ERASMUS UNIVERSITY ROTTERDAM
Erasmus School of Economics

Master Thesis Business Analytics & Quantitative Marketing

In collaboration with Pointlogic, a Nielsen Company

ESTIMATING THE TV AUDIENCE PER DEMOGRAPHIC SEGMENT FROM HOUSEHOLD LEVEL DATA

September 13, 2019

M.E.J. (Martijn) Pigeaud
Student ID number: 481592

Supervisor: dr. W. Wang

Abstract

In TV advertisements, recent developments such as the set-top box have made it possible for media agencies to collect detailed numbers of viewers (also called *TV ratings*) of the shows they broadcast. Ideally, this information can be used to gain insight in the demographic composition of the target group for each TV channel at each moment in time. However, this can not directly be achieved. The reason for this, is that only household-level data can be obtained, rather than individual-level data. In this work, two methodologies have been developed to obtain TV ratings per demographic segment from household viewing and composition data. In the first method, theories of group utility and choice models are combined to form a household level choice model, using individual utilities. From a number of group utility specifications, the multiplicative group utility (using the product of individuals' utilities as household utility) proved to be the most suitable to apply in a choice model. In the second method, a linear regression model and a LightGBM model are estimated on data where segment TV ratings are known, to apply this model on data from a different source. This method uses aggregated household ratings as predictors and is therefore called the *aggregated method*. To tune the LightGBM model, Bayesian Optimization of its hyperparameters is used. The LightGBM model proved to outperform the linear model by a large margin in terms of model fit. The LightGBM model also outperformed the choice model in terms of model fit and computation time. Therefore, we conclude that the LightGBM model of the aggregated method is the most suitable to estimate segment TV ratings. However, the aggregated method can only be applied to household viewing data where segment TV ratings of data from a similar source is available. In case this data is unavailable, one has to apply the choice model.

Key words: TV ratings, target group identification, segmentation, choice model, group utility, regression, LightGBM, Bayesian Optimization.

PREFACE

In front of you is my master thesis titled "Estimating the TV audience per demographic segment from household level data". This thesis describes the setup and results of my statistical research aimed at developing methodology to obtain the number of TV viewers in each demographic segment for specific channels and moments in time. I have written this thesis as part of the finalization of the MSc Business Analytics & Quantitative Marketing at the Erasmus University Rotterdam. Furthermore, I have performed this research as part of my graduation internship at Pointlogic, a Nielsen Company, based in Rotterdam, the Netherlands.

I would like to thank my university supervisor, Wendun Wang, for his feedback on my work throughout the process of writing this thesis. Furthermore, I would like to thank Xin Wang from Pointlogic, for the countless meetings, chats and discussions we have had to discuss methodologies, coding problems and preliminary results. Your tips and tricks have taught me a lot. I would like to thank Harald Hoogstrate from Pointlogic as well, for the great discussions we have had that led to the methodologies I applied in my research.

Martijn Pigeaud

Delft, September 13th, 2019.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

Symbol	Definition
a	Segment.
A	Number of segments.
c_{aj}	Percentage of individuals in segment a living in household of type j.
$c_{\hat{v}_t}$	Coefficient of Variation (CV) of segment TV ratings at time t.
d_t	Vector of binary variables denoting the date and time at t.
e_{ast}	Scaled deviation of segment TV ratings from population mean TV ratings for segment a and channel s at time t.
g_{hj}	Indicator equal to one when household h is of type j, zero else.
h	Household.
H	Number of households.
i	Individual person.
IX	index of segment TV rating.
j	Household type.
m_a	Total number of people in segment a.
n_{ah}	Number of individuals in segment a in household h.
n_h	Vector of size A with elements n_{ah}.
q_a	Percentage of total population that belongs to segment a.
s	Channel.
S	Number of channels.
t	Time unit.
T	Total time, end of time interval over which data was obtained.
u_{hst}	Utility of household for channel s at time t, estimated in choice model.
v_{ast}	Segment TV ratings for segment a, channel s at time t.
w_i	Representativity weight of individual i.
X_a	(1) Regressor matrix in Multinomial Logit choice model. (2) Regressor matrix in the linear aggregated model.
y_{hst}	Binary variable equal to one if household h watched channel s at time t.
z_{jst}	Aggregated household TV ratings for household type j, channel s at time t.

Symbol	Definition
α_a	Intercept for segment a in the linear aggregated method.
β_a	Regression parameter for date and time for segment a in the linear aggregated method.
β_{st}	Regression parameter in Multinomial Logit choice model.
γ_{aj}	Regression parameter of scaled household TV ratings $c_{aj}z_{jst}$ for segment a.
δ_{st}	Tuning parameter to align estimated segment TV ratings with total TV ratings.
ϵ_{ast}	Error term in regression of linear aggregated model.
ζ_h	Household weight if household h is a single household, zero else.
η_{ast}	Scaled error term in regression of linear aggregated model.
$\hat{\pi}_{ast}$	Estimated probability of an individual in household a watching channel s at time t.
σ_a^2	Assumed variance of η_{ast}.
τ_h	Binary variable indicating whether household h is a single household or not.
ψ_{st}	Total TV ratings of channel s at time t.
ω_{ast}	Utility perceived by an individual in segment a by watching station s at time t.
ω_{st}	Vector of size A with elements ω_{ast}.

1 INTRODUCTION

In recent years, the number of methods of online advertising have increased at a high pace. In online advertising, identifying whether an individual internet user is part of an advertiser's target group is relatively easy: the user's browser saves cookies, telling the advertiser about the user's browsing history, general interests and approximate location (Farahat & Bailey, 2012). Using this information, it is easy to tell the general demographic characteristics of this user.

Unlike online advertisers, TV advertisers have limited possibilities to identify such characteristics of the people that watch a certain TV channel. Using two-way cable television set-top boxes (STB), it is possible to track the viewing behaviour of households in the possession of such an STB (Chang, Kauffman, & Son, 2012). This so-called Return Path Data (RPD) shows the viewing behaviour on the household level. When matched to data of demographic information of this household and aggregating over all households, one can theoretically identify target groups of different channels.

The composition of these target groups is of great interest to advertisers. Using this demographic composition of each channel's viewers, they can select the channels that are likely to be watched by the target group of the product they are trying to sell, and show their advertisements on these channels. By doing so, advertisers can reach a larger part of their target group with less advertisements, increasing their advertising efficiency. For TV channels, knowing what their target group is, is interesting as well, as it may yield an increase in revenue. Advertisers are obviously willing to pay more to show their advertisements if they have a larger probability of reaching their actual target group (J. Webster & Phalen, 1997). Furthermore, TV can be seen as a *mass medium*, as almost all people watch TV on a regular basis (Sharp, Beal, & Collins, 2009). Therefore, TV is a medium that can be used to reach an advertiser's full target group. Besides, there are more possibilities than ever for TV advertisers to focus their advertisements on channels watched by their target group, because the number of TV channels individuals can choose from has increased massively in recent years (J. G. Webster, 2005).

However, using RPD data only, this target group can not be obtained directly, as that would imply treating the household as a single entity. As Alderman, Chiappori, Haddad, Hoddinott, and Kanbur (1995) concluded, treating a household as a single entity ia likely to lead to incorrect conclusions. To make predictions of TV ratings (the number of individuals watching a specific channel at a specific time) per segment, individual level data is desired rather than household level data (Gensch & Shaman, 1980). This is reflected in the RPD data as this only states that at least one person in a household watched a certain channel. It does not contain any information on which members of the household are watching television. This implies that the TV viewing behaviour can not directly be linked to a certain demographic characteristic. It is therefore not possible to directly obtain TV ratings per demographic group from the sole observation of RPD data.

This research focuses on developing models to make estimates of the demographic segment TV ratings, using household level data. Traditionally, predictions on consumer behaviour are made using statistical choice models. However, these choice models are developed under the assumption that individuals choose independently of each other. In the case of TV channel selection though,

different members of a household have to decide together on the channel to watch. Therefore, in this paper, we develop a choice model (CM) that can model the influence individuals have on the household decision, and be used to estimate segment TV ratings as well. This household choice model estimates the utilities individuals from different segments perceive from watching each channel from the observed household behaviour. The choice model is estimated separately per quarter hour, to allow for different utilities over time in the data. The downside of this, is that the choice model is likely to require a long time to estimate in a large number of time periods.

Besides choice models, it is also possible to model the relation between segment TV ratings and aggregated TV ratings of households directly, and use the output of this model to estimate segment TV ratings on observed viewing behaviour data of other sources, such as RPD data. To do so, we develop a model estimating the relation between segment TV ratings and aggregated household TV ratings directly. This model is named the *aggregated model*.

The aggregated model has two theoretical advantages over the choice model. Firstly, it can be estimated over a large period of time at once, yielding a relatively low total computation time. Secondly, the aggregated model is much more similar to a traditional regression model. Therefore, analytical methods such as Ordinary Least Squares can be applied. However, these analytical techniques only yield accurate models in case a number of assumptions hold. For example, the relation between the dependent variable and independent variables should be linear. Nonlinear machine learning techniques such as Gradient Boosting Decision Trees do not require this assumption (James, Witten, Hastie, & Tibshirani, 2013). However, these machine learning methods have complicated algorithms, possibly leading to long computation times (Ke et al., 2017). Therefore, we implement both a linear and a nonlinear model for the aggregated method to compare each model's advantages and disadvantages.

These choice and aggregated models should be able to replicate the TV ratings per segment as close as possible. Besides accuracy though, computation time is an important performance measure as well: datasets of TV viewing behavior are large, possibly leading to long computation times. Therefore, we compare these methods in terms of prediction accuracy and computational feasibility.

The choice model and aggregated model are applied to Nielsen TV panel data, obtained during four weeks in Chicago, US. For this data, both household-level viewing behavior and segment-level TV ratings are known.

The main objective of this research is therefore to answer the following research question:

> *What is the best method to estimate TV viewing behaviour of demographic segments from household level data?*

To answer this main research question, it is split up into several parts. First of all, we need to develop methodologies to obtain segment TV ratings from household TV viewing data. We develop two methodologies: a method estimating these segment ratings using a choice model, and a method estimating segment ratings from aggregated household TV ratings.

In the choice model, we use the observed household viewing behaviour and household characteristics to estimate utilities per channel for each demographic segment. Using these utilities, proportions

of TV ratings per demographic segment can be obtained, which are used to estimate the segment TV ratings.

The aggregated model takes a different approach. Here, the observed household viewing behaviour is aggregated to TV ratings per household type. Here, the household type is the number of individuals per segment in each household. These aggregated household TV ratings are used as explanatory variables in a regression model with the segment TV ratings as dependent variable. We use two different regression setups: a linear one using Least Squares, and a nonlinear one using Gradient Boosting Decision Trees. To optimize the hyperparameters of the Gradient Boosting Decision Tree model, Bayesian Hyperparameter Optimization is applied.

Next, the aforementioned methods should be compared. We do this according to two performance measures. The first is model fit: the implemented model should perform well at predicting segment TV ratings. The second is computational feasibility. the model should have an acceptable computation time.

To assess the suitability of the two models, as well as answer the main research question, we split up the main research question into four subquestions:

Q1: How can TV ratings per demographic segment be estimated using a choice model?

Q2: How can TV ratings per demographic segment be estimated directly from aggregated household TV viewing behavior?

Q3: Which model works best in terms of model fit?

Q4: Which model works best in terms of computational feasibility?

This thesis is organized as follows. Section 2 will shortly discuss relevant literature, and the contribution this research makes. Next, section 3 gives a brief introduction to the available data. Section 4 describes the methodology of choice models and utility specifications that are used. Furthermore, it contains a description of the aggregated model, which estimates a direct relation between segment TV ratings and aggregated household TV viewing behaviour. It also describes how the output of these models is used to estimate TV ratings per segment, and how model performance will be measured. Section 5 will describe the outcomes per model, as well as a comparison of each model's performance. For readability, only the main findings are presented in section 5, more figures are available in the appendices. Section 6 will answer the research questions mentioned above. Furthermore, it will also present some suggestions for further research in this field.

2 LITERATURE & BACKGROUND

This section will start with a review on analysis of TV ratings in section 2.1. Next, literature on the proposed methodologies is discussed. Some literature background of the choice model specification used in this research can be found in section 2.2. The background behind the techniques used for the aggregated model specification is summarized in section 2.3.

2.1 TV ratings analysis

The problem of estimating TV ratings per demographic segment using household level data has received limited attention in literature. However, there has been extensive research in the field of identifying TV audiences and segmentation of TV audience among channels, using individual data and choice models.

In their review of literature on the recent developments in TV viewing behaviour, Sharp et al. (2009) noted that in recent years, slightly more fragmentation among TV channels has occurred due to the increased number of channels. However, they also concluded that the major channels attract similar audiences from all segments of society. Only some smaller channels in their research have less varied audiences.

In agreement to the conclusions of Sharp et al. (2009), J. G. Webster (2005) found limited fragmentation in TV audiences for established channels. This is less the case for smaller, newer channels, which are often focused around a single theme. Even though these channels deliberately focus on one theme to attract a homogeneous audience, individuals from all demographic segments are still part of their audience.

Rust and Alpert (1984) used a choice model with utilities per TV program to estimate probabilities for individuals from different demographic segments to watch different channels over time. They proposed a model to estimate TV ratings per demographic segment as well. However, their focus was, contrary to the research proposed in this paper, on predicting TV viewing behaviour rather than evaluating TV ratings per segment. Furthermore, their model was based on TV viewing data from 1978 from Simmons Media Studies (1978) and due to the limited computer power of their era, the model was built on a limited dataset in terms of number of channels, time span and number of segments.

Meyer and Hyndman (2005) performed a ratings analysis study as well, studying somewhat similar data as in this research. However, Meyer and Hyndman (2005) used individual level data rather than household level data. Unlike our work, there was no uncertainty about the demographics of the viewers at each moment in time in their research. Furthermore, their focus was on predicting future TV ratings, whereas we focus on estimating actual TV ratings per segment.

This research can make a contribution to literature on analysis of TV ratings in a practical and academic manner. Knowing the TV ratings per demographic segment makes it easier for TV channels to identify their target groups, which is of great interest to their advertisers. Furthermore, this specific application of statistical choice model methods has until now received limited attention in literature. This research will try to combine statistical choice models based on observed individual behaviour and group utility methods that are used to estimate household utilities from individual utilities.

2.2 Choice model methodology

Extensive research has been done to estimating TV ratings using choice models. Among others, Meyer and Hyndman (2005), Rust and Alpert (1984), Rust, Kamakura, and Alpert (1992) and Shachar and

Emerson (2000) attempted to create choice models to predict viewing behaviour. However, none of these researches included the aggregation step from household to individual preferences. Hence, research performed so far is focused on predicting household level behaviour from household level viewing data or on predicting individual behaviour using individual data.

Choice models assume that each individual's gain from each possible choice (in this case the TV stations) is measurable and expressible as a single number, the utility (Heij, de Boer, Franses, Kloek, & van Dijk, 2004). Group choice models require that the utility for each choice for the entire group is expressed in a single number. However, extensive research has not lead to consensus in how this group utility should be calculated from individual utilities (Corfman & Gupta, 1993).

To be able to estimate a household choice model from individual utilities, a function $u_h = f(\omega_i)$ is needed that translates the individual's utility functions ω_i to household utility functions u_h. In literature, different options for this function are researched. The most well-known possibilities for $f(\omega_i)$ are the *additive*, *multiplicative*, *maximin* and *maximax* functions (Brock, 1980; Curry, Menasco, & Ark, 1991). The additive function takes the sum of each individual's utility as group utility for a specific option (Harsanyi, 1955), whereas the multiplicative takes the product of the individual utilities (Nash, 1950). The maximin and maximax functions take the maximum or the minimum, respectively, of the individuals' utilities as group utility (Rawls, 1971).

Each of the aforementioned group utility functions has received much attention in literature on group behaviour and utilitarianism. However, there has been limited attention to the application of these functions in choice models. Therefore, this research makes an academical contribution by assessing the suitability of each of these utility functions for use in econometric choice models.

2.3 Aggregated model methodology

In the aggregated model, a regression model is developed regressing the segment TV ratings on household TV viewing behaviour and several other explanatory variables (details will follow in section 4.3.3). Traditionally, regression models are estimated using linear regression, making the assumption of the existence a linear relation between the dependent variable and the predictors (Heij et al., 2004). Estimating these models analytically using least squares methods furthermore requires the assumption that the dependent variable is normally distrubuted as function of the independent variables.

To avoid the assumption of normality, various tree-based regression methods have been developed, such as Random Forests (RF) (Liaw & Wiener, 2002), Boosted Regression Trees (BRT) (Elith, Leathwick, & Hastie, 2008) and Gradient Boosting for Decision Trees (GBDT) (Friedman, 1999). Each of these tree-based regression models has very good overall predictive performance, although the predictive power of each method differs per case (Caruana & Niculescu-Mizil, 2006). Moreover, these ensemble methods are often more accurate in their predictions than linear regression. However, each of these models also have a major downside: they can be quite costly to estimate due to their complex nature (Ke et al., 2017).

In order to make these decision tree algorithms less costly, several propositions have been made. Specifically in the field of GBDT, many algorithms have been proposed to reduce computation time. The main bottleneck in the implementations of GBDT algorithms lies in the identification of the most useful split point in each tree (Ke et al., 2017). Therefore, several propositions have been made to make selection of split points quicker. Shafer, Agrawal, and Mehta (1996) attempted to speed up split point selection by enumerating all possible split points on the subset belonging to the branch where the split is to be made. However, their algorithm still appeared to be time- and memory-consuming. Another option is the histogram-based approach of Wu, Landgrebe, and Swain (1975), which selects split points by evaluating the shape of attributes' histograms. However, according to Perner and Trautzsch (1998), this method does not perform too well in terms of predictive power. Furthermore, in large datasets, the creation of the histograms takes up so much computation time that the algorithm is not very fast either (Ke et al., 2017).

Ke et al. (2017) have developed the LightGBM algorithm, combining two novel techniques to speed up GBDT. Firstly, they use Gradient-based One Side Sampling (GOSS), in which unimportant data instances are discarded when updating the gradient used to build new trees. Secondly, they use EFB (Exclusive Feature Bundling) to combine mutually exclusive variables into one variable to reduce dimensionality. Their algorithm has proven to be both efficient and accurate in several studies (Fonseca et al., 2017; Ke et al., 2017; Ma et al., 2018). However, like any decision tree algorithm, the LightGBM algorithm by Ke et al. (2017) contains a large number of hyperparameters that should be tuned by the user to obtain an optimal result.

To find the optimal hyperparameter settings of machine-learning algorithms like LightGBM, multiple possibilities exist. The first, most obvious method, is to manually select hyperparameters based on previous experience. This however either requires much manual labour or expert experience (Snoek, Larochelle, & Adams, 2012). A different option is grid search (J. Bergstra & Bengio, 2012), in which all possible combinations of a user-specified selection of values per hyperparameter are tried out. The downsides of this are that only these user-specified values are taken into account and that the model has to be trained a huge amount of times, once for each combination of hyperparameter values (J. Bergstra & Bengio, 2012). This number of evaluations increases exponentially as the number of hyperparameters increases (Bellman, 1961). A second option is therefore random search, in which only a subset of random combinations of the grid are tried out (J. Bergstra & Bengio, 2012). This is somewhat faster than grid search, but still requires many evaluations of the model to be trained. A third method is automated hyperparameter tuning. A common way to automate hyperparameter tuning is using Bayesian Optimization, in which a seperate probability model is created using the performance metric of the LightGBM algorithm as dependent and hyperparameter settings as independent variables (Snoek et al., 2012). This seperate probability model is much easier to optimize and therefore leads to the optimal hyperparameter setting using less evaluations of the model (Snoek et al., 2012; Thornton, Hutter, Hoos, & Leyton-Brown, 2013).

3 DATA DESCRIPTION

The data studied is from Nielsen Media Research and consists of three interconnected datasets: a dataset of household characteristics, a dataset of household viewing behaviour and a dataset containing the true segment TV ratings per segment. Together, these three datasets define the TV viewing behaviour of a TV panel in Chicago, Illinois, USA. The data was obtained during four weeks at the beginning of 2019: from January 3rd, 2019 at 4am until January 31st, 2019 at 4am. The data in the TV panel only contains data of adults, data of children younger than 18 years is not taken into account.

Besides the data of the Chicago TV panel, data is also available for the TV panel of Cleveland, Ohio, USA. This data has the same structure as the Chicago data. The Cleveland data serves to test the *out-of-market* predictive power of the aggregated model.

This section will describe the structure and descriptive statistics of the three datasets of Chicago. First, the composition data is described in section 3.1, followed by the viewing data in section 3.2 and the truth data in section 3.3. A brief overview of the Cleveland data is in section 3.4.

