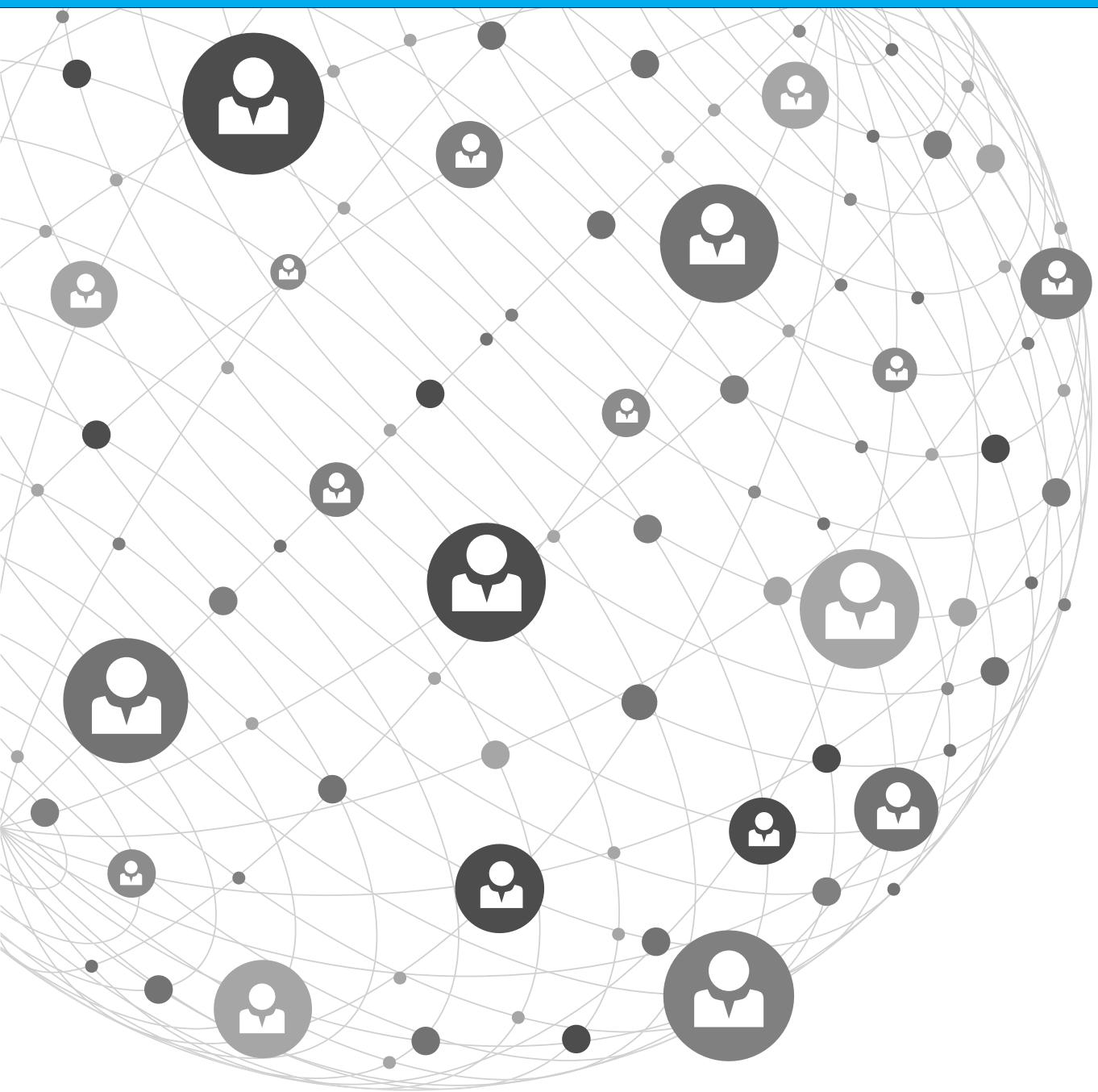
NIELSEN JOURNAL of MEASUREMENT

JULY 2016-VOL 1 ISSUE 1

THE VALUE OF PANELS IN MODELING BIG DATA

by Paul Donato, Chief Research Officer, Nielsen



The world of measurement is changing.

