Prediction of TV ratings with dynamic models

Roberto Pagano
Politecnico di Milano, DEIB
Milan, Italy
roberto.pagano@polimi.it

Massimo Quadrana
Politecnico di Milano, DEIB
Milan, Italy
massimo.quadrana@polimi.it

Paolo Cremonesi
Politecnico di Milano, DEIB
Milan, Italy
paolo.cremonesi@polimi.it

Sergio Bittanti
Politecnico di Milano, DEIB
Milan, Italy
sergio.bittanti@polimi.it

Simone Formwentin
Politecnico di Milano, DEIB
Milan, Italy
simone.fomentin@polimi.it

Andrea Mosconi
Politecnico di Milano, DEIB
Milan, Italy
andrea.mosconi@polimi.it

ABSTRACT

This paper describes and compares different time-series forecasting algorithms for the prediction of television ratings (i.e., the percentage of TV households tuned to a specific TV station). The prediction is based on historical viewing habits, schedule and attributes of TV shows, and other contextual information such as hour and day. The importance of predicting TV audience extends beyond mere curiosity. With the billions of dollars spent annually on TV programs and commercials, reliable TV audience information is required to evaluate and maximize the effectiveness of this investment. The cost of an advertisement is strongly related to the audience size for the program in which it is embedded, with higher rating programs costing more. Projections of audience size are made before TV programs have been aired and a mismatch might exists between estimated and true audience. This mismatch leads to potentially lost of revenues for broadcasters (when the true audience is larger than the estimated one) or to dissatisfaction of advertisers (when the true audience is smaller than the estimated one). The accuracy of the algorithms is evaluated on a dataset containing 21 millions TV viewing events collected from 5,666 households over 217 channels during a period of 4 months.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Information filtering*

General Terms

Audience, Ratings

Keywords

Television ratings, Television audience, Dynamic models, Time-series analysis, Context-aware recommender systems

RecSysTV 2015
2nd Workshop on Recommendation Systems for TV and Online Video

1. INTRODUCTION

Predicting accurate information about TV audiences is important for broadcasters, advertisers and advertising agencies. Television networks sell time to advertisers at a price that depends on the projected television audience [31]. As an example, the average cost for a 30 seconds TV commercial in the U.S. ranges between $100 thousand and $2.4 million, depending on the expected audience of the TV program [7]. According to recent data presented in [14, 23], worldwide television advertising expenditure was $178 billion in 2014 ($78 billion in the U.S. alone) and is expected to reach $236 billion in 2020.

The size of television audience is measured either in terms of *television ratings* (percentage of households tuned to a specific TV station) or *rating points* (one ratings point represents 1% of viewers). Rating points are the currency for the TV commercials. As an example, in the U.S. the average value of one rating point is $780 million per year and can be as high as $70 thousand per minute [24].

Television advertising time is purchased some time in advance, with advertisers planning to achieve a target number of ratings points during their commercials [15][17]. Hence, accurate ratings forecasts are vital to both advertisers and TV networks, with a difference of just one rating point resulting in a substantial gain or loss for either a broadcaster or an advertiser [21][10].

Rating forecasts are based on historical measurements of TV viewing habits. Measurements rely on TV meters installed in a representative sample of households by independent media companies (e.g., Nielsen, Auditel, AGF, BARB) As an example, in 2014 Nielsen was using a sample of more than 5,000 households, containing over 13,000 people, to be representative of 116.3 million TV homes in the U.S. [24]. Recently, some works suggested the usage of social networks as an alternative way to measure television ratings [5][29].

There are two possible approaches to the problem of predicting television ratings, differing on the granularity of the data used by the models [21].

- Algorithms in the first family build a model to estimate future viewing habits from raw household measurements, which are later aggregated into television ratings (Figure 1.a). Examples are collaborative-filtering algorithms for the watch–next problem [12].
- Algorithms in the second family aggregate measured viewing habits into historical television ratings which are later used to build a model (Figure 1.b). Recent

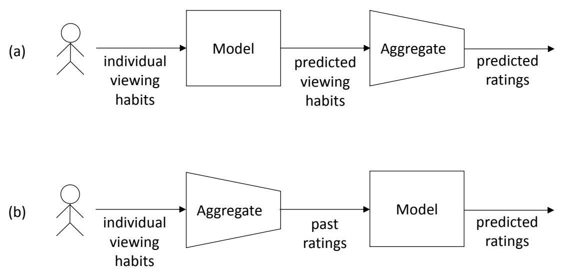

Figure 1: Data aggregation and alternative models

works show that data aggregation simplifies non linearity and improves accuracy of predictions [21].