3.1 Composition

The first dataset describes the composition of the 904 households in the panel. For each household, information is known about its demographic composition: for each individual i belonging to a household h, the education level and demographic group (containing information on age and gender) is known. To make sure the panel of 904 households is a good representation of the actual US population, each individual has been assigned a weight w_i based on the number of similar individuals it represents. This way, individuals that are more typical for the US population or are less represented in the TV panel have larger weights than individuals that are rarer, or well represented in the TV panel. The same holds for households, each household has a weight w_h denoting the number of households it represents.

Each of the 904 households in the data contain a number of 1 up to 7 people aged 18 and above. The average number of people per household is 2.01, summing up to 1,820 individuals in total. As figure 1 shows, most households contain no more than two people. Only 194 households consist of three or more people (approximately 21.5% of the households). This is not surprising, as only household members of grown-up age are in the dataset.

The composition data, adjusted using the representation weights w_i, contains a fairly equal division of males and females: the percentage of females is slightly higher with 51.9%. As figure 2 shows, the observations are pretty evenly spread among the age groups as well. Furthermore, the spread of observations among education levels in 3 shows that the majority of individuals has education level around 4, with some outliers at 0. Unfortunately, as the definitions of the education levels are unknown, it is not possible to give an interpretation to the different education levels.

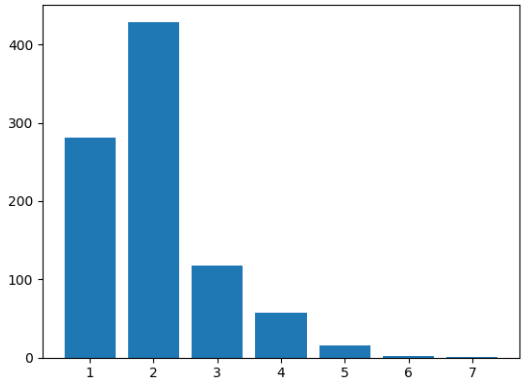


Figure 1: Histogram of the number of households (x-axis) per number of individuals per household (y-axis). The distribution shows a large skew to the left.

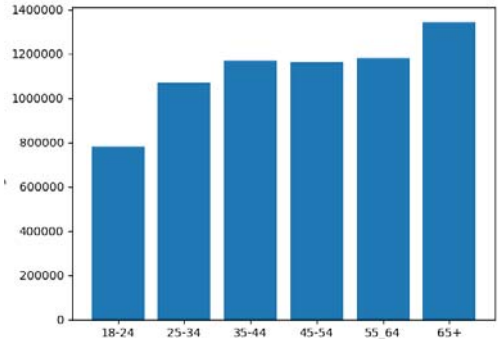


Figure 2: Weighted sum of observations in each age category

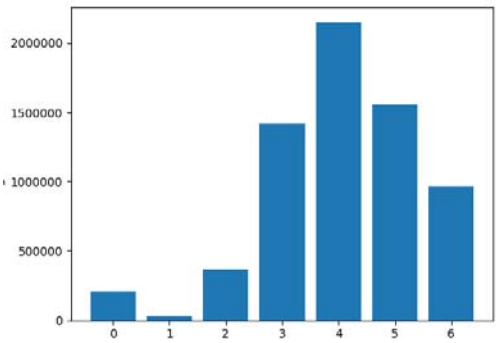


Figure 3: Weighted sum of observations in each education category

3.2 Household viewing

The second dataset contains the actual TV viewing behaviour of the households. Each observation stores three variables: the household ID (h), the channel (s) being watched and the date and time (t) at which channel s was watched. Time is measured in quarters of an hour; a TV viewing session is counted if channel s was watched during at least 5 of these 15 minutes. Therefore, each observation represents the viewing behaviour of one household h at a point in time t. In total, there are 1,127,953 observations of TV viewing moments in the viewing dataset.

As figure 4 shows, the number of viewing sessions per household varies greatly. The mean number of television viewings per household per month is 1,256. As the viewing sessions are measured per 15 minutes, this would imply that the average household watches TV 10 hours per day. However, this number is likely to be much lower, as a TV watching session is already counted when the household has watched a certain channel for 5 of the 15 minutes. Furthermore, the distribution of the amount of time spent watching TV is heavily skewed to the left. This implies that there is a large number of light

TV users and a smaller number of very heavy TV users. This is in agreement with Sharp et al. (2009), who concluded that in general, the small group of heavy TV viewers cause the mean TV time to be larger than expected.

Furthermore, figure 4 also shows a number of observations for which the number of TV viewing sessions is larger than 2,688, the number of quarter hours in the observation period. This can be explained in two ways. First of all, as mentioned before, a TV viewing session is counted if in 15 minutes at least 5 were spent watching a specific channel. Therefore, up to three viewing sessions can be created in one quarter hour. Additionally, it is possible that households possess more than one TV set, therefore watching multiple channels at the same time.

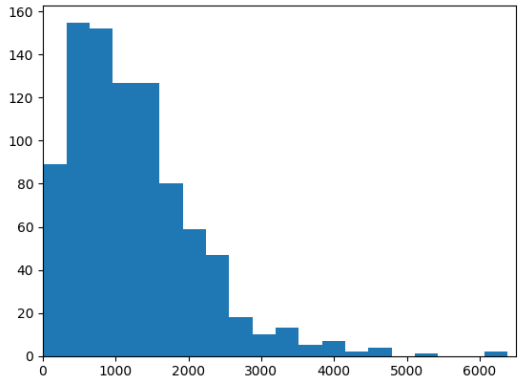


Figure 4: Distribution of the number of TV viewing sessions per household.
x-axis: no. viewed quarter hours.
y-axis: no. households.

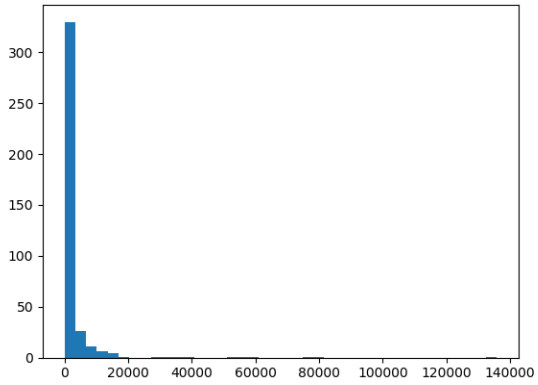


Figure 5: Distribution of TV viewing sessions per channel.
x-axis: no. quarter hours.
y-axis: no. channels

As figure 5 shows, the popularity of channels varies largely. The mean of the number of times a channel has been watched during the data collection period is approximately 2,900. However, this mean is influenced largely by a small number of very popular channels, having as much as 135,000 views. More than half the channels can be seen as small channels, having less than 250 views in total. This may impose some issues in estimation of models, as very little information is available for these channels. For approximately 25% of the channels, less than 10 observations are available. Especially for these channels, estimation of TV ratings can be hard due to lack of data.

As figure 6 shows, the number of households watching TV varies greatly over time. There is however a *seasonal pattern* visible: During prime time (evening hours), almost 700 of the 904 households are watching TV. During the night and early morning, this number is just approximately 150. This can cause trouble in model estimation, as for these morning hours, limited data is available.

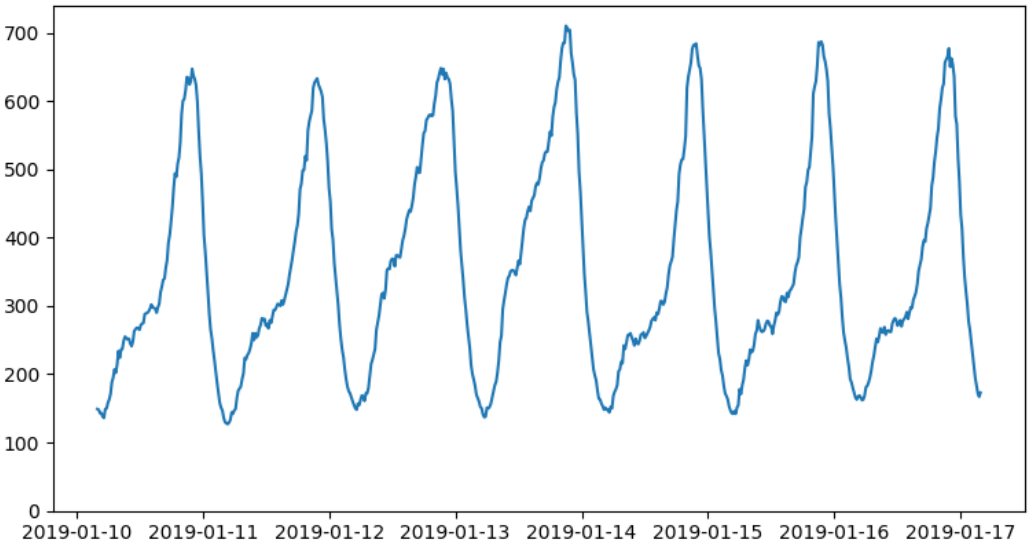


Figure 6: Number of households watching TV over time

3.3 Truth

The last dataset contains the actual TV viewing numbers of each channel s at time t per segment a of demographic group and education level. This is the dataset that is aimed to be reproduced using the first two datasets. The total viewing number v_{ast} of channel s for each segment a is defined as the sum of the weights of all individuals belonging to segment a who watched channel s at time t. Mathematically, this can be displayed as

$$v_{ast} = \sum_{i \in a} w_i I\left[y_{it} = s\right], \tag{1}$$

where $I[\cdot]$ is the indicator function. Hence, v_{ast} can be seen of the sum of the weights of the people in segment a who watched channel s at time t.

3.4 Cleveland data

Besides these four datasets of a TV panel in Chicago, data of a TV panel in Cleveland, US is available as well. This data will be used to assess the extent to which the models developed on the Chicago data are applicable to data from other regions. The data for Cleveland again consists of a composition, viewing, household weight and truth dataset. The Cleveland panel is somewhat smaller than the Chicago panel: it consists of 534 households with in total 968 people. The total number of TV viewing sessions is 723,699, spread over a number of 330 channels. 147 of these channels are local Cleveland channels, that are not available to TV viewers in Chicago.

4 METHODOLOGY

This section describes the methodology that is used to estimate TV ratings per demographic segment. First, section 4.1 discusses the way segmentation is imposed. Then, section 4.2 describes the choice models and the different household utility specifications that are used. Furthermore, it will also show the calculations needed to transform the choice model outcome to TV ratings per demographic segment. Section 4.3 explains the model specification estimating segment ratings directly using aggregated household TV ratings. It presents two techniques for this model: an analytic, linear solution and a nonlinear solution based on decision tree algorithms. At last, the measures used to validate and compare the results of each model are introduced in section 4.4.

4.1 Segment definition

Using the individual level demographic characteristics on education, gender and age, it is possible to define a segment for each possible combination of these. As the number of categories of the gender, education and age variables is two, seven and eight respectively, the total number of segments would then be 112. Theoretically, defining the segments as all possible combinations of demographic variables will lead to estimates of TV ratings for very specific demographic segments. However, splitting the data into 112 segments will mean that the number of observations per segment can be very small, especially for the segments of people that are less represented in our data. This would likely reduce the accuracy of the models.

Therefore, we initially impose segmentation per (group of a) demographic variable. Hence, of the variables age, gender and education, only one variable is used to segmentate the data. This implies that when segmenting on gender, the number of segments in the data is equal to two. When segmenting on age or education, we combine some categories to obtain three segments. Based on the results of the models discussed in section 5, the most applicable method can be selected for each type of segmentation. Furthermore, based on the results of the different models, the demographic property that has the clearest distinction in TV viewing behaviour can be identified as well.

The models will also be tested on data segmented on both age and gender, to make a judgement on how the models perform on a larger number of segments.

4.2 Choice model

The choice of TV channel in a household can be modelled with a choice model. This choice model aims to estimate the probability household h is watching channel s at time t, denoted as $P\left(y_{hst}=1\right)$. These probabilities can be estimated using utilities u_{hst} of household h for channel s at time t. As Heij et al. (2004) described, the probability that household h prefers channel s over channel r can be defined as the probability of the utility of channel s being higher than the utility of channel r. Let S be the total number of channels households can select from. Then, $P\left(y_{hst}=1\right)$ can be estimated using (2) (Luce, 1959).

$$
\begin{aligned}
P\left(y_{hst}=1\right) &= P\left(u_{hst} > u_{hrt} \;\; \forall r \neq s\right) \\
&= \frac{u_{hst}}{\sum_{r=1}^{S} u_{hrt}}
\end{aligned}
\tag{2}
$$

As any statistical model, using the choice model implies making a few assumptions. The most important assumptions of the choice model are the independence of irrelevant alternatives and the identically and independently distributed errors (Solgaard & Hansen, 2003).

The assumption of independence of irrelevant alternatives means that the choice between two options is unaffected by the other options (Domenich & McFadden, 1975). In the case of TV channel selection, this implies that the choice of a household between two channels s and r is independent of all other channels. This assumption should generally hold for TV channel selection, as it is intuitively unlikely that an individual's choice between channels s and r is affected by a third channel q.

The assumption of identically and independently distributed error terms implies that individuals are relatively homogeneous, in the sense that two individuals with the same characteristics have the same choice behaviour (Solgaard & Hansen, 2003). In the sense of the TV channel choice model, this means that two individuals i and j frome the same demographic group have the same choice probabilities for each channel. This is a strong assumption: it is very likely that two individuals with the same characteristics contain *unobserved heterogeneity*, causing their preferences to be different. For model simplicity though, we assume that individuals are homogeneous within segments.

Viewing preferences are likely to differ over time. For example, viewers interested in sports are likely to obtain higher utility from watching a sports channel in the evenings and weekends, as major games are played during these moments. As viewing behavior changes over time, the choice model should be able to capture dynamic behavior. Therefore, the choice model is estimated per time step, such that household utilities can change over time. Hence, for each time t between 1 and T, the choice model in (2) is estimated separately.

Theoretically, the choice model has two major downsides. Firstly, the fact that the model is estimated separately for each quarter hour may cause some computation issues. The data available covers viewing behaviour of four weeks. Estimating the choice model for this full month of data implies estimating the model 2,688 times. Secondly, as figure 6 shows, there is a large number of quarter hours with only a small number of households watching. This is mostly the case for the nightly quarter hours, when most people are asleep. For these quarter hours, few observations are available to estimate the model on. This data limitation is likely to reduce model accuracy.

4.2.1 Household utility specifications

The model in (2) assumes that each household has a single utility u_{hst} for each stattion s from which the choice for a specific channel is made. However, in this case, the decision is made by the members of the household h. To model this household decision, we assume that the household utility is a

function of the utilities of the individuals in this household.

This section therefore describes several methods that can be used to obtain household utilities for a channel s from individual level utilities. For simplicity, we assume that utilities for individuals from the same segment a are equal. Furthermore, we assume that all individuals' utilities in a household h are equally important in the TV channel selection process.

There are several possibilities to estimate household utilities u_{hst} from the segment utilities of the individuals in a household ω_{ast} and the number of individuals from segment a in a household n_{ah}. As Curry et al. (1991) denoted, most estimation methods fall in one of two major categories. The first category consists of additive utility functions, also called Harsanyi solutions, named after its initial proposer, Harsanyi (1955). The second category consists of multiplicative utility functions, synonymous for Nash solutions because they satisfy the conditions in Nash (1950) and Nash (1953). The remainder of this subsection briefly describes these methods and shows how they can be implemented in a choice model.

The Harsanyi solution is the most straightforward type of group utility function. It simply takes the sum of all individuals' utilities as household utility (Harsanyi, 1955). Let n_h denote an $A \times 1$ vector where the ath element, n_{ah}, denotes the number of individuals from segment a in household h, and let ω_{st} denote the $A \times 1$ vector with the utilities ω_{ast} as elements. The household utility can then be estimated using (3).

$$u_{hst} = \omega'_{st} n_h \tag{3}$$

The Nash solution uses a multiplicative function to estimate household utilities out of its individuals' utilities. It can be denoted as the product of the individual utilities. Therefore, the household utility u_{hst} is equal to the product of all segment utilities ω_{ast}, to the power n_{ah}, the number of individuals per segment in household h. It's mathematical definition is stated in (4).

$$u_{hst} = \prod_{a=1}^{A} (\omega_{ast})^{n_{ah}} \tag{4}$$

Besides the Nash and Harsanyi solutions, other approaches exist where not all individuals' utilities are used in estimating the household utility. The most well known is the *Maximin* rule by Rawls (1971), which maximizes the minimum of individual's utilities. Its mathematical representation is displayed in (5), where $I[\cdot]$ denotes the indicator function. The rationale behind this rule is that a decision is made that is best for the worst off individual, in order to make sure every individual has at least some level of utility (Brock, 1980). In terms of a household decision for a TV channel, this can be seen as the decision to watch a certain channel that not everybody particularly likes, but nobody dislikes either.

$$u_{hst} = \min_{a} \left(\omega_{ast} I[n_{ah} > 0] \right) \tag{5}$$

Similar to the Maximin rule, we can also apply the *Maximax* rule in (6), which takes the maximum

of a household's individual utilities as household utility (Rawls, 1971). This can be seen as a household to choose a TV channel because one individual really wants to see it, even though some others may dislike it.

$$u_{hst} = \max_{a} \left(\omega_{ast} I \left[n_{ah} > 0 \right] \right) \tag{6}$$

To assess which method works best when estimating demographic segment TV ratings, we use all four methods described above used and compare their performance, to assess which group decision method is most appropriate in TV ratings estimation.

4.2.2 Estimation with Maximum Likelihood

Substituting the methodology in section 4.2.1 in the choice model of (2), we obtain a model that can be used to estimate individual utilities. Under the assumption that individuals from the same segment of demographic group and education have equal preferences, we can see each individual's utility as the utility of its segment: ω_{ast}. Each household utility can then be estimated using the estimation techniques in 4.2.1 and the segment utilities ω_{ast}.

Then, for the additive household utility estimation method of Harsanyi, the choice model to be estimated is as follows:

$$P\left(y_{hst} = 1\right) = \frac{\omega'_{st} n_h}{\sum_{r=1}^{S} \omega'_{rt} n_h}. \tag{7}$$

Similarly, using the Nash method to estimate household utilities multiplicatively, the choice model to be estimated is as follows:

$$P\left(y_{hst} = 1\right) = \frac{\prod_{a=1}^{A} \left(\omega_{ast}\right)^{n_{ah}}}{\sum_{r=1}^{S} \prod_{a=1}^{A} \left(\omega_{art}\right)^{n_{ah}}}. \tag{8}$$

For the Minimax and Maximax approaches, the probabilities are estimated as below in (9) and (10), respectively.

$$P\left(y_{hst} = 1\right) = \frac{\min_a \left(\omega_{ast} I \left[n_{ah} > 0 \right] \right)}{\sum_{r=1}^{S} \min_a \left(\omega_{art} I \left[n_{ah} > 0 \right] \right)} \tag{9}$$

$$P\left(y_{hst} = 1\right) = \frac{\max_a \left(\omega_{ast} I \left[n_{ah} > 0 \right] \right)}{\sum_{r=1}^{S} \max_a \left(\omega_{art} I \left[n_{ah} > 0 \right] \right)} \tag{10}$$

The values of the parameters ω_{ast} are estimated using Maximum Likelihood, treating each moment in time t as a separate model. The likelihood function of (2) can be written as

$$L\left(p_{hst}, y_{hst}\right) = \prod_{h=1}^{H} \prod_{s=1}^{S} \left(p_{hst}\right)^{y_{hst}} \left(1 - p_{hst}\right)^{\left(1 - y_{hst}\right)}, \tag{11}$$

where $P\left(y_{hst}=1\right)$ is denoted as p_{hst}. Again, y_{hst} is equal to 1 if household h watched channel s at time t, and zero else, and H is the total number of households.

To maximize likelihood, it is often more convenient to minimize the negative log-likelihood. Taking the log of (11) and multiplying by -1, we obtain the following objective:

$$\arg\min -\sum_{h=1}^{H}\sum_{s=1}^{S} y_{hst}\log\left(p_{hst}\right)+\left(1-y_{hst}\right)\log\left(1-p_{hst}\right) \tag{12}$$

Substituting (7) in (12), the actual objective in case of additive household utilities becomes:

$$\arg\min -\sum_{h=1}^{H}\sum_{s=1}^{S} y_{hst}\log\left(\frac{\omega'_{st}n_h}{\sum_{r=1}^{S}\omega'_{rt}n_h}\right)+\left(1-y_{hst}\right)\log\left(1-\frac{\omega'_{st}n_h}{\sum_{r=1}^{S}\omega'_{rt}n_h}\right) \tag{13}$$

The log-likelihood minimization objectives for different household utility specifications can be obtained in a similar way.

The log-likelihood in (13) is complicated and its minimum can not be obtained analytically. Therefore, the optimization process is done numerically, using the L-BFGS-B (Limited memory Broyden–Fletcher–Goldfarb–Shanno with Bounds) algorithm by Zhu, Byrd, Lu, and Nocedal (1997).

