Data Construction Details

August 18, 2024

The Cost of Curbing Externalities with Market Power by Conlon and Rao

Note: Some of this may be duplicative with text in the Appendix or paper itself.

1. Data Construction Overview

The main data construction is divided into the following tasks:[1]

1. Step 1: Construct a Monthly Series for Prices and Quantity

   (a.) `1a_monthly_quantity.py`: Combines DISCUS shipment data with NielsenIQ purchase data (for our set of products) and calculates product-level sales totals by `master_prod_id` and year.
   Inputs: `discus_ct_quarterly.parquet` (DISCUS), `ct_only_sales.parquet` (Nielsen).
   Outputs: `annual_discus_raw.parquet`

   (b.) `1b_predict_discus.R`: Impute the annual shipment data when it is missing using the retail sales data. Predict with annual sales data. Our preferred specification uses a random forest and includes characteristics for (size, category, year) and average price. We obtain similar results where $\beta(x_{j,y})$ is a spline (interacted with size) and fixed effects for category, size, and year.

   $$\log q_{j,y}^{discus} \sim \beta(x_{j,y}) \cdot \log q_{j,y}^{nielsen} + \gamma \cdot x_{j,y} + \varepsilon_{j,y}$$

   Inputs: `annual_discus_raw.parquet`
   Outputs: `annual_discus_predicted.parquet`

   (c.) `1c_monthly_prices.py`: Construct price series at the level of wholesaler-product-month. Also outputs a list of manufacturer and wholesaler prices so that missing manufacturer prices can be imputed. To do this we must construct a harmonized *monthly* quantity series (Nielsen and DISCUS – similar to `1a_`) which is *NOT* used in estimation.
   Inputs: `discus_ct_quarterly.parquet` (DISCUS), `ct_only_sales.parquet` (Nielsen), `annual_discus_predicted.parquet` (Imputed), `prices_monthly_cleaned.parquet` (Wholesale), `manuf_prices_raw.parquet` (Manufacturer).
   Outputs: `all_quantity_monthly.parquet`, `price_data.parquet`, `predict_manuf.parquet`.

[1] We list the most important inputs (and outputs) for each script, but not necessarily all of the crosswalks and supporting files.

(d.) `1d_impute_manuf.R`: Impute the manufacturer prices when they are missing using both a spline with multiple fixed effects and a random forest. Fit on observed data, and predict on missing data.

$$\log p_{j,t}^{manuf} \sim \beta(x_{j,t}) \cdot \log p_{j,t}^{wholesale} + \gamma \cdot x_{j,t} + \varepsilon_{j,t}$$

Inputs: `predict_manuf.parquet`
Outputs: `imputed_manuf_prices.parquet`

(e.) `1e_combine_monthly.py`: Combine the DISCUS (wholesale shipment) and NielsenIQ (retail) data with the monthly price data to construct a monthly series of prices and quantities (including the imputed annual shipment data, and imputed manufacturer pricing data). These are primarily used for the descriptive analyses *NOT* in estimation.
Inputs: `discus_ct_quarterly.parquet` (DISCUS), `ct_only_sales.parquet` (Nielsen), `annual_discus_predicted.parquet` (Imputed), `prices_monthly_cleaned.parquet` (Wholesale), `imputed_manuf_prices.parquet` (Manufacturer).
Outputs: `all_quantity_monthly.parquet` (Quantity Only)
`official_data_monthly.parquet` (Quantity and Price).

2. Step 2: Construct a Quarterly Series for Prices and Quantity.

   `2_combine_quarterly.py`: Similar to `1e_`. Combine the DISCUS (wholesale shipment) and NielsenIQ (retail) data with the quarterly price data to construct a quarterly series of prices and quantities (including the imputed annual shipment data, and imputed manufacturer pricing data). This becomes the basis for the estimation data (constructed in step 5).
   Inputs: `discus_ct_quarterly.parquet` (DISCUS), `ct_only_sales.parquet` (Nielsen), `annual_discus_predicted.parquet` (Imputed), `prices_monthly_cleaned.parquet` (Wholesale), `manuf_prices_raw.parquet` (Manufacturer).
   Outputs: `all_quantity_monthly.parquet` (Quantity only), `official_data_quarterly.parquet` (Price and Quantity).

3. Step 3: Estimate income distribution from the Nielsen panelists

   `3_process_agents.py`: Read the NielsenIQ panelist data and construct the "agent data" for estimation. Fit a lognormal distribution for income to the NielsenIQ panelist data for all households in Connecticut and again for all spirit purchasers in Connecticut. Compute the "expanded" income bins: $\mathbb{E}[y_i \mid Inc_{lb} \leq y_i \leq Inc_{ub}]$ for each of the bins of the income distribution in the NielsenIQ panelist data. (Note that the bin definitions change in 2010). These become the distribution we draw income from in estimation. (Previous versions included age and race, but they were never significant).
   Inputs: `ct_liquor_X.parquet` (trips, purchases, panelists, stores).

Outputs: `agent_moments_all.csv`, `agent_moments_liquor.csv`, `agent_income_expanded.csv`

4. Step 4: Construct Micro-Moments from Nielsen panelists (including "second choices")

   `4_micro_moments_by_year.py`: Compute the micro moments from the NielsenIQ panelist data. These are described in more detail in the paper. Compute a separate set of "same category" moments from the panelist data.
   Inputs: `ct_liquor_X.parquet` (trips, purchases, panelists, stores) and `agent_income_expanded.csv`
   Outputs: `same_category_moments.csv`, `micro_moments_year_categories.csv`, `micro_moments_cov_year.npy`

5. Step 5: Construct the datasets used in estimation

   `5_data_prep_quarterly.py`

   (a.) Restrict the set of products (three sizes, hard alcohol, no cognac, prices $<$ \$60/L, sales rank in top 750)

   (b.) Unused: Compute 25 Grouped Fixed Effects.

   (c.) Construct the Λ or κ ownership matrix.

   (d.) Compute the Gandhi Houde IV's (quadratic): category, size, proof, flavored, p_m and processed with principal components. Combine with cost IV's (taxes and p_m).

   (e.) Draw various integration rules `draws_XXX.parquet`

2. Product Data

This is found in `master_product_info.parquet`.

We focus on the three largest and most popular bottle sizes (750mL, 1L, and 1.75L). We make the conscious decision to exclude the smaller bottle sizes (such as 50-100mL "nips", which are most frequently found on airplanes and in hotel mini-bars, and 375mL or "pint" or "flask" size bottles). The included bottle sizes comprise the overwhelming majority of sales (by volume), and the widest variety or products are available in the 750mL size.

There is some segmentation by sales channel and bottle size. The majority of sales in "on premise" locations (bars and restaurants) are the 750mL and 1L bottles. With a few exceptions (Grey Goose Vodka), 1L bottles are relatively rare in "off premise" retail liquor stores.

We organize each product into six categories (Gin, Tequila, Rum, Vodka, North American Whiskey, and UK Whisky). These follow standard industry classifications. Our North American Whiskey category includes both US manufactured whiskeys (Bourbon, Rye, etc.) as well as Canadian whiskeys. Our UK whisky definition includes both Scotch and Irish whiskies. In Table 1 we reproduce the product summary table from the paper. In Table 2 we report the share of sales (by volume) for each category-size combination as a cross-tab. The interesting thing here is that Vodka is over-represented among large bottles while Tequila and UK Whiskey are over-represented among 750mL bottles.

For each product, we also track its proof (twice the percentage of alcohol by volume ABV%) and whether or not it is "flavored". The overwhelming majority of distilled spirits are bottled at 80 proof (40% ABV), while most flavored vodkas are bottled at 60 proof. Malibu Rum is bottled at 42 proof (21% ABV), overproof whiskies at 94.6 proof, overproof vodkas at 100 proof, and Bacardi 151 at 151 proof.[2]

The overwhelming majority of flavored products are either vodkas or rums. In general, we consolidate several flavored products with the same price and proof (and size) (ie: Stolichnaya Vodka in Raspberry and Blueberry). We do this mostly because reporting for prices of flavored products can be haphazard (we see sales for Blueberry Vodka but no price is reported that month, while the prices of Raspberry Vodka are reported for each month etc.). In this case we consider *Stolichnaya Flavored Vodka 60PF 750mL* as our consolidated product. There is typically little to no variation within different flavors of a flavored vodka/rum. One exception is when limited time flavors are on clearance (in this case we try not to use clearance prices).

