Using Designed Data to Correct for Errors in Big Data

1

TODAY'S AGENDA

- → Big Data Overview
- → Sources of Error with Big Data
- → Nielsen's TV Measurement
- → Coverage Error
- → Measurement Error
- → Q&A

2

What is Big Data?

Big data refers to large data sets that are often characterized by three key attributes…

Volume
Amount of data available that is driven by the data collection methods and storage capabilities

Velocity
Speed at which data collection can occur and the pressure to manage large data sets in real time

Variety
Complexity of formats for big data that could include structured as well as unstructured data streams

Source: Laney 2001

 3

3

Big Data Examples

Apps / Websites
Social Media, Search Engines, Google Street View, Shopping, Health Apps

Transactional Data
Loyalty cards, Purchases, Reservations, Subscriptions

Device Data
Phones, Smart watches, Set top box, Smart TV data

Third-party Data
Demos, HH characteristics, email addresses, IP addresses

Tracking / Sensor Data
Cookies, Geolocation, Smartphone log data, Road sensors, Picture data

Administrative Data
Healthcare data, Housing Permits, Voter registration, Medical records

4

Big Data vs. Designed Data

Big Data

- "Found" data that typically has another primary use
- Provides stability through large sample sizes, especially for lower incidence behaviors
- Sources may have limited coverage and systematic

Designed Data

- "Made" data that is created with a specific research question in mind
- Provides high coverage and probability sampling methods ensure error is random rather than systematic

5

Relationship between Big Data and Designed Data

Supplementation

Big data can be used to complement or augment designed data

Examples:

Using Google Street View data as a source of Auxiliary Data in a Crime Survey
(Yang - AAPOR 2020)

Examining the relationship between respondent concerns and media coverage of the 2020 Census
(Virgile - AAPOR 2020)

Calibration

Designed data can be used to make corrections to big data *(or vice versa)*

Examples:

Understanding the Difference in Freight Transport Estimates With and Without Road Sensor Data
(Klingwort et al. - BigSurv 2020)

The Combination of survey and health app data: Sharing behavior, quality assessment, and validation of survey-based health indicators

 6

6

Total Error Framework for Big Data

Coverage Error	Bias introduced by under-coverage, overcoverage or duplication	Twitter account holders skew younger; One person may tweet from multiple handles
Sampling Error	The magnitude of data may lead to statistical significance and false confidence	A large probability sample of Amazon shoppers yields significant results by breaks that have no meaningful difference
Specification Error	Big data variables are pre-defined and may not exactly align with the construct of interest	You are interested in measuring a person's activity on their Android phone but are limited to the data provided by their activity log
Nonresponse / Missing Data Error	Differs from undercoverage as the reason for the missing data is different	A new housing development is not shown on Google Street View

Source: Total Error in a Big Data World: Adapting the TSE Framework to Big Data (Amaya, Biemer & Kinyon, 2020)

CONFIDENTIAL - DO NOT DISTRIBUTE 7

7

Total Error Framework for Big Data (cont'd)

Measurement/ Content Error	Various sources due to measurement, transcription, data conversion, false readings from devices, etc.	You are interested in measuring a person's heart rate but their smartwatch is providing false readings
Processing Error	Due to steps in producing a data file for analysis, linking data sources, etc.	Building permit data file is not processed frequently enough to account for updates in the data sources
Modeling Error	Results from unknown underlying mechanisms and lack of relevant variables for imputation	Trying to model missing data from medical records with limited variables to inform imputation
Analytic Error	Errors made by data users and clients in analyzing and interpreting results	Projecting the findings related to Voter Registration in a particular area to the general US population

Source: Total Error in a Big Data World: Adapting the TSE Framework to Big Data (Amaya, Biemer & Kinyon, 2020)

CONFIDENTIAL - DO NOT DISTRIBUTE 8

8

Nielsen TV Measurement

Combining panel and big data sources for media measurement

Nielsen National Panel
- Designed sample representative of the entire U.S. population.
- Panel size is ~45,000 households; smaller than most big data sources.

TV Big Data Sources
- Organic data from devices that capture tuning as people watch TV.
- Larger sample sizes; only represent a portion of the population and viewing.

CONFIDENTIAL - DO NOT DISTRIBUTE

9

Coverage Error

UNDERCOVERAGE

Bias introduced if there is a large difference in the characteristics of interest for the covered and uncovered populations

ex. What is the coverage for each TV data provider? Which homes return data?

OVERCOVERAGE

May include units that are out of scope leading to inefficiencies and increased cost and potential bias

ex. The data set could include businesses, which is outside the TV household universe

DUPLICATION

Can bias data by over representing the duplicated units, especially problematic when combining data sets

ex. A household may be included in multiple big data sources.

Source: Total Error in a Big Data World: Adapting the TSE Framework to Big Data (Amaya, Biemer & Kinyon, 2020)

CONFIDENTIAL - DO NOT DISTRIBUTE 10

10

Coverage of TV Big Data Sources

Each of these big data sources only covers some of the ways people watch TV today and often our data is further limited to certain device types or providers

77% have an internet streaming device

53% have an enabled Smart TV

41% have an enabled set top box

Source: Nielsen National U.S. TV Panel, Based on Scaled Installed Count percent of all TV households.

