Available online at www.sciencedirect.com

SciVerse ScienceDirect

International Journal of Forecasting 27 (2011) 1215–1240

international journal of forecasting

www.elsevier.com/locate/ijforecast

Forecasting television ratings

Peter J. Danaher[a,*], Tracey S. Dagger[b,1], Michael S. Smith[a]

[a] Melbourne Business School, 200 Leicester Street, Carlton, Victoria 3053, Australia

[b] Department of Marketing, Faculty of Business and Economics, Monash University, Caulfield, Melbourne, Victoria 3145, Australia

Abstract

Despite the state of flux in media today, television remains the dominant player globally for advertising spending. Since television advertising time is purchased on the basis of projected future ratings, and ad costs have skyrocketed, there is increasingly pressure to forecast television ratings accurately. The forecasting methods that have been used in the past are not generally very reliable, and many have not been validated; also, even more distressingly, none have been tested in today's multichannel environment. In this study we compare eight different forecasting models, ranging from a naïve empirical method to a state-of-the-art Bayesian model-averaging method. Our data come from a recent time period, namely 2004–2008, in a market with over 70 channels, making the data more typical of today's viewing environment. The simple models that are commonly used in industry do not forecast as well as any econometric models. Furthermore, time series methods are not applicable, as many programs are broadcast only once. However, we find that a relatively straightforward random effects regression model often performs as well as more sophisticated Bayesian models in out-of-sample forecasting. Finally, we demonstrate that making improvements in ratings forecasts could save the television industry between \$250 and \$586 million per year.

Keywords: Television audience; Regression; Random effects; Bayesian model averaging

1. Introduction

Even in the current difficult economic times, the global spend on television advertising in 2009 was estimated to be US \$173 billion (Yao, 2009). Although the total advertising spend has declined over the past year, television's share of this revenue is expected to increase from 38.6% in 2009 to a record 39.3% in 2010 (Callan, 2009). A recent study by ZenithOptimedia (2009) reports that when advertisers cut budgets across the board they often cut television last, since they remain convinced of its effectiveness. Due to the high demand for television advertising, an average 30-s national television commercial in the US now costs \$350,000 (Gaebler, 2009), while a 30-s spot during the 2010 Super Bowl cost nearly \$3 million (The Ad Class, 2010).

* Corresponding author. Tel.: +61 3 9349 8255; fax: +61 3 9349 8144.

E-mail addresses: p.danaher@mbs.edu (P.J. Danaher), tracey.dagger@buseco.monash.edu.au (T.S. Dagger), mike.smith@mbs.edu (M.S. Smith).

[1] Tel.: +61 3 9903 1927.

0169-2070/\$ - see front matter
doi:10.1016/j.ijforecast.2010.08.002

In an environment where marketing budgets are increasingly facing downward pressure, the extremely high cost of television advertising places greater scrutiny on the most expensive part of advertising, namely media planning and buying. Purchasing television commercial time requires a media planner to forecast television ratings for up to six months into the future, depending on the lead time when advertising slots are put up for sale (Horen, 1980; Katz, 2003; Napoli, 2001). The cost of advertising is directly linked to predicted program ratings, based on which, advertisers plan to achieve a target number of gross ratings points (GRPs) over the duration of the campaign (Givon & Grosfeld-Nir, 2008; Kelton & Schneider-Stone, 1998). An underachievement in actual GRPs delivered downstream places the broadcaster in the position of having to "make good" on the shortfall, thereby giving away advertising time and losing potential revenue (Consoli, 2008). An overachievement in actual GRPs means that the broadcaster could have sold the surplus to another advertiser, or that the advertiser could have spent less and still achieved their GRP objective (Beville, 1985; Friedman, 2007). Hence, accurate ratings forecasts are vital to both advertisers and broadcasters, with a swing of as little as one rating point resulting in a substantial gain or loss for either a broadcaster or an advertiser (Givon & Grosfeld-Nir, 2008; Kelton & Schneider-Stone, 1998).

The television environment is also changing rapidly, with the number of channels increasing each year, and the delivery method going from terrestrial, to cable, to satellite, and now to the Internet and cell phones. This all adds to the complexity of ratings forecasting, and yet, despite these challenges, a review of the literature reveals that this is a surprisingly neglected area of thorough academic and industry endeavor. For instance, all past ratings forecasting research has examined markets with between 3 and 6 channels, with only one study using data from the 21st century. That is, previous research is markedly out of kilter with the current television media environment. Other shortcomings of earlier works include an overreliance on basic regression models, a lack of comparisons of alternative forecasting models and even a tendency not to validate forecasting models, or do so for only a small number of programs. Even as long ago as the 1980s, Rust and Eechambadi (1989, p. 13) lamented that television rating forecasts are often "notoriously inaccurate". Furthermore, during the 1997–1998 television seasons, advertising industry forecasters predicted audience shares accurately for only 5 new programs and incorrectly for 22 new prime-time programs (Wells, 1997).

The purpose of this study is therefore to compare previous television ratings forecasting models with new methods, including state-of-the-art Bayesian model-averaging forecasts. We compare 8 forecast models, employing recent data from 2004 to 2008, in a market with over 70 channels. Some of the models also include random effects in order to capture features which are unique to each program. The Bayesian forecasting methods perform better than previous methods, but a simple regression model with program random effects also produces forecasts with accuracy similar to that of the Bayesian methods. Lastly, using the advertising cost for each program, we are able to calculate the expected financial gain from improved forecasts. For instance, using one of our best forecasting models instead of a commonly-used empirical method results in annual savings of up to $586 million for US television networks.

2. The forecasting challenge

To get a sense of the issues that need to be addressed by a television ratings forecasting model, we now examine some television ratings data. More details of the data will be provided later, but for the moment, we use Nielsen ratings, which are based on a panel of peoplemeter households, between 6:00 pm and 11:00 pm for the four and a half year period from January 2004 to June 2008.[2]

Fig. 1 displays the total ratings across all 70 channels in this market at 6, 8 and 10 pm. Twelve-period moving averages are overlaid on the time series graphs to smooth the daily figures and help exhibit the key trends. Two features are apparent. First, ratings are highly seasonal, with peaks in the winter months and troughs in summer. Second, the total viewing

[2] We use television program ratings throughout this study, as opposed to commercial break ratings. Nielsen has only been providing commercial ratings, known as C3 ratings, since the last quarter of 2007. The differences are often small, and the models we develop can use either program or commercial break ratings.

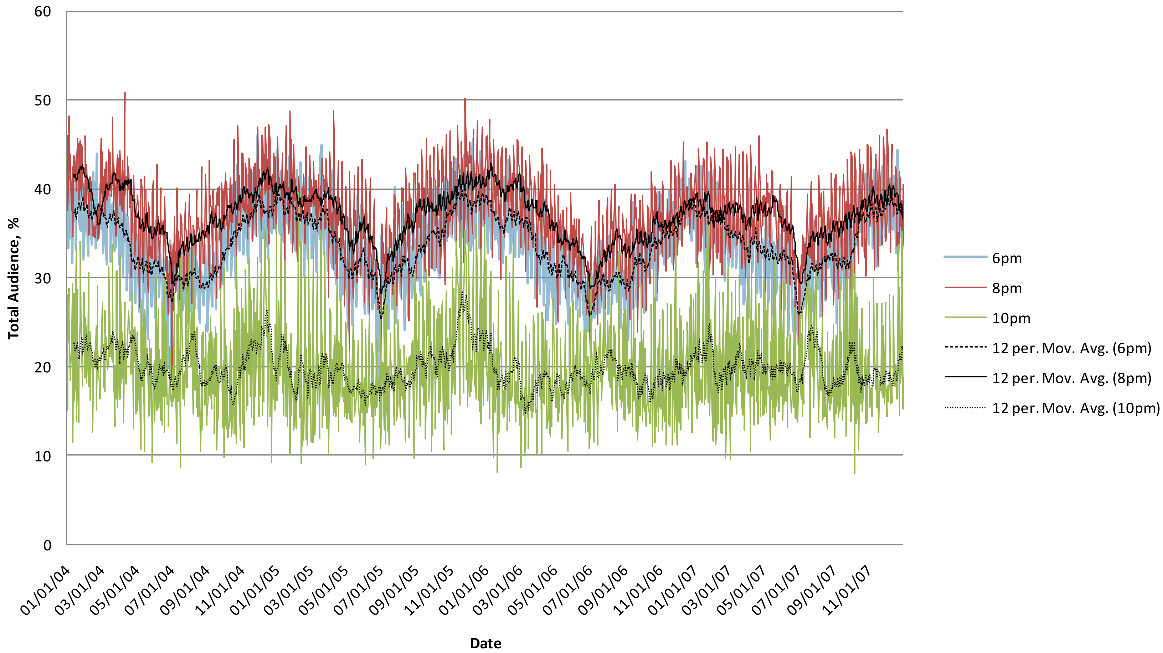

Fig. 1. Total audience from January 2004 to December 2007.

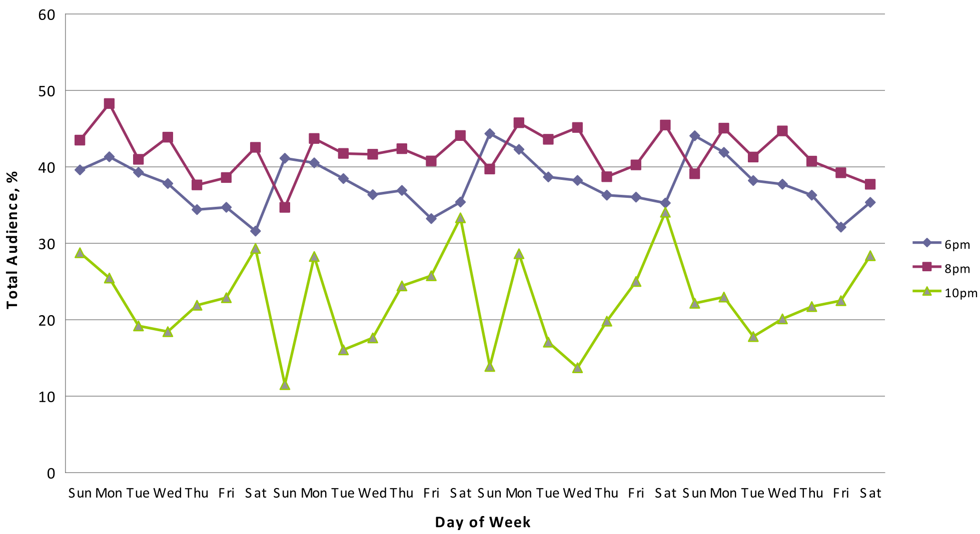

Fig. 2. Total audience by day of week for July 2004.

number varies considerably across prime time, with 8 pm having the highest total audience, and 10 pm the lowest. The seasonal variation is also much less noticeable at 10 pm than at 6 pm or 8 pm, as that time period is less affected by the number of hours of daylight, the temperature and daylight saving, which all reduce viewing in the early evening in summer.