There are several product categories we do not include. We focus solely on distilled spirits, and thus exclude beer and wine. This also means we exlcude fortified wines (vermouth) and other ready-to-drink malt beverages (hard cider, hard lemonade, etc). Hard Seltzers aren't an issue for

[2]Proof is reported for NielsenIQ products by UPC as part of its product description. Proof is also reported as part of the product descriptions with the price posting both by manufacturers and wholesalers. We cross check proof information to verify that the UPC level proof information coincides with proof information in price postings. In the rare cases of disagreement we verify the proof information from the distiller/manufacturer website.

2007-2013, as mostly don't exist in 2007-2013.

We also exclude brandies, cordials, and liqueurs. The majority of these products are much lower alcohol content (25% ABV or less on average) than the other distilled spirits (40% ABV on average). Some well known brands of liqueurs include: Triple-Sec, Cointreau, Gran Marnier, Bailey's Irish Cream, Kahlua, etc. This also includes herbal liqueurs and apertifs such as: Campari, Maraschino, Cynar, Fernet, Chartreuse, etc.

The only spirits category we exclude is Cognac, which is generally bottled at high proof (40% ABV). This category is dominated by two major brands: Hennesey and Courvoisier. The reason we exclude Cognacs is that they often include both age statement (4 years or "Very Old" VO, 8 years or "Extra Old" XO) as well as vintage statements (1998, etc.). These are not always consistently reported across datasets (particularly in wholesale prices) and price differences can be orders of magnitude. Rather than include Hennesey and Courvoisier only, we chose to exclude the entire category.

To summarize a "product" has the following pieces of information:

- Brand Name (*Smirnoff Vodka*)
- Package Size (750mL, 1L, 1.75L)
- Category (Gin, Rum, Tequila, NA Whiskey, UK Whiskey, Vodka)
- Flavored Dummy $\{0, 1\}$
- Proof/ Ethanol Content: $42, 60, 80, 94, \ldots$ (PF = 2× % Alcohol By Volume)
- Manufacturer/Distiller/Importer Name (e.g. Diageo, Bacardi, etc.)

					Manufacturer		Wholesaler		Retailer	
	# Obs	Share	Proof	% Flavored	Price	Margin	Price	Margin	Price	Margin
Gin	59	7.40	87.07	0.02	11.15	3.01	16.21	3.79	18.72	2.34
Rum	147	17.50	73.63	0.21	10.17	2.60	15.08	3.65	17.60	2.52
Tequila	92	4.90	80.04	0.00	15.17	4.07	22.05	5.60	28.51	4.70
Vodka	208	44.80	79.19	0.15	10.73	2.79	15.42	3.42	18.05	2.54
NA Whiskey	127	15.20	81.80	0.00	11.59	3.18	17.41	4.54	20.08	2.76
UK Whiskey	102	10.20	80.79	0.00	18.36	4.51	25.04	5.41	28.15	3.12
750mL	310	20.10	79.05	0.18	16.44	4.32	23.57	5.85	28.32	4.74
1L	174	23.20	79.32	0.12	13.80	3.73	19.92	4.85	24.85	4.35
1.75L	251	56.70	79.55	0.08	9.32	2.36	13.53	2.94	14.91	1.36
All	735	100.00	79.40	0.11	11.79	3.07	17.03	3.97	19.82	2.71

Table 1: Percentage of Sales by Product Size and Category

Source: Harmonized Quantity Data (Reproduced from Paper)

Category	0.750000	1.000000	1.750000	Total
Tequila	1.81	1.83	1.47	5.11
Gin	1.13	1.58	4.59	7.30
UK Whiskey	2.68	2.26	5.40	10.33
NA Whiskey	3.06	3.18	9.09	15.32
Rum	4.20	5.01	8.41	17.62
Vodka	7.96	9.49	26.87	44.31

Table 2: Percentage of Sales by Product Size and Category

Source: Harmonized Quantity Data

3. Quantity Data

This is implemented in `1_combine_quarterly.py`.
Our main quantity data comes from a combination of two sources:

1. DISCUS Shipment Data
2. NielsenIQ Retail Scanner Data

We provide our own crosswalks from the DISCUS data (`crosswalk_discus.parquet`) and NielsenIQ data (`crosswalk_nielsen.parquet`) to our harmonized product identifier (`master_prod_id`).

NielsenIQ Data

There are approximately 1,200 retail package stores in Connecticut that are licensed to sell spirits. Beer can be purchased at supermarkets, but only licensed package stores can sell distilled spirits. The NielsenIQ data track weekly retail sales at 25 stores. Of those 25 stores, 5 are independent stores and the remaining 20 belong to one of three chains. These chain stores tend to be *much* larger than the the typical store. The largest chain accounts for roughly half of the sales volume among these 25 stores. We estimate these 25 stores account for approximately 5.5% of all distilled spirits purchased in Connecticut.[3]

Each store reports UPC-level sales weekly along with the average price (revenue divided by unit sales). We use the reported retail prices only for the purposes of computing retailer markups in summary tables, and cross-state comparison figures, not in our demand model (which focuses on the decisions of wholesalers).

We construct a crosswalk between `upc` and `upc_ver_uc` in the NielsenIQ data and our own `master_prod_id`. The same product may have multiple UPC's (holiday packaging, gift packs, etc. all leading to new UPC's). This was ultimately done by hand with the aid of some fuzzy string matching libraries to start things off. At the level of a product-month (statewide) using our `master_prod_id` we have 60,758 observations (total sales and total revenue).

[3]The NielsenIQ data agreement prevents us from providing additional chain level information.

Because it comprises only 25 retail stores (and no bars/restaurants) the NielsenIQ retailer data is NOT the primary source of quantity data in our main analysis. We use it mostly for descriptive purposes and cross state comparisons. See below where NielsenIQ data is used:

- Cross state comparisons and price-indices (Figures 1,2,3)
- To construct the seasonal sales pattern in our "Harmonized Quantity Data" when DISCUS shipment data are infrequent (Step 2)
- To impute annual shipments for non-DISCUS distillers (Step 1b)

DISCUS Data

We also have access to proprietary data from the Distilled Spirits Council of the United States an industry trade group that represents most of the large distiller/manufacturers. These data are from January 2007 to July 2013, and track shipments from manufacturers to wholesalers. For the set of products sold by our manufacturer/distillers we see the universe of shipments to Connecticut wholesalers (actually nationwide). This includes not only products sold in retail package stores, but also products sold to (and consumed) in "on premise" locations such as bars and restaurants. We also observe the identity of the wholesaler to which the products are shipped. These data are technically recorded shipment by shipment, but we only see the month and year (not the exact date) for each shipment.

These data record the number of cases shipped by product and size (ie: *Smirnoff Cherry Vodka 1.75L*) (using three different measures: "Physical Cases", "Bureau of Alcohol Tobacco and Firearms Cases", and "Nine Liter Equivalent Cases"). We use the BATF cases measure which mostly coincides with the other two. Under the BATF measure, there are (12) 750mL bottles; (12) 1L bottles; and (6) 1.75L bottles per case.[4] These data also include a brand identifier and a bottle size. We construct a crosswalk from the DISCUS product identifiers to our `master_prod_id`. We see 32,422 observations at the product-month level (including shipments to all wholesalers) using our consolidated product identifier.

