DYNAMIC LINEAR MODELS FOR MOTION PICTURES BOX-OFFICE FORECASTING

by

RÉMI JEAN DESMEULES

B.A.A., Université Laval, 1999

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE IN BUSINESS ADMINISTRATION

in

THE FACULTY OF GRADUATE STUDIES

THE FACULTY OF COMMERCE AND BUSINESS ADMINISTRATION
DIVISION OF OPERATIONS AND LOGISTICS

We accept this thesis as conforming
to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

April 2001

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

~~Department~~ of FACULTY OF COMMERCE AND BUSINESS ADMINISTRATION

The University of British Columbia
Vancouver, Canada

Date APRIL 24

Abstract

We use a Bayesian dynamic forecasting model to predict the weekly gross box-office of motion pictures, particularly trying to replicate the "Top 10" charts for results at the box-office. We use multiple regression to estimate means of the prior distributions of the parameters of the exponential decay models used to model the revenue stream. We then use Dynamic Linear Models (DLM) to dynamically update the forecast as data is gained on the stream of revenue of the different movies. We compare the results of the DLM with those of an exponential smoothing model with trend, and the results from a "complete recalibration" method. We also use the "attraction model" to fine tune our Top 10 predictions and account for seasonality and competition.

The use of our model need not be restricted to movie box-office forecasting. Indeed, the model can be applied to many instances in the new product introduction framework, and could also be used for inventory control. We formulated the model so that it could be used as a component in an optimisation model, within the framework of MARKOV Decision Processes.

TABLE OF CONTENTS

List of Tables

List of Figures

Acknowledgements

I would like to thank Bruce Nash and collaborators at The-Numbers (http://www.the-numbers.com/) for providing the extensive data set used in this research. I am endlessly grateful to these people as they made this research possible and allowed me to save a lot of time and effort.

I would also take the opportunity to thank Martin L. Puterman and Charles B. Weinberg for sharing their extensive knowledge of the topics under study in this research. I am grateful to Professor Puterman for introducing me to the world of Dynamic Linear Models and "Bayesian updating". I am also indebted to Marty for all the suggestions, comments and support that I received through my time at the University of British Columbia.

I am also delighted to have had the chance to work with Charles B. Weinberg, in my mind the best and most successful researcher on marketing problems in the motion pictures industry. Professor Weinberg helped me understand the issues and reminded me of many important variables to consider in the modelling exercise.

Finally, I would like to thank my girlfriend Elisabeth, my family, fellow students and everyone who listened to and discussed areas of my thesis. I always appreciated receiving comments and suggestions on my work; these always enlarged my point of view on the problems at hand.

Chapter 1 - Introduction

The motion pictures industry reaches and touches almost everyone. Going to the theatre is surely one of the world's favourite pastime. That is why many facets of the industry are being studied. For instance, people can learn about the history of cinema or be taught filming or acting techniques. The intrinsic power of the industry to captivate is undeniable. Marketing researchers have found many problems and challenges to study, and have put their best efforts forward to help the different players and the public to better understand the industry.

There is a flourishing literature on the motion pictures industry in marketing journals and other related periodicals. High levels of investment from Hollywood studios and the large sums of money involved in the industry have further made it worthwhile to subject the industry to careful analysis and scrutiny.

Typically, the study of motion pictures fits into the new product introduction framework and models are usually formulated to be extendable to new products other than motion pictures. Since large quantities of data are publicly available, the motion pictures industry has often been selected to represent the category.

Many articles have focused on the forecast or explanation of the stream of gross box-office revenue at the national (aggregate) level. Sawhney and Eliashberg (1996) were interested in the adoption process and built a parsimonious model to forecast the total revenue of movies over their lifetime. Their model is interesting in that it looks at the individual level and then integrates over the population. The model is built from two consecutive exponential distributions representing the "time to decide" and the "time to act". The aggregation of the two leads to a family of distributions with three parameters that are able to represent most streams of revenue.

Other researchers (Krider and Weinberg 1998, Jedidi, Krider and Weinberg 1998, and Lehmann and Weinberg 2000) have found that a two-parameter exponential decay model

provides a good fit for most successful movies. According to Jedidi, Krider and Weinberg (1998), movies can be grouped in 4 clusters called "Hollywood Heroes", "Mega Movies", "Fast Fades" and "Fair Flicks". Swami, Puterman and Weinberg (1999) use the results to provide solutions for the movie screens management problem faced by motion pictures exhibitors. They use a MARKOV Decision Processes model to provide optimal replacement policies.

The movie screens management problem can be summarised as: "what movie to show and when to replace it with what other movie". Swami, Eliashberg and Weinberg (1999) successfully tackled the problem with their SilverScreener model. The model obtained good results after implementation (Eliashberg, Swami, Weinberg and Wierenga 2000). In SilverScreener, the optimal replacement problem was formulated using an integer-programming model, which made it straightforward to optimise with readily available algorithms and software.

Zufryden (1996, 2000) successfully tackled the problem of predicting a film's box-office performance. His work provides confirmation to the intuitive idea that production and advertising budget explain a fair part of the success of a movie. Zufryden (2000) explained about 90 percent of the variation in film box-office performance ($R^2 = 0.9$). The study used the following six attributes: number of screens, time since introduction, Cinemascore exit interview data, film website activity (new distinct point of origin or 'visitors'), production budget, and a seasonality measure. The large quantity of data needed for the model makes finding a more parsimonious approach appealing.

Eliashberg and Shugan study the impact and role of movie critics, finding that reviews correlate with late and cumulative box-office but are not significant predictors of early box-office receipts. Other problems studied in the marketing literature include (but are in no way limited to) the release timing game, the sequential release of movies and videos, the international performance of movies and consumer choice processes in the selection of movies. The release timing game (Krider and Weinberg 1998) provides an answer to

the studio's problem of "when to release movies". Given the competition, the release schedule that a studio adopts can prove very successful or very costly.

Studios also need to decide when to release the video into rental stores and other retail outlets. Lehmann and Weinberg (2000) provide an analysis of the problem and suggest that profits could be increased from an earlier release of videos. Also following the domestic (U.S. and Canada) release of movies is the international launch of movies. Neelamegham and Chintagunta provide a Bayesian model to forecast the box-office performance of movies at many different stages of its life (market evaluation, post-movie production, pre-domestic launch, pre-international lauch, etc.). Neelamegham and Jain (1999) provide an econometric model and analysis to study consumer choice processes for the selection of movies. Using latent psychological variables, they provide predictions for choice and post-choice behaviours, and thus obtain estimates for the box-office performance of movies.

Chapter 2 - Model Overview, Formulation, Properties

Overview

Background

The intent of this research is to provide a real-life application of Dynamic Bayesian Forecasting because we believe it has many advantages and properties that have not received enough attention in extant literature. The only other application of Dynamic Linear Models (DLM) was for estimating the extended warranty costs on durable goods, in which case the model was compared to two other strategies (Wasserman and Sudjianto 1996).

What we do in this research is develop a model to forecast the gross box-office revenue of individual movies during their run in theatres. Alternatively, we can view our task as one of predicting the gross box-office revenues of the Top 10 movies on the chart, as published every week in sources such as Variety (http://www.variety.com/). Although much work has been done on similar yet not exactly equivalent topics, our work is of

value because of its transportability. Indeed, we provide an application to the motion pictures industry because of our inherent interest in the topic and the fact that data is publicly available. However, applications need not be restricted to movie forecasting. The family of dynamic models presented here can be applied to almost any forecasting situation, and could also be used for inventory control, advertising effectiveness studies, and in conjunction with optimisation models such as dynamic programs based on MARKOV Decision Processes theory. We shall discuss potential applications and further research in the concluding section of this THESIS.

The motion pictures industry - players

The marketing literature approaches the release of motion pictures within the framework of new product introduction with limited lifetime and shelf space. The industry players are on the one side the exhibitors, who own the theatre screens, and on the other side are studios and distributors, who produce, release and distribute the movies. Movies are services, so the challenge of exhibitors is to maximise the occupancy of the seats at the different presentations of the movies. Outside of matinees and Tuesdays, exhibitors are limited in their price promotional ability. Therefore, exhibitors need to select the best movie to play for their audience. Selecting movies, show times or promotional spending is challenging for exhibitors, since they need to share profits according to complicated agreements (Squire, 1992) that may or may not favour them over studios and distributors.

For studios and distributors, the main challenges are evaluating promotional spending, choosing what movies to produce or distribute, and choosing the right time to release their movies (Krider and Weinberg 1998). For these major players, decisions involve very large sums of money and mistakes can prove costly. It is said that less than 70% of the movies released by Hollywood get to be profitable (Vogel 1994).

The motion pictures industry - Seasonality

The motion pictures industry is characterised by a strong seasonal cycle. Indeed, as can be seen in Figure 1, there is a high season for the winter holidays (December-January)

and the summer months from June to mid-August. Further, there is a fairly "stable" low season around March to May, and September to November. Three-day weekends and U.S. holidays such as Martin Luther King's day, Presidents' day, Memorial Day, Independence Day, Labour Day and Thanksgiving further influence results, since these are usually occasions for gatherings as well as movie-going, and since more people are willing to go see a movie, studios can decide to release their top movies on those dates, if not during high season. (Readers should note that the domestic box-office is composed of revenues in the United States and Canada, Canada being quite a small portion of the total revenues.)

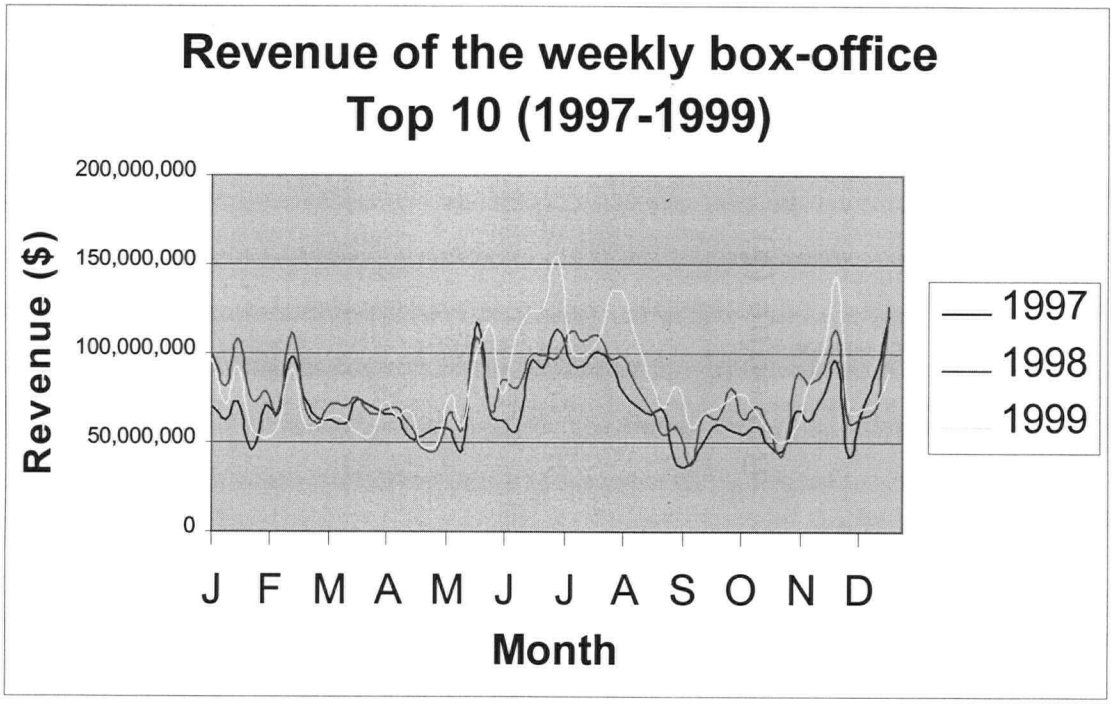


Figure 1: Seasonality of the box-office top 10 revenue (1997-1999)

Therefore, the parties interested in making gross revenue predictions should try to include seasonality effects in their model. Obviously, there will be the added task of doing predictions on the cumulative box-office, possibly on top of individual movie predictions. (Appendix 1 presents Figure 1 as a time series.)

Still, studios and distributors enhance this seasonality by releasing their best products while it is high season. However, consumers in the end decide whether or not to make it

"high season" or not. That is, Hollywood has been criticised at times for providing products (movies) that somehow lacked in "quality". We believe it is for this reason that there is a surge in the appeal of lower budget production and the "film festival" movie type. But for now we only need to recognise that a certain "quality" level must be maintained for the historical seasonality to hold into the future. That is, we believe that seasonality is based on the number of people willing to see movies *and* the quality of movies offered for people to see.

The motion pictures industry - Competition

In order to obtain an accurate forecast of the gross box-office of a movie at some point in its run, one could take advantage from the knowledge of the opposition the movie will face. Indeed, movies released along with huge blockbusters of the likes of Star Wars or Jurassic Park will fare particularly worse than if the competing features were less attractive. Blockbusters do tend to capture a large share of the market for their release weekend, but perhaps some even have the capacity of "creating" some market for themselves, on top of the projected demand for the weekend.

Modelling alternatives

Although we will revisit alternative models and formulations in a later chapter, we should first comment on why we discarded some of the initiatives that we did not choose to use.

Family of distribution

First, in creating a dynamic Bayesian forecasting model we need to select a family of distributions that will represent the stream of revenue of the movies. We need to remember that movies have a short life span of about 10 weeks in theatres (second run theatres usually take over after 10-13 weeks) and even less on the top spots on the charts (the most successful movies are out of the top 10 in 6 weeks). Furthermore, most of the revenue of movies is gathered during the first few weeks after release.

Therefore, we argue, the best choice of a family of distribution is one with very few parameters (most likely two) so that forecasting error does not accumulate while the model's parameters are still being initialised. Given that songs stay longer on the Billboard charts than movies on the box-office charts, Bradlow and Fader (2000) were able to use the "generalized gamma (GG) lifetime curve with double exponentially (DE) distributed errors" in their lifetime model for the "Hot 100" Billboard songs. Their formulation allowed for 5 parameters, which seems reasonable for songs, since not all of them begin on top spot or at their highest reached rank. In their data set of 248 songs, only 15% attained their top ranking in their debut week. However, movies decay a lot more rapidly and usually reap their highest gross at opening week. (In terms of revenues, about 95% of movies had their top-grossing weekend in their release weekend).

For movies, Sawhney and Eliashberg (1996) use a more parsimonious model that requires three parameters. Although they obtain predictions on the three parameters, they fit their model after 3 weeks of data. Their model allows for more shapes of the stream of revenue, such as two-parameter exponential, Erlang-2 probability distribution and three-parameter Gamma distribution.

Our model uses only one explanatory variable, namely *Time*, but modelling need not be restricted to simple linear regression. Thus, the use of dynamic multiple regression is yet another alternative in the framework of dynamic Bayesian forecasting. A certain number of explanatory (independent) variables must be adopted and their progress followed. Multiple regression DLMs will update the coefficients of each explanatory variable and provide a forecast that takes each variable into account. One might decide to include a seasonality "index" as an independent variable, in addition to another variable such as the time since release. Still, the model has to go through an initialisation period that includes ***at least*** two periods, one for a constant and another for an independent variable. The challenge here is also to make sure that all independent variables are meaningful at every period, and that the explanatory relationship is linear.

We believe that our modelling framework is flexible enough to account for many shapes of curves that streams of revenues may follow (Erlang-2, Gamma, etc.). That is, our

forecasts may be accurate even if movies do not follow an exponential decay pattern of revenues. We rely on the learning capability of dynamic models to provide a good fit to other curves.