1.1 Previous work

There is a lack of recent research works on the problem of predicting television ratings, mainly because most of the datasets are proprietary. The few works in the literature describe algorithms that use static regression models, with covariates based on attributes of programs and time of viewing events [20][21]. These algorithms give static predictions and the forecasts do not depend on the prediction horizons chosen, but only on covariates [7][8]. What is missing from previous works is the development of models reflecting the lead-in effect, i.e., the fact that television programs inherit viewers from the immediately preceding programs scheduled on the same TV channel [28].

Moreover, the results are validated only on 5 or fewer TV channels and, in most of the cases, only on prime-time programs (e.g., between 20:30 and 22:30). However, modern TV environment scenarios have hundreds of TV channels. As an example, a recent study by Nielsen shows that the typical household in the U.S. has almost 200 channels to choose from [24].

1.2 Our contribution

The focus of this study is on the development of models able to handle thousands of TV programs and hundreds of TV channels, and reflecting the fact that observations close together in time are more closely related than observations further apart (lead-in).

We describe and compare two different *auto-regressive* models for the prediction of television ratings. The predictions are based on historical viewing habits, past and future schedule of TV programs, attributes of programs, and time-related contextual information, such as hour and day. We use exogenous variables to capture the intrinsic appeal of new TV programs from observed characteristics of past TV programs with similar attributes.

Our dynamic models are compared with three state-of-the art approaches: (i) context-aware collaborative-filtering with fuzzy seasonal context described in [12]; (ii) static regression model with program-based covariates (e.g., genre, sub-genre) and time-based covariates (e.g., day of week, hour of day) described in [8]; and (iii) nested logistic model described in [7].

We validate our approach on a dataset containing 21 millions TV viewing events collected from 5,666 households over 217 channels during a period of 4 months, for a total of 21,194 distinct TV programs. The dataset is available for download [1].

1.3 Organization of the paper

The paper is organized as follows. Section 2 provides a review of previous works. Section 3 describes the problem and presents essential definitions. Section 4 introduces our dynamic models. Section 5 introduce a number of baseline algorithms and compare them with the dynamic models. The results of the experiments are discussed in Section 6. Finally, Section 7 draws the conclusions, pinpoints our key contribution, and outlines some challenging directions for future research.

2. RELATED WORKS

2.1 Aggregated models

Meyer and Hyndman [21] investigate the effect of aggregation in relation to the accuracy of television rating forecasts. Models are fitted using neural networks, decision trees and regression. They show that data aggregation serves to simplify non linearity.

Danaher et al.[8] describe a two-steps logistic model: the first step models the decision to switch the TV on (viewer availability), and the second step models channel selection (program choice). The model is validated against a Bayesian averaging model, focusing on the 5 most popular TV channels during prime time (20:30–22:30).

Nikolopoulos et al.[25] show that multiple regression models outperform item-based nearest neighbors approach when forecasting TV ratings for 12 sport events (test set) based on the ratings of 34 similar events (training set).

Kelton et al. [16] and Reddy et al. [26] design models for the optimal scheduling of TV programs in order to increase future TV ratings.

Patelis et al. [20] describe a first attempt to use dynamic models to predict TV ratings: ratings are predicted finding weeks that have the same type of competitiveness with the one to forecast, while a time series approach is used to model total viewership.

2.2 Direct models

A number of papers address the watch–next problem, which aims to provide recommendations for the next program to watch following the currently watched program [6][12].

One of the first paper on TV recommender systems is [9] which presents a personalized Electronic Program Guide (EPG). A number of other works focus on the concept of personalized EPG. Cotter and al. in [4] present a personalized EPG where the selected TV shows are based on a hybrid recommender system which mixes collaborative and content-based recommendations. Users manually input their preferences about channels, genres, and viewing times. This information is combined with the user's viewing activity by means of case-based reasoning and collaborative filtering techniques. Ardissono et al. in [3] present a personalized EPG where recommendations are generated locally on the client side using a hybrid approach on the basis of three information sources: user's implicit preferences represented

[1] http://recsys.deib.polimi.it/tv-audience-dataset/

in terms of program categories and channels (content attributes are downloaded from the satellite stream), (ii) user classes, and (iii) user viewing activity.