This dataset has two main limitations. The first is that it does not track shipments from independent manufacturers who are not part of DISCUS (and DISCUS membership changes over time). Still, we estimate the coverage to be between 75% and 80% of total shipments for the state of Connecticut. The second is that shipments from manufacturers to wholesalers can be irregular. The most popular products are shipped in consistent monthly quantities, while less popular products may see only a few shipments per year, and some unsuccessful products may see only a single initial shipment. This creates some challenges when we wish to apportion the sales to months or quarters.

[4]For rare and expensive products the "physical cases" measure may count single bottle "cases".

4. Harmonizing Quantity Data

This is implemented in steps (1) and (2) above.

4.1. Imputing Missing Annual Shipments

The next step is to construct a harmonized measure of sales (measured in Liters) by product/brand and bottle size. The first challenge is that we have the universe of DISCUS shipments for 77.7% of sales volume, and retail scanner data for 25 stores which account for only 5.5% of total sales volume (arguably a larger share of off-premise purchases).

We proceed in two steps: first we construct annual totals (measured in total liters) from each dataset for each product. This gives us 7,695 product-years for which we have NielsenIQ (retail scanner) data and 5,138 for which we also have DISCUS (shipment) data.

For cases where we don't have total annual shipments from DISCUS, we impute what shipments to wholesalers would have been by using products with similar retail scanner sales. We consider three cases of coverage: (a) both NielsenIQ (retail scanner) and DISCUS (shipment) data which serves as our *training data*; (b) just retail scanner data, which serves as our *imputed data*; (c) just DISCUS shipment data which we ignore for the imputation exercise; (d) "Neither" which are predominantly non-DISCUS products with sporadic or negligble $\log Q < 2$ retail sales.[5]

	Sample	Observations	Sales
Both	Training	2948	74.0%
Nielsen Only	Imputed	2369	22.2%
DISCUS Only	Ignore	2190	3.8%
Neither	Ignore	188	n/a

Table 3: Cross Coverage of Annual Totals Data

Source: NielsenIQ. Sales in thousands of bottles.
Nielsen Data: We require annual $\log(Q) > 2$; Discus Data: We require $log(Q) > 9$.
DISCUS Only and Neither are excluded from the imputation step.

We label case (a) as training data and predict the total annual discus shipments for each product as a function of total retail sales, bottle size and category dummies, and year dummies. We try a variety of prediction models (linear fixed effects, cubic splines in retail sales, and a random forest regression). All of them fit the data relatively well $R^2 > 0.81$ for the regressions and $R^2 = 0.878$ for the random forest and give highly similar predictions. The random forest seems to provide the best fit and least number of parameters, so that is our preferred specification.

We obtain similar results where $\beta(x_{j,y})$ is a spline (interacted with size) and fixed effects for category, size, and year.

$$\log q_{j,y}^{discus} \sim \beta(x_{j,y}) \cdot \log q_{j,y}^{nielsen} + \gamma \cdot x_{j,y} + \varepsilon_{j,y}$$

[5]We don't exclude these from our overall sample

We plot the relationship between the retail sales and DISCUS shipments for our training data in Figure 1. We see a nearly linear relationship (in logs) between annual retail sales and annual DISCUS shipments (with some variation across size and category). We also plot the in-sample fit of our random forest (comparing $\log q_{j,y}^{discus}$ with $\log \widehat{q_{j,y}^{discus}}$) which closely follows the 45 degree line.

In Figure 2 we plot the (out-of-sample) forecast of DISCUS shipments using the retail scanner sales (and other covariates) from our random forest. As an example to justify heterogeneous $\beta(x_{j,y})$ our predictions suggest that for 1L Vodka products (pink triangles) we infer higher shipment totals for each level of retail sales, which is consistent with our training data in Figure 1.[6] This is also consistent with industry information suggesting that 1L bottles are most commonly purchased by bars and restaurants (and are thus not as popular in the retail scanner data).

This means that we have observed annual shipments for 77.8% of our data (Both + Discus Only), and imputed annual shipments (using retail sales totals) for the remaining 22.2%.

4.2. Assigning Annual Sales to Quarters

The second task involves apportioning these annual data to each quarter. This is meant to capture both the seasonality and how sales respond to changes in prices over time.

We describe our main algorithm as follows, we apportion the annual sales to each quarter based on the retail sales totals for product-quarter-year $q_{j,t}$. We assign the sales based on the following (Empirical Bayes Shrinkage) formula:

$$w_{jt} = \lambda \cdot \frac{q_{jt}}{\sum_t q_{jt}} + (1-\lambda) \cdot \frac{\sum_{j \in \mathcal{G}} q_{jt}}{\sum_{j \in \mathcal{G},t} q_{jt}} \text{ with } \lambda = \frac{\sum_t q_{jt}}{k + \sum_t q_{jt}}. \tag{1}$$

The first term apportions sales to the seasonal pattern for the product-quarter-year $q_{j,t}$ (e.g. Smirnoff Vodka 1750mL), and the second apportions sales to the seasonal pattern for the category-quarter-year (Vodka). When the overall product sales total $\sum_t q_{jt}$ is small, we place less weight λ on the product-level estimates and more on the category estimates $(1-\lambda)$. Here k functions like a number of "pseudo-observations" from the category level "prior". The idea is that when we have complete data with seasonal patterns we use those, and for smaller products we assume they follow the season pattern for the category (ie: more Whiskey sales in Q4, or more Rum sales in summer).

We try to use the NielsenIQ Retail Scanner data to assign the seasonal patterns in Equation (1) whenever possible. This may not be possible if the product does not appear in the NielsenIQ dataset or of it does appear but has very small numbers of sales that year. In those cases we try to infer the seasonal pattern from the DISCUS shipment data. If the shipment data are also sparse (rare) we then rely primarily on the shrinkage to the category means.

A more detailed explanation of our exact algorithm is as follows:

[6] Removing category and size information seems to make our (cross-validated) forecasts substantially worse, even if we are worried about imputing larger than reasonable values for 1L products.

Figure 1: Training Data: Annual Shipment/Sales Totals by Product

Figure 2: Forecast: Annual Shipment/Sales Totals by Product

1. For each product-year and dataset (NielsenIQ or DISCUS) we construct $(\hat{w}_{jt}^{discus}, \hat{w}_{jt}^{nielsen})$ using Equation (1) using $k^{discus} = 1500$ and $k^{nielsen} = 25$ (in liters) as our "prior"

$$w_{jt} = \lambda \cdot \frac{q_{jt}}{\sum_t q_{jt}} + (1-\lambda) \cdot \frac{\sum_{j \in \mathcal{G}} q_{jt}}{\sum_{j \in \mathcal{G},t} q_{jt}} \text{ with } \lambda = \frac{\sum_t q_{jt}}{k + \sum_t q_{jt}}.$$

2. We then take a weighted average

$$\widehat{w}_{jt} = \gamma \cdot \widehat{w}_{jt}^{nielsen} + (1-\gamma)\widehat{w}_{jt}^{discus}$$

3. Then we consider the following cases:

   (a) (22% of sales): $> 50L$ of annual sales in the NielsenIQ retail data; 2 quarters or fewer w/ DISCUS shipment $\rightarrow$ set $\gamma = 1$

   (b) (5% of sales): $< 50L$ of annual sales in the NielsenIQ retail data; 3 quarters or more w/ DISCUS shipment $\rightarrow$ set $\gamma = 0$

   (c) (72% of sales): $> 50L$ of annual sales in the NielsenIQ retail data; 3 quarters or more w/ DISCUS shipment $\rightarrow$ set $\gamma = 0.9 + \min\left\{0.1, \frac{q_{jt}^{nielsen} - k^{nielsen}}{k^{discus}}\right\}$. This uses the NielsenIQ data but will mix in some of the DISCUS data when the Nielsen sales are very small.

   (d) (1.1% of sales): $< 50L$ of annual sales in the NielsenIQ retail data; 2 quarters or fewer w/ DISCUS shipment $\rightarrow$ set $\gamma = 0.5$ (in practice this is the simple average between the category means for both datasets since $\lambda \approx 0$).