Fig. 2 shows the total audience for different days of the week for the month of July, 2004. Again, the variation by the time of the evening is evident, but,

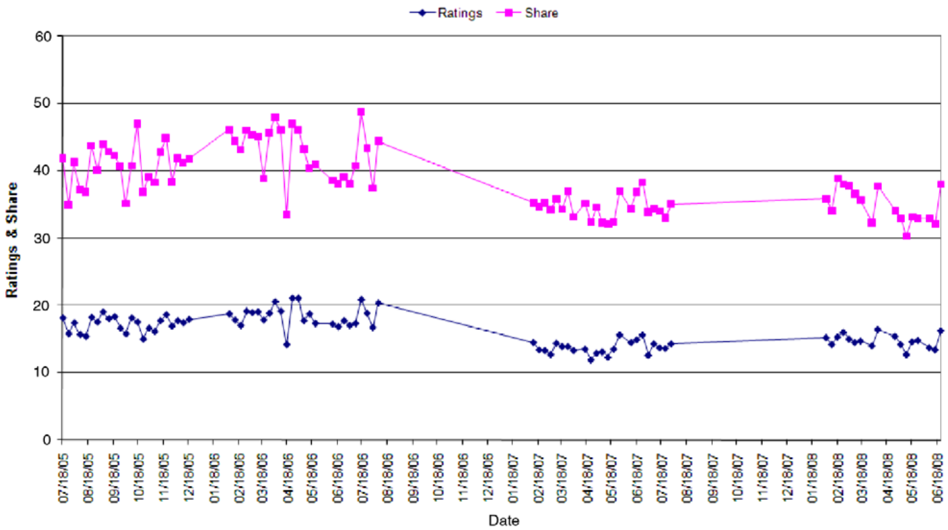

Fig. 3a. Ratings and share for *Desperate Housewives*.

more importantly, there is also variation across days. The largest total audiences at 6 pm and 8 pm are generally on Sundays and Mondays, while viewing at 10 pm is highest on a Saturday. The observation of a strong variation in ratings by day, time and month is consistent with the findings of previous studies (e.g., Gensch & Shaman, 1980), and demonstrates the need to include these factors in any reasonable forecasting model.

Modeling the total audience is only part of the process for predicting television ratings, with the other part being forecasting of channel shares, since program ratings are the product of the total audience and the channel share (Henry & Rinne, 1984). Hence, we now look at both ratings and channel share for two programs. *A priori*, we expect the channel share to be more variable than the total audience, since the share is calculated from among just those people who are *viewing*, whereas the total audience is the percentage of *all* panelists in the peoplemeter sample who are viewing anything.

The upper time series in Fig. 3a is the channel share and the lower series is the ratings for *Desperate Housewives* for July 2005 to July 2008. This show's share averaged about 40% in its first season and was reasonably steady from week to week. When *Desperate Housewives* returned in 2007, at the same time on the same day, its share decreased to about 35%, and remained at this level during 2008. There is also a reasonable degree of consistency in ratings for this show within each of the viewing seasons. Thus, forecasting either the share or ratings for this show should not be difficult.

Fig. 3b gives the share and ratings for *CSI:NY*. An interesting feature of this show is that it changed day from Wednesday to Tuesday in the 2008 season but retained its 8:30 pm start time. Previous work by Henry and Rinne (1984) and Napoli (2001) has demonstrated the difficulty of predicting the ratings of rescheduled shows. While the average share for *CSI:NY* decreased in 2008, its overall ratings stayed about the same, since the total audience was higher on a Tuesday during this time period. There was also a dramatic rise in both the share and ratings at the ends of the 2005 and 2006 seasons, as well as at the beginning of the 2006 season.

This sequence of graphs for the total audience, share and ratings gives an indication of the challenges faced when forecasting television ratings. Although the seasonality, day-of-the-week and time-of-day factors are consistent and are well correlated with the total audience, the channel share and ratings show evidence of much greater variation. This is probably related to additional factors, most of which are associated with the dynamics of the program schedule and the quality of the programs. For instance, many previous studies have shown the importance of lead-in, where viewing a program on the same channel

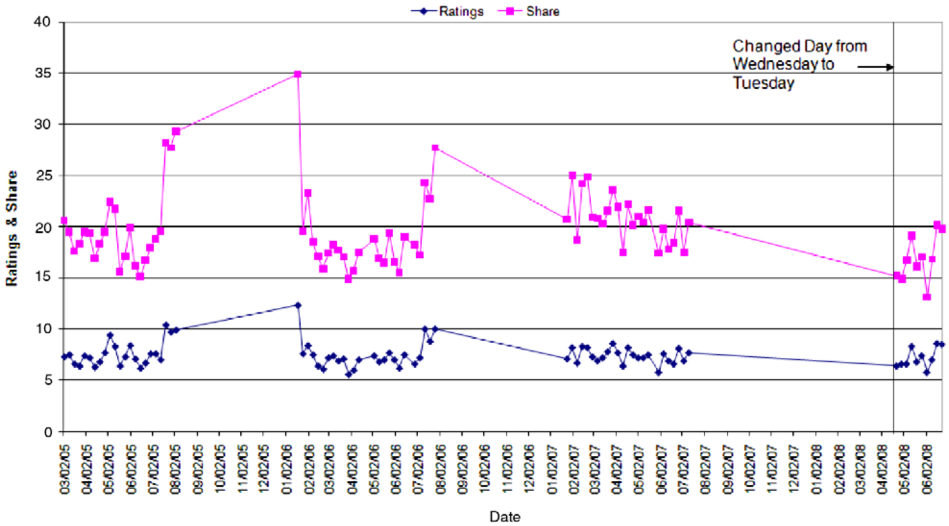

Fig. 3b. Ratings and share for *CSI:NY*.

prior to the current program enhances the viewing of the current program (Ehrenberg & Wakshlag, 1987; Goodhardt, Ehrenberg, & Collins, 1975; Henry & Rinne, 1984; Napoli, 2001). Another key determinant of viewing is the program type, often called the “genre”, where viewers might form preferences for program styles such as drama or comedy (Henry & Rinne, 1984; Rust & Alpert, 1984). Interactions between lead-in and genre have also been observed. For example, Henry and Rinne (1984) found that the ratings of the current show are enhanced if the previous program was of the same program genre. We now review the literature on television ratings forecasting, which will guide us as to the most accurate previous methods and the important factors to include in an econometric model.

3. Literature review

Table 1 lists all directly relevant previous studies that have forecast television ratings. Although this topic received a considerable amount of attention in the 1980s, there is a dearth of recent literature. This is surprising, given that the costs of television advertising have risen markedly over the past decade (Gaebler, 2009), which increases the demand for accurate forecasts. Part of the reason for the small number of previously published studies is that most of the forecasting models used in the media industry are proprietary (Patelis, Metaxiotis, Nikolopoulos, & Assimakopoulos, 2003).

As most previous studies were conducted some time ago, they use old data (with the oldest being 1966), and therefore include only a small number of channels, typically 3 to 4. This appears to be at odds with today’s television environment, where 50–100 channels would be normal. However, it must be remembered that even today most television markets are dominated by a handful of networks. This is even true in the US, where the four main networks still command a combined channel share of 40% (Nielsen Research, 2008) despite the prevalence of cable homes with numerous channels. The key issue in the modern multichannel television environment is not so much forecasting ratings for all channels, but the fact that a larger number of channels adds greatly to the viewing options, thus increasing the ratings variability and making forecasting more demanding. That is, whereas 15 years ago the competition was largely restricted to just a handful of networks, today’s viewers have many more choices, which probably increases the difficulty of predicting network ratings.

Regarding the prediction objective, many of the studies aim to predict program ratings, but a reasonable number predict quarter-hour or half-hour ratings, and some predict both. The dependent variable in the prediction model can be either program ratings or channel share. Webster and Wakshlag (1983) make

Table 1
Comparison of the previous television ratings forecasting literature.

Study	Chans	Country	Date	Time	Predict. object	Model dep. var.	Data points Est	Data points Val	R^2 %	MAPE[a]	MAD[b]	Forecast length	Data method	Samp size	Genre	Forecast method	Important variables	Opt schd
Cattin et al. (1994)	4	France	10/92–7/93	All day	Qtr hrs	PUT × shr[c]	7056	480	n/a[d]	13.0	1.1	5 wks	*p*-meter	n/a	Y (n/a)	Historical	Prog type, day, time, week	N
Darmon (1976)	3	Canada	1/73 2 wks	6 pm–12 am	Progs	Ratings	180	180	72	n/a	3.4	2 wks	Diary	250	Y (11)	Regression	Channel, prog type	N
Danaher and Mawhinney (2001)	3	NZ	6/94	6–10 pm	Progs	Ratings	108	6	n/a	45.7	1.8	2 wks	Diary	164	N	Logit model	Lead-in, program dummies, program length	Y
Gensch and Shaman (1980)	3	US	11/66–8/69	7:30–10 pm	Qtr hrs & progs	PUT × shr	916	20	89	n/a	2.5	4 wks	Diary	n/a	N	Regression	Day, time, month	N
Henry and Rinne (1984)	3	US	10/81–3/82	8–11 pm	Progs	Shr only	525	102	81	18.2	2.35	9 wks	Diary & survey	1500	Y (12)	Logit model	Lead-in, lead-out, years on TV, changed time, prog type & preference	N
Horen (1980)	3	US	9/69–4/74	8–11 pm	Half hr ratings	PUT × shr	4130	0	70	n/a	n/a	n/a	Diary	n/a	Y (3)	Regression	Last year's ratings, competing progs, day, time, lead-in	Y
Kelton and Schneider-Stone (1998)	3	US	81–89 12 wk	7–10 pm	Progs	Ratings	53	55	70	n/a	3.6	1 wk	Diary	n/a	Y (8)	Regression	Lead-in, prog type, day, time, channel, prog attributes	Y
Napoli (2001)	4	US	1/93–12/98	8–11 pm	Progs	Shr only	140	0	19	21.4	15.2	N/A	Diary	n/a	N	Regression	Lead-in, lead-out, year	N
Patelis et al. (2003)	6	Greece	4/99–3/00	All day	Progs	PUT × shr	n/a	36 269	n/a	13.6	n/a	12 wks	n/a	n/a	Y (8)	Time series	Day, time, month, holidays, prog type	N
Reddy, Aronson, and Stam (1998)	1	US	1/90–3/90	8–11 pm	Progs	Ratings	338	0	93	n/a	n/a	n/a	n/a	n/a	Y (5)	Regression	Day, time, prog type, duration of show, attractiveness	Y
Rust and Alpert (1984)	3	US	9/77–11/77	5–11 pm	Half hr & progs	Ratings	34	34	93	n/a	2	2 days	Survey–Simmons	5434	Y (5)	Logit model	Prog type, audience flow	N
Tavakoli and Cave (1996)	4	UK	3/90 1 wk	5:30–10:30 pm	5 mins	Shr only	500	24	92	12.1	2.0	2 wks	*p*-meter & diary	n/a	Y (29)	Logit model	Lead-in, prog type, channel, audience appreciation	N

Table 1 (*continued*)

Study	Chans	Country	Date	Time	Predict. object	Model dep. var.	Data points Est	Data points Val	R^2 %	MAPE[a]	MAD[b]	Forecast length	Data method	Samp size	Genre	Forecast method	Important variables	Opt schd
Van Meurs (1994)	5	Holland	12/92–1/94	All day	n/a	PUT × shr	4200	500	n/a	31.0	0.9	1 month	*p*-meter	n/a	Y (n/a)	Historical	Prog type, channel, day, time, month	N
Weber (2002)	1	Germany	1/95–4/97	6–11 pm	Progs	Ratings	n/a	0	81	25.2	n/a	2 months	*p*-meter	12 000	N	Neural net	n/a	N
Yoo and Kim (2002)	4 US 3 Kr	US Korea	9/94–9/97	8–11 pm	Progs	Shr only	0	105 US 88 Kor	n/a	20.2 US 31.2 Kr	n/a	12wks	n/a	10 US 6 Kor	N	Judgment	Lead-in, day, prog type	N

[a] Mean absolute percentage error.

[b] Mean absolute deviation.

[c] PUT = People using Television, and shr = channel share.

[d] n/a = not available or not reported in the study.

a case for a theory of television program choice in two steps, the first being the decision to switch the set on (viewer availability), and the second being channel selection. This is consistent with a Total audience × Channel share approach to ratings prediction. However, a number of studies predict ratings directly, so we will investigate forecasts made by both the two-step and direct methods.

Observe that the majority of previous studies examine prime time only (7/8 pm–11 pm), which is in keeping with the large share of the ad spend which is directed at this time period. Also note the wide variation in the number of programs or quarter hours used for model estimation, ranging from a low of 34 up to 4200. Equally, the range of validation points — when validation is conducted — goes from only 6 up to 36,269. Only 4 studies exceed 100 validation points, while 4 other studies do not validate their models at all. This is an area which has generally been weak in previous television forecasting efforts.

