STATISTICAL METHODS IN ECOMMERCE RESEARCH

Chapter: Differential Equation Trees to Model Price Dynamics in Online Auctions

Wolfgang Jank
Department of Decision and Information Technologies, R. H. Smith School of Business, University of Maryland, College Park, MD 20742, USA

Galit Shmueli
Department of Decision and Information Technologies, R. H. Smith School of Business, University of Maryland, College Park, MD 20742, USA

Shanshan Wang
Modeling and Analytical Services, DemandTec Inc., San Carlos, CA 94070, USA

WILEY-INTERSCIENCE

A JOHN WILEY & SONS, INC., PUBLICATION

CHAPTER 1

DIFFERENTIAL EQUATION TREES TO MODEL PRICE DYNAMICS IN ONLINE AUCTIONS

1.1 INTRODUCTION

Empirical research of online auctions has been growing steadily in the last several years. Online auctions are different from offline auctions in several important ways: They are usually much longer, bidders and sellers are anonymous, and the barriers of entry are much lower for both bidders and sellers. These differences lead to auction dynamics that can be very different from those in offline auctions. One important aspect of these dynamics is their effect on the auction price. In this work, we are particularly interested in the price path of an online auction and its dynamics, that is, we are interested in the speed of the price increases and how it changes over the entire auction duration.

Within the rich empirical online auction literature there has been very little in the way of studying the price dynamics. Most price-related studies have focused on the final price alone. However, in a series of recent papers by Jank & Shmueli and co-authors [2, 8, 9, 10, 11, 18, 19], they show that dynamics matter, that even auctions for the same product can have very different price paths and dynamics [9], and that incorporating the information contained in the price dynamics of an ongoing auction greatly improves the

ability to forecast its final price [19]. In particular, [19] find that the availability of dynamics greatly improves the forecasting error compared to powerful competitors such as double exponential smoothing. Furthermore, the relationship between the price path and other auction-related information (e.g., seller rating, auction duration, opening bid, and item properties) changes during the auction [2]. One example is the effect of the opening bid on the price at different times during the auctions. [2] and [18] find that although there is a positive relationship between the price and the opening bid at any point in the auction, the strength of this relationship declines as the auction progresses, implying that bidders derive less and less information from the opening bid.

In order to estimate the price path and its dynamics from the discrete observed bids, Jank & Shmueli take a functional data analytic approach. In that approach, the price path of each auction is represented by a smooth continuous curve. The derivatives of this curve capture price dynamics: the first derivative captures the price velocity, indicating when the price increases fast and when the increase slows down. Similarly, the second derivative captures price acceleration.

The estimation of smooth continuous price curves is achieved via smoothing methods, as is customary in functional data analysis [17]. [20] build upon this and identify a family of differential equation models (DEM) that parsimoniously capture auction price dynamics. However, it is not quite clear how to incorporate external auction-related information into the DEM. In this chapter we build upon the work of [20] and show how regression trees can be extended to model the relationship between price dynamics and other auction-related variables. Specifically, we propose a novel tree-based approach for DEM based on recursive partitioning. We want to point out that this chapter is rather technical in that its primary focus is on developing methodology for differential equation trees.

The roadmap for the remainder of the paper is as follows (see also Figure 1.1.). We first describe online auction data and the features that make them amenable to functional data models. We then describe the steps necessary for identifying and fitting a DEM to auction data. These include data smoothing to create a smooth functional object in the first step. The smooth functional object allows for a gauging of the price dynamics. We describe a rather novel approach of exploring price dynamics via phase-plane plots. The observed price dynamics are our main motivation for studying DEMs. We describe the general nature of DEMs and their shortcomings in that they do not allow for an easy incorporation of external, non-price related covariates. This shortcoming is addressed via our new method-

ology of differential equation trees. We describe our method and illustrate it on a dataset of eBay auctions.

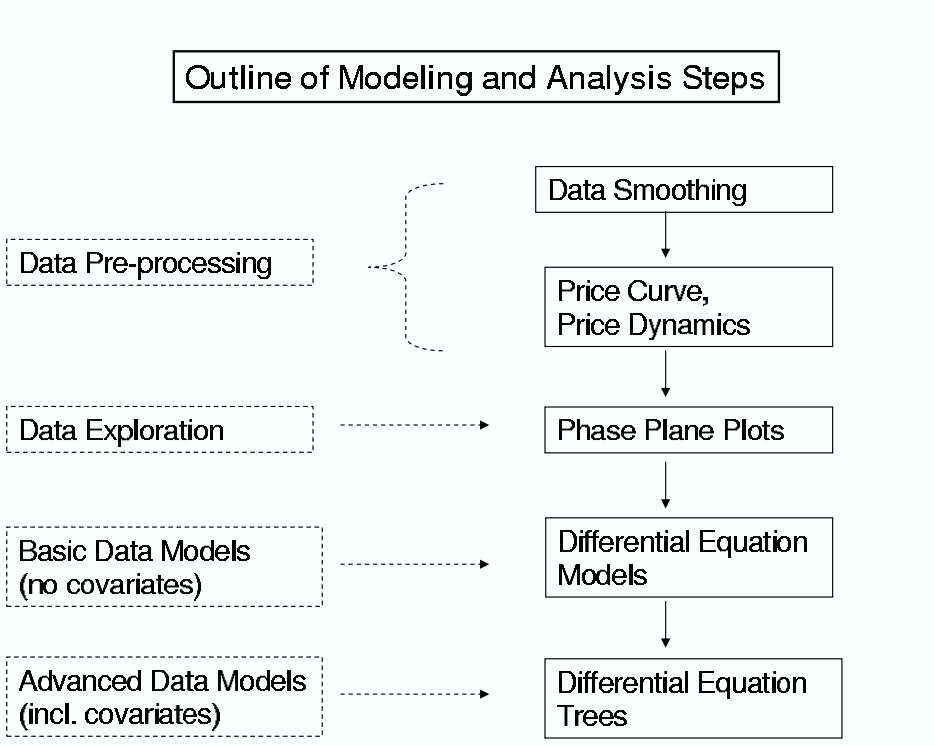

Figure 1.1. *Roadmap for the remainder of this paper.*

1.2 DATA

The data used in this chapter consist of closed 7-day auctions for two different products, *Microsoft Xbox* gaming systems and *Harry Porter and the Half-Blood Prince* books. All auctions transacted and included at least 2 bids. Xbox systems were popular items on eBay at the time of data collection and had a market value of $179.98 (based on Amazon.com). Harry Potter books were also very popular items and sold for $27.99 on Amazon.com. In that sense, we can consider Xbox systems high-valued items and can contrast them with the lower-valued Harry Potter books.

