

README

Accessing the Data

All of the data used in this study were provided by Nielsen and accessed through the Kilts Center for Marketing at the University of Chicago Booth School of Business. Interested researchers may register to gain access to the RMS scanner data, the HMS panel data and the Ad Intel Data online at <https://www.chicagobooth.edu/research/kilts/datasets/nielsen>. Upon approval, Kilts will provide instructions on how to access each of these data sets. Our code will run on these data files from 2010-2014. For additional questions not addressed on the website, please contact marketingdata@chicagobooth.edu.

Advertising-Generalizability

Project Overview

The objective of this project is to estimate the effect of TV advertising on retail sales for a wide range of brands. To do that, we need the following data for each brand:

- Weekly volume, price, promotion, and feature/display at store or market level.
- Weekly advertising (GRP, duration, or spending) at market level.

We create the data we want in the following steps:

- (1.1) The ad occurrences and viewerships are separate in the raw Ad-Intel data. We need to merge them in order to find the GRP for each advertisement.
- (1.2) There are some discrepancies between the national and local records of Network TV ads. We need to resolve those discrepancies.
- (2) Ad-Intel and RMS use different brand definitions, so for each RMS brand, we need to find all the corresponding Ad-Intel brands.
- (3.1) RMS data comes in UPC-Store-Week level, so we need to aggregate it to Brand-Store-Week level.
- (3.2) Ad-Intel data comes in {AdIntel Brand}-Market-Channel-Second level, so we need to aggregate it to {RMS Brand}-Market-Week level.
- (4) We need to remove some small or problematic stores from the RMS data.

(1) Build Ad-Intel Data

General Concepts

Media Types

Ad-Intel covers 4 TV media types: Cable, Network, Syndicated, and Spot.

- For Cable TV, ads are purchased at national level. Impressions are observed only nationally.
- For Network and Syndicated TV, ads are purchased at national level. The programs are broadcast at local TV stations. Impressions are observable locally through the Network Clearance and Syndicated Clearance impressions files.
 - The local TV stations are typically affiliated to a national network. For example, WBZ is the Boston affiliate of CBS.
- For Spot TV, ads are purchased at DMA level. The programs are also broadcast at local TV stations.

Since Network and Syndicated TV ads are purchased nationally but broadcast locally, the Ad-Intel record them in two ways:

- The Network TV and Syndicated TV occurrence files record them at national level.
 - i.e. the date and time each ad is *supposed* to be broadcast at every local station
- The Network Clearance Spot TV and Syndicated Clearance Spot TV occurrence files record them at local channel level.
 - i.e. the date and time each ad is *actually* broadcast at every local station
- It is important to emphasize that the Network (Syndicated) TV occurrences and the Network (Syndicated) Clearance Spot TV occurrences are meant to represent the exact same ad occurrences. The main distinction here is that the “clearance” occurrences are measured at the local DMA level, allowing for us to match them with local impressions data. They do not represent two distinct forms of advertising. For our purposes, the value derived from the Network (Syndicated) TV occurrences (measured nationally) is (1) to measure the estimated cost of the ad, which is purchased nationally, and (2) to audit the advertising occurrence schedule in the “clearance” data to detect times when a local station might override a nationally purchased ad and distinguish it from local measurement error.
- To further explain the previous point, The local channels have some authority to replace or move nationally scheduled ads, and it possible that the Nielsen local measurement devices data are also not perfect. Hence there are discrepancies between those national and local files not every “Network TV” advertising occurrence will match perfectly with a “Network Clearance Spot TV” occurrence, even though they represent the same ad buys. Any incidence of “Network TV” occurrences which do not have associated “Network Clearance Spot TV” occurrences in a local market will be referred to as the “Missing Network Discrepancy.”

Occurrence Data

The occurrence data provides detailed information for each advertisement, including:

- Date [AdDate]
- Time [AdTime]
- Media Type [MediaTypeID]
- Channel [DistributorCode, DistributorID]

- Market (can be national) [MarketCode]
- Primary, Secondary, and Tertiary Brands [PrimBrandCode, ScndBrandCode, TerBrandCode]
- Duration [Duration]
- The associated TV program [NielsenProgramCode, TelecastNumber]
- Other time-related info [TVDayPartCode, DayOfWeek, TimeIntervalNumber]

Note that Ad-Intel does not capture any local ad between 2AM and 5AM.

Impression (Viewership) Data

For the national media types (Cable, Network, and Syndicated), Ad Intel provides the estimated number of impressions for each TV program—defined as a pair of NielsenProgramCode and TelecastNumber.

For the local media types (Network Clearance, Syndicated Clearance, and Spot), Ad Intel provides the estimated impressions at {Local Station}-Month-{Day of Week}-{5 Minute Time Interval} level.

Note: There are only 25 markets (the “Local People Meter” markets) for which the local impressions are available in all months. In those markets, impressions are measured using set top boxes. For the rest of the markets, local impressions data are only available in four “sweeps months”: February, May, July, and November and are measured by Nielsen households filling out diaries. Therefore, we need to impute the impressions for the non-sweeps months in non-LPM markets. Now we use an average between the two closest available months, weighted by the time difference. For example, for June we use 1/2 May + 1/2 July, and for March we use 2/3 February + 1/3 May.

Universe Estimates

Ad-Intel also provides the estimated total number of TV audience at national and market level. Those universe estimates are updated yearly.