As the log-likelihood in (13) is a complex and non-convex function, there is no guarantee that the optimum found by the L-BFGS-B algorithm is the global optimum. The same holds for the log-likelihoods of the other utility specifications. Therefore, it is difficult to judge whether a found solution is optimal or not. We attempt to resolve this issue by using different random initializations of the segment utilities in ω_{ast} as starting values for the optimization algorithm, in order to find the lowest minimum and to identify which utility specification has the most stable model fit.

Using the estimated segment utilities $\hat{\omega}_{ast}$, we can obtain estimated probabilities $\hat{\pi}_{ast}$ for individuals from segment a to watch channel s at time t, using (14).

$$\hat{\pi}_{ast}=\frac{\hat{\omega}_{ast}}{\sum_{r=1}^{S}\hat{\omega}_{art}} \tag{14}$$

4.2.3 Multinomial logit model

Besides estimating the vector of segment utilities ω_{st} directly in (13), an approach defining ω_{st} as a linear function of the characteristics of individuals in segment a is used. The advantage of this approach is the reduced number of parameters: when estimating the utilities as a function of a segment's characteristics x_a as in (15), the numbers of parameters is reduced from AS to $4S$.

$$\begin{aligned}\omega_{ast} &= x'_a\beta_{st}\\ &= \beta_{0st}+\beta_{1st}\text{Education}_a+\beta_{2st}\text{Gender}_a+\beta_{3st}\text{Age}_a\end{aligned} \tag{15}$$

When using the utility specification in (15), it is more appropriate to use the Multinomial Logit (MNL) model of McFadden et al. (1973). The simple reason for this, is that (15) could yield negative

values of ω_{ast}, leading to negative probability estimates. The MNL model uses exponentials to handle negative utilities.

The additive household utility specification of (3), consisting of individual utilities according to (15) can be rewritten by creating a matrix X, of which the ath row contains the characteristics of an individual in segment a:

$$u_{hst} = \omega'_{st} n_h = \beta'_{st} X' n_h. \tag{16}$$

Using this specification of u_{hst}, we can rewrite (2) to the following:

$$\begin{aligned} p_{hst} &= \frac{\exp\{u_{hst}\}}{\sum_{r=1}^{S} \exp\{u_{hrt}\}} \\ &= \frac{\exp\{\beta'_{st} X' n_h\}}{\sum_{r=1}^{S} \exp\{\beta'_{rt} X' n_h\}}. \end{aligned} \tag{17}$$

Similarly, the log-likelihood in (13) can be rewritten using the utility specification of (15). The log-likelihood of the additive MNL model is then as follows:

$$\ell(\beta) = \sum_{h=1}^{H} \sum_{s=1}^{S} y_{hst} \log\left(\frac{\exp\{\beta'_{st} X' n_h\}}{\sum_{r=1}^{S} \exp\{\beta'_{rt} X' n_h\}}\right) + (1 - y_{hst}) \log\left(1 - \frac{\exp\{\beta'_{st} X' n_h\}}{\sum_{r=1}^{S} \exp\{\beta'_{rt} X' n_h\}}\right). \tag{18}$$

Rearranging terms and taking the denominator out of the first logarithm (which is independent s) out of the summation over channels, we obtain the following simplified log-likelihood:

$$\begin{aligned} \ell(\beta) = \sum_{h=1}^{H} \Bigg[&\sum_{s=1}^{S} \left(y_{hst} \beta'_{st} X' n_h\right) - \log\left(\sum_{r=1}^{S} \exp\{\beta'_{rt} X' n_h\}\right) \\ &+ \sum_{s=1}^{S} (1 - y_{hst}) \log\left(1 - \frac{\exp\{\beta'_{st} X' n_h\}}{\sum_{r=1}^{S} \exp\{\beta'_{rt} X' n_h\}}\right)\Bigg]. \end{aligned} \tag{19}$$

The log-likelihoods for other household utility specifications can be derived in a similar way. Using the estimated $\hat{\beta}_{st}$, we obtain estimated probabilities for individuals to watch channel s from each segment using equation 20.

$$\widehat{\pi}_{ast} = \frac{\exp\{x'_a \hat{\beta}_{st}\}}{\sum_{r=1}^{S} \exp\{x'_a \hat{\beta}_{rt}\}} \tag{20}$$

4.2.4 Derivation of TV ratings per segment

Using the output of the choice model, the estimated ratings $\hat{v}_{ast}$ can be obtained in two steps. In the first step, the TV ratings per segment from single-person households are calculated. As we know the demographic characteristics of these people and know they are the only ones in the households,

and therefore have to be the people watching, the segment TV ratings from single-person households are not stochastic. In the second step, the segment TV ratings from households with more than one person are estimated using the choice model probabilities.

The estimated TV ratings $\hat{v}_{ast}$ can therefore be decomposed as

$$\hat{v}_{ast} = v^{(s)}_{ast} + \hat{v}^{(m)}_{ast}, \tag{21}$$

where $v^{(s)}_{ast}$ denotes the TV ratings from single person households and $\hat{v}^{(m)}_{ast}$ denotes the estimated ratings from mixed households.

The single ratings $v^{(s)}_{ast}$ can then, as first step, be calculated as the sum of all the weights of individuals in single person households, watching channel s at time t and with the characteristics of segment a. Mathematically, this is displayed as

$$v^{(s)}_{ast} = \sum_{h=1}^{H} y_{hst} n_{ah} \tau_h \zeta_h, \tag{22}$$

where τ_h is a binary variable indicating whether household h is a single person household or not:

$$\tau_h = \begin{cases} 1 & \text{if } \sum_{a=1}^{A} n_{ah} = 1 \\ 0 & \text{else} \end{cases} \tag{23}$$

and ζ_h is equal to the individual weight w_i of the individual in household h if household h is a single person household, and zero else:

$$\zeta_h = \begin{cases} w_i & \text{if } i \in h \text{ and } \tau_h = 1 \\ 0 & \text{else} \end{cases} \tag{24}$$

Using the numbers of $v^{(s)}_{ast}$, we can also find the total TV ratings from single person households $\psi^{(s)}_{st}$, as

$$\psi^{(s)}_{st} = \sum_{a=1}^{A} v^{(s)}_{ast}. \tag{25}$$

Using (25) we can obtain $\psi^{(m)}_{st}$, the total TV ratings from mixed households, using (26), where ψ_{st} is the total TV ratings for channel s at time t. This $\psi^{(m)}_{st}$ can be used in the next step, where we estimate $\hat{v}^{(m)}_{ast}$.

$$\psi^{(m)}_{st} = \psi_{st} - \psi^{(s)}_{st} \tag{26}$$

In the second step, we estimate $\hat{v}^{(m)}_{ast}$ using the fact that the sum of $\hat{v}^{(m)}_{ast}$ over a should be equal to $\psi^{(m)}_{st}$. We can then estimate each $\hat{v}^{(m)}_{ast}$ using the number of individuals in segment a, denoted as m_a, estimated probabilities $\hat{\pi}_{ast}$ from (14) or (20) and a tuning parameter δ_{st}, which makes sure the sum

of $\hat{v}_{ast}^{(m)}$ aligns $\psi_{st}^{(m)}$. We achieve this by estimating the value of the parameter δ_{st} using (27).

$$\begin{aligned} \psi_{st}^{(m)} &= \sum_{a=1}^{A} \hat{v}_{ast}^{(m)} \\ \psi_{st}^{(m)} &= \delta_{st} \sum_{a=1}^{A} \hat{\pi}_{ast} m_a \\ \delta_{st} &= \frac{\psi_{st}^{(m)}}{\sum_{a=1}^{A} \hat{\pi}_{ast} m_a} \end{aligned} \tag{27}$$

In (27), the parameter m_a is the total number of individuals in segment a and can be obtained as the sum of the weights w_i of all individuals in segment a:

$$m_a = \sum_{i \in a} w_i. \tag{28}$$

Using the estimated probabilities $\hat{\pi}_{ast}$, the number of individuals in the segment m_a and the tuning parameter δ_{st}, we can obtain estimates of the mixed segment TV ratings $\hat{v}_{ast}^{(m)}$:

$$\hat{v}_{ast}^{(m)} = \delta_{st} \hat{\pi}_{ast} m_a. \tag{29}$$

Finally, we add up the calculated $v_{ast}^{(s)}$ and estimated $\hat{v}_{ast}^{(m)}$ to obtain $\hat{v}_{ast}$:

$$\hat{v}_{ast} = v_{ast}^{(s)} + \hat{v}_{ast}^{(m)}. \tag{30}$$

4.3 Aggregated model

To overcome the two limitations of the choice model (long computation time and limited data in some quarter hours), a second method is developed. This different approach is to model the relation between the segment ratings v_{ast} from the truth set, and the observed household viewing behavior directly in a regression model. This method will be denoted as the *aggregated model*. The aggregated model has as advantage over the choice model, that all quarter hours can be modelled at once. Hence, only one model based on a very large number of observations needs to be estimated. This will likely be computationally lest costly than the choice model. However, other than the choice model, the aggregated model does require the availability of true segment TV ratings in at least a part of the data.

This section describes two methods for this approach: a linear model estimated with Least Squares and a nonlinear tree-based method. Before describing the specifics of these two methods, the basics of the aggregated method itself are introduced.

4.3.1 Basics of aggregated method

In the aggregated method, the segment TV ratings v_{ast} are attempted to be calculated directly from aggregated viewing behaviour in the data, without using intermediate steps of individual probabilities.

Therefore, the goal is to find a function $f(\cdot)$, that calculates the segment TV ratings from a number of predictors. The predictors included in this function $f(\cdot)$ can be obtained from the viewing and composition datasets. The predictors used are the household type TV ratings z_{jst}, the day and time d_t, the percentage of the population that is part of segment a, denoted as q_a, the total ratings of channel s at time t, denoted as ψ_{st}, and the percentage of all individuals in segment a living in a household of type j, denoted as c_{aj}. Hence, in the aggregated model, the goal is to find a function $f(\cdot)$ that best approximates (31).

$$v_{ast} = f\left(z_{jst},\, d_t,\, c_{aj},\, q_a,\, \psi_{st}\right) \tag{31}$$

The predictor z_{jst} is the total TV rating of channel s at time t of households from a certain type j. It can be seen as the total number of people living in a certain household type, where at least someone in their household is watching channel s at time t.

The household types are chosen to be *Mutually Exclusive* and *Collectively Exhaustive* (MECE) and are based on the variable chosen to do segmentation on. All combinations of people from each segment present in the data are considered a separate household type j. To illustrate this, we will discuss an example. In case segmentation is done on gender, each individual belongs to either segment 0, of females, or segment 1, of males. Every possible combination of number of males and females in a single household present in the data is then used as a separate household type.

Intuitively, one would define z_{jst} as the number of individuals in households of type j watching channel s at time t. However, as each individual represents a larger number of similar individuals not included in the TV panel, each household also represents a larger number of similar households not included in the TV panel. Therefore, the TV ratings per household type z_{jst} are calculated through the weights of individuals in the households. Let the household weight of household h be the sum of the weights of its individuals $\left(\sum_{i\in h} w_i\right)$. Furthermore, let g_h be a vector of size J where the jth item is equal to one if household h is of type j, and zero else. Then, the observed TV ratings for household type j for channel s at time t is the sum of the weights of all households in j that watch s at t. The quantities of z_{jst} are obtained using (32).

$$z_{jst} = \sum_{h=1}^{H} \left(y_{hst} g_{hj} \sum_{i\in h} w_i \right) \tag{32}$$

The variable c_{aj}, denoting the percentage of individuals in segment a that live in households of type j, is then specified as the sum of weights w_i of individuals belonging to segment a and living in a household of type j, divided by the total number of individuals in segment a, m_a, as displayed in (33).

$$c_{aj} = \frac{\sum_{h=1}^{H} g_{hj} \sum_{i\in h \cap i\in a} w_i}{m_a} \tag{33}$$

The aggregated model is trained on the last three weeks of data of Chicago, using the first week from Chicago as validation set and the data from Cleveland as test set. This split into three sets of data

is made to be able to test for two types of fit: *within-market* fit and *out-of-market* fit.

Using the model estimated on the last three weeks of Chicago, *within-market* estimates can be made for the first week of segment TV ratings. The *within-market* fit is then accuracy of these estimates for the first week of Chicago data. It is a measure of how well the model can describe the relation between segment TV ratings and the predictor.

The *out-of-market* fit is the accuracy of the estimated segment TV ratings for the Cleveland data, again obtained using the model estimated on the last three weeks of Chicago data. By comparing these to the actual segment ratings in Cleveland, we can assess the *out-of-market* fit. This way, we can measure model fit in two ways. Firstly, we can asses how the model fits on unseen data, and check whether the model has not been overfitted on our training data. Secondly, we can test whether the aggregated model is applicable across markets. The intended application of the methods developed in this research is to be able to estimate segment TV ratings on RPD data, which comes from a different source than the panel data used to estimate these models. Therefore, it is important to measure whether the model estimated on the panel data is applicable on data from different sources or different markets as well.

The applicability of the aggregated method on different markets depends on the overlap between the household types. The AM is estimated on one market, yielding parameters for the variables (household types) present in this market. Then, these parameters are applied on a test market dataset to obtain out-of-sample estimates of the segment TV ratings. In case a large proportion of the household types in this training market is not present in the test market, the model is very likely to produce less accurate ratings for the test market, as it is missing many variables there. The other way around, a large number of missing household types in the test set has adverse effects as well: in that case, there is information in the test set that is not used in the estimation of household types.

As figures 12a - 12d and table 11 in appendix A display, the overlap between household types present in both the Chicago and Cleveland data is dependent on the number of segments used. When segmenting only on gender, there are 20 distinct household types in Chicago. Of these 20 types, 15 (75%) are also present in Cleveland, which also contains one household type that is not in the Chicago data. When imposing both gender and age as segments, yielding 6 segments, the overlap reduces to 52% of the household types in Chicago.

Whereas there are a large number of variables in Chicago missing in Cleveland, this is less the case the other way around. This can largely be explained by the fact that the Cleveland data is smaller than the Chicago data in terms of number of households and individuals. As 12a - 12d and table 11 in appendix A show, the number of household types existent in Cleveland but non-existent in Chicago is only one when segmentation is done on gender. For age and education segments, all household types apparent in Cleveland are also apparent in Chicago. Hence, when imposing segmentation on age or education, all information in the Cleveland data can be used. This is less the case when segmentation is done on gender and age at the same time. Then, for a number of 6 segments, 13 of 65 household types in Cleveland are missing in Chicago, an overlap of 80%.

The aggregated method has one major downside over the choice model. Because it uses parameters estimated on one TV market to make predictions on another market, it assumes that the function $f(\cdot)$ is the same across markets. Therefore, it can only be applied to markets of which similar markets exist, for which individual-level TV viewing behaviour is available. In this research, the assumption of $f(\cdot)$ being similar between the TV viewing markets is likely to hold, as the cities from which the data are taken (Chicago and Cleveland) are both northern US cities, approximately 500 kilometres apart.

4.3.2 Linear aggregated model

Under the assumption of linearity and normality, the relation between the household type ratings z_{jst} and the actual ratings v_{ast} can be modelled as in (34). The term $q_a\psi_{st}$ is the 'benchmark' estimation of v_{ast}: it is simply the population mean of TV ratings at time t for channel s. The remainder of (34) measures the deviation from the population mean of the segment TV ratings. The term ϵ_{ast} is the error term in the model as in traditional regression models. As in any regression model, it is assumed to be identically and independently normally distributed. The parameter α_a is the segment-specific intercept. Furthermore, the γ_{aj} parameters capture the impact of the TV ratings of households of type j on the TV ratings of demographic segment a.

$$v_{ast} = q_a\psi_{st} + (1-q_a)\left(\alpha_a + \sum_{j=1}^{J}\gamma_{aj}c_{aj}z_{jst}\right) + \epsilon_{ast} \tag{34}$$

The model in (34) is a basic linear representation of a relation between TV ratings per household type and TV ratings per demographic segment. It has a limited number of parameters to be estimated: only $J+1$ per segment. However, the model in (34) can not capture any dynamic behavior because of the time-invariant parameters γ_{aj} and α_a. As TV viewing behavior is likely to be different for different times of the day, we should extend the model to be able to capture this dynamic behaviour.

As figure 6 showed, TV viewing behavior clearly shows trend-wise behaviour. This trend can be represented in (34) by incorporating the term d_t, as displayed in (35). This term d_t is a vector of binary variables, indicating what day of the week and what part of the day it is at time t.

$$v_{ast} = q_a\psi_{st} + (1-q_a)\left(\alpha_a + d_t'\beta_a + \sum_{j=1}^{J}\gamma_{aj}c_{aj}z_{jst}\right) + \epsilon_{ast} \tag{35}$$

The aggregated approach has a computational advantage over the choice model approach. Firstly, it can be estimated with a large number of methods, such as simple traditional methods as Ordinary Least Squares (OLS). Secondly, the number of parameters to estimate is just the number of segments times the number of parameters per segment, or $A(J+L+1)$, where L is the length of d_t. Lastly, because of the term c_{aj} in (35), the model estimated on TV panel data can directly be used to estimate the unobserved segment ratings of RPD data as well.

To estimate the model in (35), we rewrite it such that it has the form of a general linear model:

$$\begin{aligned} v_{ast} &= q_a \psi_{st} + (1-q_a)\left(\alpha_a + \sum_{l=1}^{L} \beta_{al} d_{lt} + \sum_{j=1}^{J} \gamma_{aj} c_{aj} z_{jst}\right) + \epsilon_{ast} \\ \frac{v_{ast} - q_a \psi_{st}}{1-q_a} &= \alpha_a + \sum_{l=1}^{L} \beta_{al} d_{lt} + \sum_{j=1}^{J} \gamma_{aj} c_{aj} z_{jst} + \eta_{ast}, \end{aligned} \tag{36}$$

where $\eta_{ast} = \epsilon_{ast}/(1-q_a)$. Rewriting this in different terms yields

$$e_{ast} = x'_{ast}\theta_a + \eta_{ast}, \tag{37}$$

where the terms e_{ast}, x_{ast} and θ_a are defined as

$$\begin{aligned} e_{ast} &= \frac{v_{ast} - q_a \psi_{st}}{1-q_a}; \\ x_{ast} &= \begin{pmatrix} 1 & d_{1t} & \cdots & d_{Lt} & c_{a1} z_{1st} & \cdots & c_{aJ} z_{Jst} \end{pmatrix}'; \\ \theta_a &= \begin{pmatrix} \alpha_a & \beta_{a1} & \cdots & \beta_{aL} & \gamma_{a1} & \cdots & \gamma_{aJ} \end{pmatrix}'. \end{aligned} \tag{38}$$

Stacking all observations of all stations of a single segment, for all moments in time t on top of each other, we obtain the following, regular vectorized OLS function:

$$e_a = X_a \theta_a + \eta_a. \tag{39}$$

The estimate $\hat{\theta}_a$ is now just the regular OLS estimate:

$$\hat{\theta}_a = \left(X'_a X_a\right)^{-1} X'_a e_a. \tag{40}$$

Using this $\hat{\theta}_a$, out-of-sample predictions $\hat{e}_{ast}$ can be obtained as $\hat{e}_{ast} = x'_{ast}\hat{\theta}_a$. These estimates of $\hat{e}_{ast}$ can then be transformed to estimated TV ratings $\hat{v}_{ast}$ using (41).

$$\hat{v}_{ast} = q_a \psi_{st} + (1-q_a)\,\hat{e}_{ast} \tag{41}$$

We know the sum of the estimated segment TV ratings over segments should be able to the total TV ratings:

$$\sum_{a=1}^{A} \hat{v}_{ast} = \psi_{st}. \tag{42}$$

Therefore, for each channel and moment in time, the estimated segment ratings $\hat{v}_{ast}$ should be adjusted such that (42) holds. Similar to (27), this is achieved by multiplying each predicted rating with a multiplication factor δ_{st} to obtain corrected estimates $\hat{v}^*_{ast}$:

$$\hat{v}^*_{ast} = \delta_{st}\,\hat{v}_{ast}, \tag{43}$$

where

$$\delta_{st} = \frac{\psi_{st}}{\sum_{a=1}^{A} \hat{v}_{ast}}. \tag{44}$$

When estimating the model using (40), we make a few assumptions. First, we assume that the relation between the deviation of the mean tv ratings e_{ast} and the predictors is linear. Furthermore, estimating the model using least squares implies that the residuals η_{ast} are identically and independently normally distributed with mean zero and variance σ_a^2: $\eta_{ast} \sim N\left(0,\, \sigma_a^2\right)$ (Heij et al., 2004). To test whether this assumption holds, we perform a Jarque-Bera test (Jarque & Bera, 1980).

