Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

When Kerry Met Sally: Politics and Perceptions in the Demand for Movies

Jason M. T. Roos, Ron Shachar

To cite this article:
Jason M. T. Roos, Ron Shachar (2014) When Kerry Met Sally: Politics and Perceptions in the Demand for Movies. Management Science 60(7):1617-1631. http://dx.doi.org/10.1287/mnsc.2013.1834

MANAGEMENT SCIENCE
Vol. 60, No. 7, July 2014, pp. 1617–1631
ISSN 0025-1909 (print) | ISSN 1526-5501 (online)

http://dx.doi.org/10.1287/mnsc.2013.1834

When Kerry Met Sally: Politics and Perceptions in the Demand for Movies

Jason M. T. Roos
Rotterdam School of Management, Erasmus University, 3000 DR Rotterdam, The Netherlands, roos@rsm.nl

Ron Shachar
Arison School of Business, Interdisciplinary Center Herzliya, Herzliya, 46150 Israel, ronshachar@idc.ac.il

Movie producers and exhibitors make various decisions requiring an understanding of moviegoer's preferences at the local level. Two examples of such decisions are exhibitors' allocation of screens to movies and producers' allocation of advertising across different regions of the country. This study presents a predictive model of local demand for movies with two unique features. First, arguing that consumers' political tendencies have an unutilized predictive power for marketing models, we allow consumers' heterogeneity to depend on their voting tendencies. Second, instead of relying on the commonly used genre classifications to characterize movies, we estimate latent movie attributes. These attributes are not determined a priori by industry professionals but rather reflect consumers' perceptions, as revealed by their moviegoing behavior. Box-office data over five years from 25 counties in the U.S. Midwest provide support for this model. First, consumers' preferences are related to their political tendencies. For example, we find that counties that voted for congressional Republicans prefer movies starring young, Caucasian, female actors over those starring African American, male actors. Second, perceived attributes provide new insights into consumers' preferences. For example, one of these attributes is the movie's degree of seriousness. Finally, and most importantly, the two improvements proposed here have a meaningful impact on forecasting error, decreasing it by 12.6%.

Keywords: marketing; leisure industries; new products; forecasting
History: Received April 27, 2011; accepted September 15, 2013, by Pradeep Chintagunta, marketing. Published online in *Articles in Advance* February 24, 2014.

1. Introduction

Predicting box-office success still excites the media, academic scholars, and, of course, movie producers, exhibitors, and other executives involved in the motion picture industry. For example, a CNN blogger recently cited an expert, saying, "Everyone would love to have a crystal ball to predict box office success. If someone came up with a crystal ball like that, they would be a billionaire" (LaMagna 2012). That said, we have come a long way in this direction in recent decades; scholars have discovered various box-office predictors ranging from critics rating and script analysis to *Wikipedia* activity (Boatwright et al. 2007, Eliashberg et al. 2013, Mestyán et al. 2013). However, almost all these advances have been made at the national level. Until recently, local variation in box-office returns was largely ignored.[1]

Yet predicting box-office outcomes at the local level is critically important because many managerial decisions are made at this level. Consider the decision of a local exhibitor on the number of screens for each movie. For example, say that the exhibitor has three screens with the same audience capacity on which she intends to present two movies. She certainly prefers to allocate two screens to the movie with higher appeal to fill as many seats possible and maximize her profit. But which film appeals the most to her local audience? This is not a trivial question, because as we show here, local variation is not only meaningful but also systematic. For example, we show that in our data there is a sharp difference in movies' popularity between Republican- and Democratic-leaning counties. Furthermore, the allocation of screens to movies is not the only important decision with local-level considerations. Another such decision, discussed in §5.2, relates to the allocation of movies' advertising budget across various regions of the country.

Thus, a model that can predict local box-office outcomes should be useful and instrumental for managers in this industry. Predicting these outcomes is a challenging task because each movie is unique, it is difficult to compare movies, and thus one cannot easily rely on the success of previous movies in predicting the fate of a new one (Eliashberg et al. 2000).

[1] Recent advances along these lines include Davis (2005, 2006), Venkataraman and Chintagunta (2008), and Chintagunta et al. (2010).

This study presents such a predictive model that not only fits the data well but also does a good job predicting the performance of movies in the holdout sample. The successful predictions are due to two unique features of our model. The first is the inclusion of political explanatory variables. Specifically, as in previous studies (Berry et al. 1995), we formulate the demand for movies as a function of the match between the products' attributes and consumers' preferences. Unlike previous studies, we allow consumers' preferences to depend not only on their sociodemographic and unobservable characteristics but also on their political characteristics (e.g., turnout rates and vote shares at the county level). Furthermore, we argue that the predictive value of the political characteristics should go beyond the movie industry and that they should be used more often in marketing models.

The second unique feature of our model is remodeling the movies' attributes. Specifically, whereas previous studies have relied on predetermined movie genres, we estimate movie attributes as they are perceived by the consumers. We show that estimating movies' attributes (rather than assigning them a priori) is both more insightful and more effective (to practitioners). Furthermore, although we apply this modeling strategy at the local level, our findings suggest that such an approach is likely to be insightful at the national level as well.

The rest of the introduction is organized as follows. First, we present the rationale behind (1) the predictive power of political data and (2) the advantages of estimated attributes over predetermined genres. Second, we briefly describe our data, model, and main empirical results. Third, we discuss the relevant literature.

1.1. Rationale Behind the Predictive Power of Political Data

Political information might be predictive of consumers' choices for at least two reasons. First, it is correlated with individuals' personalities, which play a significant role in their choices. Specifically, recent work has shown that political variables and personality traits are closely related (Gerber et al. 2010, Mondak et al. 2010), and, of course, it is well established that personality traits play an important role in various consumption decisions (Kassarjian 1971, Baumgartner 2002). Thus, political variables can reflect personality information that may be useful in explaining consumption behavior. Second, individuals express their identity via their political choices (Granberg and Holmberg 1990) and through their consumption (Belk 1988), including their choice of movies (Chernev et al. 2011). Thus, political variables can reflect the identity that the individual desires to express, and this information might be valuable in predicting consumption behavior. For example, an individual who wishes to express himself as a rugged individualist might prefer movies with protagonists with similar characteristics; such a consumer might also vote for candidates who are portrayed in a similar light. Although we do not present a full theoretical framework for the predictive power of the political data, the two mechanisms highlighted here certainly suggest that such data may be valuable because they are reflecting something that sociodemographic variables by themselves cannot: consumer's personalities and the identities they wish to express.

Another reason for the possible added value of the political data can be attributed to its superiority as a measure. Researchers frequently rely on sociodemographic measures that are collected by the Bureau of the Census once per decade. Political data, on the other hand, are "collected" every two years. Given that the demographic composition of counties changes all the time, it is possible that political data reflect such changes and thus draw a much more precise picture of the counties. Furthermore, although the sample used by the Bureau of the Census is large, it is still a sample. The results of the election are not a sample in the sense that every eligible voter is measured—after all, even abstaining from voting is a choice. Last, we note that individual states typically report election results by precinct (which on average represents about 1,200 voters).[2] Such levels of disaggregation are clearly suitable for many marketing decisions.

Although we use the arguments above to suggest the predictive power of political variables in the demand for movies, they can also be used to claim that such variables are likely to be effective in predicting consumers' taste in variety of marketing models. Interestingly, so far, with the exception of a small number of individual-level studies (e.g., Baumgarten 1975, Crockett and Wallendorf 2004)—none of which explains actual market outcomes—political data have not been found in the marketing literature.[3]

1.2. Rationale Behind the Advantages of Estimated Attributes over Predetermined Genres

Entertainment and artistic products, such as movies, are very different from, say, automobiles, for which it is (1) easy to identify the attributes that affect consumers' choices (e.g., miles per gallon, horsepower)

[2] See http://www.census.gov and http://www.eac.gov for more information about censuses and elections, respectively.

[3] Baumgarten (1975) has shown that political attitudes are related to innovativeness (i.e., early adopters), and Crockett and Wallendorf (2004) use an ethnographic approach to suggest that political ideologies normatively influence consumer decisions.

and (2) easy to "measure" each product on these attributes (e.g., the Honda Prelude B20A5 has 135 horsepower). In contrast, even an expert would find it hard to determine the level of romance in, say, the *Dark Knight*. For this reason, films are not characterized on continuous measures, but rather they are categorized by genres. The Internet Movie Database (IMDb), for example, categorizes the *Dark Knight* as action, crime, and thriller. Accordingly, previous studies have relied on such genres to describe movies. However, such descriptors have three disadvantages: (1) it is unclear that they describe all the main attributes that affect consumers' choices, (2) they are discrete rather than continuous, and (3) they are based on the coder's perception of the movie rather than the moviegoers'.[4] An attractive alternative to this approach is to estimate the attributes rather than define them a priori. In other words, one can estimate the perception of products' attributes from the market results (e.g., correlations in popularity across heterogeneous markets), as Goettler and Shachar (2001) do in their analysis of the TV industry. Interestingly, the idea of using multidimensional scaling (MDS) to study movies has been suggested in the past (Wierenga 2006) but not followed so far. A "structural MDS" approach, such as the one used by Goettler and Shachar, does not suffer from the disadvantages discussed above. Because such estimated attributes reflect the perceptions of moviegoers, we will refer to them below as "perceived attributes." This term is also more closely related to the term "perceptual maps" used frequently in marketing.