Build the Regular Parts

The logic of the regular build is very simple. For each media type in each month, we need to do the following:

1. Merge occurrences with impressions
 - (a) For national data (Cable TV), merge on NielsenProgramCode and TelecastNumber
 - (b) For local data (Spot TV, Network Clearance Spot TV, Syndicated Clearance Spot TV), merge on DistributorID, DayOfWeek, and TimeIntervalNumber
 - (c) Remember to do the imputation for non-LPM markets in non-sweep months.
2. Merge the result with universe estimates
3. Calculate the GRP as $100 * \text{Impression} / \text{Universe}$ for each ad occurrence

The script Build-Regular.R follows exactly those steps.

Resolve the “Missing Network” Discrepancy

As mentioned previously, each Network ad that is actually shown on TV in a local market should be associated with a Network Clearance Spot TV ad. These represent the exact same showing of the exact same ads. For the most part,

ads purchased as Network TV ads are shown “simultaneously” (in the same slot during the same television program) across the country. However, in rare circumstances, a Network TV ad may not actually run in one or more local markets. Additionally, in even rarer circumstances, a Network TV ad may run in a local market but not get recorded in the Network Clearance occurrences due to an error in the recording devices at the local level. Our goal in this part of the build is to distinguish the explanation when Network TV and Network Clearance Spot TV ads do not match in a given market at a given time.

We do this by reconstructing the TV schedule, down to the second. From the observed schedule, we infer whether or not it was a recording error by the local measurement device or whether the local station likely displaced the Network ad. If, for example, there is a “gap” in the schedule that is 30 seconds long, a 30 second long Network TV ad that purportedly ran in that time interval, we infer that the Network TV ad actually ran on the local station in that slot but the recording device failed in that moment. In these cases, we “insert” the Network TV ad occurrences into the advertising schedule into the “gap” where we inferred that it belonged. Alternatively, if there is no “gap” in the schedule (i.e. we observe ads back-to-back every second followed by programming), we then infer that the local station displaced the Network ad either for extended programming (e.g., local news alert, sporting event gone long) or for an additional locally purchased Spot TV ad. Network Clearance TV ads also are always missing between 2AM and 5AM, as local ratings are not recording during those time intervals.

The procedure of this part, however, is **much more complicated**. Basically, we need the following steps:

1. Find the information for each local station, including:
 - The market (MarketCode) and network (Affiliation) for each local station (DistributorCode).
 - The DistributorID for each DistributorCode.
 - This is in fact an one-to-one relationship, but we have to record that because the “Station Affiliation” data only has DistributorCode, while the impressions data only has DistributorID.
 - *Corresponding scripts:* 1-Extract-Local-Distributors.R and 2-Process-Station-Information.R
2. For each network and each local station, we stack all the monthly data.
 - We cannot use the raw monthly data because the national and local files have different dates.
 - Stacking also prevents errors at month boundaries. For example, a national ad at 2012/05/31 23:30:00 may be distributed locally at 2012/06/01 00:30:00. This will not be captured if we process the data month-by-month.
 - *Corresponding scripts:* 3-Extract-Station-Level-Data.R
3. For each local station, we find the “unexpectedly missing” occurrences. Please see the relevant code for details, but in short we categorize all the national ads as following:
 - A national ad is **directly matched** to the local data if its closest local occurrence has the same primary brand code.
 - A national ad is **indirectly matched** to the local data if there’s a local occurrence that is aired within some time limit before or after the scheduled air-time.
 - This step accounts for the ads that are moved around. The time limit is 3 hours for ETZ/CTZ, 6 hours for MTZ, and 7 hours for PTZ.
 - A national ad is **replaced by another ad** if another spot / network clearance / syndicated clearance ad runs into its scheduled time slot.
 - A national ad is **not captured locally** if its scheduled air-time is between 2AM and 5AM.
 - We mark all remaining national ads as **unexpectedly missing** at this local station. These are the “gaps” described above.
 - *Corresponding scripts:* 4-Categorize-Stations.R and 5-Run-Station-Level.R
4. We get all the “unexpectedly missing” occurrences at each station, and we re-organize them into monthly files. We then merge those monthly files with the monthly local impressions data.
 - *Corresponding scripts:* 6-Station-Level-To-Month-Level.R and 7-Merge-Viewership-On-Monthly-File

An important note: We must be careful to account for the “broadcast delay” for mountain and pacific time zones.

- A nationally scheduled program or ad can be broadcast with a delay of 0/1/2/3 hours in pacific-time markets or 0/1 hours in mountain-time markets. This delay can be seemingly arbitrary.
- In step 3, we say a national ad is “unexpectedly missing” only if it’s “unexpectedly missing” under *all the possible delays*, i.e. 0/1 hour in MTZ and 0/1/2/3 hours in PTZ.
- In step 4, for PTZ/MTZ markets we average the impressions at the airtime and 3/1 hours after the airtime.

(2) String Matching and RMS Module/Flavor Aggregation

General Concepts

String matching means matching the brands in RMS and Ad-Intel. We classify the matches in 4 “tiers” below.

Please see Shapiro, Hitsch & Tuchman (2020) for more details (“Research Note: Procedure for Building Nielsen Ad Intel Data and Merging with Nielsen RMS Scanner Data”)

In theory, tier-1 and tier-2 advertising should have a positive effect on sales, while the effect of tier-3 and tier-4 ads can be either positive or negative.