For each auction, we collected the bid history, which reveals the temporal order and magnitude of bids, and that forms the basis of the differential equations model. Figure 1.2. shows an example of a typical bid history for an Xbox auction. In addition to the bid history, we collected information on a wide variety of other auction characteristics such as

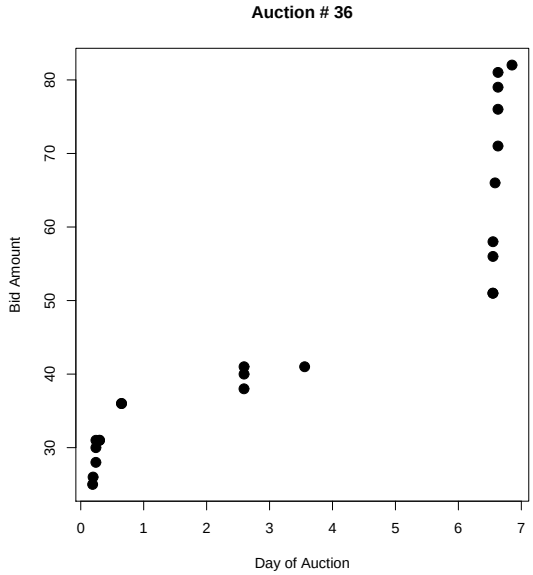

Figure 1.2. *The bids placed in auction number 36 of a Microsoft Xbox auction. The horizontal axis denotes time (in days); the vertical axis denotes bid amount (in £).*

the opening bid and the final price, the number of bids, and the seller and bidder ratings (summary statistics of these continuous variables are given in the top of Table 1.1.). We also recorded item condition (used vs. new), whether or not the seller set a secret reserve price, and whether or not the auction exhibited early bidding or jump bidding (see summary statistics in the bottom of Table 1.1.). For further details on these data, see [19].

1.3 PRICE CURVES AND DIFFERENTIAL EQUATION MODELS

Although observed bidding data are discrete, we prefer a smooth continuous representation. The reason for this is that, from a conceptual point of view, price during an auction is a continuous process. Moreover, from a practical point of view, we want to estimate price dynamics which can be done by calculating derivatives of the price process. For that reason, we also prefer a smooth representation. This lends itself to the use of functional data analysis (FDA). In FDA the object of analysis is a continuous (functional) object rather than a scalar or vector. In the recent years there has been a surge of research in FDA (mainly due to the monographs by [16, 17]), both in application areas as well as in theoret-

Variable	Mean	Median	Min	Max	StDev.
Opening Bid	19.84	5.99	0.01	175.00	31.07
Winning Bid	71.78	17.75	7.00	405.00	76.99
Number of bids	14.12	11.00	2.00	75.00	11.05
Seller Rating	280.00	85.5	0	9515.00	829.96
Bidder Rating	61.72	34.23	0	758.00	85.31

Variable	Case	Count	Proportion (%)
Value	High	93	48.95
	Low	97	51.05
Reserve Price	Yes	5	2.63
	No	185	97.37
Condition	New	60	31.58
	Used	130	68.42
Early Bidding	Yes	81	42.63
	No	109	57.37
Jump Bidding	Yes	34	17.89
	No	156	82.11

Table 1.1. Summary statistics for all continuous (top) and categorical (bottom) variables.

ical research. FDA is especially suitable in the online auction context, since our object of interest is the price path and we have many replications of the same (or similar) path (i.e. a sample of auctions for the same or similar product). The standard approach for estimating a functional object from discrete data is via smoothing which we describe next.

1.3.1 Fitting price curves via smoothing

Our goal is to measure, for each individual auction, the dynamics of its associated price process. To that end, we first need a smooth representation of the process itself. We refer to this process as the smooth functional object. After creating the functional object we obtain estimates for its dynamics via the first and second derivatives.

More specifically, let $\tilde{y}_i^{(j)}$ denote the jth ($j = 1, \ldots, n_i$) bid in auction i ($i = 1, \ldots, N$), on day t_{ij} ($0 \leq t_{ij} \leq 7$)[1]. Note that because the bids arrive at irregularly spaced times, the t_{ij}'s (and n_i's) vary from one auction to another. To account for the irregular spacing, we sample the current price step function at a common set of time points $t_j, 0 \leq t_j \leq 7, j =$

[1]In our application, bid-values are transformed into log-scores to better capture common price surges, especially towards the auction end.

$1, \ldots, n$. Thus the observed price path for auction i can be represented by a vector of fixed length n:

$$y_i(t) = (y_i^{(1)}, \ldots, y_i^{(n)}), \tag{1.1}$$

where $t = (t_1, \ldots, t_n)$ and $y_i^{(j)} = y_i(t_j)$ denotes the value of the bid sampled at time t_j.

In order to arrive at a smooth representation, we approximate y_i using basis functions. We write

$$y_i(t) = f_i(t) + \epsilon_i(t), \tag{1.2}$$

where the error term $\epsilon_i(t)$ is assumed to be the only cause of roughness for an otherwise smooth object. Using an appropriate basis functions expansion, we can represent $f_i(t)$ as

$$f_i(t) = \sum_{k=1}^{K} c_{ik}\phi_k(t) \tag{1.3}$$

for a set of known basis functions $\phi = (\phi_1(t), \ldots, \phi_K(t))$ and a coefficient vector $c_i = (c_{i1}, \ldots, c_{iK})^T$. Then the $K \times N$ estimated coefficient matrix $\hat{c} = (\hat{c}_1, \ldots, \hat{c}_N)$ minimizes the penalized sum of squares

$$PENSSE_\lambda(c) = \sum_{i=1}^{N}\sum_{j=1}^{n}(y_i(t_j) - f_i(t_j))^2 + \lambda \int (Lt)^2 dt. \tag{1.4}$$

Using $Lt = f''(t)$, $\hat{c}$ is given by

$$\hat{c} = (c^T c + \lambda K)^{-1} c^T Y(t), \tag{1.5}$$

where c is the $n \times K$ basis matrix, $Y(t)$ is the $n \times N$ matrix of responses, and λ is a smoothing parameter that controls the trade-off between data fit and smoothness. Note that the elements of K are given by $K_{kl} = \int c_k''(t)c_l''(t)dt$. In this work, we use B-splines of order 6 to allow for a reliable estimation of at least the first three derivatives of f. The selection of the knots and smoothing parameter are driven by visual inspection of the resulting functional objects [2].