The Jarque-Bera test statistic is calculated as

$$JB = \frac{n-k+1}{6}\left(S^2 + \frac{1}{4}(C-3)^2\right), \tag{45}$$

where n is the number of observations, k is the number of predictors, S is the residuals' skewness and C is the residuals' kurtosis. Under the null hypothesis of normally distributed residuals, the test statistic follows a $\chi^2(2)$ distribution. Using a confidence level of 0.05, that means the null hypothesis of normally distributed error terms should be rejected if $JB > 5.991$

In case the Jarque Bera test is rejected, the assumption of normally distributed errors η_{ast} does not hold. Generally, this is also an indication that the linear regression model is not the optimal choice to model the available data. Often, the distribution of the residuals is more heavy-tailed than that of a normal distribution (Jarque & Bera, 1987). The regular OLS estimate is likely to be influenced largely by observations in the tails of this distribution (Hogg, 1979), therefore leading to bad model fit on the majority of the observations (Jarque & Bera, 1987).

4.3.3 Nonlinear aggregated model

The linear aggregated model in the previous subsection has a number of drawbacks. It imposes the assumption of a linear relation between dependent variable and predictors, with identically and independently normally distributed residuals. Furthermore, the relation is assumed to be without any cross-effects.

To overcome the aforementioned drawbacks, we also use a nonlinear method. Common methods for this are tree-based. As Sutton (2005) concluded, regression tree methods have several advantages over linear regression. Firstly, they do not require assumptions on the form of the underlying distribution. Secondly, regression tree methods automatically capture relations with interaction or cross effects as well. These interaction effects are hard to uncover in linear models. Thirdly, it allows for heterogeneity in the extent to which the different predictors affect the outcome. In other words, in regression tree models, a variable can be a very important factor in one subset of the predictor space, while having limited influence in another subset. This way, the influence of for example the TV ratings of a specific household type can be different for different times of the day.

As fitting a single decision tree on a set of data often leads to overfitting, we use Gradient Boosting Decision Trees. Using a boosting algorithm for decision trees generally leads to good model fit, as the model creates new trees based on the performance of previous trees it created (James et al., 2013). However, as multiple trees are created and combined into one model, computation time can be long and the resulting model is hard to interpret. A recent development in boosting is *LightGBM*, which is an advanced boosting algorithm that requires relatively limited computation time (Ke et al., 2017).

LightGBM is a boosting algorithm derived from the Gradient Boosting Decision Tree (GBDT) method (Ke et al., 2017). GBDT algorithms are boostings algorithm for forming regression or classification trees, using gradients to improve the decision trees in every iteration (Friedman, 2001). As explained by Li (2016), the general idea behind the algorithm is as follows. We assume that there exists a set of data points y that one wants to predict, a set of known variables x, and an imperfect model $F^{(m)}$ that attempts to replicate y. Using this information, we can obtain predictions $\hat{y}_i = F^{(m)}(x_i)$ and residuals $e_i = y_i - \hat{y}_i$. To *boost* the model $F^{(m)}$, we can add an estimator $h(x)$ that tries to fit the residual e_i. This estimator is found by minimizing the *loss function* of the residuals, which can for example be the mean squared error $\sum_{i=1}^{n} \left(\hat{y}_i - F^{(m)}(x_i)\right)^2$ Adding this estimator to the existing model, we obtain a new model $F^{(m+1)}$ and we can obtain new residuals, to fit a new extra estimator on. In terms of decision trees, the algorithm tries to find the split that minimizes the loss function.

Although GBDT models can lead to very accurate models (Friedman, 2001), it does have an obvious downside. When seeking the split with the most information gain, the GBDT-algorithm estimates the information gain for all possible splits for all observations (Ke et al., 2017). This leads to a very large number of computations per iteration, especially for data with a large number of observations and/or large number of features. Therefore, GBDT can be quite costly to apply.

To overcome this computational downside, Ke et al. (2017) have developed LightGBM, an algorithm derived from the GBDT algorithm that requires a smaller number of computations. This is achieved by using two techniques to reduce dimensions and data. Firstly, observations with small gradients, having little room for improvements, are discarded in the estimation of the new model. Secondly, predictors that seem mutually exclusive are combined to reduce the number of dimensions. These two techniques have in previous research led to up to 20 times shorter computation time than conventional GBDT techniques, while retaining similar model accuracy (Fonseca et al., 2017; Ke et al., 2017).

LightGBM is a complex algorithm, with many parameters that need to be fine-tuned. These parameters include the maximum depth and number of leaves of the decision trees, but also the size of the learning rate or shrinkage, a correction factor reducing the impact of individual trees, and therefore reducing the risk of overfitting (Friedman, 1999). Furthermore, parameters as the minimum number of observations in each leaf and a regularization measure can be imposed as well. The performance of the LightGBM model is therefore highly dependent on the value of these hyperparameters (Snoek et al., 2012; Thomas, 2019). Setting these parameters by hand either requires expert experience (to have knowledge on which parameter settings might work) or a lot of time (to perform a grid search: try many different combinations of parameters to see which combination

yields the best result) (Snoek et al., 2012).

To overcome these issues, we have employed Bayesian Optimization of the hyperparameters, a technique proposed by Snoek et al. (2012) to automate the tuning of parameters. The main idea of this technique is to build a new, simple probability model of the objective based on the values of the hyperparameters. By finding the optimum of this surrogate probability distribution, we can identify parameter values that are likely to yield good results on the actual optimizer. Mathematically, Bayesian Optimization attempts to solve the following:

$$x^* = \arg\min_{x \in \mathcal{X}} f(x), \tag{46}$$

where x^* denotes the hyperparameters that yield the lowest value of the performance metric used, and $\mathcal{X}$ denotes the domain of possible values for the hyperparameters (Dewancker, McCourt, & Clark, 2015). The outcome of the function $f(x)$ is the value of the performance metric of the nonlinear model for hyperparemeters x.

The most common choice of modelling objective functions in Bayesian Optimization are Gaussian Processes (GPs) (Dewancker et al., 2015; Snoek et al., 2012). In GPs, the outcome of the function f is assumed to follow a Gaussian distribution with mean function $\mu(x)$ and covariance function $K(x)$: $f \sim \mathcal{N}\big(\mu(x),\, K(x)\big)$. The distribution of f and the optimal value x^* are built up in a series of iterations, called Sequential Model-Based Optimization (J. S. Bergstra, Bardenet, Bengio, & Kégl, 2011). In each iteration, the distribution of f is updated based on the observations $\mathcal{D} = \{x_i, y_i\}_{n=1}^{N}$ evaluated so far, where x_i is a vector of hyperparameter values and y_i is the value of the performance metric of the LightGBM model evaluated with hyperparameters x_i. We model $y_i \sim \mathcal{N}\big(f(x_i),\, \nu\big)$, where ν is the variance in the actual model not captured by the GP. Next, the parameter values that should reduce f are obtained by maximizing the *acquisition function*. This set of hyperparameter values is then applied to the original model to obtain new data for a new iteration (J. S. Bergstra et al., 2011; Snoek et al., 2012).

An essential step is the maximization of the acquisition function. The acquisition function is basically a function that defines a maximization objective using $f(x)$ to *acquire* the next set of hyperparameter values x to evaluate. The acquisition function mainly used in Gaussian Processes is the Expected Improvement (EI) (Jones, Schonlau, & Welch, 1998), (47), which seeks to maximize the difference between the current optimal hyperparameter values x_{best} and the new candidate values x_{new}. With the improvement function $I(x) = \max\big(f(x_{\text{best}}) - f(x_{\text{new}}),\, 0\big)$ (Jones et al., 1998), the EI is obtained as in (47) (Snoek et al., 2012). The second step is obtained by integration by parts Jones et al. (1998). In (47), $f' = f(x_{\text{best}})$ and $f = f(x_{\text{new}})$. Furthermore, $\phi(\cdot\,; \mu, \sigma^2)$ and $\Phi(\cdot\,; \mu, \sigma^2)$ denote the pdf and the CDF, respectively, of a normal distribution with mean μ and variance σ^2.

$$
\begin{aligned}
E\left[I(x) \mid \mathcal{D}\right] &= \int_{-\infty}^{f'} \left(f' - f\right) \phi\left(f;\, \mu(x), K(x)\right) df \\
&= \left(f' - \mu(x)\right) \Phi\left(f';\, \mu(x), K(x)\right) + K(x) \phi\left(f';\, \mu(x), K(x)\right)
\end{aligned} \tag{47}
$$

A new set of parameter values to estimate the model on is then obtained by maximizing (47). This can either be done by reducing the mean function $\mu(x)$ or increasing the variance function $K(x)$ (Jones et al., 1998). Hence, there is a trade-off between staying in the currently "optimal" area of parameters and exploring a different area in search of new, even better optima. The EI criterion automatically makes a trade-off between these options.

The process of evaluating the EI and evaluating the new parameter values in the actual model is repeated until convergence or for a maximum number of 50 iterations. Furthermore, for each evaluation, the number of iterations of the LightGBM algorithm is capped at 100.

Bayesian Optimization with GPs has been applied to a number of hyperparameters of the LightGBM model, that are intuitively difficult to set by hand. Table 1 summarizes which hyperparameters have been tuned using Bayesian Optimization and which values are used as the domain of these hyperparameters. The maximum depth controls how many consecutive leaves in the trees can be grown, and thus how 'deep' the tree can become. The learning rate controls the impact of new changes to a tree. A high learning rate may lead to overfitting. The leaf ratio controls how complex the model can become. The leaf ratio is defined as the theoretical maximum of leaves (which is $2^{\text{num\_leaves}}$) divided by the maximum depth. It therefore defines what percentage of the maximum number of leaves the trees should have. The minimum data in leaf defines the minimum number of observations that should be in a leaf for it to be included in the tree. Setting this value low can yield a good fit but may cause overfitting. The L2-lambda is a regularization measure to reduce the tree complexity.

Name	**Type**	**Value range**
Maximum depth	Discrete	{4, 5, 6, 7, 8, 9, 10, 11, 12}
Learning rate	Continuous	[0.001, 0.200]
Leaf ratio	Continuous	[0.4, 1.0]
Minimum data in leaf	Discrete	{10, 20, 50, 100, 150, 200, 500}
L2-lambda	Continuous	[0.0, 0.1]

Table 1: Hyperparameters used in Bayesian Optimization

The predictors used in the nonlinear model are slightly different than in the linear model. The weighted household types ratings $c_{aj} z_{jst}$ are again used as predictors for v_{ast}. Furthermore, the day of the week and the hour of the day are used as predictors, to allow for heterogeneity over time. To allow for heterogeneity among channels, the channel size is included as predictor as well. This channel size is defined as follows. For every station s, the absolute channel size is the sum over time of the number of households watching s at time t. The categorical variable channel size of s is then either *small*, *medium* or *large*, based on this absolute channel size of s. Lastly, the total rating for station s at

time t, ψ_{st}, is also included as predictor.

Like the linear aggregated model, the nonlinear aggregated model does not necessarily generate estimates $\hat{v}_{ast}$ that align the total ratings ψ_{st}. Therefore, the estimated $\hat{v}_{ast}$ have to be rescaled, using the steps in (42) - (44) again.

4.4 Performance measures

In order to validate the estimates of $\hat{v}_{ast}$ obtained using the models in sections 4.2 and 4.3, we compare them to the actual viewing numbers v_{ast} in the truth dataset. Performance measurement of continuous variables is often done using the Mean Squared Error (MSE) and the Mean Absolute Error (MAE). However, as these traditional performance measurements often have the downside of taking into account absolute deviations from the true value, a relative performance measurement is needed.

A measurement not taking these absolute measures into account is the Kullback-Leibler (KL) divergence, named after Kullback and Leibler (1951). The KL divergence compares whether two discrete distributions P and Q with the same n possible outcomes are similar, using (48). The lower the outcome of (48), the more similar the distributions are.

$$KL = \sum_{i=1}^{n} P_i \log\left(\frac{P_i}{Q_i}\right) \tag{48}$$

To assess model performance in this research, we use KL divergence to compare the distributions of the segment TV ratings as percentage of the total TV ratings between the estimated segment TV ratings and the true segment TV ratings. For each time t and channel s, the KL-divergence KL_{st} is then calculated as:

$$KL_{st} = \sum_{a=1}^{A} P_{ast} \log\left(\frac{P_{ast}}{Q_{ast}}\right); \tag{49}$$

where P_{ast} and Q_{ast} are the estimated and true percentages of viewers of channel s at time t from segment a, respectively. For example, if at time t channel s is viewed by two individuals from segment a of 10 viewers in total, then $Q_{ast} = \frac{2}{10} - .20$. P_{ast} and Q_{ast} can be obtained using (50)

$$P_{ast} = \frac{v_{ast}}{\psi_{st}} \quad \text{and} \quad Q_{ast} = \frac{\hat{v}_{ast}}{\psi_{st}} \tag{50}$$

To put more emphasis on the accuracy of the ratings of larger channels, we use a weighted version of the KL-divergence. This is achieved by scaling each channel's KL-divergence KL_{st} by the total ratings ψ_{st} for that channel. The overall weighted KL-divergence KL^* is then obtained using (51).

$$KL^* = \frac{\sum_{t=1}^{T}\sum_{s=1}^{S} KL_{st}\psi_{st}}{\sum_{t=1}^{T}\sum_{s=1}^{S} \psi_{st}} \tag{51}$$

Another way of evaluating the model fit is to measure the Mean Percentage Error (MPE), which measures the mean difference between the estimated and true percentage of TV ratings per segment.

The outcome is therefore the average number of percentage points between the estimated and true ratings. Mathematically, this can be calculated as in (52).

$$MPE = \frac{100\%}{AST} \sum_{t=1}^{T} \sum_{s=1}^{S} \sum_{a=1}^{A} \frac{v_{ast} - \hat{v}_{ast}}{\psi_{st}} \tag{52}$$

Furthermore, as measure of feasibility, the computation time needed to estimate the models is used as performance measure as well.

To visualize the fit of the two models, *index plots* have been created. in these index plots, each segment TV ratings estimate and true value has been transformed to a relative value. This way, the estimates and truth values are more convenient to plot in the same plot. To create such a plot, each true ratings index is obtained by dividing the relative segment rating, calculated as v_{ast}/m_a by the overall relative segment rating $\psi_{st}/\sum_a m_a$:

$$IX = \frac{\frac{v_{ast}}{m_a}}{\frac{\psi_{st}}{\sum_a m_a}} = \frac{v_{ast} \sum_{a=1}^{A} m_a}{\psi_{st} m_a}. \tag{53}$$

Similarly, the estimated segment ratings index $\widehat{IX}$ can be obtained by replacing v_{ast} in (53) by $\hat{v}_{ast}$.

These indices are a relative measure of size of the TV ratings. When the segment ratings v_{ast} are equal to the 'benchmark rating' $q_a \psi_{st}$, the index will be equal to 100. If the v_{ast} is much lower than $q_a \psi_{st}$, the index will be close to zero. Moreover, if v_{ast} is much larger than $q_a \psi_{st}$, it will be larger, approximately up to 100 times the number of segments if $v_{ast} = \psi_{st}$.

An index plot is then created by creating a scatter plot of indices, with the true index IX on the x-axis, and the estimated index $\widehat{IX}$ on the y-axis. The closer these scatters are to the line $y = x$, the better the model fit.

5 Results

This section will describe the results of the models described in section 4 applied on the data discussed in section 3. First, general results of the choice model approach are discussed. Next, the results of the aggregated model are described. Lastly, the two approaches are compared with each other to distinguish each method's advantages and disadvantages.

5.1 Results of choice model

In this subsection, the results of the choice model are laid out. The different household utility specifications as discussed in sections 4.2.2 and 4.2.3 are compared on their accuracy of the segment TV ratings estimates, their model stability and their computation times.

5.1.1 Model fit

To assess the suitability of the different utility specifications, we apply them to a subset of the Chicago the. Table 2 shows the performance of the different utility specifications in estimating TV ratings for a Thursday at prime time, on a random subset of 50 channels, using gender as the only segment. From this table, we can denote that all but the maximax approach yield estimates of TV ratings that are better than random guessing in terms of both KL-divergence and Mean Squared Error (MSE). Furthermore, multiplicative models appear to be numerically quicker to solve, as the multiplicative models have computation times of about half the computation times of other models.

Model	MSE ($\times 10^6$)	KL-divergence	Computation time (s)
MNL additive	28.0	0.048	1.292
MNL multiplicative	28.9	0.048	0.578
Choice additive	58.3	0.079	1.337
Choice multiplicative	30.5	0.048	0.603
Choice maximax	56.2	1.457	0.784
Choice maximin	27.9	0.038	1.373
Choice random	361.0	0.434	

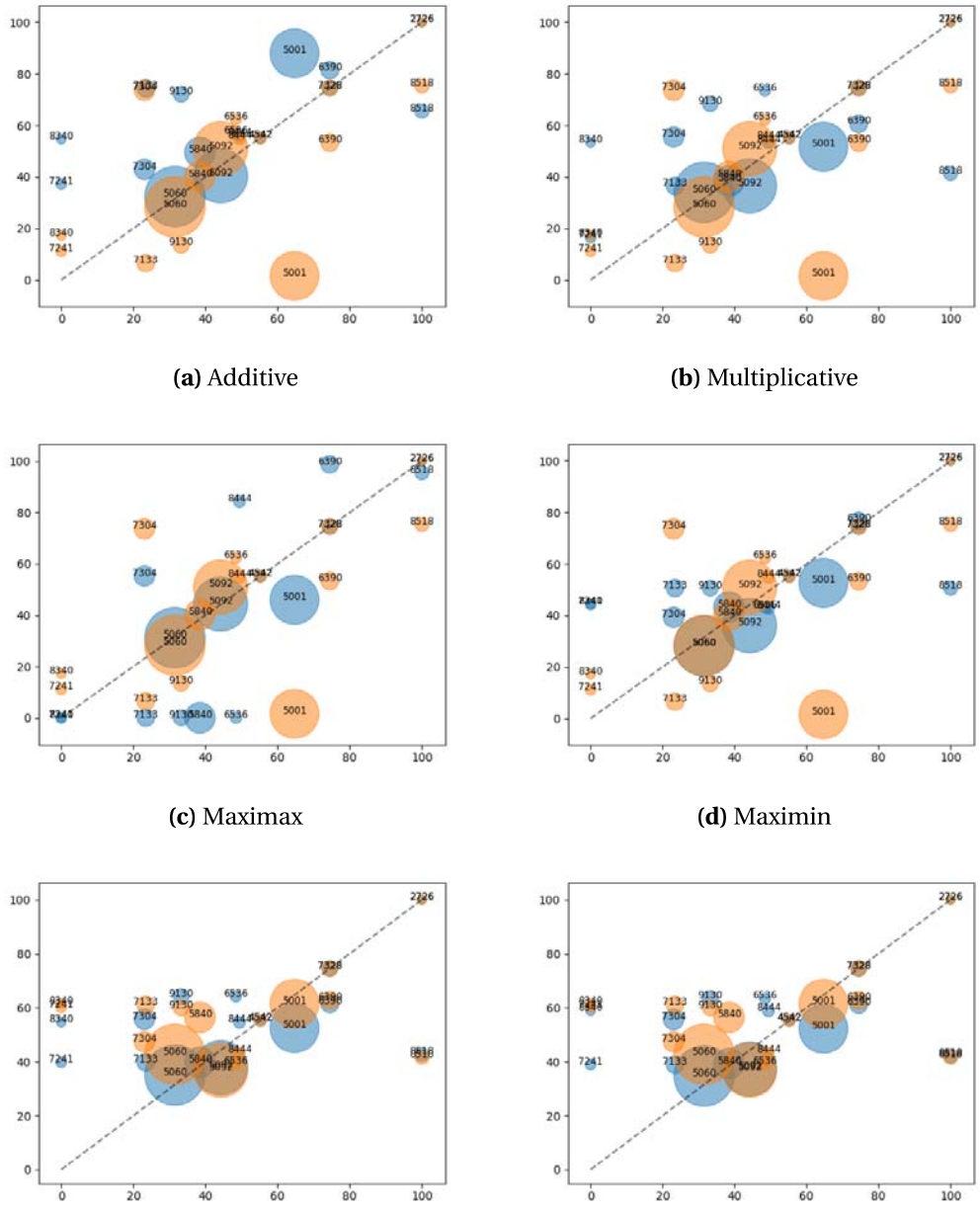
Table 2: Results of the different utility specification on estimation of TV ratings in Chicago on Thursday January 3rd, 2019, at 20:00 local time, on a subset of 50 channels, and using gender as segmentation factor.

The actual percentage and estimated percentage of males per household utility specification are displayed in figures 7a - 7f. Each specification is initialized with the same set of random utilities. The maximax approach takes on many extremes, estimating the TV viewers of multiple channels to be either 0% or 100% males. The multiplicative approach seems to be a bit more conservative, having dots spread around the line of 50%. The two MNL-approaches push all estimates to the mean of 50%: both its optimal values as the initial random utilities yield estimates close to the mean. The additive specification seems to be less drawn to the population mean than most other specifications. However, the additive specification is therefore also a bit less accurate than the other specifications.