4. Finally, we construct $q_{j,quarter} = a_{year} \cdot \widehat{w}_{j,quarter} \cdot Q_{j,year}$.

One big advantage of the overall shrinkage procedure is that it guarantees that $w_{jt} > 0$ for all quarters within a year, and that the annual sales match our estimates from the previous section. Otherwise we follow the seasonal patterns from the NielsenIQ retail data whenever possible and use annual totals from the DISCUS data whenever possible.

4.3. Harmonizing with TTB Data

We compare our estimated sales volume (in liters) to that produced by the U.S. Treasury Alcohol and Tobacco Tax Trade Bureau (TTB). The TTB data compute "apparent consumption" as calculated by from tax records.[7] One advantage of these data is that because they are derived from tax records they include both on- and off- premise consumption. On one hand these should be close to

[7] See `https://pubs.niaaa.nih.gov/publications/surveillance102/CONS13.pdf` for a description.

"ground truth" on the other hand there are some assumptions that may make them different from our dataset.

We report the total sales for our set of products from the DISCUS and Nielsen data (in millions of liters) and from our "Harmonized" sample that combines the two data sources in Table 4. We also use the TTB data to report total "apparent consumption" (in millions of liters) and the per-capita consumption of individuals of legal drinking age. Our set of products explains about 55% of total apparent spirits consumption in the early year and closer to 50% in the later years. This is likely driven by the expansion of non-DISCUS members, and some members ceasing to report data in later years.

We need to make sure that our "harmonized" quantity estimates reflect the fact that the overall spirits category is growing even while (due to coverage issues) our DISCUS and NielsenIQ samples may be declining. To address this, we inflate the annual sales year by year to match the annual "apparent consumption" numbers reported by TTB. This requires scaling by a factor $a_{year} \in (1.80, 2.03)$ for each year so that the totals of the "Harmonized" data match the TTB totals in Table 4. The level doesn't matter as much as the variation across years. We could take $a_{year}/2$ for each year and this would simply be equivalent in the model to modifying the outside good share or arrival rate of consumers.

The main differences are that the TTB data include a wider range of distilled spirits products (including lower alcohol content Cordials and Liqueurs) which can be a significant amount of volume even though they represent a smaller amount of ethanol. In the TTB data the average ABV % is 37% (in our data it is 39.7%). We've also intentionally excluded Cognac and Brandy products, which are often full ABV of 40% or more. We estimate that combined this represents less than 10% of volume.

The TTB data also include other products we've intentionally excluded from our analysis including: smaller package sizes (100 and 150mL "nips" and 375mL flasks). We don't have the ability to estimate the share of lower-proof Cordials and Liqueurs, but industry data suggests that 10% of ethanol would be represent an upper bound. Our sample (see the product descriptions above) also excludes very expensive products (over \$60/L) but we estimate this to be a small fraction of overall sales.

The other discrepancy might arise from the sales of spirits products that are neither DISCUS members, nor are available in the 25 retail stores in the NielsenIQ dataset. For some of these products we observe their posted wholesale prices, but no retail sales or shipment data. It is impossible to estimate this number, but none of these brands are in the top 500 of national sales, and thus we expect this to be a very small number of niche products.

A likely reason that TTB data might be overstate spirits volume is methodological differences in how the TTB arrives at the overall consumption numbers. Federal taxes are paid based on ethanol content for spirits, but conversion between ethanol content and volume in the TTB dataset

	DISCUS	Nielsen	Harmonized	TTB	Per Capita (TTB)
2007	10.60	0.83	12.45	22.37	8.77
2008	10.74	0.81	12.72	22.88	8.89
2009	9.31	0.82	11.69	23.01	8.87
2010	9.80	0.82	12.68	23.82	9.11
2011	9.61	0.83	12.51	24.07	9.17
2012	9.53	0.62	12.30	25.03	9.49
2013	4.87	0.52	12.66	25.31	9.55

Table 4: Annual Totals by Data Source (Millions of Liters)

Per-Capita Numbers are Liters per legal drinking age adult.
DISCUS data end in July 2013 (Harmonized numbers adjusted for rest of year).

assumes a fixed ABV for product type, and does not account for the fact that flavored Vodka and Rum is generally significantly lower alcohol content. Because flavored prodcuts are 22% of Rum (by Volume) and 15% of Vodka this tends to inflate the ethanol totals by as much as 30% for these products, and might lead to overstating implied overall volume.

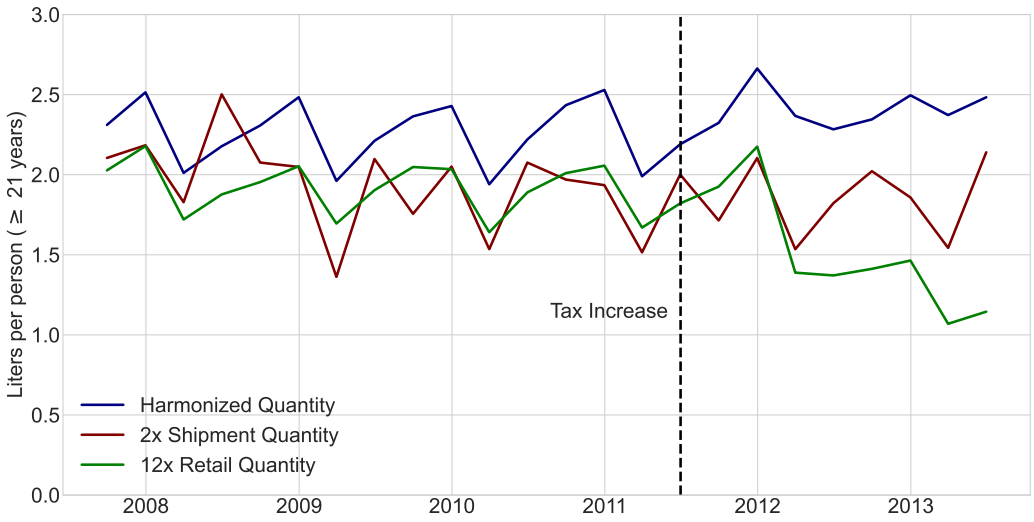

Figure 3: Sales and Shipments: Per Legal Drinking Age Pop

Source: DISCUS Shipments; NielsenIQ Scanner Sales, our Harmonized Calculations
Population is legal drinking age (LDA) pop (over 21 years) for Connecticut. (NIAAA).

In Figure 3 we report total quarterly quantity (measured in liters) from our DISCUS shipment data (scaled by 2×, the NielsenIQ retail scanner data (multiplied by 12×) and our harmonized quantity data (calibrated to match the annual TTB totals). We scale the plots to facilitate comparing the seasonal trends. Seasonal patterns arehighly similar across the DISCUS wholesale shipment data and the NielsenIQ retail sales data. This suggests that (at least at the quarterly level) there

is little lag between wholesale shipments and retail purchases (such as buildup of inventories).

Both Table 4 and Figure 3 illustrate that the spirits category is growing over time. Even after the tax increase in July 2011, sales continue to rise, but slightly slower than the previous trend. By adjusting our overall totals to match the TTB totals, we ensure we don't miss the overall growth of the spirits categoy over time. The discrepancy in the retail sales at the end of the sample seems to come from a decline in the number of stores at one of the large chains, while the decline in DISCUS sales is because certain members de-affiliate and stop sharing shipment data.

5. Price Data

This is implemented in `1c_monthly_prices.py`. We provide our own crosswalks from the wholesale product identifiers (`crosswalk_wholesale.parquet`). One key challenge is that each wholesaler uses its own convention for labeling products so these crosswalks must be constructed by hand. We encounter similar issues with manufacturer data, which rarely uses unique product identifiers and often uses product names instead, we provide the corresponding crosswalk in `manufacturer_info.parquet`. These crosswalks may not be perfect, but represent several hundred hours of work.