1. RMS and Ad-Intel brand names are identical.
2. Ad-Intel brand is more specific than the RMS brand.
 - Example: Ad-Intel brand LAYS POTATO CHIPS CHICKEN AND WAFFLE is a tier-2 match to RMS brand LAY 'S.
3. Ad-Intel brand is more general than the RMS brand.
 - Example: Ad-Intel brand COCA-COLA SOFT DRINKS is a tier-3 match to RMS brand COCA-COLA R.
4. Ad-Intel brand is an “associate” to the RMS brand.
 - Example: Ad-Intel brand COCA-COLA ZERO DT is a tier-4 match to RMS brand COCA-COLA R.

Module aggregation means aggregating some too specific RMS modules together. For example, the RMS modules NUTS-BAGS, NUTS-CANS, NUTS-JARS, and NUTS-UNSHELLED are essentially the same thing, and advertisements never distinguish between them.

Flavor aggregation means aggregating many sub-brands together. For example, the brand “Lean Cuisine Frozen Entree” has 50 sub-brands in RMS (e.g. LEAN CUISINE ONE DISH FAVORITE or LEAN CUISINE SPA COLLECTION). Aggregating them together makes the matching easier, and it creates more tier-3 matches and fewer tiers-2/4 matches.

Folder Structure

Exported-Files/

Universe-For-Estimation-180508.csv: Detailed information about our final universe of 358 brands.

Consolidated-UPC-Dictionary-For-Aggregation-180508.rds: A revised version of the original Products meta-data that reflects all module and flavor aggregations.

All-Main-Matches-180508.csv: String matches for the 358 estimation brands.

All-Comp-Matches-180508.csv: String matches for the top competitors not included in the main universe.

Information-Files/ and Meta-Data/

Information-Files/: Contains intermediate files for brand matching and flavor aggregation.

Meta-Data/: Meta data from RMS data build and version correction.

(3) Brand Aggregation

Ad-Intel

The Ad-Intel data build comes at {AdIntel Brand}-Channel-Time level, and in the end we want to aggregate it to {RMS Brand}-Market-Week level.

- To make things more flexible, in this section we only aggregate the ad data to {AdIntel Brand}-{Media Type}-Market-Week level and save the data for each RMS brand. The further aggregation is done on-the-fly in section 6.
- The aggregation here only involves adding up Duration and GRP.
 - Some ad occurrences come with 2/3 brands, but those brands are mostly the same product (e.g. Snapple Black Tea and Snapple Green Tea, which we eventually combine to a single “brand” as per above). To avoid double-counting the ads, we use the following trick: if an occurrence has two/three brands, we treat it as two/three occurrences with half/one-third of the Duration and GRP.

The corresponding script is 1-Ad-Intel.R.

RMS

The RMS data build comes at UPC-Store-Week level, and we want to aggregate it to Brand-Store-Week level. As mentioned before, UPCs are generally much more specific than brands, as they reflect many different sizes and presentations..

- One RMS brand may contain hundreds of UPCs with different sizes (size1_amount, say 12 OZ or 24 OZ) and different multi-pack status (multi, say 6-pack or 12-pack).
 - Therefore, instead of using the units field in RMS data, we need calculate the volume in equivalency units: $\text{volume} := \text{units} * \text{multi} * \text{size1_amount}$. We adjust price accordingly.
- For each store-week, the brand-level variables are calculated as follows:
 - Volume: sum of UPC-level volumes
 - Price: weighted average of UPC-level prices. The weight for a UPC is its average weekly revenue in this store.
 - Promotion: weighted average of UPC-level promotion indicators ($\text{price} / \text{base_price} < 0.95$).
 - Feature/Display: weighted average of UPC-level feature/display indicators (remove missing values).

Corresponding scripts:

- 2-RMS-Extract-Data-From-All-Other.R extracts small UPCs that are not stored individually in the RMS data build.
- 3-RMS-Run-Brand-Aggregation.R does the aggregation. It uses a pseudo-parallelization trick so that it can be submitted in parallel to different nodes.
- 4-RMS-Job-Submission.sh submits script No.3 to different nodes on the Grid.

(4) Select Stores and Panelists

This is a housekeeping step that deals with the inconsistencies between various location-related files. Note that `Panelist.RData` comes from Homescan data build in `0-Build-Nielsen-Data/`. The exported files are `Processed-Data/Selected-Stores.RData` and `Processed-Data/Selected-Panelists.RData`.

Remove Shifting FIPS

For the implementation of the border strategy, we use the Nielsen provided mapping between counties and DMAs. From this, we constructed a data set that flags the counties that lie on a border between DMAs. However, some counties change DMAs over time, since the borders are re-drawn periodically. Therefore, we removed all the counties that did not stay in a single DMA, and we removed the borders that were re-drawn.

Select Usable Panelists

We removed the panelists in the HMS data that switch between different FIPS-codes.

Select Usable Stores

We removed the stores in the RMS data that switch between different FIPS-codes, or stores that are not continuously covered in 2010-2014. We then rank the stores by the total 2010-2014 revenue (across all products), and find the stores that constitute 90% of total revenue. We will use those stores for all estimations.

(5) Basic Descriptions

Screening Based on Tiers 1+2 Advertising

- Scripts: `Grid-Scripts/Ad-Intel-Screening-and-Share-Analysis.R` and `Local-Scripts/Inspect-Shares.R`
- Results: `Processed-Data/Screening.rds`
- Description: For each of the 358 RMS brands in our universe, calculate the fraction of market-weeks with positive tiers 1+2 GRP, and the mean tiers 1+2 GRP conditional on it being positive.
 - We kick out 70 brands whose “positive fraction” is below 5%, or whose “positive mean” is below 10 points.

