The accuracy of television network rating forecasts: The effects of data aggregation and alternative models

Denny Meyer\textsuperscript{a,∗} and Rob J. Hyndman\textsuperscript{b}
\textsuperscript{a}Faculty of Life and Social Sciences, Swinburne University of Technology, PO Box 218, Hawthorn VIC 3122, Australia
\textsuperscript{b}Department of Econometrics and Business Statistics, Monash University, Australia

Abstract. This paper investigates the effect of aggregation in relation to the accuracy of television network rating forecasts. We compare the forecast accuracy of network ratings using population rating models, rating models for demographic/behavioural segments and individual viewing behaviour models. Models are fitted using neural networks, decision trees and regression. The most accurate forecasts are obtained by aggregating forecasts from segment rating models, with neural networks being used to fit these models. The resulting models allow for interactions between the variables and the non-linear carry-over effect is found to be the most important predictor of segment ratings, followed by time of day and then genre. The analysis differs from those of previous authors in several important respects. The AC Nielsen panel data considered stretches over 31 days, 24 hours per day, 60 minutes per hour, making it necessary for ratings to be appropriately transformed prior to the fitting of the rating models and for non-viewing time periods to be under-sampled when fitting the models for individual viewing. For the first time individual viewing within each 15 minute time period is defined by network choice and proportion of viewing time.

Keywords: Aggregation, discrete choice models, neural networks, decision trees, two-stage models

1. Introduction

According to recent data published by AdWatch [1], television advertising makes up 44.8\% of the total advertising spend in the United States of America. Television executives sell time to advertisers at a price that is estimated using forecast television network ratings; however, these audience projections are often made before any program episodes have been aired. Forkan [9] and Rust and Eechambadi [22] maintain that these forecasts have been very inaccurate in the past. Indeed, Napoli [18] has found that forecasting error has increased significantly over time.

Networks typically reschedule programs when actual audience sizes do not reach projected levels. Advertisers are refunded when this happens, but they are not compensated for the disruption to their media plan. Accurate forecasts are therefore important and advertisers are prepared to pay for more reliable audience information [10,31]. This research adds to the literature on improving the accuracy of short-term rating forecasts. In particular it investigates the influence of forecast aggregation on forecast accuracy and the importance of allowing for non-linearity and interaction effects when developing rating models. It also investigates the relative importance of carry-over effects from one time period to the next, genre, scheduling and audience segment in determining rating forecasts.

Aggregation is an important research topic in time series analysis. A book on temporal aggregation for ARMA processes written by Quenouille in 1957 is an early example. More recently Granger and Lee [13] have found that data aggregation serves to simplify non-linearity, Tiao [29] and Breitung and Swanson [8] have
found that causality information is lost as a result of data aggregation, while Zellner and Tobias [32] have shown that data aggregation reduces forecasting accuracy. In the context of agricultural economics Shumway and Davis [25] claim that inferential errors due to data aggregation are small relative to model estimation errors. However, in other areas such as labour demand, it has been found that aggregation bias does occur [17]. In this paper we consider the effect of aggregation on forecasting accuracy for television ratings. The research question is “To what extent should the data be aggregated before forecasting network ratings?”. Another way to ask this question is “To what extent should forecasts for network ratings be aggregated as an alternative to aggregation of the data prior to model estimation and forecasting?”. Linear regression models for the prediction of population ratings have been used by several authors [3,4,16,21]. These models are all short-term in nature, being typically derived from prime time viewing for only a few weeks of the year, often only one week, using AC Nielsen panel data for viewing in 15 minute or 30 minute periods. Most of these rating models have considered network, time of day, day of the week, and the rating for the previous period as the predictor variables in their models. However, Shachar and Emerson [24] have incorporated program dummies, cast demographics and switching costs in their rating models. The fitting of such rating models has been criticised by Rust and Echambadi [22] because audience demographics are not taken into account. In answer to this criticism several authors have used choice models (nominal logistic regression) to analyse individual network choices, obtaining ratings by aggregating the predictions from their models [12,22,24,27,28]. Some authors have included a behavioural and/or demographic segment in their model with Rust et al. [23] recommending that segments with homogeneous viewing are preferable to a segmentation based on demographics or life style. These authors developed network choice models for each of three demographic/geographic segments and then combined the predicted network probabilities for each segment using the relative size of each segment as a weight. Models for individual network choice have also included network, time of day, day of the week and genre. All the above models for network choice are fitted using logistic regression methods. Swann and Tavakoli [27] allow for interactions between program genre and demographics (gender, age and socio-economic level) with product terms and, in the same way, Kelton and Stone [16] allow for genre interactions with time of day and day of the week.