Many previous forecasting models have used regression or a logit model for channel share, with no models using the state-of-the-art methods available today. Two industry studies (Cattin, Festa, & Le Diberder, 1994; Van Meurs, 1994) have used historical data to match future programs with previously broadcast programs on the basis of program genre and various attributes. Yoo and Kim (2002) report the accuracy of expert judge forecasts for ten ad agency executives in the US and six in Korea. In both countries the expert forecasts were among the worst in terms of validation performance. Weber's (2002) neural network method alone could be considered to be in the class of modern forecasting techniques, but it too performs relatively poorly. The appeal of traditional econometric methods is apparent and justified, since they often fit very well, typically having R^2 values exceeding 80%. Out-of-sample prediction errors are not always reported, but when they are, there is a high variance. The mean absolute prediction error (MAPE = average of |actual − predicted|/actual) ranges from a low of 12.1% up to 45.7%. Another commonly-reported forecast error measure is the mean absolute deviation (MAD = average of |actual − predicted|), which ranges from a very low 0.9 rating points to a very high 15.2 share points. Five of the studies use forecast models as the basis for the optimal scheduling of television programs, and therefore some do not report out-of-sample validation forecasts. If they do, they have higher errors than forecasting-only studies because they have to predict in the more challenging environment where a program changes time (e.g., Danaher & Mawhinney, 2001).

Some studies (e.g., Gensch & Shaman, 1980; Henry & Rinne, 1984; Yoo & Kim, 2002) disclose that they specifically exclude "nonregular" programs, such as movies or variety shows (e.g., Oscars), and avoid holidays such as 4th July and days when dramatic news events occur. Restricting the forecasting to just regular programs with a steady audience from week to week clearly makes forecasting less challenging, and therefore detracts from the versatility of the method. Broadcasters and ad agencies require "industrial strength" forecasting for every program and every day of the week.

The forecast horizons for all previous studies are very short, with the longest being 12 weeks, but with many being just a few days or weeks. Given that the forecasting horizon for television media planning can be as long as 6 months (Katz, 2003), it is rather surprising that previous models have produced forecasts for such short horizons.

The list of significant independent variables in the econometric models shows that day, time, month, program type and lead-in consistently appear as significant determinants of television ratings or share. The program type, or genre, is included in the majority of previous studies, with genre categories varying in number from 3 to 29, but with most having 8 or fewer.

In summary, Table 1 highlights the need for a comprehensive ratings forecasting study that uses several years of data rather than just a few weeks or months. A weakness of nearly all previous studies is the lack of model comparison. The exceptions are Gensch and Shaman (1980), though they compare only minor variations of historical-share forecasts, and Henry and Rinne (1984), who compare their method with Gensch and Shaman's model. In addition, Cattin et al. (1994) and Patelis et al. (2003) only compare their models with naïve models. A concern with many previous forecasting methods is either that out-of-sample validation has been omitted or that validation has been conducted using only a small number of programs. Finally, almost all previous studies have used a forecast horizon that is shorter than what is required in the television industry.

The limitations of earlier television ratings forecasting studies invite a much more thorough effort. Our study addresses the limitations of past work by using a very recent time period for estimation, namely 2004–2007, and using the first half of 2008 for validation. In the particular market under study there are 70 channels, including the major networks and subscription channels. We develop forecasting models for all of the networks and for one of the satellite channels which is partially supported by advertising. Our data include almost 48,000 program ratings for estimation and over 5000 program broadcasts for validation. Eight different forecasting models are compared, ranging from a simple naïve method to a state-of-the-art Bayesian model averaging method. We also compare models that combine the total audience and channel share with ones that predict the ratings directly.

4. Model preliminaries

4.1. Aggregate or individual forecasts

Traditionally, television ratings data have been collected at the individual level via diaries or peoplemeter panels, then aggregated to produce ratings (Katz, 2003). This begs the question as to whether the forecasts should be made at the individual or aggregate level. Three factors are strongly in favor of aggregate forecasts. First, forecasting at the individual level is inherently less accurate than at the aggregate level (Greene, 2003). Second, providers such as Nielsen report television audiences only at the aggregate level, while the forecast objective, namely ratings or share, is an aggregation of individual-level information. Third, broadcasters and advertising agencies generally only have access to television audience data at the aggregate level, making it impossible to develop models at the individual level. Meyer and Hyndman (2005) investigate another forecasting option, namely forecasting at a segment level, as was suggested by Rust, Kamakura, and Alpert (1992). They find that, due to the homogeneity of viewers within a segment, segment-level forecasts that are weighted up to the population perform better than individual-level forecasts. The methods which we employ in this study could also be used for this purpose, but we leave this to future research.

4.2. Available data

Before discussing the alternative models, we detail the data which were used to fit and validate the models, as this helps explain our modeling decisions. As was mentioned above, we have data from January 2004 to the end of June 2008. For proprietary reasons we cannot disclose the exact location of the market, but we can say that it is a Western market with four main free-to-air television networks, which we label Channels 1 to 4. All networks are totally supported by advertising. There is also a satellite television provider which has been operating in this market since 1991, with subscribers who pay a monthly fee for a suite of channels. Since its launch, the number of satellite channels has grown from 4 to 70, and their combined share exceeds 30%. This parallels the evolution of many other markets closely, where previously dominant networks have lost a substantial amount of ground to pay television.

Since the vast majority of the ad revenue is devoted to early evening and prime time, we restrict our estimation and forecasting to the period 6–11 pm, in keeping with most previous studies. We have program ratings for each of the networks and for six of the satellite channels. Moreover, we have the program name, rerun and genre information for the entire period 2004–2008, as well as the rate card cost for adverting in each program. During this time period, the ratings data came from a peoplemeter panel operated by Nielsen Media Research. The panel size is about 1150 people, aged five or more, spread representatively across the entire market. Although we have ratings for all of the networks and six of the satellite channels, we produce forecasts for just the four networks and one satellite sports channel, since these five channels dominate the market in terms of advertising revenue.[3] Restricting our forecasts to just the networks is reasonable, as they have the greatest need for good forecasts. This is also consistent with the most recent study in Table 1 (Patelis et al., 2003), which reports forecasts for only the 6 main channels in a market where 40 channels are available.

[3] Many of the satellite channels carry little or no advertising, e.g., movie channels, and therefore forecasts of their ratings are unnecessary.

4.3. Explanatory versus predictive models

Before embarking on model development, it is worth emphasizing that, in our forecasting context, the criterion for a good model is that it predicts well, as opposed to having a good explanatory power. The two goals are often not completely compatible: models with a good explanatory power seek to avoid misspecification and demand a low bias. By contrast, good predictive models aim to minimize the sums of the squared bias and variance (Greene, 2003). It is possible to have a biased predictive model that predicts better than a good explanatory model because the predictive model has a lower variance (Hastie, Tibshirani, & Friedman, 2001, pp. 197–200).

The usual measure of explanatory power in a linear regression is the R^2 statistic, where the fit criterion is calculated from the same data for which the model is estimated. However, a predictive model is generally assessed by the MAPE or MAD, calculated in an out-of-sample validation dataset. Wu, Harris, and McAuley (2007) give the conditions when an underspecified linear regression model can have a higher prediction accuracy than a better-specified model. These conditions are: when the dependent variable has a high variance; when the true value of an omitted independent variable is close to zero; when there is collinearity among the independent variables; and when the range of the omitted variables is small.

Another consideration that distinguishes explanatory from predictive models is that the independent variables (predictors) must be known for a future time period. If they are not, they too will have to be predicted, which usually diminishes the prediction accuracy (Danaher, 1994). This is very relevant to our situation, where a number of variables that have previously been shown in the literature to be significantly related to program ratings may not be known with certainty in the future, for example the genre of competing programs, the match in genre between two successive programs and the lead-in ratings. Such variables are typically significant in an explanatory regression model, but introduce further "noise" in a predictive setting. Indeed, Wu et al. (2007) show that removing these statistically significant variables can actually increase the predictive accuracy.

The trade-off between a good explanatory power and good predictive ability has received much attention in the statistics literature (see, e.g., Zheng & Agresti, 2000). It is acknowledged that an increase in model complexity often results in a better in-sample fit, and therefore a greater explanatory power, but can also result in a reduced predictive ability relative to simpler models (Hastie et al., 2001, p. 197). Thus, in our quest for a good television ratings forecasting model we will consider models of varying degrees of complexity, making comparisons on the basis of the predictive ability in a sizable holdout sample.

5. Forecasting models

There are a number of choices for forecasting models. Previous efforts have used relatively straightforward methods, ranging from simple naïve forecasts based on historical data to regression models estimated using OLS. We also introduce new television ratings forecasting models using random effects and the Bayesian model averaging methodology.

5.1. Notation

Let Y_t be the total ratings (a.k.a. "people using television", or PUTs) across all channels for month $m(t)$, day $d(t)$, year $y(t)$ and half-hour $h(t)$. That is, implicit within the t subscript is the month, day, year and broadcast time. R_t^c and S_t^c are, respectively, the rating and share for channel c at time t. The notation for the rating and share of a specific program p is $R_t^c(p)$ and $S_t^c(p)$, respectively. The covariate vector used to predict these ratings or shares is denoted by $X_t^c(p)$, and comprises the overall intercept, month, day, year and holidays, as well as program information for each channel, such as the duration and genre, and whether the program is a rerun.

5.2. Previous forecasting models

As a starting point, industry forecasts have been based on historical data and do not use a statistical model. This method is popular among practitioners, as is exemplified by Cattin et al. (1994) and Van Meurs (1994), and the typical forecast for a program broadcast today was whatever the rating was for the equivalent time, day and channel in the previous year.[4] This ignores program effects and any "shocks" such as special sporting events and most statutory holidays, which generally do not fall on the equivalent day in the previous year. However, this method does implicitly allow for seasonal, day of the week and time of day effects, which have repeatedly been shown to influence television ratings. Due to the widespread popularity of this "historical forecast" (Patelis et al., 2003), it serves as a robust benchmark for model comparison. Hence, our first forecast method, denoted by HIST, uses the historical rating from the equivalent, day, time and channel in the previous year, and is defined as

$$\hat{R}_t^{c,\mathrm{HIST}}(p) = R_{t_{y-1}}^c, \tag{1}$$

where t_{y-1} denotes the equivalent time point in the previous year (see footnote 3). This forecast does not attempt to match programs or program genres, but simply uses the rating for whatever program was broadcast in the equivalent time slot the previous year.

Our first econometric model for forecasting television ratings regresses program *ratings* on a covariate vector comprising program, seasonal, day of week, yearly-trend and statutory holiday variables. The covariates are described in detail below. Separate regression models are fit for each channel over the estimation period using the model $R_t^c(p) = \beta_R^c X_t^c(p) + \varepsilon_t^c(p)$, with Gaussian channel-specific disturbances $\varepsilon_t^c(p)$. Since ratings are predicted directly, we label this model as Ratings Direct (RD), with forecasts

$$\hat{R}_t^{c,\mathrm{RD}}(p) = \hat{\beta}_R^c X_t^c(p), \tag{2}$$

where $\hat{\beta}_R^c$ is the vector of estimated regression coefficients for each channel c.