An example of stream of revenue

For clarity, we now present what is meant by "stream of revenue" and "gross box-office revenues during a movie's theatre run". We choose to present the movie ***Erin Brockovich*** because it will also serve to point out the effects of seasonality and competition. The stream of revenue for the movie can be found in the third column of Table 1 (Figure 3 also provides a graphical view of the series). This data is our dependent variable and the series that we are trying to forecast. The number figure represents the revenue from the Friday, Saturday and Sunday only, as opposed to the whole week.

Wk #	Date	Gross	% chg.	Screens	Per Screen	Top 10 cumulative $	Index	Notable openings
1	3/17/00	*$28;138,465*		2,848	$9,880	*$71,227,498*	*90.9*	Erin Brockovich
2	3/24/00	*$18,545,755*	-34%	2,851	$6,505	*$70,001,108*	*89.4*	Romeo must die
3	3/31/00	*$13,798,460*	-25%	2,903	$4,753	*$72,074,699*	*92.0*	Road to El Dorado
4	4/7/00	*$9,808,065*	-29%	3,265	$3,004	*$69,411,543*	*88.6*	Rules of Engagement
5	4/14/00	*$7,030,315*	-28%	3,070	$2,290	*$62,237,328*	*79.5*	28 Days
6	4/21/00	*$5,500,790*	-22%	3,056	$1,800	*$70,397,639*	*89.9*	U-571
7	4/28/00	*$3,622,105*	-34%	2,498	$1,450	*$64,213,479*	*82.0*	Flintstones: VRVegas
8	**5/5/00**	*$2,184,770*	**-40%**	**1,943**	**$1,124**	*$73,974,852*	*94.4*	**Gladiator**
9	5/12/00	*$1,722,120*	-21%	1,491	$1,155	*$67,300,737*	*85.9*	Battlefield Earth
10	5/19/00	*$1,104,330*	-36%	952	$1,160	*$99,438,784*	*126.9*	Dinosaur
11	**5/26/00**	*$1,057,355*	**-4%**	**798**	**$1,325**	*$168,582,598*	*215.2*	**Mission: Impossible 2**
12	6/2/00	*$615,395*	-42%	737	$835	*$95,077,357*	*121.4*	Big Momma's house
13	6/9/00	*$527,260*	-14%	643	$820	*$90,433,502*	*115.5*	Gone in 60 seconds
14	6/16/00	*$332,235*	-37%	485	$685	*$94,155,016*	*120.2*	Shaft
15	6/23/00	*$269,205*	-19%	411	$655	*$94,088,127*	*120.1*	Me, Myself & Irene

Table 1: Box-office results for Erin Brockovich (March 17th to June 23rd)

Since people mostly go to theatres during the weekend (except maybe for Tuesdays and holidays) these three days represent a substantial part of the revenue for the week. The period from Monday to Thursday might add about a third (up to about a half later on) of the weekend revenue to the cumulative gross of the movie (75% weekend, 25% rest of the week). We also present the weekend date, the percentage change from the preceding

week, the number of screens the movie played on, the revenue per screen, the cumulative box-office revenue of the top 10 grossing movies the same weekend, the index of the week strength (100 being the mean of the top 10 in the data set), and the most notable movie opening that weekend.

One could do many analyses from Table 1 but we pinpoint a selection that illustrates what this section was arguing about up to this point. First, one can note how quickly the revenues decline over the first few weekends. Up to week 7, the movie seems to be losing a steady 30% of its revenues every weekend. That observation is supported by a steady season (cumulative box-office of the top 10 movies) and the fact that the new movies opening during these weeks wrapped up combined grosses of about 20-30 million dollars. In week 8, however, the blockbuster ***Gladiator*** was released and gathered 47% of the cumulative box-office for that weekend, with 35 million dollars. However, holidays weren't really a factor that weekend, so other movies took a loss, which explains why the movie ***Erin Brockovich*** fell by 40% instead of its average 30% up to that weekend. Then, in week 11, the movie drops by a mere 4%, because of a holiday. Indeed, the 11^{th} weekend in the movie's run happened to be the ***Memorial Day*** long weekend. More people were willing to see a movie and could now go on Sunday night. The boost in total revenue was helped by the release of the blockbuster ***Mission: Impossible 2***. The revenue of the movie fell abruptly on the 12^{th} weekend since people went back to a regular working week.

Of course, not all streams of revenues behave in the way that the one for ***Erin Brockovich*** does. Indeed, some movies can be called "sleepers", meaning that their highest revenue is not in the release weekend (e.g. *The Tigger Movie*). Other streams of revenue might call for irregular "bumps" that are explained by some external factor. Figure 2 presents some of the cases that deviate from the linear pattern on the log scale. These challenges make the forecasting assignment more difficult and warrant the need for a flexible and somehow "intelligent" model, perhaps one with some "learning" capability.

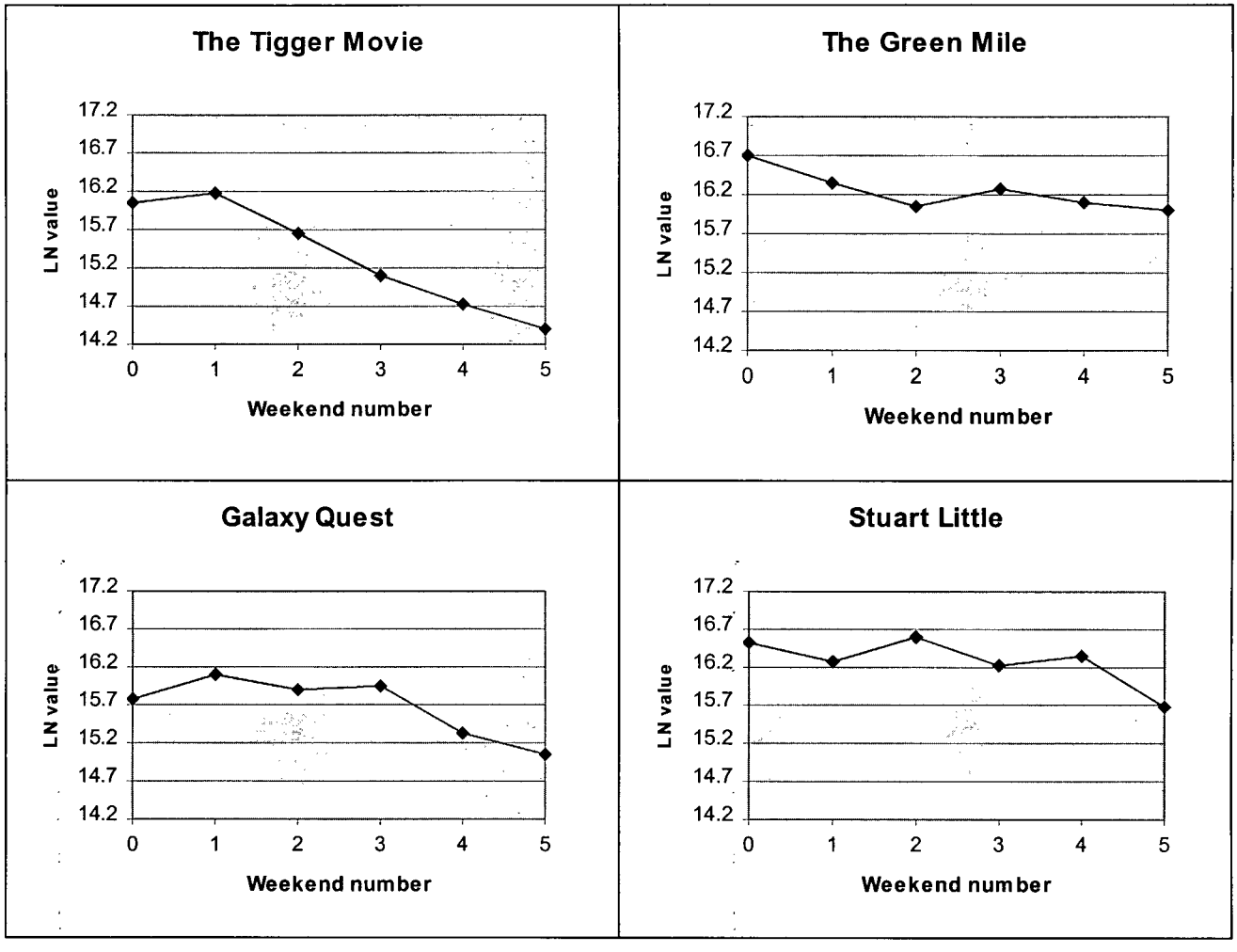


Figure 2: Challenging streams of revenues for forecasting

An example of the Top 10

A reason why one would obtain predictions for movies' gross box-office is to determine the top movies on the charts. Top positions on the charts are important because they represent good selling arguments, which are used in pieces of advertisement. Indeed, phrases like "the most popular movie in America" or "the number one comedy in America" are often used by Hollywood to convince people of choosing their movie. (One could note that these phrases would be mostly used in the absence of critics' rave reviews or in conjunction with them…)

#	Week-end of:	3/17/00	Age
1	Erin Brockovich	$28,138,465	1
2	Mission to Mars	$11,385,709	2
3	Final Destination	$10,015,822	1
4	My Dog Skip	$5,268,241	3
5	Ninth Gate, The	$3,531,618	2
6	Whole Nine Yards	$3,274,453	5
7	American Beauty	$3,159,322	25
8	Cider House Rules	$2,477,000	11
9	Snow Day	$2,205,015	6
10	Tigger Movie	$1,771,853	6
	TOP 10 TOTAL :	$71,227,498	

#	Week-end of:	3/24/00	Age
1	Erin Brockovich	$18,545,755	2
2	Romeo Must Die	$18,014,503	1
3	Final Destination	$7,218,840	2
4	Mission to Mars	$5,720,133	3
5	Here on Earth	$4,510,705	1
6	Whatever it Takes	$4,104,298	1
7	American Beauty	$4,024,983	26
8	My Dog Skip	$3,058,430	4
9	Cider House Rules	$2,799,556	12
10	Whole Nine Yards	$2,003,905	6
	TOP 10 TOTAL :	$70,001,108	

#	Week-end of:	3/31/00	Age
1	Erin Brockovich	$13,798,460	3
2	Road to El Dorado	$12,846,652	1
3	The Skulls	$11,034,885	1
4	Romeo Must Die	$9,378,376	2
5	High Fidelity	$6,429,107	1
6	American Beauty	$5,407,279	27
7	Final Destination	$5,355,284	3
8	Mission to Mars	$3,324,853	4
9	Here on Earth	$2,262,462	2
10	Whatever it Takes	$2,237,341	2
	TOP 10 TOTAL :	$72,074,699	

#	Week-end of:	4/7/00	Age
1	Rules of Engagmnt	$15,011,181	1
2	Erin Brockovich	$9,808,065	4
3	Road to El Dorado	$9,085,803	2
4	Return to Me	$7,820,836	1
5	The Skulls	$6,450,720	2
6	Ready to Rumble	$5,257,778	1
7	Romeo Must Die	$4,552,754	3
8	High Fidelity	$4,241,028	2
9	Final Destination	$3,835,071	4
10	American Beauty	$3,348,307	28
	TOP 10 TOTAL :	$69,411,543	

#	Week-end of:	4/14/00	Age
1	Rules of Engagmnt	$10,933,627	2
2	28 Days	$10,310,672	1
3	Keeping the Faith	$8,078,671	1
4	Erin Brockovich	$7,030,315	5
5	Road to El Dorado	$6,156,329	3
6	Return to Me	$5,008,744	2
7	American Psycho	$4,961,015	1
8	The Skulls	$4,023,025	3
9	Final Destination	$3,049,212	5
10	Ready to Rumble	$2,685,718	2
	TOP 10 TOTAL :	$62,237,328	

#	Week-end of:	4/21/00	Age
1	U-571	$19,553,310	1
2	Love and Basketball	$8,139,180	1
3	Rules of Engagmnt	$8,007,551	3
4	28 Days	$7,301,753	2
5	Keeping the Faith	$7,233,699	2
6	Erin Brockovich	$5,500,790	6
7	Road to El Dorado	$5,225,727	4
8	Return to Me	$3,961,664	3
9	Final Destination	$2,761,900	6
10	The Skulls	$2,712,065	4
	TOP 10 TOTAL :	$70,397,639	

Table 2: Top 10 charts (3/17/2000 to 4/21/00)

We see some interesting facts from the study of the top 10 during the first 6 weeks of ***Erin Brockovich***'s (***E.B.***) run in theatres. The movie kept the first spot for three weeks before it had to relinquish it to the hands of the movie ***Rules of Engagement***. It's interesting to see that ***Romeo must die*** almost beat ***E.B.*** but had to settle for the number 2 spot. The effects of competition were to create quite a clash between ***E.B.***, ***The Road to El Dorado*** and ***The Skulls***. Had a single one of the two newcomers been released instead of both, the remaining one would have probably taken a comfortable lead on top of the chart. But releasing movies at the right time is serious business, which has been studied by Krider and Weinberg (1998).

Worth noting is the fact that movies that do not make it to the top three spots on the chart at their opening week are quite brutally thrown out of the top 10 in no more than 3 weeks. Examples of this are the movies ***Here on Earth***, ***Whatever it takes***, and ***High Fidelity***, which appear on the top 10 in 2 consecutive weeks. Opening at the seventh spot is the movie ***American Psycho***, which doesn't even make it to a second week on the top 10. Perhaps the study of competition on the charts is more important than it seems.

To study the top 10, we will use the share attraction framework suggested in Krider and Weinberg (1998). We shall present the formulation of the framework in a later section but for now we do mention that the model needs inputs of total demand and "attraction" powers, which need to be provided by a forecasting model that we need now describe.

Formulation

Model overview

The model we use to obtain our forecasts is the *Univariate Dynamic Linear Model* (DLM) (West and Harrison 1997). This family of models offers the advantages of Bayesian updating. That is, the model is based on conditional expectations and features dynamic updating of forecasts as data is acquired. Perhaps the model's greatest advantage is simplicity. Indeed, the model is quite parsimonious in that it uses only two parameters, which are inherently easy to conceptualise. Furthermore, the model uses closed-form solutions, which adds to ease of use and implementation, especially with spreadsheets.

The fundamental assumption in our modelling efforts is that the stream of revenues of movies follows a two-parameter exponential model (exponential decay model) of the following form:

$$Y_t = \alpha e^{-\beta t} e^{\varepsilon} \qquad (1)$$

As we suggested, the parameters α and β of the distribution represent opening and decay factors. Therefore, one could describe a movie's revenue by finding what will be its opening week revenues (α), and then assessing by what percentage it will decline weekly thereafter (β).

Log-scale transformation

Working on a Bayesian forecasting model with the exponential distribution can become a arduous task and warrant the use of numerical estimation techniques, such as MARKOV Chain Monte Carlo methods (MCMC). Since closed-form solutions for linear models are

available in extant literature, we use a log transformation to obtain a corresponding linear formulation.