Thanks to recent advances in data collection, transfer, storage and analysis, there's never been more data available to research organizations. But 'Big Data' does not guarantee good data, and robust research methodologies are more important than ever.

Measurement Science is at the heart of what we do. Behind every piece of data at Nielsen, behind every insight, there's a world of scientific methods and techniques in constant development. And we're constantly cooperating on ground-breaking initiatives with other scientists and thought-leaders in the industry. All of this work happens under the hood, but it's not any less important. In fact, it's absolutely fundamental in ensuring that the data our clients receive from us is of the utmost quality.

These developments are very exciting to us, and we created the Nielsen Journal of Measurement to share them with you. This paper is part of VOL1 ISSUE 1 of the Journal.

WELCOME TO THE NIELSEN JOURNAL OF MEASUREMENT

SAUL ROSENBERG

The Nielsen Journal of Measurement will explore the following topic areas in 2016:

BIG DATA - *Articles in this topic area will explore ways in which Big Data may be used to improve research methods and further our understanding of consumer behavior.*

SURVEYS - *Surveys are everywhere these days, but unfortunately science is often an afterthought. Articles in this area highlight how survey research continues to evolve to answer today's demands.*

NEUROSCIENCE - *We now have reliable tools to monitor a consumer's neurological and emotional response to a marketing stimulus. Articles in this area keep you abreast of new developments in this rapidly evolving field.*

ANALYTICS - *Analytics are part of every business decision today, and data science is a rich field of exploration and development. Articles in this area showcase new data analysis techniques for measurement.*

PANELS - *Panels are the backbone of syndicated measurement solutions around the world today. Articles in this area pertain to all aspects of panel design, management and performance monitoring.*

TECHNOLOGY - *New technology is created every day, and some of it is so groundbreaking that it can fundamentally transform our behavior. Articles in this area explore the measurement implications of those new technologies.*

THE VALUE OF PANELS IN MODELING BIG DATA

BY PAUL DONATO *Chief Research Officer, Nielsen*

OVERVIEW

Hardly a day goes by without an industry report on audience fragmentation. Of course, it's not a new phenomenon. With the rise of cable in the 80s, digital broadcast satellite in the 90s, Internet video in the 2000s and, more recently, over-the-top options, television audiences have enjoyed a steady stream of new programming choices year after year: more networks, more niche programs, and more ways to watch them.

For the research community however, that increased diversity has come at a price, and the accelerating pace of change in recent years is straining the panel-based measurement capabilities that the industry has historically relied on to monitor viewing activity. It has simply become a challenge to assemble panels large enough to provide stable measurement for programs with small audiences.

Return path data from television set-top boxes (RPD) represents an opportunity to overcome that problem, but only if the limitations and biases in these data can be corrected and validated. This document describes how panels can effectively correct for these limitations and help validate the ratings derived from RPD datasets.

Panels and RPD together represent a winning combination for accurate and stable video audience measurement.

LIMITATIONS OF RETURN PATH DATA

What is return path data? It's viewing activity collected by the very infrastructure that delivers media content to viewers – content (TV shows, program guides, etc.) goes one way, and usage activity (channel and program selections, clickstream, remote control activity, etc.) is returned to the distributor along a technical path that includes set-top boxes, middleware systems and headend servers.

With modern equipment, RPD data collection can now be activated across entire media delivery systems. As a result, RPD datasets typically cover viewing activity from millions of households, and ratings generated from RPD homes can be highly stable. However, RPD-based measurement is hampered by four important limitations that can result in highly inaccurate audience figures in a video environment where a tenth of a ratings point can determine the success of a program, network or station.

