Washington State Marijuana Traceability System Data Cleaning Appendix

August 2017

The retail sales prices reported to the tracking system may vary from the price paid by the consumer. In particular, during our sample period, many firms report prices that exclude one or more of the applicable taxes. We clean the retail prices by taking advantage of two characteristics of retail prices. First, publicly advertised prices (or 'list' prices) are nearly universally all tax-inclusive. Second, retailers nearly always choose to set prices in whole-dollar or (rarely) quarter-dollar increments.[1] We clean prices algorithmically, by finding, for each firm and month, the price multiplier within the range bounded by the size of the taxes, that, when applied to each retail transaction, results in the greatest fraction of whole-number prices. We take the modal multiplier for the months before and after July 1, 2015[2] as the true difference between the reported price and the price faced by the consumer. We verify these multipliers by collecting historical menu data through The Internet Archive and matching advertised prices to specific inventory lots and transactions. Though the Archive does not contain historical menus for every firm, our algorithm matches the manual process for each firm with available menus.

[1] We verified this through conversations with retailers as well as using historical menus available through The Internet Archive.

[2] There was a tax change on July 1, 2015, so the multipliers will generally be different before and after this date.

We clean a few additional price errors stemming from misplaced decimals (e.g. the price for a gram of marijuana in a given inventory lot is usually $9.50, but for one transaction it is $95 or $0.95.). To address this, if a given price was off from the modal price in an inventory lot[3] by approximately a factor of 10, we replaced the reported price with the modal price. This only affected several thousand individual transactions.

We then drop some remaining extreme outliers in the data. In particular, we drop all observations at the producer level if the number of plantings, harvestings, or days from plant-to-harvest are outside the 0.5th or 99.5th percentiles of their respective distributions. We drop all wholesale transactions with a usable weight above 2,500 grams[4] and all retail transactions if the usable weight was above 28.5 grams.[5] We also drop all wholesale or retail price per grams above $42.[6] We censor the THC content data if it is zero or above 40 in both the processor and retailer data.[7] We do not trim prices or weights at the low end of the distribution because there were no clear outliers at this end of the data.

Lastly, we drop some firms or firm-days in our data set. In particular, we drop the first 14 days in operation for all firms because this data tends to be very noisy. We also require for each firm that the first activity date (e.g. planting, harvesting, or sale transaction) occurs such that we have at least one month of data for each firm before the policy change we analyze. We also require that all producers have one planting or harvesting (chosen to match the outcome we are examining) in the two months prior to the policy change to be included in our data set. Similarly, we drop all processors that did not sell at least once in

[3]The modal price was calculated separately pre and post tax change for each weight amount in the inventory lot.

[4]This is about 0.025% of wholesale transactions.

[5]The maximum legal sale was one ounce. This step drops 0.15% of retail transactions.

[6]This is less than 0.03% of wholesale transactions and less than 0.04% of retail transactions.

[7]This affects 0.2% of wholesale transactions and 5% of retail transactions.

the two months before the policy change (either because they had not yet opened or because they took a long hiatus from selling any marijuana). A few retail firms open briefly, and then close for more than a month before re-opening for good. In these cases, we drop the first brief selling period and consider their first activity date the first date upon re-opening in our data.