Many works present recommender systems based on hybrid collaborative and content based approaches [2]. The work in [13] describes a TV recommender systems which combines together content-based and collaborative filtering by means of Neural Networks. The system also uses information on the users, such as demographic information, interests, and moods. Martinez et al. in [19] exploit a hybrid approach to solve new-item, cold-start, sparsity, and overspecialization problems. The methods mix together in the same interface the outcome of content-based (computed using the cosine similarity among item feature vectors) and collaborative filtering (using Singular Value Decomposition to reduce the size of item's neighborhood).

3. DEFINITION OF THE PROBLEM

The size of television audience is described in terms of television ratings. Television ratings can be measured either with *rating points* (percentage of households tuned to a specific TV station) or with *viewing time* (average number of minutes a household is tuned to a specific TV station). It is always possible to convert rating points in viewing time, and viceversa, by knowing the number of households in the sample.

Throughout the rest of the paper we will use the term **ratings** with the meaning of **television ratings**, measured in terms of average **viewing time** per household and channel (the viewing time is an implicit measurement of the users interest on a specific program). Moreover, we will use the terms **user** and **household** interchangeably.

In this paper we address the problem of predicting TV ratings based on users' viewing habits and on some characteristic of the TV programs. The raw input data to the system are:

1. the EPG (Electronic Programming Guide) containing program-id, genre, sub-genre, start-time, end-time, and TV channel for all the scheduled programs;

2. the set of viewing events containing house-id, start-time, end-time and TV channel for all the viewing events (e.g., whenever a user watched a TV channel).

In order to have a data structure easier to manage, we pre-process and simplify the raw data. We divide the time span of a day into time slots of equal duration L. We denote by $\mathcal{S}$ the set of possible time slots. In our experiments we have used time slots of $L = 60$ minutes.

We denote by $\mathcal{C}$ the set of TV channels, by $\mathcal{A}$ the attributes of a TV program (genre and sub-genre) and by $\mathcal{I}$ the set of TV programs. Additionally, we denote by $\mathcal{F}$ the set of all possible features as the union of channels, attributes and ids

$$\mathcal{F} = \mathcal{C} \cup \mathcal{A} \cup \mathcal{I}$$

We define the *television rating tensor* $r_{ufs}(t)$ as the total number of minutes user u spent watching programs with feature f in time slot s during day t.

For instance, let's assume that on day t user u watches 20 minutes of the TV program "Mickey Mouse" (feature f_1) of genre "Children" (feature f_2), sub-genre "Cartoon" (feature f_3) on "Disney Channel" (feature f_4) spanning slots s_1 (e.g., Sunday, 2.00pm–3.00pm) and s_2 (e.g., the same day, 3.00pm–4.00pm). This event increments by 20 minutes the 8 elements of the user-preference tensor corresponding to the four features $f_1 \dots f_4$, two slots s_1 and s_2 and program i_1. More formally, for each $f \in \{f_1, f_2, f_3, f_4\}$ and $s \in \{s_1, s_2\}$ we have that $r_{ufs}(t) \leftarrow r_{ufs}(t) + 20$.

Elements of $r_{ufs}(t)$ can be easily computed from the raw data (by joining EPG and tuning events) and can be used to estimate how much a user prefers to watch TV programs of specific genres, on specific channels, or during specific time slots. The user-preference tensor resembles the user-item-context tensor of traditional context-aware recommender systems [1].

The main goal is, given user u and time slot s, to estimate the number of minutes r_{us} (e.g. the rating of the user) for all the programs in time slot s.

3.1 Baseline algorithms

The dynamic models are compared with three static state-of-the art approaches, one based on individual viewing habits and two on aggregated television ratings.

CARS: we implemented the context-aware collaborative-filtering model with fuzzy seasonal context described in [12]. The model is solved by using the implicit tensor-based factorization algorithm iTALS described in [11]. We first define the *user-context-item* tensor $R = \{r_{ufs}\}$ where r_{ufs} is the total amount of minutes that user u spent watching programs with features f during time slot s

$$r_{ufs} = \sum_t r_{ufs}(t)$$

The tensor is approximated as the product of three latent factor matrices $A \in \Re^{l \times u}$,$B \in \Re^{l \times i}$ and $X \in \Re^{l \times s}$ (one for each dimension, l is the number of latent features)

$$r_{ufs} \approx \sum_l \alpha_{lu} \cdot \beta_{li} \cdot \gamma_{ls} \tag{1}$$

where $\alpha \in A$, $\beta \in B$ and $\gamma \in X$. The estimated television ratings for user u on item i in time slot s is

$$\hat{r}_{usi} = \sum_l \alpha_{lu} \cdot \beta_{li} \cdot \gamma_{ls} \tag{2}$$

STAT: a static regression model considering program-based covariates (e.g., genre, sub-genre) and time-based covariates (e.g., day of week, hour of day) described in [8].