From figures 7a - 7f, we can also conclude that in general, the larger the channel, the closer the estimated ratings are to the actual ratings. This is not surprising, as there is more data available to estimate a large channel's TV ratings on.

5.1.2 Sensitivity analysis

To assess model stability and the proneness of the different utility specifications to end up in local minima of their log-likelihood, we perform a sensitivity analysis. The model is estimated for each utility specification, using 100 different initial values of the parameters. The reason for this test, is that it is hard to tell whether the optima of the objective functions from which the TV ratings in figures 7 and 2 are obtained, are local or global optima. A function of multiple variables is convex if the Hessian is positive semi-definite (Lau, 1987). However, the objective functions are hard to differentiate, and therefore it is very difficult to obtain the Hessian matrix analytically. Because of this, we can not assess



(a) Additive

(b) Multiplicative

(c) Maximax

(d) Maximin

(e) MNL additive

(f) MNL multiplicative

Figure 7: Accuracy plots of the different household utility specifications. For each dot. the location on the x-axis is the true percentage of males of the channel, whereas the location on the y-axis represents the estimated percentage of males. Orange dots represent the 'random' benchmark estimate (when each channel's utility is random), whereas the model estimates are displayed in blue. The size of the dot is the size of the channel in terms of total number of viewers. The four-digit number in each dot is the channel code of the corresponding channel. Ideally, all dots are located at the diagonal dashed line, indicating that the estimated percentage of males is equal to the true percentage.

whether the log-likelihoods are convex, and thus whether the obtained optima are local or global.

Model stability across seeds is tested using the Coefficient of Variation (CV), also known as Relative Standard Deviation. The CV is simply the sample standard deviation rescaled by the sample mean, to be able to generalize standard deviations of TV ratings of different channels. Mathematically, the CV is defined as

$$c_{\hat{v}_t} = \frac{s_{\hat{v}_t}}{\hat{\bar{v}}_t}, \tag{54}$$

where $\hat{v}_t$ is the mean and $s_{\hat{v}_t}$ is the standard deviation of the TV ratings estimates $\hat{v}_{ast}$ at time t (Abdi, 2010). The mean CV is then the the mean of the CVs of TV ratings over the different seeds of random parameter initialization. When the CV of seed k is denoted as $c^{(k)}_{\hat{v}_t}$, the mean of K CVs, $\bar{c}_{\hat{v}_t}$, can be calculated as

$$\bar{c}_{\hat{v}_t} = \frac{1}{K}\sum_{k=1}^{K} c^{(k)}_{\hat{v}_t}. \tag{55}$$

Sensitivity to initialization of parameters is not necessarily a bad thing. Normally, we would estimate such a sensitive model multiple times and select the model and initialization with the lowest outcome of the objective function, in this case the negative log-likelihood. However, as table 3 shows, the difference in values of the log-likelihood is not so large among seeds. The standard deviation of the optimal log-likelihood value found using 100 different initializations is close to zero for the multiplicative models and approximately 0.3 for the additive models. The maximax and maximin model show more sensitivity in its log-likelihood values to the random parameter initialization. For the maximax, maximin and additive specifications, the CV of the estimated TV ratings is larger though, indicating that the log-likelihoods have many local optima close to each other, with different parameter values. Therefore, it is hard to say which of these estimates is 'best' based on log-likelihoods only. Therefore, it is in this case better not to select a model based on outcome of the log-likelihood, but on other measures such as computational feasibility, model fit and sensitivity to different utility initializations.

When assessing the sensitivity of the models using the numbers in table 3, we can conclude that the two multiplicative specifications are more stable than the other specifications. With CVs close to zero, they are fairly constant in their predictions. Apparently, the log-likelihoods of the multiplicative and MNL-multiplicative models have few local minima. The computation times of multiplicative models is lower as well. This is an indication that for the log-likelihood of the multiplicative specification, it is easier to find a (local) minimum of the negative log-likelihood.

In terms of model fit and computational feasibility, the multiplicative specification seems to outperform the other specifications as well. The multiplicative choice model has one of the lowest average weighted KL-divergence over the 100 runs. The maximin approach has a similar model fit, but is more difficult to optimize: the process of optimization took on average almost one and a half second per model initialization, whereas this was only half a second for the multiplicative model. A possible reason for this, is that the log-likelihood of the maximin approach is not a smooth function, whereas the log-likelihood of the multiplicative specification is smooth. Furthermore, the maximin approach

Model	CV	Comp. time (s)	Mean KL-div	Mean LL	LL SD
MNL additive	0.002	121.6	0.0475	1160.65	0.0018
MNL multiplicative	0.000	175.1	0.0485	1159.18	0.0000
Choice additive	0.127	121.0	0.0971	1149.45	0.2900
Choice multiplicative	0.000	70.1	0.0476	1159.19	0.0000
Choice maximax	0.389	120.1	0.5256	1149.39	0.7257
Choice maximin	0.110	176.7	0.0450	1148.18	0.8142
Choice random	0.339		0.4611	1343.39	68.7856

Table 3: Sensitivity analysis of household utility specifications in the choice model. LL is short for log-likelihood, CV for Coefficient of Variation, and SD for Standard Deviation. Each model is estimated with 100 different random initializations of its parameters. The model is estimated on a subset of 50 channels, for TV ratings in Chicago on Thursday January 3rd, 2019, at 20:00 local time, using gender as only segment.

seems to get stuck in several local minima, leading to a large standard deviation in its estimates of TV ratings. The multiplicative model is more stable. The two MNL-specifications yield relatively stable results as well. However, as figure 7 showed, these specifications yield estimates of TV ratings that are more or less the same as the estimates of the multiplicative model. Their likelihood is less convenient to optimize though, according to the computation times in table 3.

5.1.3 Comparison of utility specifications

All in all, the estimations of different household utility specifications show quite some differences in model fit, computation costs and sensitivity to random initialization. The multiplicative utility specification seems to be the most suitable one, having limited differences in estimated TV ratings, the lowest computation time and the lowest KL-divergence. Apparantly, the log-likelihood of the multiplicative model is a bit convex and therefore relatively convenient to optimize. For this reason, we will use the multiplicative definition in the performance comparison of the choice model and aggregated model.

5.2 Results of aggregated model

In this subsection, the results of the linear and nonlinear models of the aggregated method are discussed. They are compared in terms of model fit and computation time. Furthermore, the suitability of the linear model is tested using a Jarque-Bera test on normality of its residuals.

To check the fit of the aggregated method, two different sets of data are used. The models are trained on the data of last three weeks of the Chicago market. Their model fit is then evaluated on the first week of Chicago (the validation set), to test the *within-market* fit. Then, the models are applied to the Cleveland market (the test set) to assess the *out-of-market* fit. This way, we can assess the extent to which TV viewing behaviour is different across markets. Furthermore, the *out-of-market* fit will indicate how suitable the aggregated model is for estimating the segment TV ratings of data for which individual level viewing behaviour is unavailable (e.g. in RPD data).

5.2.1 Linear model

The parameters of the linear aggregated model, displayed in tables 12 - 15 in appendix B, are mostly as expected. Household ratings of household types containing individuals of only one segment, only affect the ratings of that segment and not of other segments. From the time/date variables, we can gain some interesting insights, though. According to the linear model, men generally watch more TV during the night and morning than women. Furthermore, older people watch more TV during weekdays, whereas middle-aged people watch more TV during weekends. Highly educated people watch more TV during the night and morning hours, but less during weekdays than people with low or medium levels of education.

The Jarque-Bera test on normality of the residuals is rejected for all linear models (every segment and every type of segmentation). This indicates that the residuals η_{ast} are apparently not normally distributed. The residuals' distributions seem to be too heavy-tailed for a normal distribution. A possible explanation for this, is the large number of observations (quarter hourly TV ratings per channel) where a channel is watched by only one individual in the data. Hence, in these cases, the total TV rating ψ_{st} is equal to the individual weight w_i of the single individual watching channel s. Furthermore, the segment rating v_{ast} is equal to w_i for the segment of this single individual, and zero for all other segments. Hence, for these quarter hours, it can be argued that the TV ratings follow a type of discrete distribution, instead of a continuous normal distribution. For this reason, the linear model, assuming normally distributed TV ratings, will not perform well for these quarter hour observations.

5.2.2 Nonlinear model

The variable importance of the nonlinear aggregated model, displayed in tables 17 - 20 in appendix D show similar results as the linear parameters. The numbers in tables 17 - 20 are the number of times each variable was used as split in each of the 100 decision trees, and thus give an indication of the extent to which their value influences segment TV ratings. It is hard to draw any conclusions from these numbers, as variable importance does not say anything about the direction of the split. However, we can conclude that the time of the day, day of the week and benchmark rating play a large role in all nonlinear models. Hence, the nonlinear model captures some dynamic behaviour. Furthermore, the channel size has relatively high importance in many models as well, indicating some heterogeneity exists among channels.

The hyperparameters of the nonlinear aggregated method, identified with Bayesian Hyperparameter Optimization, are displayed in table 16 in appendix C. The hyperparameters obtained by Bayesian Optimization are in general quite similar. Most models are built with a relatively high number of leaves and a high learning rate, but low minimum number of observations per leaf. Apparantly, the LightGBM model does not overfit too quickly on the training set (the last three weeks of Chicago data). Two exceptions here are the models for young females when segmenting on age and gender, and for middle-aged people when segmenting on age. To avoid overfitting on training data, Bayesian

Optimization has yielded a relatively low number of leaves and maximum depth here. Furthermore, the regularization factor (the Lambda L2-parameter) is relatively high for these models as well.

The Bayesian Optimization technique ensures that the model is not overfitted on the training data, by testing each model on the validation set (the first week of data in Chicago). However, it does not account for *out-of-market* overfitting. Therefore, in the next subsection, the linear and nonlinear aggregated method will be tested for fit on the Cleveland data as well.

5.2.3 Comparison

The fit on the Chicago data (table 4) differs largely between the two implementations of the aggregated method. The LightGBM model fits very well on the training data, having KL divergences below 0.1 for all of the possible segmentations. The KL-divergence on the validation data (the first week of data) is a bit higher, but nevertheless the model does not seem to be overfitted. By just looking at table 4, the nonlinear model seems to outperform the linear model for every possible type of segmentation in terms of *within-market* model fit. This better training fit and *within-market* fit can be explained by the fact that the nonlinear model captures interaction effects that explain part of the segment TV ratings. The linear model does not capture these effects.

Segmentation		**Training**		**Validation**	
Type	Number	AM linear	AM nonlinear	AM linear	AM nonlinear
Gender	2	0.2599	0.0424	0.2490	0.0592
Age	3	1.0455	0.0368	1.1385	0.0672
Education	3	0.6587	0.0273	0.6780	0.0581
Age-Gender	6	1.2465	0.0727	1.3370	0.1526

Table 4: KL divergences of the two specifications of the aggregated model on the Chicago market.

To assess the *out-of-market* fit, the out-of-sample estimates for the TV ratings in Cleveland have been split up into four buckets of different channel sizes. Table 5 shows how these buckets have been defined. The *out-of-market* fit is evaluated for each channel size bucket separately, to assess differences in the strengths and weaknesses between the two methods. The splitting is done per quarter hour t, based on the number of households viewing a channel at t. For example, if a channel s is viewed by 7 households at time t, it would for time t belong to channel size *Small*. If at time $t+1$ four more households tune in, bringing the number of households viewing s to 11, channel s will belong to channel size *Large* at time $t+1$.

Channel size	**Abbreviation**	**Minimum**	**Maximum**
Very large	XL	20	∞
Large	L	10	20
Small	S	5	10
Very small	XS	0	5

Table 5: Definition of household sizes. The columns indicate the minimum and maximum number of households viewing a station per quarter hour, respectively, for a station to fall into that channel size bucket.

As table 6 shows, the *out-of-market* fit of the linear and nonlinear methods is similar when the number of segments is limited to two. However, as the number of segments increases, the linear model becomes unstable and increasingly less accurate. The nonlinear model can handle a larger number of segments better: when segmenting on age or education, the weighted KL-divergence of the *large* and *very large* channels is still below 0.1. When the number of segments increases to six, the accuracy of the nonlinear aggregated model drops as well.

Segments			**Model**	
Type	Number	Channel size	Linear	Nonlinear
Gender	2	XL	0.017	0.013
		L	0.036	0.036
		S	0.059	0.059
		XS	0.482	0.607
Age	3	XL	0.373	0.017
		L	1.046	0.051
		S	2.094	0.088
		XS	1.568	0.291
Education	3	XL	0.080	0.019
		L	0.666	0.044
		S	1.916	0.076
		XS	3.749	0.364
Gender & Age	6	XL	2.525	0.067
		L	2.185	0.150
		S	2.826	0.255
		XS	3.525	2.789

Table 6: KL divergences per channel size of the different aggregated model specifications, using different types of segmentation, on the *out-of-market* Cleveland data.

The index plots of the *within-market* fit for the two aggregated models when imposing segmentation on age in figure 8 reflects the general results obtained in table 4. The cloud of estimated indices of the nonlinear method follows the optimal diagonal line quite closely, even for small channels. The linear model seems to suffer from regression to the mean, meaning that the estimated ratings $\hat{v}_{ast}$ are generally closer to the benchmark ψ_{st} than the actual ratings v_{ast} (Galton, 1886). The cloud of the estimates of the linear model is a bit tilted, it is a bit more horizontal, indicating that the linear estimates are generally closer to the population mean.

The index plot of the *out-of-market* fit in figure 9 shows a similar pattern. Again, for the results in this figure, segmentation is performed on age. The indices for the estimates of the nonlinear model are again scattered around the diagonal. Except for the largest channels, the spread is a bit larger than in the plots of the *within-market* fit. The linear model is again a bit tilted towards the population mean. For the smallest channels, neither model produces accurate results, with widely scattered points for both models. The estimates of channels in the *small* category are more accurate, but nevertheless not too trustworthy. For this reason, a rule of thumb is that only segment TV ratings of observations (a single station in a single quarter hour) with at least 10 households viewing can be estimated with an acceptable accuracy.

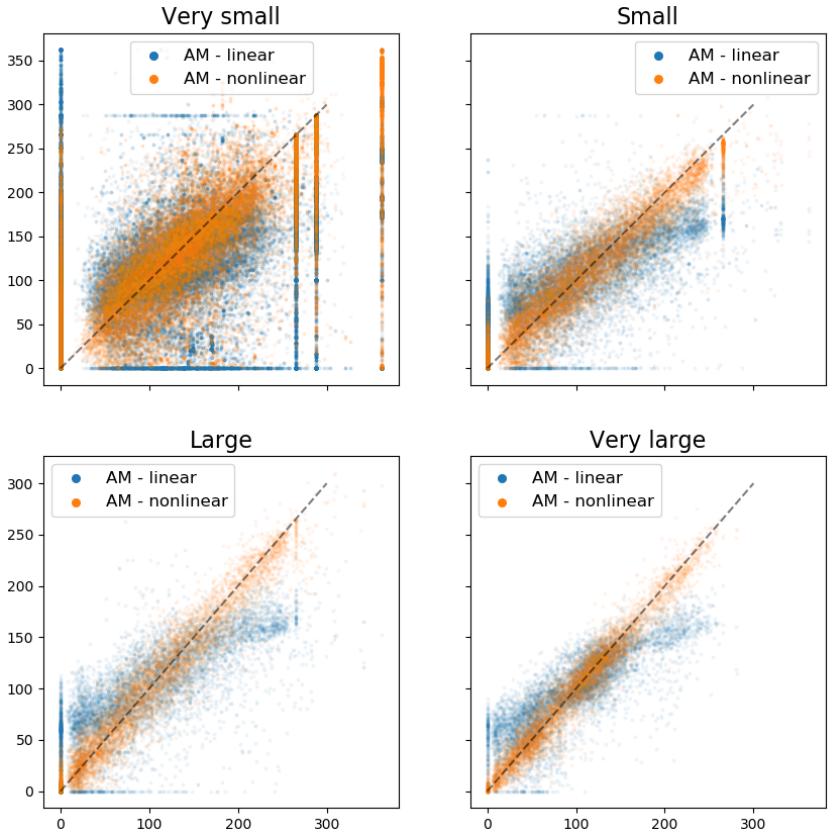


Figure 8: Index plots displaying the *within-market* model fit of the aggregated method when segmenting on age, on the first week of the Chicago data (the validation data), per channel size. Segments are defined as Young (ages 18-34), Middle-aged (ages 35-54) and Old (ages 55+). Channel sizes are defined as in table 5. True indices are on the x-axis, estimated indices on the y-axis.

Comparing the *out-of-market* fit of the models with the Venn plots of household types in figure 12 and the according overlap percentages in table 11 in appendix A, the importance of the overlap in household types becomes apparent. This is especially the case when comparing the models using education and age as segments. The overlap in household types between Cleveland and Chicago is higher when imposing segmentation on education (71%) than when imposing segmentation on age (63%). As a result, the fit of the linear model is far worse for the age segments than for education segments. This difference is less apparent for the nonlinear model, but the fit is still slightly better for the *large* and *small* channel sizes. Furthermore, the drop in performance of both methods when increasing the number of segments to six can partially be explained by figure 12d. It shows that when imposing segmentation on gender and age, 48% of household types are missing in Cleveland. Furthermore, 20% of the household types in Cleveland are not present in Chicago. This means, that for 20% of the household types in Cleveland, their TV viewing behaviour is not used in the estimation of these *out-of-market* segment TV ratings.

The computation times in table 7 show large differences in estimation time between the different

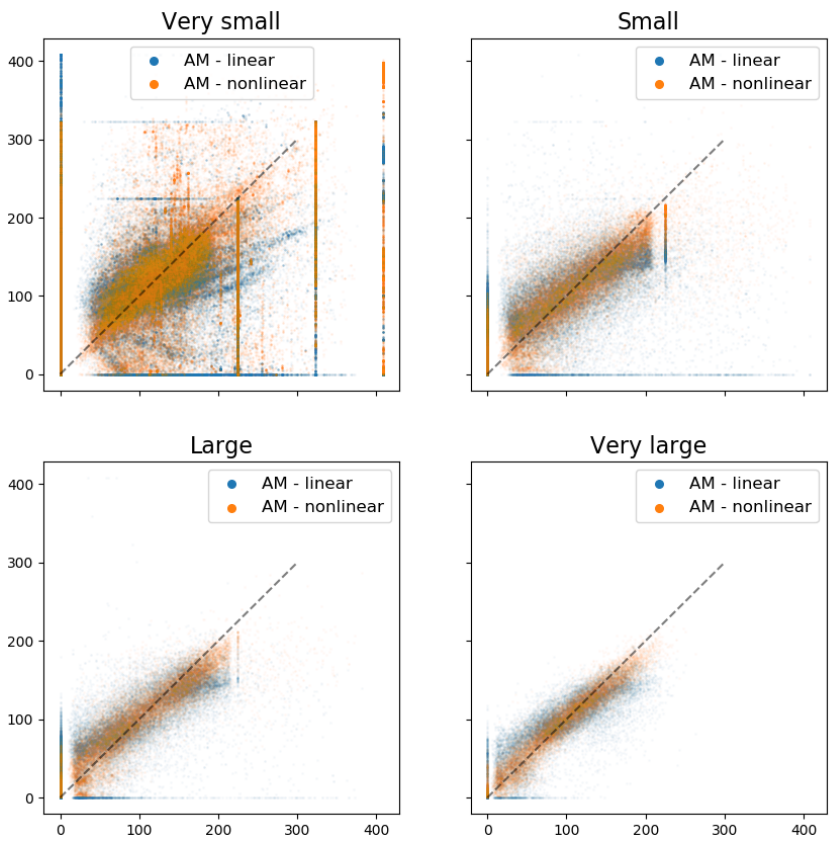


Figure 9: Index plot displaying the *out-of-market* fit of the estimates of segment TV ratings of both aggregated model on the full month of Cleveland data. Segmentation is done on age. Segments are defined as Young (ages 18-34), Middle-aged (ages 35-54) and Old (ages 55+). Channel sizes are defined as in table 5. True indices are on the x-axis, estimated indices on the y-axis.

methods. The linear model is by far the fastest: for every type of segmentation, the whole process of estimating parameters on the training set and obtaining predictions for the validation and test sets takes less than one minute per segment. For the nonlinear model, this is substantially longer: ranging from just shy of 11 minutes for the three segments when segmenting on age to more than 32 minutes when segmenting on age and gender. This result is hardly surprising as the nonlinear model is a far more complex model than the analytically solvable linear model. Most of the computation time of the nonlinear model is spent in the process of identifying optimal hyperparameters. During the process,

Segments		**Model**	
Type	Number	Linear	Nonlinear
Gender	2	00:01:04	00:13:53
Age	3	00:01:47	00:10:48
Education	3	00:02:14	00:19:09
Age-Gender	6	00:05:47	00:32:52

Table 7: Computation times per model and segmentation. The times are in the format HH:MM:SS.

the LightGBM model has to be evaluated multiple times to find the optimal model.