These are the limitations:

1. **Inability to detect whether the television is on**

   The average set-top box is turned on anywhere from 50% to 80% of the time, while the television to which it is connected is turned on only about half that time. This means that a model must be developed to determine when the television set is actually on.

2. **Bias within the home and within a market of using return path data**

   Not every TV set in an RPD home can return RPD data. For example, in satellite homes (approximately 35% of RPD homes), only those sets with an Internet or telephone connection can return data. By definition, these sets are more likely to be used to access video on demand, so they will be a poor reflection of activity from other sets in the home. Further, not all homes in a market will provide RPD data: In most markets, a limited number of multichannel video programming distributors (MVPDs) will provide RPD data, and the corresponding datasets may cover anywhere between 15% and 60% of the homes in each market. Because of the way systems are licensed, these MVPDs usually are skewed to certain counties, and as a result, demographics and even channel availability differ from one MVPD to another in the market.

3. **Lack of knowledge about household members**

   Generally, researchers who use RPD measurement do not know the demographic composition of each household, and they attempt to address that problem via matches with third-party datasets. Those third-party datasets, in turn, aren't always accurate, and privacy policies may prevent such matches altogether.

4. **Inability to determine who in the household is actually watching**

   RPD is associated with a box, not a person, and it's impossible, from the RPD alone, to determine precisely who is actually watching the TV set connected to that box.

PANELS CAN IMPROVE RETURN PATH DATA

Fortunately, panels can address all these limitations. Let's take them one by one

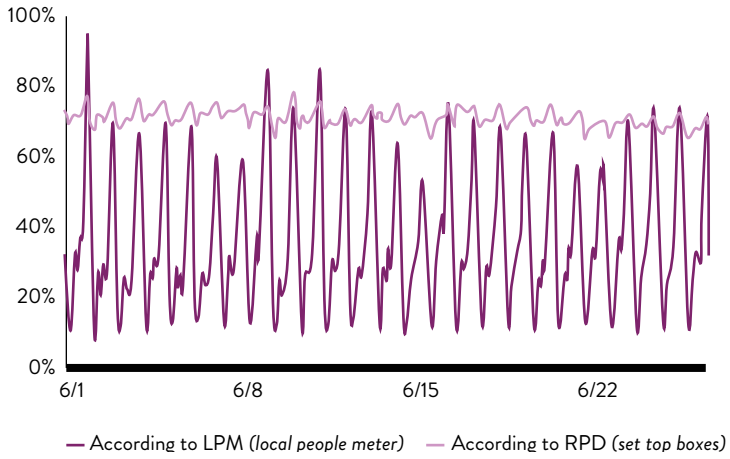
Using panels to address RPD's inability to detect whether the television is on.

The chart immediately below illustrates the relation between Set On as measured by a representative Nielsen Local People Meter panel (LPM) and the apparent Set On from the raw data registered by RPD for a typical MVPD in a major people meter market over a period of one month. From the RPD standpoint, the TV sets in that market appear to be on around the clock: There's never a moment over the course of that month, day or night, when less than 60% of the households were using television, a conclusion that is clearly inaccurate. Both lines peak during prime time each day, as expected, but even those peaks aren't well correlated – notice for instance how the prime time HUT level (% Households Using Television) rises in the LPM line between June 7 and June 8, while it drops for the RPD line.

This chart illustrates the challenges of eliminating artificial viewing captured by the RPD when the set is not actually on.

IS THE TV SET ALWAYS ON?

Fluctuations in the percentage of Household Using Television (HUT)over the course of 1 month

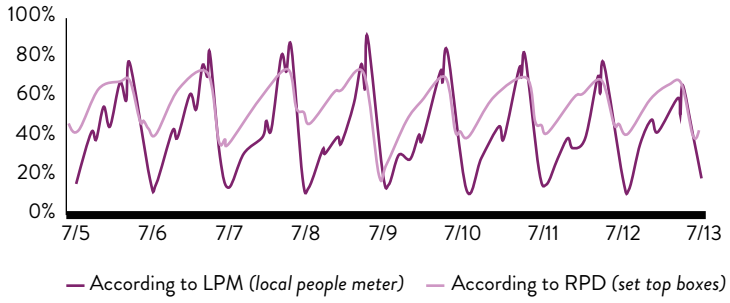


If all systems looked the same—if, that is, there were a relation between when the TV set is on and when it appears to be on—a single model might be developed to adjust RPD figures with those found in panels in the same market, but unfortunately that's not the case.