Our price data comes from the Connecticut Department of Consumer Protection (DCP) price postings. Both manufacturer/distillers (upstream) and wholesaler/distributors (downstream) are required to post prices each month. Manufacturer/Distillers are often large multinationals like Bacardi or Diageo, but include some smaller independent firms such as Black Prince (Dubra Vodka). Wholesalers are single-state entities that are not allowed to ship across state lines and retailers (both liquor stores and bars and restaurants) are required to purchase from an in-state wholesaler. These firms are not household names and include: Brescome-Barton, Allan S. Goodman, Eder Brothers, Hartley Parker. In some cases (CT Distributors inc.), wholesalers are commonly owned by multi-state holding companies (Charmer-Sunbelt), but required (by law) to operate independently in each state.

The price posting procedure is detailed in the paper. By the 12th day of each month, every manufacturer and wholesaler is required to post a list of all of the products it plans to sell the following month and a uniform case price, which we convert into the equivalent price per liter. The important issue for the data construction is that each firm posts both a full price list for every product they sell, and then after a lookback period posts "amendments" to the initial price list. In theory, we are supposed to observe both the initial price list and the amendments. In practice, we generally observe one of:

(a.) prices as amendmened (good)

(b.) initial un-amended prices (ok)

(c.) only the amendments (bad).

Moreover, for many firms it is not clear whether we observe the "as amended" prices or the "pre-amendment" prices. This prevents us from observing whether both initial prices and amendments are set in accordance with the subgame perfect strategy in the paper. The second challenge is that only a small fraction of prices are amended, so in case (c) above we don't observe prices for the majority of products. This is particularly problematic for a single wholesaler (*Connecticut Distributors Inc.*) which is among the largest in our dataset. Because this doesn't happen every month, we are mostly able to fill prices using adjacent months.

We are not completely certain, but evidence suggests that from 2010-onwards we observe only case (b) for another wholesaler (*Eder Bros*). In cases where prices alternate between "high" and "low" price points, we observe Eder setting only the "high" price point (see Figure 5 in the paper) and Figure 8 below. This is consistent with price amendments that are not observed in the data. Furthermore in some cases the dates and headers suggest these are "pre-amendment" rather than "as amended" prices. Practically speaking this is of little significance because our estimates depend on the lowest price across wholesalers.

The price postings themselves are occasionally delivered in Microsoft Excel (XLS) format, but are most often delivered in machine-readable PDF's, and in some cases are delivered as PDF images of scanned faxes (which are analyzed via OCR). The formatting is not necessarily consistent within a seller from month to month, so the resulting parsing task is non-trivial. Matters are further complicated by the fact that many wholesalers also distribute wine, which adds thousands of products whose prices we aren't necessarily interested in. The parsing task is non-trivial and represents around one thousand hours of work.

The majority of coverage gaps arise when price postings are missing, unreadable, or cover only "amendments". These cases represent approximately 3.5 million (monthly) sales of 75.2 million in our entire dataset. We are able to fill 98.2% of the missing prices (4.6% of total sales) using prices in adjacent months. We are able to fill a 0.08% of overall sales using prices for the same product in periods that are more than 12 months removed from the missing period.[8] The remaining 0.001% of sales are cases where we have no price information ever for that product and exclude it from our sample.

The manufacturer/distiller price data are even more sparse. In some cases, filings only cover products when prices are changed, rather than exhaustive list of products. For these cases we fill prices forward from the most recent price change. For products with regular monthly manufacturer filings (or those we've filled) we observe prices corresponding to 95.5% of sales in the overall dataset.

For the remaining products (4.47% of sales) we cannot find the price filings from the corresponding manufacturer. For these missing observations we impute the manufacturer prices as a function of the wholesale price, bottle size, product category, month of year, and year. We split the data into a training sample where we observe both manufacturer and wholesaler prices $N = 59{,}550$. The in-sample $R^2 = 0.96$ within $R^2 = 0.94$ for a fixed effects regression with a cubic spline (with knot points at wholesale prices of \$10, \$20, \$30 per liter), and the out-of-bag $R^2 = 0.985$ for a random forest with the same covariates. We plot the training data in Figure 4. We plot the in-sample fit in Figure 5, and the out of sample forecast for $N = 7{,}387$ observations in Figure 6.

For several top products in Figure 8 we demonstrate the intertemporal variation in the prices at the wholesale and manufacturer level. For our best selling product (1.75L Smirnoff Vodka), we see that it is sold by three wholesale firms (Goodman, Barton, and Eder) and the prices co-move in a

[8] The majority of these products do not appear to experience any change in prices during our sample.

