Using Designed Data to Correct for Errors in Big Data

TODAY'S AGENDA

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

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
**Big Data Examples**

- **Apps / Websites**
  - Social Media, Search Engines, Google Play Store, Shopping, Health Apps

- **Device Data**
  - Phones, Smartwatches, Set-top box, Smart TV data

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

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

- **Third-party Data**
  - Demographics, Household characteristics, email addresses, IP addresses

- **Transactional Data**
  - Loyalty cards, Purchases, Reservations, Subscriptions

- **Third-party Data**
  - Demographics, Household characteristics, email addresses, IP addresses

**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

**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

**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 between, quality assessment, and validation of survey-based health indicators
  - **Kapousouz et al. - BigSurv 2020**
### Total Error Framework for Big Data

<table>
<thead>
<tr>
<th>Error Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coverage Error</td>
<td>Bias introduced by under-coverage, over-coverage or duplication. Twitter account holders view younger people more than actual.</td>
</tr>
<tr>
<td>Sampling Error</td>
<td>The magnitude of data may lead to statistical inferences or false confidence. A large probability sample of Amazon shoppers provides an accurate view.</td>
</tr>
<tr>
<td>Specification Error</td>
<td>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 Amazon Prime, but it is based on data provided by their phone.</td>
</tr>
<tr>
<td>Nonresponse / Missing Data Error</td>
<td>Differs from undercoverage as the reason for the missing data is different. A true housing development is not shown on Google Street View.</td>
</tr>
</tbody>
</table>

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

### Total Error Framework for Big Data (cont'd)

<table>
<thead>
<tr>
<th>Error Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Measurement/Content Error</td>
<td>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.</td>
</tr>
<tr>
<td>Processing Error</td>
<td>Due to steps in producing a data file for analysis, linking data sources, etc. Building permits data is not processed frequently enough in the data sources to account for updates.</td>
</tr>
<tr>
<td>Modeling Error</td>
<td>Results from unknown underlying mechanisms and lack of relevant variables for imputation. Trying to model housing data from medical records with limited information on medications.</td>
</tr>
<tr>
<td>Analytic Error</td>
<td>Errors made by data users and clients in analyzing and interpreting results. Processing clients need to keep registration of events in particular areas not general to a population.</td>
</tr>
</tbody>
</table>

### 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.
Coverage Error

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

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

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

- Bias introduced if there is a large difference in the characteristics of interest for the covered and uncovered populations.
- May include units that are out of scope leading to inefficiencies and increased cost and potential bias.
- Can bias data by over representing the duplicated units, especially problematic when combining data sets.

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

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.

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.
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 sample size within the portions of the population they cover.

Big Data as the Foundation

- Big data is the basis of measurement where behavioral modeling adjustments based on panel data account for areas not covered by big data sources.

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.

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 disproportionate 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.
Big Data as the Foundation

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

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.

Measurement Error

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

<table>
<thead>
<tr>
<th>SOURCES</th>
<th>BIG DATA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Survey</td>
<td>Question wording</td>
</tr>
<tr>
<td></td>
<td>Social desirability bias</td>
</tr>
<tr>
<td></td>
<td>Interviewer administration</td>
</tr>
<tr>
<td></td>
<td>Recall bias</td>
</tr>
<tr>
<td></td>
<td>Measurement process</td>
</tr>
<tr>
<td></td>
<td>Transcription Errors</td>
</tr>
<tr>
<td></td>
<td>Data conversion errors</td>
</tr>
<tr>
<td></td>
<td>False or outdated readings/estimates</td>
</tr>
</tbody>
</table>
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
2. Match common devices in Panel and provider data
3. Compare tuning collected through Nielsen Meter versus

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

**CLEANING THE RAW TUNING DATA**
Identify data issues and develop corrections for those limitations using "common home" (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.

Types of Refinement Opportunities

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

- Missing of eKRBES tuning minutes
- Incorrect station or multiple stations identified
- Incorrect time credited
- Tuning without a station identified, credited time, or other relevant information

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

Source: Nielsen Data Science Common Homes Analysis
Invalid Tuning Due to Provider Initiated Event

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

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

Determining Household Demographics

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

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?

1. Nielsen's representative panels put us in a unique position to analyze this data.
2. Compare the demographics provided by the party in question to what we know of the occupants of the home.
3. Check the accuracy and completeness of the party's demographic information.
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 its own.

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