NESTED: the nested logistic model described in [7].

4. DYNAMIC MODELS

When dealing with time series analysis, the theory of stationary stochastic processes becomes a powerful tool to analyze and predict the main dynamics of a signal. In our case, the TV ratings $r(t)$ that we observe can be seen as a realization of a stochastic process $R(t)$ assumed to be sum of three components [30]

$$R(t) = Z(t) + T(t) + S(t), \tag{3}$$

where $Z(t)$ is a *Stationary Stochastic Process* (SSP), $T(t)$ denotes a trend (e.g. $T(t) = \alpha t + \beta$ would be a linear trend) and

$S(t)$ describes a nonrandom cyclic influence like a seasonal component. The additive structure is appropriate (rather than the multiplicative one) because the seasonal variation does not vary with the level of the series [22].

The innovative feature of this representation is that $Z(t)$ can be a function of its past values, e.g. $Z(t-1)$, $Z(t-2)$, etc.

In order to predict $R(t)$, one first needs to provide the mathematical description of the three components. In particular, the identification process is divided into two steps: the identification and removal of $T(t)$ and $S(t)$, representing the deterministic parts of the process, and the identification of the model structure of the stochastic part of the process, namely $Z(t)$.

4.1 Trend and seasonality removal

The trend identification $\hat{T}(t)$, by assuming a linear trend, can be performed by minimizing via least squares the sum of residuals

$$\min_{\alpha,\beta} \sum_{i=1}^{N} \left[r(t) - (\alpha t + \beta)\right]^2 . \qquad (4)$$

where $r(t)$ are the observed ratings in $R(t)$ and N is the total number of observations. In the de-trended series $R_{ZS}(t) = R(t) - \hat{T}(t)$, the seasonal component $S(t)$ can be estimated as

$$\hat{S}(t) = \frac{1}{(N/\tau)-1} \sum_{j=1}^{(N/\tau)-1} r(t + j\tau) \qquad (5)$$

where τ is the period of the seasonality, and $(N/\tau) - 1$ is the total number of observations in a single period.
The period τ can be easily detected by analyzing the Fourier transform of the signal (in our case, a peak at a frequency corresponding to 7 days would be a weekly seasonality).

Once the seasonal part is removed with

$$R_Z(t) = R_{ZS}(t) - \hat{S}(t) = R(t) - \hat{T}(t) - \hat{S}(t) \qquad (6)$$

an estimation of $Z(t)$ can be obtained from the data $R_Z(t)$, as explained in the next section. Elements $r_Z(t)$ of $R_Z(t)$ are the television ratings once trend and seasonality have been removed.

4.2 AR model identification

In this work, Auto-Regressive (AR) models are selected for the estimator of $Z(t)$ as they are simple and linear in the parameters. In what follows, we will also show that such a model class is flexible enough to capture the main TV rating dynamics. Formally, an AR model is defined as

$$r_Z(t) = \varphi(t)'\theta + \xi(t) \qquad (7)$$

where $\theta = [\vartheta_1\ \vartheta_2\ \dots\ \vartheta_{n_\vartheta}]'$ is a vector of unknown parameters, $\varphi = [r_Z(t-1)\ \dots\ r_Z(t-n_\vartheta)]'$ is the regressor with past data and n_ϑ is the model order. In simple words, an AR model is a linear regression over the past samples of the time series at hand, after removing trend and seasonality.

The *Minimum Description Length (MDL)* criterion can be used to estimate the model order. It relies on the *"modeling by shortest data description"* principle [27] and gives consistent estimates of the order of AR models, assuming that the number of samples is much larger than the number of estimated model parameters (which is always the case in this application). The MDL formula reads as follows:

$$MDL(n_\vartheta) = (\log n_\vartheta)\frac{n_\vartheta}{N} + \log\left(\frac{1}{N}\sum_t \epsilon(t)^2\right) \qquad (8)$$

where $\epsilon(t) = r_Z(t) - \varphi(t)'\hat{\theta}$ is the error between the measured value and the model output. An alternative method is cross-validation. For this, we need to divide the time series in two parts, and use only the first one to identify the parameters while the second one assesses the model quality to establish the correct model order.