1.3. Model and Data

We assess (1) the predictive power of political data and (2) the advantages of perceived attributes over predetermined genres using data provided by an anonymous exhibitor who operates in 25 counties across four states: Minnesota, Wisconsin, Illinois, and Ohio. The data include quarterly movie ticket sales spanning 21 quarters between 2000 and 2005. (We estimate our model using the first 14 quarters and reserve the last seven for holdout predictions.)

We model box-office performance in each market (i.e., county/quarter) as a function of the match between products' attributes and consumers' preferences. We allow these preferences to depend on (1) the county's sociodemographic characteristics (e.g., racial composition, income levels); (2) the county's political behavior (e.g., vote shares and turnout) in the presidential and congressional elections of 2000, 2002, and 2004; and (3) unobservable factors. We estimate two versions of this model, one with predetermined genres and the other with perceived attributes. In the predetermined genres version, movie characteristics are based on data from IMDb and the Motion Picture Association of America (MPAA). In the perceived attributes model, movies are located in a latent attribute space and tastes are represented as ideal points. We estimate each version of the model with and without political data and compare the predictive power of these specifications in the holdout sample to address our research questions.

1.4. Results

As expected, we find that including political data improves the performance of our models. The improvements we see in holdout predictions are meaningful. To illustrate them, we return to the example of the exhibitor we introduced above. Recall that for her screen allocation decision, she must estimate the number of ticket sales each movie will attract. The closer her prediction is to the actual number of tickets demanded, the better her decision. For the period of our analysis, the inclusion of the political variables closes the gap between predicted and actual number of tickets by 7.2 million tickets per year for the entire U.S. theatrical market.

We also show that political data can provide new insights about customer tastes. Indeed, many of the political variables used in our estimation have significant relationships with tastes for certain kinds of movies. For example, we find that counties that voted for congressional Republicans prefer movies starring young, Caucasian, female actors over those starring African American, male actors. Thus, political data provide new insights about the nature of heterogeneity in consumer tastes.

Remodeling movies' characteristics as perceived attributes rather than predetermined genres improves prediction even further. The gap between the predicted and actual number of tickets sold in the holdout sample decreases by another 26 million tickets. The predictive power of the unobserved perceived attributes in the holdout sample suggests that they are not a mere model contrivance; rather, it appears that they reflect real and fundamental aspects of movies that actually play an important part in consumers' movie choices.

Another indication that the perceived attributes reflect real and fundamental aspects of movies is the ease with which we can interpret them. For example, movies are differentiated on whether they are better suited for families or adults and whether they include

[4] Eliashberg and Sawhney (1994, p. 1168) point out that assuming consumers have tastes for such classifications is problematic: "Besides poor predictive power, genre preferences are too 'generic,' and cannot distinguish between different movies within the same genre. Further, movie critics (and obviously less expert average moviegoers) often disagree on the genre classification for a particular movie."

elements of science fiction and horror versus comedy and romance. Such distinctions seem quite natural. Furthermore, these attributes also convey information about movies that predetermined genres do not, such as cast demographics and whether the movie is "serious" or "light." Thus, our results show that not only do the seven perceived attributes we estimate predict box-office performance better than the 22 predetermined genres but they seem to be more insightful about the nature of movies. The straightforward nature of these dimensions will make it easier for managers to make use of this model.

Finally, in percentage terms, combining the improvements as a result of both the inclusion of the political data and the remodeling of movies' characteristics as perceived attributes decreases forecasting error for tickets sold by 12.6%.

1.5. Literature Review

The movie industry has received attention from marketing scholars for many good reasons (see, e.g., Eliashberg et al. 2006).[5] Until recently, however, none of these studies has looked at local (i.e., geographic) variation in demand for movies (i.e., the studies above have modeled aggregate demand only).[6] More recently, papers by Davis (2005, 2006), Venkataraman and Chintagunta (2008), Chintagunta et al. (2010), and Gopinath et al. (2013) have shown that demand variation at this level is important. Here, we contribute to this growing part of the literature by showing that both political data and perceived attributes can greatly improve demand forecasts at the local level.

Although our study is the first to predict box-office performance by estimating the heterogeneous tastes of viewers with respect to the latent movie attributes, two earlier studies have also estimated movie characteristics rather than define them a priori. Jedidi et al. (1998) identify movie clusters using the decay of sales during the lifetime of the movies, and Peress and Spirling (2010) estimate movies' locations in a latent attribute space using critical reviews. These studies differ from ours in various ways. For example, neither of them (1) estimates viewers' preferences with respect to the movies characteristics nor (2) explains variation in box-office performance across markets. Note also that the locations estimated by Peress and Spirling (2010) are not based on the perceptions of moviegoers as in our study, but rather on those of professional movie reviewers.

The remainder of this paper is structured as follows: In §2 we describe the data used in our study, in §3 we present our model, and in §4 we discuss issues related to its estimation. We present our results in §5, and §6 concludes.[7]

2. Data

We now describe the three data sets used in the analysis: box-office returns, election results, and demographics.

2.1. Movie Data

An anonymous theater chain provided us with movie revenue data. The data are aggregated by quarter, spanning 21 such periods between 2000 and 2005, and cover theaters in 25 counties across four U.S. states: Minnesota, Wisconsin, Illinois, and Ohio. We report results using the first 14 quarters for estimation (the training sample) and the remaining 7 periods for predictions (the holdout sample).

We know the names and gross quarterly revenues of the top 20 performing movies at each theater in each period. We further aggregate these theater-level data by county because this is the unit of observation in the demographic and political data. Because of the large number of movies (1,075), we focus our attention on films that were exhibited in at least 16 counties (a subset of movies accounting for about 90% of revenues in our data). The resulting data set contains 354 unique movies. Because many movies were exhibited in more than one quarter, the number of unique movie–period combinations is 744 and the total number of observations—i.e., the combination of movie, time, and county—is 9,926.

Some movies were not exhibited in every county, raising the issue of selection. If and how such selection affects the estimates is not immediate, and resolving this issue is beyond the scope of this paper. Our estimation should therefore be interpreted as if it is for the subsample of all the observations for which the exhibitor found the movie potentially attractive to her local audience.

[5] This attention has produced a number of important insights and tools in the areas of forecasting aggregate demand (e.g., Sawhney and Eliashberg 1996, Neelamegham and Chintagunta 1999, Swami et al. 1999, Eliashberg et al. 2000, Simonoff and Sparrow 2000, Swami et al. 2001, Sharda and Delen 2006) and timing new releases (e.g., Krider and Weinberg 1998, Ainslie et al. 2005, Einav 2007), and it has improved our understanding of how advertising (e.g., Elberse and Eliashberg 2003, Elberse and Anand 2007), critical reviews (e.g., Eliashberg and Shugan 1997, Basuroy et al. 2003, Ravid et al. 2006, Boatwright et al. 2007), and word of mouth (e.g., Moul 2007) influence aggregate demand for movies.

[6] Many researchers have paid attention to variation in demand at the micro (i.e., individual, screen, or theater) level (Eliashberg and Sawhney 1994, Swami et al. 1999, Eliashberg et al. 2000, Swami et al. 2001), but such studies have not employed variation in consumer behavior across theaters, as we do here, although such data are easier to obtain (for the decision maker) and can enrich predictions.

[7] Finally, note that the title of this paper is a play on the title of the 1989 romantic comedy *When Harry Met Sally* and the last name of the U.S. Democratic presidential candidate in 2004 (a year covered by our sample), John Kerry, which rhymes with "Harry." Thus, the title, like our study, joins politics and movies.

Table 1 **IMDb Genre Labels and MPAA Ratings in Our Data Set**

Genre	Movies	Genre	Movies	Genre	Movies	Genre	Movies	Genre	Movies	Rating	Movies
Comedy	164	Adventure	106	Crime	56	Horror	28	History	12	G	22
Drama	122	Romance	79	Sci-fi	48	Sport	20	Biography	9	PG	68
Thriller	119	Family	76	Mystery	43	Music	16	Western	4	PG-13	180
Action	111	Fantasy	58	Animation	31	War	13			R	84

Note. IMDb typically assigns more than one genre to each movie.

