

ABSTRACT

Title of dissertation: REPRESENTING, VISUALIZING,
AND MODELING
ONLINE AUCTION DATA

Valerie Hyde
Doctor of Philosophy, 2007

Dissertation directed by: Professor Galit Shmueli
Department of
Decision and Information Technologies
and
Professor Wolfgang Jank
Department of
Decision and Information Technologies

The wide and growing popularity of online auctions creates enormous amounts of publicly available bid data providing an important topic for research. These data pose unique statistical challenges because of their special structures. This research focuses on methods for representing, visualizing, and modeling such data.

We find semi-continuous data manifested in auction consumer surplus data. Semi-continuous data arise in many applications where naturally continuous data become contaminated by the data generating mechanism. The resulting data contain several values that are “too-frequent”, a hybrid between discrete and continuous data. The main problem is that standard statistical methods, which are geared towards continuous or discrete data, cannot be applied adequately to semi-continuous data. We propose a new set of two transformations for semi-continuous data that “iron-out” the too-frequent values into completely continuous data. We show that

the transformed data maintain the properties of the original data but are suitable for standard analysis.

We are also interested in the effect of concurrency not only on the final price of an auction but also on the relationship between the current bid levels and high bids in simultaneous ongoing auctions. We suggest several ways to visually represent the concurrent nature of online auction prices. These include “rug plots” for displaying the price-evolution and price dynamics in concurrent auctions, time-grouped box plots, and moving statistics plots. We find price trends and relationships between prices in concurrent auctions and raise new research questions.

Finally, bids during an online auction arrive at unequally-spaced discrete time points. Our goal is to capture the entire continuous price-evolution function by representing it as a functional object. Various nonparametric smoothing methods exist to estimate the functional object from the observed discrete bid data. Previous studies use penalized polynomial and monotone smoothing splines; however, these require the determination of a large number of coefficients and often lengthy computational time. We present a family of parametric growth curves that describe the price-evolution during online auctions. Our approach is parsimonious and has an appealing interpretation in the online auction context. We also provide an automated fitting algorithm that is computationally fast.

REPRESENTING, VISUALIZING, AND MODELING
ONLINE AUCTION DATA

by

Valerie Hyde

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2007

Advisory Committee:

Professor Galit Shmueli, Chair/Advisor

Professor Wolfgang Jank, Co-Chair/Co-Advisor

Professor Benjamin Kedem

Professor Ben Shneiderman

Professor Paul Smith

© Copyright by
Valerie Hyde
2007

Dedication

I would like to dedicate this dissertation to my father, my mother, and Bill.

Acknowledgments

I wish to express my deep appreciation to my advisor, Professor Galit Shmueli, for her time, sage advice, meticulous reading of all of my papers, and encouragement. Her enthusiasm for research is contagious, and her concern for the development of her students is exceptional. It has been a wonderful experience to have a talented, sharp mentor and role model in a field with so few females.

I would also like to thank my co-advisor, Professor Wolfgang Jank. His creativity, advice, and prodding for me to take risks and look at the big picture greatly enhanced my research as well as this final document.

Thanks are also due to the rest of my thesis committee: Professor Benjamin Kedem, Professor Ben Shneiderman, and Professor Paul Smith (alphabetically) for their time and reading of my dissertation. Professor Kedem provided me with a full university fellowship my first two years at the university which gave me the much needed time to devote to coursework and prepare for qualifying examinations. Professor Smith has made himself available to me, and all students, since the first day of classes. I especially appreciate that he conducted an independent study with me to help me gauge my interest in data mining and increase my statistical toolbox. I would also like to thank Professor Shneiderman for his comments and helping me advance this research.

My statistics study group was recognized university-wide for our ability to work together as a team and to translate that teamwork into success in the classroom. Without my fellow “geese”: Eulus Moore, Aliza Kwiat, Andie Hodge, and

Olga Karles, I have no doubt that I would not have come this far or this fast. I appreciate their continued support long after they have left the university.

I am especially grateful to the numerous people who provided me with real-world data sets. I would like to thank Ravi Bapna from the Indian School of Business for the consumer surplus data, Claudia Perlich from the IBM T.J. Watson Research Center for the customer wallet data, Shanshan Wang for the Xbox data, and Sharad Borle for the luxury wristwatch data.

I owe a special debt of thanks to my family and friends for their never-ending patience and support. Without them I never would have succeeded.