Once the model order is determined, the model parameters can be estimated according to the prediction error method [18]. Following such an approach, the best model using N data (parameterized by $\hat{\theta}_N$) is the one minimizing the output prediction error, that is

$$\hat{\theta}_N = \arg\min_\theta \frac{1}{N}\sum_{t=1}^{N}\left[r_Z(t) - \varphi(t)'\theta\right]^2 . \qquad (9)$$

The resulting n_θ-th order AR model can be rewritten as a first-order, n_θ-dimensional vector autoregression model:

$$\begin{bmatrix} r_Z(t) \\ r_Z(t-1) \\ \dots \\ r_Z(t-n_\theta-1) \end{bmatrix} = \begin{bmatrix} \vartheta_1 & \dots & \vartheta_{n_\theta} \\ 1 & 0 & 0 \\ \dots & \dots & \dots \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} r_Z(t-1) \\ r_Z(t-1) \\ \dots \\ r_Z(t-n_\theta) \end{bmatrix} + \begin{bmatrix} \xi(t) \\ 0 \\ \dots \\ 0 \end{bmatrix}$$

or, in matrix notation,

$$R_Z(t) = AR_Z(t-1) + E(t)$$

with obvious definitions for A and $E(t)$. Finally, the K-step predictor of the stochastic component Z is

$$\hat{R}_Z(t+K|t) = A^K R_Z(t),$$

whereas the total predicted TV rating reads

$$\hat{R}(t+K|t) = \hat{R}_Z(t+K|t) + \hat{T}(t+K) + \hat{S}(t+K). \qquad (10)$$

where K is the number of steps in the future (in our case, days) for the prediction of TV ratings.

4.3 ARX model identification

When some exogenous variables act on the dynamics of the process, the analysis of the time series alone is not able to describe the underlying data generation mechanism nor to accurately forecast future values of the time series. This is the case when we want to include information about which kind of program is broadcast (genre and sub-genre) or whether the considered period is of vacation or not. **Exogenous** inputs in our auto-regressive model are equivalent to **side information** in collaborative-filtering.

Exogenous inputs can be easily embedded into AR models by considering their ARX extension (where ARX stands for Auto-Regressive models with eXougenous inputs). Formally, an ARX model is defined as

$$r_Z(t) = \varphi(t)'\theta + \xi(t) \qquad (11)$$

where $\theta = [\vartheta_1\ \vartheta_2\ \dots\ \vartheta_{n_\vartheta}]'$ is a vector of unknown parameters, but $\varphi = [r_Z(t-1)\ \dots\ r_Z(t-n_\vartheta), u(t)\ \dots\ u(t-n_u)]'$ now contains both exogenous input u (i.e., genre, sub-genre,

...) and measured ratings r_Z, with n_u being the degree of the exogenous part.

In this work, we fix $n_u = 0$, since we assume only algebraic dependency between input and output. This observation comes from the nature of the system at hand: e.g. we suppose that the program broadcast at a given time does not affect the future rating. Once the additional parameter n_u is fixed, the procedure remains the same as the one given for AR models.

The resulting n_θ-th order ARX model can be rewritten as a first-order, n_θ-dimensional vector autoregression model:

$$R_Z(t) = AR_Z(t-1) + U(t) + E(t)$$

where $U(t) = [u(t), \quad 0, \ldots, 0]^T$. The corresponding K-step predictor of the stochastic component Z is

$$\hat{R}_Z(t+K|t) = A^K R_Z(t) + U(t),$$

as $U(t)$ is assumed to be perfectly known, whereas the formula for the total predicted TV rating is again (10).

5. EVALUATION AND RESULTS

5.1 Dataset

We used a dataset containing the TV viewing habits of 5,666 households over 217 channels, either cable, over-the-air, or satellite, free or pay-TV. The dataset has been collected by an independent media agency during a period of 4 months in 2013, for a total of 21 million viewing events and 56,101 EPG entries.

The EPG entries describe 21,194 distinct programs (1754 of which are TV series), classified into 7 genre categories and 117 sub-genre categories. Repeated shows as TV series or daily news are counted only once. The catalog of programs is very dynamic: on average there are 82 novel programs each day (about 0.3% of the total number of programs). The available data for the viewing events are channel, start time, and end time. All entries in the dataset (viewing events and EPG) have a one-minute resolution.

The four months of the dataset have been partitioned into three subsets: training set (first two months), validation set (one month) and test set (last month).

The dataset is available for download from `www.xxx.xxx`[2].