An intuitively appealing method for forecasting ratings is to first model the total audience (PUTs), then multiply the forecast PUTs by a forecast of the share for each channel. Indeed, Gensch and Shaman (1980) suggest that this is generally superior to forecasting the ratings directly, as is done in Eq. (2). The appeal of this two-step method comes from the observation that the total ratings are highly dependent on the season, time of day, day of the week and holidays, which are easy to monitor both in the past and into the future. That is, the total ratings are less dependent on programs (Webster & Wakshlag, 1983). By contrast, the share of the audience watching each channel is expected to be related more closely to the programs on offer at the time. Thus, our second econometric forecasting model uses regression to initially forecast PUTs by regressing the total ratings onto the non-program specific covariates to obtain $\hat{Y}_t = \hat{\beta}_Y X_t$, where $\hat{\beta}_Y$ are the estimated regression coefficients.[5]

Subsequently, two methods are used to forecast the share for each channel. The first share forecast is a naïve one, which simply uses the historical share for the previous year for the equivalent time, day and channel.[6] It essentially ignores programming information, assuming that a particular channel retains its share regardless of the program on offer. Although this might seem like a brave assumption, there is some evidence to support it (e.g., Zubayr, 1999). For example, viewers might be loyal to a particular channel and watch it no matter what is broadcast (Goodhardt et al., 1975). This is more likely to occur when channels go to some effort to "brand" or stereotype themselves to a target audience.[7] Since this forecast is PUTs times the historical share, we label it as PUT × HS and define it to be

$$\hat{R}_t^{c,\mathrm{PUT}\times\mathrm{HS}}(p) = \hat{Y}_t S_{t_{y-1}}^c, \tag{3}$$

where $S_{t_{y-1}}^c$ is the observed channel share at the equivalent time period in the previous year.

[4] For example, the ratings forecast for Channel 1 at 6 pm on Sunday 1 January, 2008, can be obtained from the historical rating for Channel 1 at 6pm on Sunday 2 January, 2007. That is, the forecast rating is the historical rating lagged by 364 days. It is not lagged 365 days because it is important to match the day of the week, and since 7 × 52 = 364, not 365, the lag is 364 days. A further adjustment is needed for leap years.

[5] We tested for autocorrelated errors in this PUTs model using the Durbin-Watson test and found no significant autocorrelation.

[6] Of course, there are many possible variations in the way the naïve share prediction could be calculated. We have chosen the channel share at the equivalent time the previous year. Alternatively, the average share at the equivalent time over the past three years could be used, or the share averaged over several weeks rather than for just one day. We tried these variations in the calculation of the historical channel share and found that they did not improve the prediction accuracy.

[7] For example, Fox targets people aged 18–49 and airs shows which are appealing to younger people, such as *American Idol* and *Lost*. Hence, a young person who is wanting to watch television but is not aware of what is being broadcast at the time is likely to try Fox first, as their experience may have "trained" them that Fox is the channel that is most likely to be showing something that will appeal to them.

The second variant of this method forecasts the channel share using regression, with separate models for each channel, using the predictive model $\hat{S}_t^c(p) = \hat{\beta}_S^c X_t^c(p)$.[8] Here, $X_t^c(p)$ are just the intercept, program, day of week, season and year variables. The holiday variables are not included, as they are common across all channels and hence are used only in the model for total audience. Since this forecast is PUTs times the predicted share, we denote it as PUT × PS and define it as

$$\hat{R}_t^{c,\mathrm{PUT}\times\mathrm{PS}}(p) = \hat{Y}_t \hat{S}_t^c(p). \tag{4}$$

The model in Eq. (3) is very similar to that used by Gensch and Shaman (1980), while the share component of Eq. (4) is akin to Henry and Rinne's (1984) model.[9]

5.3. Random effects model

Over the past decade, random effect models have been being used increasingly, as they accommodate individual-level heterogeneity (Allenby & Rossi, 1998). We do not have individual-level data, but we do have a large number of programs, with repeated observations (i.e., weekly ratings) on most of them. Ansari, Essegaier, and Kohli (2000) developed a model for movie recommendations that considered the recommendations to be a combination of observed and unobserved factors. In our application we also have observed program characteristics, such as genre, but it is likely that there are also unobserved unique program effects. These unobserved effects can be captured by a random effects model. The random effects version of the Ratings Direct model (denoted RD_RE) simply enhances the point forecast in Eq. (2) by adding a program specific effect, denoted by $\hat{\delta}^c(p)$, to give

$$\hat{R}_t^{c,\mathrm{RD\_RE}}(p) = \hat{\beta}_R^c X_t^c(p) + \hat{\delta}^c(p). \tag{5}$$

Verbeke and Molenberghs (1997, p. 115) show how $\hat{R}_t^{c,\mathrm{RD\_RE}}(p)$ can be calculated by using empirical Bayes methods to compute $\hat{\delta}^c(p)$ for each program on channel c.

5.4. Bayesian model averaging with random effects

Over the past two decades, increasingly flexible statistical methods have been being developed rapidly. They aim to capture the relationship between a potentially large number of input variables (i.e., covariates) and a response variable, while guarding against overfitting the data, thereby striking a balance between flexibility and the degree of smoothness in the form of the relationship. While there are a number of such methods, Bayesian model averaging (George & McCulloch, 1993; Raftery, Madigan, & Hoeting, 1997; Smith & Kohn, 1996) has proven to be one of the most successful for forecasting purposes, with recent applications in economics (Wright, 2008) and meteorology (Raftery, Gneiting, Balabdaoui, & Polakowski, 2005), for example. Hence, we select it to represent the state-of-the-art in forecasting methods. Bayesian model averaging has not previously been used in the context of forecasting television ratings, although it has been employed in the marketing literature to study print advertising (Smith, Mathur, & Kohn, 2000) and to account for the heterogeneity that often arises in the analysis of individual-level marketing data (Gilbride, Allenby, & Brazell, 2006).

We now outline the Bayesian model averaging (BMA) method that we employ, with a more detailed exposition given in the Technical Appendix. There are a number of variants of BMA, depending on the priors that are adopted; Clyde and George (2004) provide a recent overview. We use an approach which was initially suggested for a linear regression model by Smith and Kohn (1996) and which is widely used (e.g., Fernandez, Ley, & Steel, 2001), but extend it to also account for additive random effects. BMA introduces binary indicators $\gamma = (\gamma_1, \ldots, \gamma_m)$ to represent the possibility that each of the coefficients of m fixed effects is exactly equal to zero, so that

$$\beta_j = 0 \quad \text{iff } \gamma_j = 0 \quad \text{and} \quad \beta_j \neq 0 \quad \text{iff } \gamma_j = 1 \tag{6}$$

[8] Note that we use the logit transformation for both the ratings and the share, in order to ensure that the predicted values range between 0 and 1. The reverse transformation (the logistic function) has a slight positive bias, averaging 0.1%. Correcting for this bias using a Monte Carlo method results in the overall prediction errors differing only in the third decimal place.

[9] Another candidate model is one which includes the historical forecast as a covariate. We tried this for Eq. (2) with the RD model and the forecast errors barely changed. Lastly, time series methods could be used, but such methods are difficult to apply here as the vast majority of television programs do not run continuously every day or week, and thus there are big gaps in the time series data at the program level. Moreover, many programs, such as movies, are broadcast only once, meaning that they have no time series information.

for all $j = 1, \ldots, m$ fixed effects. By introducing the indicator γ and setting the prior probability $p(\gamma_j = 0) > 0$, a nonzero probability is placed on the possibility that β_j is exactly equal to zero. This would not be possible using a regular continuous distribution as a prior for the fixed effect coefficients $\beta = (\beta_1, \ldots, \beta_m)$, because the prior probability of any specific value is always zero. The prior on the vector of m indicator variables is given by Cripps, Carter, and Kohn (2005), and places equal weight on subsets of covariates with different numbers of terms. Following Smith and Kohn (1996), a conjugate Gaussian proper prior is employed for the nonzero regression coefficients. This prior is often called a "g-prior", and is used extensively in Bayesian model averaging and variable selection (Clyde & George, 2004).

Using this Bayesian framework, the point estimate of the fixed effects coefficients is the Bayesian posterior mean

$$E(\beta \mid \text{data}) = \sum_{\gamma} E(\beta \mid \gamma, \text{data}) p(\gamma \mid \text{data}). \tag{7}$$

This is called a Bayesian model average because it averages over the posterior distribution of the possible configurations of the exogenous input variables. BMA is likely to substantially out-perform traditional least squares style estimators of β when there is a high level of uncertainty regarding the appropriateness of many input variables. In the case where there are no superfluous input variables, or only a few, the Bayesian model average in Eq. (7) is known to be only slightly less efficient than ordinary least squares.

When applied to channel ratings directly, forecasts $\hat{R}_t^{c,\text{BMA\_RD}}(p) = \hat{\beta}^c_{\text{BMA\_RD}} X_t^c(p)$ are obtained, where $\hat{\beta}^c_{\text{BMA\_RD}}$ are the BMA estimates of channel-specific coefficients β^c, computed using Bayesian Markov chain Monte Carlo (MCMC) methods, as detailed in the Appendix. When applied to channel share, we get forecasts $\hat{S}_t^{c,\text{BMA}}(p) = \hat{\beta}^c_{\text{BMA\_S}} X_t^c(p)$, which are then used to construct ratings forecasts $\hat{R}_t^{c,\text{BMA\_PUT}\times\text{PS}}(p) = \hat{Y}_t \hat{S}_t^{c,\text{BMA}}(p)$.

Our final variant of this model allows for additive Gaussian random effects in addition to the fixed effects that are subject to Bayesian model averaging. While this is a simple extension of the model, it requires the development of a more complex MCMC sampling scheme for estimation. The Appendix shows how this is undertaken, with a sampling scheme that explicitly generates both the random effects and the indicator variables γ. When applied to program ratings, the BMA forecasts with random effects (denoted BMA_RE) are

$$\hat{R}_t^{c,\text{BMA\_RE}}(p) = \hat{\beta}^c_{\text{BMA\_RE}} X_t^c(p) + \hat{\delta}^c_{\text{BMA}}(p), \tag{8}$$

where $\hat{\beta}^c_{\text{BMA\_RE}}$ are the BMA estimates computed using the Monte Carlo iterates from the sampling scheme with random effects included.

6. Covariates

The covariates used in some of the predictive models fall into two groups, the first being time-based, such as the day of the week, season, year and holidays, and the second being program-specific covariates like genre and duration. We now give details of the way in which the covariates are operationalized in our forecast models.

6.1. Time-based covariates

We follow Gensch and Shaman (1980) and use a linear combination of trigonometric functions to capture seasonal effects, as Fig. 1 clearly shows that PUTs have an annual cyclical pattern. As there are 12 months in a year, we use 6 sine and 6 cosine variables, $\sum_{j=1}^{6} \phi_j \cos(2\pi jk/365) + \varphi_j sin(2\pi jk/365)$, where k is the day of the year, ranging from 1 to 365 (or 366 for 2004). For instance, the first cosine term in this linear combination corresponds to an annual cycle, as observed in Fig. 1. Since the highest PUTs occur in December and January, we expect $\phi_1 > 0$. Gensch and Shaman (1980) note that the other five cosine and sine terms represent harmonics with periods of 6, 4, 3, 2.4 and 2 months. A similar construction could be used to model days of the week, but Fig. 2 shows that each day is likely to have a distinct effect, so we simply use daily dummy variables instead, with Friday being the baseline.

We allow for 10 different half-hour time periods in prime time, from a start time of 6 pm through to the last half-hour, which commences at 10:30 pm. These are dummy coded, with the 10:30–11:00 pm slot being the baseline, as it usually has the lowest ratings within this time period. Recent trends in television viewing show a declining pattern year-on-year (Helm, 2007), so we capture this trend using a linear covariate, "year", which is coded as 4–8, corresponding to the years 2004–2008. In case this downward trend is nonlinear, we also include a quadratic covariate, the square of the "year" variable.

Lastly, there is ample evidence that holidays such as Christmas, Thanksgiving and New Year's Day (and surrounding days) have lower PUTs than other periods in the year (Patelis et al., 2003). We therefore include dummy variables for the key holidays in this market.