The movie exhibitor operates theaters in counties that vary widely in terms of their demographics and tastes for movies. Furthermore, we have good reasons to assume that the distribution of revenues across movies at this exhibitor is a close approximation to the distribution of revenues across the entire country. These reasons are (1) the list of movies presented by this exhibitor represents well the list of movies in the United States in the relevant years, and (2) the locations of its theaters within each county are diverse. (See additional information in the online technical appendix, available at http://ssrn.com/abstract=2395746.)

The total size of each county's market for movie tickets in each period is approximated by its spending on entertainment. Specifically, we use Bureau of Labor Statistics (BLS) data describing the "average annual expenditure on entertainment: fees and admissions" for the Midwest, as well as the population of each county, to calculate each county's total spending on entertainment per quarter.[8] This entertainment product category in the BLS data includes movies, sports events, and club memberships. As will become clearer after the presentation of the model, any definition of the total market has no effect on our substantive results (it merely shifts the estimated value of the non-movie outside alternative). For each county i, in each period t, we calculate the market share of each movie j by dividing its total revenue by total spending on entertainment. We denote these market shares as s_{ijt}.

Revenues are somewhat skewed across movies, with a small number of movies earning a lot and the remainder earning relatively little. For example, focusing on the median theater (in terms of total revenue), we find that half of its revenue was generated by a quarter of the movies. Revenues are also highly variable among theaters; the median film in each county earned as little as $1,205 and as much as $134,359. Clearly, the box-office performance of movies in this sample is highly heterogeneous, even after limiting our sample to a subset of the top movies. In other words, although we have limited our sample to just the most popular movies, we still see a striking degree of variation among movies and counties.

Finally, we collected information on movies' genres and ratings from IMDb and the MPAA, respectively.[9] The four ratings and 19 genres are listed in Table 1. For each movie j, we represent these attributes in the vector $x_j \in \{0, 1\}^{22}$ (we omit PG-13 as the base category). Table 2 provides summary statistics for revenues by movie and county.

2.2. Political Data

Our political data were compiled from public sources. Presidential election results for 2000 and 2004, as well as congressional election results for 2000, 2002, and 2004, were downloaded from state election websites.[10] We limit our attention to results for the Democratic and Republican parties. Each party's share of the vote is defined as its vote total divided by the total number of votes cast. The turnout rate is the total number of votes cast divided by the population aged 18 or older.

Our election data exhibit a great degree of variation across counties and elections. Turnout differs across counties by as much as 50% within a single election. The counties with the lowest and highest participation had turnout rates of 28% and 98%, respectively (both were for congressional ballots). Turnout also varies over time and was lowest for the 2002 congressional election (averaging 44% across all counties) and highest for the 2004 presidential election (averaging 75%).[11]

Counties also vary by their political preferences. Counties in our sample preferred Republicans over Democrats by a margin of 55%–45% across all elections. However, in terms of total votes, Democrats were preferred in each of the five elections in our sample by the majority of voters, reflecting the greater popularity of Democrats in areas with higher population density (e.g., Chicago). Counties also differ in

[8] Information collected from http://data.bls.gov/data (accessed August 12, 2008).

[9] See http://www.imdb.com and http://www.mpaa.org, respectively (accessed August 20, 2009).

[10] See http://www.elections.il.gov, http://electionresults.sos.state.mn.us, http://www.sos.state.oh.us, and http://gab.wi.gov (accessed July 28, 2008).

[11] This is the simple (not weighted) average across counties, which can explain part of the difference between this number and the national average of 60.1%, based on the United States Elections Project (http://elections.gmu.edu, accessed April 18, 2013).

Table 2 Summary Statistics for Select Demographic, Movie, and Political Data, Comparing the 25 Exhibitor Counties Used for Estimation with All Counties in the United States

	Exhibitor counties						U.S. counties
Factor	Min	25%	Median	Mean	75%	Max	Median
Population (1,000's)	31	95	160	500	410	5,400	25
Under 18 (%)	23	25	27	27	29	31	25
African American (%)	0.3	0.9	1.2	4.7	4.4	26	1.9
Per-capita income ($1,000's)	20	21	22	24	27	32	17
Bachelor's degree or higher (%)	13	20	24	26	34	42	14
Median age, female	32	34	35	36	37	40	39
Vote, Bush 2004 (%)	29	50	55	43	60	68	61
Presidential turnout, 2004 (%)	51	69	74	75	85	98	58
Movie revenue per county/quarter ($1,000's)	11	330	510	740	960	3,700	
Total revenue per movie ($1,000's)	140	510	770	1,000	1,200	5,200	

Note. As expected, counties with movie theaters tend to have greater populations and per-capita incomes than those without movie theaters, and their population is better educated.

how partisan they are, with some counties voting for one of the parties with great consistency. For example, the share of votes going to Republican candidates for Congress has a standard deviation across all elections ranging from 2% to 12% of total votes. So although some counties were highly persistent in their votes, and thus had little variation (2%), others were less consistent in their choices (e.g., the county with a standard deviation of 12%).

Because we have just 25 counties in our sample, we cannot use all of the political variables (in addition to the demographic variables presented below). Therefore, we must reduce the dimensionality of our data. We use factor analysis for this purpose. We select a three-factor solution (the third and fourth eigenvalues are 1.19 and 0.53, respectively) and generate factor scores for each county. The three factors are as follows.

Political Factor 1: *Low Turnout*. This variable reflects variation in turnout in general, but it is most closely associated with turnout in the 2000 and 2004 presidential elections. Counties with lower turnout load positively on this factor. Because George Bush gained a higher share of the vote in counties with an unusually high turnout, his performance loads negatively on this factor.

Political Factor 2: *Congressional Democrats*. This variable captures a basic Democratic–Republican preference (congressional races, which load more positively on this factor, tend to be more partisan than presidential elections).

Political Factor 3: *Gore and Kerry*. Counties where Democratic presidential candidates Al Gore and John Kerry received an unusually high vote share (relative to the baseline level of turnout and partisanship captured by the first two factors) load positively on this factor.

Because these variables are generated through factor analysis, they have mean 0, have unit variance, and are orthogonal. The factor scores for each county i are represented by the 1×3 row vector p_i. Table 2 provides summary statistics for vote share and turnout in 2004.

2.3. Demographic Data

The demographic data comprise 19 variables gathered from the 2000 U.S. Census describing the age, gender, race, family status, income, geography, and education of the counties in our sample.[12] For the same reason we reduced the number of political variables (i.e., the small number of counties), we need to represent the demographic characteristics of counties with a small number of variables. We again rely on factor analysis for this purpose. We select a four-factor solution (the fourth and fifth eigenvalues are 1.43 and 0.87, respectively) and generate factor scores for each county. The four factors are as follows.

Demographic Factor 1: *Large Families, High Income*. This factor relates to the size and composition of families and to income. Counties loading high on this dimension have more married couples with many children and moderate-to-high income. Counties with more single-person households and poverty load low.

Demographic Factor 2: *African Americans, Low Income, Unmarried*. Factor 2 captures elements of race, income, and family composition. Counties loading high on this dimension have greater proportions of African Americans, Hispanics, and people identifying as multiracial; more poverty; greater reliance on public assistance income; fewer married couples; and more single mothers.

Demographic Factor 3: *Educated, Urban, High Income*. This factor reflects education, income, and urbanization. Counties loading high on this dimension

[12] http://www.census.gov/main/www/cen2000.html (accessed August 19, 2009).

are more urbanized, have greater shares of college-educated adults, and have moderate-to-high incomes.

Factor 4: *Older, Retired*. Factor 4 is associated with age. Counties with older populations—either retired or nearing retirement—load higher.

These variables also have mean 0, have unit variance, and are orthogonal. For each county i, we collect these variables in a 1×4 row vector denoted y_i. Table 2 provides summaries of some of the demographic variables that went into the four factors.

Our approach of using factor analysis to reduce the dimension of the demographic and political variables raises a concern that the political data may merely replace the information lost while generating the demographic factor scores. The following exercise demonstrates that this is not the case.

To identify potentially lost information, we have added a fifth factor for the demographic data. This additional factor is supposed to capture the most important information that might be lost by using a four-factor model. We regress this factor on the political factors and find none of the coefficients is significantly different from zero. This suggests that the political factors are not capturing lost information as a result of factorizing the demographic variables.