Share of Advertising by Media Type

- Scripts: `Grid-Scripts/Ad-Intel-Screening-and-Share-Analysis.R` and `Local-Scripts/Inspect-Shares.R`
- Results: `Processed-Data/Shares-By-Media-Type.rds`
- Description: For each brand, we calculate the share of GRP/Duration in each media type.

R2 of Advertising Against Fixed Effects

- Scripts: `Grid-Scripts/Ad-Intel-R2-Analysis.R` and `Local-Scripts/Inspect-R2.R`
- Results: `Processed-Data/Advertising-R2.rds`

- Description: For each brand, regress GRP/Duration on fixed effects and get the R2. The rationale is to see how much variation we are left with in the regressor after controlling for the FEs. The list of FEs is:
 - Store + Quarter/Month + Week of Year
 - Store + Border-[Quarter/Month] + Week of Year
 - DMA + Quarter/Month + Week of Year
 - Border-DMA + Border-[Quarter/Month] + Week of Year
 - Store + Week
 - Store + Border-Week
 - DMA + Week
 - Border-DMA + Border-Week

(6) Basic Estimations

Fixed Configurations

- Advertising Control: Own Tiers 1+2, Own Tiers 3+4, and Top Competitor (all tiers)
- Number of Lags: 52
- Stores: those constituting top 90% of total revenue

Variable Configurations

1. Data Level: Store / Market
2. Strategy: Border / Regular
3. Own Feature-Display Control: Yes / No
4. Price-Promotion Control:
 - None
 - Own brand only
 - Up to 1 competitor (subject to the 80% revenue share limit)
 - Up to 2 competitors
 - Up to 3 competitors
5. Time FE: Quarter + Week-of-Year / Month + Week-of-Year
6. Advertising Measure: GRP / Duration
7. Goodwill Formula: Log of Sum / Sum of Log
8. Delta: $\text{seq}(0, 1, 0.01)$

Therefore, we need to run up to $2*2*2*5*2*2*2*101 = 32320$ regressions for each brand, and that amounts to 9308160 regressions across all 288 brands. Note that we decided to try all the possible price/promo controls and impose the maximum-observation-loss restrictions *ex post*.

Basic Steps for Running A Regression

For a given brand, fixing parts (1), (2), and (3) above, we do the following:

1. Load RMS data, merge in DMA and border information, and do the proper aggregation.
2. Load Ad-Intel data and do the proper aggregation.
3. Loop through $\text{delta} = \text{seq}(0, 1, 0.01)$ (at most 40 regressions for each delta):
 - Calculate goodwill (4 possible ways: log-sum GRP, log-sum Duration, sum-log GRP, sum-log Duration)

- Loop through 5 kinds of price-promo controls:
 - Run the regression with quarter or month fixed effects.

(7) Semi-parametric Model

Estimating Advertising Elasticities Using Semi-parametric Estimation

- Scripts: 1-Baseline-Residualized.R and 2-Baseline-Ad-Elasticities.R
- Source: Regression-Functions-Residualized.R
- Results: Processed-Data/Elasticities.rds
- Description: Refer to Appendix E. Flexible functional form for advertising response.

Predicting Quantities from Semi-parametric Model

- Scripts: 3-Predict-Q.R and 4-Collect-Predicted-Q.R
- Results: Processed-Data/cvspline_knots.rds and Processed-Data/cvspline_Q_hat.rds
- Description: Refer to Appendix E. Flexible functional form for advertising response.

(8) ROI Analyses

Advertising Cost

- Scripts: Ad-Cost/0-Ad-Cost.R
- Source: Ad-Cost/Ad-Cost.R
- Results: Processed-Data/Elasticities.rds
- Description: Refer to Data-Build-Document.pdf, section 2 Advertising Cost.

Estimating Weekly Advertising ROI

ROI calculation

- Scripts: Weekly/1-Weekly-ROI.R and Weekly/2-Weekly-ROI-se.R
- Source: Weekly/Weekly-ROI.R and Weekly/Weekly-ROI-se.R
- Results: Weekly/Processed-Data/base-average-ROI-observed.RData and Weekly/Processed-Data/base-
- Description: Refer to Appendix G. ROI calculation.

Break-even ad effects

- Scripts: Weekly/3-Weekly-ROI-breakeven.R
- Source: Weekly/Weekly-ROI-breakeven.R
- Results: Weekly/Processed-Data/average-breakeven-observed.RData
- Description: Refer to Appendix G.3. Break-even ad effects.

Estimating Overall Advertising ROI

ROI calculation

- Scripts: Overall/1-Overall-ROI.R and Overall/2-Overall-ROI-se.R
- Source: Overall/Overall-ROI.R and Overall/Overall-ROI-se.R
- Results: Weekly/Processed-Data/base-avg-ROI-overall-observed.RData and Weekly/Processed-Data/b
- Description: Refer to Appendix G. ROI calculation.

Break-even ad effects

- Scripts: Overall/3-Overall-ROI-breakeven.R
- Source: Overall/Overall-ROI-breakeven.R
- Results: Weekly/Processed-Data/avg-overall-breakeven-observed.RData
- Description: Refer to Appendix G.3. Break-even ad effects.

(9) Additional Notes

An intuitive guide to building the data is provided in Online Appendix D of the paper. For your convenience, it is also duplicated below.