There is clearly a large literature dealing with models for television viewing. However, there are some notable omissions from this literature. Interestingly, perhaps on account of the emphasis on prime time viewing, the distribution of rating model errors is assumed normally distributed despite the fact that ratings are proportions and therefore heteroscedastic. It is expected that an arcsine square root transformation will produce more reliable rating models while still allowing for zero ratings [26, p. 386]. All authors appear to ignore the relatively high percentage of non-viewer records in the data sets used to estimate network choice models for individuals. In our data set the percentage of records for non-viewers was nearly 90%. In these circumstances it is necessary to under-sample the non-viewers or oversample the viewers for each time period in order to obtain reliable models [5, p. 78]. Our models learn by example so we need a reasonable proportion of cases in each of the categories of interest.

In addition it appears that no authors have made allowance for network switching within a time period, perhaps because they did not have this information. Also there is no comparison of the performance of non-linear models fitted using decision trees or neural networks with the performance of simpler models fitted using regression or logistic regression. Finally, to the best of the authors’ knowledge the literature contains no comparison of the accuracy of rating forecasts obtained from data at different aggregation levels (e.g. population, segment or individual). This paper addresses these issues directly while investigating the following expectations in the current context.

1) It is expected that for rating models neural networks will perform best because they allow for interaction effects and for any non-linear rating carry-over effects from one period to the next. Despite their ability to model interaction effects, trees are not expected to manage the carry-over effect as well as neural networks, requiring too many branches in order to achieve the same accuracy as neural networks. Regression models are expected to perform worse than neural networks when they are not specifically designed to describe interaction effects or non-linear effects.

2) For individual viewing models it is expected that decision trees will provide more accuracy than neural networks or regression models. The independent variables in this case are all nominal, allowing the use of relatively simple trees to de-
scribe interaction effects. Again regression models must be specifically designed to incorporate these interactions, while the power of neural network models in describing non-linear relationships is not utilised in these models.

3) In all our models, carry-over effects from one time period to the next are expected to be most important predictors. Genre is expected to be an important predictor only for the individual viewing models, with genre effects obscured when ratings are considered.

4) It is expected that the population rating forecasts will be relatively inaccurate because no demographic information is used while forecasts based on individual viewing will also be relatively inaccurate because of idiosyncratic individual behaviour. The most accurate rating predictions are therefore expected from the aggregated segment forecasts. In addition the segment forecasts will allow advertisers to better match their advertising to their audiences.

This paper is expected to advance our understanding of rating prediction on several fronts. Firstly, it will compare the accuracy of three model estimation tools (regression, neural networks and decision trees) in the context of television rating models. Secondly it will determine the relative importance of the predictor variables (genre, time of day, network, carry-over, demographic/viewing segment) in rating predictions. Finally it will compare the accuracy of three aggregation strategies for rating forecasts.

2. Methodology

This study predicts TV ratings from individual viewing and demographic data, collected in New Zealand by Nielsen Media Research during July 2003. The data were recorded by people meters located in 470 households containing approximately 1226 people. We consider viewing data for 15 minute time periods, only the 3 most popular networks (TV1, TV2 and TV3) and 14 program genres. In all there were approximately 3.2 million records of individual data that were compressed into ratings for 2976 fifteen minute time periods for each of the three networks. Ratings were calculated for each of these time periods taking into account the viewing time for each person in the panel (as a fraction of 15 minutes) and the weight assigned to each person by AC Nielsen Media Research. If the panel were to reflect the demographics of the New Zealand population exactly all these weights would be one. In practise this is not possible so people in the panel must be weighted more highly if they have demographic characteristics that are under represented in the panel, while people whose demographic characteristics are over represented in the panel must be weighted lower. These weights must of course sum to one for each time period.