Table of Contents

| | |
|--|------|
| List of Tables | viii |
| List of Figures | ix |
| List of Abbreviations | xi |
| 1 Introduction to Online Auctions | 1 |
| 1.1 Online Auctions | 1 |
| 1.2 eBay Data Structure | 2 |
| 1.2.1 eBay Bid Level Data Sets | 5 |
| 1.3 Online Auction Literature | 5 |
| 1.4 Contributions of this Dissertation | 10 |
| 1.4.1 Semi-Continuous Transformation | 11 |
| 1.4.2 Visualizing Concurrency | 12 |
| 1.4.3 Growth Models | 13 |
| 1.4.4 Summary of Dissertation Contributions | 15 |
| 2 Transformations for Semi-Continuous Data | 16 |
| 2.1 Introduction and Motivation | 16 |
| 2.2 Identifying Semi-Continuous Data | 24 |
| 2.3 Transforming Semi-Continuous Data | 25 |
| 2.3.1 The Jittering Transform | 26 |
| 2.3.2 The Local Regeneration Transform | 28 |
| 2.3.3 Goodness of Fit and Deviation | 30 |
| 2.4 Performance Evaluation Via Simulation | 32 |
| 2.4.1 Data Simulation | 32 |
| 2.4.2 Transforming the Contaminated Data: Jittering | 36 |
| 2.4.3 Transforming the Contaminated Data: Local Regeneration | 36 |
| 2.4.4 Choosing The Transformation Parameters | 40 |
| 2.4.4.1 Jittering | 44 |
| 2.4.4.2 Local Regeneration | 47 |
| 2.5 Transforming the Surplus Data | 49 |
| 2.6 Conclusions | 56 |
| 3 Functional Data Analysis | 59 |
| 3.1 What is Functional Data Analysis? | 59 |
| 3.2 Representing Auction Price Evolution as a Continuous Curve | 60 |
| 3.3 Representing a Functional Object Nonparametrically | 62 |
| 3.3.1 Basis Series Expansion | 62 |
| 3.3.2 Kernel Smoothers | 65 |
| 3.3.3 Roughness Penalty | 65 |
| 3.4 FDA Methodology | 67 |

| | | |
|---------|---|-----|
| 4 | Investigating Concurrency in Online Auctions Through Visualization | 69 |
| 4.1 | Introduction and Motivation | 69 |
| 4.2 | Visualizing Concurrent Auctions | 74 |
| 4.2.1 | Rug Plots | 74 |
| 4.2.1.1 | Curve Shapes | 79 |
| 4.2.1.2 | Temporal Groupings of Curve Types | 80 |
| 4.2.1.3 | Other Groupings of Curves | 82 |
| 4.2.2 | Time Grouped Box Plots | 83 |
| 4.2.2.1 | Comparing Medians of Adjacent Auctions | 87 |
| 4.2.3 | Moving Statistics Plots | 88 |
| 4.2.4 | Price Autocorrelation | 90 |
| 4.3 | Conclusions | 92 |
| 5 | A Family of Growth Models for Representing the Price Process in Online Auctions | 94 |
| 5.1 | Introduction and Motivation | 94 |
| 5.2 | Wristwatch Data | 97 |
| 5.3 | Representing Auction Price Nonparametrically | 98 |
| 5.3.1 | Smoothing Splines | 98 |
| 5.3.2 | Monotone Splines | 101 |
| 5.4 | Representing Price Evolution Parametrically | 102 |
| 5.4.1 | Exponential Growth | 103 |
| 5.4.1.1 | Exponential Model | 103 |
| 5.4.1.2 | Logarithmic Model | 105 |
| 5.4.2 | Logistic Growth | 106 |
| 5.4.2.1 | Logistic Model | 106 |
| 5.4.2.2 | Reflected-Logistic Model | 108 |
| 5.4.3 | Fitting Growth Models | 109 |
| 5.4.3.1 | Fitting Exponential Growth | 110 |
| 5.4.3.2 | Fitting Logarithmic Growth | 110 |
| 5.4.3.3 | Fitting Logistic Growth | 111 |
| 5.4.3.4 | Fitting Reflected-Logistic Growth | 111 |
| 5.5 | Selecting the Best Growth Model | 112 |
| 5.5.1 | Model Selection Metrics | 113 |
| 5.5.2 | Model Selection Procedure | 114 |
| 5.6 | Smoothing Method Comparison | 117 |
| 5.7 | Using Growth Curves | 123 |
| 5.7.1 | Rug Plots | 124 |
| 5.7.2 | Integrating Growth Model Parameters Into Analyses | 128 |
| 5.8 | Conclusions | 129 |
| 6 | Future Research | 132 |
| 6.1 | Semi-Continuous Transformation | 132 |
| 6.2 | Visualizing Concurrency | 133 |
| 6.3 | Growth Models | 134 |

| | | |
|-----|----------------------------------|-----|
| A | eBay Bid Level Data Sets | 138 |
| A.1 | Luxury Wristwatch Data | 138 |
| A.2 | Palm Pilot M515 Data | 138 |
| A.3 | Xbox Data | 140 |