The charts below illustrate the same relationship between Set On as measured by the LPM panel and the apparent Set On from the raw data registered by a certain RPD provider in market 'A' (top) and in a different market, market 'B' (bottom), where a different model of set-top boxes is in use.

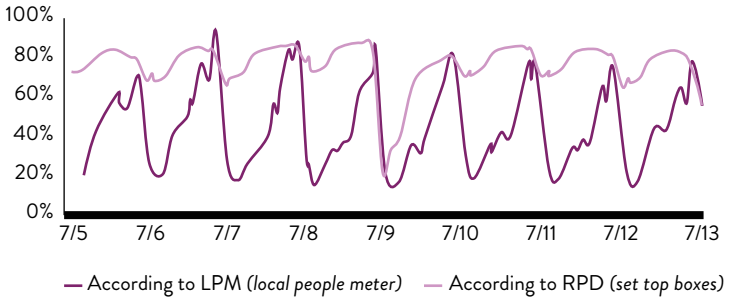
'SET ON' DIFFERENCES IN MARKET 'A'

With RPD provider 'X' and set-top model 'M'



'SET ON' DIFFERENCES IN MARKET 'B'

With RPD provider 'X' and set-top model 'N'



The only way to adjust for a problem of this scope is to have access to panels that measure across many markets and many RPD providers. This provides all the information necessary to adjust for these substantial biases. Without the size, scope and accuracy of these panels, it would be impossible to estimate the degree to which ratings are overstated by RPD.[1]

Addressing home- and market-level bias introduced because we do not have return path data from all sets in a market.

Often, one or more sets in an RPD household do not have the ability to return data. Even if all MVPDs cooperated and made RPD available across all their markets, only about half the television sets in use in the U.S. would be able to send data back. As noted, however, the datasets available cover only 15% to 60% of the homes in each market.

Our panels allow us to examine the differences in behavior between homes that do and don't have RPD data, and consequently to construct models to calibrate RPD readings to determine what's actually happening. However, the effectiveness of the models is a function of 1) the coverage of the RPD homes in the target market, 2) the coverage of the RPD homes used to develop the model and 3) the consistency of the bias between RPD sets and homes in the market(s) on which the model is developed and the market to which the model is applied.

As you would expect, when markets have high RPD coverage, there is less bias in the RPD data and it is not necessary to make large adjustments to it. However, in markets where RPD

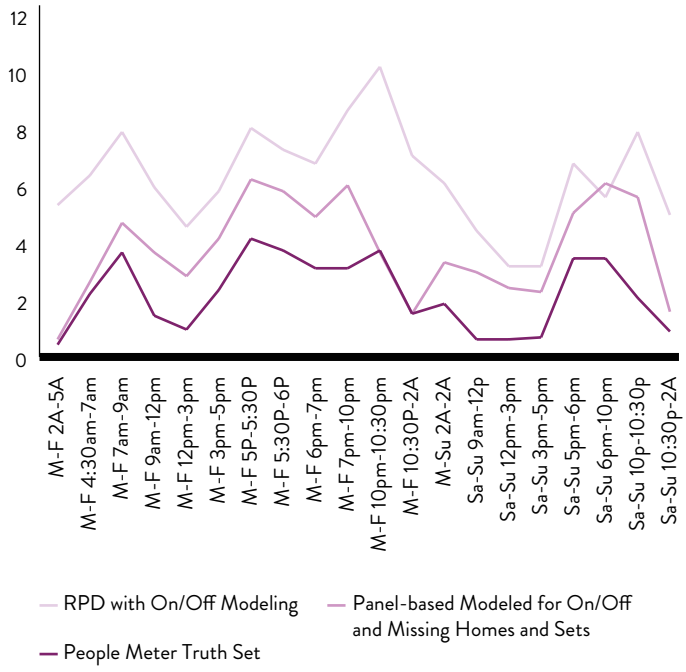
[1] *In some cases, and with some providers, some information about Set On may be available through an HDMI back channel, though this in itself creates another bias as these are a unique set of television sets and providers.*