Re-writing (1):

$$Y_t = e^{\alpha} e^{-\beta t} e^{\varepsilon} = e^{\alpha + \beta t + \varepsilon}$$
$$Ln\ Y_t = Ln\ [\ e^{\alpha + \beta t + \varepsilon}\ ]$$

$$Ln\ Y_t = \alpha + \beta t + \varepsilon \qquad (2)$$

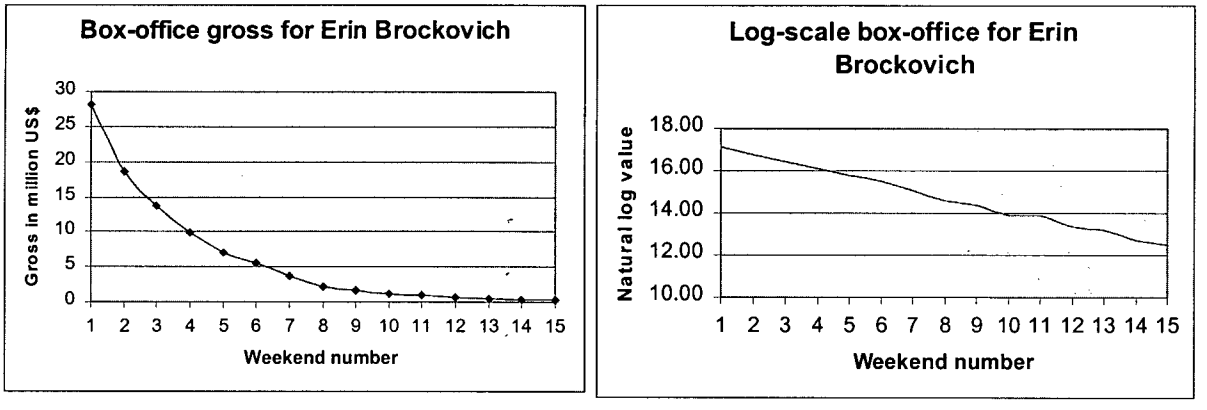


Figure 3: Scatter plot of the gross box-office revenue in both dollar and log scale

Therefore, our data points (Y_t) will be the log-transformed stream of revenue. Figure 3 presents the dollar scale and log-scale graphs of the stream of revenue for the movie ***Erin Brockovich***. From the graphs, both the exponential and the linear relationships are quite evident. The linear formulation can now fit the requirements of the general univariate DLM. We begin our exposition with some useful notation. We use D_t to represent the information set available at time t. The information set contains elements such as movie attributes, and is further enriched with data. Specifically, the initial information set (movie attributes, such as genre and number of screens) is denoted D_0. The vector of parameters will be denoted θ_t (where $\theta_t = (\alpha_t , \beta_t)$). In general, boldface notation will represent vectors or matrices. Conditional probabilities, such as the one for parameters given the initial information set will be denoted $(\theta_t \mid D_0)$. These conditional probabilities will mainly follow a Normal distribution with mean μ and variance σ^2, as in $(\theta_t \mid D_0) \sim N[\mu, \sigma^2]$.

The univariate DLM

Our formulation of the univariate DLM will use the quadruple {**F, G,** V, **W**}. Since the set is not time-indexed, we call this formulation a *constant DLM*. The matrices (**F, G**) are used as design matrices for our modelling task, and variance scalar V and matrix **W** (V, **W**) represent the variability in the conditional probabilities involving forecasts and distribution parameters. The quadruple will be used to relate Y_t and θ_t at time t in the following way:

$$(Y_t \mid \theta_t) \sim \mathrm{N}[\, \mathbf{F}'\theta_t, \mathrm{V}\,] \qquad (3)$$
$$(\theta_t \mid \theta_{t-1}) \sim \mathrm{N}[\, \mathbf{G}\theta_{t-1}, \mathbf{W}\,] \qquad (4)$$

Readers should note that D_{t-1} should be present in the conditioning of equations (3-4), but we shall keep it implicit for the sake of notational simplicity. Specifically, the relationships expressed in (3)-(4) can be represented as follows:

$$Y_t = \mathbf{F}'\theta_t + \nu_t, \qquad \nu_t \sim \mathrm{N}[\, 0, \mathrm{V}\,] \qquad (5)$$
$$\theta_t = \mathbf{G}\theta_{t-1} + \varpi_t, \qquad \varpi_t \sim \mathrm{N}[\, \mathbf{0}, \mathbf{W}\,] \qquad (6)$$

We only get one data point for the revenue each week, so Y_t is a scalar. In equations (3)-(6), **F** is called the design matrix of known values of independent variables. Therefore, **F** is a column vector multiplying a column vector of parameters θ_t. Because we wanted to keep both parameters positive in sign, we assign the values ***(1, -t)*** to the column vector **F**. Doing so, we obtain a forecast that is based on a linear model of the form specified in (2) (omitting error terms):

$$F'\theta_t = [1, -t] \times \begin{bmatrix} \alpha_t \\ \beta_t \end{bmatrix} = \alpha_t - \beta_t t$$

We could assume that the stream of revenues follows a stationary distribution. That would mean that we believe that a set of two parameters describes the stream of revenues of the movie (almost perfectly). That is, if we could see into the future, we would use the parameters that best fit the curve at the end of a movie's run. To represent this idea, we

would assign values of 0 to each cells of the variance matrix **W**. We would thus obtain a "static regression" model.

However, we believe that we could profit from adding some flexibility to the model. That is, we assume that the streams of revenue are only locally constant, and that some flexibility could help capture some "local trends". Hence, we use non-zero values for W, and update the parameters according to a random walk. The matrix **G**, called the evolution (also system, transfer or state) matrix, need not affect the previously obtained parameters (θ_{t-1}). Hence **G** is a 2x2 identity matrix (I_2). Therefore, the estimates and values of alpha (α) and beta (β) are allowed to change in time.

Thus, we have the following:

$$Y_t = \alpha_t + \beta_t t + \nu_t \qquad \nu_t \sim N[0, V]$$

$$\begin{bmatrix} \alpha_t \\ \beta_t \end{bmatrix} = \begin{bmatrix} \alpha_{t-1} \\ \beta_{t-1} \end{bmatrix} + \begin{bmatrix} \omega_t^{\alpha} \\ \omega_t^{\beta} \end{bmatrix} \qquad \varpi_t \sim N[\,\mathbf{0}\,, \mathbf{W}\,]$$

Constant DLM

Our formulation of the constant DLM is defined by:

Observation equation:	$Y_t = \mathbf{F}'\theta_t + \nu_t\,,$	$\nu_t \sim N[\,0\,, V\,]$	(5)
System equation:	$\theta_t = \mathbf{G}\theta_{t-1} + \varpi_t\,,$	$\varpi_t \sim N[\,\mathbf{0}\,, \mathbf{W}\,]$	(6)
Initial information:	$(\theta_0 \mid D_0) \sim N[\mathbf{m_0}\,, \mathbf{C_0}]$		(7)

In equation (7), the prior is specified using moments $\mathbf{m_0}$ and $\mathbf{C_0}$. We could mention that the information set at time 0 (D_0), used to obtain the prior mean $\mathbf{m_0}$ and variance $\mathbf{C_0}$, is based on historical data found in a database. We will present our initial database in detail in the following chapter.

Updating equations

Our formulation of the constant univariate DLM is closed to external information at times $t \geq 1$. Thus, any information acquired at any future time t is only the observed revenue

(data) stream Y_t. That is, $D_t = \{Y_t, D_{t-1}\}$ and the model has the MARKOV property. Indeed, at any time the existing information about the system is represented and sufficiently summarised by the posterior distribution for the current state vector θ_t.

The system is initialised with some vector of parameters θ_0, following a Normal distribution with mean $\mathbf{m_0}$ and variance $\mathbf{C_0}$. More generally, the system evolves from posterior, to prior, one-step forecast, and posterior at time t as follows:

Posterior at time t-1: $(\theta_{t-1} \mid D_{t-1}) \sim N[\mathbf{m}_{t-1}, \mathbf{C}_{t-1}]$ (8)

Prior at time t: $(\theta_t \mid D_{t-1}) \sim N[\mathbf{a}_t, \mathbf{R}_t]$ (9)

Where $\mathbf{a}_t = \mathbf{G}\mathbf{m}_{t-1}$ and $\mathbf{R}_t = \mathbf{G}\mathbf{C}_{t-1}\mathbf{G}' + \mathbf{W}$

One-step forecast: $(Y_t \mid \mathbf{D}_{t-1}) \sim N[f_t, Q_t]$ (10)

Where $f_t = \mathbf{F}'\mathbf{a}_t$ and $Q_t = \mathbf{F}'\mathbf{R}_t\mathbf{F} + V$

Posterior at time t: $(\theta_t \mid D_t) \sim N[\mathbf{m}_t, \mathbf{C}_t]$ (11)

With $\mathbf{m}_t = \mathbf{a}_t + \mathbf{A}_t\mathbf{e}_t$ and $\mathbf{C}_t = \mathbf{R}_t - \mathbf{A}_t Q_t \mathbf{A}'_t$
Where $\mathbf{A}_t = \mathbf{R}_t\mathbf{F}Q_t^{-1}$ and $e_t = Y_t - f_t$

Note that the proof for these updating equations is obtained by the use of the multivariate normal distribution theory. The derivation can be found in West and Harrison (1997).

Discussion

Many extensions to our model are discussed in West and Harrison (1997). However, we keep our formulation very simple because we believe that it could potentially become a forecasting component to an optimisation model based on MARKOV Decision Processes theory. That is, the model is state-based and MARKOVIAN, so optimality equations and probability matrices could potentially be based on the updating equations found in (8-11).

One such application of this kind of optimisation models is to creation of a dynamic decision-making tool for the movie screens management problem faced by motion pictures exhibitors. A deterministic model solving the problem has already been

developed and implemented by Swami, Eliashberg and Weinberg (1999). Owners of movie theatres and multiplexes face the movie screens management problem every week. The decision to be made is which movie to play, and when to replace it. Additionally, the problem can be extended to which screen to play it on, when screen capacity is varying.

Implementation

Since all results of the model are obtained via closed-form solutions, the model can be easily implemented on spreadsheets, or else written in most programming languages.

Chapter 3 - Data Set, Initialisation (parameters)

Database

The (Microsoft Access) database from which we drew our data set was created and compiled by Bruce Nash and collaborators at the Internet web site http://www.the-numbers.com/. It features 18091 entries of weekly box-office data mostly gathered from 1997 to August 2000 (some entries are for earlier movies). Listed are the date, film name, weekend gross, percentage change from the previous week, number of screens, cumulative gross, date of release, rank, previous rank and revenue per screen. The same information is available in 15016 entries of daily box-office figures.

The database also includes numerical data on the number of movies and the corresponding average and total box-office of 697 stars (actors and actresses). Additionally, the database lists the rating of each Star on the *Hollywood Stock Exchange* Internet game (http://www.hsx.com/). The same information is also available for 123 directors: number of movies, average gross, total gross and HSX rating. For movies, the database includes specific data on the studio, date of release and total box-office receipts, in addition to the data that can be obtained from the weekly box-office table.

Data preparation

Because our interest was mainly on the top grossing movies, we restricted our attention to movies that made their way to at least the 5th spot on the box-office rankings for at least one weekend. Jedidi, Krider and Weinberg (1998) used a similar procedure in their

clustering effort. We kept all movies that ran completely from 1997 to August 2000. We had to focus on box-office successes in order to get a more homogeneous set. Additionally, the marketing literature on movies (Sawhney and Elisashberg 1996, Jedidi et al. 1998) suggests that only those movies which do well on the box-office charts have streams of revenue that can be well represented by the exponential decay model (eq. 1). Only some rare "sleepers" make their highest grossing week past their wide-release date (controlling for seasonality).

As a result, our database comprised of 319 movies. Interestingly enough, the yearly mean of the first week of box-office for these movies did not differ too much so as to suggest a control for inflation. A one-way ANOVA for the interaction of year on the first weekend box-office is not significant *($F=.951$, $p=.416$)*. Still, there seems to be a trend, but our analysis will focus on a single year (2000). Readers should note that there has been a decrease in the number of consumers in the motion picture industry, but increases in ticket prices have accounted for it.

	Year	N	Mean	Median	Std. Deviation
Week 1 Box-Office	1997	87	$14,194,607	$11,660,216	$10,774,763
	1998	91	$14,671,155	$11,633,495	$9,181,250
	1999	90	$16,381,849	$13,031,125	$12,323,326
	2000	51	$16,728,438	$12,846,652	$12,378,280
	Total	**319**	**$15,352,736**	-	**$11,088,355**

Table 3: Descriptive statistics for Week 1 Box-Office (1997-2000)

For each of one of these 319 movies, we collected the first 6 weeks of box-office revenue, making sure that the first week was the movie's wide-release weekend. Some movies, such as "Disney" movies are released on a small number of screens some number of weeks prior to their wide release (1 week ahead, "Disney" movies were released on 2 screens). We decided to cut at 6 weeks because many movies do not really make it past that mark. Most movies are out of the Top 10 before the seventh week and some problems can also arise later on with the addition of second run theatres adding to the revenue of the movies. Movies seem to be released in second-run theatres after their tenth week.

In addition to these 6 weeks of data, we collected data on the movies' stars and directors. Since clear relationships were not established between movies and their stars, the first part of the task was to reconcile movies and stars. We used a spreadsheet to obtain a table listing movies and their two biggest stars. We then reconciled the stars' HSX rating, average gross and total gross of the movies they were involved in. When available, we got the data on the movies' director. We kept data on the director's number of movies, average gross and total gross, then calculated the average of the first weekend's revenue of their last 3 movies.

In order to capitalise on existing research, we also included movie attributes that had been successfully used to predict movie gross box-office. We decided to also collect data on the genre and MPAA (*Motion Pictures Association of America*) rating, and also created a dummy variable assigning a value of 1 to sequels, and 0 otherwise. For the genre, we used the first qualifier used for the movie at the Internet Movie Database web site (http://www.imdb.com/). The genres coded into dummy variables were action, comedy, drama, thriller, horror, romance, sci-fi, mystery and children. The MPAA ratings, namely G, PG, PG-13, R and NC-17 (no movie in the data set was rated NC-17), were easily accessible from the organisation's web site (http://www.mpaa.org/).

From the database, we also calculated the cumulative gross of the Top 10 movies every week from January 3, 1997 to August 11, 2000. This was done to study the seasonality of box-office revenues and also to de-seasonalise data (express revenue in terms of share of the box-office). Such a table is available at the box-office "guru's" web site (http://www.boxofficeguru.com/). Readers should note that the Top 10 movies on the chart gather about 75-80% of the total weekly box-office for the industry (the Top 15 make for about 85-90% of the total for the industry in a given week).

Data set

In the end, the data set contained 319 entries of movie name, release date, wide-release date, highest rank on the charts, week 1-6 of dollar revenue, week 1-6 log-transformed

(ln) revenue, release week strength, week 1-6 of de-seasonalised data (divided by week strength), log-transformed de-seasonalised data, information on directors and stars, studio (dummy variables), MPAA rating (dummy variables), genre (dummy variables), and the number of screens in weekend 1 and 2.

Initialisation

The reason why this data set was constructed was to provide a first week forecast of the two parameters of the exponential decay model. Our data set can be viewed as D_0, the initial data set which parameters are conditioned upon at time 0. Since the most important weeks of a movie's life are the first few, it is important to obtain a good prediction of the two parameters. Indeed, the prediction in the first week is based entirely on the alpha (α) derived from the initial data set (D_0) and the prediction in the second week is largely based on the forecast of the beta (β) parameter obtained from D_0. Therefore, the success of the forecasting task is dependent on the accuracy of the opening (week 1) box-office forecast ***AND*** on the forecast of the "percentage" loss from week 1 to week 2.

Therefore, we needed to develop a tool that would provide accurate forecasts of the box-office in week 1 and the slope of the loss in revenue from week 1 to week 2. It should be noted here that we choose to predict the loss from week 1 to week 2 instead of the beta (β) parameter itself (the "average" over the whole stream of revenue) because of the capital importance of the first two weeks of revenue in terms of the total revenues of a movie.

Regression analyses

To obtain initial point forecasts and variances for the two parameters of the exponential decay model, we used multiple linear regression. This method tries to fit a "line" (at least in two-dimensional space) or a plane through points by minimising the sum of squares of the distance between the curve and the different data points. Using a scatter plot, one can readily see what the method achieves.