The left panel in Figure 1.3. shows the smooth price curves for the 190 auctions in our dataset. As mentioned earlier, an advantage of having smooth price curves is that we can readily obtain estimates for their dynamics via their derivatives. First and second derivatives of the price curve correspond to the price-velocity and price-acceleration, respectively. The middle panel in Figure 1.3. shows that most price velocities are close to zero, especially during mid-auction, implying a process with linear growth. In contrast,

[2]For more on the quality of fit of the smooth curves to the observed data see [20]

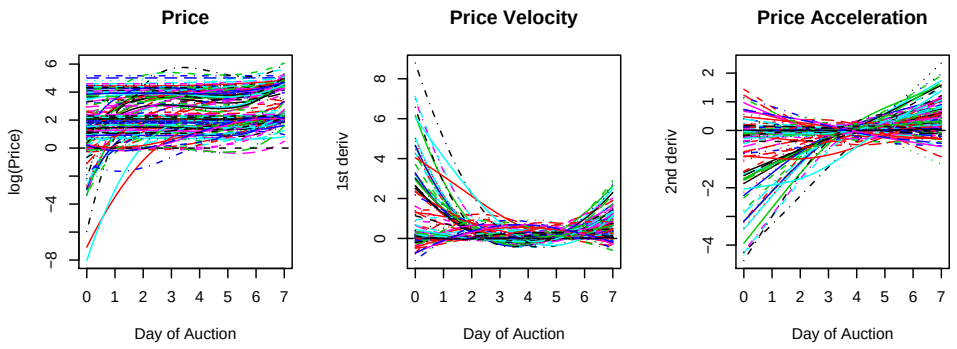

Figure 1.3. *Price curves for the 190 7-day auctions on Xbox play stations and Harry Potter books, together with their estimated first two derivatives.*

velocities are often very high at the auction-start and especially at the end. However, the magnitude of the dynamics differs quite significantly from auction to auction. (This can also be seen from the range of the price accelerations in the right panel of Figure 1.3..) Thus, auction dynamics can be quite heterogeneous.

In the following we use *phase-plane plots* (PPPs) to further investigate auction dynamics. In particular, we use PPPs to investigate the relationships between dynamics of different order. This investigation will motivate our subsequent efforts to model dynamics via differential equations.

1.3.2 Exploring Auction Dynamics via Phase Plane Plots

At the heart of differential equations are models that relate the function and its derivatives to one another. A preliminary step to choosing an appropriate differential equation model is to examine plots of pairs of derivatives one vs. the other. Such plots are called phase plane plots (PPPs). In the functional context, where one has repeat observations at each derivative level, we plot the *averages* versus one another.

Figure 1.4. shows a PPP for the average price-acceleration vs. the average price-velocity. The numbers along the curve indicate the day of the auction (for 7-day auctions). We see that price velocity is high at the beginning of the auction and then the dynamics slow down: price acceleration becomes negative and there is a slow-down in the price-velocity.

This continues until about day 4, after which the dynamics reverse: acceleration becomes positive and velocity increases rapidly until the auction-end.

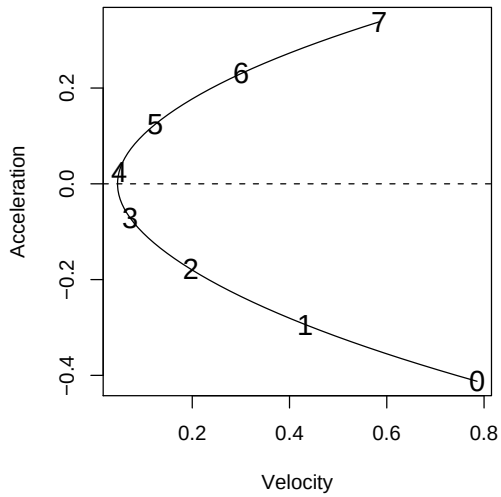

Figure 1.4. *Phase Plane Plot for the average price curve of the data: the second derivative (acceleration) versus the first derivative (velocity).*

There are several interesting aspects that appear in Figure 1.4.. First, the "C"-shape of the PPP is typical of an online auction: a phase of decrease in dynamics followed by a transitional phase of change, and finally a phase of increase in dynamics. Second, the magnitude/importance of each phase varies as a function of different auction characteristics. This can be seen in the series of *conditional* PPPs in Figure 1.5., where the average derivatives are conditional on the auction characteristics described in Table 1.1.. Here, pairs of plots can be compared to see the effect of the two different levels (e.g., new vs. used items in the top left, or high vs. low bidder rating in the bottom right). While the general "C"-shape persists in all PPPs, the *magnitude* of the dynamics varies. Moreover, the different size of the "C"-shapes also indicates that the magnitude of the *relationship* between velocity and acceleration differs.