coverage is low and therefore bias is more significant, models are crucial to make more aggressive adjustments for bias.

We will illustrate this by looking at two people meter markets (i.e. markets where we have the benefit of panels with person-level viewing data): St. Louis, where we have 500,000 RPD households covering 42% of the market and Dallas, where we have 150,000 RPD households covering only 6% of the market.

The following figure illustrates household rating levels for the FOX affiliate KDFW in Dallas, modeled using only non-Dallas people meter data from St. Louis (also in the Central time zone).

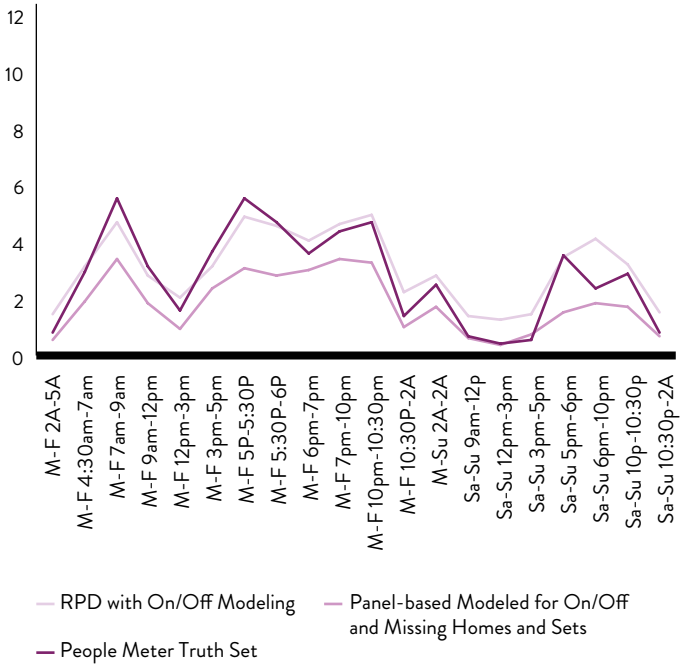
DALLAS KDFW (FOX) HH RATINGS



The "RPD with On/Off Modeling" is modeled only for whether the set is on. This shows that, when only 6% of the market is covered by available RPD data, as is the case in Dallas, significant bias can exist; note the difference from the people meter data for that market. However, using machine-learning algorithms developed by looking at the relation between St. Louis RPD data and St. Louis people meter data, we developed a model to correct this overstatement. The chart shows that the model pulls the significantly overstated RPD data much closer to the levels of the people meter truth set.

The significant overstatement occurs because the Dallas RPD data covers only 6% of the Dallas households. The model corrects a significant amount of the resulting bias, but cannot correct all of it. The residual error may be due in part to the fact that St. Louis RPD households cover 42% of the population, which results in a good model, but one that is perhaps less aggressive than what is required for Dallas. The solution is to create pseudo-markets as similar to the target market as possible in order to develop the necessary adjustment models.

ST. LOUIS KTVI (FOX) HH RATINGS



What happens when we reverse the modeling process (i.e. develop the model based on Dallas RPD and people meter data and apply it to the St. Louis RPD data)? The figure above shows that, at the 42% level of RPD coverage, the RPD data corrected only for On/Off does a good job of estimating the people meter truth set. However, as we can see from the fact that the model pulls the raw data to and beyond the people meter data, the bias adjustment model is overly aggressive:

The bias in the RPD data is modest in St. Louis, and it does not need so large an adjustment.

