Data Construction

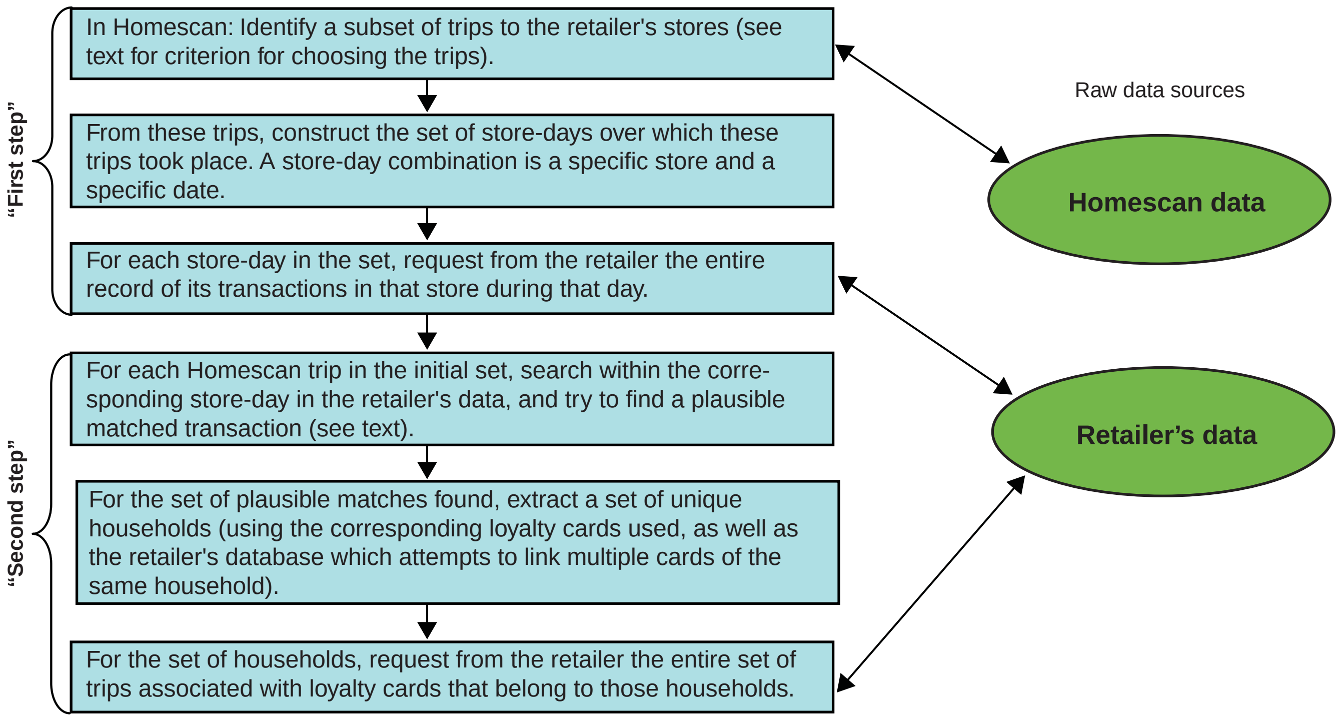
The main challenge in constructing the data was to match transactions and households, as recorded by Homescan, with corresponding transactions and loyalty card numbers, as recorded by the retailer. We matched records between the two data sets in two steps. We first obtained complete transaction level data from the retailer for stores and dates when a household in the Homescan data recorded a visit to the retailer store, and we developed a simple algorithm to match between the purchases recorded in the Homescan data and one of the many transactions recorded in the retailer's data (on that day at that store). We then asked the retailer for the full set of transactions recorded by the holders of the loyalty cards associated with these matched transactions. Figure 1 provides a schematic chart that sketches the key steps in the data construction process. Below we describe this process in more detail. Some readers may find it useful to skip these nitty-gritty details, and go directly to the end of the section, where we summarize the final data set we ended up using.

First Step

For the first step, the objective was to maximize the number of matched transactions given size limitations. These size limitations arise because, without additional information, we needed to have a complete transaction record from a particular store on a particular date for each potential matched transaction. The size of the data file containing this information was about three megabytes, and due to constraints imposed by the retailer, we had to limit this step to roughly 1,500 store-day transaction-level records.

Figure 1

Stages in the data construction process


"First step"
In Homescan: Identify a subset of trips to the retailer's stores (see text for criterion for choosing the trips).
From these trips, construct the set of store-days over which these trips took place. A store-day combination is a specific store and a specific date.
For each store-day in the set, request from the retailer the entire record of its transactions in that store during that day.
"Second step"
For each Homescan trip in the initial set, search within the corresponding store-day in the retailer's data, and try to find a plausible matched transaction (see text).
For the set of plausible matches found, extract a set of unique households (using the corresponding loyalty cards used, as well as the retailer's database which attempts to link multiple cards of the same household).
For the set of households, request from the retailer the entire set of trips associated with loyalty cards that belong to those households.
Raw data sources
Homescan data
Retailer's data

Source: Authors' calculations using Homescan and retailer data.