List of Tables

| | | |
|-----|---|-----|
| 2.1 | Simulated Semi-Continuous and Transformed Data Descriptive Statistics | 35 |
| 2.2 | Simulated Semi-Continuous and Jittering Transformed Lognormal Data GOF Statistics | 41 |
| 2.3 | Simulated Semi-Continuous and Jittering Transformed Weibull Data GOF Statistics | 42 |
| 2.4 | Simulated Semi-Continuous and Jittering Transformed Normal Data GOF Statistics | 43 |
| 2.5 | Simulated Semi-Continuous and Local Regeneration Transformed Data GOF Statistics | 50 |
| 2.6 | Surplus and Transformed Data Descriptive Statistics | 54 |
| 2.7 | Surplus and Transformed Data GOF Statistics | 54 |
| 5.1 | Growth Models Formulas | 104 |
| 5.2 | Parametric and Nonparametric Comparison | 119 |
| 5.3 | Distribution of Chosen Smoothing Method | 120 |
| 5.4 | Elapsed Run Time for Different Smoothing Methods | 123 |
| A.1 | Wristwatch Descriptive Statistics | 139 |
| A.2 | Palm Pilot M515 Descriptive Statistics | 141 |
| A.3 | Xbox Descriptive Statistics | 143 |

List of Figures

| | | |
|------|--|----|
| 1.1 | Bid History and Auction Attributes for Rolex Auction | 3 |
| 1.2 | Proxy and Live Bid Plot for Rolex Auction | 4 |
| 2.1 | Max-Bin Histograms for Real World Semi-Continuous Data | 18 |
| 2.2 | Surplus Probability Plots | 20 |
| 2.3 | Surplus Residuals | 21 |
| 2.4 | Lognormal Max-Bin Histograms | 37 |
| 2.5 | Weibull Max-Bin Histograms | 38 |
| 2.6 | Normal Max-Bin Histograms | 39 |
| 2.7 | Simulated Semi-Continuous and Jittering Transformed Data GOF Scree Plots | 45 |
| 2.8 | Simulated Semi-Continuous and Jittering Transformed Data Kernel Density Estimates | 48 |
| 2.9 | Simulated Semi-Continuous and Local Regeneration Transformed Data GOF Scree Plots | 51 |
| 2.10 | Transformed Surplus Max-Bin Histograms and Weibull Probability Plots | 52 |
| 2.11 | Transformed Surplus GOF Scree Plots | 55 |
| 2.12 | Transformed Surplus Residuals | 55 |
| 2.13 | Transformed Surplus Probability Plots | 56 |
| 4.1 | eBay Website Resulting From Palm Pilot Search | 71 |
| 4.2 | Live Bids and Monotone Smoothed Price Curves for 3 Palm Pilot Auctions | 75 |
| 4.3 | Palm Pilot Price Rug Plots (Monotone Smooth) | 77 |
| 4.4 | Palm Pilot Velocity Rug Plots (Monotone Smooth) | 78 |

| | | |
|------|--|-----|
| 4.5 | Palm Pilot Price Rug Plot (Monotone Smooth) Coded By Auction Duration | 83 |
| 4.6 | Palm Pilot Closing Price Time Grouped Box Plots (Width Proportional to Number of Auctions) | 85 |
| 4.7 | Palm Pilot Time Grouped Closing Price Box Plots and Auction Frequency Histograms | 86 |
| 4.8 | Palm Pilot Time Grouped Closing Price Box Plots (Notched, Overlapping, and Zoomed-In) | 89 |
| 4.9 | Palm Pilot Closing Price Moving Mean Plot | 90 |
| 4.10 | Palm Pilot Closing Price Autocorrelation Plot | 92 |
| 5.1 | Daily Bid Distribution for Wristwatch Data | 98 |
| 5.2 | Wristwatch Exponential Dynamics | 105 |
| 5.3 | Wristwatch Logarithmic Dynamics | 105 |
| 5.4 | Wristwatch Logistic Dynamics | 108 |
| 5.5 | Wristwatch Reflected-Logistic Dynamics | 108 |
| 5.6 | Wristwatch Model Selection Overweighting Y Dimension | 115 |
| 5.7 | Wristwatch Model Selection Overweighting X Direction | 115 |
| 5.8 | Price Curves for Growth Models Fit to Wristwatch Auctions | 116 |
| 5.9 | Growth Model Distribution for Wristwatch Data | 117 |
| 5.10 | Growth Models and Nonparametric Smoothing of Sample Wristwatch Auctions | 121 |
| 5.11 | Palm Pilot and Xbox Rug Plots (Growth Models), Jittered Closing Models, and Temporally Grouped Model Bar Plots | 125 |
| 5.12 | Palm Pilot Rug Plots (Growth Models) Broken Out By Duration | 127 |
| 5.13 | Xbox Rug Plots (Growth Models) Broken Out By Condition | 128 |