To summarize the findings on the relationship between price dynamics and individual auction-related variables from Figure 1.5.: For item value, high-valued items appear to have a larger range of dynamics compared to lower-valued items. Early bidding seems to

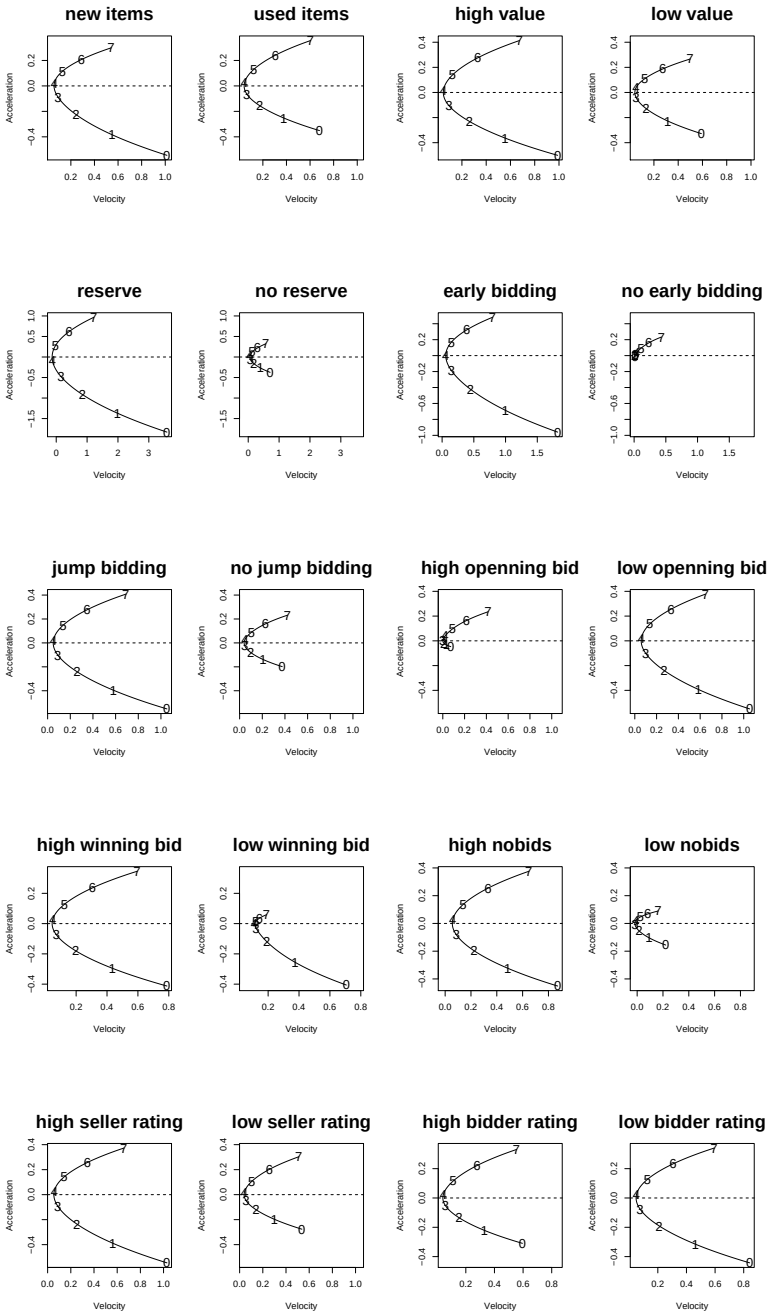

Figure 1.5. *Conditional Phase Plane Plots (PPP) for the average price curve of the data, conditional on 10 auction characteristics from Table 1.*

have a large effect on the dynamics not only at the auction-start but throughout the entire auction. For jump-bidding, we see that auctions that experience jump bids have a different relationship between velocity and acceleration compared to auctions without jump bids. Additional observations are that the opening bid and the number of bidders both have an impact on the dynamics. Interestingly, while different bidder ratings do not appear to make much of a difference, seller ratings do. Note however that several of these variables are correlated, e.g., closing price and item value. Our goal is therefore to look at the effect of the combined information on price dynamics. For this purpose we formulate a differential equation model that relates price dynamics to a set of predictor variables.

The previous analysis shows that dynamics exist and that heterogeneity among auction dynamics are due to different auction sub-populations determined by the product, the auction format, the seller or the bidders. Moreover, while dynamics vary, we see a persistent "C" shape for the relationship between acceleration and velocity. We take this as motivation that online auction dynamics can be captured by a single family of differential equation models (DEM). In the following we derive such a class of DEM.

1.3.3 Differential Equation Models for Price

Differential equations are widely used in the areas of engineering and physics. A differential equation describes a process with changing dynamics by specifying relationships among the function and its derivatives. In the context of online auctions we view the price process as a dynamic system with many observed and unobserved factors acting upon it. Our exploratory analysis using PPPs indicates a general structure of price dynamics, and we now set out to find a DEM that can capture this structure.

Let y_i be the price function for auction i $(i = 1, \ldots, N)$ and let $D^m y_i$ be the m^{th} derivative of y_i. Our goal is the identification of a linear differential operator (LDO) of the form

$$L = \omega_0 I + \omega_1 D + \cdots + \omega_{m-1} D^{m-1} + D^m \tag{1.6}$$

that satisfies the homogeneous linear differential equation $Ly_i = 0$ for each observation y_i. In other words, we seek a linear differential equation model so that our data satisfy

$$D^m y_i = -\omega_0 - \omega_1 D y_i - \cdots - \omega_{m-1} D^{m-1} y_i. \tag{1.7}$$

In practice, due to the prevalence of noise, it is impossible to find a model that satisfies (1.7) *exactly*. Hence, principal differential analysis adopts a least squares approach to the fitting of the differential equation model. For details on parameter estimation see [20].

Next, we focus on a second-order differential equation since the PPPs indicated varying relationships between the first and second derivatives of price. The general second-order differential equation is of the form

$$Ly_i = \omega_0 y_i + \omega_1 Dy_i + D^2 y_i = 0. \tag{1.8}$$

However, to obtain a strictly monotone, twice-differentiable function as required from a price process in online auctions, we must set $\omega_0 = 0$ [15]. This leads to the differential equation

$$Ly_i = \omega Dy_i + D^2 y_i = 0, \tag{1.9}$$

Note that coefficient function $\omega^* = -D^2 y/Dy$ measures the relative curvature of the monotone function in the sense that it assesses the size of the curvature of $D^2 y$ relative to the slope Dy. A constant value of ω^* implies an exponential process of the form $y(t) = C_0 + C_1 \exp(\omega t)$, with $\omega^* = 0$ the special case of a linear function. Thus, small or zero values of ω^* correspond to locally linear functions, whereas very large values correspond to regions of sharp curvature. In mechanical systems, the latter type is generally caused by internal or external frictional forces or viscosity. In the context of online auctions, sharp curvature in the price process can be related to jump bids caused by bidders attempting to apply external force to the bidding process, and locally linear motion is observable during the middle of the auction.