The formula used to calculate these ratings is given below in Eq. (1). If we define \( V_{kjt} \) as the proportion of viewing time for individual \( k \) on network \( j \) in time period \( t \) and \( W_{kt} \) as the weighting for individual \( k \) in time period \( t \), then we can calculate \( R_{jt} \), the rating for network \( j \) in time period \( t \), by summing over all \( n \) individuals who view network \( j \) in time period \( t \).

\[
R_{jt} = \sum_{k=1}^{n} W_{kt} \cdot V_{kjt}
\]  

(1)

Population rating models were derived using ratings for the 2976 fifteen minute time periods for each of the three networks. In view of the size of our data sets, standard data mining approaches, as described in Berry and Linoff [5], were applied in the fitting and testing of models. Models were fitted, validated and tested using independent samples of appropriate size and composition. Forty percent of the data was used for training, 30% was used as a validation data set in order to prevent overfitting and error variances were estimated using a 30% holdout (test) sample. The models for individual viewing were derived using a data set describing individual behaviour in each of the 2976 time periods. However, time periods when individuals were not viewing networks TV1, TV2 or TV3 were under-sampled in order to provide a more balanced data set in terms of network choice. In these “non-viewing” time periods there was either no viewing or the chosen network was not TV1, TV2 or TV3. Only 6% of the “non-viewing” time periods were selected using a systematic sample, after sorting the data by day of the month and time of day. The reduced data set, consisting of 562 thousand observations, contained 31% TV1, 22% TV2, 18% TV3 and 29% “Nonviewer” records.

2.1. Variable transformation

In order to avoid heteroscedasticity in the residuals the ratings were transformed using the arcsine square root transformation. This transformation is recommended for responses that are proportions as explained previously [26]. In addition the genre and time of day
variables were collapsed into fewer categories by the SAS Enterprise Miner variable selection node. This node collapses the categories in such a way as to maximise the correlation with the dependent variables. This meant that the number of genre and time of day categories varied in response to changes in the dependent variable. In the case of the 14 genre categories, five collapsed genre categories were normally used while the 96 fifteen minute time of day blocks were usually replaced by only six time of day categories. This reduction in the number of categories for the genre and time of day variables served to simplify all our models, thereby improving predictive accuracy.

2.2. Model estimation

All analyses were performed using SAS or SAS Enterprise Miner. A simple least squares approach was used to fit all the models except the individual network choice models. We allowed for network switching within a time period by fitting individual network choice models and the proportion of viewing time models simultaneously, using a two-stage procedure [2], with the network choice probabilities, in particular the probability of the “nonviewing” network, incorporated in the proportion of viewing time model. This approach permitted correlation between probabilities for network choice and proportion of viewing time, with higher times expected for the “nonviewing” network.

2.3. Types of model

Neural networks, decision trees and regression models have been used to estimate our models. Multilayer perceptrons, described by Bishop [6], provided the architecture for the neural networks with hyperbolic tangent functions producing values for a single layer of hidden nodes and linear functions producing output values. F-tests were used to determine the optimum splitting points when producing decision tree models [7]. CHAID decision trees, first published by Hartigan [15], with chi-squared tests to obtain the optimum splitting points, were used for predicting network choice. In all trees only binary splits were permitted. A backward selection of variables was used in the estimation of the regression models and there was no attempt to incorporate interaction terms in these models.