References

Shapiro, Bradley T., Günter J. Hitsch, and Anna E. Tuchman. 2021. "TV Advertising Effectiveness and Profitability: Generalizable Results from 288 Brands." *Manuscript*.

A Data construction

This appendix details exactly how we construct our data. We need the following data for each brand:

- Weekly volume, price, promotion, and feature/display at store or market level.
- Weekly advertising (GRP, duration, or spending) at television market (DMA) level.

We construct the data using the following steps:

1. Build Ad Intel Data
 - (a) The ad occurrences and viewerships are separate in the raw Ad Intel data. We need to merge them in order to find the GRPs for each advertisement.
 - (b) There are some discrepancies between the national and local records of Network TV ads. We need to resolve those discrepancies.
2. Create brand map between Ad Intel and RMS data sets.
 - (a) Ad Intel and RMS use different brand definitions, so for each RMS brand, we need to find all the corresponding Ad Intel brands.
3. Aggregate Data
 - (a) RMS data come in UPC-Store-Week level, so we need to aggregate it to Brand-Store-Week level.
 - (b) Ad Intel data come in {Ad Intel Brand}-Market-Channel-Second level, so we need to aggregate it to {RMS Brand}-Market-Week level.
4. Identify RMS stores to be used in estimation
5. Identify products to be used in estimation

Each of these steps is described in more detail below.

A.1 Build Ad Intel Data

A.1.1 General Concepts

Media Types Ad Intel covers 4 TV media types: Cable, Network, Syndicated, and Spot.

- For Cable TV, ads are purchased at a national level.
- For Network and Syndicated TV, ads are purchased at a national level. The programs are broadcast at local TV stations.
 - The local TV stations are typically affiliated to a national network. For example, WBZ is the Boston affiliate of CBS.
- For Spot TV, ads are purchased at the DMA level. The programs are also broadcast at local TV stations.

Since Network and Syndicated TV ads are purchased nationally but broadcast locally, the Ad Intel data record them in two ways:

- The Network TV and Syndicated TV occurrence files record them at a national level.
 - i.e., the date and time each ad is supposed to be broadcast at every local station
- The Network Clearance Spot TV and Syndicated Clearance Spot TV occurrence files record them at a local channel level.
 - i.e., the date and time each ad is actually broadcast at every local station
- It is important to emphasize that the Network (Syndicated) TV occurrences and the Network (Syndicated) Clearance Spot TV occurrences are meant to represent the exact same ad occurrences. The main distinction here is that the “clearance” occurrences are measured at the local DMA level, allowing us to match them with local impressions data. They do not represent two distinct forms of advertising. For our purposes, the Network (Syndicated) TV occurrences (measured nationally) allow us to (1) measure the estimated cost of the ad, which is purchased nationally, and (2) audit the advertising occurrence schedule in the “clearance” data to detect times when a local station might override a nationally purchased ad and distinguish it from local measurement error.
- To further explain the previous point, local channels have some authority to replace or move nationally scheduled ads, and it is possible that the Nielsen local measurement devices are not perfect. Hence, not every “Network TV” advertising occurrence will match perfectly with a “Network Clearance Spot TV” occurrence, even though they represent the same ad buys. Any incidence of “Network TV” occurrences that do not have associated “Network Clearance Spot TV” occurrences in a local market will be referred to as the “Missing Network Discrepancy.”

Occurrence Data The occurrence data provide detailed information for each advertisement, including:

- Date [AdDate]
- Time [AdTime]
 - Note that Ad Intel does not capture any local ads between 2AM and 5AM.
- Media Type [MediaTypeID]
- Channel [DistributorCode, DistributorID]
- Market (can be national) [MarketCode]
- Primary, Secondary, and Tertiary Brands [PrimBrandCode, ScndBrandCode, TerBrandCode]
- Duration [Duration]
- The associated TV program [NielsenProgramCode, TelecastNumber]
- Other time-related info [TVDayPartCode, DayOfWeek, TimeIntervalNumber]

Impression (Viewership) Data For the national media types (Cable, Network, and Syndicated), Ad Intel provides the estimated number of impressions for each TV program, defined as a pair of NielsenProgramCode and TelecastNumber.

For the local media types (Network Clearance, Syndicated Clearance, and Spot), Ad Intel provides the estimated impressions at the {Local Station}-Month-{Day of Week}-{5 Minute Time Interval} level.

There are only 25 markets (the “Local People Meter” markets) for which the local impressions are available in all months. In those markets, impressions are measured using set top boxes. For the rest of the markets, local impressions data are only available in four “sweeps” months: February, May, July, and November and are measured by Nielsen households filling out diaries. Therefore, we need to impute the impressions for the non-sweeps months in non-LPM markets. We use an average between the two closest available months, weighted by the time difference. For example, for June we use $1/2$ May + $1/2$ July, and for March we use $2/3$ February + $1/3$ May.

Universe Estimates Ad Intel also provides the estimated total number of TV households at the national and market level. Those universe estimates are updated yearly.