Figure 4: Training Data: Monthly Manufacturer and Wholesale Prices

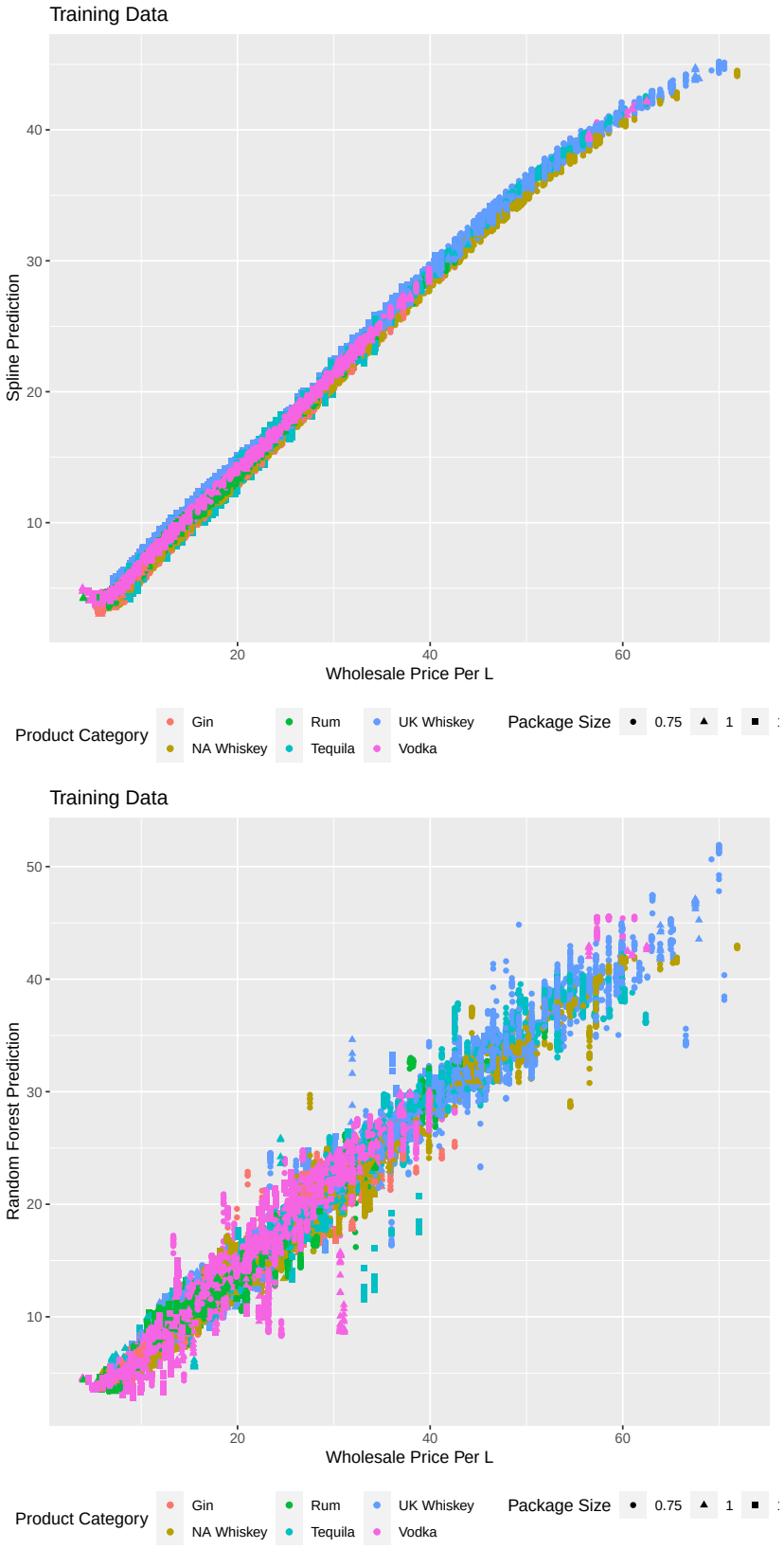

Figure 5: In Sample Fit: Piecewise Cubic Spline and Random Forest

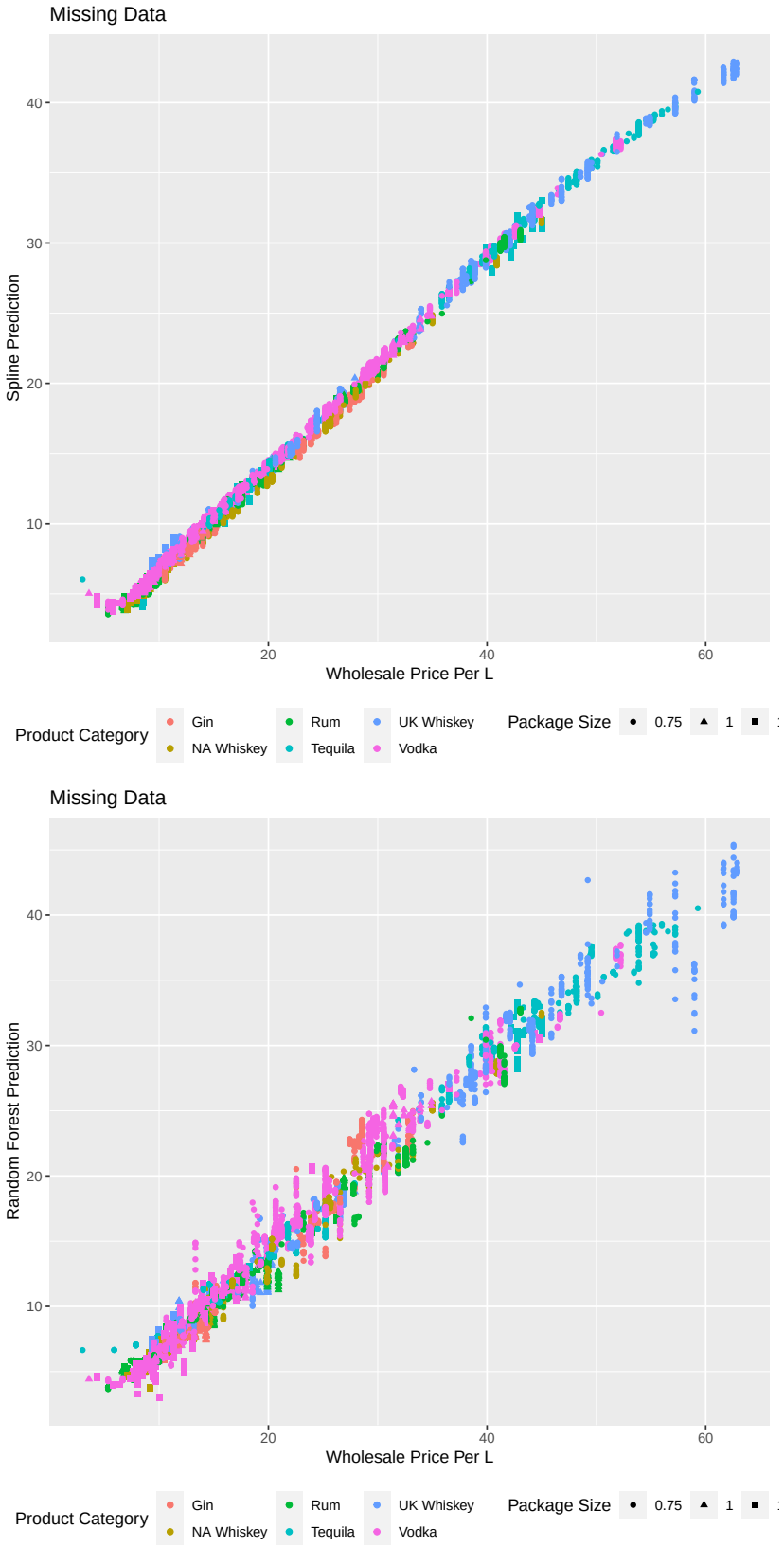

Figure 6: Out of Sample Forecast: Piecewise Cubic Spline and Random Forest

complex seasonal pattern. Starting in 2009, it appears that one of the three firms stays at the high price and does not engage in temporary price cuts. This may be because they actually don't cut prices to match competitors, or because we don't see the amended prices for Eder. We see a nearly identical pattern for 1.75L bottles of Bacardi Rum (the top selling rum), and Johnny Walker Black Label (1.75L), though there are four wholesalers instead of three. We also plot the manufacturer prices in solid black. These prices are much less volatile than wholesale prices, suggesting that something other than manufacturer price changes drive changes in wholesale price.

For Absolut Vodka (1.75L) another top product, we provide a similar plot in Figure 9. We see that the prices of the two sellers Hartley-Parker and Connectict Distributors Inc (CT Dist) largely coincide. However, we fill Hartley-Parker for parts of 2009 and 2010, in the first case the price we fill ends up being below the price of the other wholesaler while in the other case it is above. An alternative would be to use only the observed prices and only fill forward or backward when there is a single seller of the product. Towards the end of the sample (in 2013) we see the wholesale prices do not perfectly align, this may be because prices were amended but we only observed the "pre-amendment" price. This example is meant to illustrate some imperfections in the pricing data.

6. Product Level Summary Statistics

This is implemented in `Plots- Sales and Prices.ipynb`.

Each of the 1502 products in our dataset is uniquely identified by a `master_prod_id`. We have some degree of price and quantity data for 1360 of them. We rank the products by total sales volume (liters) and plot them in Figure 10. In the top panel we plot the cumulative sales by product ranking and see that 99% of sales are explained by 910 products, 95% of sales are explained by the top 512 products. The second panel plots the total sales for each product by rank and the average price for each product by rank. This suggest that price per liter is increasing in sales rank.

In Table 6 we report the share of each product category, number of unique products, and product characteristics. We also report the average (per liter) price at the retail (NielsenIQ), wholesale, and manufacturer levels (Connectict DCP posting data). For the purposes of this table only we assume that retailer marginal costs are wholesale prices; wholesaler marginal costs are manufacturer prices plus state excise taxes; and that manufacturer marginal costs are only the federal excise tax. This represents an extreme lower bound on the manufacturer costs (and an upper bound on $P - MC$). We report both average price per liter P and the average margin $P - MC$ for each size and category. Overall the suggests a wholesale Lerner markup $\frac{P-MC}{P} = 0.22$ and a retailer Lerner markup of $\frac{P-MC}{P} = 0.14$. These markups are meant to be a back of the envelope and ignore that both retailers and wholesalers face additional marginal costs involved in actually stocking and selling products.