Based on table 4 and figure 8, we can conclude that the nonlinear model outperforms the linear model in terms of *within-market* fit. The *out-of-market* fit displayed in table 6 and figure 9 also favour the nonlinear model. There are two straightforward reasons for the better model fit of the nonlinear model. Firstly, the Jarque-Bera test on normality of the residuals was rejected. This indicates that the normality assumption of the linear model does not hold, leading to bad model fit. Secondly, the nonlinear model was able to capture nonlinear relations (interaction effects between predictors). The linear model is per definition not able to capture this type of relation, whereas the nonlinear model can. By modelling these nonlinear relations, the nonlinear model can account for variable importance that varies between subsets of the data.

The better fit of the nonlinear model comes at a cost of a computation time that is at least five times as large as the computation time of the linear model. However, as this computation time is still manageable, with just over half an hour of computation for the largest model tried out, the nonlinear model seems to be the most suitable one of the two.

Furthermore, based on the *out-of-market* fit of the nonlinear model, we can conclude that the aggregated model can be used to estimate segment TV ratings for markets where no individual TV viewing behaviour is known (like RPD data). The nonlinear aggregated model is able to produce relatively accurate estimates of the segment TV ratings for different types of segmentation, despite the imperfect overlap in household types and the fact that TV viewing behaviour may be different in different markets.

5.3 Comparison of choice model and aggregated model

This subsection will compare the suitability of the choice model and aggregated model. Specifically, the multiplicative choice model and the nonlinear aggregated model will be compared, as these specifications yielded the best results. These models are compared in terms of model fit, measured by weighted KL-divergence and MPE, and feasibility, measured by computation time. The models are compared by making estimates of the segment TV ratings on the full month of Cleveland data.

The estimates of the choice model are obtained by estimating the TV ratings per segment for each time step independently, as explained in section 4.2. In section 5.1, the model was estimated per subset of 50 channels. Including all 330 channels in a single model estimation drastically increases the number of segment utilities ω_{ast} to estimate in the log-likelihood in (12). As table 8 shows, the estimation procedure takes approximately 10 times more time when the number of stations used in the choice model increases from 50 to 330. Furthermore, table 8 also shows that the efficiency of the choice model is dependent on the number of stations. Using 50 stations at a time seems to produce the most accurate estimates.

For the aforementioned reasons, the data is split when applying the choice model to the full month of Cleveland data. The data is split into 8 subsets of approximately 50 channels each. The choice model is then estimated for each channel subset and quarter hour separately. This should in the end

No. stations	KL-divergence	Comp. time (s)
25	0.144	4.27
50	0.090	4.50
75	0.110	4.87
100	0.117	6.16
330	0.319	44.16

Table 8: Model fit and computation times of the multiplicative CM on different subset sizes of stations for a single quarter hour. Results are obtained from the quarter hour starting at January 3rd, 2019 at 19:00 local time in the Cleveland data (*out-of-market* data).

yield reliable estimates within a reasonable computation time.

Similar as for the results described in section 5.2, the estimates of the aggregated model are obtained by training the nonlinear AM on the last three weeks of the Chicago data. The model obtained on the Chicago data is then applied on the Cleveland data set to obtain *out-of-market* estimates of the TV ratings per segment.

5.3.1 Model fit

From the accuracy of both models for different channel sizes and different types of segmentation, displayed in table 9, we can draw a number of conclusions regarding the fit of the channels for different channel sizes, segment types and performance measures.

First of all, when segmenting on gender, keeping the number of segments as low as two, the two models seem to yield relatively similar results. The fit of the aggregated model and choice model are the same for the largest channels. For the other three channel sizes, the choice model estimates yield slightly lower KL-divergences and MPEs, but the difference is mostly small.

However, as the number of segments increases, a difference in performance between the two models is noticeable. As expected, both models perform worse when the number of segments increases. Nevertheless, this decrease in performance is much larger for the choice model than for the aggregated model. When segmenting on age or education, the aggregated model still produces reasonable results, with KL-divergences close to that of the model segmented on gender. The choice model is able to produce satisfactory results when segmenting on education. However, the results for age show a much larger difference between the KL divergence of the choice model and that of the aggregated model. This trend is also reflected in the MPE, which shows that especially in the age segments, the aggregated method is able to produce estimates that are much closer to the true segment ratings than the choice model estimates.

A possible explanation for this, is the limited number of data points used to estimate the choice model on. The choice model is estimated for each quarter hour seperately. For each quarter hour, utilities for each segment and channel have to be estimated based on the data available for that quarter hour. In case the number of segments increases, the number of utilities to estimate increases as well. As a result, more variables have to be estimated with the same limited number of observations. This

Segments			KL Divergence		MPE (in %)	
Type	Number	Channel size	AM	CM	AM	CM
Gender	2	XL	0.013	0.013	6.4	6.3
		L	0.036	0.029	10.4	9.0
		S	0.059	0.055	12.9	11.8
		XS	0.607	0.310	25.7	23.5
Age	3	XL	0.017	0.064	4.3	8.1
		L	0.051	0.118	6.4	10.9
		S	0.088	0.207	8.0	13.8
		XS	0.291	0.656	10.2	21.9
Education	3	XL	0.019	0.029	5.2	5.7
		L	0.044	0.058	7.3	8.4
		S	0.076	0.130	9.3	11.0
		XS	0.364	0.580	12.7	20.7
Gender & Age	6	XL	0.067	0.116	4.0	5.4
		L	0.150	0.200	5.5	7.1
		S	0.255	0.360	7.0	9.1
		XS	2.789	0.920	9.6	15.0

Table 9: Comparison of prediction accuracy of the nonlinear aggregated model (AM) and the multiplicative choice model (CM). These numbers represent the *out-of-market* fit for each model. Model fit is assessed in terms of weighted KL-divergence and Mean Percentage Error (MPE), as described in section 4.4.

lack of data is likely to have a negative effect on the accuracy of the estimated utilities. The aggregated model does not suffer so much from this problem, as it uses all quarter hour observations to estimate its parameters. Increasing the number of segments will also increase the number of parameters to estimate in the aggregated model, but this does not have a drastic effect as there are still enough observations available.

Table 9 also shows that both models perform significantly worse for smaller channels. For the largest channel size, the KL divergence is much smaller than for the smaller channel sizes. This holds for both models, for all types of segmentation. The difference in MPE is not extreme, but nonetheless significant. For the smallest channels, the estimates are close to random guesses.

The aforementioned results are also observable in the index plots in figure 10, displaying the indices of the estimates of both the choice model and aggregated model estimates relative to the true index of TV ratings of the different age segments. The cloud of scatter points displaying the indices of the aggregated model follows the optimal diagonal relatively closely for the largest channels. For the *large* channels, the indices also follow the diagonal, albeit not as closely. The cloud of the choice model indices is much more horizontal: its estimates tend more to the mean. The estimates of the smallest category of channels are spread across the entire plot, indicating that for those channels, neither model produces accurate results. The estimates of the second smallest channels are neither very accurate. Therefore, we can again conclude that the models only work for observations (a single channel in a single quarter hour) where at least 10 households watched that channel.

According to table 9, the choice model and aggregated model produce relatively similar results in terms of accuracy, when segmentation is imposed on gender. However, as the index plots in 11 show,

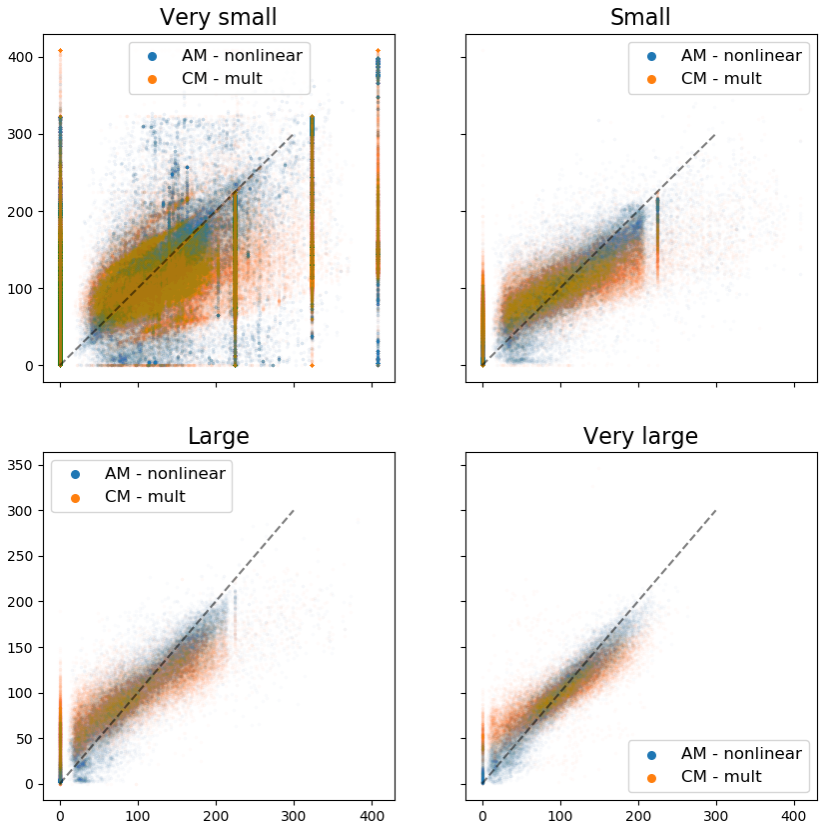


Figure 10: Index plot displaying the *out-of-market* fit of the nonlinear aggregated model and the multiplicative choice model, imposing segmentation based on age groups. The plot shows indices of estimates of the full month of Cleveland data. Segments are defined as Young (ages 18-34), Middle-aged (ages 35-54) and Old (ages 55+). Channel sizes are defined as in table 5.

their estimates differ quite a bit in location. Here, the aggregated model produces estimates that are closer to the mean, while the choice model takes on more extreme values. In this case, the estimates of the choice model actually seems to fit a bit better than the nonlinear aggregated model, especially for the *small* and *large* channel categories.

The accuracy numbers in table 9 slightly favour the aggregated model over the choice model when segmentation is done on education levels. As figure 13 in appendix E shows, this is mostly because the estimates of the choice model tend a bit towards the population mean. Nevertheless, the difference between the two methods here is not very large in terms of model fit.

Similar conclusions can be drawn for segmentation imposed on gender and age, plotted in figure 14. The fit of the aggregated model is slightly better than that of the choice model. However, the models have scatter clouds that are similarly shaped. The scatter clouds of both models are much more widely spread across the diagonal, indicating that the estimated TV ratings are generally further of the actual TV ratings than for other segmentation types. Some accuracy is obviously lost due to the increase in number of segments. Compared to the index plots with lower number of segments, the

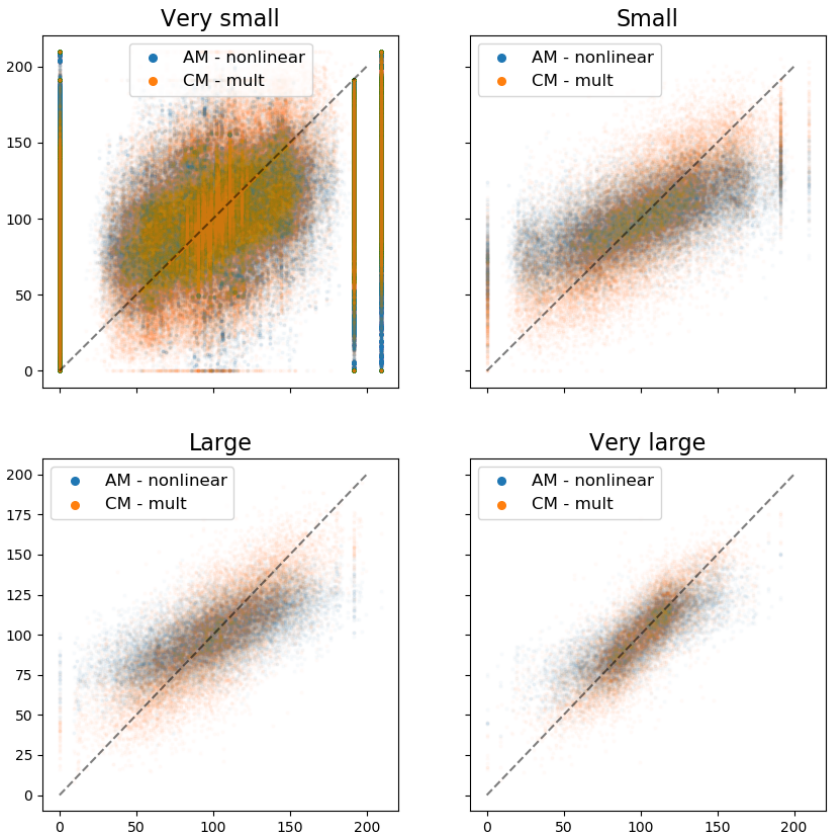


Figure 11: Index plot displaying the *out-of-market* fit of the nonlinear aggregated model and the multiplicative choice model, imposing segmentation based on gender. The plot shows indices of estimates of the full month of Cleveland data. Channel sizes are defined as in table 5.

observations in the *small* and *large* channel size buckets appear to be much less accurate. It seems that when the number of segments is 6, only the observations in the *very large* category can reasonably be estimated. The minimum number of households viewing a channel to estimate segment TV ratings with a reasonable accuracy therefore doubles from 10 to 20 when the number of segments doubles from 3 to 6.

Comparing the shapes of the plots in figures 10, 11 and 13, the conclusion can be drawn that the largest difference in viewing behaviour seems to be between age segments. For gender segments, many of the true indices of segment TV ratings are around 100, indicating that often the total quarter hour channel TV ratings ψ_{st} are relatively evenly split between males and females. For education, a similar conclusion can be drawn, especially for the larges channels. For age groups however, the distinction is more clear. The true and estimated indices in the bottom plots of figure 10 show indices ranging from close to zero (indicating almost no TV viewers for that channel in that segment) to above 200 (indicating that almost all viewers from that station are from a single age segment).

5.3.2 Computational feasibility

The computation times of the aggregated model and choice model in table 10 shows a large difference between the two models. The process of making estimates of the segment TV ratings for all quarter hours and channels in the data of Cleveland using the choice model takes two to three hours when imposing segmentation based on age, gender or education, yielding 2 or three different channels. This is no surprise, as the choice model has to be evaluated multiple times (once for each set of 50 channels) per quarter hour. The model estimation and out-of-sample prediction process of the aggregated model when segmenting on age, gender or education separately takes approximately 10 to 20 minutes, significantly shorter than the computation time of the choice model. Moreover, this difference increases when the number of segments is larger. When segmenting on both gender and age, yielding 6 segments in total, the aggregated model requires approximately half an hour computation time. For the choice model, this is approximately half a day. Therefore, the straightforward conclusion to draw from table 10, is that the aggregated model is to be preferred over the choice model in terms of computational feasibility.

Segments		**Model**	
Type	Number	AM	CM
Gender	2	00:13:53	02:36:34
Age	3	00:10:48	02:40:19
Education	3	00:19:09	02:09:46
Age-Gender	6	00:32:52	12:05:05

Table 10: Comparison of computation times between the nonlinear aggregated model and multiplicative choice model for different types of segmentation.

The computational advantage of the aggregated model over the choice model is not very surprising. The nonlinear aggregated model is a far more complicated model than the choice model and generally requires more time per estimation. However, choice model has the disadvantage of having to be estimated for each quarter hour seperately, per subset of 50 channels. This implies that estimating segment TV ratings for the full month of data in Cleveland requires $2,688 \cdot 7 = 18,816$ numerical optimizations of the log-likelihood in (12). For this same period, the aggregated model's parameters only have to be estimated A times.

6 Conclusions and Discussion

In this research, methodology has been developed to estimate TV ratings per demographic segment, using only household level TV viewing behaviour. We have identified two different methods to obtain these segment TV ratings: The choice model and the aggregated method. Furthermore, we assessed for each method which model or technique is the most suitable.

The first method applied a choice model to estimate probabilities for individuals from each segment to watch a certain channel, and used these probabilities to estimate segment TV ratings.

The observed household choices for a certain channel were modelled by a trade-off between each household's members' preferences. This is achieved by modelling the household utilities as a function of the utilities of the household's individuals' utilities. Maximizing the log-likelihood of this household-individual choice model yielded estimates of the utilities individuals in a segment gain from watching a certain channel. Using these segment utilities for all segments and the total TV ratings of the channel, the TV ratings per channel per segment could be obtained through a small number of straightforward calculation steps.

The choice model is a well-known and researched econometric model. Estimating group utility from individual utilities has received a lot of academic attention as well. However, combining these two by implementing group utility functions in the choice model has received relatively limited attention in literature so far. Therefore, we have tried numerous known group utility functions to find out which one is the most suitable. Four straightforward options were to take the sum, product, maximum or minimum of the individuals' utilities as household utility. Two other utility specifications involved the use of the additive and the multiplicative individual utility specifications, respectively, in a multinomial logit (MNL) model.

The multiplicative household utility specification has proven to be the most useful. Taking the maximum or the sum of individual utilities proved to be suboptimal, yielding a choice model that produced inaccurate TV ratings with a relatively long computation time. Taking the minimum of individual utilities as household utility produced an unstable choice model with again a long computation time. The two MNL-approaches and the multiplicative choice model produced similar results as the multiplicative specification in terms of accuracy, but the multiplicative approach appeared to require less computation time. In addition, the estimates of the MNL-models were pushed too much towards the population mean.

The second method, the *aggregated method*, attempted to directly formulate a function to obtain the segment TV ratings, using aggregated household type TV ratings, information on the date and time, and channel size as predictors. A household's type is defined by the number of individuals from each segment that are part of it. The aggregated household type TV ratings then are the number of people living in households of that type, watching the TV channel of interest at that point in time. To be able to apply this model on other markets, these household type ratings were scaled by the percentage of people from the segment of interest living in that specific household type.

The aggregated method is estimated with two different estimators. First, a linear model regression model was specified, with the household ratings and time-date information as predictors and the deviation of the segment TV rating of a channel from the population channel rating at that time as dependent variable. Second, the relation between the segment channel ratings and the predictors was modeled with the nonlinear tree-based method LightGBM. The hyperparameters of this LightGBM model were optimized using Bayesian Hyperparameter Optimization. Both estimators were trained on the observed household viewing behaviour and known segment TV ratings of the last three weeks of the Chicago data. Using these models, *within-market* estimates were obtained for the first week of

Chicago, and *out-of-market* estimates were obtained for the Cleveland data.

Of the two estimators of the aggregated method, the nonlinear model (using LightGBM) proved to be generally better than the linear model. Its estimates were much closer to the actual segment TV ratings in terms of *within-market* fit. In terms of *out-of-market* fit, the two models achieved similar results for the gender segment ratings. For other segments though, the nonlinear model produced more accurate estimates than the linear model. A plausible reason for the better model fit, is the fact that the nonlinear model can capture relations that only hold for a subset of the data, by using complicated interaction effects. The downside of the nonlinear model as opposed to the linear model is its computation time. Due to the model complexity and the fact that it is not analytically solvable, the nonlinear model requires up to ten times more computation time than the linear model. However, as the nonlinear model's computation time never went much above 30 minutes to estimate the model parameters and calculate *within-market* and *out-of market* estimates, this does not seem to be too problematic.

When comparing the two best configurations of the two methods, the nonlinear aggregated model generally outperforms the multiplicative choice model. In terms of accuracy, the two methods had similar results for the gender segments. For the age segments, the aggregated model performed much better than the choice model. For other two segment specifications, education and gender & age, the aggregated model also produced more accurate results than the choice model. However, for these segmentation types, the difference was not very large.

The aggregated model also has a computational advantage over the choice model. Estimating the choice model on a full month requires much time, as the model has to be estimated multiple times for every quarter hour separately: once for every subset of 50 channels. This adds up to over 10,000 estimations for the four weeks of available data. Therefore, obtaining estimates for the segment TV ratings for the full month of Cleveland data required several hours of computation times for the gender, education and age segments (separately), and almost half a day when imposing segmentation on age and gender together. This is mostly more than 10 times longer than the aggregated model's computation time, which is only estimated A times.

The aggregated model does have a practical downside over the choice model. The aggregated model needs a dataset with the true segment TV ratings to initialize its parameters. It therefore captures relations between variables in this estimation market that do not necessarily hold in other markets. Furthermore, the aggregated model estimates segment TV ratings based on household type TV ratings. Application of an aggregated model estimated on a different market therefore only works if there is a large overlap in household types between these markets.

