)	-3.635***		-4.873***	-1.822***		-2.385***
	(0.667)		(0.856)	(0.471)		(0.56)
Housing ownership rate	0.045***		0.056***	0.041***		0.041***
	(0.013)		(0.018)	(0.009)		(0.012)
Retail emp. share		-0.134***	-0.119***		-0.053***	-0.052**
		(0.03)	(0.043)		(0.019)	(0.026)
Construction emp. share		0.057**	0.096***		0.006	-0.001
		(0.025)	(0.035)		(0.018)	(0.025)
Finance/insurance emp. share		-0.229***	-0.407***		-0.142***	-0.188***
		(0.062)	(0.089)		(0.036)	(0.048)
Manufacturing emp. share		-0.019**	0.015		-0.009	-0.001
		(0.01)	(0.017)		(0.006)	(0.01)
Public admin. emp. share		0.055***	0.071***		-0.007	-0.004
		(0.018)	(0.025)		(0.014)	(0.019)
Real estate emp. share		0.297***	0.068		0.178**	-0.001
		(0.097)	(0.17)		(0.081)	(0.115)
Obs.	1343	1463	1056	1639	1833	1300

Note:
*** p<0.01, ** p<0.05, * p<0.1. Percentages and fractions were multiplied by 100, so coefficients represent a 1 pp increase. Estimation is by logistic regression. Increase only counties are excluded from the "ever increase" estimation and vice versa for the "ever decrease" estimation. Standard errors in parentheses.

industry employment shares; and ϵ_c encapsulates all other factors impacting any county's decision to change sales taxes. Note that the first specification excludes $EmpShare_{c2006}$, while the second specification excludes all other variables.

Results are found in Table 3. We see that higher populations, higher unemployment rates, a more educated population, lower incomes, and higher home prices are associated with a higher likelihood of *decreasing* sales taxes; a more educated population, lower incomes, less retail employment, and less manufacturing are associated with a higher likelihood of *increasing* sales taxes.

Interestingly, when comparing the results between the "ever decrease" and "ever increase" groups, the signs of the coefficients on a majority of the predictive characteristics are the same: percent of the population with no college (negative), log median income (negative), and finance/insurance employment share (negative). This indicates that these counties which are increasing or decreasing sales tax rates are different from the counties that do not change their sales tax rates, but different in a similar way.

When interpreting these results, it is important to remember that the states and time periods mentioned in the **Distribution of Reforms** section make up a significant portion of the treatment groups, so some results may be reflecting characteristics of those states or those time periods that may not be relevant for tax reforms. As a check, I re-run the regressiosn with dummies for California and North Carolina ("ever decrease") and New Mexico and Kansas ("ever increase"). Coefficients on these dummies are insignificant (z-scores very close to 0). Most of the previously significant factors remain significant; however, all coefficients exhibit significant shrinkage. These results are available upon request.

Product-specific seasonality

The goal here is to identify which goods exhibit the most seasonality over time. I do this on the server due to the size of the dataset. I estimate the following model on the product-store-quarter-year ($psqy$) level:[1]

$$y_{psqy} = \alpha_{qp} + \gamma_p time_{t(q,y)} + \epsilon_{psqy},$$

where

$$y_{psqy} = \ln\left(\frac{sales_{psqy}}{sales_{psQ_12008}}\right),$$

γ_p is a product-specific linear time effect, and α_{qp} is a product-quarter fixed effect. I then calculate a seasonality range for each product,

$$SR_p = \max_q(\alpha_{qp}) - \min_q(\alpha_{qp}).$$

SR_p is then essentially the product-specific ratio of normalized log sales of the season with the highest average sales (peak-season) to the season with the lowest average sales (low-season).[2] Taking e^{SR_p} gives a rough

[1] For computational purposes, I instead residualize out the linear time trend to obtain $\widetilde{y}_{psqy} = y_{psqy} - \gamma_p time_{t(q,y)}$ and then take the mean of $\widetilde{y}_{psqy}$ over all stores s and years y for each product-quarter pair to obtain α_{qp}. This is functionally equivalent to what the estimated product-quarter fixed effects would be (without an intercept).