2.4. Preliminary Analysis

Before we formulate our model and use structural estimation to identify relationships between tastes for politics and movies, it makes sense to look at the raw data to see whether Republican- and Democratic-leaning counties prefer different types of movies. To this end, we generate a score indicating which counties are the most or least Republican-leaning and consider all sets of the most or least Republican-leaning counties ranging from the top 1 to top 10 counties.[13] For each set of counties, we identify the top 10 movies performing unusually well after controlling for their average popularity across counties and time. In every case, no movie appeared in the top 10 lists for both the most and the least Republican-leaning counties. By way of comparison, if the ordering of counties were random (rather than based on the shares of votes for the Republican candidates), the probability of finding no overlapping movies in all sets of counties would be 0.06. Table 3 lists the top 10 movies based on this criterion for the three most and three least Republican-leaning counties.

Table 3 **Top 10 Disproportionately Successful Movies in the Three Most Republican-Leaning and Three Most Democratic-Leaning Counties**

Rank	Three most Republican counties	Three most Democratic counties
1	*Pirates of the Caribbean*	*Fahrenheit 9/11*
2	*Star Wars: Episode II: Attack of the Clones*	*The Talented Mr. Ripley*
3	*The Matrix Reloaded*	*A Beautiful Mind*
4	*Shrek 2*	*The Beach*
5	*Chicago*	*Road Trip*
6	*The Lord of the Rings: The Return of the King*	*Gladiator*
7	*Spider-Man*	*Cold Mountain*
8	*Harry Potter and the Sorcerer's Stone*	*The Last Samurai*
9	*Harry Potter and the Chamber of Secrets*	*Kill Bill: Vol. 2*
10	*Finding Nemo*	*Love Actually*

There are clear differences between the two lists of movies in Table 3. Republicans seem to have preferred action-adventure and children's movies, whereas Democrats seem to have preferred dramas and thrillers. Interestingly, half of the entries in the Republican list are sequels, perhaps because action-adventure and children's movie tend to produce movie franchises. This preliminary analysis, crude though it may be, lends support to our idea that there may be correlation in tastes for politics and movies.

3. Model

This section describes a model of the market share of movies in each county. In formulating our model, we allow preferences for movies to be correlated not only with demographic characteristics but also with preferences for political candidates as reflected by voting behavior.

Our formulation of market shares follows the vast literature in industrial organization and marketing (starting with Berry et al. 1995) that allows for a match between products' attributes and heterogeneous consumers' preferences, with a rich structure of unobservable taste components (e.g., an unobserved random match between a specific county and a particular movie). We deviate from this common approach in one significant way: we use an ideal point setting in a latent attribute space to model the match between movies and individuals (i.e., counties), and accordingly, we estimate the perceived movies attributes rather than define them a priori.

We described the potential advantages of the perceived attributes (over predetermined ones) in §1. The rationale behind formulating the match as an ideal point is that such a setting seems well suited for entertainment products in general, and movies in particular, because individuals frequently have different views about the optimal level of the attributes of these products (e.g., how much action or romance should

[13] For each county and election period, we calculate the difference between the share of votes going to Republicans and the average vote share earned by all Republicans in that election; we then sum these differences across elections to create a score that is higher (lower) in counties that consistently voted for (against) Republicans.

appear in a movie?). To assess the value of perceived attributes in describing movies, we wish to compare it with the standard approach—i.e., using random effects and predetermined attributes—and we will sometimes refer to this standard approach as the "benchmark model." In other words, we formulate and estimate two versions of the model: in one version, the match follows the standard approach, and in the other, it is based on an ideal point, perceived attribute framework. We compare the performance of our approach with this baseline and demonstrate the usefulness of political data in both settings.

We begin with a discussion of the portion of our model common to both approaches (i.e., all elements of the model other than the match) and then describe both formulations of the match: (1) the standard approach and (2) our approach.

3.1. Movie Demand

Each film is uniquely indexed by $j = 1, \ldots, J$, where J is the total number of unique movies offered across all n counties and T time periods. We denote the set of all movies exhibited in county i at time t by $\mathcal{J}_{it}$. Conditional on $j \in \mathcal{J}_{it}$, the market share of movie j in county i at time t is

$$s_{ijt} = \frac{\exp(u_{ijt})}{\exp(u_{i0t}) + \sum_{k \in \mathcal{J}_{it}} \exp(u_{ikt})}, \tag{1}$$

where u_{ijt} denotes the "mean utility" from movie j in county i at time t and is specified as

$$u_{ijt} = \eta_{jt} + \xi_i + \delta_{ij} + \epsilon_{ijt}. \tag{2}$$

The normalized mean utility from the outside good (indexed by $j = 0$) is $u_{i0t} = \epsilon_{i0t}$. One can think of this formulation as resulting from a more fundamental structure at the individual level (hence we speak of mean utility in this model). In such a case, the utility of each individual consumer h is $u_{ijt} + v^h_{ijt}$, the v^h_{ijt}'s follow a Type I extreme value distribution, and we assume consumer h chooses the option with the highest utility. Because we do not have data at the individual level, however, we formulate the market share directly.

Movies are differentiated both horizontally and vertically. The vertical attribute denoted by η_{jt} does not imply a high degree of achievement on some cultural scale, but rather a high level in terms of overall execution (e.g., good directing, an attractive cast). Because some movie genres may be more popular than others, we allow η_{jt} to depend on the observable features of movies, x_j. Moreover, for each movie, we allow η_{jt} to also vary over time to capture the "hipness" effect of new movies. Accordingly, we specify the following distribution for η_{jt}:

$$\eta_{jt} \sim \mathcal{N}(\psi_{r(j,\,t)} + x_j \phi^{\eta}, \sigma^2_{\eta}), \tag{3}$$

where x_j is a 1×22 row vector containing the observable attributes of movie j (MPAA ratings and IMDb genres), and the index $r(j, t)$ indicates the number of periods movie j has been exhibited as of period t.

Although we provide a rich specification for vertical differentiation, the horizontal variation is the focus of this paper. Such a match could reflect, for example, greater preference for family movies and lower preference for R-rated movies in a county with a high proportion of children. We consider two alternative ways to formulate δ_{ij}, which we describe in the following subsections.

Finally, the utility function also includes two random variables, ξ_i and ϵ_{ijt}. The first, ξ_i, is county specific and might be related to the quality of the moviegoing experience in different locations—for example, the amenities at the theater, the cost of parking—or they might reflect differences in the quality of the outside option in each county. The second, ϵ_{ijt}, is a county-movie-time effect that is observed by the individuals in the county but not by the researcher. This random variable accounts for the effects of various unobservables, including price and advertising (Elberse and Anand 2007).

3.2. Match Between Movies and Counties: Predetermined Genres (Benchmark)

We first introduce the benchmark model in which the match between movies and viewers is a function of predetermined attributes and county-specific preferences (i.e., random effects; see Neelamegham and Chintagunta 1999, Venkataraman and Chintagunta 2008). Formally stated, it is

$$\delta_{ij} = x_j \beta_i, \tag{4}$$

where the row vector x_j contains the observable attributes of the movie (e.g., science fiction indicator), and the column vector β_i is the county's specific taste for these attributes. The preferences parameter, β_i, is a function of both the county's observable and unobservable characteristics:

$$\beta_i' \sim \mathcal{N}(y_i \bar{\beta}^{\text{demo}} + p_i \bar{\beta}^{\text{pol}}, \Sigma^{\beta}). \tag{5}$$

The row vectors y_i and p_i contain the four demographic and three political factor scores for county i. The matrices $\bar{\beta}^{\text{demo}}$ and $\bar{\beta}^{\text{pol}}$ (4×22 and 3×22, respectively) represent differences in tastes among counties with different demographic and political compositions, and Σ^{β} is a 22×22 matrix reflecting covariation in unobserved preferences for predetermined genres and MPAA ratings.

Venkataraman and Chintagunta (2008) have already shown that interactions between demographic variables (in their study, income, African American, and

Hispanic) and genres account for a significant portion of heterogeneity in demand for movies. Thus, we expect $\bar{\beta}^{\text{demo}}$ to contain at least some nonzero entries. One of the main themes of this study is that some of the parameters in the $\bar{\beta}^{\text{pol}}$ matrix are also different from zero. In other words, we suggest that the political variables can improve our understanding of the preferences of a county for predetermined movie genres, even after accounting for the demographics of the county.

3.3. Match Between Movies and Counties: Perceived Attributes

This subsection formally presents our suggested formulation of δ_{ij}—i.e., an ideal point structure over a K-dimensional latent attribute space. We formulate the match between county i and movie j as a decreasing function of their squared distance in the latent attribute space (Goettler and Shachar 2001):

$$\delta_{ij} = -(z_j - \nu_i)(z_j - \nu_i)'. \tag{6}$$

The $1 \times K$ row vector ν_i denotes county i's K-dimensional ideal point, whereas z_j denotes the K-dimensional location of movie j. Under this formulation, movies that are closest to county i's ideal point in the latent space have the highest utility.