Next, we estimate the second order differential equations model for our 190 auctions (for details on model fit see [20]). Figure 1.6. shows the estimated coefficient curve ω^*. We can see that ω^* has three phases: negative, zero, and finally positive. These correspond to the three typical bidding phases during an auction: early activity, little mid-auction activity, and high late activity. Recall that a value of zero indicates linear motion of the price process (i.e., no dynamics), whereas large positive or negative values are indicative of changes in the dynamics (oppressing them or increasing them, respectively). The first phase (up to day 3) is characterized by a negative ω^*, with a dip on day 2. This negative dip marks the change from early bidding to "bidding draught", when velocity decreases. Then, we see that $\omega^* = 0$ during the bidding draught, until price starts to increase again with a peak on day 6, in transition to high-intensity last moment bidding.

In summary, we find that a second order differential equation model fits the data reasonably well. It captures the three typical phases of bidding and the interplay of dynamics that change over the course of the auction. However, [20] also find that the degree of model fit varies at different periods of the auction and that a considerable amount of variation is

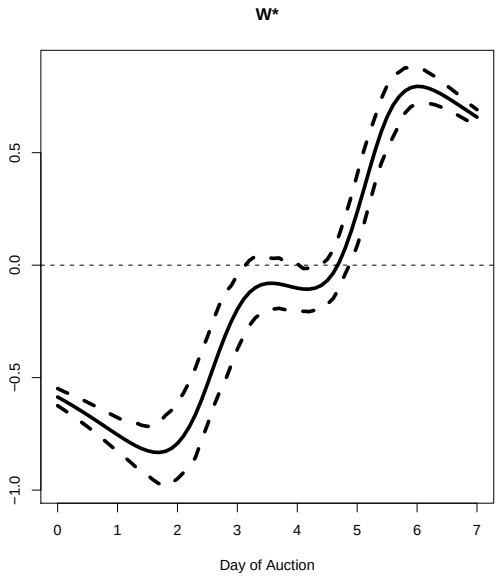

Figure 1.6. *Estimated coefficient function of the monotone 2nd order differential equation fitted to online auction data.*

remained unexplained. The next step is therefore to incorporate external auction-related information. [20] fit separate models to different levels of the categorical covariates and find differences between pairs of models. However, we strive for a more general approach that allows for the incorporation of covariate information directly into the dynamic model. For that purpose we develop a novel approach based on regression trees. More specifically, we propose a new modeling approach for dynamic data via functional differential equations trees.

1.4 FUNCTIONAL DIFFERENTIAL EQUATION TREES

The previous section shows that differential equation models can capture the changing dynamics in online auctions, but also that a significant amount of the variation in the dynamics is left unexplained. The roadmap for the rest of the paper is thus as follows. In the following we set out to explain some of the residual variation using covariate information. In the auction context, plenty of covariate information is available (see, e.g., Table 1). However, incorporating covariate information into differential equation models is not straightforward. [17] suggest a way of incorporating covariate information via the forcing

function, thereby creating a non-homogenous differential equation but this approach has not received much traction to date. We thus propose an alternative and innovative approach, borrowing ideas from recursive partitioning. In particular, we propose a novel tree-based approach for differential equation models. We discuss how to estimate differential equation trees and we subsequently illustrate their performance on our online auction data. We are quick to point out that our description is rather technical and that the data example is for illustrative purposes only. The full power of our approach still remains to be investigated.

Tree models give simple descriptions of often complex, nonlinear relationships between several predictors and a univariate or multivariate response. A classical reference is the monograph *Classification and Regression Trees* by [3]. While tree models are generally very powerful, they encounter problems when the response is high dimensional. [21] explore two ways of fitting trees to high dimensional data. Both approaches proceed by first reducing the dimensionality of the data and then fitting a standard multivariate tree to the reduced response. In the first approach, the dimensionality is reduced by representing the response as a linear combination of spline basis functions, while in the second one, the dimensionality is reduced using principal component analysis, retaining only the first several principal components.

Note that classical trees fit a scalar in each node of the tree. Since this tends to produce rather large and complex trees (see e.g., [4]), research on incorporating (simple) parametric models into trees has recently received considerable attention. Such approaches are often referred to as *functional trees* ([5]) with the most notable being *M5* ([14]). (See also [13, 12, 4] for related approaches.) One particularly noteworthy approach is that of [22] who take the integration of parametric models into trees one step further by embedding recursive partitioning into the model estimation and variable selection framework. Within that framework, every leaf is associated with a conventionally fitted model such as, e.g., a maximum likelihood model or linear regression. The model's objective function is used for estimating the parameters as well as the split points. The appeal of this approach is that the same objective function is used for partitioning and for parameter estimation. Building upon these ideas, we propose a novel model-based functional differential equation tree which allows the incorporation of covariate information into dynamic models. We first briefly review the main ideas of model-based recursive partitioning. We then use these ideas to derive our functional differential equation tree methodology.

1.4.1 Model-Based Recursive Partitioning

Let $M(y, \theta)$ be a parametric model where $y = (y_1, \ldots, y_N)$ are (possibly vector-valued) observations and $\theta \in \Theta$ is a k-dimensional vector of parameters. We assume that $M(\cdot)$ can be estimated by minimizing some objective function, say, $\Psi(y, \theta)$, yielding the parameter estimate $\hat{\theta}$ where

$$\hat{\theta} = \text{argmin}_{\theta \in \Theta} \Psi(y, \theta). \tag{1.10}$$

Estimators of this type are based on many well-known estimation techniques, the most popular ones being ordinary least squares (OLS) and maximum likelihood (ML). In the case of OLS, Ψ is given by the error sum of squares, and in the case of ML, it is the negative log-likelihood.

Let $(Z_1, \ldots, Z_L)$ be a set of partitioning variables (i.e., covariates). We assume that there exists a partition $\{\beta_b\}_{(b=1,\ldots,B)}$ of the space $Z = Z_1 \times \cdots \times Z_L$ into B cells (or segments) such that in each cell β_b, a model $M(y, \theta_b)$ with a cell-specific parameter θ_b holds.