A.1.2 Build the Regular Parts

The logic of the regular build is very simple. For each media type in each month, we need to do the following:

1. Merge occurrences with impressions
 - (a) For national data (Cable TV), merge on NielsenProgramCode and Telecast-Number
 - (b) For local data (Spot TV, Network Clearance Spot TV, Syndicated Clearance Spot TV), merge on DistributorID, DayOfWeek, and TimeIntervalNumber
 - (c) Do the imputation for non-LPM markets in non-sweep months.
2. Merge the result with universe estimates
3. Calculate the GRP as $100 * \text{Impression} / \text{Universe}$ for each ad occurrence

A.1.3 Resolve the "Missing Network" Discrepancy

As mentioned previously, each Network ad that is actually shown on TV in a local market should be associated with a Network Clearance Spot TV ad. These represent the exact same showing of the exact same ads. For the most part, ads purchased as Network TV ads are shown “simultaneously” (in the same slot during the same television program) across the country. However, in rare circumstances, a Network TV ad may not actually run in one or more local markets. Additionally, in even rarer circumstances, a Network TV ad may run in a local market but not get recorded in the Network Clearance occurrences due to an error in the recording devices at the local level. Our goal in this part of the build is to distinguish the explanation when Network TV and Network Clearance Spot TV ads do not match in a given market at a given time.

We do this by reconstructing the TV schedule, down to the second. From the observed schedule, we infer whether or not it was a recording error by the local measurement device or whether the local station likely displaced the Network ad. If, for example, there is a “gap” in the schedule that is 30 seconds long and a 30 second long Network TV ad purportedly ran in that time interval, we infer that the Network TV ad actually ran on the local station in that slot but the recording device failed in that moment. In these cases, we “insert” the Network TV ad occurrence into the advertising schedule in the “gap” where we inferred that it belonged. Alternatively, if there is no “gap” in the schedule (i.e. we observe ads back-to-back every second followed by programming), we then infer that the local station displaced the Network ad either for extended programming (e.g., local

news alert, sporting event gone long) or for an additional locally purchased Spot TV ad. Network Clearance TV ads also are always missing between 2AM and 5AM, as local ratings are not recorded during those time intervals.

While in principle this exercise is intuitively simple, in practice, this procedure is complicated to implement. We take the following steps:

1. Find the information for each local station, including:
 - (a) The market (MarketCode) and network (Affiliation) for each local station (DistributorCode).
 - (b) The DistributorID for each DistributorCode.
 - i. This is in fact a one-to-one relationship, but we have to record that because the “Station Affiliation” data only have DistributorCode, while the impressions data only have DistributorID.
2. For each network and each local station, we stack all the monthly data.
 - (a) We cannot use the raw monthly data because the national and local files have different dates.
 - (b) Stacking also prevents errors at month boundaries. For example, a national ad at 2012/05/31 23:30:00 may be distributed locally at 2012/06/01 00:30:00. This will not be captured if we process the data month-by-month.
3. For each local station, we find the “unexpectedly missing” occurrences. In short, we categorize all the national ads as following:
 - (a) A national ad is directly matched to the local data if its closest local occurrence has the same primary brand code.
 - (b) A national ad is indirectly matched to the local data if there is a local occurrence that is aired within some time limit before or after the scheduled air-time. This step accounts for the ads that are moved around. The time limit is 3 hours for ETZ/CTZ, 6 hours for MTZ, and 7 hours for PTZ.
 - (c) A national ad is replaced by another ad if another spot / network clearance / syndicated clearance ad runs into its scheduled time slot.
 - (d) A national ad is not captured locally if its scheduled air-time is between 2AM and 5AM.
 - (e) We mark all remaining national ads as unexpectedly missing at this local station. These are the “gaps” described above.

4. We get all the “unexpectedly missing” occurrences at each station, and we reorganize them into monthly files. We then merge those monthly files with the monthly local impressions data.

Note: We must be careful to account for the “broadcast delay” for mountain and pacific time zones.

- A nationally scheduled program or ad can be broadcast with a delay of 0/1/2/3 hours in pacific-time markets or 0/1 hours in mountain-time markets. This delay can be seemingly arbitrary.
- In step 3, we say a national ad is “unexpectedly missing” only if it is “unexpectedly missing” under all the possible delays, i.e. 0/1 hour in MTZ and 0/1/2/3 hours in PTZ.
- In step 4, for PTZ/MTZ markets we average the impressions at the airtime and 3/1 hour(s) after the airtime.

A.2 Create Brand Map between RMS and Ad Intel

We merge the advertising and sales data sets at the store-brand-week level. This merge is non-trivial and non-obvious. In particular, the brand variables in the Ad Intel and RMS data sets are not always specified at the same level. For example, UPCs in the RMS data are sometimes much more specific than the generic brands in the Ad Intel data. Our matching procedure results in three distinct “types” of advertising variables to be used in our models. First, we specify advertising that directly corresponds to the RMS product in question. Second, we create a variable that captures advertising for affiliated brands, including potential substitutes, that may affect the focal RMS product. Third, we include advertising for the top competitor. For example, for the Diet Coke brand, own advertising includes ads for Diet Coke, whereas affiliated advertising includes advertising for Coca-Cola soft drinks, Coke Zero, Coca-Cola Classic, and Cherry Coke. Furthermore, we include advertising for Diet Pepsi, the top competitor of Diet Coke.

The procedure for creating this match is as follows. First, we create a map between the brands in the RMS and Ad Intel data sets using an automated string matching procedure. Second, the three authors and two research assistants hand-audited the matches to classify them into tiers that are associated with the above description. Finally, any brands for which the authors could not come to an agreement on classification of matches were thrown out as “failed matches” and not included in the sample. We classify the matches in 4 “tiers,” which are described below. In theory, tier-1 and tier-2 advertising should have

a positive effect on sales, while the effect of tier-3 and tier-4 ads can be either positive or negative. Based on this logic, we construct our measure of own advertising by grouping tier-1 and tier-2 advertising together and our measure of “affiliated brands” advertising by grouping tier-3 and tier-4 advertising together.