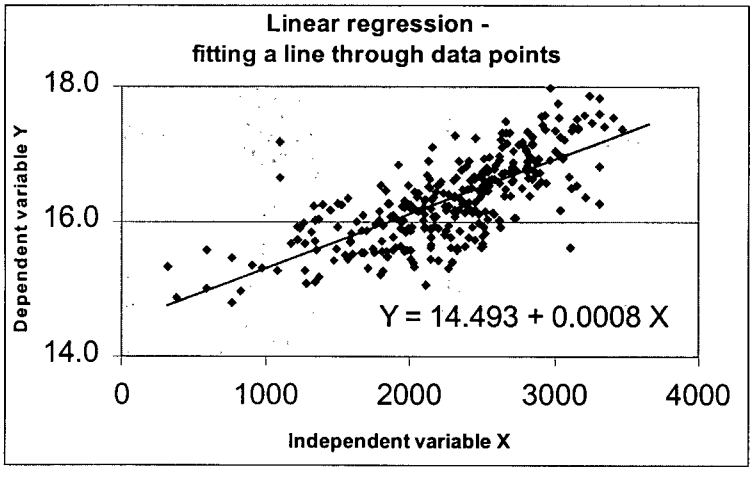


Figure 4: Regression in two-dimensional space

In Figure 4, an independent variable ***X*** is used to predict the value of a dependent variable ***Y***. The equation is of the form $Y = b_0 + b_1(X)$, where b_0 is a constant, which is the intercept of the curve at $X=0$. The other parameter, b_1, is called the *coefficient* of the variable ***X***. A positive coefficient means larger values of ***X*** lead to larger values of ***Y***, a negative coefficient to ***X*** leads to smaller ***Y***s as ***X*** increases. To obtain a prediction, one would add the constant, 14.493, to the result of the value of ***X*** times its coefficient (***0.0008***). For example, if $X = 1000$, then we think that the value of ***Y*** will be ($14.493 + 0.0008\ (1000) =$) ***15.293***.

We tested many attributes (independent variables) for their significance in predicting the two parameters but would have liked to have more data on some attributes that are intuitively appealing and have been confirmed in the extant literature to be good predictors of box-office performance. Those are, for instance, advertising budget, production budget, film web site activity, grades from exit interviews, etc.

Unfortunately, we had to work with publicly available data. However, we have to mention that players in the motion pictures industry have access to many sources of data. Extant research has obtained excellent results by using advertising spending (Zufryden 1996). A study by Zufryden (2000), the most comprehensive study in terms of the number and quality of the sources of data, has succeeded in explaining about 90 percent

of the variation in film box-office performance (R-square is 0.9). The study used the following six attributes: number of screens, time since introduction, Cinemascore exit interview data, film website activity (new distinct point of origin or 'visitors'), production budget, and a seasonality measure.

That said, the movie attributes that were available for our use are the following: star power, director power, number of screens, genre, MPAA rating, sequel and studio. To make sure that our regression analysis represents true relationships, we hypothesise on the direction of the relationship (sign of the coefficient) between our independent and dependent variables.

Effects on α: the opening revenue

Star Power

The term "star power" refers to the "quality" of actors and actresses or their effectiveness in attracting and getting people to see (and enjoy) their movies. This is intrinsically measured in the salary that actors receive for participating in movies. For example, stars like Tom Cruise, Mel Gibson, Julia Roberts or Harrison Ford attract the masses to their movies and command large salaries (some now ask for more than US$ 20 million for a single movie). Unfortunately, we don't have access to data on the salaries of most actors. However, two measures of star power were available from our database. We collected the data on the 2 biggest stars of the different movies, and double-checked by searching the Internet Movie Database (http://www.imdb.com/). Some studies have used a dummy variable for the presence of big stars, while others have added the total "star value" of the entire cast of the movie. We hypothesised that people could identify the 2 biggest stars of a movie and somehow evaluate their interest based on that information.

First, we could use the rating of actors/actresses on the Internet's Hollywood Stock Exchange (HSX). These ratings are the value that traders give to the different actors. These ratings are of the following form. With AAA being the best, stars are assigned one of the following rating: AAA, AA, A, BBB, BB, B, CCC, CC, C, and D. These had to be converted to 10 dummy variables. The hypothesised relationship was that higher ratings

had significant positive relationships and that lower ratings would have either smaller or negative coefficients.

As it turns out, only the AAA, AA, A and CC ratings were significant for the biggest star (Star 1) and ratings A and CC were significant for the second star. However, the coefficient for AA was higher than AAA for Star 1, and both CC ratings received highly negative coefficients. Since this wasn't exactly what we expected, we decided to turn to our second option, which was to use the average gross of movies in which stars played.

We chose to add the average grosses of the two biggest stars to form a "star power" variable. This average measure included the cumulative box-office of a movie into an actor's average if that actor was in the "first-billed" part of the cast. Thus, the actor didn't have to be a "star" in a movie for the average to include that movie's revenues. Furthermore, the average was calculated over the whole career of an actor (all movies for which actor was "first-billed"). In the end, we had a single quantitative variable, so we could look at its relationship with the box-office by using a scatter plot (see Figure 5). A linear relationship explained 14.76% of the variation in the opening week revenue (W1), while an exponential relationship explained 16.29% of the variation in W1. Taking the natural log of the opening week revenue (LN W1), the linear relationship is now exponential with relation to W1.

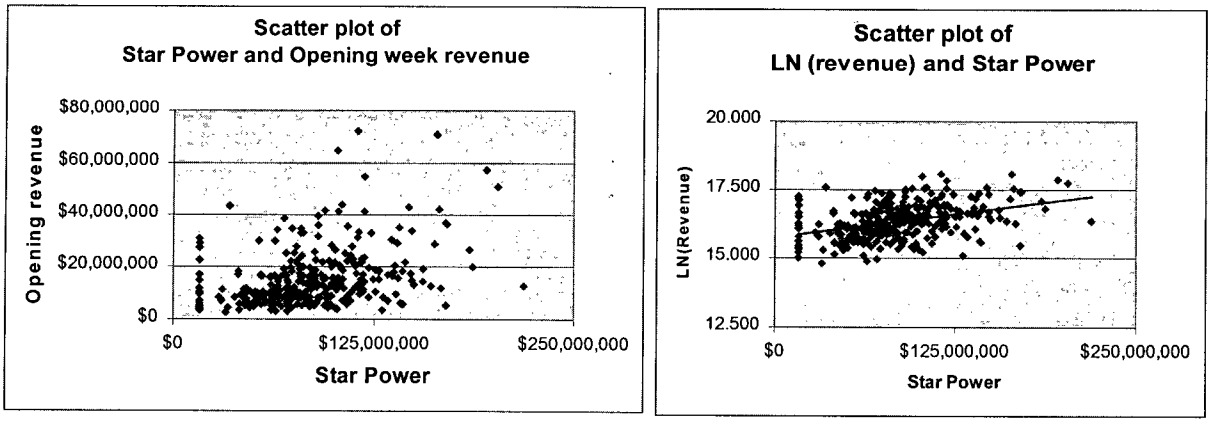


Figure 5: Scatter plots of the relationship between Star Power and revenue (dollar and log scale)

Discussion of Star Power

It seems that stars realise their worth quite quickly. The talent of actors is recognised very fast, which allows stars to play in successful movies right from the start of their career. We believe that the average gross over all movies by the two biggest stars is a good representation of the Star Power involved in a motion picture. Further, we believe that Star Power performs a dual purpose. First, movies with big stars are worth more because people will go see movies in which their favourite stars play. Second, big stars make "big bucks", so movies that have big stars also have large budgets for both production and advertising. Those two variables are positively correlated with movie box-office success.

Director Power

A very important attribute of a movie is its director. Among the most famous names in this category are Steven Spielberg (Indiana Jones series, Jurassic Park, etc.), Robert Zemeckis (Back to the Future series, Forrest Gump, Cast Away, etc.) or James Cameron (Titanic, Terminator 2, etc.). Directors give movies their flavour and also participate in the movie's promotion. They receive an increasingly important place in a movie's advertising and promotion. Readers may recall hearing the catch phrase "from the director of the movie "..." comes a new movie about..." Some of the biggest directors even get listed before the movie's stars. Of course, our hypothesis is that better directors positively influence the success of a movie at opening week. We thus expect positive regression coefficients.

The data available on directors was similar to that of stars, except that it was possible to track the weekly revenue of directors in addition to the total gross of their movies. As it did for stars, the HSX rating of director explained some of the variation in the first week revenue, but results were unsatisfactory when compared to the quantitative variables. The only significant HSX ratings were AAA and AA with high positive coefficients. However, these two variables explained only 8% of the variation in the opening week revenue (W1) (7% of LN W1).

Thus, we pitted the movies' average gross and the first week's average gross measures against each other to determine which one was worthier of receiving attention. We believed that the average first week's revenue (of the director's last 3 movies) was different and would explain more variation since it could provide more information on the type of movie that directors usually make. That is, some directors make movies that are more complex and rely more on word-of-mouth, but attract people for a longer period of time (Ang Lee is an example, with its latest movie *Crouching Tiger, Hidden Dragon*). Others use all of Hollywood's flashes and attract people right out of the gate, making the top gross for their movies in the opening week (George Lucas with 'Star Wars' or Wolfgang Petersen with 'The Perfect Storm' are examples).

The results of the simple regressions on the two possibilities was that both were significant and explained about 10% of the variation in the opening box-office. We chose to keep the average first week's revenue for further consideration since it explained slightly more variation in our dependent variable.

Number of screens

The number of theatre screens on which a movie is showing greatly influences a movie's box-office revenue. The foremost explanation suggested in the literature is accessibility. Indeed, more screens increase the likelihood that people will be close to a theatre showing the movie. It is considered a marketing variable in that it represents the effective distribution strategy for a movie.

We believe that in addition to accessibility, the number of screens represents the opinion of experts on the quality and marketability (and playability) of a movie. Indeed, exhibitors must choose which movies they show on their screens, and the fact they decide to show a movie speaks volumes about their opinion on its quality and market potential. It must be said that exhibitors receive a lot of information about the different movies being released during a season, and their experience and expertise drives their decision to show a movie or not.

The current trend in theatres is to build large multiplexes featuring about 20 screens and multiple quality concession stands. Since they have many screens, we believe and can see that these multiplexes play every movie that is being released. Thus, it is possible that a movie slated for a large audience will open on a certain number of screens almost *by default*. Therefore, it is interesting to look at the relationship between the number of screens and the opening box-office of a movie. Perhaps an exponential relationship would confirm that there now exists a "floor" on the number of screens on which movies get released.

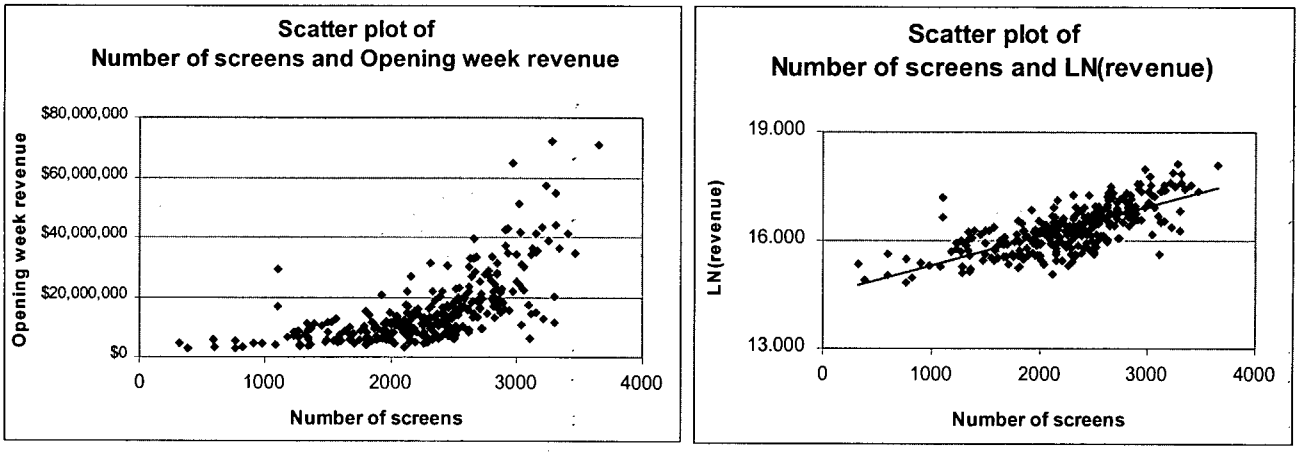


Figure 6: Scatter plots of the relationship between the number of screens and revenue (dollar and log scale)

Figure 6 presents quite convincingly how the relationship between the number of screens and revenue of a movie is exponential (thus linear on a log scale). Even all by itself, this relationship is so strong that it explains 53.7% of the variation in the natural log of the revenue in week 1.

Summary of other variables

At this point, the remaining explanatory factors are *genre, MPAA rating, sequel and studio*. For the *genre* category of variables, we find that action, comedy and drama are significant, with comedy and drama posting negative coefficients, while action receives a positive coefficient. We explain the result for "action" movies by the fact that this type of movie is fairly easy to understand (story and advertising) so people can rapidly decide

whether or not to see an action movie (Sawhney and Eliashberg 1996). The results for comedy and drama are harder to explain by a theoretical model, but we believe that they are due to the increased reliance on word-of-mouth and critics, and the fact that these movies usually have smaller budgets.

Then, we tested the significance of the *MPAA rating* as an independent variable predicting the opening box-office. Our goal was mostly to test the impact of the "R" rating on the success of movies. Since the rating restricts access to the movie only to adults, we expected a negative coefficient to be assigned to the variable. However, the relationship is not significant (and would go in the other direction if it were).

We also tested the impact of the *studio* on the success of a movie. The effects of the studio category were not significant because we restricted our attention to box-office successes that made it to at least the fifth spot on the charts. Therefore, our sample only contained movies by large studios with ample means, or else included the biggest box-office successes of smaller studios. It seems that studios do get both hits and misses, but box-office successes will almost exclusively come from the most important studios. There were some positive effects of a *sequel* variable but these effects disappeared when other variables were included. Thus, we are now ready to present the multiple linear regression equation that we will use to forecast the opening week revenue of movies.

Results: the regression equation on α

The attributes that were used for our regression equation on α were the number of screens, the director power (average first week revenue of last 3 movies) and the star power (combined average gross of the 2 biggest stars). Thus the model we ran was:

$$\textit{LN(week 1 revenue)} = b_0 + b_1 * \textit{(Screens)} + b_2 * \textit{(Davg1)} + b_3 * \textit{(S12Combo)} \qquad (12)$$

Where: ***Screens*** = number of screens in week 1
Davg1 = average first week revenue of the director's last 3 movies
S12Combo = combined average gross of the 2 biggest stars

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.785	.617	.613	.3968

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	79.832	3	26.611	169.028	.000
	Residual	49.592	315	.157		
	Total	129.424	318			

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	14.388048	.09490		151.607	6.2E-21
	Dir avg 1st wk rev	1.739E-08	3.0E-09	.219	5.821	1.4E-08
	Weekend 1 screens	6.827E-04	4.1E-05	.620	16.484	6.2E-21
	S12 Combo	2.707E-09	6.5E-10	.156	4.175	3.9E-05

Table 4: Regression results

From Table 4, one can find (under R Square) that the model explains 61.7% of the variation in the dependent variable (LN(W1), the natural log of the first week's revenue). That is pale in comparison with a study by Zufryden (2000) that succeeded in explaining 90% of the variation in the dependent variable. However, our study used publicly available data and still outperformed a study by Sawhney and Eliashberg (1996) that obtained an R^2 of .419 for their prediction on the maximum possible box-office potential (similar variable).