[2] Consider this measure restricted to only one store s in just one year, $y = 2008$. See that

$$\alpha_{qp} = \ln\frac{sales_{pq}}{sales_{pQ_1}}.$$

Then we have

$$\begin{aligned} SR_p &= \max_q(\alpha_{qp}) - \min_q(\alpha_{qp}) \\ &= \max_q\left(\ln\frac{sales_{pq}}{sales_{pQ_1}}\right) - \min_q\left(\ln\frac{sales_{pq}}{sales_{pQ_1}}\right) \\ &= \ln\left(\frac{\max_q\left(sales_{pq}/sales_{pQ_1}\right)}{\min_q\left(sales_{pq}/sales_{pQ_1}\right)}\right) \\ &= \ln\left(\frac{\max_q sales_{pq}}{\min_q sales_{pq}}\right) \end{aligned}$$

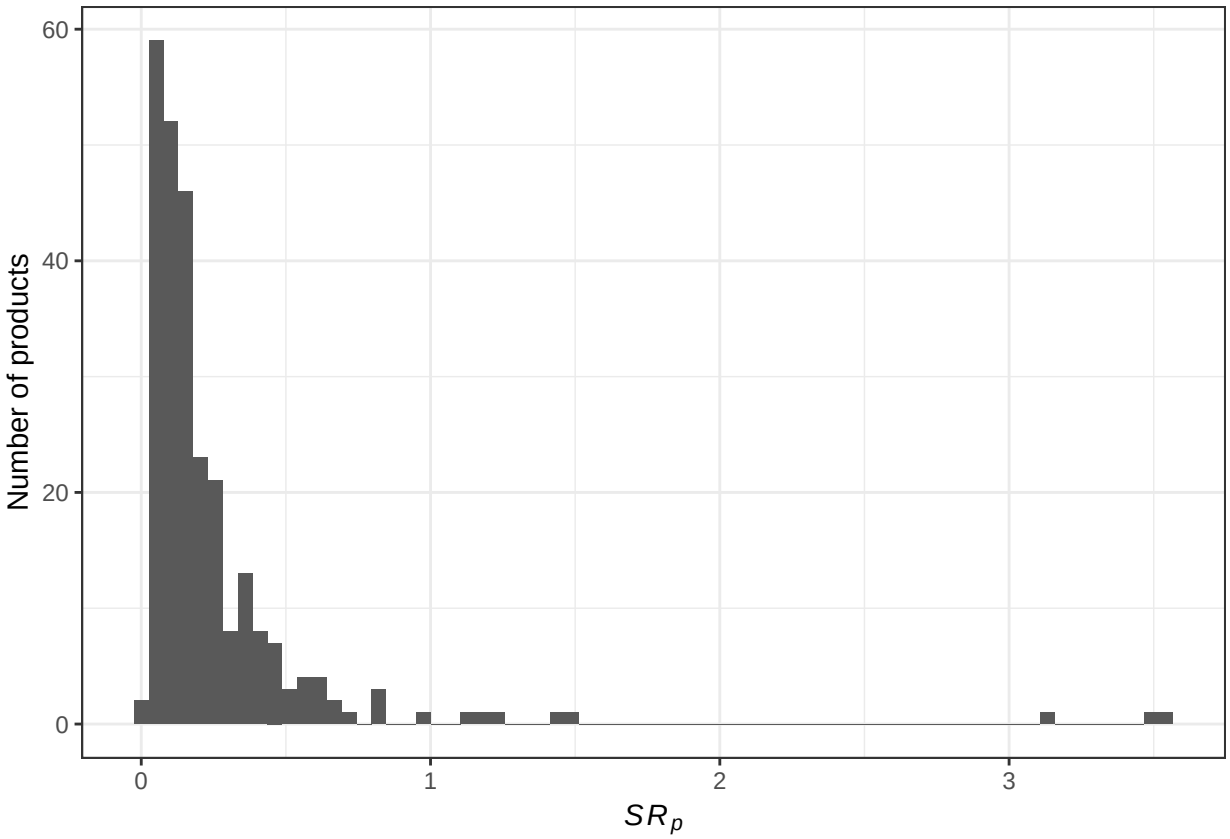

Figure 3: Distribution of Seasonality Range across Products

estimate of the actual ratio of peak-season sales to low-season sales.[3] We use SR_p primarily as an index to compare seasonality across goods.

We can use these resulting SR_p indicators to examine the distribution of seasonality within the 265 best-selling products left in the balanced panel.

Most products appear to exhibit some sort of seasonality (Figure 3). However, seasonality across products exhibits considerable heterogeneity, and there are clearly some outliers.

The most seasonal goods by far are specialty chocolates, electric fans, and sunscreens/sunblocks (see Table 4). Not surprisingly, these are goods one would expect to be correlated with seasons. Prices of fans and sun products peak as summer begins, in Q_2, and reach their lowest point as winter begins, in Q_4. Specialty chocolates have highest sales during the quarter that contains Valentine's day.