2.4. Population rating model approach

The rating for network j in time period t as defined in Eq. (1) can be described in terms of a set of dummy variables for network, genre, day of the week and time of day as well as carryover effects, measured using lagged terms for the previous period’s rating, and random error (e). The \( f(.) \) function represents an arcsine square root transformation, used to ensure that the random errors have a more constant variance and a more normal distribution, while the Greek letters \((\alpha, \delta, \gamma, \beta, \eta, \tau)\) represent parameters that must be estimated. The dummy variables are defined with \( N_{jt} \) set equal to one for network \( j \) in time period \( t \), \( G_{jxt} \) set equal to one for genre \( x \) showing on network \( j \) in time period \( t \), \( D_{rt} \) set equal to one for \( t \) on day \( r \) and \( B_{qt} \) set equal to one for time of day category \( q \) in time period \( t \).

\[
f(R_{jt}) = \alpha + \delta R_{jt-1} + \gamma_j N_{jt} + \sum_{x} \beta_{jt} G_{jxt} + \eta_j D_{rt} + \sum_{q} \tau_q B_{qt} + e_{jt} \tag{2}
\]

However, the predictions from this model need to be constrained to allow for the choice nature of television viewing and the competition between the networks. As observed by Webster and Lichty [30], Patelis et al. [19] and others, a better way to estimate ratings is to first obtain predicted television ratings for period \( t \) \((\hat{R}_t)\) and then multiply by the estimated network shares for each network \( j \) for time period \( t \) \((\hat{S}_{jt})\) to obtain a constrained predicted network rating for each network \( j \) for time period \( t \) \((\hat{R}_{jt})\). Model (3) can be used to estimate the overall television rating for all networks with the constrained network ratings calculated from Eq. (4).

\[
f(R_t) = \lambda + \phi R_{t-1} + \sum_{r} \varphi_r D_{rt} + \sum_{q} \theta_q B_{qt} + e_t \tag{3}
\]

\[
\hat{R}_{jt} = \hat{R}_t \frac{\hat{R}_{jt}}{\sum_{i=1}^{N} \hat{R}_{it}} = \hat{R}_t \hat{S}_{jt} \tag{4}
\]

If models (2) and (3) are fitted using regression trees or neural networks, non-linear carry-over effects and factor interactions can be automatically included.
2.5. Segmentation

It was expected that more accurate rating forecasts could be obtained by splitting the people in the AC Nielsen panel into segments with similar demographics and viewing behaviour, and then developing separate rating models for each segment. The dendrogram from an agglomerative hierarchical segmentation was constructed using Ward’s linkage method [14, p. 496]. Four segments loosely defined in terms of age and viewing behaviour were suggested and the characteristics of these segments are described in Table 1. The first segment was the largest, containing mostly middle-aged people, while the second segment contained mostly young people under the age of twenty. The third segment had a large percentage of retired people while the fourth segment contained a high percentage of people who watch other networks, in particular the SKY networks, which unlike TV1, TV2 and TV3 have a monthly rental.

2.6. Segment rating model approach

The above models (2) and (3) can also be developed for each of the four segments \((i = 1, 2, 3, 4)\), and the rating forecasts \((\hat{R}_{j_it})\) from these four models can then be aggregated to provide an alternative population rating forecast for each network \(j\) in each time period \(t\). Following the approach of Rust et al. [23], if the weight for the \(i\)th segment is \(\omega_i\), with \(\sum_{i=1}^{4} \omega_i = 1\), we obtain estimated ratings for network \(j\) by summing over the four segments as shown below.

\[
\hat{R}_{jt} = \sum_{i=1}^{4} \omega_i \hat{R}_{j_it} \quad (5)
\]

2.7. Individual viewing model approach

The forecasts from the individual viewing models can also be aggregated to provide a further population rating forecast for each network in each time period. We use Discrete Choice Models to describe the network choices of an individual, including “No Viewing” as one of the networks. Possible input variables include day of the week \((D)\), time of day \((B)\), demographic characteristics as measured by the segmentation category \((C)\), program genre \((G)\) and carry-over viewing behaviour from the network viewed in the previous time period. All these variables are nominal and must therefore be recoded using sets of dummy variables. The target variable was the individual network choice. This model can be fitted using conventional logistic regression analyses or it can be fitted using classification trees or neural network analyses if interaction effects are to be included.