Consequently, when applying the aggregated model to data from sources such as RPD, where no individual or segment level viewing behaviour is available, we have to be sure that the market the aggregated model is trained on is similar to the RPD market. In case there is no similar market with individual level data for the RPD market for which segment ratings are desired, one has to resort to the choice model.

In this research, we estimated the aggregated method on supervised data from Chicago, US and applied it to data from Cleveland, US. These two cities appeared to be similar enough in terms of culture (two cities in Northern US, approximately 500 kilometres apart) and in terms of household type overlap (which is 50-75%, depending on the segments imposed): the estimates of the segment TV ratings obtained using the model trained on the Chicago market were reasonably close to the true segment TV ratings.

A downside of both models is the fact that they can only be used on a limited number of segments. In this research, the models estimated on each of the gender, education or age segments separately, using a maximum of 3 segments per model, yielded relatively good results. The quality of the TV ratings estimates reduced fairly much when increasing the number of segments to 6. For both models, a major reason for this is the smaller data that is available to estimate these ratings. Besides, the number of quarter hours in which a channel is watched by only one individual from a segment is much higher when the number of segments is higher. For the aggregated model, an additional factor that reduces model performance is the decreased overlap in household types. Because of this smaller overlap, the model does not have access to the same amount of information it was trained on in the test set, yielding less accurate results.

Furthermore, both models do not yield accurate estimates of the segment TV ratings for every channel and every quarter hour. Depending on the number of segments, the models need at least 10 to 20 households viewing the specific channel at the given quarter hour to be able to make reasonably accurate estimates of the segment TV ratings.

6.1 Suggestions for future research

Based on the results from this research, a number of suggestions for future research in this field can be made.

Firstly, the linear aggregated model in (35) could be extended to a linear dynamic panel model, including previous (estimated) segment TV ratings in its predictors. To estimate the resulting model, an estimator like the estimator of Arellano and Bond (1991) could be used. This might lead to a better predictive accuracy of the linear model. However, it may also lead to longer computation times (as the predictions have to be made per time step). Furthermore, it is doubtful whether the Arellano-Bond estimator will improve the linear aggregated model so much that it will match the nonlinear model's predictive power.

Secondly, the aggregated model could be trained on data from a larger number of geographical regions, to apply it on a dissimilar market. This way, two issues of the aggregated model might be resolved. Firstly, the probability that all household types in the test market are in the model trained on such a large scale dataset. Furthermore, the model will not capture market-specific behaviour and is therefore generally better applicable on differing markets. Therefore, the necessity for a similar training market could be resolved. However, the model fit on the test market might also be reduced because the number of household types in the training data missing in the test data is likely to be

large.

Lastly, it could be interesting to explore possibilities to improve the *out-of-market* fit of the nonlinear aggregated model. In this research, there appeared to be a significant difference between the *within-market* fit and the *out-of-market* fit of the nonlinear aggregated model. Hence, the model was slightly overfitted on the training market. A possible measure to improve this *out-of-market* fit would be to use data from a different market as validation set in the Bayesian Hyperparameter Optimization algorithm. Whether this improved *out-of-market* fit could then be assessed by applying the resulting model on data from a third market.

REFERENCES

Abdi, H. (2010). Coefficient of variation. *Encyclopedia of research design*, *1*, 169–171.

Alderman, H., Chiappori, P.-A., Haddad, L., Hoddinott, J., & Kanbur, R. (1995). Unitary versus collective models of the household: is it time to shift the burden of proof? *The World Bank Research Observer*, *10*(1), 1–19.

Arellano, M., & Bond, S. (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *The review of economic studies*, *58*(2), 277–297.

Bellman, R. E. (1961). *Adaptive control processes: a guided tour* (Vol. 2045). Princeton University Press.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, *13*(Feb), 281–305.

Bergstra, J. S., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for hyper-parameter optimization. In *Advances in neural information processing systems* (pp. 2546–2554).

Brock, H. W. (1980). The problem of "utility weights" in group preference aggregation. *Operations Research*, *28*(1), 176–187.

Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine Learning* (pp. 161–168).

Chang, R. M., Kauffman, R. J., & Son, I. (2012). Consumer micro-behavior and TV viewership patterns: data analytics for the two-way set-top box. In *Proceedings of the 14th Annual International Conference on Electronic Commerce* (pp. 272–273).

Corfman, K. P., & Gupta, S. (1993). Mathematical models of group choice and negotiations. *Handbooks in operations research and management science*, *5*, 83–142.

Curry, D. J., Menasco, M. B., & Ark, J. W. V. (1991). Multiattribute dyadic choice: Models and tests. *Journal of Marketing Research*, *28*(3), 259–267.

Dewancker, I., McCourt, M., & Clark, S. (2015). *Bayesian optimization primer.*

Domenich, T., & McFadden, D. (1975). *Urban travel demand: a behavioural approach.* North-Holland Publishing Co.

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, *77*(4), 802–813.

Farahat, A., & Bailey, M. C. (2012). How effective is targeted advertising? In *Proceedings of the 21st international conference on world wide web* (pp. 111–120).

Fonseca, E., Gong, R., Bogdanov, D., Slizovskaia, O., Gómez Gutiérrez, E., & Serra, X. (2017). Acoustic scene classification by ensembling gradient boosting machine and convolutional neural networks. In *Detection and Classification of Acoustic Scenes and Events, 2017 workshop (DCASE2017)* (pp. 37–41).

Friedman, J. H. (1999). Stochastic gradient boosting. *Computational statistics & data analysis, 38*(4), 367–378.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232.

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland, 15*, 246–263.

Gensch, D., & Shaman, P. (1980). Models of competitive television ratings. *Journal of Marketing Research, 17*(3), 307–315.

Harsanyi, J. C. (1955). Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of political economy, 63*(4), 309–321.

Heij, C., de Boer, P., Franses, P. H., Kloek, T., & van Dijk, H. K. (2004). *Econometric methods with applications in business and economics.* Oxford, United Kingdom: Oxford University Press.

Hogg, R. V. (1979). An introduction to robust estimation. In *Robustness in statistics* (pp. 1–17). Elsevier.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

Jarque, C. M., & Bera, A. K. (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics letters, 6*(3), 255–259.

Jarque, C. M., & Bera, A. K. (1987). A test for normality of observations and regression residuals. *International Statistical Review/Revue Internationale de Statistique*, 163–172.

Jones, D. R., Schonlau, M., & Welch, W. J. (1998). Efficient global optimization of expensive black-box functions. *Journal of Global optimization, 13*(4), 455–492.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems* (pp. 3146–3154).

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics, 22*(1), 79–86.

Lau, L. J. (1987). Testing and imposing monoticity, convexity and quasi-convexity constraints. *Production Economics: A Dual Approach to Theory and Applications, 1*, 409-453.

Li, C. (2016). A gentle introduction to gradient boosting.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news*, *2*(3), 18–22.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. Courier Corporation.

Ma, X., Sha, J., Wang, D., Yu, Y., Yang, Q., & Niu, X. (2018). Study on a prediction of P2P network loan default based on the machine learning lightGBM and XGboost algorithms according to different high dimensional data cleaning. *Electronic Commerce Research and Applications*, *31*, 24–39.

McFadden, D., et al. (1973). Conditional logit analysis of qualitative choice behavior.

Meyer, D., & Hyndman, R. J. (2005). The accuracy of television network rating forecasts: The effects of data aggregation and alternative models. *Model Assisted Statistics and Applications*, *1*(3), 147–155.

Nash, J. F. (1950). The bargaining problem. *Econometrica: Journal of the Econometric Society*, 155–162.

Nash, J. F. (1953). Two-person cooperative games. *Econometrica: Journal of the Econometric Society*, 128–140.

Perner, P., & Trautzsch, S. (1998). Multi-interval discretization methods for decision tree learning. In *Joint IAPR international workshops on statistical techniques in pattern recognition (SPR) and structural and syntactic pattern recognition (SSPR)* (pp. 475–482).

Rawls, J. (1971). *A theory of justice*. Harvard university press.

Rust, R. T., & Alpert, M. I. (1984). An audience flow model of television viewing choice. *Marketing Science*, *3*(2), 113–124.

Rust, R. T., Kamakura, W. A., & Alpert, M. I. (1992). Viewer preference segmentation and viewing choice models for network television. *Journal of Advertising*, *21*(1), 1–18.

Shachar, R., & Emerson, J. W. (2000). Cast demographics, unobserved segments, and heterogeneous switching costs in a television viewing choice model. *Journal of Marketing Research*, *37*(2), 173–186.

Shafer, J., Agrawal, R., & Mehta, M. (1996). SPRINT: A scalable parallel classifier for data mining. In *VLDB* (Vol. 96, pp. 544–555).

Sharp, B., Beal, V., & Collins, M. (2009). Television: Back to the future. *Journal of Advertising Research*, *49*(2), 211–219.

Simmons Media Studies, N. Y. (1978). Selective markets and the media reaching them [Computer software manual]. Simmons Media Studies New York.

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems* (pp. 2951–2959).

Solgaard, H. S., & Hansen, T. (2003). A hierarchical Bayes model of choice between supermarket formats. *Journal of retailing and Consumer Services*, *10*(3), 169–180.

Sutton, C. D. (2005). Classification and regression trees, bagging, and boosting. *Handbook of statistics*, *24*, 303–329.

Thomas, J. (2019). *Gradient boosting in automatic machine learning: feature selection and hyperparameter optimization* (Unpublished doctoral dissertation). Ludwig-Maximilians-Universität München.

Thornton, C., Hutter, F., Hoos, H. H., & Leyton-Brown, K. (2013). Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms. In *Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 847–855).

Webster, J., & Phalen, P. F. (1997). *The mass audience: Rediscovering the dominant model.* Routledge.

Webster, J. G. (2005). Beneath the veneer of fragmentation: Television audience polarization in a multichannel world. *Journal of Communication*, *55*(2), 366–382.

Wu, C., Landgrebe, D. A., & Swain, P. (1975). The decision tree approach to classification.

Zhu, C., Byrd, R. H., Lu, P., & Nocedal, J. (1997). Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software (TOMS)*, *23*(4), 550–560.

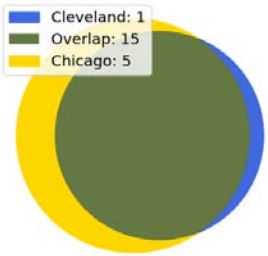
Appendices

A Overlap between household types

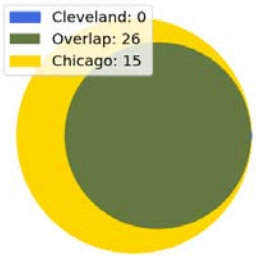
The following Venn diagrams show the overlap between household types, as defined in section 4.3, existing in both Chicago and Cleveland for a number of segment specifications.

A.1 Venn diagrams

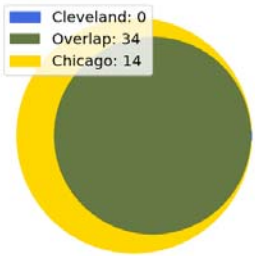
From these Venn diagrams, two conclusions can be drawn. Firstly, the number of household types in Chicago (displayed in yellow) is generally significantly larger than in Cleveland. This is rather unsurprising, as the number of households itself is larger in Chicago than in Cleveland as well. Secondly, the larger the number of segments, the smaller the relative overlap in household types. When gender is used as segment, the overlap between household types is 75% of the household types appearing in Chicago. However, when both age and gender are used as segments, this number drops to only 52%.



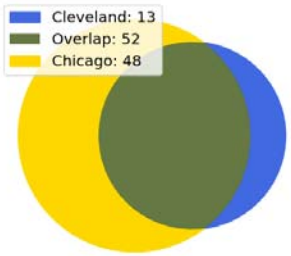
(a) Gender (2 segments). 75% of



(b) Age (3 segments): overlap 63%



(c) Education (3 segments): overlap 71%



(d) Gender & age (6 segments): overlap 52%

Figure 12: Venn diagrams of overlap between the household types existing in Chicago (displayed in yellow) and Cleveland (blue). The larger the number of segments gets, the smaller this overlap is, relatively.

A.2 Table of overlap

Segmentation	No. segments	% overlap Chicago-Cleveland	% overlap Cleveland-Chicago
Gender	2	75	94
Age	3	63	100
Education	3	71	100
Gender & Age	6	52	80

Table 11: Percentage overlap in household types between Chicago and Cleveland. The *% overlap Chicago-Cleveland* column contains the percentage of household types present in Chicago that also exist in Cleveland. The *% overlap Cleveland-Chicago* column contains the percentage of household types present in Cleveland that also exist in Chicago

B PARAMETER ESTIMATES OF LINEAR AGGREGATED MODEL

This appendix contains tables of parameter estimates of the linear aggregated model in (35) for different types of imposed segmentation. The variables with numerical variable names are each household-type ratings. Their variable names specifies the composition of the households in that household type. E.g. when segmentation is done on gender, each household type rating variable's name is of the form xy, where x is the number of females (segment 0) in that household type, and y is the number of males (segment 1) in that household type. For example, the household type '12' would in this case contain all household with exactly one female and two males.

The household types variables are scaled by the percentage of individuals from segment a living in a household type j, hence the parameters of the household types are for the scaled household types $c_{aj}z_{jst}$

B.1 Gender

Variable	Female (0)	Male (1)
intercept	-2.34	-121.51
01	5.83	0.00
02	8.39	0.00
04	0.00	0.00
10	0.00	9.57
11	-0.26	0.00
12	0.20	-0.92
13	3.12	-7.46
14	-1.61	4.53
15	86.66	-425.18

Continued on next page

Table 12 – continued from previous page

Variable	Female (0)	Male (1)
20	0.00	11.98
21	-2.07	0.57
22	-0.85	0.36
23	-3.33	2.81
24	11.57	-17.75
30	0.00	108.78
31	-10.83	2.43
32	-34.84	22.72
34	-20.50	27.63
41	-143.30	28.04
42	31.53	-16.49
tuesday	6.03	13.82
wednesday	13.37	-9.25
thursday	6.12	10.44
friday	20.19	-12.80
saturday	5.06	-18.87
sunday	-31.37	24.61
night	-70.57	105.05
morning	-29.45	82.43
evening	-66.02	34.32

Table 12: Parameters of the linear aggregated model segmented on gender. For the household types, their variable names are denoted in the format xy, where x is the number of females and y the number of males, respectively, in the household.

B.2 Age

.

Variable	Young(0)	Middle-age (1)	Old (2)
intercept	-100.16	-414.55	-74.71
001	10.43	0.00	0.00
002	2.88	0.00	0.00
003	87.42	0.00	0.00
004	18.25	0.00	0.00
010	0.00	5.47	0.00

Continued on next page

Table 13 – continued from previous page

Variable	Young (0)	Middle-age (1)	Old (2)
011	-2.90	0.89	0.00
012	7.60	-9.00	0.00
013	30.69	-50.99	0.00
020	0.00	0.93	0.00
021	-7.09	2.52	0.00
022	-2.89	1.62	0.00
023	-2.07	3.65	0.00
024	30.99	-3.96	0.00
030	0.00	76.30	0.00
031	-144.51	71.31	0.00
100	0.00	0.00	4.60
101	-20.73	0.00	-11.82
102	-2.34	0.00	-113.49
103	-13.53	0.00	-80.07
104	44.02	0.00	-276.42
110	0.00	-8.70	-5.61
111	-27.82	-10.22	-14.27
112	-11.53	-1.66	-14.70
113	-11.66	-24.94	-54.40
115	0.10	-20.14	15.39
120	0.00	-2.66	-59.41
121	-26.14	2.74	-19.97
122	-23.27	-1.67	-33.34
132	29.94	42.68	144.55
200	0.00	0.00	1.46
201	-13.97	0.00	1.08
202	-4.63	0.00	-13.43
203	-95.27	0.00	-45.72
204	-1.88	0.00	5.42
210	0.00	-16.43	-0.25
211	-10.67	-6.17	-17.65
212	-14.30	-64.89	0.72
220	0.00	24.22	-164.21
221	142.38	66.65	101.98
300	0.00	0.00	17.66

Continued on next page

Table 13 – continued from previous page

Variable	Young (0)	Middle-age (1)	Old (2)
301	-93.36	0.00	37.93
tuesday	-57.49	-79.39	89.37
wednesday	-55.85	-14.35	19.65
thursday	-40.75	-3.14	81.80
friday	-35.45	-68.17	118.71
saturday	-48.18	54.51	-77.49
sunday	-47.45	158.99	-165.20
night	185.46	397.63	-196.40
morning	-57.14	243.28	-76.55
evening	-156.43	71.53	-186.75

Table 13: Parameters of the linear aggregated model segmented on age. The household types variables here have the name format xyz, where x is the number of old, y is the number of middle-aged and z is the number of young people in the household. Segments are defined as follows: Young (0): ages 18-34. Middle-aged (1): ages 35-54. Old (2): Ages 55+

B.3 Education

Variable	Low education (0)	Medium education (1)	High education (2)
intercept	139.43	-121.20	-572.79
001	2.71	0.00	0.00
002	2.33	0.00	0.00
003	7.41	0.00	0.00
004	4.85	0.00	0.00
005	7.96	0.00	0.00
006	65.24	0.00	0.00
010	0.00	3.50	0.00
011	-2.26	-0.01	0.00
012	1.39	-7.53	0.00
013	4.50	-31.88	0.00
014	114.63	-393.76	0.00
015	25.22	-31.19	0.00
020	0.00	2.13	0.00
021	-17.31	5.84	0.00
022	-2.16	4.80	0.00
023	84.46	-129.26	0.00
030	0.00	9.95	0.00

Continued on next page

Table 14 – continued from previous page

Variable	Low education (0)	Medium education (1)	High education (2)
031	-17.21	8.01	0.00
032	-21.51	11.70	0.00
034	15.41	-13.99	0.00
040	0.00	6.90	0.00
042	13.33	6.52	0.00
100	0.00	0.00	4.61
101	-12.51	0.00	-4.99
102	4.58	0.00	-24.25
103	61.25	0.00	-235.48
110	0.00	-0.76	-1.79
111	-11.36	-8.52	-19.93
112	-4.55	-15.98	-43.05
113	-3.21	-25.37	-36.60
120	0.00	2.31	-2.80
121	-9.03	0.77	-15.69
122	-57.32	-12.26	-74.46
130	0.00	1.30	-30.17
131	-227.62	-47.73	-0.23
200	0.00	0.00	1.46
201	-16.47	0.00	2.52
202	8.70	0.00	-7.10
203	-13.92	0.00	3.32
210	0.00	-5.22	1.11
220	0.00	-4.57	0.89
221	-136.78	-39.80	9.75
240	0.00	-73.07	226.43
300	0.00	0.00	7.51
302	-17.17	0.00	3.64
310	0.00	-19.67	-1.00
400	0.00	0.00	-23.65
410	0.00	-55.05	23.90
tuesday	11.10	7.31	-45.18
wednesday	-27.18	34.14	5.29
thursday	25.25	43.67	-51.04
friday	-4.20	27.05	-29.56

Continued on next page

Table 14 – continued from previous page

Variable	Low education (0)	Medium education (1)	High education (2)
saturday	-73.98	32.46	0.53
sunday	-65.99	17.91	-28.96
night	-85.35	80.95	240.50
morning	-101.37	43.85	263.57
evening	-194.14	-133.77	82.71

Table 14: Parameters of the linear aggregated model when segmenting on education. The household types variable names are composed as xyz, where x is the number of highly educated, y is the number of medium educated and z is the number of low educated people in the household. Segments are defined as follows: Low education (0): education levels 0-3. Medium education (1): education level 4. High education: education levels 5 and 6.