The predetermined genres have their flaws, but we believe they might have some informative value. We therefore relate movie locations to predetermined genres through the following distributions:

$$z_j \sim \mathcal{N}(x_j \phi^z, \Sigma^z), \tag{7}$$

where ϕ^z is a $22 \times K$ matrix of parameters relating each genre to the K latent attributes, and Σ^z is a $K \times K$ diagonal variance matrix.[14] Note that movie locations in our model are constant over time.

In the perceived attributes model, consumers' preferences for movies are represented by ν_i. A central point of this study is that political data can help us understand these preferences over and above the information contained in the demographic variables. We therefore assume the following relationship:

$$\nu_i \sim \mathcal{N}(y_i \gamma^{\text{demo}} + p_i \gamma^{\text{pol}}, \Sigma^{\nu}), \tag{8}$$

where γ^{demo} and γ^{pol} are $4 \times K$ and $3 \times K$ matrices relating demographic and political factors to ideal point locations, and Σ^{ν} is a diagonal variance matrix. We test various specifications of our models in which elements of γ^{demo} and γ^{pol} are set to zero.

To summarize, we have proposed two ways to formulate the match between counties and movies and to test the predictive power of political data. One formulation uses predetermined genres in a random effects setting, and the other relies on latent attributes in an ideal point setting. We estimate three versions of each of these models: one with political data, one with demographic data, and one with both.

4. Estimation Issues

We first describe the likelihood function and then our estimation and prediction procedures.

4.1. Likelihood Function

We build our likelihood function by first taking the log of each side of Equation (1) and then subtracting $\log s_{i0t}$. These operations yield the following expression:

$$\log s_{ijt} - \log s_{i0t} = u_{ijt} - u_{i0t} = \eta_{jt} + \xi_i + \delta_{ij} + (\epsilon^u_{ijt} - \epsilon^u_{i0t}). \tag{9}$$

We assume that differences in county-level unobservables for movies and the outside good follow a normal distribution; $\epsilon^u_{ijt} - \epsilon^u_{i0t} \sim \mathcal{N}(0, \sigma^2_u)$. Thus, conditional on the model parameters (including the unobservables in η_{jt} and δ_{ij}), differences in log shares follow a normal distribution:

$$\log s_{ijt} - \log s_{i0t} \mid \eta_{jt}, \xi_i, \delta_{ij}, \sigma^2_u \sim \mathcal{N}(\eta_{jt} + \xi_i + \delta_{ij}, \sigma^2_u). \tag{10}$$

After integrating over the unobservables in η_{jt} and δ_{ij}, the marginal likelihood for the model with predetermined genres is

$$\begin{aligned} L(\theta_{1,PG}) & \\ \propto \prod_i \prod_t \prod_{j \in \mathcal{J}_{it}} \int_{\epsilon^\eta_{jt}} \int_{\epsilon^\beta_i} & \mathcal{N}\big(\log s_{ijt} - \log s_{i0t} \mid \{\psi_{r(j,t)} + x_j \phi^\eta + \epsilon^\eta_{jt}\} \\ & + \xi_i + x_j \{y_i \bar{\beta}^{\text{demo}} + p_i \bar{\beta}^{\text{pol}} + \epsilon^\beta_i\}', \sigma^2_u\big) \\ & \cdot \mathcal{N}(\epsilon^\beta_i \mid 0, \Sigma^\beta) \mathcal{N}(\epsilon^\eta_{jt} \mid 0, \sigma^2_\eta)\, d\epsilon^\beta_i\, d\epsilon^\eta_{jt}, \end{aligned} \tag{11}$$

where $\theta_{1,PG} = \{\psi, \phi^\eta, \xi, \bar{\beta}, \Sigma^\beta, \sigma^2_\eta, \sigma^2_u\}$. Note, however, that we pursue a data augmentation approach during estimation and thus sample both the ϵ^β_i's and ϵ^η_{jt}'s directly. Moreover, we condition on the sampled ϵ^β_i's during prediction. In the model with perceived attributes, the marginal likelihood is

$$\begin{aligned} L(\theta_{1,PA}) \propto \prod_i \prod_t \prod_{j \in \mathcal{J}_{it}} \int_{\epsilon^\eta_{jt}} \int_{\epsilon^z_j} \int_{\epsilon^\nu_i} & \mathcal{N}\Big(\log s_{ijt} - \log s_{i0t} \\ & \mid \{\psi_{r(j,t)} + x_j \phi^\eta + \epsilon^\eta_{jt}\} + \xi_i \\ & - \big[(\{x_j \phi^z + \epsilon^z_j\} - \{y_i \gamma - \epsilon^\nu_i\}) \\ & \cdot (\{x_j \phi^z + \epsilon^z_j\} - \{y_i \gamma - \epsilon^\nu_i\})'\big], \sigma^2_u\Big) \cdot \mathcal{N}(\epsilon^z_j \mid 0, \sigma^2_z) \\ & \times \mathcal{N}(\epsilon^\nu_i \mid 0, \sigma^2_\nu) \cdot \mathcal{N}(\epsilon^\eta_{jt} \mid 0, \sigma^2_\eta)\, d\epsilon^z_j\, d\epsilon^\nu_i\, d\epsilon^\eta_{jt}, \end{aligned} \tag{12}$$

[14] We use diagonal variance matrices in defining our ideal point structures because we assume the tastes represented by the various dimensions are independent.

where $\theta_{1,PA} = \{\psi, \phi^{\eta}, \xi, \phi^{z}, \gamma, \sigma_z^2, \sigma_\nu^2, \sigma_\eta^2, \sigma_u^2\}$. Note also that we use a data augmentation approach here as well and thus sample the ϵ_j^z's, ϵ_i^ν's, and ϵ_{jt}^η's directly. As above, we condition on the sampled ϵ_i^ν's during prediction.

We denote the hyperparameters of the distributions of η, β, z, and ν (e.g., γ^{demo}) by θ_2 and define $\theta \equiv \{\theta_{1,PG}, \theta_2\}$ for the model with predetermined genres and $\theta \equiv \{\theta_{1,PA}, \theta_2\}$ for the model with perceived attributes. The marginal posterior distribution of the model parameters is proportional to the product of the marginal likelihood (Equation (11) or (12)) and various prior distributions: $p(\theta \mid s) \propto L(\theta \mid s)\pi(\theta)$.

4.2. Estimation Strategy

We estimate our model using the Metropolis-Hastings algorithm. In the online technical appendix, we discuss the prior distributions and normalizations needed for estimation. Note that the selection of distributions and normalizations are standard.

We also employ a number of reparameterizations that improve the efficiency of our Markov chain Monte Carlo sampler but have no bearing on our results. These reparameterizations, along with the full posterior conditional distributions used for sampling, are described in the online technical appendix. Finally, before proceeding with the estimation, we tested our model and estimation code using the method of Cook et al. (2006).

4.3. Predictive Distributions and Measures of Fit

We judge the relative performance of our models on the basis of their predictions of the holdout sample. To that end, we now describe the holdout sample, our measures of fit, and our approach to predicting market shares in the holdout sample.

4.3.1. Holdout Sample and Measures of Fit. The holdout sample for our main results includes all observations in the last 7 of the 21 periods. For the purpose of testing the results' robustness through cross validation, we use a shorter holdout sample. Specifically, in the cross validation, we replicate the estimation and prediction procedures 10 times with holdout samples of two periods. (Each period appears in no more than one holdout sample.)

We compare models using the root mean squared error (RMSE) of the movies' predicted market shares in each county/period in the holdout sample.[15] Furthermore, to provide a more intuitive comparison, we also calculate for each model the prediction error in terms of the number of tickets sold. For this purpose we use (1) our model and data and (2) data on box-office returns, and we calculate for each county, in each period, both the actual and the expected tickets sold.[16] The number that we eventually report is the sum over counties and time scaled to the entire U.S. market for the year 2004, during which 1.5105 billion tickets were sold.