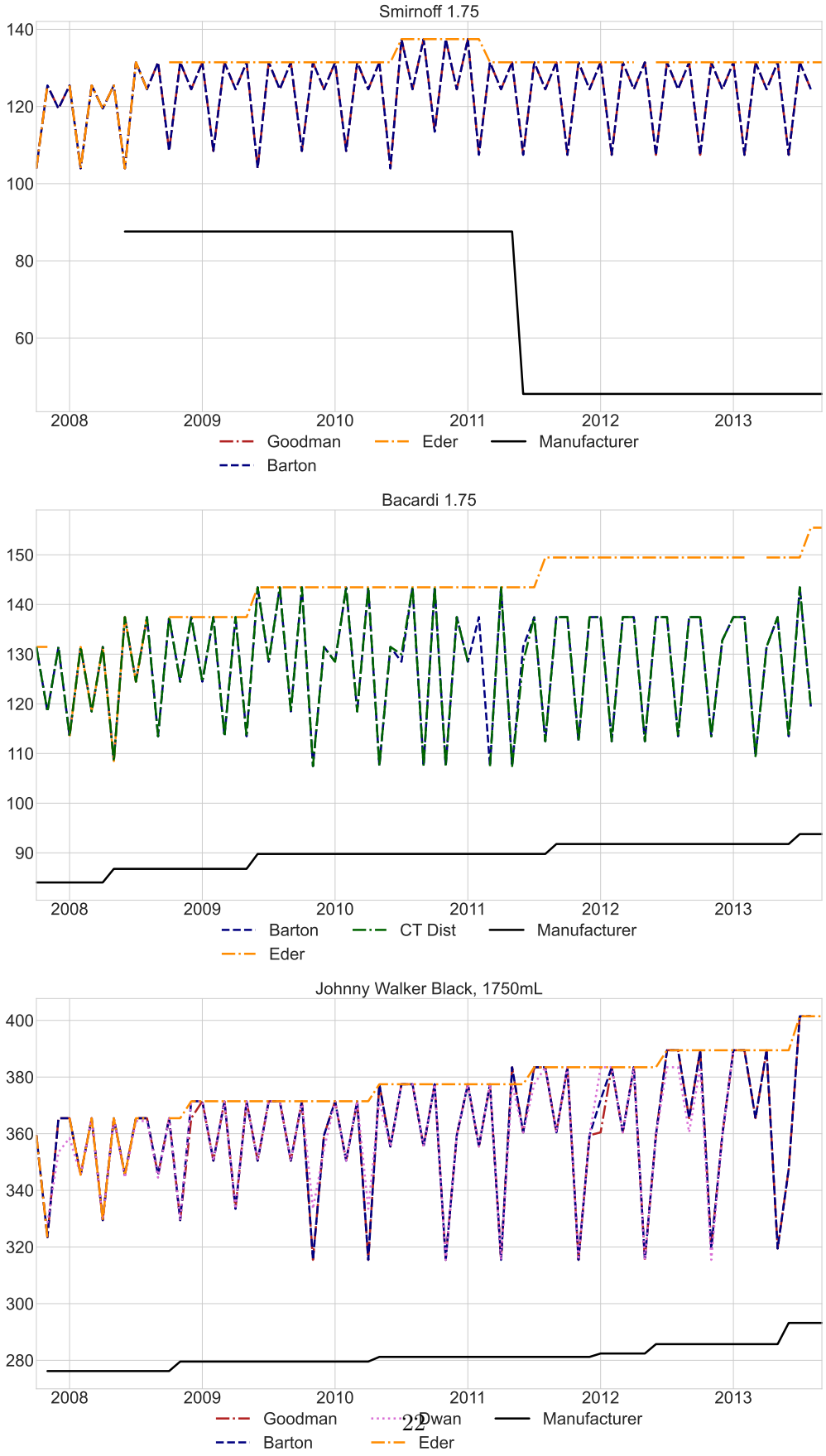

Figure 7: Co-movement of Wholesale Prices: Top Products

Figure 8: Co-movement of Wholesale Prices: Aggregated to Quarters

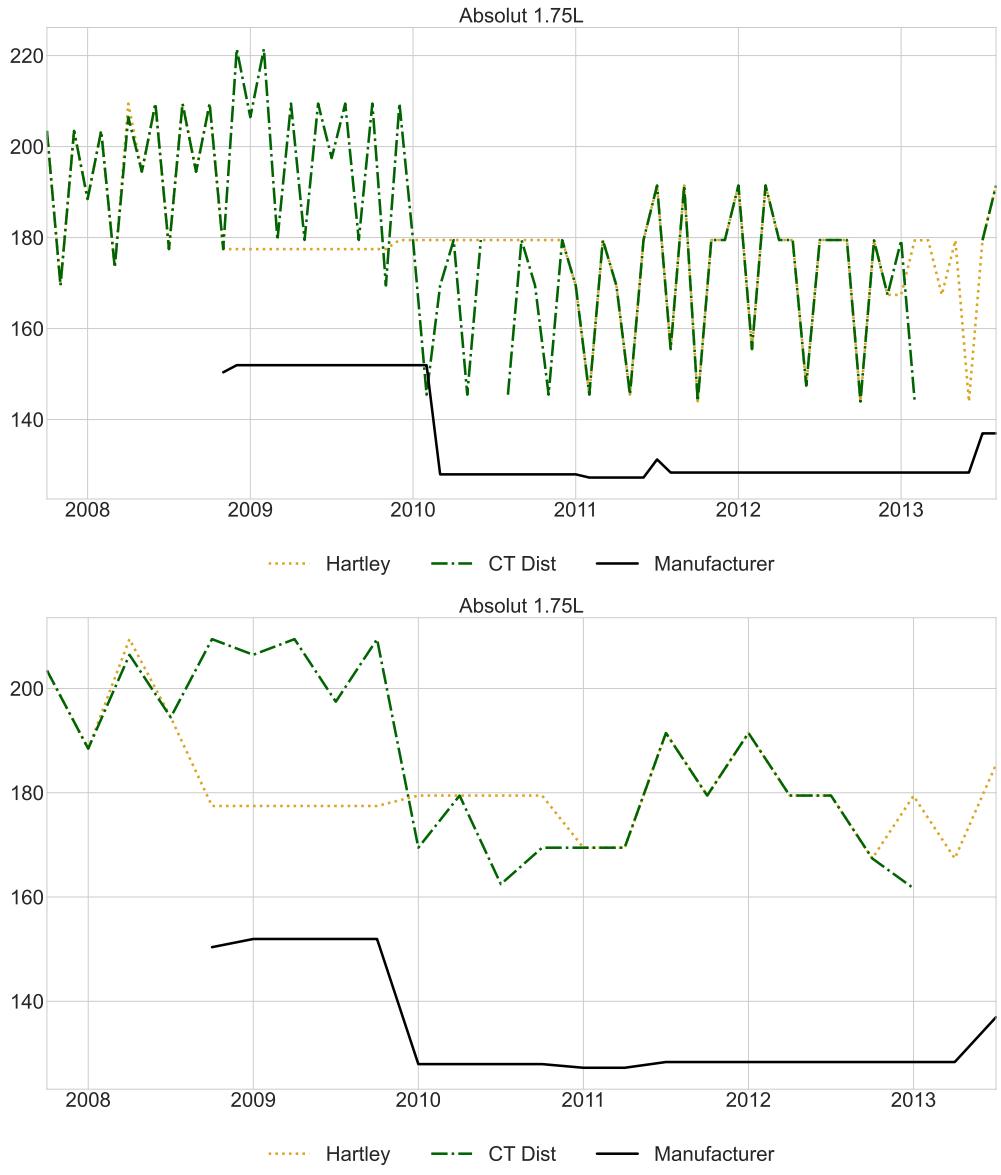

Figure 9: Co-movement of Wholesale Prices: Absolut Vodka

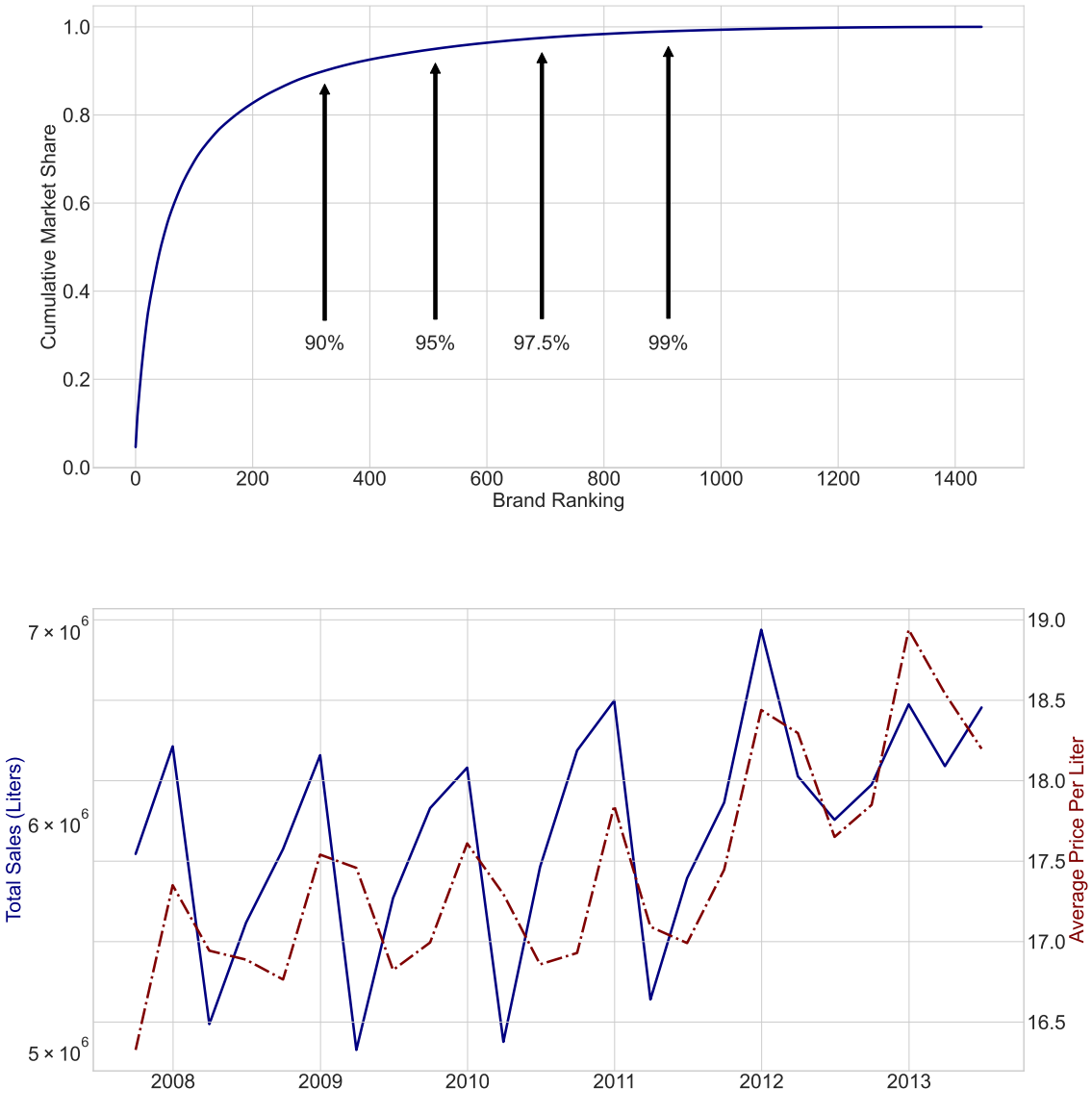

Figure 10: Average Price and Sales by Product Ranking

	# Obs	Share	Proof	% Flavored	Manufacturer Price	Manufacturer Margin	Wholesaler Price	Wholesaler Margin	Retailer Price	Retailer Margin
Gin	92	7.30	87.06	0.02	11.13	7.44	16.06	3.65	18.75	2.54
Rum	285	17.60	73.71	0.22	10.19	6.81	14.98	3.52	17.65	2.69
Tequila	199	5.10	79.99	0.00	15.46	10.17	22.35	5.62	29.14	5.00
Vodka	323	44.30	79.17	0.15	10.80	8.11	15.37	3.31	18.12	2.67
NA Whiskey	265	15.30	81.92	0.01	11.64	7.63	17.34	4.42	20.18	2.95
UK Whiskey	196	10.30	80.83	0.00	18.46	13.34	24.98	5.25	28.44	3.49
750mL	637	20.80	79.19	0.18	16.56	12.50	23.62	5.79	28.66	5.01
1L	355	23.30	79.32	0.12	13.79	10.04	19.86	4.80	24.80	4.38
1.75L	368	55.80	79.54	0.08	9.32	6.19	13.34	2.75	14.91	1.54
All	1360	100.00	79.42	0.11	11.87	8.40	17.00	3.86	19.97	2.89

Table 5: Product Summary by Size and Category

Source: Harmonized Price and Quantity Data
Prices are per Liter.
Federal Taxes: $2.85/L; State Taxes $1.19/L before July 1 2011; $1.43/L after.

	# Obs	Share	750mL	1L	1.75L	Manufacturer Price	Manufacturer Lerner	Wholesaler Price	Wholesaler Lerner	Retailer Price	Retailer Lerner
diageo	225	31.50	0.16	0.21	0.63	11.81	0.67	16.87	0.22	19.29	0.12
bacardi	75	13.70	0.21	0.34	0.45	14.32	0.69	19.88	0.22	22.67	0.12
pernod	108	13.40	0.21	0.34	0.46	15.29	0.76	20.82	0.20	24.29	0.14