B.4 Gender & Age

Variable	F 18-34	F 35-54	F 55+	M 18-34	M 35-54	M 55+
intercept	-99.85	-165.13	-29.53	-141.21	-300.57	-117.72
000001	10.53	0.00	0.00	0.00	0.00	0.00
000002	17.35	0.00	0.00	0.00	0.00	0.00
000010	0.00	9.46	0.00	0.00	0.00	0.00
000011	0.49	12.50	0.00	0.00	0.00	0.00
000013	11.33	-15.07	0.00	0.00	0.00	0.00
000020	0.00	20.84	0.00	0.00	0.00	0.00
000100	0.00	0.00	4.24	0.00	0.00	0.00
000101	-11.98	0.00	26.95	0.00	0.00	0.00
000110	0.00	-17.06	41.89	0.00	0.00	0.00
000200	0.00	0.00	16.00	0.00	0.00	0.00
001000	0.00	0.00	0.00	21.48	0.00	0.00
001001	0.95	0.00	0.00	1.33	0.00	0.00
001003	39.67	0.00	0.00	-22.47	0.00	0.00
001010	0.00	8.48	0.00	-6.01	0.00	0.00
001011	1.37	2.00	0.00	-3.47	0.00	0.00
001100	0.00	0.00	-20.57	-22.80	0.00	0.00
001101	390.72	0.00	-797.62	-297.48	0.00	0.00
001102	-5.49	0.00	4.59	-10.82	0.00	0.00
001103	17.07	0.00	-165.62	-139.67	0.00	0.00

Continued on next page

Table 15 – continued from previous page

Variable	F 18-34	F 35-54	F 55+	M 18-34	M 35-54	M 55+
001110	0.00	-7.77	0.50	-24.96	0.00	0.00
001200	0.00	0.00	102.68	-67.49	0.00	0.00
001210	0.00	29.26	5.32	-65.19	0.00	0.00
002000	0.00	0.00	0.00	8.60	0.00	0.00
002010	0.00	10.26	0.00	62.51	0.00	0.00
002100	0.00	0.00	14.06	-60.17	0.00	0.00
002110	0.00	27.94	-13.88	-1.82	0.00	0.00
002202	-3.96	0.00	12.44	2.57	0.00	0.00
003000	0.00	0.00	0.00	34.00	0.00	0.00
003001	2.27	0.00	0.00	12.24	0.00	0.00
003010	0.00	-33.19	0.00	32.22	0.00	0.00
010000	0.00	0.00	0.00	0.00	10.37	0.00
010001	-0.14	0.00	0.00	0.00	2.56	0.00
010010	0.00	-0.02	0.00	0.00	0.02	0.00
010011	-2.45	0.12	0.00	0.00	-0.08	0.00
010012	-8.16	11.24	0.00	0.00	-4.32	0.00
010100	0.00	0.00	14.14	0.00	-17.57	0.00
010101	-36.03	0.00	16.18	0.00	-32.49	0.00
010102	-42.42	0.00	51.43	0.00	-96.38	0.00
010110	0.00	-1.17	-25.62	0.00	5.13	0.00
010111	-5.70	0.51	10.22	0.00	-4.14	0.00
010112	-14.35	-0.29	9.46	0.00	-21.48	0.00
010122	13.27	21.31	54.91	0.00	0.44	0.00
011000	0.00	0.00	0.00	15.34	11.09	0.00
011001	12.55	0.00	0.00	0.35	3.06	0.00
011010	0.00	5.40	0.00	-5.05	-0.89	0.00
011011	-4.89	-1.21	0.00	-4.64	-0.71	0.00
011012	-0.86	-0.94	0.00	-1.29	3.87	0.00
011020	0.00	28.02	0.00	-33.18	21.87	0.00
011100	0.00	0.00	-3.40	-29.90	-2.91	0.00
011101	-205.57	0.00	24.39	-182.29	-134.75	0.00
011110	0.00	2.82	-19.40	-21.22	-8.06	0.00
011111	-1.29	9.23	-122.93	-4.52	8.08	0.00
012010	0.00	-9.28	0.00	2.62	7.86	0.00
013010	0.00	-11.79	0.00	-6.41	5.79	0.00

Continued on next page

Table 15 – continued from previous page

Variable	F 18-34	F 35-54	F 55+	M 18-34	M 35-54	M 55+
013011	2.04	-2.49	0.00	12.89	-3.94	0.00
020000	0.00	0.00	0.00	0.00	16.84	0.00
020010	0.00	-1.17	0.00	0.00	41.90	0.00
020100	0.00	0.00	-0.32	0.00	-7.29	0.00
100000	0.00	0.00	0.00	0.00	0.00	5.75
100010	0.00	-1.02	0.00	0.00	0.00	2.23
100011	-149.22	86.83	0.00	0.00	0.00	-110.58
100012	-85.02	-4.85	0.00	0.00	0.00	42.59
100100	0.00	0.00	0.23	0.00	0.00	0.31
100101	-7.97	0.00	-7.08	0.00	0.00	-0.91
100102	-0.06	0.00	15.01	0.00	0.00	-66.80
100110	0.00	-2.43	-0.34	0.00	0.00	-1.17
100111	15.40	15.34	-9.19	0.00	0.00	-30.79
100112	-6.47	-40.09	-19.61	0.00	0.00	19.44
100200	0.00	0.00	14.73	0.00	0.00	-8.91
101000	0.00	0.00	0.00	-23.65	0.00	-181.29
101001	-4.12	0.00	0.00	22.33	0.00	17.20
101010	0.00	-22.94	0.00	-34.75	0.00	-3.34
101011	-29.52	8.10	0.00	-29.39	0.00	-9.19
101012	-8.03	-8.55	0.00	-18.11	0.00	-1.04
101100	0.00	0.00	0.49	-6.54	0.00	-1.74
101101	-2.78	0.00	-15.38	-3.13	0.00	-0.72
101103	-69.69	0.00	-244.23	-209.50	0.00	146.48
101110	0.00	7.65	-133.65	3.68	0.00	-81.11
101111	-40.13	-73.55	-251.66	-26.64	0.00	37.10
101200	0.00	0.00	46.58	-82.28	0.00	-20.39
102002	24.27	0.00	0.00	37.29	0.00	-40.80
102010	0.00	1.53	0.00	-54.87	0.00	-27.53
102011	9.28	-58.36	0.00	5.87	0.00	-65.82
102013	1.23	4.98	0.00	-7.54	0.00	13.85
102100	0.00	0.00	-72.43	-25.35	0.00	-57.91
103100	0.00	0.00	-58.80	-15.05	0.00	6.44
110000	0.00	0.00	0.00	0.00	19.59	9.73
110001	-144.00	0.00	0.00	0.00	208.83	279.33
110010	0.00	141.06	0.00	0.00	-43.25	-150.47

Continued on next page

Table 15 – continued from previous page

Variable	F 18-34	F 35-54	F 55+	M 18-34	M 35-54	M 55+
110011	-111.49	-54.00	0.00	0.00	60.40	21.29
110100	0.00	0.00	-9.16	0.00	-8.35	-8.96
110101	-99.56	0.00	-147.47	0.00	15.75	126.73
110110	0.00	-19.05	-103.73	0.00	28.75	-32.24
111001	-68.05	0.00	0.00	-72.03	-41.63	-18.64
111100	0.00	0.00	-5.18	-18.33	-4.55	13.07
120101	-1.60	0.00	-83.34	0.00	-7.44	12.98
200000	0.00	0.00	0.00	0.00	0.00	-7.25
200011	-37.31	-82.33	0.00	0.00	0.00	14.46
200100	0.00	0.00	4.99	0.00	0.00	10.44
201100	0.00	0.00	14.24	-29.45	0.00	16.87
tuesday	-35.44	-0.68	14.59	-14.41	-43.38	55.34
wednesday	-12.36	37.23	-50.53	-29.38	-31.50	36.39
thursday	0.28	9.89	-16.89	-27.39	3.76	43.58
friday	-11.95	-18.62	-15.47	-13.45	7.83	63.81
saturday	-19.52	-9.98	-126.10	-34.16	84.85	16.72
sunday	-22.96	-16.73	-197.44	-28.40	160.63	11.82
night	57.56	98.66	-127.10	131.25	218.95	-75.52
morning	-7.04	81.61	-41.94	29.43	157.61	-6.80
evening	-121.33	-2.02	-120.71	-115.53	-21.73	-76.04

Table 15: Parameter values for linear aggregated model segmented on gender and age. The household types variable names are of the form $abcdef$, where a are males aged 55+, b males aged 35-54, c are males aged 18-34, d are females aged 55+, e are females aged 35-54 and f are the number of females aged 18-34 in the households in that type. E.g. household type 200100 are households with 2 males aged 55+ and one female aged 55+.

C HYPERPARAMETERS OF THE NONLINEAR AGGREGATED METHOD

The table in this appendix contains the hyperparameters selected by the Gaussian Processes of the nonlinear aggregated model, as described in section 4.3.3.

Segmentation		Hyperparameters				
Type	Segment	Max depth	Learning rate	MDIL	Lambda L2	No. leaves
Gender	0 (female)	9	0.200	20	0.100	204
	1 (male)	12	0.200	20	0.000	2255
Age	0 (young)	11	0.107	20	0.043	1739
	1 (middle-age)	6	0.200	20	0.000	25
	2 (old)	12	0.200	20	0.000	3355
Education	0 (low)	12	0.200	20	0.012	4059
	1 (medium)	11	0.200	20	0.066	1646
	2 (high)	12	0.200	20	0.089	3578
Gender & Age	0 (young female)	5	0.200	10	0.100	31
	1 (middle-aged female)	10	0.183	20	0.000	893
	2 (old female)	12	0.200	20	0.069	3112
	3 (young male)	12	0.178	20	0.100	3711
	4 (middle-aged male)	12	0.196	20	0.100	3845
	5 (old male)	12	0.200	20	0.100	3060

Table 16: Hyperparameters as selected by the Gaussian Processes for the nonlinear aggregated model. 'minimum data in leaf' is abbreviated to 'MDIL'

D VARIABLE IMPORTANCE OF NONLINEAR AGGREGATED METHOD

Tables 17 - 20 display the feature importance per predictor in each LightGBM model as described in section 4.3.3. The feature importance here is defined as the number of times the variable has been used as split in a tree. The variable names of the household ratings z_{jst} are the same as in the tables of linear parameter estimates in appendix B.

D.1 Gender

Variable	Female (0)	Male (1)
01	1655	0
02	1076	0
04	22	0
10	0	2791
11	2079	3165
12	1797	2347
13	595	726
14	259	293

Continued on next page

Table 17 – continued from previous page

Variable	Female (0)	Male (1)
15	47	20
20	0	1408
21	1519	2472
22	1276	1976
23	499	714
24	71	82
30	0	111
31	781	1372
32	96	134
34	26	42
41	124	230
42	85	156
day_of_week	601	575
hour	1199	1175
benchmark_rating	2353	3157
channelsize	483	983

Table 17: Feature importance of the nonlinear aggregated model, when segmenting on gender.

D.2 Age

Variable	Young(0)	Middle-age (1)	Old (2)
001	936	0	0
002	2259	0	0
003	143	0	0
004	431	0	0
010	0	156	0
011	2233	148	0
012	1130	73	0
013	266	12	0
020	0	211	0
021	1510	148	0
022	1171	129	0
023	491	58	0

Continued on next page

Table 18 – continued from previous page

Variable	Young (0)	Middle-age (1)	Old (2)
024	311	49	0
030	0	21	0
031	0	0	0
100	0	0	3692
101	917	0	882
102	364	0	196
103	110	0	53
104	207	0	38
110	0	136	2013
111	714	79	904
112	773	87	785
113	274	39	277
115	51	2	43
120	0	75	411
121	278	70	324
122	153	32	256
132	64	19	34
200	0	0	3389
201	1511	0	2211
202	930	0	1014
203	76	0	42
204	109	0	119
210	0	110	2070
211	304	42	485
212	168	8	178
220	0	18	139
221	14	4	29
300	0	0	954
301	84	0	376
day_of_week	650	93	576
hour	1184	200	1274
benchmark_rating	5366	286	5615
channelsize	980	47	1019

Table 18: Feature importance of the nonlinear aggregated model, when segmenting on age.

D.3 Education

Variable	Low (0)	Medium (1)	High (2)
001	3223	0	0
002	3166	0	0
003	1398	0	0
004	699	0	0
005	477	0	0
006	79	0	0
010	0	2340	0
011	4093	2636	0
012	1716	1023	0
013	364	171	0
014	30	4	0
015	192	102	0
020	0	2105	0
021	999	1029	0
022	333	315	0
023	45	32	0
030	0	523	0
031	262	256	0
032	58	98	0
034	56	19	0
040	0	139	0
042	39	88	0
100	0	0	2322
101	2225	0	1559
102	770	0	420
103	59	0	25
110	0	2374	2411
111	720	522	491
112	645	352	333
113	313	117	149
120	0	1005	876
121	551	488	388
122	159	101	91
130	0	305	190
131	42	24	31

Continued on next page

Table 19 – continued from previous page

Variable	Low (0)	Medium (1)	High (2)
200	0	0	2381
201	729	0	865
202	130	0	123
203	32	0	25
210	0	867	1248
220	0	486	739
221	52	43	103
240	0	16	40
300	0	0	630
302	89	0	157
310	0	267	537
400	0	0	45
410	0	26	62
day_of_week	906	558	431
hour	1698	1206	945
benchmark_rating	7707	4034	3760
channelsize	1348	888	654

Table 19: Feature importance of the nonlinear aggregated model, when segmenting on education.

D.4 Gender & Age

Variable	F 18-34	F 35-54	F 55+	M 18-34	M 35-54	M 55+
000001	80	0	0	0	0	0
000002	66	0	0	0	0	0
000010	0	1512	0	0	0	0
000011	110	624	0	0	0	0
000013	15	2	0	0	0	0
000020	0	220	0	0	0	0
000100	0	0	4496	0	0	0
000101	44	0	475	0	0	0
000110	0	125	169	0	0	0
000200	0	0	950	0	0	0
001000	0	0	0	424	0	0

Continued on next page

Table 20 – continued from previous page

Variable	F 18-34	F 35-54	F 55+	M 18-34	M 35-54	M 55+
001001	323	0	0	3193	0	0
001003	20	0	0	24	0	0
001010	0	675	0	955	0	0
001011	109	390	0	605	0	0
001100	0	0	577	473	0	0
001101	19	0	34	52	0	0
001102	15	0	70	78	0	0
001103	26	0	28	14	0	0
001110	0	123	182	84	0	0
001200	0	0	122	29	0	0
001210	0	47	53	42	0	0
002000	0	0	0	483	0	0
002010	0	47	0	76	0	0
002100	0	0	63	58	0	0
002110	0	70	32	40	0	0
002202	8	0	112	70	0	0
003000	0	0	0	108	0	0
003001	33	0	0	237	0	0
003010	0	77	0	136	0	0
010000	0	0	0	0	1435	0
010001	169	0	0	0	1388	0
010010	0	3171	0	0	3971	0
010011	113	1173	0	0	1436	0
010012	44	147	0	0	184	0
010100	0	0	1229	0	958	0
010101	26	0	72	0	79	0
010102	13	0	66	0	66	0
010110	0	222	196	0	266	0
010111	14	88	107	0	114	0
010112	11	159	244	0	146	0
010122	19	50	43	0	23	0
011000	0	0	0	484	442	0
011001	70	0	0	228	230	0
011010	0	793	0	669	987	0
011011	91	639	0	872	895	0

Continued on next page

Table 20 – continued from previous page

Variable	F 18-34	F 35-54	F 55+	M 18-34	M 35-54	M 55+
011012	31	154	0	178	218	0
011020	0	0	0	0	0	0
011100	0	0	192	121	220	0
011101	0	0	36	27	27	0
011110	0	126	168	96	122	0
011111	14	61	67	70	74	0
012010	0	216	0	234	227	0
013010	0	108	0	105	98	0
013011	27	147	0	206	122	0
020000	0	0	0	0	328	0
020010	0	14	0	0	48	0
020100	0	0	189	0	107	0
100000	0	0	0	0	0	2567
100010	0	1170	0	0	0	1527
100011	20	147	0	0	0	110
100012	9	76	0	0	0	85
100100	0	0	5752	0	0	3849
100101	115	0	1165	0	0	1067
100102	13	0	28	0	0	17
100110	0	952	1453	0	0	1287
100111	15	33	29	0	0	33
100112	30	53	99	0	0	92
100200	0	0	273	0	0	190
101000	0	0	0	62	0	61
101001	19	0	0	77	0	114
101010	0	147	0	146	0	146
101011	24	171	0	89	0	152
101012	38	219	0	217	0	298
101100	0	0	1313	984	0	972
101101	85	0	658	686	0	719
101103	4	0	28	17	0	25
101110	0	55	43	49	0	74
101111	10	34	28	29	0	28
101200	0	0	191	92	0	127
102002	23	0	0	72	0	26

Continued on next page

Table 20 – continued from previous page

Variable	F 18-34	F 35-54	F 55+	M 18-34	M 35-54	M 55+
102010	0	87	0	80	0	98
102011	12	40	0	58	0	35
102013	20	28	0	22	0	41
102100	0	0	211	160	0	178
103100	0	0	45	71	0	36
110000	0	0	0	0	127	148
110001	9	0	0	0	51	36
110010	0	20	0	0	20	8
110011	4	38	0	0	85	45
110100	0	0	861	0	652	743
110101	12	0	87	0	38	44
110110	0	85	120	0	170	119
111001	16	0	0	140	155	149
111100	0	0	85	51	38	62
120101	1	0	13	0	8	27
200000	0	0	0	0	0	112
200011	8	24	0	0	0	36
200100	0	0	461	0	0	505
201100	0	0	82	22	0	158
day_of_week	109	720	962	763	764	749
hour	186	1302	1791	1254	1374	1407
benchmark_rating	309	4946	7763	6477	5559	5246
channelsize	71	959	1365	1000	1118	1091

Table 20: Feature importance of the nonlinear aggregated model, when segmenting on gender and age.

E INDEX PLOTS OF COMPARISON

This appendix contains two index plots comparing the fit of the choice model and aggregated model when imposing segmentation based on education (section E.1) or gender and age (section E.2).

E.1 Education

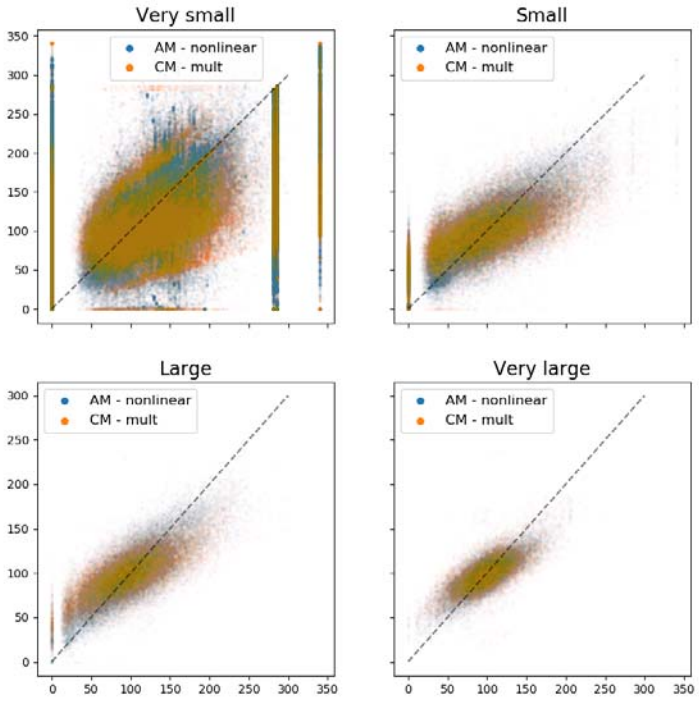


Figure 13: Index plot displaying the *out-of-market* fit of the nonlinear aggregated model and the multiplicative choice model, imposing segmentation based on education. The true index is on the x-axis, the estimated index on the y-axis. The plot shows indices of estimates of the full month of Cleveland data. Channel sizes are defined as in table 5.

E.2 Gender & age

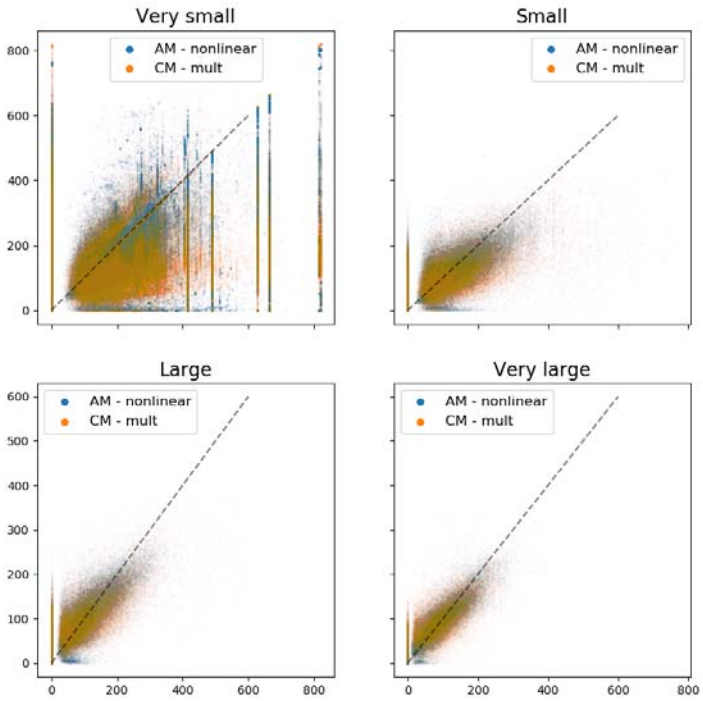


Figure 14: Index plot displaying the *out-of-market* fit of the nonlinear aggregated model and the multiplicative choice model, imposing segmentation based on age and gender. The true index is on the x-axis, the estimated index on the y-axis. The plot shows indices of estimates of the full month of Cleveland data. Channel sizes are defined as in table 5.