5.2 Model parameters

Based on the results presented in [8], we have chosen to apply the AR model (3) to each channel c and to each time slot s by defining

$$r(t) = r_{cs}(t) = \sum_{f \notin c,u} r_{ufs}(t) \qquad (12)$$

where $r_{cs}(t)$ contains the television ratings per channel and slot. Overall we created $217 \times 24 = 5208$ models (one per channel and time slot). The models can be easily trained in parallel, each one on a different partition of the dataset.

The sampling time t has been set to one day. Smaller sampling times (e.g., one hour) have been avoided because of the lead-in anchoring effect: with one-hour sampling time the model overfits the ratings of the previous hour.

Ratings (12) have been detrended by using (4) with $\alpha = 0$. Frequency analysis detected a weekly seasonality, as expected, which has been removed by using (6) with $\tau = 7$.

[2]Anonymized for blind review

For each channel and time slot we have applied the MDL formula (8) to identify the model order n_ϑ. The MDL has been computed on the validation set, once each model has been trained on the test set. Table 1 shows that most of the models are first order models.

Order n_ϑ	**Percentage**
1	75.3 %
2	10.3 %
3	5.4 %
4	2.8 %
5	1.1 %

Table 1: Orders of the auto-regressive models (only the most frequents)

We used nine exogenous inputs u as side information in the ARX model (11): seven for *genre* (“young”, “sport”, “movies”, “life-style”, “entertainment”, “news”, “undefined”), one for *holiday*, and one for *live-event*. Each of them is a binary input: 1 if the attribute is present, 0 otherwise.

The genre inputs are later normalized with respect to the number of non zero genres in the inputa.

The holiday input is set to 1 if the day is a national vacation other than Saturday or Sunday.

The live-event input is set to 1 if the program scheduled in the time slot is a live event.

5.3 Results

In this section we present the accuracy of our dynamic models compared with respect to the accuracy of the baseline algorithms.

The root mean squared error (RMSE) has been used to measure the accuracy of the different methods. As we are measuring television ratings in terms of avergae number of viewing time, we normalize the error with respect to the number of users in the sample (5,666), in order to make it independent of the population size

$$e = \frac{\sqrt{\frac{\sum_{c,s,t}\left[r_{c,s}(t) - \hat{r}_{c,s}(t)\right]^2}{N_{\text{channels}\times\text{slots}\times\text{days}}}}}{N_{\text{users}}} \qquad (13)$$

Error e is measured in terms of

$$\frac{\text{minutes}}{\text{hour} \times \text{household} \times \text{channel}}$$

With a preliminary set of experiments we compared the accuracy of the baseline methods CARS, STAT and NESTED. The STAT method, despite its simplicity, outperformed both the CARS and NESTED methods. This can be explained by considering that, when aggregating television ratings (either measured or predicted), the seasonal effect becomes stronger and context becomes the main effect.

Table 2 shows the error for the STA, AR and ARX methods when predicting television ratings at one day of distance. The first raw of the table shows the value of e for the three different models. The errors seem relatively small for all the models. But we should keep in mind that these errors have to be scaled with respect to the number of households, which are 116.3 million in the U.S. [24]. The scaled errors are shown in the second row of the table.

Figure 2 shows the improvement in accuracy with respect to the static method, as a function of the prediction hori-

Error	AR	ARX	STAT
per household and channel [min/hour]	0.8218	0,8039	0,8670
per channel [millions of min/hour]	90	88	95