The basic idea of model-based recursive partitioning is now that each node is associated with a single model. In the first step, the designated model $M(y, \theta)$ is fit to all observations by estimating $\hat{\theta}$ via minimization of the objective function Ψ; this yields a model in the top node. Then in the second step, a fluctuation test for parameter instability is performed to assess whether splitting of the node is necessary. Generally speaking, determining the necessity of a split based on a partitioning variable Z_l is based on comparing the estimated parameters of the post-split resulting daughters to determine whether they come from the same mean (and thus the split is unnecessary), or not. If there is significant parameter instability with respect to any of the partitioning variables Z_l $(1 \leq l \leq L)$, then we select the variable Z_l that is associated with the highest parameter instability. In the third step, we compute the split point(s) of Z_l that locally optimize Ψ. Finally, we split the node into B locally optimal segments and repeat the procedure. If no more significant instabilities can be found, the recursion stops and returns a tree where each terminal node is associated with a model of type $M(y, \theta_b)$.

The steps of the algorithm are as follows:

1. Fit the model $M(y, \theta)$ to all observations in the current node by estimating $\hat{\theta}$ via minimization of the objective function Ψ.

2. Assess the stability of the parameters w.r.t. every ordering $Z_1, \ldots, Z_L$. If there is some overall instability, choose the variable Z_l associated with the highest parameter instability for partitioning; otherwise stop.

3. Search for the locally optimal split point(s) in Z_l by minimizing the objective function of the model ψ.

4. Split the node into daughter nodes and repeat the procedure.

The details for steps 1-3 are described below. To keep notation simple, dependence on the current segment $b \in \{1, \ldots, B\}$ is suppressed.

1.4.1.1 Testing for Parameter Instability We assess parameter instability adopting the method of [22]. The basic idea is to check whether the score functions $\hat{\psi}_i$ ($\hat{\psi}_i = \hat{\psi}(y_i, \hat{\theta})$, where $\psi = \frac{\partial \Psi(y,\theta)}{\partial \theta}$) fluctuate randomly around their mean of 0 or exhibit systematic deviations from 0 over Z_l. These deviations can be captured by the empirical fluctuation process

$$W_l(t) = \hat{J}^{-1/2} n^{-1/2} \sum_{i=1}^{\lfloor nt \rfloor} \hat{\psi}_{\sigma(Z_{il})}, (0 \leq t \leq 1) \tag{1.11}$$

where $\sigma(Z_{il})$ is the ordering permutation which gives the antirank of the observation Z_{il} in the vector $Z_l = (Z_{1l}, \ldots, Z_{nl})^T$. Thus, $W_l(t)$ is simply the partial sum process of the scores ordered by the variable Z_l, scaled by the number of observations n and a suitable estimate $\hat{J}$ of the covariance matrix $COV(\psi(y, \hat{\theta}))$, e.g., $\hat{J} = n^{-1} \sum_{i=1}^{n} \psi(y_i, \hat{\theta}) \psi(y_i, \hat{\theta})^T$. This empirical fluctuation process is governed by a functional central limit theorem under the null hypothesis of parameter stability: it converges to a Brownian bridge. We can derive a test statistic by applying a scalar functional $\lambda(\cdot)$ capturing the fluctuation in the empirical process to the fluctuation process $\lambda(W_l(\cdot))$ and determine its limiting distribution. Then, in order to test whether there is instability in the current node, we check whether the observed significance level falls below a certain threshold α (e.g. $\alpha = 5\%$) and we consequently split the variable Z_l with the smallest p-value. In the following we describe this process in more detail for assessing numerical and categorical variables.

Assessing Numerical Variables: For capturing instabilities of a numerical variable Z_l, we use the sup*LM* statistic proposed by [1]:

$$\lambda_{supLM}(W_l) = max_{i=\underline{i},\ldots,\overline{i}} \left(\frac{i}{N} \frac{N-i}{N} \right)^{-1} \left\| W_l(\frac{i}{N}) \right\|_2^2. \tag{1.12}$$

This statistic returns the maximum of the squared L_2 norm of the empirical fluctuation process scaled by its variance function. This type of statistic first appeared in [1]; it can be interpreted as the *LM* statistic[3] against a single change point alternative where the potential change point is shifted over the interval $[\underline{i}, \overline{i}]$. The interval is defined by requiring some minimal segment size $\underline{i}$ and then $\overline{i} = N - \underline{i}$. The limiting distribution of (1.12), as shown in [1], is given by the supremum of a squared, m-dimensional tied-down Bessel process $sup_t(t(1-t))^{-1}\|W^0(t)\|_2^2$, where W^0 denotes a Brownian bridge.

Assessing Categorical Variables: For capturing instabilities of a categorical variable, we need a different statistic. A categorical variable Z_l with G levels (or categories) has ties and a total ordering of the observations is not available. We therefore need a different statistic. An appropriate statistic is one that is insensitive to the ordering of the G levels and of the ordering of observations within each level. One such statistic is given by (see [7]):

$$\lambda_{\chi^2}(W_l) = \sum_{g=1}^{G} \frac{|I_g|^{-1}}{N} \|\Delta_{I_g} W_l(\frac{i}{N})\|_2^2 \tag{1.13}$$

where $\Delta_{I_g} W_l$ is the increment of the empirical fluctuation process over the observations in category $g = 1, \ldots, G$ (i.e., essentially the sum of the scores in category g). The test statistic is then the weighted sum of the squared L_2 norm of the increments which has an asymptotic χ^2 distribution with $m \cdot (G-1)$ degrees of freedom. For more details, see [7, 22].

One last problem remains. Recall that the empirical fluctuation process depends on the score function $\hat{\psi}$ which is given by the derivative of $\Psi(y, \theta)$ w.r.t θ. Differentiating with respect to the infinite dimensional parameter function $\theta = \theta(t)$ is not straightforward. We solve it via basis expansion of the form $\theta \approx \sum_k c_k \phi_k = c'\phi$, where $\phi = (\phi_1, \ldots, \phi_K)^T$ is a K-vector of basis functions and $c = (c_1, \ldots, c_K)^T$ is a vector of associated coefficients.