So a prediction for the first week box-office of a movie is obtained in the following way:

Predicted LN(W1) = 14.388048 + 1.739E-08 * (***Davg1***)
+ 6.827E-04 * (***Screens***) + 2.707E-09 * (***S12Combo***)

Readers should note that the regression model is highly significant at $p<0.0001$. So are the coefficients of the explanatory variables, with the least significant variable still in the order of $p<0.00001$. Further, one will see from the standardised coefficients (beta) that

the model relies heavily on the number of screens, and somehow equally less on the director power and star power. The correlation matrix for the predictors used in this regression equation can be found in Table 5. Also note that a split half validity test shows that all predictors remain significant and that our regression equation is quite stable.

Correlations

		Davg1	S12Combo	Screens	Action?	LN (W1)	Decay rate
Pearson Correlation	Davg1	1.000	.295**	.306**	.205**	.455**	-.071
	S12Combo	.295**	1.000	.295**	.101	.404**	-.062
	Screens	.306**	.295**	1.000	.201**	.733**	.126*
	Action?	.205**	.101	.201**	1.000	.216**	.165**
	LN (W1)	.455**	.404**	.733**	.216**	1.000	.060
	Decay rate	-.071	-.062	.126*	.165**	.060	1.000

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Table 5: Correlation matrix for regression analyses

Effects on β: the decay rate

We still need to obtain a prediction for the decay rate, or more specifically the difference between the first and second week on the log (ln) scale. We use the difference between first and second week because these two weeks represent a large percentage of the total revenue and there is a wider margin for error since numbers are larger than in following weeks.

It is extremely difficult to explain much of the variation in the decay rate for movies because we expect this rate to vary widely depending not only on the internal attributes of the movie, but mainly on external factors such as critics review, new competition and seasonality. However, it is still better to use a regression forecast then to use the same mean value of β for all movies. Note that β is positive since time provides the negative sign. Therefore, higher positive values mean quicker decay and larger difference between the revenue in the first and second week.

For the decay rate, we tried to use the same attributes as for the opening strength, with the following relationships in mind. *Number of screens* would provide an idea of the competition the movie might face, as well as its starting point. We thought higher number

of screens would mean faster decay rates, or a positive coefficient. Perhaps the *Star Power* and *Director Power* would represent the quality of the movie, and these attributes would thus receive negative coefficients, slowing down the decay rate with higher "power". Additionally, we thought certain genres would influence the decay rate. As it turned out, the *Star Power* was not significant and the only genre that made it to significance was "*action*".

Results: the regression equation on β

The attributes that were used for our regression equation on β were the number of screens, the director power (average first week revenue of last 3 movies) and the action dummy variable. Thus the model we ran was:

$$\textit{Decay rate (LN(W1) - LN(W2))} = b_0 + b_1 * \textit{(Screens)} + b_2 * \textit{(Davg1)} + b_3 * \textit{(Action?)} \quad \textit{(13)}$$

Where: ***Screens*** = number of screens in week 1
Davg1 = average first week revenue of the director's last 3 movies
Action? = dummy variable for the action genre (1 if action, 0 otherwise)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.235	.055	.046	.2472

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.130	3	.377	6.162	.0004
	Residual	19.250	315	6.111E-02		
	Total	20.379	318			

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.356753	.057		6.297	.000
	Action?	.117557	.040	.168	2.964	.003
	Avg# 1st wk	-4.64E-09	.000	-.147	-2.525	.012
	Weekend 1 screens	6.01E-05	.000	.137	2.364	.019

Table 6: Regression results for β

It might appear to the reader that the results in Table 6 are unsatisfactory for a regression analysis, since all the model explains is a mere 5.5% of the variation in the decay rate. However, the model is highly significant (***p=0.0004***) as well as are all coefficients (***p<0.02***). Predictions are obtained in the following manner:

Predicted decay rate (β) = 0.356753 + (-4.64E-09) * (***Davg1***)
+ 6.01E-05 * (***Screens***) + 0.1175 * (***Action?***)

Discussion on the results for β

A prediction for the decay rate is only really needed before the second week of the movie's run in theatres, since time (the multiplier of the decay rate) takes the value ***0*** at the opening of a movie. But our formulation is closed to external information at time ***t>0***, hence we need to include a prediction of the decay rate in the initial vector of parameters $\mathbf{m}_0$. The model was closed to external information to be applicable as a component to an optimisation model such as a dynamic program (MDP). If the model were open to external information, we could include data such as critic reviews after the release of the movie, and would perhaps obtain a better forecast on the second week.

The decay rate is somewhat less important since one point of data has been acquired by the model, so the decay rate will be applied to a representative starting point. Furthermore, the forecast is only used for the second week's forecast, since data takes over at the third week, and the difference between first and second week is now available to the model, from 2 data points.

Once again, the correlation matrix can be found in Table 5. Note that there are no colinearity problems between the predictors. Validity tests show this regression analysis to be somewhat conclusive, although some of the coefficients lose significance on some runs of the regression on a half of the data set.

Chapter 4 - DLM execution, results, comparison

DLM specifications

When implemented, the constant DLM closed to external information needs no manipulation other than the addition of the weekly data. The benefits of closed-form solutions and Bayesian updating are the ease of use and simplicity of the resulting forecasting tool. As it did for this application, a DLM can handle very large quantities of data rapidly and efficiently. Modifications, scenarios and what-if simulations can be tested almost instantly, especially when the model is running on user-friendly interfaces such as spreadsheets. We now show how the DLM was implemented and tested using a spreadsheet (Microsoft Excel).

The first elements to include for the model are specifications of the variance (V) in the observation equation, both the design (**G**) and variance (**W**) matrices for the system equation, and the prior, namely the $\mathbf{m_0}$ vector and $\mathbf{C_0}$ matrix.

Design and variance matrices						
V	1	G	1	0	W	4 0
			0	1		0 2

Prior				
$\mathbf{m_0}$	16.645 0.4250	$\mathbf{C_0}$	3 0	0 1

Table 7: Model specifications

Table 7 shows the different values of V, **G**, **W** and $\mathbf{C_0}$ that were used to do the forecasting for all movies in the test sample. The only movie-specific values are the alpha (α) and beta (β) coefficients found in $\mathbf{m_0}$. **G** is a 2x2 identity matrix (I_2). The variance V is associated to the observation equation and represents the error usually associated with measurement. Here, it has a value of 1 to account for the variability created by seasonality and competition. The variance matrix **W** represents the variance in the system equation, which represents the correctness of the assumption of linear constancy. Higher values will react to local trends, whereas θ values define a static regression model. The

variance matrix **W** assumes that there is no a priori covariance between the two parameters and takes on values that affect the adaptive coefficient A_t, which indicates the fraction of the error that the model will take into account in its parameter updating.

The prior variance C_0

The prior variance $\mathbf{C_0}$ was the same for all movies since all forecasts were obtained from the same two regression equations. The specifications for this matrix can be based on expert opinions or some calculations. In this case, the variances have been chosen to cover the range of possible values for the two parameters. Note that the prior variance also influences the adaptive coefficient A_t, especially for the first update of the parameters (after the first data point).

DLM Deployment

The closed-form solutions make it quite easy to implement the model on spreadsheets. Table 8 presents the implementation of the DLM for the first 3 weeks of the movie ***28 Days***. Our study was extended in a similar fashion to include 6 weeks of data for all movies. Sections of the table present the week, design vector $\mathbf{F_t}$ $(1, -t)'$, posterior, prior, forecast mean and variance, adaptive coefficient, weekly data, error $(Y_t - f_t)$ and posterior information. These sections represent the equations (8-11) found in the second chapter of this THESIS.

		Posterior at time t-1			Prior at time t			Forecast Mean	Forecast Var.	Adaptive Coeff.	Data	Error	Posterior information		
Wk	F_t	m_{t-1}	C_{t-1}		a_t	R_t		F_t	Q_t	A_t	Y_t	e_t	m_t	C_t	
1	1	16.65	3	0	16.65	7	0	**16.65**	8.00	0.88	**16.15**	-0.50	16.21	0.88	0
	0	0.43	0	1	0.43	0	3			0			0.43	0	3
2	1	16.21	0.88	0.00	16.23	4.88	0.00	**15.81**	10.88	0.45	**15.80**	-0.01	16.22	2.69	2.24
	-1	0.43	0.00	3.00	0.43	0.00	5.00			-0.46			0.42	2.24	2.70
3	1	16.22	2.69	2.24	16.23	6.69	2.24	**15.38**	17.53	0.12	**15.20**	-0.18	16.20	6.41	3.14
	-2	0.42	2.24	2.70	0.42	2.24	4.70			-0.41			0.49	3.14	1.78

Table 8: Implementation of the DLM for the movie ***28 Days***

It should come as no surprise that the posterior at time ***t-1*** and the prior at time ***t*** are based on design matrices multiplied by parameters or mean vectors. What is especially interesting to study is the impact of the adaptive coefficient (A_t), which expresses what

fraction (percentage) of the forecasting error is to be included in the new parameter specification.

It can be inferred from equations (9-11) that the adaptive coefficient (A_t) is based on the errors in the system equation (**W**), observation equation (V) and the initial information variance ($\mathbf{C_0}$). The impact of larger variances is a higher reliance on data. Indeed, higher values for the adaptive coefficient make the model more responsive to data, whereas smaller variances, and a smaller value for A_t, mean more reliance on the prior specification.

This need not be the case in other applications, but box-office forecasting demands a swift response and adaptation to data. If our initial forecast were better (higher r-square measure), we would put more faith in it and restrain the reliance on data. However, since there are some challenging streams of revenues to forecast (see figure 2), our model needs to be flexible and respond to new trends as they develop.

Box-office forecast results from the DLM and two competing approaches

The predictions obtained by the DLM assume that the external environment is stable from period to period. That is, we do not feed data on competition or seasonality into the model. Thus, the model could be used in its present form as a forecasting component to an optimisation routine or model. The model is also fast and flexible enough to be used as a simulation tool for the planning of a release schedule. We will see in the next section that it can be combined with other models and achieve even better results.

We will compare the results obtained by the DLM with two other models. We will look at the results from the use of an exponential smoothing model with trend, and also compare our model to the use of a "complete recalibration" method. We now describe the two competing approaches.

"Complete recalibration" (RCL)

This is the simplest way of obtaining a forecast. For the first period (when no data is available), the forecast is based on the estimate for the mean of α, obtained by the regression equation based on (eq. 12). The prediction in the second week is calculated using the data point as intercept, and the slope is defined by the estimate for the mean of the decay rate beta (β) obtained from the regression based on (eq. 13).

Subsequently, forecasts are obtained by extending the linear fit of the data points for one period ahead. For instance, the third week's forecast is the y-coordinate at period 3 (e.g. x-coordinate is 3) of a line connecting the two prior data points. Specifically, the line fit is obtained via linear regression, based on a least-squares method (or the Excel function LINEST). Note that the method uses all available data points in later periods. It fits a line through all available data points. (An alternate method could have been using the latest two data points.)

Exponential smoothing model with trend

There exist many formulations for exponential smoothing models with trend. We selected the one that obtained the best results, after "optimizing" the smoothing constants. Note that we used the same smoothing constants for all movies. Thus, the second competing approach is as follows:

$$\mathbf{FIT}_t = \mathbf{F}_t + \mathbf{T}_t$$
$$\mathbf{F}_t = \mathbf{FIT}_{t-1} + \gamma(\mathbf{A}_{t-1} - \mathbf{FIT}_{t-1})$$
$$\mathbf{T}_t = \mathbf{T}_{t-1} + \delta(\mathbf{A}_{t-1} - \mathbf{FIT}_{t-1}) \qquad \textbf{(14)}$$

The variable $\mathbf{FIT}_t$ is the forecast at time **t**, composed of the time-indexed level $\mathbf{F}_t$ and the trend $\mathbf{T}_t$. The model is completed with actual data in the previous period ($\mathbf{A}_{t-1}$), and two smoothing constants denoted γ and δ. Note that the initial level ($\mathbf{F}_0$) and trend ($\mathbf{T}_0$) are based on the regression equations described in the previous section. Thus, this model is similar to the DLM in the sense that it incorporates a fraction (defined by the smoothing constants γ and δ) of the error into the new specification of the level and trend. However,

the DLM uses an *adaptive* coefficient that is evaluated dynamically and based on conditional probabilities. Further, the DLM specifies the intercept (with y-axis) and subtracts a certain number of slope increments (defined by the time period), whereas the exponential smoothing model keeps a "level" that is independent of the period. Again, note that the smoothing constants were the same for all movies in the test sample.

Error measure

The error measure used to compare the models is calculated by using the following equation:

$$\textit{Percentage error} = MIN\left(\left| \frac{(Real - Predicted)}{Real} \right| , 1 \right) \qquad (15)$$

This measure calculates the absolute value of the difference between the real value and the prediction, this difference divided by the real value. Thus, it obtains a percentage error in terms of the real value. We also take the smallest value (MIN) between this percentage and 1. This value of 1 represents a percentage error of 100%. Higher percentage errors (more than 100%) would only confuse the results. Note that these do occur, for instance when if the prediction was 12 and the real value was 4, then the error would be 200%, but we would replace that with 100%.

This measure has some definite disadvantages such as the difficulty of obtaining good results for small denominator values. But, to its defense, we must say that it has been previously used in extant research (Sawhney and Eliashberg 1996) and it also serves as an equal comparison basis with web-based box-office forecasters such as the "Box-Office Mojo" (http://www.boxofficemojo.com/). Furthermore, the error measure keeps the forecasting task challenging even in the tail weeks, when it should be easier considering more data is available, and variability is much smaller. Thus, we keep the error measure and work even harder to obtain better results.

Samples

We tested the three approaches on the 59 movies that made it to at least the 5th spot on the box-office charts for the period running from January 7, 2000 to July 7, 2000. For the

Top 10 forecast results that will be presented in the next section, the sample consisted of 27 weeks. This cross-section is fairly representative of the full spectrum in terms of revenues, seasonality and competition. The opening week revenue for the movies in the sample ranged from $4.5 to $70.8 million with an average of $16.9 million. The cumulative top 10 in the period ranged from $42 to $168 million, with an average of $82 million. In terms of competition, the number of new movies opening in a given week ranged from 0 to 3, with 1 week without new openings, 8 with 1 new movie, 11 with 2, and 7 with 3 new movies being released. Other studies were mainly concerned with the high seasons, and focused on the summer and winter holiday seasons (e.g. Krider and Weinberg 1998).

Results

We "optimized" the parameters of the DLM to minimise the average "percentage error" (eq. 15) over our sample. The values were $V = 1$, $W = 2.5, 0; 0, 3$ and $C_0 = .5, 0; 0, 0$. We also found the optimal set of smoothing constants for the exponential smoothing model with trend. The smoothing constant for the level was $\gamma = .8$ and the smoothing constant for the trend was $\delta = .35$. (Values were found using the Excel Solver (non-linear)).

	MODEL		
Measure (in percentage error)	**Dynamic Linear Model (DLM)**	**Exponential smoothing with trend**	**Complete recalibration (RCL)**
Average	24.71%	24.81%	26.84%
Minimum	7.73%	6.68%	7.62%
5th percentile	9.95%	9.29%	9.72%
25th percentile	14.92%	14.86%	13.87%
Median	22.29%	22.54%	24.68%
75th percentile	33.45%	31.68%	35.81%
95th percentile	49.77%	46.76%	52.80%
Maximum	57.08%	54.74%	65.32%

Table 9: Distribution of percentage errors (59 movies, 6 weeks) for competing approaches

As can be seen in Table 9, the DLM outperformed the two competing approaches. However, each model was best for at least some cases, but on average the DLM was the best model. Appendix 2 presents the complete results for the DLM, whereas Appendix 3 presents results for the exponential smoothing model. Results for the complete recalibration method are found in Appendix 4.