Table 2: Error for predictions at 1-day distance

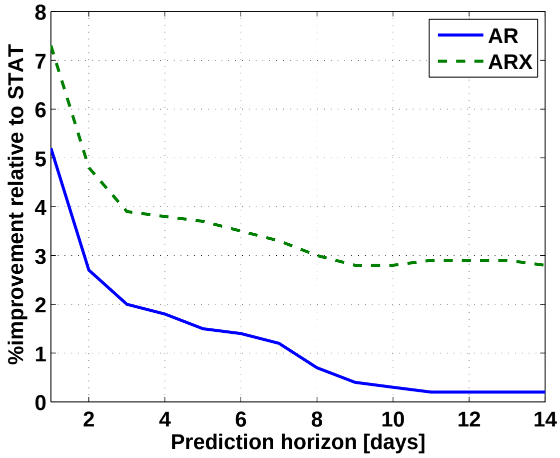

Figure 2: Improvement with respect to STAT method

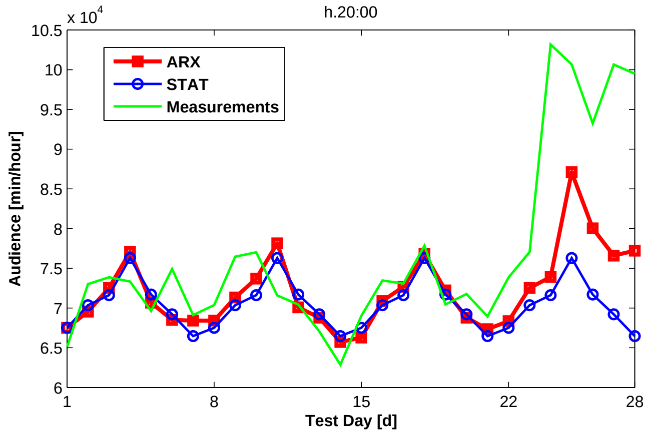

Figure 3: Predicted vs. measured television ratings, for the most popular channel, between 20:00 and 21:00.

zon. As expected, the dynamic models perform much better than the static one for short-term predictions (almost 8% improvement for the ARX model). When the prediction horizon increases, the accuracy of the simple AR models tends to the accuracy of the static model. However, the ARX model with exogenous inputs performs significantly better than the static model even for predictions up to 14 days in the future with an improvement of 3%.

Figure 3 shows the predicted ratings for the most popular TV channel in the test month, by using both the ARX and STAT methods. Predictions are computed by using the previous three months for training, and are compared with the measured ratings. From the figure we observe that during the last week the television ratings deviated from the "normal" behavior. The static model is not able to react to this change, while the dynamic model better adapts to the new trend.

6. DISCUSSION

The analysis of the results presented in the previous section suggest a number of interesting considerations: (i) models working on aggregated television ratings have a better accuracy than models working directly on viewing events, (ii) dynamic models have a better accuracy than static models for short term predictions, and (iii) dynamic models with exogenous input have better accuracy even for long-term predictions.

1. Context-aware collaborative-filtering based on tensor factorization is consistently the worst performer, even with respect to less detailed and sophisticated methods. Given the fact the the algorithm is working on detailed and non-aggregated viewing data, we did not expect this result. This may have to do with the fact that aggregated viewing present strong seasonal effects that are easy describe with even linear models.

2. Dynamic models outperform static regression models for short term predictions. This is not surprising as dynamic model are able to quickly react to lead-in effects or to other unexpected changes in the values. With long-term predictions, auto-regressive models reduce their advantages over static models as they asymptotically became equivalent to static models.

3. Exogenous inputs are able to easily capture features of television programs and to boost the dynamism of the model even on long-term predictions, when the correlation with past ratings is less strong.

7. CONCLUSIONS

With the billions of dollars spent annually on TV commercials, reliable audience predictions are required to evaluate the effectiveness of this investment.

In this paper we explore a new approach for the prediction of television audience (e.g., viewing time) based on time series analysis and auto-regressive dynamic models. The inclusion of side information as exogenous variables is able to capture the intrinsic appeal of each program based on observable attributes. The dynamic models have been proved to outperform collaborative-filtering and regression-based models in predicting television ratings for short and long term horizons.

We have validated our results on a large-scale dataset containing 21 millions TV viewing events collected from 5,666 households over 217 channels during a period of 4 months. The dataset has been released for free download.

The unexpected good results of auto-regressive models with respect to context-aware collaborative filtering techniques suggest that auto-regressive models could be efficiently applied to more traditional recommender systems domains as a valid alternative – or as an integration – to collaborative-filtering, especially in the case of time-dependent recommendations.

8. REFERENCES

[1] G. Adomavicius and A. Tuzhilin. Context-aware recommender systems. In *Recommender systems handbook*, pages 217–253. Springer, 2011.

[2] K. Ali and W. van Stam. TiVo: Making show recommendations using a distributed collaborative filtering architecture. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 394–401, New York, NY, USA, 2004. ACM.

[3] L. Ardissono, C. Gena, P. Torasso, F. Bellifemine, A. Difino, and B. Negro. User modeling and recommendation techniques for personalized electronic program guides. *Personalized Digital Television*, 6(1):3–26, 2004.

[4] P. Cotter and B. Smyth. Ptv: Intelligent personalised tv guides. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, pages 957–964. AAAI Press, 2000.

[5] P. Cremonesi, R. Pagano, S. Pasquali, and R. Turrin. TV program detection in tweets. In *Proceedings of the 11th european conference on Interactive TV and video*, pages 45–54. ACM, 2013.