jim beam	209	8.30	0.19	0.24	0.57	9.71	0.50	14.69	0.23	17.76	0.14
brown forman	67	5.10	0.23	0.30	0.47	14.91	0.71	22.30	0.27	26.03	0.14
imputed	206	4.50	0.39	0.09	0.51	11.87	0.63	17.21	0.22	21.23	0.19
skyy	29	2.70	0.29	0.06	0.65	11.51	0.73	16.50	0.21	19.16	0.13
constwines	7	2.60	0.19	0.11	0.71	7.43	0.65	12.38	0.29	14.37	0.14
constellation	65	2.10	0.08	0.13	0.79	5.13	0.21	8.35	0.22	10.18	0.15
star industries	17	2.10	0.13	0.29	0.58	4.67	0.37	7.86	0.24	9.51	0.17
imperial	27	2.00	0.19	0.12	0.70	5.74	0.53	9.19	0.23	12.17	0.24
mhw	60	1.90	0.44	0.15	0.40	12.31	0.65	17.75	0.23	21.98	0.19
black prince	8	1.90	0.10	0.28	0.61	3.98	0.30	5.95	0.11	7.17	0.17
white rock	7	1.20	0.25	0.00	0.75	7.08	0.64	10.51	0.20	13.49	0.22
william grant	23	1.20	0.24	0.12	0.65	10.52	0.45	16.06	0.24	18.95	0.13
remy cointreau	36	1.00	0.36	0.14	0.50	17.91	0.74	24.46	0.20	28.81	0.15
heaven hill	24	1.00	0.23	0.02	0.75	6.49	0.49	9.90	0.19	12.05	0.16
sazerac	50	0.70	0.37	0.22	0.41	10.23	0.48	14.94	0.20	19.36	0.22
usdist	6	0.70	0.23	0.00	0.77	7.01	0.65	9.46	0.11	14.50	0.33
moet hennessy	23	0.60	0.42	0.36	0.22	25.00	0.87	32.06	0.17	38.46	0.17
luxco	43	0.50	0.18	0.42	0.40	7.50	0.40	11.62	0.24	15.90	0.23
proximo	9	0.50	0.99	0.00	0.01	15.46	0.62	24.58	0.32	29.46	0.16
mccormick	22	0.20	0.08	0.46	0.46	5.83	0.26	8.36	0.17	12.73	0.23
ms walker	8	0.10	0.16	0.31	0.53	5.45	0.32	7.58	0.10	11.00	0.19
duggans	3	0.10	0.10	0.24	0.66	8.11	0.40	13.09	0.28	15.63	0.16
castle brands	3	0.00	0.91	0.00	0.09	18.93	0.78	26.78	0.22	34.91	0.22

Table 6: Manufacturer Summary

Source: Harmonized Price and Quantity Data
Prices are per Liter.
Federal Taxes: $2.85/L; State Taxes $1.19/L before July 1 2011; $1.43/L after.