4.3.2. Predictions in the Holdout Sample. The challenge of predicting holdout sample outcomes can be illustrated by two movies from the Bourne trilogy. *The Bourne Identity* is included in our training sample, and thus we have an estimate of its vertical attribute η_{jt} and (in the model with perceived attributes) its location z_j. But this is not the case for its sequel, *The Bourne Supremacy*, which only appears in the holdout sample; its perceived attributes and η_{jt} remain unobserved even after we estimate the model for the training sample. While we do not have an estimate of η_{jt} and z_j for movies in the holdout sample, we do have their predictive distributions (conditional on the parameters) and, of course, the posterior distributions of the parameters. Thus, to predict box-office performance, we integrate over the posterior distribution of the parameters the predictive distributions of η and z and the error term in the likelihood function (11 or 12). Letting the hat symbol denote variables pertaining to movies in the holdout sample, the expected market shares in the model with perceived attributes are given by the following expression:

$$
\begin{aligned}
&\mathbb{E}[\hat{s}_{ijt} \mid s] \\
&\quad = \int \frac{\exp\{\hat{\eta}_{jt} + \xi_i - (\hat{z}_j - \nu_i)(\hat{z}_j - \nu_i)' + \hat{\varepsilon}_{ijt}\sigma_u\}}{1 + \sum_{k \in \mathcal{J}_{it}} \exp\{\hat{\eta}_{kt} + \xi_i - (\hat{z}_k - \nu_i)(\hat{z}_k - \nu_i)' + \hat{\varepsilon}_{ikt}\sigma_u\}} \\
&\qquad \cdot \mathcal{N}(\hat{\varepsilon} \mid 0, 1)\, p(\hat{z}, \hat{\eta}, \theta \mid s)\, d\hat{\varepsilon}\, d\hat{z}\, d\hat{\eta}\, d\theta, \qquad (13)
\end{aligned}
$$

where s denotes the observed market shares in the training sample.

We approximate this expression through Monte Carlo integration. The full algorithm for predicting market shares is given in the online technical appendix, and here we provide a brief description of it. Conditional on the lth draw from the posterior distribution of the parameters, $\theta^{(l)}$, we sample $\hat{\eta}_{jt}^{(l)}$ from Equation (3) and $\hat{z}_j^{(l)}$ from Equation (7). Note that these predictions are defined conditional on knowing movies' observable attributes, x_j, prior to their release,

[15] To account for Monte Carlo error when comparing model fit, we calculate standard errors around the RMSEs via bootstrap resampling. Significance tests assume a normal distribution around the estimated RMSE. For cross validation, we calculate an average RMSE weighted by the square root of the number of holdout observations in each replication.

[16] To predict the number of tickets sold, we multiply predicted movie market shares by the size of the local market, then sum across all movies in a given period. We divide these numbers by their sum across all counties for the periods (quarters) used for this prediction and then multiply by the number of tickets sold in the United States.

which is reasonable in practice. Next, conditional on the draws of $\hat{\eta}_{jt}^{(l)}$ and $\hat{z}_j^{(l)}$, we sample predicted differences in the log market shares from Equation (10) and transform them into predicted market shares.

The above discussion demonstrates that, in a sense, we are using the same data, x_j, for holdout predictions in both models (predetermined genres and perceived attributes). How can the perceived attributes model do better in the holdout sample than the predetermine genres model if both are based on the same data? The only way for it to succeed is by capturing (in the training sample) something fundamental about viewers' behavior and using it in the holdout sample. We believe that this is possible by building δ_{ij} directly on individuals' perceptions rather than on industry experts coding. The empirical analysis will, of course, provide the only valid answer to this issue.

It is important to note that the perceived attributes model is likely to perform even better (than it does here) when used by practitioners. Specifically, as discussed in the next section, the perceived attribute dimensions we uncover are easy to interpret. Thus, we anticipate that film exhibitors, who can draw on their vast knowledge of movies, might do an even better job identifying the locations of new movies along the various perceived attribute dimensions (once identified via estimation of this model) when predicting market shares.

5. Results

Because of the predictive nature of this study, we start this section with a discussion of fit and prediction. This is followed by the managerial implications—and only then do we describe the estimates.

5.1. Fit and Predictions

This subsection starts with our benchmark model—the one with the predetermined genres. We show that the inclusion of the political variables improves the predictions of this model. We then move to the model with the perceived attributes and compare it with the benchmark. We find that it outperforms the benchmark model, and further, that even within this flexible modeling framework, the inclusion of the political variables is valuable.

We estimate the benchmark model with three different characterizations of counties. In the first case, we use factors based only on the demographic attributes of each county. This is the pure benchmark case—movies' attributes are the predetermine genres, and counties' characteristics are only the demographics variables. In the second case, the factors are based on the demographics and the political variables, and in the third, the factors are only based on the political variables.

Table 4 **RMSE of Predicted Market Shares (in Percent) in the Holdout Sample (Bootstrap Standard Errors in Parentheses)**

Data	Predetermined genres	Perceived attributes[b]
Demographic only	0.3136	0.2993
	(0.00030)	(0.00055)
Demographic and political	0.3143	0.2993
	(0.00038)	(0.00042)
Political only[a]	0.3128	0.2979
	(0.00030)	(0.00049)

Note. The p-values were derived assuming errors around the reported RMSEs are normally distributed (see §4.3.1).

[a]Models with only political data have lower prediction error than models with demographic data ($p < 0.05$).

[b]Models with perceived attributes have lower prediction error than models with predetermined genres ($p < 0.001$).

Comparing the models' predictions (see Table 4), we find that the models that include only political data outperform the benchmark model (i.e., the one with demographic information only). For the model with predetermined genres, this improvement is significant at the 5% level and corresponds to an improvement in forecasting error of 7.2 million tickets (in 2004). A robustness check via cross validation (as described above) confirms these findings. Note that the model with political data predicts better than the model with both political and demographic data.[17] This finding suggests that tastes for movie genres may be explained well enough by political characteristics and that adding demographic information does not improve the model's out-of-sample predictions. Of course, it is possible that the predictive power of the political data will not be so dramatic for other product categories.

Moving to the model with perceived attributes, Table 4 shows that the model with only political data provides better predictions ($p < 0.001$) than the model with only demographic data.[18] These results are also confirmed by cross validation. As in the case of the predetermined genres, the evidence illustrates

[17] Note that because our estimation is not based on minimizing a specific object—for example, the RMSE—the fit and certainly the predictions of the model with the political data only can be better than the one with both political and demographics.

[18] The number of latent dimension is seven for the models with only demographic or political data and eight for the model with both. The number of dimensions is model specific, as driven by our wish to be true to the way our model will be used by practitioners. Specifically, say that a manager adopts our approach and characterizes counties only based on their political attributes. She will determine the number of latent dimensions based on the training sample and, of course, employ this number for the holdout data. According to our estimation, in such a case the number would be seven on the basis of RMSE of market shares for the training sample. She would reach the same number if she adopts the more traditional approach and characterizes counties only by their demographics. However, if she decides to use both demographic and political data, the number of dimensions that she will end up using is eight.

the advantages of the political data and the redundancy of the demographic information when political data are included.

Having presented evidence demonstrating the benefits from using political data, we now turn to the value of remodeling movies characteristics as perceived attributes. From Table 4, it is evident that perceived attributes improve predictions for the models with both demographic and political data (these improvements are also confirmed by the cross validation). These differences in prediction error are not only statistically significant ($p < 0.001$) but meaningful—the RMSE of market shares decreases by 4.6% and the prediction error with respect to tickets sold improves by an additional 26 million (in 2004). This improvement in holdout predictions suggests that practitioners and scholars who wish to predict box-office outcomes should stop using the traditional approach and instead adopt the latent attributes modeling strategy.

5.2. Managerial Implications

The main purpose of this study is to provide practitioners with a predictive model to improve their decision making. Predicting box-office success is critical in this industry, because movies with greater market potential receive larger support and resources from producers and exhibitors. These resources include, among others, advertising money and distribution slots (i.e., screens allocated). Below we focus on the distribution issue, but before we do so, it is important to note that advertising money is a major factor in this industry, and on average, for each dollar spent on the production of a movie, 50 cents are spent on its advertising (Vogel 2007). The advertising budget is not distributed equally across various parts of the country, and thus a predictive model that identifies the regions that are best suited for a specific movie can be instrumental for executives in this industry.

Focusing on the distribution decision, a predictive model can improve exhibitors' allocation of screens. Consider the example in §1: an exhibitor has three screens with the same audience capacity on which she intends to present two movies. Say that the audience capacity of each screen is 250. Table 5 presents the actual demand and two scenarios, each corresponding to a different demand prediction. First, the actual demand is 300 tickets for movie A and 400 for movie B, and thus the best allocation of screens is one for movie A and two for B. With such an allocation, the exhibitor will sell 650 tickets (250 + 400). Of course, in practice, the exhibitor does not know the demand and thus makes a forecast. Such expectations are illustrated in Cases 1 and 2. Comparing these cases, we find that an improvement in the prediction that leads to a decrease of 20 tickets in the forecasting error (from 110 to 90) ends up increasing the number of tickets sold by 100 (from 550 to 650). The details of this calculation are included in the table, but the logic is simple: a small change in prediction is significant enough to change the exhibitor's screen allocation decision so that the third screen is allocated to movie B rather than A.