6.2. Program-based covariates

Genre. Several studies (e.g., Henry & Rinne, 1984) have demonstrated the importance of program genre for ratings forecasts. The provider of our data categorized each program into one of 15 program genres (comedy, current affairs, documentary, drama, magazine, movie, news, soap, variety, sport, game, music, reality TV, science and travel). These genre categories are very similar to a typology developed by Henry and Rinne (1984). We simplify these genres into four broad categories, similar to those used by Horen (1980), namely "light content" (comedy, soap, variety, game, music, reality TV), "heavy content" (current affairs, magazine, documentary, drama, news, science, travel), sport and, lastly, movies. In addition, we include covariates for the program genre on the competing channels.[10] We do this for the 4 networks, but not the sports channel, as its genre is always the same. As there are 4 genres and 4 networks, there are $(4 - 1) \times 4 = 12$ genre dummy variables.

Lead-in. As was mentioned in the literature review, many previous studies have revealed the importance of the lead-in as a determinant of the rating for a program. In particular, Henry and Rinne (1984) define the lead-in for a program as the channel share for the program broadcast immediately prior to the current program. This is consistent with Rust and Alpert's (1984) audience flow model. While the lead-in is undoubtedly a key factor in describing television audience behavior, it presents problems when included in a forecast model, since its use as a covariate means that the channel share for the previous program also has to be predicted in order to get a rating forecast for the current program. This is likely to compound the forecast error (Danaher, 1994). Thus, we examine program rating forecasts both with and without the lead-in. A covariate related to the lead-in is the genre-match, which is coded as 1 if the current and preceding shows on a particular channel are of the same genre, and 0 otherwise.

Duration and Reruns. An additional covariate is the program length (measured in minutes), which was also included in Henry and Rinne's (1984) model. Finally, another factor which might influence the ratings for a program is whether or not it is a rerun. This information is available in our data, so we include a dummy variable to indicate whether the program is a rerun.

7. Results

7.1. Summary statistics

Table 2 reports summary statistics for the program-level covariates for each network and for one of the satellite sports channels. Channel 2 broadcasts the most programs, while the other channels show about the same number of programs. The first three networks have the highest ratings and shares, while the sports channel has the lowest average rating, although it does have a high variability relative to its mean. This is due to it showing live sports occasionally, at which times it obtains reasonably high ratings. Such a high variability is likely to make forecasting more challenging for this channel. In addition to the standard deviation, we also report the coefficient of variation (standard deviation divided by the mean) for the ratings and shares. They confirm the relatively higher variability in both ratings and share for the sports channel.

The percentage of programs which are reruns is 6.5% overall, but Channels 2 and 3 are much higher than this average. Across the four networks there is quite a lot of variation in program genre. Channels 2 and 4 are noted more for light entertainment, such as comedies and reality TV, while Channels 1 and 3 steer towards more serious viewing, such as drama, news and documentaries. Channel 1 is particularly strong

[10] As networks announce their upcoming program schedules well prior to the actual broadcast of the programs, a knowledge of the program genre within the forecast horizon can be assumed.

Table 2
Program covariate summary statistics by channel, 2004–2007.

	All channels	Channel				
		1	2	3	4	Sport
No. of programs	47 880	9501	10 308	9383	9385	9303
Program duration, mins	51.2 (34.6)[a]	51.3 (24.0)	49.1 (32.7)	53.1 (28.5)	52.5 (25.7)	50.0 (53.8)
Average rating,%	5.46 (5.0)	11.26 (5.1)	7.03 (3.7)	6.54 (2.4)	1.55 (1.0)	0.65 (1.5)
Coeff. variation, rating	0.91	0.45	0.53	0.37	0.65	2.31
Average share,%	16.3 (13.5)	29.9 (11.3)	18.1 (8.5)	17.0 (6.1)	4.7 (2.7)	2.0 (4.4)
Coeff. variation, share	0.83	0.38	0.47	0.36	0.57	2.20
Rerun,%	6.5	2.4	13.9	9.6	6.0	0
Genre_light,%	35.2	21.5	57.2	29.1	45.8	0
Genre_heavy,%	39.5	73.9	24.8	59.2	41.9	0
Genre_sport,%	21.7	2.9	1.5	1.9	7.6	100.0
Genre_movies,%	3.6	1.6	12.8	9.8	4.6	0
Genre match,%	54.7	50.0	54.2	42.0	39.8	100.0

[a] Standard deviations are in parentheses.

on "heavy" programs, while Channels 2 and 3 show more movies than the other two networks. The genre match variable indicates the percentage of occasions on which there is a match between the current and previous program genres, and thus is a measure of genre consistency. Channels 1 and 2 are the most consistent in genre, while Channels 3 and 4 are the least consistent.

7.2. Regression model for total audience

Table 3 gives the fitted model $\hat{Y}_t = \hat{\beta}_Y X_t$ for the total audience, based on the calibration period of 2004–2007. There are five main groupings of covariates: year (with both linear and quadratic terms), holidays, day of the week, season and time of the evening.[11] The final row of Table 3 gives the R^2 value for the model, which is a very high 85%. Gensch and Shaman (1980) also observed high R^2 values for their models of the total audience. We can therefore conclude that this basic set of covariates does a good job of explaining the total ratings.

Turning to the estimated regression coefficients, there is a downward annual linear trend, but the significant positive quadratic term indicates that this is leveling out. For the holidays, Labor Day Sunday, Christmas Day and the day after, New Year's Eve, New Year's Day and the long-weekend Friday and Saturday, all have lower total audiences than non-holidays. The day of the week effect exhibited in Fig. 2 is very apparent in the regression model coefficients. All days except Saturday have significantly higher viewings than the baseline day, Friday. As expected, there are strong seasonal effects which are captured by the harmonic variables, particularly for the annual and six month periods. Also, recall from Figs. 1 and 2 that the ratings are generally higher at 6 pm and 8 pm than at 10 pm. The coefficients in Table 3 reveal that the total audience peaks at around 7:30 pm to 8:00 pm and declines steadily thereafter.

7.3. Regression model for program ratings

Table 3 also gives the fitted regression models using Eq. (2), where the program ratings are regressed directly on the full set of time-based and program-based covariates, with separate models for each channel.[12] Overall, notice that the R^2 values for the

[11] Note that we give some holidays generic names to prevent the identification of the market location. In addition to the holidays listed we also initially tried some state holidays and high school holiday periods, but they were generally not significant and so were dropped from the model. We also tried a dummy variable to indicate periods of daylight saving, but this was highly correlated with the seasonal variables and was therefore removed. Moreover, the market is geographically large enough that rainfall and temperature do not influence the ratings over and above the seasonal effects which are already modeled explicitly.

[12] These estimated regression coefficients are broadly similar to those obtained for the random effects model in Eq. (5) with ratings

Table 3
Regression models for the total audience and program ratings.

	Total audience	Channel				
		1	2	3	4	Sport
Intercept	**−1.471**[a]	**−2.568**	**−4.258**	**−3.833**	**−3.638**	**−5.466**
Year	**−0.086**	**−0.207**	**0.262**	**0.102**	**−0.793**	**−0.262**
Year_squared	**0.006**	**0.010**	**−0.023**	**−0.010**	**0.074**	**0.023**
National day	**0.041**	**−0.144**	0.023	0.010	−0.029	0.178
Long weekend—Thu	**0.051**	−0.068	0.063	**0.124**	−0.094	**0.532**
Long weekend—Fri	**−0.126**	−0.062	0.059	**−0.270**	−0.204	−0.083
Long weekend—Sat	**−0.214**	**−0.191**	−0.125	**−0.135**	−0.203	−0.188
Long weekend—Sun	−0.025	0.000	−0.060	0.068	−0.115	0.030
War memorial day	**0.093**	0.069	0.045	−0.015	0.008	−0.148
Labor day Sunday	**−0.172**	**−0.163**	−0.018	**−0.167**	−0.127	**−0.471**
Labor day Monday	**0.033**	0.086	−0.018	−0.008	0.114	**−0.438**
Christmas Day	**−0.435**	**−0.399**	−0.145	**−0.285**	−0.202	−0.388
Day after Christmas	**−0.260**	**−0.251**	**−0.365**	**−0.275**	−0.215	**0.471**
New Year's Eve	**−0.477**	−0.119	**−0.384**	**−0.477**	**−0.375**	**−0.637**
New Year's Day	**−0.134**	0.022	−0.041	−0.011	−0.209	−0.183
Day after New Year	−0.034	−0.060	**0.249**	−0.060	−0.228	−0.274
Monday	**0.199**	**0.181**	**0.294**	**0.246**	**0.169**	**−0.653**
Tuesday	**0.175**	**0.222**	**0.208**	**0.232**	**0.123**	**−0.505**
Wednesday	**0.094**	**0.180**	**0.126**	**0.256**	−0.031	**−0.704**
Thursday	**0.069**	**0.132**	**0.212**	**0.075**	−0.009	**−0.838**
Saturday	**−0.067**	**−0.078**	**−0.108**	**−0.094**	**−0.332**	**0.536**
Sunday	**0.175**	**0.160**	**0.192**	**0.237**	−0.021	**−0.457**
cos1	**0.133**	**0.083**	**0.109**	**0.081**	**0.102**	**0.138**
cos2	**0.011**	−0.006	**0.033**	**0.012**	**0.023**	**0.099**
cos3	**0.004**	0.005	**0.014**	**0.011**	0.011	**−0.045**
cos4	**0.007**	**0.008**	**0.037**	−0.006	**0.018**	−0.009
cos5	**0.020**	0.002	**0.027**	0.007	0.005	**0.044**
cos6	**0.011**	−0.002	**0.018**	0.006	0.007	0.022
sin1	**0.010**	**−0.022**	−0.007	**0.024**	**0.077**	**−0.081**
sin2	0.001	−0.003	0.008	0.007	0.007	**−0.068**
sin3	**−0.010**	**−0.013**	−0.009	−0.005	**0.022**	**−0.055**
sin4	−0.001	−0.004	0.005	−0.003	0.004	−0.016
sin5	0.000	**−0.010**	**0.019**	0.007	−0.001	−0.018
sin6	0.001	**−0.010**	**0.009**	−0.001	−0.009	0.007
1800	**0.992**	**1.666**	0.004	**0.920**	**0.415**	0.067
1830	**1.086**	**1.534**	**0.424**	0.000	**0.430**	0.011
1900	**1.173**	**1.352**	**1.308**	**0.588**	**0.907**	**0.463**
1930	**1.211**	**1.332**	**1.187**	**0.691**	**1.128**	**0.686**
2000	**1.196**	**1.152**	**1.170**	**0.664**	**1.056**	**0.572**
2030	**1.093**	**1.000**	**1.163**	**0.842**	**1.009**	**0.910**
2100	**0.811**	**0.641**	**0.885**	**0.651**	**1.164**	**0.697**
2130	**0.711**	**0.475**	**0.893**	**0.557**	**0.610**	**0.651**
2200	**0.468**	**0.175**	**0.622**	**0.195**	**0.228**	**0.220**
Chan1_light	–	0.023	−0.018	**−0.056**	−0.028	**−0.205**

(*continued on next page*)

models predicting ratings directly, without the lead-in, are reasonably high for the three main networks, but much lower for Channel 4 and the sport channel.

All of the channels have significant time trends. Channels 1, 4 and Sport have declining ratings over

predicted directly. We report only this model, as it permits the calculation of an R^2.