Assuming that network 4 is the “No Viewing” network choice, we define \(P_{kjt}\) as the probability that individual \(k\) will watch network \(j\) during time period \(t\). Then

\[
\ln\left(\frac{P_{kjt}}{P_{k4t}}\right) = \nu + \kappa_j N_{j,t-1} + \sum_{i} \xi_i C_i + \sum_{x} \psi_x G_{jxt} + \sum_{r} \nu_r D_{rt} + \sum_{q} \zeta_q B_{q,t} + r_{kjt} \quad (6)
\]

In this model the Greek letters \((\nu, \kappa, \xi, \psi, \nu, \kappa, \zeta)\) are parameters which must be estimated and \(r_{kjt}\) is the error term. The variable \(C_i\) is set equal to one if person \(k\) belongs to segment \(i\), zero otherwise, and the other

---

**Table 1**

<table>
<thead>
<tr>
<th>Segment characteristics</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>% Viewers</td>
<td>38.0</td>
<td>14.5</td>
<td>28.1</td>
<td>19.4</td>
<td>100</td>
</tr>
<tr>
<td>Favourite network</td>
<td>TV2</td>
<td>TV2</td>
<td>TV1</td>
<td>TV1</td>
<td>TV1</td>
</tr>
<tr>
<td>Average Daily Viewing(hrs)</td>
<td>2.3</td>
<td>1.8</td>
<td>3.0</td>
<td>2.2</td>
<td>2.4</td>
</tr>
<tr>
<td>% under 20 years</td>
<td>21.5</td>
<td>91.4</td>
<td>3.8</td>
<td>19.6</td>
<td>26.3</td>
</tr>
<tr>
<td>% 20-60 years</td>
<td>71.5</td>
<td>8.6</td>
<td>53.9</td>
<td>62.6</td>
<td>55.7</td>
</tr>
<tr>
<td>% 60+ years</td>
<td>7.0</td>
<td>0.0</td>
<td>42.3</td>
<td>17.8</td>
<td>18.0</td>
</tr>
<tr>
<td>% SKY viewing</td>
<td>1.9</td>
<td>2.2</td>
<td>1.9</td>
<td>34.1</td>
<td>8.2</td>
</tr>
<tr>
<td>% viewing 5pm–8pm</td>
<td>30.8</td>
<td>35.5</td>
<td>45.0</td>
<td>33.2</td>
<td>35.9</td>
</tr>
<tr>
<td>% viewing 8pm–11pm</td>
<td>40.3</td>
<td>18.3</td>
<td>35.3</td>
<td>34.7</td>
<td>34.6</td>
</tr>
<tr>
<td>% viewing 8am–5pm</td>
<td>18.1</td>
<td>39.6</td>
<td>15.9</td>
<td>21.7</td>
<td>21.3</td>
</tr>
<tr>
<td>% Male</td>
<td>49.8</td>
<td>57.7</td>
<td>33.1</td>
<td>57.5</td>
<td>47.7</td>
</tr>
<tr>
<td>% University Graduate</td>
<td>14.3</td>
<td>1.8</td>
<td>14.8</td>
<td>12.8</td>
<td>11.3</td>
</tr>
<tr>
<td>% European Descent</td>
<td>60.3</td>
<td>50.3</td>
<td>81.4</td>
<td>62.1</td>
<td>65.1</td>
</tr>
<tr>
<td>% without income</td>
<td>7.0</td>
<td>35.6</td>
<td>2.8</td>
<td>2.7</td>
<td>9.1</td>
</tr>
<tr>
<td>% income &gt; $50000</td>
<td>12.4</td>
<td>0.0</td>
<td>13.3</td>
<td>23.3</td>
<td>13.0</td>
</tr>
</tbody>
</table>

Segment Name
- Middle-Aged Kids
- Older Pay-TV Patrons
- Complete Panel
variables \((N, G, D\) and \(B)\) are dummy variables defined as in Eq. (2). In fitting this model we are assuming that all viewers can choose among all available networks but can choose only one of these networks to view in any time period. This assumption was obviously false for our 15 minute periods because people often switch networks inside a 15 minute period. We therefore also needed to estimate the proportion of time spent viewing any network and we needed to distribute the weights for each person between the networks in accordance with this proportion when predicting ratings.