We ran the same analysis for three other affiliates, and in all cases, the Dallas estimate of the people meter ratings significantly improved. Nevertheless, in three out of four St. Louis affiliates, the model still over-adjusted. The simple On/Off-adjusted RPD data more closely matched the people meter ratings.

The implication is that a model developed using RPD data with reasonable coverage (in this case 42%) can correct for much of the bias in markets where RPD coverage is as low as 6%. However, the reverse is not true: Unmodeled (On/Off only) RPD data grows more accurate as coverage grows, and further modeling it based on a model where the biases are more significant (because of the low 6% coverage) over-adjusts the data.

Importantly, without panels used for modeling and for measuring accuracy, there is no way to understand or manage any of these factors when using RPD data to enhance television ratings.

How panels may help in understanding household composition.

Who lives in each RPD home? In some cases, third-party databases are available to estimate household composition. But some MVPD's do not allow direct matches of RPD households for privacy reasons. We frequently test third-party database accuracy against our panels, and we have found that "fingerprint modeling" of household composition is a more accurate representation than much of the third-party data, anyway.

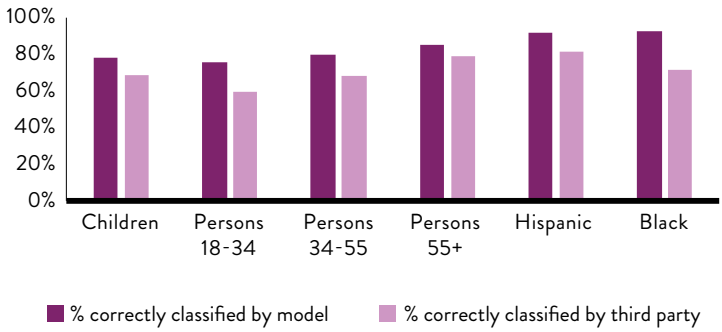
Fingerprinting is a technique used to model demographic composition of a household through its overall set-top box tuning activity. Its principle is to look for viewing patterns across sets in an RPD home and for similar patterns in our panel households, where we know the household composition.

We use variables linked to viewing activity in our models to estimate the presence of certain demographic groups in the household. Third-party suppliers sometimes rely on credit card transactions, vehicle registration, shopping activity, etc. to accomplish the same task.

It's instructive to compare the results of the two methods. The following chart shows the accuracy of household classifications of our fingerprint model based on our television panel versus a typical third-party data supplier.