[6] P. Cremonesi and R. Turrin. Time-evolution of IPTV recommender systems. In *Proceedings of the 8th international interactive conference on Interactive TV&Video*, pages 105–114. ACM, 2010.

[7] P. Danaher and T. Dagger. Using a nested logit model to forecast television ratings. *International Journal of Forecasting*, 28(3):607–622, 2012.

[8] P. J. Danaher, T. S. Dagger, and M. S. Smith. Forecasting television ratings. *International Journal of Forecasting*, 27(4):1215 – 1240, 2011.

[9] D. Das and H. ter Horst. Recommender systems for TV. In *Workshop on Recommender Systems, Proceedings of 15th AAAI Conference, Madison, Wisconsin*, pages 35–36, 1998.

[10] P. W. Farris, N. T. Bendle, P. E. Pfeifer, and D. J. Reibstein. *Marketing metrics: The definitive guide to measuring marketing performance*. Pearson Education, 2010.

[11] B. Hidasi and D. Tikk. Fast als-based tensor factorization for context-aware recommendation from implicit feedback. In *Machine Learning and Knowledge Discovery in Databases*, pages 67–82. Springer, 2012.

[12] B. Hidasi and D. Tikk. Approximate modeling of continuous context in factorization algorithms. In *Proceedings of the 4th Workshop on Context-Awareness in Retrieval and Recommendation*, pages 3–9. ACM, 2014.

[13] S. H. Hsu, M.-H. Wen, H.-C. Lin, C.-C. Lee, and C.-H. Lee. AIMED: A personalized TV recommendation system. In *Proceedings of the 5th European Conference on Interactive TV: A Shared Experience*, EuroITV'07, pages 166–174, Berlin, Heidelberg, 2007. Springer-Verlag.

[14] M. Joo, K. C. Wilbur, B. Cowgill, and Y. Zhu. Television advertising and online search. *Management Science*, 60(1):56–73, 2013.

[15] H. Katz. *The Media Handbook: A Complete Guide to Advertising Media Selection, Planning, Research, and Buying: A Complete Guide to Advertising Media Selection, Planning, Research, and Buying*. Routledge, 2014.

[16] C. M. Kelton and L. G. S. Stone. Optimal television schedules in alternative competitive environments. *European Journal of Operational Research*, 104(3):451 – 473, 1998.

[17] M. Lee and C. Johnson. *Principles of advertising: a global perspective*. Routledge, 2013.

[18] L. Ljung. *System identification: theory for the user*. Prentice-Hall., 1999.

[19] A. Martinez, J. Arias, A. Vilas, J. Garcia Duque, and M. Lopez Nores. What's on TV tonight? an efficient and effective personalized recommender system of TV programs. *IEEE Trans. on Consum. Electron.*, 55(1):286–294, Feb. 2009.

[20] K. Metaxiotis, K. Nikolopoulos, V. Assimakopoulos, and A. Patelis. FORTV: decision support system for forecasting television viewership. *Journal of Computer Information Systems*, 43(4):100–107, 2003.

[21] D. Meyer and R. J. Hyndman. The accuracy of television network rating forecasts: The effects of data aggregation and alternative models. *Model Assisted Statistics and Applications*, 1(3):147–155, 2006.

[22] D. C. Montgomery, C. L. Jennings, and M. Kulahci. *Introduction to time series analysis and forecasting*, volume 526. John Wiley & Sons, 2011.

[23] S. Murray. TV advertising forecasts. `http://www.digitaltvresearch.com/products/product?id=107`, Aug. 2014. (accessed April 2015).

[24] Nielsen. Nielsen's 2015 advance national tv household universe estimate (UE), Apr. 2015.

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

[26] S. K. Reddy, J. E. Aronson, and A. Stam. Spot: Scheduling programs optimally for television. *Management Science*, 44(1):83–102, 1998.

[27] J. Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978.

[28] J. T. Tiedge and K. J. Ksobiech. The "lead-in" strategy for prime-time tv: Does it increase the audience? *Journal of Communication*, 36(3):51–63, 1986.

[29] S. Wakamiya, R. Lee, and K. Sumiya. Towards better tv viewing rates: exploiting crowd's media life logs over twitter for tv rating. In *Proceedings of the 5th International Conference on Ubiquitous Information Management and Communication*, page 39. ACM, 2011.

[30] W. W.-S. Wei. *Time series analysis*. Addison-Wesley Redwood City, California, 1994.

[31] K. C. Wilbur. A two-sided, empirical model of television advertising and viewing markets. *Marketing Science*, 27(3):356–378, 2008.