Viewing time for network \(j\) by person \(k\) in time period \(t\) is expressed as a proportion of the possible 15 minutes in any time period. Models for this variable \((T_{kj,t})\) have a similar form to Eq. (6) above, but the estimated probabilities for network choice \((\hat{P}_{kj,t})\) were also included as predictor variables, necessitating the use of a Two-Stage procedure for fitting the \(P_{kj,t}\) and \(T_{kj,t}\) models simultaneously. Estimated viewing time proportions were then adjusted proportionately in order to ensure that they added to one for every 15 minute time period. Forecasts for \(P_{kj,t}\) and \(T_{kj,t}\) \((\hat{P}_{kj,t}\) and \(\hat{T}_{kj,t}\) respectively) were easily converted into estimated ratings \((\hat{R}_{jt})\) for network \(j\) at time interval \(t\) using the following formula:

\[
\hat{R}_{jt} = \sum_{k=1}^{n_{jt}} \hat{P}_{kj,t} \hat{T}_{kj,t} W_{kt}
\]

where \(W_{kt}\) is the individual \(k\) weighting in time period \(t\) and \(n_{jt}\) is the number of network \(j\) viewers at this time.

2.7. Forecast aggregation

Three levels of forecast aggregation were considered for the production of network ratings, with higher levels of data aggregation corresponding to lower levels of forecast aggregation (if any). Firstly, in the case of the models for population ratings in Eq. (3) above, there was no need to aggregate forecasts because the data used for the model had already been fully aggregated in the form of population network ratings.

Secondly, when separate network rating models were fitted for each of the segments, the segment forecasts had to be aggregated across the four segments in order to produce forecasts for network ratings. In this case there was initial aggregation of the data in order to produce segment ratings and then further aggregation of the forecasts across the segments in order to produce population network ratings. As indicated in Eq. (5) above each segment forecast was weighted according to the size of the segment before this final aggregation took place.

Thirdly, models for individual viewing behaviour were fitted for all the individuals in the panel, producing individual forecasts for network choice and viewing time within each time period. In this case there was no data aggregation so maximum forecast aggregation was required in Eq. (7) above in order to produce population network ratings while taking into account the AC Nielsen weights assigned to each person.

2.9. Relative importance of predictors and predictive accuracy

For each of the above models the relative importance of the predictor variables was assessed using an R-square contribution in the case of the rating and proportion of time models and using a normed chi-square value contribution \((\chi^2/df)\) in the case of individual network choice. Variables which contribute more of the explanatory power in a model are considered to be more important.

Finally, in this study predictive accuracy is measured using expected bias, error standard deviations, mean absolute error and mean absolute relative error. In particular, models which produce lower mean square errors are preferred to models which produce higher mean square errors. Wherever possible errors are calculated from hold-out (test) samples rather than just the data used to fit the models.

3. Results

In this section we fit the various models using neural networks, decision trees and regression, choosing the best fitting models in terms of the test data to provide our rating forecasts for each network in each time period. We then compare the above three aggregation approaches using the complete data for July 2003 to compute the average squared prediction errors.