Own Advertising

- Tier 1: RMS and Ad Intel brand names are exact matches.
- Tier 2: Ad Intel brand is more specific than the RMS brand.
 - Example: Ad Intel brand LAYS POTATO CHIPS CHICKEN AND WAFFLE is a tier-2 match to RMS brand LAY’S.
- For the median brand in our data, tier-1 matches make up 43% of own advertising GRPs, while tier-2 matches make up the remaining 57% of own advertising GRPs. The distribution of GRPs coming from identical matches is shown in Table 1.

Affiliated Brands Advertising

- Tier 3: Ad Intel brand is more general than the RMS brand.
 - Example: Ad Intel brand COCA-COLA SOFT DRINKS is a tier-3 match to RMS brand COCA-COLA R.
- Tier 4: Ad Intel brand is an “associate” to the RMS brand.
 - Example: Ad Intel brand COCA-COLA ZERO DT is a tier-4 match to RMS brand COCA-COLA R.

We also carry out some module aggregation, which amounts to aggregating some very specific RMS modules together. For example, the RMS modules NUTS-BAGS, NUTS-CANS, NUTS-JARS, and NUTS-UNSHELLED are essentially the same thing, and advertisements never distinguish between them.

Finally, we do some aggregation across flavors and sub-brands. For example, the brand “Lean Cuisine Frozen Entree” has 50 sub-brands in RMS (e.g., LEAN CUISINE ONE DISH FAVORITE or LEAN CUISINE SPA COLLECTION). Aggregating them together makes the matching easier and creates more tier-2 matches and fewer tiers-3/4 matches.

Table 1: Fraction of Own Advertising GRPs from Different Matches

	Median	Mean	Percentiles			
			10%	25%	75%	90%
Exact Matches	43.3214	47.8728	0	0.4212	97.2967	100
Inexact Matches	56.6786	52.1272	0	2.7033	99.5788	100

A.3 Aggregate Data

Ad Intel The Ad Intel data build comes at the {Ad Intel Brand}-Channel-Time level, and in the end we want to aggregate it to the {RMS Brand}-Market-Week level.

First, we aggregate the ad data to the {Ad Intel Brand}-{Media Type}-Market-Week level. The aggregation here only involves adding up Duration and GRP.

- Some ad occurrences come with 2/3 brands, but those brands are mostly the same product (e.g., Snapple Black Tea and Snapple Green Tea, which we eventually combine to a single “brand” as per above). To avoid double-counting the ads, we use the following trick: if an occurrence has two/three brands, treat it as two/three occurrences with half/one-third of the Duration and GRP.

RMS The RMS data build comes at the UPC-Store-Week level, and we want to aggregate it to the Brand-Store-Week level. As mentioned before, UPCs are generally much more specific than brands, as they reflect many different sizes and presentations.

- One RMS brand may contain hundreds of UPCs with different sizes (size1_amount, say 12 OZ or 24 OZ) and different multi-pack status (multi, say 6-pack or 12-pack).
 - Therefore, instead of using the units field in the RMS data, we need to calculate the volume in equivalent units: $\text{volume} = \text{units} * \text{multi} * \text{size1_amount}$. We adjust price accordingly.
- For each store-week, the brand-level variables are calculated as follows:
 - Volume: sum of UPC-level volumes
 - Price: weighted average of UPC-level prices. The weight for a UPC is its average weekly revenue in this store.
 - Promotion: weighted average of UPC-level promotion indicators ($\text{price} / \text{base_price} < 0.95$).
 - Feature/Display: weighted average of UPC-level feature/display indicators (remove missing values).

A.4 Store and Border Selection

We remove the stores that switch between different counties and stores that are not continuously tracked by Nielsen between 2010–2014. We then rank the stores by the total 2010–2014 revenue (across all products) and find the stores that constitute 90% of total revenue. We use those stores for all of our analyses.

For the implementation of the border strategy, we use the Nielsen provided mapping between counties and DMAs. From this, we construct a data set that flags the counties that lie on a border between DMAs. However, some counties change DMAs over time, since the borders are re-drawn periodically. Therefore, we remove all the counties that did not stay in a single DMA, and we remove the borders that were re-drawn.

A.5 Product Selection

We began our analysis with the top 500 national brands in the RMS data based on sales revenue between 2010–2014. The above flavor and module aggregation steps reduce the count of unique brands somewhat. We are able to match 358 of these aggregated RMS brands to brands in the Ad Intel data.

Screening Based on Own Advertising For each of the 358 RMS brands in our universe, we calculate the fraction of market-weeks with positive own advertising GRPs, and the mean own advertising GRPs conditional on it being positive. We drop 70 brands that have positive GRPs in less than 5% of observations, or whose “positive mean” is below 10 GRPs. In Table 2, we show the frequency of departments and the total revenue share. In Table 3 we show the frequency of different categories.

A.6 Advertising cost

We estimate the cost of buying an ad GRP in DMA d in week t for each manufacturer using data on advertising expenditure, impressions, and audience size contained in the Nielsen Ad Intel data set. We use these cost estimates, together with advertising effect estimates, to compute advertising return on investment (ROI).

Expenditure Data

- For Cable, Network, and Syndicated TV, ads are purchased at the national level.