Table 3 (*continued*)

	Total audience	Channel				
		1	2	3	4	Sport
Chan1_heavy	–	−0.009	0.080	**−0.071**	−0.045	**−0.159**
Chan1_sport	–	**0.256**	0.012	**−0.087**	−0.075	**−0.408**
Chan2_light	–	**0.126**	**−0.095**	**0.050**	**−0.073**	**0.105**
Chan2_heavy	–	**0.135**	**−0.128**	**0.033**	**−0.113**	**0.141**
Chan2_sport	–	**0.100**	**−0.243**	**0.130**	**0.133**	**−0.192**
Chan3_light	–	**0.075**	**−0.119**	0.025	−0.009	**0.126**
Chan3_heavy	–	**0.023**	−0.015	**0.152**	**−0.127**	−0.007
Chan3_sport	–	**−0.172**	**−0.109**	−0.017	0.024	**0.357**
Chan4_light	–	**−0.063**	0.013	−0.018	**0.441**	**0.146**
Chan4_heavy	–	−0.007	0.032	**−0.034**	**0.461**	**0.111**
Chan4_sport	–	**−0.099**	0.025	**−0.150**	**0.206**	**0.246**
Live sport	–	–	–	–	–	**2.056**
Program length	–	**−0.002**	**−0.001**	0.000	**0.007**	**0.004**
Rerun	–	**−0.042**	**−0.069**	**0.089**	**−0.054**	−0.254
Genre match	–	**0.018**	**0.062**	**0.044**	0.020	–
R^2,% (without lead-in)	84.8	83.2	74.2	57.1	41.4	42.6
Lead-in parameter	–	**1.019**	**1.453**	**1.094**	**3.469**	**6.324**
R^2,% (with lead-in)	–	84.2	76.3	58.5	42.7	47.4

[a] Coefficients which are statistically significant at the 5% level are bolded.

the period 2004–2007, but the positive quadratic trend indicates that the decline is leveling out. Holidays have different effects on the different channels, with no one holiday having a consistent impact on all of the channels. Day of the week effects are very pronounced for all of the channels, with the four networks having significantly higher ratings on Sunday to Thursday, but significantly lower ratings on Saturday than on the baseline day of Friday. By contrast, the sport channel does best on Saturdays, a day when a lot of live sport is aired. Annual seasonal effects consistently influence the ratings for all five channels, with a smattering of shorter time periods (6, 3 and 2 months) being associated with ratings for some channels. As for the total audience, the time of day effects are significant, with all channels exhibiting their highest ratings in mid-evening.

Regarding the program genre, Table 3 provides some interesting insights regarding the most attractive genre for a channel, and how the genres on competing channels influence a focal channel. For example, sports do significantly better than movies on Channel 1, while the predominant genre for Channel 1, “heavy” programs, does no better than movies.

The sport channel experiences a large increase in ratings when it shows live sport.[13] This is consistent with the findings of previous work by Nikolopoulos, Goodwin, Patelis, and Assimakopoulos (2007), who also find that forecasting ratings for live sport is extremely challenging.

The program length has a significant effect on the ratings for the majority of the channels, being sometimes negative (Channels 1 and 2) and sometimes positive (Channel 4 and Sport). Reruns generally result in lower ratings, as would be expected. Matching the genre from one program to the next results in higher ratings for the three largest networks.

Due to the previously-observed importance of including the lead-in for models of channel ratings (Napoli, 2001), we also ran separate models with the lead-in variable included, where the lead-in is defined as the channel share for the previous program on the same channel. The last two lines of Table 3 report just the estimate of the lead-in parameter (the other parameter estimates did not change all that much, so they are not reported again) and the R^2 values

[13] Note that this variable is unique to the sport channel.

Table 4
Average forecast errors as groups of covariates are added to the model.

	No. of programs forecast	Variables in model			
		Time trend, holidays, day of the week season, time of day	+ Genre	+ Duration, rerun, genre match	+ Lead-in
No. covariates		43	55	58	59
Channel 1	1045				
MAPE: RD		0.2052	0.2034	0.1984	0.1919
MAD: RD		2.0864	2.0644	2.0068	1.8842
Channel 2	1130				
MAPE: RD		0.2444	0.2345	0.2389	0.2476
MAD: RD		1.6153	1.5495	1.6076	1.5286
Channel 3	1051				
MAPE: RD		0.2199	0.2104	0.2104	0.2257
MAD: RD		1.1715	1.1273	1.1273	1.1764
Channel 4	1017				
MAPE: RD		0.6836	0.6983	0.6694	0.7854
MAD: RD		0.8118	0.8238	0.7884	0.9060
Sports	969				
MAPE: RD		0.9905	0.9897	1.0060	0.9297
MAD: RD		0.5902	0.5892	0.5890	0.5932
Average over all channels	5212				
MAPE: RD		0.4687	0.4673	0.4646	0.4761
MAD: RD		1.2550	1.2308	1.2238	1.2335

when the lead-in is included. Its significant positive influence is apparent for every channel, in keeping with previous studies. Moreover, the R^2 values are always significantly higher, as is confirmed by F-tests. This indicates that the lead-in variable always improves the fit of a ratings model. While it might be assumed that this should mean that including the lead-in should improve the predictions, we now demonstrate that it does not.

7.4. Selection of covariates for forecasting

As was mentioned above, the fact that a variable is significant in a model does not guarantee that it will improve prediction. Therefore, rather than using all of the covariates, as reported in Table 3, we introduce groups of covariates to see whether their addition improves or worsens the predictions. We use the ratings-direct regression models to predict the ratings for every program in the validation period, namely the first six months of 2008. There are 1057 unique programs in this validation period, but many are broadcast multiple times, making the total number of program forecasts 5212. The prediction errors are judged by two measures, MAPE and MAD, which have been used extensively for gauging the prediction accuracy of television ratings (e.g., Napoli, 2001; Nikolopoulos et al., 2007; Patelis et al., 2003).

Table 4 reports the prediction errors for each channel, and the average over all 5 channels, for four groupings of covariates. The first grouping contains the time-based variables, namely time trend, holidays, day-of-the-week, season and time of day. The second grouping is the collection of 12 program genre variables, the third grouping is the program-specific covariates of duration, rerun and genre match, and the last grouping is the single covariate of the lead-in. It should be noted that in progressing from the time-based to the genre column of Table 4, we add the covariates to the model cumulatively, so that the column headed "+ genre" means that the covariates included are the time-based ones plus the 12 program genre dummy variables. It can be seen that the lowest average prediction errors across all channels occur when the program-specific covariates are added, and that they increase again when the lead-in covariate is included. Hence, even though the lead-in variable is significant in the fitted model, it does not improve the predictions. This is probably due to the fact that predicting using the lead-in requires the channel share for the program previously shown to be forecast as well. In effect, this means that the covariate for the lead-in is itself subject to prediction errors, thus compounding the overall forecast error. We therefore do not include the lead-in in any of the subsequent forecast models, but we do retain all of the other covariates listed in Table 3.

7.5. Accuracy of rating forecasts

Total audience forecasts. Table 5 reports the prediction errors for the 8 forecast methods detailed above. Before discussing the program forecasts for each channel, we examine the predictions of the total audience, which are reported in the first four rows of Table 5. Given that some of the forecast methods depend on the accuracy of these PUT forecasts, it is helpful to gauge the relative size of the errors when forecasting the total audience. The forecast errors are reported for both the calibration and validation data, and the two sets of errors are very similar, indicating little degradation in forecast performance in the validation period. This is encouraging, and suggests that the covariates which we have chosen capture the variation in the total audience extremely well. The average MAPE error across the validation period is only 8%, with a MAD of 2.4 GRPs. This compares favorably with the MAPE errors from previous studies (of ratings or shares), as exhibited in Table 1.[14]

Program rating forecasts. Table 5 also reports the average program rating prediction errors for each channel. The MAPEs are lowest for the three largest networks, but are somewhat higher for Channel 4 and the sport channel. In contrast, the MADs reveal the opposite effect. The reason for the difference is that the ratings for Channel 4 and the sport channel are rather low, and since the denominator in the MAPE is the actual rating, the MAPE is inflated for these channels.

Since many previous studies have reported the prediction MAPE, this could potentially enable a comparison across studies. However, a clear limitation of this measure is that it is sensitive to the size of the actual ratings, with the MAPE increasing as actual ratings decline. Another consideration is that advertisers and broadcasters tend to measure the differences between actual and forecast ratings in absolute rather than relative terms. This suggests that the MAD is a more practical measure of the forecast accuracy in the television industry, as was also recommended by Nikolopoulos et al. (2007). We therefore place more emphasis on the MAD in our model comparison.

Several patterns emerge when comparing the eight models. First, naïve forecasts based only on historical data from the previous year generally do the worst. Second, comparing the econometric models that do not include random effects (RD, PUT $\times$ HS and PUT $\times$ PS) reveals that RD and PUT $\times$ PS perform similarly, with both being much better than PUT $\times$ HS. Third, there are mixed findings when comparing models that forecast ratings directly with those that forecast using the product of PUT and channel share predictions. Direct forecasts are better when using the MAD criterion and vice versa for the MAPE criterion. Fourth, adding program random effects provides a significant improvement in forecast accuracy, for both the RD and BMA methods. Fifth, the three Bayesian forecasting methods, despite being state-of-the-art, do not consistently outperform the corresponding traditional econometric methods.

Overall, the RD_RE and BMA_RE models produce the most accurate forecasts, with RD_RE doing best according to our favored MAD criterion. Given the added complexity of estimating the BMA_RE model relative to the RD_RE model (which we estimated using just SAS software), the RD_RE model also has pragmatic appeal.

7.6. Forecasts for regular programs

The top panel of Table 5 reports the forecast errors for all 5212 program slots in the validation period. Many of these programs are once-offs, such as movies and sports programs. Others last just one season or were only introduced in 2008. For all such programs, the estimate of their random effect is necessarily 0. For this reason we select a subset of these programs, namely just those which had at least 2 episodes for

[14] The MAD for the total audience cannot be compared with the previous studies listed in Table 1, as none of them report the forecast errors for the total audience.

Table 5
Average forecast errors for each channel.

						Average
Total audience						
MAPE: calibration						0.0833
MAD: calibration						2.3931
MAPE: validation						0.0801
MAD: validation						2.4396
All programs specific channels	Channel					All channels
	1	2	3	4	Sport	wtd average
MAPE: HIST	0.2506	0.3232	0.2815	0.5861	1.7677	0.6201
MAPE: RD	0.1984	0.2389	0.2117	0.6694	1.0051	0.4517
MAPE: PUT×HS	0.2174	0.2956	0.2553	0.5667	1.7103	0.5877
MAPE: PUT×PS	0.2066	0.2461	0.2138	0.6369	0.8105	0.4129
MAPE: RD_RE	0.1787	0.2132	0.2160	0.5349	0.7480	0.3690
MAPE: BMA_ RD	0.1988	0.2400	0.2258	0.6605	0.8903	0.4318
MAPE: BMA_ PUT×PS	0.2073	0.2438	0.2166	0.6578	0.8530	0.4250
MAPE: BMA_ RE	0.1792	0.2124	0.2175	0.5175	0.7548	**0.3672**
MAD: HIST	2.0710	1.9080	1.4119	0.7715	0.6578	1.3864
MAD: RD	2.0068	1.6076	1.1491	0.7884	0.5891	1.2460
MAD: PUT×HS	1.9139	1.7697	1.3217	0.7759	0.6465	1.3055
MAD: PUT×PS	2.0681	1.6418	1.1591	0.7638	0.5719	1.2597
MAD: RD_RE	1.8397	1.4675	1.1399	0.6613	0.5116	**1.1410**
MAD: BMA_RD	2.0195	1.6285	1.1574	0.7769	0.6779	1.2690
MAD: BMA_ PUT×PS	2.0986	1.6487	1.1510	0.7667	0.6813	1.2866
MAD: BMA_ RE	1.8497	1.4670	1.1482	0.6415	0.5907	1.1522
Regular programs specific channels	Channel					All channels
	1	2	3	4	Sport	wtd average
MAPE: HIST	0.1661	0.2647	0.2198	0.4692	–	0.2462
MAPE: RD	0.1244	0.2082	0.1898	0.5919	–	0.2286
MAPE: PUT×HS	0.1594	0.2488	0.2012	0.4394	–	0.2303
MAPE: PUT×PS	0.1287	0.2085	0.1899	0.5533	–	0.2247
MAPE: RD_RE	0.1215	0.1832	0.1699	0.5630	–	0.2119
MAPE: BMA_ RD	0.1259	0.2040	0.1912	0.7851	–	0.2557
MAPE: BMA_ PUT×PS	0.1333	0.2051	0.1891	1.0458	–	0.2941
MAPE: BMA_ RE	0.1204	0.1746	0.1704	0.5154	–	**0.2034**
MAD: HIST	1.7176	2.5005	1.2850	0.7638	–	1.5868
MAD: RD	1.4030	2.1810	1.1770	0.7469	–	1.3842
MAD: PUT×HS	1.6818	2.3394	1.1969	0.7403	–	1.5103
MAD: PUT×PS	1.4565	2.1649	1.2061	0.7217	–	1.4048
MAD: RD_RE	1.3370	1.9075	1.0932	0.7809	–	1.2858
MAD: BMA_RD	1.4219	2.1204	1.1761	0.7459	–	1.3780
MAD: BMA_ PUT×PS	1.4888	2.1217	1.1808	0.9798	–	1.4341
MAD: BMA_ RE	1.3217	1.7896	1.0821	0.5995	–	**1.2288**

at least two seasons prior to 2008 and were also broadcast in the 2008 validation period. These can be considered as "regular programs", which might be expected to be easier to forecast since they may have a more stable audience. For regular programs, the number of program slots is reduced to 1896, comprising 88 unique programs, including the two depicted in Figs. 3a and 3b.