Table 5 **Improvement in Actual Sales and Forecast Error as a Result of Better Predictions in a Hypothetical Screen Allocation Problem**

	Movie A	Movie B	Total
Actual demand	300	400	
Best seat allocation	250	500	
Maximum tickets sold	250	400	650
Case 1—Inaccurate prediction			
Predicted demand	355	345	
Seats allocated	500	250	
Tickets sold	300	250	550
Forecast error	55	55	110
Case 2—Improved prediction			
Predicted demand	345	355	
Seats allocated	250	500	
Tickets sold	250	400	650
Forecast error	45	45	90

Note. Comparing Case 2 with Case 1, sales improved by 100 tickets whereas forecast error improved by 20 tickets.

This example illustrates the following sequence: (1) an improvement in prediction (and thus a decrease in the forecasting error), followed by (2) a change in the allocation of screens, leading to (3) an increase in the number of tickets sold. This means that including political variables and using perceived attributes in the model is likely to increase sales. Unfortunately, since we do not observe screen allocations, we cannot determine the improvement in sales in our data. However, our estimates enable us to determine the decrease in the forecasting error in terms of tickets. Specifically, we calculate for each county at each period both the actual and the expected number of tickets and the absolute difference between them—i.e., $|\text{Tickets}_{it} - \text{Predicted Tickets}_{it}|$. For the last year of our sample (i.e., the last four periods of the holdout sample), the decrease in the forecasting error is 7.2 million tickets due to political data and an additional 26 million due to perceived attributes; the combined improvement is a decrease in forecasting error of 12.6%. Note, however, that as pointed out above this number is not the increase in sales but rather the decrease in the forecasting error.[19]

[19] Although we have focused on predicting revenues in established markets for movies that have yet to be released, one can also use our model to predict revenues in new markets (e.g., when making entry decisions). It seems reasonable to expect that the power of political data will be even greater when used in this way. We thank an anonymous reviewer for highlighting this potential use of our model.

5.3. Estimates

We now briefly discuss the coefficient estimates for consumer preferences for movies.

5.3.1. Predetermined Genres. Recall that this model was estimated in three versions that differ in which variables were used to characterize individuals (demographic, political, or both). The $\bar{\beta}$ parameters (tastes for predetermined genres) and the posterior distributions of all coefficients are reported in the online technical appendix. In all cases, we find significant interactions between county-level descriptors and predetermined genres. On the whole, these interactions appear coherent. For example, in the demographic-only model, counties with larger families do not like R-rated movies, and in the model with only political data, R-rated movies are disliked in counties that prefer Republicans (for both congressional and presidential races). These results might give exhibitors and distributors new insights into which movies will perform better across different local markets.

5.3.2. Perceived Attributes. The model with perceived attributes not only outperforms the model with predetermined genres, it also provides a characterization of the movies in our sample that is both concise and insightful. Here, we briefly discuss our interpretation of the seven latent attributes revealed by the data. Latent attributes are listed in order of their importance (as measured by their standard deviations). Because the model with political data has the best fit, we present results from that model. (Our interpretation of the perceived attributes dimensions are the same when considering the model with only demographic data.) Our interpretation of these dimensions is aided by the various coefficient estimates, which are presented in the online technical appendix, as well as exploratory analysis of the movie at the extremes of the dimensions, including cast listings and trailers from IMDb.

Dimension 1: *Adult vs. Family*. This dimension, which has the highest standard deviation, differentiates movies on the basis of whether they are more suitable for families (e.g., *Rugrats Go Wild*, *Center Stage*, *The Adventures of Rocky & Bullwinkle*) or adults (e.g., *Chicago*, *Traffic*, *Gangs of New York*). It is encouraging to note that Peress and Spirling (2010) also identify a latent dimension that separated adult-oriented films from more family-friendly movies, as do Gazley et al. (2011) in a study factor analyzing stated preferences for predetermined genres (not individual movies). With respect to the ideal point locations, counties won by George Bush tend to have a stronger preference for family-oriented movies.

Dimension 2: *Demographics of Lead Actor*. On a superficial level, this dimension appears to separate thrillers from romantic dramas. However, closer inspection reveals that differences in cast member race and gender provide a more plausible explanation. Looking at the casts (as listed by IMDb), we find 15 of the 20 movies loading highest on one side feature African Americans in lead or supporting roles, compared with just 8 at the other end. More significant, 12 of those 15 had African American males in lead roles (e.g., *Big Momma's House*, *Nutty Professor II*: *The Klumps*, and *Training Day*, starring Martin Lawrence, Eddie Murphy, and Denzel Washington, respectively)—compared with just three at the other end.[20] We also find 5 of the 20 movies loading furthest from the African American cluster featured Caucasian teenage females in lead roles (e.g., *A Walk to Remember*, *Blue Crush*, and *What a Girl Wants*, starring Mandy Moore, Kate Bosworth, and Amanda Bynes, respectively)—but none at the other end. With respect to ideal points, we find counties that (1) voted more for George Bush and (2) had relatively low voter turnout tended to prefer movies without African American male lead actors.[21] Finally, even though this dimension has the second largest scale, and is therefore important to moviegoers, we note that IMDb does not identify the race of movie actors, nor does it label movies as being targeted to African Americans or Caucasians.

Dimension 3: *Light vs. Serious*. This dimension reflects differences between "light" or "easy" movies (e.g., *Two Weeks Notice*, *Driven*, *A Knight's Tale*), which are less intellectually and emotionally demanding, and "serious" or "emotional" movies (e.g., *The Beach*, *Vanilla Sky*, *Enemy at the Gates*), which are more intellectually or emotionally demanding. Interestingly, Peress and Spirling (2010), in a spatial analysis of movie reviews, find a similar dimension, which they characterize as action and adventure versus "deep" or emotional movies. The estimates indicate that the more serious movies are more popular among counties with higher support for congressional Democrats.

Other Dimensions:

- Dimension 4 differentiates movies with elements of science fiction and horror (e.g., *Blade II*, *The Cell*, *Hollow Man*) from those that are romantic, family-oriented, or funny (e.g., *Toy Story 2*, *The Princess*

[20] In general, "lead" means either the actor's name was billed above the movie title in promotional material (which was located through IMDb and Google Images) or he or she played a significant role in a cast without any identifiable star (i.e., was a member of an ensemble cast).

[21] Shachar and Emerson (2000) show that television audiences prefer casts with demographic characteristics similar to their own; however, when we estimate the model with both demographic and political data, demographic factor 2 (associated with higher proportions of African Americans) has a smaller and less statistically significant coefficient than the political factor representing preference for George Bush.

Diaries, *Shrek*). The former are more popular among counties voting for Democratic presidential candidates.

- Dimension 5 separates family-oriented comedies (e.g., *Kangaroo Jack*, *The Santa Clause 2*, *102 Dalmatians*) from R-rated dramas (e.g., *Vanilla Sky*, *American Beauty*, *The Hurricane*). The latter are more popular among counties supporting congressional Democrats.
- Dimension 6 differentiates thrilling mysteries from PG-rated movies (especially those involving sports).
- Dimension 7 differentiates science-fiction fantasy (including movies with comic-book heroes) from R-rated comedies.

Generally speaking, each of the seven dimensions we describe here has an identifiable, if not statistically significant, relationship with at least one genre, but on the whole, these relationships are weak: predetermined genres explain on average only half of the variation in movie locations. And yet despite these weak connections with predetermined genres, the latent attributes we have uncovered are easily interpreted and seem to make sense—both individually and collectively. As a result, we believe these will be easy for practitioners to work with. Specifically, identifying the likely location of a new movie along each of these seven axes for the purposes of predicting demand should be straightforward for exhibitors and distributors (thus potentially eliminating the need to use predetermined attributes entirely).

These results show that perceived attributes are useful, outperforming the predetermined genres when it comes to holdout predictions. Not only do we see statistical improvements, these improvements are likely to be economically significant as well. Indeed, the strength of this result is somewhat unexpected, because although we actually know precisely what the predetermined genres are for each of the movies in the holdout sample, we must "guess" their locations along each of the seven perceived attribute dimensions. And yet, despite this disadvantage, the model with perceived attributes makes much better predictions. We take this to be an indication that these perceived attributes capture significant, real, and fundamental aspects of what consumers actually see in movies. Furthermore, these latent attributes are easy to interpret, occur naturally, and offer insights above and beyond the predetermined genres.

6. Conclusion

This study presents a predictive model of box-office outcomes at the local level. Such a model can enhance various decisions of movie producers and exhibitors, such as allocating (1) screens to movies and (2) advertising spending across different regions of the country.

We believe that, on the top of this, the study makes two additional contributions. First, it highlights the predictive power of a previously unused source of data—political tendencies. We find that political data improve holdout predictions. Furthermore, we show that political data can reveal new insights into consumer tastes. For example, tastes for movies starring either African American men or Caucasian teenage girls interact more strongly with political variables than with demographic variables.