Referring back to the challenges presented in Figure 2, we note that the DLM was especially better than "complete recalibration" for cases where the movie was a "sleeper". That is, the RCL method was inadequate for movies that made more revenue in their second week than their opening week (4 movies out of 59*: The Tigger Movie, Galaxy Quest, Snow Day, My Dog Skip*). Figure 7 presents a comparison of the RCL approach versus the DLM. We can see that the RCL predictions actually go upwards at the third period and take several periods to "converge" back to the real data stream. However, the DLM stays much closer to reality.

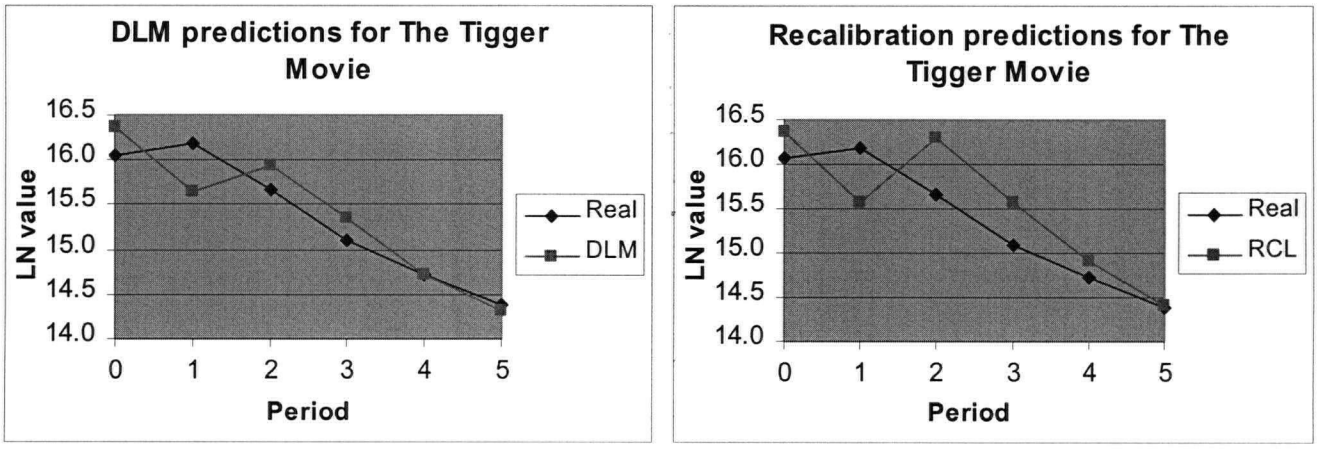


Figure 7: Comparison of predictions for "The Tigger Movie" (DLM and RCL)

The difference between the DLM and the exponential smoothing with trend forecasts resides in the adaptive coefficient. Indeed, the smoothing constants are the same for all periods whereas the adaptive coefficient of the DLM is calculated each period. Still, the DLM should be favored over the exponential smoothing model because of all the information it provides on the distribution of the forecast and parameters. Furthermore,

the DLM need not be closed to external information. Indeed, if the forecaster is aware of a factor that would influence the performance of a movie, then there are many different ways to incorporate that external knowledge into the model (see West and Harrison 1997).

Additionally, the forecaster could include more information in a systematic fashion. For instance, the DLM could be extended into a *multiple linear regression DLM* that would include perhaps the number of screens, data on critics' reviews, or any other explanatory variable that would be used in conjunction with time to explain variations in the revenue of a movie. Thus, we believe that the DLM framework is much more interesting than the two competing approaches.

Short-term vs. long-term forecasting with DLM

It must be noted that the DLM provides good one-step ahead forecasts, but is not intended to provide long term forecasts. In our study, the long-term predictions vary widely from period to period.

		Prediction by week number					
Based on	**Parameters**	**1**	**2**	**3**	**4**	**5**	**6**
Week 1:	(16.37 , 0.498)	16.37	15.88	15.38	14.88	14.38	13.88
Week 2:	(16.14 , 0.498)	16.14	15.64	15.14	14.64	14.15	13.65
Week 3:	(16.31 , 0.184)	16.31	16.12	15.94	15.76	15.57	15.39
Week 4:	(16.30 , 0.312)	16.30	15.98	15.67	15.36	15.05	14.74
Week 5:	(16.29 , 0.395)	16.29	15.90	15.50	15.11	14.71	14.32
Week 6:	(16.29 , 0.394)	16.29	15.90	15.50	15.11	14.72	14.32

Table 10: DLM Predictions based on the sequence of parameters for *The Tigger Movie*

Table 10 presents the forecasts based on the specification of the parameters in each week. The shaded cells represent the one-step-ahead forecasts that we actually compiled. Thus the line starting with "Week 1" represents our initial view. At that time we needed a forecast for the first week, which in this case is 16.37, and had we wanted a forecast for the 6th week, it would have been 13.88. As we can see, the 6th week forecast varied dramatically across the six parameter specifications. In real dollars, the 6th week forecast

in week 2 (13.65) equals $847,461 whereas the same forecast in the third week (15.39) equals $4,828,276.

Weekly breakdown of average error

The weekly breakdown of the average error for the DLM is also worth studying. As can be seen in the first row of Table 11, the first week forecast is rather off target (36.73%) but with one data point in, the model achieves much better results (20.78%). The third week achieves very good results (18.33%), since the model is past its initialisation period. The increasing error in the tail weeks is due in part to the denominator-dependent measure used to calculate the forecast error. Indeed, as values of the denominator decrease, so does the margin for error. Results for the two competing approaches follow a somewhat similar pattern.

	Weekly percentage error					
Model	***Week 1***	***Week 2***	***Week 3***	***Week 4***	***Week 5***	***Week 6***
Dynamic Linear Model (DLM)	36.73%	20.78%	18.33%	21.20%	24.65%	26.60%
Exponential smoothing	36.73%	20.62%	19.43%	20.60%	24.64%	26.83%
Recalibration method (RCL)	36.73%	19.53%	22.04%	24.16%	26.91%	31.70%

Table 11: Weekly breakdown of forecast percentage error averages

Validation

An interesting feature of the DLM is the possibility of evaluating the error as a function of the forecast's standard deviation. The model, as it is built, assumes normality of the system and observation errors, with ***0*** means and predetermined variances (V and **W**). Indeed, from equations (5-6), we see that the error terms v_t and ϖ_t should be normal and have a mean of ***0***. We use the forecast variance Q, as found in Table 8, to calculate the number of standard deviations that the error deviates from. That is, we take the error and divide it by the standard deviation to obtain a number. From normal theory, we know that most errors will fall between *1* and *-1* standard deviations away from reality, and have a mean of *0*.

$$\textit{Stdev away} = \quad (Yt - ft) / (Q)^{.5} \qquad (16)$$

We collected the ***Stdev away*** of equation (16) for the 59 movies. To verify the normality of errors, we created a histogram with the full 354 data points (59 movies x 6 weeks) of ***Stdev away***. Such a graph would have to adopt the well-known "bell" shape of the normal curve. We present the histogram in Figure 8 and conclude that the normality assumption holds.

Bin	Frequency
$-\infty$, -1.1]	3
]-1.1 , -0.9]	2
]-0.9 , -0.7]	8
]-0.7 , -0.5]	22
]-0.5 , -0.3]	33
]-0.3 , -0.1]	70
]-0.1 , 0.1]	107
]0.1 , 0.3]	69
]0.3 , 0.5]	25
]0.5 , 0.7]	11
]0.7 , 0.9]	3
]0.9 , 1.1]	0
]1.1 , $+\infty$	1

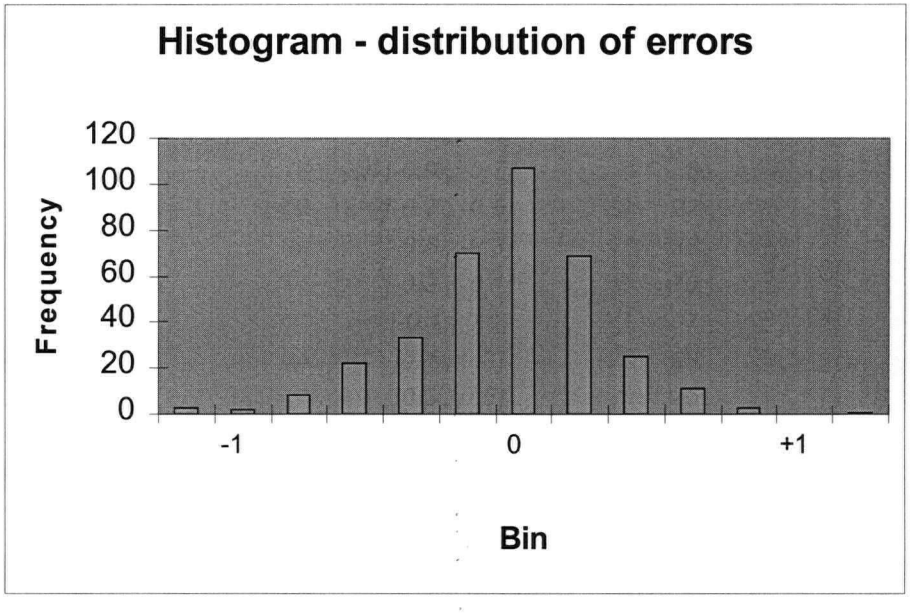


Figure 8: Histogram of the distribution of errors - normality assumption

We also looked at the weekly breakdown of these errors and constructed histograms for each individual week. We found that all adopted the bell shape, but some weeks were not centred at *0*. We present the weekly means in Table 11.We expect those averages to be *0*, so that equation (5) is respected. Negative means denote a systematic tendency to predict higher than reality. Thus, it appears that the model overshoots the real value in weeks 5 and 6, when many theatres replace movies for new ones thereby accelerating the descent of the revenue of movies.

Stdev away					
Week 1	Week 2	Week 3	Week 4	Week 5	Week 6
-0.07197	0.022297	0.020477	-0.0387	-0.11963	-0.13627

Table 12: Weekly breakdown of *Stdev away* means

Fine tuning

A DLM modeller faced with the results shown in Table 12 has the opportunity to incorporate a non-zero mean to the error term v_5 and v_6 to correct the systematic over-prediction. On average, the prediction is expected to improve. In our case, we wouldn't have known this fact to begin with, but could have applied the correction on a subsequent study.

Chapter 5 - Top 10 forecasting: attraction model

Attraction model

In this section, we try to improve on our results by incorporating the effects of seasonality and competition on top of the strength prediction. For that we use a share attraction framework, which has been suggested for movies by Krider and Weinberg (1998). The attraction model suggests that the revenue of a movie i at time t (R_{it}) is a function of the total demand (D_t) slated for a weekend and the total strength (summed) of other competing movies (strength A_{jt} for movie j at time t). That is, a movie gets a share of the total demand that is evaluated from its comparative strength relative to all other movies playing. The revenues for movie i at time t is calculated from equation (17):

$$R_{it} = D_t \frac{A_{it}}{\sum_{j \in J_t} A_{jt}} \qquad (17)$$

The strengths are summed over J_t, the set of all movies playing at time t. Thus, seasonality is represented by the total demand D_t, which we studied in Figure 1, and competition is tracked in the sum of strengths of the movie set J_t. Now, these strengths

need to be calculated, and Krider and Weinberg suggest that we use the following equation for the strength of movie i at time t:

$$A_{it} = e^{\alpha_i - \beta_i (t - t_{0i})} \qquad (18)$$

Where the result of $(t-t_{0i})$ is just the age of the movie (t_{0i} is the opening time of movie i). Thus, it is exactly the formulation of the dependent variable in our DLM framework. Hence, movie strengths are readily available and are the output of the DLM. We still need to correctly assess the total demand in order to be accurate.

Demand forecasting

The total demand (D_t) is the cumulative box-office of the top 10 in week t. Historical data is readily computed by summing the grosses of the 10 leading movies in prior weeks for some number of years. Our database featured reliable box-office results starting from the year 1997. We thus compiled the cumulative box-office for every week starting from January 3rd 1997 to August 11th 2000.

We used data from prior years (1997, 1998, 1999) to make demand forecast for the year 2000. Making sure that holiday weekends were matched, we found the weighting scheme that minimised the mean absolute deviation (MAD). This history-based prediction suggested respective weights (summing to 1) of 0.041 for 1997, 0.536 for 1998 and 0.423 for 1999.

Weights for each year		
1997	1998	1999
0.041	0.536	0.423

Table 13: Smoothing weights for year 2000 forecast

The MAD was about 13%, which is a fair result given the complexity of the series. The corresponding R^2, or the percentage of the variation explained by the forecast, was 0.675 or 67.5%. We kept the resulting predictions and used them to test an idea that seemed fairly intuitive. That is, we wanted to test whether the sum of strengths (denominator in (17)) had an effect on the total demand. We thus hypothesised that the total demand was

a function of both a historical-seasonal component, or periods in the year when people are more inclined to consider going to theatres, and a "strength recognition" component whereas the masses are more attracted by a strong set of movies, even though each consumer might only see one movie. That is somehow the answer of consumers to Hollywood, positive if studios release quality films, and negative when only weak movies are launched.

We constructed a regression equation based on the history-based prediction and the sum of strengths produced by the DLM. As a result, the percentage of the variation explained by the forecast went up to 77%. The regression nearly escapes colinearity problems since the prediction and sum of strength have a pair-wise correlation of .665, which is high but not excessively, since problems occur at correlation levels of more than .8.

Results

The results were quite surprising and somehow not as good as expected. We tested the proposed models for their sensitivity to different values of the total demand and to the prior definition for the parameters. The idea was to test the validity of the conceptual model itself. Our approach was to use an accurate specification of the prior for the parameters, thus verifying the assumption that an accurately initialised model would perform admirably. Since the error in the first two weeks was due to somehow inadequate regression equations, we corrected the problem by using real data for the prior specification, which ensured at least two weeks of perfect results.

Similarly, we tested the potential of the attraction model, coupled with our DLM predictions, and used real data to represent total demand. Thus, if the output of the DLM created the correct rankings, then the attraction model would scale the predictions and lead to excellent results. As a final test, we used both the accurate prior specification and total demand to really identify the best achievable results with the use of our conceptual model. Table 14 presents results for the average error percentage as calculated in equation ***(15)***.

	Base case	Accurate prior (parameters)	Accurate total demand	Accurate prior and total demand
DLM	23.5%	14.6%	23.5%	14.6%
Attraction model - History-based	26.0%	22.7%	21.5%	14.9%
Attraction model - Regression	24.2%	18.1%		

Table 14: Results of the box-office Top 10 forecasting

The aim of Table 14 is to enable us to answer the question of what to put more efforts on. That is, should we improve on the regression equations (prior specification), or else try to get a better forecast for the total demand? The short answer is that finding a better prediction for the parameters (α, β) is where work should be done (or money invested). We now discuss the results found in Table 14.

For the base case (first column of results), the prior specification for the parameters (α, β) is obtained from the regression equation presented in Chapter 4 and the total demand is predicted by either the history-based prediction or the regression equation using history and sum of strength. The results are that the raw output of the DLM (no attraction model modification) leads to an average percentage error of 23.5%. Using the attraction model to scale the results, we find that a total demand forecast based on history blurs the results and achieves an average error of 26%. A better prediction achieves an average error of 24.2%. If the prior specification is not so great, we conclude that it is best not to use the attraction model.