The forecast errors for regular programs are reported in the lower panel of Table 5. The smallest average MAPE is markedly lower than that for all programs, being 0.2034 rather than 0.3672. While this appears encouraging, the smallest average MAD for regular programs is actually higher, at 1.2288, compared with 1.1410 for all programs. The reason for the large reduction in MAPE but little change in MAD, is that regular programs tend to have higher ratings, and therefore the percentage error is lower. Comparing the eight models, we again see the benefit of incorporating random effects, as the two random-effect models have the lowest forecast errors. This time, however, the BMA_RE model is consistently the best. Even though it might be anticipated that forecasting ratings would be easier for regular programs than for all programs, our results show otherwise. Furthermore, it is heartening to see that our best forecast models are very robust, performing slightly better for all programs combined than for programs that have an established audience.

7.7. *Financial impact of poor forecasts*

As was mentioned above, inaccurate forecasts have cost implications for both broadcasters and advertisers. For example, network television advertising time is normally purchased "upfront" in May for the fall season, with the ad costs being based on forecast ratings. If the forecast total-campaign GRPs are higher than the actual GRPs delivered, then advertisers demand a "make-good" to compensate for the shortfall (Kelton & Schneider-Stone, 1998). The make-good is for the amount of the GRP deficit and has an opportunity cost to the broadcaster, since the make-good advertisements are free and cannot be sold to another advertiser (Consoli, 2008). This cost is significant. For example, in the fourth quarter of 2007 the estimated cost to the US television networks of make-goods was $200 million (Lafayette, 2007).

Of course, the advertisers do not expect the predicted GRPs to be exactly the same as the eventual GRPs, and, moreover, both are subject to sampling error, as they are derived from a panel. Hence, some tolerance is permitted, with typical thresholds for under-delivery in total campaign ratings being 3%–5% (Friedman, 2007).[15] Also note that the GRP comparisons are made for the entire campaign, not on a program-by-program basis.

As it happens, we have the rate card cost for every program in our dataset. For proprietary reasons we cannot divulge the actual costs, but we can calculate the expected cost of a campaign GRP shortfall and give this as the percentage of the total campaign cost. For example, if the GRP deficit is 10% for a campaign that was predicted to achieve 200 GRPs, then the expected cost of the GRP under-delivery is (200–180) × Potential Audience × CPM, where the Potential Audience is the market size (in thousands) and CPM is the "cost per thousand" for an advertisement—a standard industry measurement of ad cost. Since the CPM varies between programs, we take the average ad cost over all of the ad slots comprising the campaign. Suppose that the expected cost of the 20 GRP shortfall is $200,000 for a campaign costing $5 million; the percentage loss to the broadcaster on subsequent make-goods to the advertiser is therefore 4%.

To obtain a reliable estimate of the cost of inaccurate forecasts, we simulate 1000 typical advertising campaigns using predicted GRPs from the common industry model (historical forecasts) and one of the best-performing econometric models (RD_RE). We use predicted and actual GRPs only for the validation period in 2008, and compute the percentage loss to the broadcaster in the event of the actual campaign GRPs being short of the predicted amount by at least a threshold tolerance of 3%, 5% or 7%. Table 6 illustrates the financial impact of the forecasting errors, where we multiply the average percentage loss by the combined network television ad revenue for the US in 2008, namely $46.4 billion (Advertising Age, 2009).

At a demanding 3% threshold level, the expected cost of make-goods to networks when using the prevalent historical model is $706 million. This is broadly in line with the present estimated cost of $200 million per quarter, as reported by Friedman (2007). If the RD_RE model is used instead, then the expected make-good costs reduce markedly to $120 million, a saving to the network television industry of $586 million per annum. Even at the least-demanding threshold of 7%, the effect on the television industry of bad forecasts based on the historical model is substantial, being $264 million. The RD_RE model does much better in this case, with costs of only

[15] A more statistically-based approach, taking into account the sample size of the peoplemeter panel and the actual rating value, indicates that a more realistic tolerance threshold would be closer to 7%.

Table 6
Expected network "make-good" costs for forecasting models.

	Forecast method	Campaign GRP shortfall threshold		
		3%	5%	7%
Expected total cost to United States network television advertising industry, $million	HIST	706	451	264
	RD_RE	120	38	14
Difference between HIST and RD_RE models, $million		586	413	250

$14 million, a saving of $250 million. Thus, going to the effort of improving ratings forecasts is financially beneficial, as the last row of Table 6 shows that the better model can potentially save the television advertising industry between $250 million and $586 million per annum.

8. Conclusion

Forecasting television ratings remains one of the most important challenges in media planning and buying. Despite its importance, forecasting methods to date have been sparse in number and inadequately tested. In addition, there have been no studies using data from a current, complex television environment with numerous channels and competing broadcast platforms. Moreover, the previous forecasting models have been relatively straightforward regression-style models, with no modern methods employed. This study addresses these shortcomings by looking at data from a market in a very recent time period, where 70 channels were available. We also contrast the forecasting abilities of eight different models in validity tests over a six month period and comprising over 5000 program forecasts. No previous studies have been this comprehensive or rigorous.

In making overall recommendations, three key issues arise: (1) which covariates to include in the model; (2) whether to model ratings directly or as the product of the total audience and channel share; and (3) which is the most accurate model. We now address each of these issues in turn.

First, it is almost always important to include exogenous covariates such as holidays, season, day of the week, time of day and yearly trends in a model. They are particularly important for a model of the total audience, which is where our forecast model is particularly good, having a MAPE of just 8%. We reiterate the fact that seasonal factors are best captured using sinusoidal variables, rather than dummy variables for each month or quarter. The inclusion of a quadratic term for the yearly trend also improved the forecasts noticeably. Two potentially important variables are the program genre and lead-in. Both variables significantly improve the model fit, but only the program genre helps to improve the ratings forecasts. This may not be too surprising, since the lead-in must itself be forecast in future time periods, thus adding variation to one of the predictor variables. This is a useful managerial finding, since it means that the complications arising from having to also forecast the channel share for the previous program can be avoided, and all that is needed is a model that uses information for the current programs. In contrast, including information about the program genre on the focal and competing channels does significantly reduce the forecast errors.

Second, our finding regarding the superiority of forecasting ratings directly is contrary to the established wisdom, which recommends the development of separate models for the total audience and channel share, with the models subsequently being recombined via multiplication (Gensch & Shaman, 1980; Patelis et al., 2003). Furthermore, although the results are not reported in detail, we found that developing separate models for each channel results in much better forecasts than having a combined model with channel dummy variables.

Third, the forecast accuracy across the eight models was assessed by the MAPE and MAD measures, and the best model often varied depending on which error measure was used. As was recommended by Nikolopoulos et al. (2007), if we consider the MAD as the primary measure of forecast accuracy, then our best model is a linear regression model that predicts ratings directly and incorporates random effects for each program. However, the BMA model is a close second; it also predicts the ratings directly, with the

MAD values being 1.14 and 1.15 for the non-Bayes and Bayes models, respectively. A common feature of these models is that they both predict ratings directly and include program random effects. Our lowest MAD is 1.14, which, out of all previous studies (see Table 1), has only been bettered by Van Meurs (1994). Given that his prediction method uses just historical ratings, and that our model consistently outperforms this naïve method using our data, we are confident that our RD_RE model would have outperformed his method had the data been the same.

Although we have endeavored to be comprehensive in developing a range of forecast models, we acknowledge that other models are possible, and that some of these may improve on our chosen models. However, Table 5 clearly shows that substantial gains in forecast accuracy are unlikely to eventuate, even from a highly complex model. Our use of Bayesian model averaging, which is among the best forecasting techniques in the present literature, shows that large reductions in forecast errors are difficult to achieve. Probably the most helpful addition to existing models is the inclusion of random effects, as the forecast accuracy increases markedly in both cases. Beyond that, additional improvement looks to be difficult. Nonetheless, we encourage further research that can improve ratings forecasts even more.

Finally, even though the forecasting improvements from more sophisticated models may not seem very substantial in absolute error terms, we demonstrate that, in economic terms, using our best model can save the network television industry somewhere between \$250 million and \$586 million per annum, as improved forecasting results in less make-good advertising compensation. At a time when network advertising revenues are falling, savings of this magnitude are certain to be welcome.

Appendix. Technical appendix for TV forecasting models

In this Appendix we show how to undertake Bayesian model averaging in a linear regression model with additive random effects. The method can also be used to compute BMA estimates without random effects by setting the effects to zero throughout.

The BMA linear regression model with random effects can be defined for a vector of n observations y on a dependent variable as:

$$y = X(\gamma)\beta(\gamma) + Z\eta + e, \quad e \sim N(0, \sigma^2 I_n). \quad \text{(A.1)}$$

Here, γ are the indicators defined in Eq. (6); $\beta(\gamma)$ are the non-zero coefficients with corresponding design matrix $X(\gamma)$; $\eta = (\eta(1)', \ldots, \eta(K)')'$ is a vector of K sets of random effects; and Z is a block diagonal matrix $Z = b\mathrm{diag}(Z_1, \ldots, Z_K)$ of design matrices Z_p for each random effect $\eta(p)$. For example, if only a constant random effect is to be considered for program p, then $Z_p = (1, 1, \ldots, 1)'$ and $\eta(p)$ is a scalar. The random effects are assumed to be independently distributed $\eta(p) \sim N(0, \Sigma)$, for $p = 1, \ldots, K$, where $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_L^2)$ is a diagonal matrix of length L.

Following Smith and Kohn (1996), we assume the following prior for the nonzero fixed effect regression coefficients:

$$\beta(\gamma) \mid \gamma, \sigma^2, \eta \sim N\left(\hat{\beta}(\gamma, \eta), n\sigma^2 \left(X(\gamma)'X(\gamma)\right)^{-1}\right), \quad \text{(A.2)}$$

which is the widely used g-prior centered around $\hat{\beta}(\gamma, \eta) = \left(X(\gamma)'X(\gamma)\right)^{-1} X(\gamma)'(y - Z\eta)$. The prior on the vector of m indicator variables is given by Cripps et al. (2005), and places equal prior weight on subsets of exogenous variables with different numbers of terms.

To estimate the model, we use the following MCMC sampling scheme, which generates the random effect vectors η directly.

MCMC sampling scheme

Step (1): For $i = 1, \ldots, m$, generate from $f(\gamma_i \mid \gamma_{k \neq i}, \eta, y)$.

Step (2): Generate from $f(\sigma^2, \beta(\gamma), \{\sigma_1^2, \ldots, \sigma_L^2\} \mid \eta, \gamma, y) = f(\sigma^2 \mid \eta, \beta(\gamma), y) f(\beta(\gamma) \mid \sigma^2, \eta, \gamma, y) \prod_p f(\sigma_p^2 \mid \eta(p))$.