We can extend this to an arbitrary number of stores and years (assuming a balanced panel for each product). Let $\bar{q}$ denote the quarter that maximizes α_{qp}, let $\underline{q}$ denote the quarter that minimizes α_{qp}, and let $N_{s,y}$ denote the number of store-years in the data. It is simple to show that

$$SR_p = \frac{1}{N_{s,y}} \sum_{s,y} \ln\Big(\frac{sales_{ps\bar{q}y}}{sales_{ps\underline{q}y}}\Big),$$

which is clearly the average over all store-years of the log ratio of sales in $\bar{q}$ to sales in $\underline{q}$.

[3] Continuing notation from footnote 4, we have

$$e^{SR_p} = \Big(\prod_{s,y} \frac{sales_{ps\bar{q}y}}{sales_{ps\underline{q}y}}\Big)^{\frac{1}{N_{s,y}}}$$

Table 4: Products with highest/lowest seasonality range

Module description	SR_p	e^{SR_p}	Low season	Peak season
Highest SR_p				
Candy-chocolate-special	3.56	35.15	3	1
Fan and ceiling fan appliance	3.51	33.47	4	2
Suntan preparations - sunscreens & sunblocks	3.13	22.84	4	2
Charcoal	1.51	4.51	4	2
Video and computer games	1.41	4.11	2	4
Personal planners binders and folders	1.22	3.40	2	3
Fresh strawberries	1.18	3.26	4	2
Ice	1.13	3.09	1	3
Video products prerecorded	0.95	2.59	3	4
Cough syrups & tablets	0.84	2.32	3	1
Frozen meat - ground beef	0.84	2.31	4	3
Cameras	0.82	2.27	1	4
Candy-chocolate-miniatures	0.69	2.00	3	4
Dough products - cookies & brownies - refrigerated	0.67	1.96	2	4
Soup-canned	0.66	1.93	2	4
Lowest SR_p				
Dinners-frozen	0.04	1.04	4	1
Baby milk and milk flavoring	0.04	1.04	3	1
Soap - bar	0.04	1.04	1	3
Razor blades	0.04	1.04	4	3
Cosmetics-eyebrow & eye liner	0.04	1.04	3	4
Dairy-flavored milk-refrigerated	0.04	1.04	2	4
Adult-incontinence	0.03	1.03	3	1
Detergents - heavy duty - liquid	0.03	1.03	2	1
Detergents-packaged	0.03	1.03	4	3
Bakery-muffins-fresh	0.03	1.03	4	2
Bakery - bread - fresh	0.03	1.03	2	3
Cheese-processed slices-american	0.03	1.03	2	3
Bags - tall kitchen	0.03	1.03	2	1
Dog food - wet type	0.02	1.02	1	4
Tooth cleaners	0.02	1.02	4	1

This serious seasonality is only exhibited for 3 out of the 265 products we examine, and is thus unlikely to be the driver of seasonality issues in the data. However, it is worth noting that items with more seasonal sales seem more likely to be non-food items (and thus taxable), while items with less seasonal sales are food items.

County-specific seasonality

I also estimate county-specific seasonality, using the same method I used to estimate product-specific seasonality (but replacing product-level season effects with county-level season effects). Note that this implicitly weights all store-product level sales equally within each county and does not account for interactions between product and county seasonality. I denote the resulting seasonality range SR_c.

We can see in (Figure 4) that county-specific seasonality follows a similar pattern as product-specific seasonality, but with a shorter right tail (note the differences in the x-axis scale between Figure 3 and Figure 4). There are some outliers on the right tail of the distribution.

A t-test shows that the mean populations of counties with $SR_c < 0.3$ and those with $SR_c >= 0.3$ are