As expected it was found that the residuals for the rating models were well behaved, exhibiting homoscedasticity as a result of the arcsine square root transformation. The results in Table 2 show the standard errors for the test data when the models described in Eqs (2) and (6) were fitted to the data. Note that the standard errors for the rating models were defined in terms of the transformed ratings while the standard errors for the viewing time model were defined as a proportion of
D. Meyer and R.J. Hyndman / The accuracy of television network rating forecasts

Table 2
Standard Errors (SE) and Misclassification Rates (MR) for Linear and Non-Linear Models for Network Ratings

<table>
<thead>
<tr>
<th>Approach</th>
<th>Regression</th>
<th>Decision Tree</th>
<th>Neural Net.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population Ratings (SE)</td>
<td>0.022</td>
<td>0.024</td>
<td>0.021</td>
</tr>
<tr>
<td>Middle-Aged Segment Ratings (SE)</td>
<td>0.025</td>
<td>0.026</td>
<td>0.023</td>
</tr>
<tr>
<td>Kids Segment Ratings (SE)</td>
<td>0.033</td>
<td>0.035</td>
<td>0.032</td>
</tr>
<tr>
<td>Older Segment Ratings (SE)</td>
<td>0.038</td>
<td>0.039</td>
<td>0.036</td>
</tr>
<tr>
<td>Pay-TV Segment Ratings (SE)</td>
<td>0.032</td>
<td>0.032</td>
<td>0.031</td>
</tr>
<tr>
<td>Individual Network Choice (MR)</td>
<td>0.24</td>
<td>0.23</td>
<td>0.30</td>
</tr>
<tr>
<td>Individual Viewing Time (SE)</td>
<td>0.28</td>
<td>0.28</td>
<td>0.28</td>
</tr>
</tbody>
</table>

Table 3
Effect of Forecast Aggregation on Prediction Accuracy for Network Ratings

<table>
<thead>
<tr>
<th>Aggregation Level for forecasts</th>
<th>Expected Error</th>
<th>MSE</th>
<th>MAE</th>
<th>MARE</th>
</tr>
</thead>
<tbody>
<tr>
<td>No aggregation Population Ratings</td>
<td>1.299</td>
<td>1.698</td>
<td>0.648</td>
<td>0.418</td>
</tr>
<tr>
<td>Aggregation of Segment Ratings</td>
<td>1.022</td>
<td>1.114</td>
<td>0.536</td>
<td>0.275</td>
</tr>
<tr>
<td>Aggregation of Individual Viewing</td>
<td>1.216</td>
<td>1.508</td>
<td>0.726</td>
<td>0.939</td>
</tr>
</tbody>
</table>

a 15 minute time period. The neural network models were fitted using 2, 3 and 4 hidden nodes with the best results, usually for 3 hidden nodes, reported in Table 2.

The results were remarkably similar for regression, decision trees and neural networks suggesting that it does not much matter which type of model is used. However, as expected, the population and segment rating models were a little more accurate when fitted using neural networks, suggesting that there are non-linear carry-over effects and interactions. Also, as expected, the models for network choice were a little more accurate when decision trees were used for handling the interaction effects for the nominal predictors. However, there were no obvious differences in the case of the individual viewing time models.

Interestingly the results for the total population and for the “Middle-Aged” segment were better than the results for the “Kids”, “Older” and “Pay-TV” segments. There was a moderate positive correlation between the forecast errors for the “Middle-Aged” and “Kids” segments ($r = 0.57$), suggesting underlying environmental effects not captured in the explanatory variables in these models. However, the forecast errors for the “Older” segment showed a strong negative correlation with the forecast errors for the “Middle-Aged” and “Kids” segments ($r = -0.76$ and $r = -0.73$), suggesting improved accuracy for a stratified forecast based on all four segments, with positive errors for the “Middle-Aged” and “Kids” segments cancelled by negative errors for the “Older” segment.

Table 3 compares the results for the best of the above models when applied to the whole data set. In this case errors were computed from the untransformed rating data allowing a comparison of the three different aggregation approaches. As expected from the error correlations for the segment forecasts, the aggregation of segment ratings produced a good result. Despite the larger bias in the aggregated segment forecast, the error standard deviation is small producing the lowest mean square error. This was confirmed by the average relative absolute errors, computed by ignoring all zero ratings, and the mean absolute errors, suggesting that this approach will produce the best results on average. A comparison of the population rating forecasts and the aggregated individual viewing forecasts showed mixed results in terms of the various error measures, with a higher MSE but lower MAE and MARE for the population rating forecasts.