These results have important implications for marketing researchers. Marketers are comfortable thinking about customers in terms of their common demographic traits. We suggest political data, which are available at quite disaggregated levels (at the local precinct level in many states), updated every two years, and disseminated free of charge, provide a new way to characterize consumers. We also believe that political data may be useful for characterizing customers in other product categories—obvious candidates include books, video games, and other types of entertainment, as well as other industries such as apparel and maybe even automobiles.

The second contribution relates directly to the movie industry. Although previous studies have predicted box-office success using categorical variables of movie characteristics as determined by experts, we present a model in which movie attributes are based on the perceptions of moviegoers. It should come as no surprise, then, that perceived attributes improve out-of-sample predictions. In addition to improving predictions, the perceived attributes uncovered in our study are easy to interpret, providing some evidence that they may represent the way consumers actually think about movies.

Acknowledgments

The authors thank Preyas Desai, Ron Goettler, Joel Huber, Wagner Kamakura, Carl Mela, Rick Staelin, and Andrew Sweeting, who were kind enough to share some of their expertise. They are especially grateful to Bob Markley for his essential role in obtaining data for this project. J. M. T. Roos's work was sponsored by the BiG Grid project for the use of computing and storage facilities, with financial support from the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (Netherlands Organisation for Scientific Research, NWO). R. Shachar's work was financed by the Henry Crown Institute of Business Research in Israel and the Israel Science Foundation [Grant 1461/10].

References

Ainslie A, Drèze X, Zufryden F (2005) Modeling movie life cycles and market share. *Marketing Sci.* 24(3):508–517.

Basuroy S, Chatterjee S, Ravid SA (2003) How critical are critical reviews? The box office effects of film critics, star power, and budgets. *J. Marketing* 67(4):103–117.

Baumgarten SA (1975) The innovative communicator in the diffusion process. *J. Marketing Res.* 12(1):12–18.

Baumgartner H (2002) Toward a personology of the consumer. *J. Consumer Res.* 29(2):286–292.

Belk RW (1988) Possessions and the extended self. *J. Consumer Res.* 15(2):139–168.

Berry S, Levinsohn J, Pakes A (1995) Automobile prices in market equilibrium. *Econometrica* 63(4):841–890.

Boatwright P, Basuroy S, Kamakura W (2007) Reviewing the reviewers: The impact of individual film critics on box office performance. *Quant. Marketing Econom.* 5(4):401–425.

Chernev A, Hamilton R, Gal D (2011) Competing for consumer identity: Limits to self-expression and the perils of lifestyle branding. *J. Marketing* 75(3):66–82.

Chintagunta PK, Gopinath S, Venkataraman S (2010) The effects of online user reviews on movie box office performance: Accounting for sequential rollout and aggregation across local markets. *Marketing Sci.* 29(5):944–957.

Cook SR, Gelman A, Rubin DB (2006) Validation of software for Bayesian models using posterior quantiles. *J. Comput. Graphical Statist.* 15(3):675–692.

Crockett D, Wallendorf M (2004) The role of normative political ideology in consumer behavior. *J. Consumer Res.* 31(3):511–528.

Davis P (2005) The effect of local competition on admission prices in the US motion picture exhibition market. *J. Law Econom.* 48(2):677–707.

Davis P (2006) Spatial competition in retail markets: Movie theaters. *RAND J. Econom.* 37(4):964–982.

Einav L (2007) Seasonality in the US motion picture industry. *RAND J. Econom.* 38(1):127–145.

Elberse A, Anand B (2007) The effectiveness of pre-release advertising for motion pictures: An empirical investigation using a simulated market. *Inform. Econom. Policy* 19(3–4):319–343.

Elberse A, Eliashberg J (2003) Demand and supply dynamics for sequentially released products in international markets: The case of motion pictures. *Marketing Sci.* 22(3):329–354.

Eliashberg J, Sawhney MS (1994) Modeling goes to Hollywood: Predicting individual differences in movie enjoyment. *Management Sci.* 40(9):1151–1173.

Eliashberg J, Shugan SM (1997) Film critics: Influencers or predictors? *J. Marketing* 61(2):68–78.

Eliashberg J, Elberse A, Leenders M (2006) The motion picture industry: Critical issues in practice, current research, and new research directions. *Marketing Sci.* 25(6):638–661.

Eliashberg J, Hui S, Zhang JZ (2013) Assessing box office performance using movie scripts: A kernel-based approach. Working paper, The Wharton School, University of Pennsylvania, Philadelphia.

Eliashberg J, Jonker JJ, Sawhney MS, Wierenga B (2000) MOVIEMOD: An implementable decision-support system for prerelease market evaluation of motion pictures. *Marketing Sci.* 19(3):226–243.

Gazley A, Clark G, Sinha A (2011) Understanding preferences for motion pictures. *J. Bus. Res.* 64(8):854–861.

Gerber AS, Huber GA, Doherty D, Dowling CM, Ha SE (2010) Personality and political attitudes: Relationships across issue domains and political contexts. *Amer. Political Sci. Rev.* 104(1): 111–133.

Goettler RL, Shachar R (2001) Spatial competition in the network television industry. *RAND J. Econom.* 32(4):624–656.

Gopinath S, Chintagunta PK, Venkataraman S (2013) Blogs, advertising, and local-market movie box office performance. *Management Sci.* 59(12):2635–2654.

Granberg D, Holmberg S (1990) The intention-behavior relationship among US and Swedish voters. *Soc. Psych. Quart.* 53(1):44–54.

Jedidi K, Krider RE, Weinberg CB (1998) Clustering at the movies. *Marketing Lett.* 9(4):393–405.

Kassarjian HH (1971) Personality and consumer behavior: A review. *J. Marketing Res.* 8(4):409–418.

Krider RE, Weinberg CB (1998) Competitive dynamics and the introduction of new products: The motion picture timing game. *J. Marketing Res.* 35(1):1–15.

LaMagna M (2012) Can an equation predict box office success? *Marquee Blog* (blog), June 15, http://marquee.blogs.cnn.com/2012/06/15/can-an-equation-predict-box-office-success.

Mestyán M, Yasseri T, Kertész J (2013) Early prediction of movie box office success based on Wikipedia activity big data. *PLoS ONE* 8(8):e71226, doi: 10.1371/journal.pone.0771226.

Mondak JJ, Hibbing MV, Canache D, Seligson MA, Anderson MR (2010) Personality and civic engagement: An integrative framework for the study of trait effects on political behavior. *Amer. Political Sci. Rev.* 104(1):85–110.

Moul CC (2007) Measuring word of mouth's impact on theatrical movie admissions. *J. Econom. Management Strategy* 16(4): 859–892.

Neelamegham R, Chintagunta PK (1999) A Bayesian model to forecast new product performance in domestic and international markets. *Marketing Sci.* 18(2):115–136.

Peress M, Spirling A (2010) Scaling the critics: Uncovering the latent dimensions of movie criticism with an item response approach. *J. Amer. Statist. Assoc.* 105(489):71–83.

Ravid SA, Wald JK, Basuroy S (2006) Distributors and film critics: It takes two to tango? *J. Cultural Econom.* 30(3):201–218.

Sawhney MS, Eliashberg J (1996) A parsimonious model for forecasting gross box-office revenues of motion pictures. *Marketing Sci.* 15(2):113–131.

Shachar R, Emerson JW (2000) Cast demographics, unobserved segments, and heterogeneous switching costs in a television viewing choice model. *J. Marketing Res.* 37(2):173–186.

Sharda R, Delen D (2006) Predicting box-office success of motion pictures with neural networks. *Expert Systems Appl.* 30(2): 243–254.

Simonoff JS, Sparrow IR (2000) Predicting movie grosses: Winners and losers, blockbusters and sleepers. *Chance* 13(3):15–24.

Swami S, Eliashberg J, Weinberg CB (1999) SilverScreener: A modeling approach to movie screens management. *Marketing Sci.* 18(3):352–372.

Swami S, Puterman ML, Weinberg CB (2001) Play it again, Sam? Optimal replacement policies for a motion picture exhibitor. *Manufacturing Service Oper. Management* 3(4):369–386.

Venkataraman S, Chintagunta PK (2008) Investigating the role of local market and exhibitor characteristics on box-office performance. Working paper, University of North Carolina, Chapel Hill.

Vogel HL (2007) *Entertainment Industry Economics: A Guide for Financial Analysis* (Cambridge University Press, Cambridge, UK).

Wierenga B (2006) Motion pictures: Consumers, channels, and intuition. *Marketing Sci.* 25(6):674–677.