1.4.1.2 Splitting In the third step the model is split with respect to the variable Z_l into B segments (typically, $B = 2$). Two rival segmentations are compared by comparing the segmented objective function $\sum_{b=1}^{B} \sum_{i \in I_b} \Psi(y_i, \theta_b)$. The optimal partition is found by performing an exhaustive search over all conceivable partitions into B segments. See [22] for more details.

[3]The *LM* (Lagrange multiplier) statistic is based on the generalized method of moments (GMM) estimators (see [6]).

1.4.2 Model-Based Functional Differential Equation Trees

We now adopt this methodology to differential equation trees: we apply recursive model-based partitioning to the differential equation context by suitably defining the objective functions $M(y, \theta)$ and $\Psi(y, \theta)$. Consider again a differential equation model of the form

$$D^m y_i = -\omega_0 y_i - \omega_1 D y_i - \ldots - \omega_{m-1} D^{m-1} y_i. \tag{1.14}$$

Following the notation from the previous section, we will denote this model by $M(y, \omega)$. As pointed out in [20], we can estimate this model by minimizing the sum of squared norms

$$\Psi(y, \omega) = \sum_{i=1}^{N} \int \left[\sum_{j=0}^{m} \omega_j(t)(D^j y_i)(t) \right]^2 dt. \tag{1.15}$$

over the weight function $\omega = (\omega_0, \ldots, \omega_m)$.

Parameter estimation can be accomplished via basis expansion. First, approximate the coefficients ω_j via a linear combination of basis functions:

$$\omega_j \approx \sum_k c_{jk} \phi_k. \tag{1.16}$$

Using this approximation, we can write $\Psi(y, \omega)$ as a quadratic form in c

$$\Psi(y, \omega) \approx C + c' R c + 2c' s, \tag{1.17}$$

where the constant C does not depend on c, and hence the estimate $\hat{c}$ is given by the solution of the equation

$$\hat{c} = -R' s, \tag{1.18}$$

where the symmetric matrix R consists of an $m \times m$ array of $K \times K$ submatrices R_{jk} of the form

$$R_{jk} = N^{-1} \int \phi(t)\phi(t)' \sum_i D^j y_i(t) D^k y_i(t) dt \tag{1.19}$$

for $j = 0, \ldots, m-1$. For more details on parameter estimation, see e.g. [17].

1.4.3 Application to Online Auction Data

Figure 1.7. shows the fitted functional differential equation tree for the data described in Section 1.2. This tree was obtained by employing a stopping criterion of a statistical significance level of $\alpha = 0.01$ and minimum number of at least 10 observations within each final node. The resulting tree uses three different splitting variables to arrive at 8 different price

curves: The opening bid (Obid), the winning bid (Wbid) and the number of bids (Numbid). Returning to Figure 1.4 (the conditional PPP plots), the tree confirms the differences between price dynamics observed for each of the variables from Table 1.1. Moreover, as noted earlier, many of the variables are correlated, and therefore it is not surprising that some of them are absent from the tree. For instance, item value and item condition are associated with the winning bid; reserve price and early bidding are associated with the opening bid. As in the PPP plots, the average bidder rating does not appear to lead to different price dynamics and is absent from the tree. In contrast, seller rating, which did appear to yield different PPP plots, is conspicuously missing from the tree. One possible reason for this is that seller rating (which proxies for experience) is indirectly associated with the opening bid, as the seller determines the opening bid[4].

With respect to the resulting 8 leaf nodes, the estimated price curves cover three main shapes: shape 1 is characterized by a fast initial price-increase followed by an end-of-auction slow-down (L1-3); shape 2 is characterized by almost linear increase (L4); and shape 3 shows little to moderate initial activity followed by late price spurts (L5-8). We also note that these three main shapes differ by the opening bid only: For auctions with an opening bid lower than $3.99, the price curves follow the first shape; for those with opening bid higher than $6.5, the curves follow the third shape; and for auctions with opening bid between $3.99 and $6.5, the price curves follow the second shape which is almost linear increase. Further segmentation of these basic shapes is achieved via the winning bid and the number of bids.

1.5 CONCLUSIONS

We propose a novel tree-based approach for incorporating covariate information into differential equation models. Our recursive model-based approach defines an objective function that is also used for determining the splits of the tree.

This work fills a void in the literature in that it offers a direct and practical method for modeling the relationship between process dynamics and external factors. By applying our methodology to online auction data, we show that dynamic models can capture the

[4]Using a more lenient stopping criterion results in a larger tree. For example, using $\alpha = 0.1$ and a minimum number of at least 5 observations within a final node results in a tree that also incorporates the seller rating as splitting variable. Note that this tree is the same as earlier except for the addition of seller rating. We leave it up to future research to determine the optimal settings for the differential equation tree.