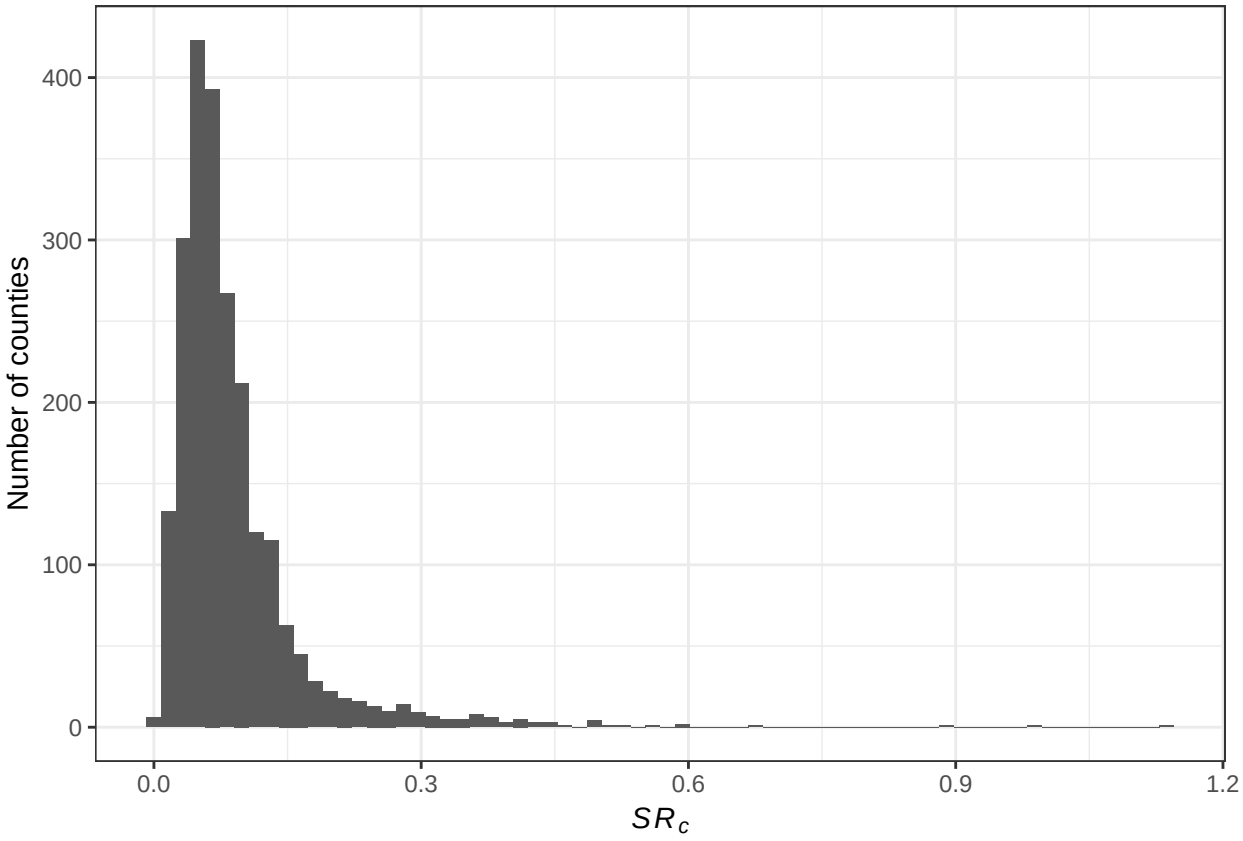

Figure 4: Distribution of Seasonality Range across Counties

Table 5: Products with highest/lowest seasonality range

Module description	SR_p	e^{SR_p}	Low season	Peak season
Highest SR_p				
Currituck County NC	1.14	3.13	1	3
Cape May County NJ	0.99	2.69	1	3
Dare County NC	0.88	2.42	1	3
Worcester County MD	0.68	1.97	1	3
Mackinac County MI	0.60	1.82	1	3
Dickinson County IA	0.59	1.80	1	3
Hancock County ME	0.56	1.75	1	3
Barnstable County MA	0.53	1.71	1	3
Door County WI	0.51	1.67	1	3
La Paz County AZ	0.50	1.65	3	1
Lincoln County ME	0.50	1.65	1	3
Custer County SD	0.50	1.65	1	3
Vilas County WI	0.49	1.63	1	3
Montmorency County MI	0.46	1.58	1	3
Collier County FL	0.45	1.56	3	1
Lowest SR_p				
Inyo County CA	0.15	1.16	1	3
Nevada County CA	0.15	1.16	2	3
Anderson County KS	0.15	1.16	3	4
Willacy County TX	0.15	1.16	3	1
Erie County OH	0.15	1.16	1	3
Wexford County MI	0.15	1.16	1	3
Fentress County TN	0.15	1.16	2	4
Crittenden County KY	0.15	1.16	1	4
Lewis County NY	0.15	1.16	2	4
Hughes County SD	0.15	1.16	1	4
Modoc County CA	0.15	1.16	1	3
Garfield County CO	0.15	1.16	1	3
McDonald County MO	0.15	1.16	4	3
Larue County KY	0.15	1.16	2	4
Barbour County WV	0.15	1.16	1	3

significantly different (t-stat = 6.1) - counties with abnormally large SR_c are on average smaller counties.[4]

[4]These results hold when population is replaced with an ordered population-rank variable (t-stat = -3.5)