Table 2: Frequency of Departments and Revenue Share

Department	No. of brands	Homescan revenue share
DRY GROCERY	127	52.19
NON-FOOD GROCERY	50	13.47
HEALTH & BEAUTY CARE	33	4.39
FROZEN FOODS	23	10.75
DAIRY	21	9.90
ALCOHOLIC BEVERAGES	19	3.49
PACKAGED MEAT	11	3.83
DELI	5	2.24
FRESH PRODUCE	1	0.14
GENERAL MERCHANDISE	1	0.20

Note: Three brands in our sample have products in two departments.

- For network ads, Nielsen obtains expenditure data from the networks. If expenditure data are unavailable, Nielsen derives estimates of expenditures using supplementary industry data and proprietary models.
- For cable ads, Nielsen’s source for expenditure data is SQAD’s NetCosts database. SQAD compiles occurrence-level data on actual purchases reported by contributing ad agencies. The measures SQAD shares with Nielsen are averages at the monthly-network-daypart level. The reported figures are believed to reflect the true weighting of upfront and scatter buys.
- Expenditure data are originally at the {Month}-{Network}-{Daypart} level for national and cable ads. Ad Intel further prorates expenditure and records the data at the {AdTime}-{Network}-{Daypart}-{Program}-{Duration} level.
- For Spot TV, ads are purchased at the DMA level.
 - Nielsen estimates spot TV expenditures by blending cost-per-point data supplied by SQAD with Nielsen’s local market ratings data. SQAD’s cost-per-point data are based on actual spot television buys reported by contributing ad agencies.

Build Advertising Cost

For each manufacturer, we do the following:

Table 3: Frequency of Categories

Category	No. of brands	Category	No. of brands
PAPER PRODUCTS	16	VEGETABLES-FROZEN	3
SNACKS	13	CHEESE	3
CARBONATED BEVERAGES	13	LAUNDRY SUPPLIES	3
BEER	11	SANITARY PROTECTION	3
DETERGENTS	11	WRAPPING MATERIALS AND BAGS	3
CANDY	11	DEODORANT	3
JUICE, DRINKS - CANNED, BOTTLED	10	NUTS	3
PACKAGED MEATS-DELI	10	BABY FOOD	2
SOFT DRINKS-NON-CARBONATED	9	PREPARED FOOD-DRY MIXES	2
CEREAL	9	COOKIES	2
PREPARED FOODS-FROZEN	7	UNPREP MEAT/POULTRY/SEAFOOD-FRZN	2
SALAD DRESSINGS, MAYO, TOPPINGS	6	COT CHEESE, SOUR CREAM, TOPPINGS	2
PET FOOD	6	PACKAGED MILK AND MODIFIERS	2
BREAKFAST FOOD	6	WINE	2
LIQUOR	6	HOUSEHOLD SUPPLIES	2
VITAMINS	6	PET CARE	2
MEDICATIONS/REMEDIES/HEALTH AIDS	6	SKIN CARE PREPARATIONS	2
DISPOSABLE DIAPERS	6	SEAFOOD - CANNED	1
CONDIMENTS, GRAVIES, AND SAUCES	5	PREPARED FOOD-READY-TO-SERVE	1
CRACKERS	5	JAMS, JELLIES, SPREADS	1
COFFEE	5	DESSERTS, GELATINS, SYRUP	1
PIZZA/SNACKS/HORS D'OEUVRES-FRZN	5	TEA	1
DRESSINGS/SALADS/PREP FOODS-DELI	5	SPICES, SEASONING, EXTRACTS	1
YOGURT	5	FRESH MEAT	1
COUGH AND COLD REMEDIES	4	PUDDING, DESSERTS-DAIRY	1
ICE CREAM, NOVELTIES	4	EGGS	1
BUTTER AND MARGARINE	4	FRESH PRODUCE	1
MILK	4	PERSONAL SOAP AND BATH ADDITIVES	1
ORAL HYGIENE	4	CHARCOAL, LOGS, ACCESSORIES	1
HAIR CARE	4	STATIONERY, SCHOOL SUPPLIES	1
FRESHENERS AND DEODORIZERS	4	TOBACCO & ACCESSORIES	1
BREAD AND BAKED GOODS	4	FIRST AID	1
SOUP	3	PASTA	1
GUM	3	VEGETABLES - CANNED	1
BREAKFAST FOODS-FROZEN	3	DOUGH PRODUCTS	1

Note: Four brands in our sample have products in two categories.

1. Merge expenditure with impressions for each ad occurrence;
2. Aggregate expenditure and impressions to the {National}-{Year} level. This involves adding up expenditure and impressions across media type, date and markets;
 - We calculate advertising cost at the annual level since expenditure fluctuates across weeks. Hence, advertising cost for all weeks in the same year y remains the same.
3. Calculate national advertising cost per GRP in year y as:

$$\text{adcost per GRP}_{\text{national},y} = \frac{\sum_d \sum_{t \in y} \text{Expenditure}_{dt}}{100 \times \sum_d \sum_{t \in y} \text{Impression}_{dt} / \text{Universe}_{\text{national},y}}$$

4. Calculate DMA-level factor for national advertising cost using:

$$\text{Factor}_{dt} = \frac{\text{Universe}_{dt}}{\text{Universe}_{\text{national},t}}$$

5. Estimate advertising cost per GRP in DMA d in week t :

$$\text{adcost per GRP}_{dt} = \text{adcost per GRP}_{\text{national},y} \times \text{Factor}_{dt}$$