We therefore proceeded as follows. First, we restricted the data set to two metropolitan areas in which the retailer has high market share. This resulted in a total of 265 different retailer stores (147 in one area, and 118 in the other). The focus on two areas helped in obtaining more data, given the way the retailer organizes its data. Using areas with high market share of the retailer was also useful, as it could raise the probability that a single store-day record would help to match more than a single shopping trip. This would happen if two households in the Homescan panel visited the same store on the same day, which is more likely when the market share of the retailer is high. Since we identify the store by its ZIP code, we also restricted attention to retailer stores that are the only retailer stores in the same ZIP code. This eliminated 76 stores (29 percent), and left a total of 189 stores (101 in one area, 88 in the other).

We then searched the Homescan data for shopping trips at these stores, with the additional conditions that: (1) the trip includes purchase of at least five distinct UPCs (to make a match easier); (2) the trip occurred after February 15, 2004 (to guarantee that the retailer, who deletes transaction-level data older than 2 years, still had these data); and (3) the household shopped at the retailer's stores more than 20 percent and less than 80 percent of its trips, according to Homescan. These trips were made by 342 distinct households in the Homescan data. For 240 of those households, a single trip was randomly selected for each of them. For the remaining 102 households, which included households with at least 10, and not more than 20, reported trips in Homescan data, we selected all their trips. We then requested from the retailer the full transaction records for the store-days that matched those 1,779 trips. Since 74 of these trips were to the same store on the same date, we expected to get 1,705 store-day transaction-level records.

We eventually got 1,603 of those 1,705 requested store-days[5] (1,247 in the first area, 356 in the other). They accounted for 4,080,770 shopping trips and included 122 distinct stores (74 in the first area, and 48 in the other). The 1,603 store-days are associated with 1,675 trips from the sample of 1,779 shopping trips described above. However, as already mentioned, since the retailer enjoys high market share in both areas, it is not surprising that the 1,603 store-day transaction-level data records we obtained are associated with additional 904 trips in Homescan. Given the way we constructed the sample, however, many of these additional trips include a small number of items, or households that rarely shop at the retailer's stores.

Second Step

After obtaining the data from the first step, we developed a simple algorithm to find likely matches between transactions in the Homescan data with transactions in the retailer's data. These likely matches were only used to speed up the data construction process. The data were analyzed later using a more systematic matching procedure. The algorithm used the first five UPCs in the Homescan trip, and declared a match if at least three of these five were found in a given trip in the retailer's data. This algorithm was used with the data we obtained in the first step and found 1,372 likely matches that, according to the Homescan data, were associated with 293 distinct households. Of these

[5] The 102 store-days we did not get were missing at random, due to computer-related technical reasons at the retailer's end.

households, 166 were associated with more than one likely match, and 105 with four or more.

We then asked the retailer to use the loyalty card used in these 1,372 shopping trips and to provide us with all the transactions available for the households associated with these cards (in the retailer's data during the year 2004). Only two of the requested trips were not associated with loyalty cards. For the rest, we obtained all the transactions associated with the same loyalty card and additional transactions that were associated with loyalty cards used by the same household, as classified by the retailer. Since associating multiple cards with the same household did not necessarily match perfectly, the analysis used both the card level and the household level.

In this step, we obtained a total of 40,036 shopping trips from the retailer. These 40,036 trips were associated with 384 distinct stores (139 in the first area, 109 in the second, and 136 in other areas), with 682 distinct loyalty cards (472 in the first area, 203 in the second, and 7 in other areas), and with 529 distinct households, according to the retailer's definition (380 in the first area, 140 in the other).[6] Finally, the 40,036 trips were associated with 34,316 unique store-date-loyalty card combinations, 33,744 unique store-data-household combinations (using the retailer's definition of a household), and 27,746 unique store-date-household combinations, using the Homescan definition. Of these trips, 3,884 (9.7 percent) occurred in a store-day already appearing in the data we obtained earlier, and therefore were one of the 4,080,770 trips obtained in the first step. The 3,884 trips were associated with 3,514 unique store-date-loyalty card combinations, 3,477 unique store-date-household combinations, using the retailer definition, and 2,838 unique store-date-household combinations, using the Homescan definition. The algorithm used to request these data was geared to find likely matches, and therefore may have also found wrong matches. This is one reason that the number of households we intended to match (291, the original 293 minus two that had no associated loyalty cards) was less than the number of households associated with these trips. A second reason may be multiple cards used by the same household that are not linked to each other by the retailer.

Summary

To summarize, we have two different data sets from the retailer. The first data set included full transaction records of 1,603 distinct store-days. In these data, transactions were not associated with a loyalty card. The second data set included 40,036 transactions, which were associated with particular loyalty cards and households. 3,884 of these transactions overlap and appear in both data sets. The first data set was designed to match multiple transactions of 102 households in the Homescan data, and isolated transactions of other households. The second data set was designed to match all transactions of a few hundred households.

[6] Most households and card holders shopped in more than a single store. We associated a household or a card holder with the main store where either shopped. Card holders could be in other areas and appeared in the data only if they were associated with households that are in one of the two areas we focused on.