Next we consider the relative importance of the predictor variables in each of the models. Strangely, day of the week was not a significant variable in any of the models and was therefore dropped. As shown in Table 4, the carry-over effect was the most important predictor variable for all the models except viewing time. For the population and segment rating models time of day was more important than genre. However, genre and segment were more important than time of day in the case of individual network choice models. Clearly the importance of genre was obscured when more aggregated data models were used. The model for individual viewing time within each 15 minute time period was somewhat disappointing, explaining only a small proportion of the variation in this variable.
Table 4
Predictor Importance Based on Percentage Variation Explained (Normed Chi-Square for Network Choice, R-Square otherwise)

<table>
<thead>
<tr>
<th>Predictors</th>
<th>Population ratings</th>
<th>Segment Ratings for segment categories</th>
<th>Network choice ($P_{ijk}$)</th>
<th>Time ($T_{ijk}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Carry-over Rating</td>
<td>0.95</td>
<td>0.94</td>
<td>0.91</td>
<td>0.94</td>
</tr>
<tr>
<td>Carry-over Network</td>
<td>n.a.</td>
<td>n.a.</td>
<td>n.a.</td>
<td>n.a.</td>
</tr>
<tr>
<td>Time of day (B)</td>
<td>0.78</td>
<td>0.75</td>
<td>0.55</td>
<td>0.56</td>
</tr>
<tr>
<td>Genre (G)</td>
<td>0.20</td>
<td>0.24</td>
<td>0.23</td>
<td>0.20</td>
</tr>
<tr>
<td>Network Choice (N)</td>
<td>0.04</td>
<td>0.03</td>
<td>0.10</td>
<td>0.19</td>
</tr>
<tr>
<td>Segment Category (C)</td>
<td>n.a.</td>
<td>n.a.</td>
<td>n.a.</td>
<td>n.a.</td>
</tr>
</tbody>
</table>

4. Conclusions

Three aggregation approaches have been used to forecast television program ratings, with trees and neural network fitting procedures used in order to allow for interactions between the input variables and for nonlinear relationships. A carry-over effect from the previous time period was the most important predictor, with time of day relatively important in the rating models and genre and segment more important in the network choice model. For all levels of aggregation linear regression models tended to perform worse than neural networks or trees when interaction effects and/or nonlinear carryover were ignored. Neural networks performed best in the case of rating models while decision trees performed best in the case of the individual viewing models.

The results suggested that the best forecast accuracy is obtained from segment rating models, using a weighted average of the segment forecasts. This suggests that aggregation bias will be small provided that sufficient demographic data is incorporated via the segmentation. It is recommended that this approach be used to produce the rating forecasts required by program schedulers and advertisers, with neural networks being used to fit these models. In addition it is recommended that an arcsine square root transformation be applied to the ratings in order to ensure that model errors show homoscedastic behaviour.

However, the above models are based on data only for July 2003 and it is well known in the television industry that television viewing is affected by economic cycles and the seasons. In order for a television rating forecast to be useful to program schedulers the model would need to be fitted to recent appropriate data (say last month’s data) in order to obtain reasonably accurate forecasts and prediction intervals. Alternatively the model needs to include a time and a seasonal dimension as suggested by Gensch and Shaman [11], Patelis et al. [19]. A third approach is to build separate models for every successive month and to produce rating estimates and standard errors from each of these models. Reliable forecasts for program ratings and their standard errors can then be obtained from these time series using methods such as exponential smoothing or time series decomposition.

Finally, as recommended by Patelis et al. [19], rating forecasts need to be supported by a Decision Support System which incorporates qualitative factors for the forecasting of television viewership. Such a system should allow easy access to information and “what if” queries as well as the entry of exceptional influence impacts on television viewing.

Acknowledgements

Our sincere thanks to AGB Nielsen Media Research for providing us with the data that has made this study possible. Also, our thanks to the referees for their helpful comments.

References