Step (3): Generate from $f(\eta(p) \mid \sigma^2, \beta(\gamma), \{\sigma_1^2, \ldots, \sigma_L^2\}, y)$.

The derivation of the posterior at Step (1) follows computations similar to those outlined by Smith and Kohn (1996, pp. 320–321), while those at Steps (2) and (3) are similar to those outlined by Browne and Draper (2006, p. 484). Generating from each of the steps sequentially is called a "sweep" of the sampling scheme, and the approach involves many sweeps, after which the Markov chain is assumed to have converged to the posterior distribution. Further iterations are then collected, and these form a Monte Carlo sample from which estimates are computed. In particular, if the iterates of the indicators and random effects are denoted as

$$\{(\gamma^{[1]}, \eta^{[1]}), \ldots, (\gamma^{[J]}, \eta^{[J]})\},$$

then the BMA estimate of the regression coefficients is given by

$$E(\beta \mid y) \approx \hat{\beta}_{\mathrm{BMA}} = \frac{1}{J} \sum_{j=1}^{J} E(\beta \mid \gamma^{[j]}, \eta^{[j]}, y).$$

The conditional expectation in the sum is simply $\hat{\beta}(\gamma, \eta)$. Numerically, the estimate is the average of the conditional expectation of β over the posterior distributions of the subset indicator γ and random effects η. The estimate of the random effects is also computed from the Monte Carlo iterates as

$$E(\eta(p) \mid y) \approx \hat{\delta}_{\mathrm{BMA}}(p) = \frac{1}{J} \sum_{j=1}^{J} \eta(p)^{[j]}.$$

References

Advertising Age (2009). US measured ad spending by category. http://adage.com/datacenter/datapopup.php?article_id=137426 (accessed 07.07.09).

Ansari, A., Essegaier, S., & Kohli, R. (2000). Internet recommendation systems. *Journal of Marketing Research*, *37*, 363–375.

Allenby, G. M., & Rossi, P. E. (1998). Marketing models of consumer heterogeneity. *Journal of Econometrics*, *89*(1–2), 57–78.

Beville, H. (1985). *Audience ratings*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Browne, W. J., & Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis*, *1*(3), 473–514.

Callan, J. (2009). Global ad spending to drop most in 29 years, hurting TV, Print. http://www.bloomberg.com/apps/news?pid=newsarchive\&sid=anWJDu4QmrsE (accessed 19.05.09).

Cattin, P., Festa, R., & Le Diberder, A. (1994). A model for forecasting the audience of TV programs. In *Advertising research foundation. European society for opinion and marketing research: worldwide electronic and broadcasting audience research symposium. Paris.* vol. 1. (pp. 513–524).

Clyde, M., & George, E. (2004). Model uncertainty. *Statistical Science*, *19*, 81–94.

Consoli, J. (2008). TV retains marketing dollars in hard times. *New York Times*, December 29. http://www.nytimes.com/2008/12/30/business/media/30adco.html (accessed 26.06.09).

Cripps, E., Carter, C., & Kohn, R. (2005). Variable selection and covariance selection in multivariate regression models. In D. K. Dey, & C. R. Rao (Eds.), *Handbook of statistics 25: Bayesian thinking—modeling and computation* (pp. 519–552). Elsevier, North-Holland.

Danaher, P. J. (1994). Comparing naive with econometric market share models when competitors' actions are forecast. *International Journal of Forecasting*, *10*(2), 287–294.

Danaher, P. J., & Mawhinney, D. (2001). Optimizing television program schedules using choice modeling. *Journal of Marketing Research*, *38*(3), 298–312.

Darmon, R. Y. (1976). Determinants of TV viewing. *Journal of Advertising Research*, *16*(6), 17–20.

Ehrenberg, A. S. C., & Wakshlag, J. (1987). Repeat-viewing with people-meters. *Journal of Advertising Research*, 9–13.

Fernandez, C., Ley, E., & Steel, M. (2001). Benchmark priors for Bayesian model averaging. *Journal of Econometrics*, *100*, 381–427.

Friedman, W. (2007). Crazy make-goods, fewer ratings, and now no TV writers. http://www.mediapost.com/publications/index.cfm?fa=Articles.showArticle\&art_aid=70547 (accessed 05.07.09).

Gaebler, (2009). How much do television ads cost. http://www.gaebler.com/Television-Advertising-Costs.htm (accessed 19.05.09).

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

George, E., & McCulloch, R. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, *88*, 881–889.

Gilbride, T., Allenby, G. J. M., & Brazell, J. D. (2006). Models for heterogeneous variable selection. *Journal of Marketing Research*, *43*, 420–430.

Givon, M., & Grosfeld-Nir, A. (2008). Using partially observed Markov processes to select optimal termination time of TV shows. *Omega*, *36*, 477–485.

Goodhardt, G., Ehrenberg, A., & Collins, M. (1975). *The television audience: patterns of viewing*. Lexington, Massachusetts: Lexington.

Greene, W. (2003). *Econometric analysis* (5th ed.). New Jersey: Prentice Hall.

Hastie, T., Tibshirani, R., & Friedman, J. J. (2001). *The elements of statistical learning: data mining, inference and prediction*. New York, NY: Springer.

Helm, C. (2007). *IBM consumer survey shows decline of TV as primary media device*. IBM Press Room. http://www-03.ibm.com/press/us/en/pressrelease/22206.wss (accessed 16.04.09).

Henry, M. D., & Rinne, H. J. (1984). Predicting program shares in new time slots. *Journal of Advertising Research*, *24*(2), 9–17.

Horen, J. H. (1980). Scheduling of network television programs. *Management Science*, *26*(4), 354–370.

Katz, H. (2003). *The media handbook: a complete guide to advertising, media selection, planning, research and buying* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Kelton, C. M. L., & Schneider-Stone, L. G. (1998). Optimal television schedules in alternative competitive environments. *European Journal of Operational Research*, *104*, 451–473.

Lafayette, J. (2007). Nets may run out of make-good time. http://www.tvweek.com/news/2007/11/nets_may_run_out_of_makegood_t.php (accessed 05.07.09).

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.

Napoli, P. M. (2001). The unpredictable audience: An exploratory analysis of forecasting error for new prime-time network television programs. *Journal of Advertising*, *30*(2), 53–101.

Nikolopoulos, K., Goodwin, P., Patelis, A., & Assimakopoulos, V. (2007). Forecasting with cue information: a comparison of multiple regression with alternative forecasting approaches. *European Journal of Operational Research*, *180*(1), 354–368.

Nielsen Research (2008). Season to date broadcast prime-time network average. Nielsen TV ratings. http://tvbythenumbers.com/ (accessed 12.11.08).

Patelis, A., Metaxiotis, K., Nikolopoulos, K., & Assimakopoulos, V. (2003). FORTV: decision support system for forecasting television viewership. *Journal of Computer Information Systems*, (Summer), 100–107.

Raftery, A. E., Madigan, D., & Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, *92*, 179–191.

Raftery, A. E., Gneiting, T., Balabdaoui, F., & Polakowski, M. (2005). Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, *113*(5), 1155–1174.

Reddy, S. K., Aronson, J. E., & Stam, A. (1998). SPOT: scheduling programs optimally for television. *Management Science*, *44*(1), 83–102.

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

Rust, R. T., & Eechambadi, N. V. (1989). Scheduling network television programs: a heuristic audience flow approach to maximizing audience share. *Journal of Advertising*, *18*(2), 11–18.

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.

Smith, M., & Kohn, R. (1996). Nonparametric regression via Bayesian variable selection. *Journal of Econometrics*, *75*, 317–344.

Smith, M., Mathur, S., & Kohn, R. (2000). Bayesian semiparametric regression: an exposition and application to print advertising data. *Journal of Business Research*, *49*(3), 229–244.

Tavakoli, M., & Cave, M. (1996). Modeling television viewing patterns. *Journal of Advertising*, *25*(4), 71–86.

The Ad Class (2010). Super Bowl XLIV 2010 best ads/commercials. http://www.theadclass.com/events/super-bowl-xliv-44-2010-ads (accessed 27.02.10).

Van Meurs, L. (1994). The future at your fingertips: towards an automated forecasting model for television ratings. In *Proceedings of the ESOMAR/ARF worldwide electronic and broadcast audience research symposium. Paris* (pp. 497–512).

Verbeke, G., & Molenberghs, G. (1997). *Linear mixed models in practice: a SAS-oriented approach*. New York, NY: Springer.

Weber, R. (2002). Methods to forecast television viewing patterns for target audiences. In A. Schorr, B. Campbell, & M. Schenk (Eds.), *Communication research in Europe and abroad—challenges of the first decade*. Berlin: DeGruyter.

Webster, J. G., & Wakshlag, J. J. (1983). A theory of television program choice. *Communication Research*, *10*(4), 430–446.

Wells, M. (1997). New fall shows disappointing advertisers. *USA Today*, November 17.

Wright, J. H. (2008). Forecasting US inflation by Bayesian model averaging. *Journal of Forecasting*, *28*(2), 131–144.

Wu, S., Harris, T. J., & McAuley, K. B. (2007). The use of simplified and misspecified models: linear case. *Canadian Journal of Chemical Engineering*, *85*, 386–398.

Yao, D. (2009). Global ads forecast to fall 6.9% in 2009. http://news.smh.com.au/breaking-news-business/global-ads-forecast-to-fall-69-in-2009-20090415-a6k5.html (accessed 19.05.09).

Yoo, C. Y., & Kim, H.-G. (2002). An analysis of prediction error for new prime-time television programmes: a competitive study between USA and Korea. *International Journal of Advertising*, *21*, 525–546.

ZenithOptimedia (2009). ZenithOptimedia releases April 2009 ad expenditure forecasts. http://www.zawya.com/story.cfm/sidZAWYA20090414132254/%20ZenithOptimedia%20releases%20April%202009%20ad%20expenditure%20forecasts (accessed 19.05.09).

Zheng, B., & Agresti, A. (2000). Summarizing the predictive power of a generalized linear model. *Statistics in Medicine*, *19*, 1771–1781.

Zubayr, C. (1999). The loyal viewer? Patterns of repeat viewing in Germany. *Journal of Broadcasting and Electronic Media*, *43*(3), 346–363.

Peter Danaher is the Coles Myer Chair of Marketing and Retailing at the Melbourne Business School. He was previously at the University of Auckland and has had visiting positions at London Business School, The Wharton School and MIT. Peter serves on the Editorial Boards of the *Journal of Marketing*, the *Journal of Marketing Research*, *Marketing Science* and the *Journal of Service Research*. He is also an Area Editor for the *International Journal of Research in Marketing*. His primary research interests are media exposure distributions, advertising effectiveness, television audience measurement and behavior, internet usage behavior, customer satisfaction measurement, forecasting and sample surveys.

Tracey Dagger is an Associate Professor in Marketing at Monash University, Melbourne. Her research interests include service quality and satisfaction, the economic and social outcomes of services, relationship development, co-creation in service contexts, and advertising research. She has won several international research awards and has previously published in the *Journal of Service Research* and the *Journal of Services Marketing*.

Michael Smith joined Melbourne Business School in 2007 as a Professor of Econometrics. He was previously an Associate Professor in the discipline of Econometrics and Business Statistics at the University of Sydney. During that time he also held visiting

positions at the University of Munich and The Wharton School at the University of Pennsylvania. Michael's research is evenly balanced between the development of econometric and statistical models and methods, and their application for solving problems arising in business and elsewhere. He is prominent internationally for his work on Bayesian semiparametric modeling and model averaging in cross-sectional, spatial and time series contexts. His work has appeared widely in top journals in econometrics and statistics, including the *Journal of Econometrics, Journal of the American Statistical Association, Journal of Business and Economic Statistics, Journal of the Royal Statistical Society Series B* and the *Journal of Computational and Graphical Statistics*.