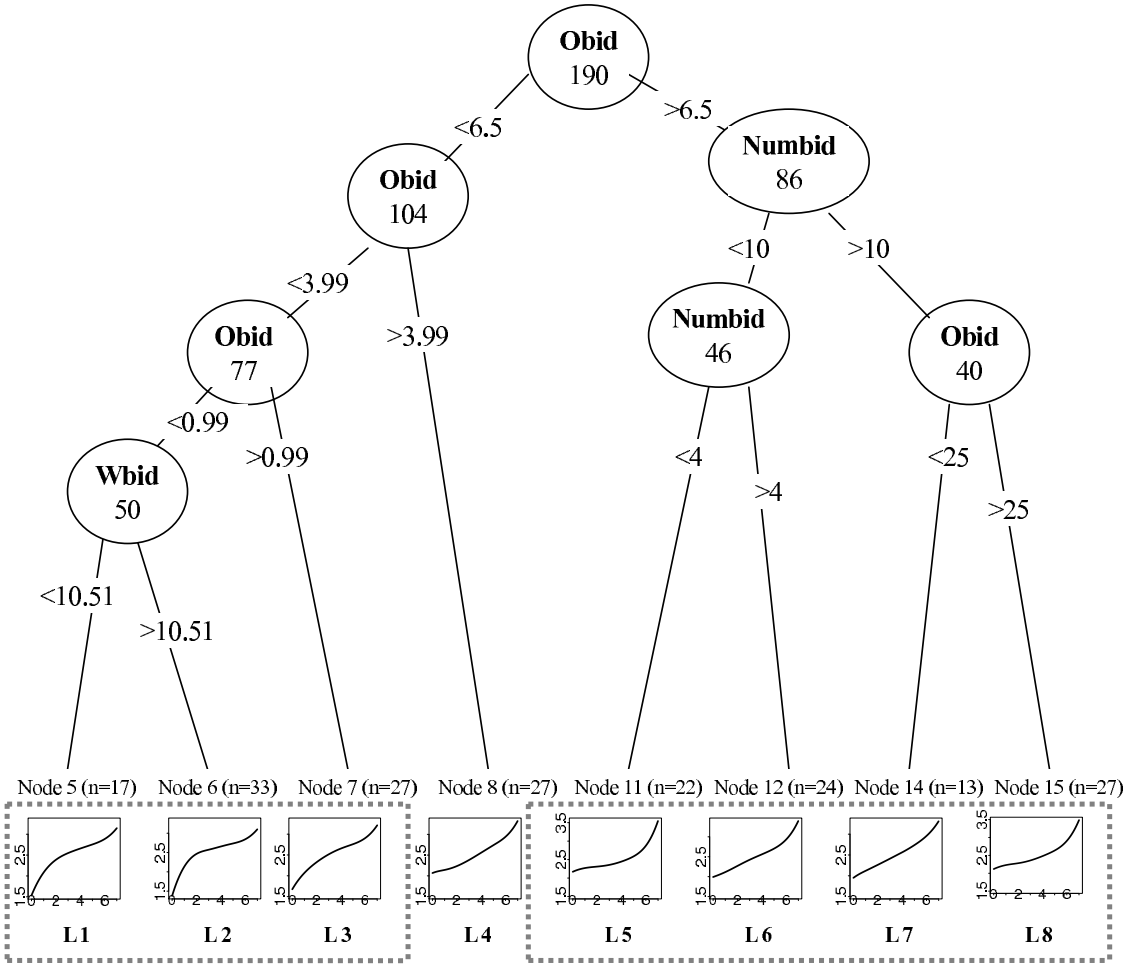

Figure 1.7. *Fitted model-based differential equation tree applied to online auction data, using* $\alpha = 0.01$ *and min(# observations in leaf node)*> 10*. The number within each splitting node is the sample size. "*$L1, \cdots, L8$*" denote the 8 terminal/leaf nodes.*

changing price dynamics in an online auction and in particular, we find that the opening bid, the number of bids, and the winning bids are the major factors in determining the shape of the price curve and its dynamics.

Bibliography

1. D.W.K. Andrews. Tests for parameter instability and structural change with unkown change point. *Econometrica*, 61:821–856, 1993.

2. Ravi Bapna, Wolfgang Jank, and Galit Shmueli. Price formation and its dynamics in online auctions. Technical report, Smith School of Business, University of Maryland, College Park, 2005.

3. L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. *Classification and Regression Trees*. Belmont, CA: Wadsworth, 1984.

4. K.Y. Chan and W.Y. Loh. Lotus: An algorithm for building accurate and comprehensible logistic regression trees. *Journal of Computational and Graphical Statistics*, 13(4):826–852, 2004.

5. J. Gamma. Functional trees. *Machine Learning*, 55:219–250, 2004.

6. L. P. Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50:1029–1054, 1982.

7. N.L. Hjort and A. Koning. Tests for constancy of model parameters over time. *Nonparametric Statistics*, 14:113–132, 2002.

8. V. Hyde, W. Jank, and G. Shmueli. Investigating concurrency in online auctions through visualization. *The American Statistician*, 60 (3):241–250, 2006.

9. V. Hyde, W. Jank, and G. Shmueli. *Statistical Methods in eCommerce Research*, chapter A Family of Growth Models for Representing the Price Process in Online Auctions. Wiley & Sons, 2008.

10. W. Jank and G. Shmueli. Functional data analysis in electronic commerce research. *Statistical Science*, 21 (2):155–166, 2006.

11. W. Jank and G. Shmueli. *Handbook on Information Series: Business Computing*, chapter Studying Heterogeneity of Price Evolution in eBay Auctions Via Functional Clustering. Elsevier, Forthcoming.

12. H. Kim and W.Y. Loh. Classification trees with unbiased multiway splits. *Journal of the American Statistical Association*, 96(454):589–604, 2001.

13. W. Y. Loh. Regression trees with unbiased variable selection and interaction detection. *Statistica Sinica*, 12:361–386, 2002.

14. J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, California, 1993.

15. J. O. Ramsay. Estimating smooth monotone functions. *Journal of the Royal Statistical Society*, Series B 60:365–375, 1998.

16. J. O. Ramsay and B. W. Silverman. *Applied functional data analysis: methods and case studies*. Springer-Verlag New York, Inc., 1st edition, 2002.

17. J. O. Ramsay and B. W. Silverman. *Functional data analysis*. New York: Springer-Verlag, 2nd edition, 2005.

18. G. Shmueli and W. Jank. *Economics, Information Systems & Ecommerce Research II: Advanced Empirical Methods, part of Advances in Management Information Systems Series*, chapter Modeling the Dynamics of Online Auctions: A Modern Statistical Approach. M.E. Sharpe, Armonk, NY, 2006. Forthcoming.

19. S. Wang, W. Jank, and G. Shmueli. Explaining and forecasting online auction prices and their dynamics using functional data analysis. *Journal of Business and Economic Statistics*, 2007 (in press).

20. S. Wang, W. Jank, G. Shmueli, and P. Smith. Modeling price dynamics in ebay auctions using principal differential analysis. Technical report, University of Maryland. Available at http://www.smith.umd.edu/faculty/wjank/Wang-Jank-Shmueli-Smith-PDA_ of_Online_Auctions.pdf, 2006.

21. Y. Yu and D. Lambert. Fitting trees to functional data, with an application to time-of-day patterns. *Journal of Computational and Graphical Statistics*, 8(4):749–762, 1999.

22. A. Zeileis, T. Hothorn, and K Hornik. Model-based recursive partitioning. Technical report, Report 19, Department of Statistics and Mathematics, Wirtschaftsuniversität Wien, 2005.