CONFIDENTIAL - DO NOT DISTRIBUTE

11

Bias in Set Top Box Data

- Set top box data **under-represents Hispanics** (by 33%, 49% for Spanish Dominant Hispanics) **and Blacks** (by 34%)
- It also **under-represents younger people** (18-34 by 17%) and over-represents older age groups (50+ by 15%)
- Substantial research shows that homes with set top boxes **view differently** than homes whose data is not returned or who view from other sources

CONFIDENTIAL - DO NOT DISTRIBUTE

12

Approaches to Integrating TV Big Data Sources

Panel as the Foundation
Designed panel data is the basis of measurement supplemented with big data sources to increase

Big Data as the Foundation
Big data is the basis of measurement where behavioral modeling adjustments based on

13

Panel as the Foundation

Nielsen's Panels
The **foundation** of measurement, providing **complete coverage** of all segments of the market

+

Smart TV and Set Top Box Data
A **supplement** to the full-coverage panel, only representing the **portion** of the market it covers

Smart TV and set top box data expand the sample size but only in the portion of the population they cover

14

Weighting Used to Combine Data Sources

Modifying weighting approach ensures that each source only represents its coverage in the population

Design weight to adjust for disportionate sampling in areas covered by big data sources (i.e., more homes in covered than non-covered areas)

Final weighting ensures different viewing sources (devices, providers, etc.) are reflected in proportion to population in the final estimates

Also include a **household tuning control** based on panel tuning levels to account for any remaining missing tuning from these devices

15

Big Data as the Foundation

Smart TV or set top box

Smart TV and/or set top box rating calibrated to account for missing data, including noncoverage of devices that don't return data

Other viewing sources

Smart TV or set top box rating calibrated to account for other viewing sources not covered by the Smart TV or set top box data

Over the air

Stabilized rating from small sample of metered homes that receive broadcast TV over the air via an antenna

Each component is weighted to represent its appropriate contribution in the market

CONFIDENTIAL - DO NOT DISTRIBUTE

16

Calibration Approach

The approach uses **behavioral** adjustments by comparing tuning for each station/network to account for differential viewing behaviors

It is **dynamic** where the learning data updates daily to reflect real changes in tuning and adjustments are by day/daypart to reflect the differences in behavior that occur throughout the day and week

Adjustments are **specific** by demographic group and computed separately by age, gender, race and ethnicity

CONFIDENTIAL - DO NOT DISTRIBUTE

17

Measurement Error

Measurement error is the difference between a measured quantity and its true value

SOURCES	
SURVEY	**BIG DATA**
➔ Question wording ➔ Social desirability bias ➔ Interviewer administration ➔ Recall bias	➔ Measurement process ➔ Transcription Errors ➔ Data conversion errors ➔ False or outdated readings from

Source: Groves (1989); Amaya, Biemer & Kinyon (2020)

CONFIDENTIAL - DO NOT DISTRIBUTE

18

Common Homes Provide a Source of Truth

Common homes provide an ongoing way for Nielsen to evaluate provider data in real homes using our meter data, enabling a side by side tuning comparison between data sources

COMMON HOMES PROCESS:

1. Identify Nielsen Panel households within provider data set
1. Match common devices in Panel and provider data
1. Compare tuning collected through

ENABLES NIELSEN TO:

- Understand differences between collection methods and data processing
- Pinpoint data quality concerns (e.g., missing or miscredited tuning)
- Examine minute-

19

Using Panel Data to Improve Measurement

CLEANING THE RAW TUNING DATA

Identify data issues and develop corrections for those limitations using "common homes" (Nielsen meter panel households within a provider data set)

DETERMINING HOUSEHOLD DEMOGRAPHICS

Determine household demographics and compositions by using third-party data as well as household tuning and known panel information

20

Types of Refinement Opportunities

Comparison of tuning from the same TV set - Set Top Box/Smart TV vs. Nielsen Meter

Missing or excess tuning minutes	Incorrect station or multiple stations identified	Incorrect time credited	Tuning without a station identified, credit time, or other relevant information

Nielsen's Common Home Analyses are critical to **identifying data measurement challenges** and also for developing and testing techniques to **correct for these limitations**

Source: Nielsen Data Science Common Homes Analysis

21

Invalid Tuning Due to Provider Initiated Event

Erratic spike in tuning data observed

Without correction, would translate to an unrealistic surge in the audience estimates for that time period

Nielsen developed a patent pending in-house model to identify and correct for provider-initiated events

Source: Nielsen Data Science Common Homes Analysis

CONFIDENTIAL - DO NOT DISTRIBUTE

22

Determining Household Demographics

Unlike panel/survey data, most Big Data (Set-Top Box, Connected TV, SmartTV) does not come with directly collected household characteristics and demographics

This information is critical for knowing household profiles, correctly representing these segments of the population, and measurement of true persons audiences

Over the years, Nielsen has developed capabilities for identifying household demographics for big data sources

CONFIDENTIAL - DO NOT DISTRIBUTE

23

Third-Party Data Alone is Not Enough

Third-party characteristic/demographic data can be **missing** or **incorrect**

~15% of the homes are **missing** all characteristics & demographics

20-50% of the time age, race, and ethnicity are **inaccurate**

How do we know?

Nielsen's representative panels put us in a unique position to analyze this data

Metrics Accuracy & Coverage

1) Receive third-party assigned char/demos for Nielsen metered homes

2) Compare char/demos assigned by 3rd-party vs. known char/demos of Nielsen panelists from the same home

3) Understand accuracy and completeness of third-party assigned characteristics + demographics

Data Sources: Nielsen's provider validation studies from 2011 - 2018. Representative panels include but not limited to Nielsen's National Television panel of 45K metered homes

CONFIDENTIAL - DO NOT DISTRIBUTE

24

Demographic Identification Technique

Nielsen's characteristic & demographic identification utilizes a state-of-the art technique that relies directly on tuning in Set-Top Box + SmartTV homes and continuously learns from known information of more than 100,000 Nielsen panelists

The technique is able to reflect unique characteristic & demographic profiles, and shows improved performance compared to using just third-party data on it's own

25

SUMMARY

Big data alone not usable for measurement

Designed panel data used to correct for errors in big data sources

Can use panel or big data as foundation based on fit for specific uses and needs

26