CLASSIFICATION ACCURACY

Among households with specific demographic characteristics



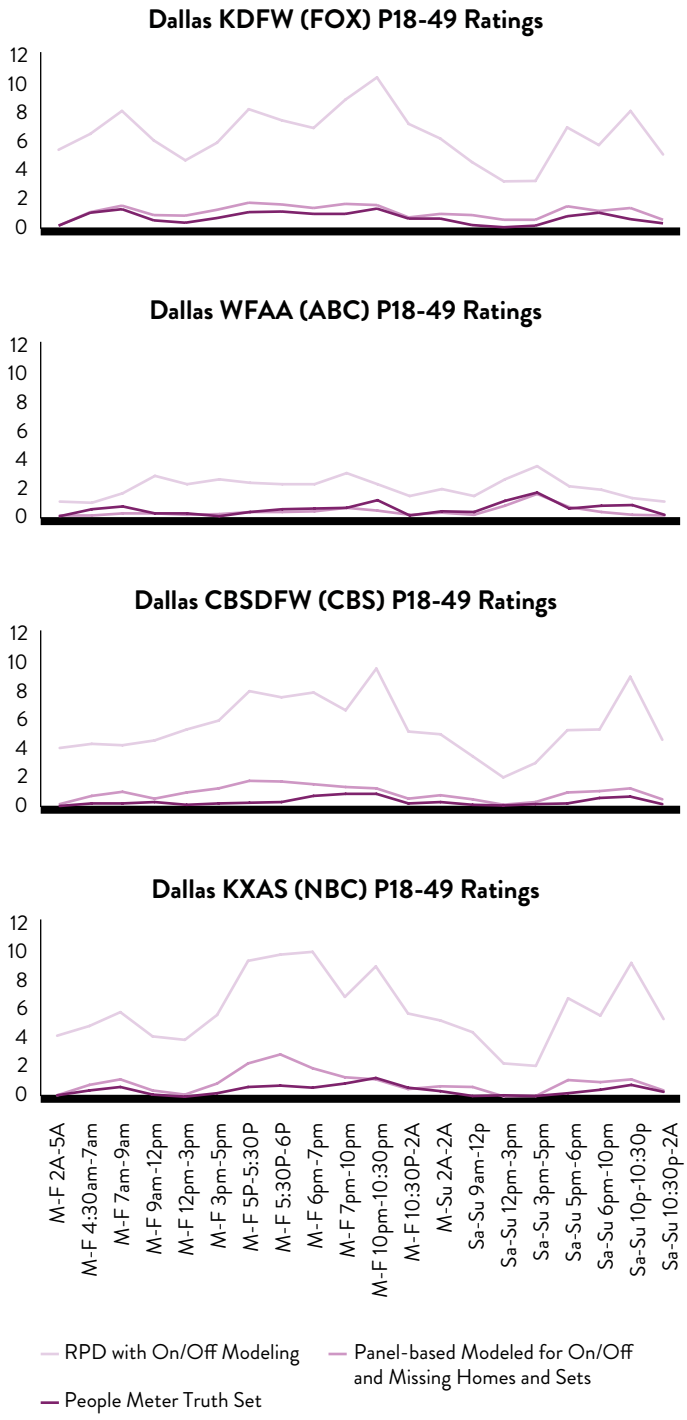
It turns out that television behavior is a very good predictor of household composition. In an environment highly focused on privacy, panels eliminate dependency on third parties. Even in cases where third-party data is available, panels are needed to serve as reference points.

How panels may help determine who in the household is actually watching.

RPD alone cannot tell us who is watching. Once we know that the set is actually on, and have a good estimate of the household composition, we turn to panels again to identify who is actually watching the television. Panels are used not only to "train" our models, but also to serve as reference points to validate the accuracy of those models when we apply them to a different set of panelists.

The following figures compare the audience of persons aged 18 to 49 in Dallas predicted by our models for each of the four primary local stations in that market with audiences actually measured by our panels. While these models are still "first generation," we can clearly see how well they perform against the LPM reference data. In each case the adjusted data is very close to the people meter truth set.

TV RATINGS



THE WAY FORWARD

This research demonstrates that models should be developed from markets likely to show RPD biases that are similar to those of the target market. Demographic differences may be controlled for by weighting or by creating demographically similar pseudo-markets as a basis for modeling, and fingerprinting techniques may be used effectively to identify who is actually watching TV in an RPD home.

This is hard work. Models need to be refined and validated regularly in order to properly reflect changes in viewing habits. But given that the available RPD in one of the markets used in this research represents only 6% of that market, the findings are remarkably encouraging.

This paper has shown how high quality panels may be used to address some important limitations and unlock the value of RPD datasets. But there are other serious limitations: For instance, new set-top boxes (those most capable of returning RPD data) are often rolled out to higher income subscribers first, further exacerbating the demographic skew we noted in the paper; not all MVPDs have set up their systems to provide time-shifted viewing data (for some programs and networks, this may account for a significant part of their viewing); and while modeling an RPD market after another RPD market is difficult, using an RPD dataset to model over-the-air (OTA) or over-the-top (OTT) viewing is another ballgame altogether. We will explore those areas in more detail in a follow-up paper, but one thing is for sure: Reference panels are not just useful but essential if one is to derive any reliable insight from RPD datasets.

nielsen