We see from the second column of results (down from "accurate prior") that the DLM alone, when initialised with real data, does not need to be accompanied by the attraction model, which only serves to blur the accuracy of the forecasts. Thus, if one were to obtain a very accurate regression equation for the parameters, the raw output of the DLM would be its best forecast for the Top 10. The average error from the DLM is an excellent 14.6%, whereas the attraction model only detracts the results to average errors of 22.7% and 18.1%.

In the case where the available regression equations (Chapter 4) were our best shot, we could improve our forecast by 2% by using real total demand within the attraction framework of equation (17). That is, we ran the model with realised demand as input to the attraction model. In the case where the DLM used relatively inaccurate forecasts, the attraction model obtained better results (21.5%) than the raw output of the DLM (23.5%).

In the rightmost column, we can see the results in the case where the model was parameterised exactly and the total demand for the attraction model (eq. 17) was also exact. In that case, it would be better to rely on the raw output of the DLM rather than use the output in conjunction with the attraction model.

Studying Table 14, we see that there is a large gain from a better parameterisation of the DLM (from 23.5% to 14.6%), whereas a better demand forecast offers a smaller margin for improvement (from 23.5% to 21.5%). Thus, it seems to be most worthwhile to find the best way to accurately predict the parameters. We believe that our model is quite valid and reliable if it is parameterised with an accurate forecast. Recall that the works of Zufryden (2000) found that the best explanatory variables were the number of screens, time since introduction, Cinemascore exit interview data, film website activity (new distinct point of origin or 'visitors'), production budget, and a seasonality measure. These explained 90% of the variation in the box-office.

Chapter 6 - Discussion of results, further research

Readers should note that the attraction model does not change the rankings obtained as output of the DLM. Thus, the attraction model is interesting for applied matters (accuracy), but in a case such as an optimisation routine, results would be less affected. This is due to the fact that the attraction model is used after the DLM, and forecasts are not fed back into the DLM. Hence the attraction model only affects the scale of the predictions, not the rankings.

That last column of Table 14 describes the error that is not explained by our conceptual framework. Since the minimum is at 14.6%, perhaps it signals that a better model might exist, or else that forecasting the box-office Top 10 is just extremely hard to do! What could be said about the attraction model is that all the movies playing are affected in the same way from the release of a new movie. However, this does not have to be the case. Perhaps the impact of a new movie is only found in movies that share the same target audience. For example, the release of a "teen" movie shouldn't affect one's intention to see a serious adult drama. Thus, there might be ways to improve the model by looking at target audiences. The knowledge and data requirements would be quite high, but results would likely be improved.

Readers should note that all analyses were conducted on the log-of-real-dollar scale, rather than on a log-of-deseasonalised data scale. The first part of the reason why is that real dollars are more in accordance with the linear assumption, since the linear fit explains 95.7% of the variation in revenues, whereas a linear fit on the deseasonalised data explains 93.6% of the variation in revenue. That alone, however, does not constitute a sufficient reason to not go forward with the analysis. However, deseasonalised data needs to be put back into real dollars, with some prediction of the total revenues for the weekend. That, in this research, was the missing link that made it impossible to conduct analyses on deseasonalised data. Our total demand forecast was not accurate enough. However, further research should still contemplate the use of deseasonalised data, as some parts of the work might be easier to manage with "market shares", as opposed to real dollars.

An interesting extension to this work and an area for further research is the study of the multivariate case of the DLM. The multivariate case would include a vector of many movies $\mathbf{Y}_t$ as the dependent variable. Perhaps all movies slated for release during a season would be included in this new dependent variable. We believe to have clearly demonstrated that there exists a dependence relationship between movies. Thus, there is a need for a model that could account for the covariance between movies. That covariance, our intuition suggests, is based on target audiences or perhaps on the genre of the movie.

As an hypothesis, it would be interesting to test whether or not competitive influences are greater within the same genre.

From a modelling standpoint, the multivariate case is quite straightforward. From equation (5), the multivariate case adopts a vector of data points Y_t, a matrix of error v_t belonging to a covariance matrix V_t. The multivariate case will be a challenge to implement in practice because of the covariance matrix V_t, which will have to be based on a study of target audiences or genres. Further, the modeller will be challenged to find a way to incorporate new movies or will have to look at a full season of movies at a time. Indeed, the model will not be closed to external information if it incorporates new movies every week. To be closed to external information, the model will need to look ahead at a full season of movies, and the covariance matrix will have to be established right from the start.

Readers should recall that closing the model to external information allowed the DLM to be a potential component of an optimisation routine. At this point, the use of dynamic programming (MARKOV Decision Processes Theory) in conjunction with a DLM seems highly promising. Indeed, the DLM shares the properties of the dynamic programming framework. The DLM is state-based and Bayesian, thus it incorporates the past into the present state (MARKOV property). Further research into this problem can be readily initiated.

Bibliography

Bradlow, Eric T. and Peter S. Fader (2000). A Bayesian Lifetime Model for the "Hot 100" Billboard Songs. *Manuscript submitted for publication.*

Chase, Richard B., Nicholas J. Aquilano and F. Robert Jacobs (1998). *Production and Operations Management: Manufacturing and Services Eighth Edition.* Irwin McGraw-Hill, New York.

Eliashberg, Jehoshua and Steven M. Shugan (1997). Film Critics: Influencers or Predictors? *Journal of Marketing* 61 (April), 68-78.

Eliashberg, Jehoshua, Sanjeev Swami, Charles B. Weinberg and Berend Wierenga (2000). Implementation and Evaluation of SilverScreener: A Marketing Management Support System for Movie Exhibitors. *Interfaces* (forthcoming).

Jedidi, Kamel, Robert E. Krider and Charles B. Weinberg (1998). Clustering at the movies. *Marketing Letters* 9 (4), 393-405.

Krider, Robert E. and Charles B. Weinberg (1998). Competitive Dynamics and the Introduction of New Products: The Motion Pictures Timing Game. *Journal of Marketing Research* 35 (February), 1-15.

Lehmann, Donald R. and Charles B. Weinberg (2000). Sales Through Sequential Distribution Channels: An Application to Movies and Videos. *Journal of Marketing* 64 (July), 18-33.

Neelamegham, Ramya and Pradeep Chintagunta (1999). A Bayesian Model to Forecast New Product Performance in Domestic and International Markets. *Marketing Science* 18 (2), 115-136.

Neelamegham, Ramya and Dipak Jain (1999). Consumer Choice Process for Experience Goods: An Econometric Model and Analysis. *Journal of Marketing Research* 36 (August) 373-386.

Sawhney, Mohanbir S. and Jehoshua Eliashberg (1996). A parsimonious Model for Forecasting Gross Box-Office Revenues of Motion Pictures. *Marketing Science* 15 (2), 113-131.

Squire, Jason E. (1992). *The Movie Business Book, Second Edition*. Fireside Editions - Simon & Schuster, New York.

Swami, Sanjeev, Jehoshua Eliashberg and Charles B. Weinberg (1999). SilverScreener: A Modeling Approach to Movie Screens Management. *Marketing Science* 18 (3), 352-372.

Swami, Sanjeev, Martin L. Puterman and Charles B. Weinberg (1999). Play it again Sam? Optimal Replacement Policies for A Motion Pictures Exhibitor. Working Paper, University of British Columbia.

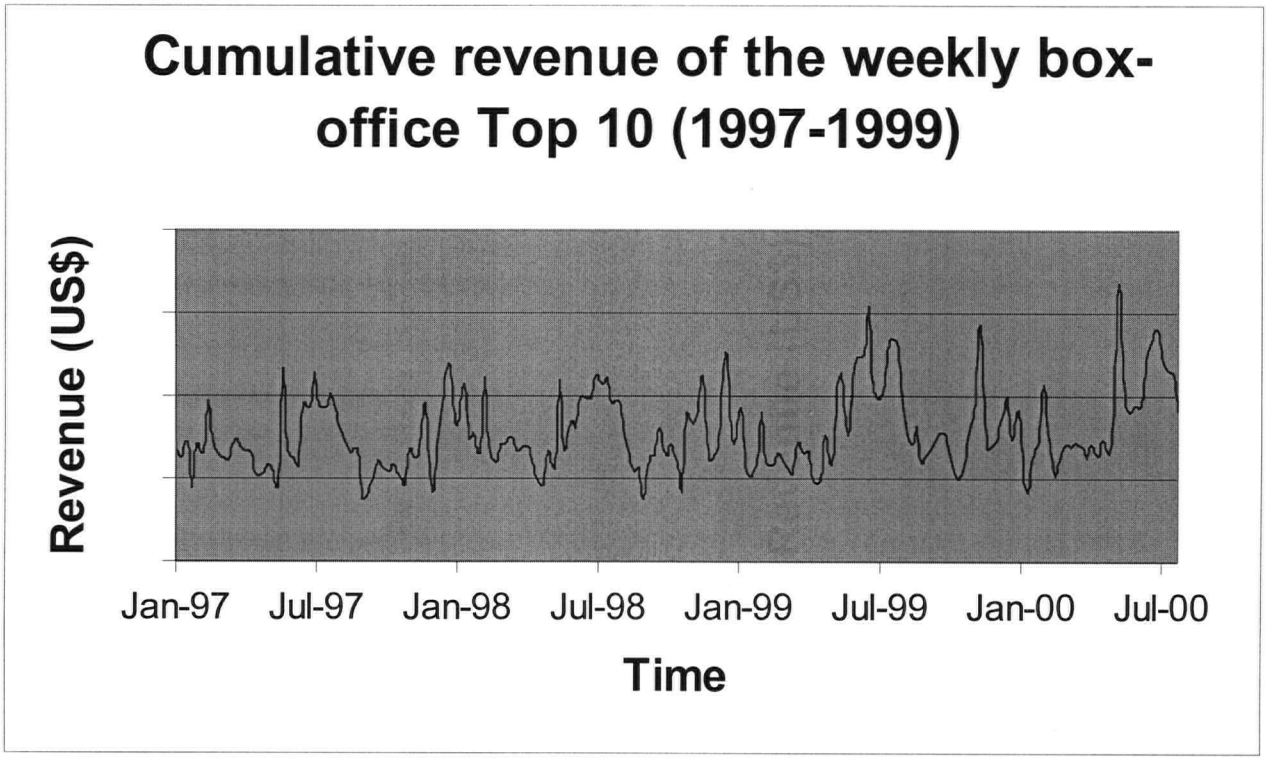
Wasserman, Gary S. and Agus Sudjianto (1996). A comparison of three strategies for forecasting warranty claims. *IIE Transactions* 28, 967-977.

West, Mike and Jeff Harrison (1997). *Bayesian Forecasting and Dynamic Models, Second Edition*. Springer-Verlag, New York.

Zufryden, Fred (1996). Linking Advertising to Box Office Performance of New Film Releases - A Marketing Planning Model. *Journal of Advertising Research* (July/August 1996), 29-41.

Zufryden, Fred (2000). New Film Website Promotion and Box-Office Performance. *Journal of Advertising Research* (April 2000), 55-64.

Appendix 1: Time Series of weekly box-office Top 10 (1997-2000)



Appendix 2: Forecast results using the DLM for the 59 movies in the test sample

		Percentage error						
#	Movie Title	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Average
1	28 Days	64.29%	4.51%	16.74%	1.73%	5.06%	52.33%	24.11%
2	Adventures Rocky Bullwinkle	100.00%	27.42%	5.41%	85.69%	35.37%	42.33%	49.37%
3	Any Given Sunday	1.63%	27.44%	3.69%	21.40%	63.64%	7.49%	20.88%
4	Battlefield Earth	100.00%	100.00%	46.98%	66.42%	22.66%	6.45%	57.08%
5	Beach, The	7.32%	12.11%	29.65%	19.29%	16.33%	4.31%	14.83%
6	Bicentennial Man	90.27%	8.76%	45.23%	34.19%	1.09%	100.00%	46.59%
7	Big Momma's House	23.96%	12.11%	0.85%	6.54%	1.65%	4.54%	8.28%
8	Chicken Run	2.33%	12.78%	7.26%	8.35%	29.22%	4.60%	10.76%
9	Deuce Bigalow: Male Gigolo	6.72%	9.11%	5.07%	38.95%	16.97%	13.18%	15.00%
10	Dinosaur	44.76%	44.53%	88.26%	20.85%	9.28%	11.59%	36.54%
11	Down to You	10.02%	23.16%	14.76%	18.89%	74.52%	70.45%	35.30%
12	Drowning Mona	58.21%	18.04%	18.80%	30.54%	48.74%	100.00%	45.72%
13	Erin Brockovich	18.49%	6.67%	10.01%	1.06%	1.06%	9.11%	7.73%
14	Eye of the Beholder	33.43%	2.08%	28.81%	100.00%	100.00%	65.99%	55.05%
15	Final Destination	17.51%	11.50%	12.35%	3.10%	11.34%	19.95%	12.63%
16	Flintstones Viva Rock Vegas	67.48%	7.50%	17.70%	12.75%	9.81%	42.19%	26.24%
17	Frequency	54.60%	5.53%	16.88%	22.70%	11.12%	37.01%	24.64%
18	Galaxy Quest	100.00%	46.49%	1.61%	16.70%	75.96%	10.33%	41.85%
19	Girl, Interrupted	12.68%	33.31%	26.43%	22.30%	28.36%	66.54%	31.60%
20	Gladiator	38.33%	28.49%	16.06%	13.69%	59.98%	16.02%	28.76%
21	Gone in 60 Seconds	27.44%	6.22%	2.86%	11.50%	10.24%	11.77%	11.67%
22	Green Mile, The	39.09%	1.18%	11.90%	46.63%	0.80%	7.90%	17.92%
23	Hanging Up	3.92%	27.88%	8.86%	11.38%	46.61%	35.07%	22.29%
24	Here on Earth	78.70%	48.83%	26.85%	48.85%	18.93%	45.77%	44.66%
25	High Fidelity	19.83%	4.37%	7.62%	20.75%	25.86%	14.44%	15.48%
26	Hurricane, The	38.44%	5.55%	22.64%	12.73%	8.90%	3.02%	15.21%
27	Keeping the Faith	20.56%	26.41%	18.68%	12.44%	1.13%	14.04%	15.54%
28	Kid, The	20.43%	28.10%	15.53%	7.70%	27.39%	11.49%	18.44%
29	Love and Basketball	28.83%	1.14%	10.42%	5.39%	6.98%	12.09%	10.81%
30	Me, Myself & Irene	0.09%	13.38%	6.99%	11.83%	17.34%	11.20%	10.14%
31	Mission to Mars	38.02%	44.46%	10.45%	11.51%	12.75%	20.27%	22.91%
32	Mission: Impossible 2	10.66%	37.91%	27.47%	22.62%	16.35%	9.40%	20.73%
33	My Dog Skip	80.30%	29.87%	15.37%	41.06%	30.72%	26.13%	37.24%
34	Next Best Thing, The	53.62%	21.33%	19.36%	52.55%	100.00%	1.55%	41.40%
35	Next Friday	71.96%	2.94%	8.05%	8.59%	6.94%	5.89%	17.40%
36	Ninth Gate, The	0.36%	23.15%	8.45%	35.29%	49.24%	29.08%	24.26%
37	Patriot, The	59.69%	13.95%	4.93%	12.13%	10.22%	25.18%	21.02%
38	Perfect Storm, The	1.15%	9.96%	1.26%	12.07%	21.06%	21.84%	11.22%
39	Pitch Black	40.21%	10.71%	14.07%	10.57%	27.80%	39.41%	23.79%
40	Reindeer Games	66.64%	24.66%	3.97%	25.78%	47.66%	86.83%	42.59%
41	Return to Me	25.73%	4.89%	21.64%	17.61%	0.32%	13.10%	13.88%
42	Road to El Dorado, The	59.11%	6.33%	10.13%	24.63%	81.96%	35.88%	36.34%

43	Road Trip	8.39%	31.12%	48.04%	10.43%	4.97%	32.75%	22.62%
44	Romeo Must Die	17.86%	11.67%	19.20%	9.14%	5.78%	5.51%	11.53%
45	Rules of Engagement	70.89%	4.88%	15.22%	15.00%	22.57%	14.87%	23.90%
46	Scary Movie	48.01%	25.65%	8.41%	13.52%	17.09%	3.23%	19.32%
47	Scream 3	0.45%	21.29%	14.43%	6.62%	6.09%	0.82%	8.29%
48	Shaft	35.47%	1.19%	23.72%	3.45%	2.87%	33.70%	16.73%
49	Shanghai Noon	21.46%	25.44%	17.85%	9.69%	12.99%	64.68%	25.35%
50	Skulls, The	8.37%	8.06%	4.41%	10.63%	34.34%	7.45%	12.21%
51	Snow Day	1.11%	42.89%	46.73%	25.88%	0.35%	4.56%	20.25%
52	Stuart Little	30.94%	16.38%	50.51%	29.50%	21.99%	78.49%	37.97%
53	Talented Mr. Ripley, The	0.81%	35.40%	3.52%	9.58%	43.25%	4.39%	16.16%
54	Tigger Movie, The	36.99%	41.56%	32.77%	29.40%	0.26%	6.25%	24.54%
55	Titan A.E.	100.00%	85.35%	38.96%	21.36%	23.28%	51.06%	53.34%
56	Toy Story 2	48.07%	15.96%	20.71%	12.42%	28.45%	44.64%	28.38%
57	U-571	20.09%	7.17%	0.14%	13.34%	17.37%	3.36%	10.24%
58	Where the Heart Is	70.47%	14.57%	26.08%	6.91%	4.97%	38.12%	26.85%
59	Whole Nine Yards, The	10.84%	2.60%	20.90%	15.33%	16.28%	9.64%	12.60%

	TOTAL	36.73%	20.78%	18.33%	21.20%	24.65%	26.60%	24.71%

Appendix 3: Forecast results using the exponential smoothing model for the 59 movies in the test sample

		Percentage error						
#	Movie Title	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Average
1	28 Days	64.29%	14.29%	4.83%	0.67%	6.71%	49.71%	23.42%
2	Adventures Rocky Bullwinkle	100.00%	3.69%	13.64%	74.53%	31.68%	45.45%	44.83%
3	Any Given Sunday	1.63%	27.91%	13.95%	11.49%	59.90%	8.23%	20.52%
4	Battlefield Earth	100.00%	50.48%	52.30%	91.25%	10.70%	8.14%	52.14%
5	Beach, The	7.32%	15.56%	37.31%	32.22%	25.49%	0.66%	19.76%
6	Bicentennial Man	90.27%	29.43%	54.70%	2.65%	7.77%	100.00%	47.47%
7	Big Momma's House	23.96%	1.95%	1.55%	5.72%	2.15%	4.72%	6.68%
8	Chicken Run	2.33%	13.57%	11.95%	13.14%	23.94%	4.24%	11.53%
9	Deuce Bigalow: Male Gigolo	6.72%	6.55%	3.49%	38.89%	22.76%	11.72%	15.02%
10	Dinosaur	44.76%	29.70%	78.56%	11.38%	8.99%	12.41%	30.97%
11	Down to You	10.02%	18.56%	10.67%	18.06%	79.56%	78.48%	35.89%
12	Drowning Mona	58.21%	1.72%	12.20%	32.36%	54.84%	100.00%	43.22%
13	Erin Brockovich	18.49%	1.27%	7.54%	2.27%	1.55%	9.27%	6.73%
14	Eye of the Beholder	33.43%	12.74%	19.11%	100.00%	100.00%	63.19%	54.74%
15	Final Destination	17.51%	17.02%	19.18%	11.00%	15.65%	22.65%	17.17%
16	Flintstones Viva Rock Vegas	67.48%	12.51%	25.75%	1.13%	13.20%	36.79%	26.14%
17	Frequency	54.60%	20.62%	26.78%	31.71%	19.76%	31.16%	30.77%
18	Galaxy Quest	100.00%	59.81%	31.11%	34.75%	47.81%	5.68%	46.53%
19	Girl, Interrupted	12.68%	40.73%	16.15%	21.68%	26.23%	72.63%	31.68%
20	Gladiator	38.33%	13.26%	15.56%	17.87%	52.90%	13.97%	25.31%
21	Gone in 60 Seconds	27.44%	6.60%	2.89%	14.35%	7.41%	11.75%	11.74%
22	Green Mile, The	39.09%	13.38%	19.12%	50.90%	13.16%	9.78%	24.24%
23	Hanging Up	3.92%	29.94%	19.48%	21.07%	56.07%	42.82%	28.88%
24	Here on Earth	78.70%	18.03%	25.74%	58.91%	11.47%	45.44%	39.71%
25	High Fidelity	19.83%	4.45%	11.95%	17.44%	24.31%	17.98%	15.99%
26	Hurricane, The	38.44%	14.65%	14.47%	13.30%	7.48%	4.45%	15.46%
27	Keeping the Faith	20.56%	31.70%	2.15%	6.51%	3.50%	10.90%	12.55%
28	Kid, The	20.43%	21.23%	9.33%	8.17%	24.55%	12.67%	16.06%
29	Love and Basketball	28.83%	13.25%	19.63%	13.32%	2.60%	10.72%	14.73%
30	Me, Myself & Irene	0.09%	13.42%	2.93%	10.91%	16.41%	13.44%	9.53%
31	Mission to Mars	38.02%	27.01%	15.52%	5.89%	14.72%	23.92%	20.85%
32	Mission: Impossible 2	10.66%	44.26%	16.85%	21.95%	17.68%	9.04%	20.07%
33	My Dog Skip	80.30%	44.58%	35.11%	14.35%	21.87%	19.99%	36.04%
34	Next Best Thing, The	53.62%	2.21%	14.61%	56.52%	100.00%	7.79%	39.13%
35	Next Friday	71.96%	71.07%	27.11%	8.64%	17.26%	0.72%	32.79%
36	Ninth Gate, The	0.36%	22.98%	16.27%	44.74%	61.79%	25.43%	28.59%
37	Patriot, The	59.69%	5.48%	11.48%	6.46%	11.22%	22.17%	19.42%
38	Perfect Storm, The	1.15%	9.54%	4.43%	9.26%	20.97%	18.50%	10.64%
39	Pitch Black	40.21%	35.95%	0.98%	17.47%	35.25%	45.79%	29.28%
40	Reindeer Games	66.64%	1.66%	8.79%	21.43%	49.92%	90.43%	39.81%
41	Return to Me	25.73%	4.28%	24.74%	8.08%	1.93%	14.95%	13.28%
42	Road to El Dorado, The	59.11%	22.19%	21.62%	32.50%	64.47%	37.73%	39.60%

43 Road Trip	8.39%	28.66%	33.35%	7.97%	3.07%	32.48%	18.99%
44 Romeo Must Die	17.86%	20.80%	29.89%	0.28%	2.14%	8.11%	13.18%
45 Rules of Engagement	70.89%	23.20%	26.97%	0.77%	16.44%	16.97%	25.87%
46 Scary Movie	48.01%	3.46%	15.25%	19.23%	22.05%	0.92%	18.15%
47 Scream 3	0.45%	21.07%	8.76%	7.04%	4.28%	1.09%	7.12%
48 Shaft	35.47%	17.70%	37.29%	8.38%	8.59%	38.39%	24.30%
49 Shanghai Noon	21.46%	38.14%	5.85%	6.24%	14.94%	69.87%	26.08%
50 Skulls, The	8.37%	4.64%	3.80%	11.12%	32.09%	10.01%	11.67%
51 Snow Day	1.11%	42.64%	21.72%	22.26%	0.04%	7.18%	15.83%
52 Stuart Little	30.94%	24.92%	56.42%	0.00%	28.53%	63.79%	34.10%
53 Talented Mr. Ripley, The	0.81%	35.19%	10.53%	0.18%	37.07%	5.10%	14.81%
54 Tigger Movie, The	36.99%	48.46%	2.43%	17.42%	3.76%	10.65%	19.95%
55 Titan A.E.	100.00%	39.09%	43.38%	36.59%	16.23%	50.52%	47.64%
56 Toy Story 2	48.07%	9.18%	12.11%	12.59%	29.29%	45.66%	26.15%
57 U-571	20.09%	1.53%	2.92%	11.97%	18.32%	4.21%	9.84%
58 Where the Heart Is	70.47%	7.42%	32.15%	17.76%	11.37%	33.42%	28.77%
59 Whole Nine Yards, The	10.84%	1.53%	22.21%	20.95%	10.15%	9.36%	12.51%

TOTAL	36.73%	20.62%	19.43%	20.60%	24.64%	26.83%	24.81%

Appendix 4: Forecast results using the complete recalibration method for the 59 movies in the test sample

		Percentage error						
#	Movie Title	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Average
1	28 Days	64.29%	7.69%	29.21%	10.50%	0.07%	54.11%	27.65%
2	Adventures Rocky Bullwinkle	100.00%	6.95%	0.45%	95.97%	18.48%	47.34%	44.86%
3	Any Given Sunday	1.63%	27.74%	13.97%	31.29%	82.89%	40.45%	33.00%
4	Battlefield Earth	100.00%	72.63%	24.30%	64.86%	12.63%	3.49%	46.32%
5	Beach, The	7.32%	14.26%	19.97%	16.62%	17.67%	0.67%	12.75%
6	Bicentennial Man	90.27%	22.31%	32.59%	46.53%	24.64%	100.00%	52.72%
7	Big Momma's House	23.96%	5.88%	1.08%	7.91%	4.85%	2.05%	7.62%
8	Chicken Run	2.33%	13.28%	0.04%	5.38%	29.57%	8.27%	9.81%
9	Deuce Bigalow: Male Gigolo	6.72%	7.52%	8.44%	38.02%	27.80%	0.57%	14.84%
10	Dinosaur	44.76%	35.66%	100.00%	13.52%	16.18%	0.99%	35.18%
11	Down to You	10.02%	20.26%	21.54%	13.55%	81.09%	100.00%	41.08%
12	Drowning Mona	58.21%	5.25%	22.39%	36.94%	64.47%	100.00%	47.88%
13	Erin Brockovich	18.49%	1.78%	11.42%	3.45%	2.30%	9.97%	7.90%
14	Eye of the Beholder	33.43%	8.90%	39.95%	100.00%	100.00%	38.57%	53.47%
15	Final Destination	17.51%	15.00%	2.84%	1.62%	8.03%	20.99%	11.00%
16	Flintstones Viva Rock Vegas	67.48%	5.50%	9.79%	19.10%	0.59%	45.62%	24.68%
17	Frequency	54.60%	15.28%	4.57%	16.79%	11.37%	36.18%	23.13%
18	Galaxy Quest	100.00%	55.27%	67.44%	9.46%	100.00%	59.74%	65.32%
19	Girl, Interrupted	12.68%	37.91%	38.62%	30.66%	13.68%	77.39%	35.16%
20	Gladiator	38.33%	19.30%	11.67%	14.62%	53.42%	1.87%	23.20%
21	Gone in 60 Seconds	27.44%	1.61%	7.34%	7.15%	9.96%	16.88%	11.73%
22	Green Mile, The	39.09%	9.01%	3.74%	43.95%	13.23%	8.73%	19.62%
23	Hanging Up	3.92%	29.16%	4.98%	4.81%	44.19%	54.17%	23.54%
24	Here on Earth	78.70%	28.72%	19.64%	51.28%	9.13%	36.61%	37.35%
25	High Fidelity	19.83%	1.06%	4.19%	22.39%	14.81%	21.15%	13.91%
26	Hurricane, The	38.44%	6.63%	29.45%	19.82%	0.63%	3.40%	16.39%
27	Keeping the Faith	20.56%	29.77%	45.50%	26.49%	9.05%	19.38%	25.12%
28	Kid, The	20.43%	23.87%	29.20%	2.74%	26.56%	23.38%	21.03%
29	Love and Basketball	28.83%	7.63%	2.02%	0.55%	9.74%	17.79%	11.09%
30	Me, Myself & Irene	0.09%	13.41%	12.81%	14.93%	11.25%	16.79%	11.55%
31	Mission to Mars	38.02%	33.28%	0.84%	14.05%	6.35%	23.05%	19.27%
32	Mission: Impossible 2	10.66%	41.86%	40.19%	31.33%	26.14%	19.67%	28.31%
33	My Dog Skip	80.30%	39.48%	17.61%	67.75%	65.87%	62.77%	55.63%
34	Next Best Thing, The	53.62%	8.98%	20.36%	58.24%	100.00%	46.32%	47.92%
35	Next Friday	71.96%	41.47%	34.08%	26.99%	11.19%	15.06%	33.46%
36	Ninth Gate, The	0.36%	23.04%	2.43%	29.39%	57.16%	16.41%	21.46%
37	Patriot, The	59.69%	1.37%	0.04%	17.02%	3.36%	25.00%	17.75%
38	Perfect Storm, The	1.15%	9.70%	3.89%	14.65%	17.09%	12.52%	9.83%
39	Pitch Black	40.21%	25.90%	28.28%	2.04%	22.31%	48.18%	27.82%
40	Reindeer Games	66.64%	9.72%	2.54%	29.42%	62.82%	100.00%	45.19%
41	Return to Me	25.73%	0.95%	19.03%	18.75%	7.98%	8.84%	13.55%
42	Road to El Dorado, The	59.11%	16.60%	4.38%	17.86%	80.35%	82.54%	43.47%

43 Road Trip	8.39%	29.59%	75.30%	0.67%	4.69%	34.92%	25.59%
44 Romeo Must Die	17.86%	17.30%	7.24%	13.36%	12.89%	2.53%	11.86%
45 Rules of Engagement	70.89%	16.80%	0.55%	25.84%	38.69%	0.93%	25.62%
46 Scary Movie	48.01%	12.45%	7.15%	10.04%	16.89%	0.75%	15.88%
47 Scream 3	0.45%	21.15%	22.40%	0.70%	6.46%	2.32%	8.92%
48 Shaft	35.47%	10.24%	11.57%	8.65%	4.39%	27.21%	16.26%
49 Shanghai Noon	21.46%	33.25%	31.14%	19.03%	3.63%	66.10%	29.10%
50 Skulls, The	8.37%	5.91%	6.27%	11.39%	30.00%	21.08%	13.84%
51 Snow Day	1.11%	42.73%	94.95%	47.91%	15.00%	0.77%	33.74%
52 Stuart Little	30.94%	21.83%	41.98%	34.78%	7.44%	81.63%	36.43%
53 Talented Mr. Ripley, The	0.81%	35.26%	29.22%	22.03%	57.26%	17.74%	27.05%
54 Tigger Movie, The	36.99%	45.98%	89.12%	58.53%	20.14%	2.75%	42.25%
55 Titan A.E.	100.00%	54.86%	21.25%	20.40%	20.72%	56.75%	45.66%
56 Toy Story 2	48.07%	1.00%	26.42%	19.39%	34.08%	53.78%	30.46%
57 U-571	20.09%	1.82%	1.92%	15.00%	22.02%	13.11%	12.32%
58 Where the Heart Is	70.47%	0.26%	21.22%	3.60%	1.23%	42.84%	23.27%
59 Whole Nine Yards, The	10.84%	0.00%	19.90%	15.78%	13.01%	18.37%	12.99%

TOTAL	36.73%	19.53%	22.04%	24.16%	26.91%	31.70%	26.84%