List of Abbreviations

| | |
|-----|-------------------------------|
| EDA | Exploratory Data Analysis |
| FDA | Functional Data Analysis |
| GCV | Generalized Cross Validation |
| LHS | Left Hand Side |
| OLS | Ordinary Least Squares |
| PCA | Principal Components Analysis |

Chapter 1

Introduction to Online Auctions

1.1 Online Auctions

With the advent of the Internet came online auction marketplaces, such as the popular eBay.com, which allow consumers and businesses to sell, buy, and bid on a variety of different goods. Ebay is the largest consumer-to-consumer (C2C) marketplace and touts net revenues that topped \$1 billion for the first time for the first quarter of 2005 and were close to \$1.4 billion (36% higher) for the first quarter of 2006. Ebay's popularity is further reflected by the amount and growth of participation. For example, at the end of the first quarter 2006, there were 192.9 million registered users (up 31% from the previous year) and 75.4 million active users: those who had bid, bought or listed in the previous 12 months (up 25% from the previous year). On any given day, several million items, dispersed across thousands of categories, are available for sale on eBay. There were 575.4 million new listings in the first quarter of 2006 (up 33% from the first quarter in 2005). Indeed, eBay's slogan, "What ever it is, you can find it on eBay," is appropriate. More information on eBay's success can be found at http://www.findarticles.com/p/articles/mi_m0EIN/is_2006_April_19/ai_n16127113.

Online auctions are different from traditional brick-and-mortar auctions in that they occur simultaneously or within close temporal proximity. Online auctions

tend to be longer, ranging over several days rather than minutes. Further, online auctions are not limited by the geographic barriers of traditional auctions. Buyers and sellers may be located on different continents and still conduct business since the online auction marketplace is always open and available.

1.2 eBay Data Structure

The number of different online auction sites is growing steadily. Despite different formats and rules, there is a common data structure that can be found across most sites. This structure comprises of a time series that describes the bids placed over time (the bid history) and an associated set of features that describe the auction setting, such as the seller rating, the auction duration, and the item category. We refer to these features as the auction attributes. Figure 1.1 provides a snapshot of a closed auction from eBay.com showing the auction attributes (top) and the bid history (bottom). We see that this is a 7-day auction for a Vintage Rolex Submariner Black Dial Men's Wristwatch. The seller *tutleandwabbit* has a feedback rating of 100 with 100% positive feedback. The closing price is \$2600.00, and there are a total of 13 bids from 9 bidders. In this case, eBay has decided to alias the bidders' user names in order to protect them from fake offers. Aliasing is performed quite often for high end merchandise.

To understand the structure of bid history data, it is necessary to understand the auction rules and bidding mechanism. On eBay, the majority of auctions are second-price auctions, which means that the winner is the bidder who places the

eBay.com Bid History for
 Vintage Rolex Submariner Black Dial Mens Wrist Watch (Item # 320068139586)
 Listed in category: [Jewelry & Watches](#) > [Watches](#) > [Wristwatches](#)

Winning bid: US \$2,600.00
 Ended: Jan-10-07 19:43:14 PST
 Starting time: Jan-03-07 19:43:14 PST
 History: [13 bids](#)
 Starting bid: US \$0.99

Winning bidder: [keefner10](#) ([120](#) ★)

Seller: [tutleandwabbitt](#) ([100](#) ★)

Feedback: 100% Positive

Member: since Jan-06-04 in United States

Item location: Bayside, New York, United States


| Bidder  | Bid Amount | Date of bid |
|--|---------------|------------------------|
| Bidder 8 ★ | US \$2,600.00 | Jan-10-07 18:22:24 PST |
| Bidder 9 ★ | US \$2,600.00 | Jan-10-07 19:35:08 PST |
| Bidder 7 ★ | US \$2,500.00 | Jan-10-07 13:07:05 PST |
| Bidder 6 ★ | US \$2,300.00 | Jan-10-07 05:46:18 PST |
| Bidder 5 | US \$2,250.00 | Jan-09-07 01:23:57 PST |
| Bidder 3 ★ | US \$2,150.00 | Jan-04-07 11:55:31 PST |
| Bidder 4 ★ | US \$2,000.00 | Jan-04-07 13:43:07 PST |
| Bidder 3 ★ | US \$1,500.00 | Jan-04-07 11:55:24 PST |
| Bidder 1 ★ | US \$1,000.00 | Jan-04-07 06:40:26 PST |
| Bidder 3 ★ | US \$1,000.00 | Jan-04-07 11:55:14 PST |
| Bidder 3 ★ | US \$500.00 | Jan-04-07 11:55:06 PST |
| Bidder 3 ★ | US \$100.00 | Jan-04-07 11:54:54 PST |
| Bidder 2 ★ | US \$55.56 | Jan-04-07 09:38:09 PST |

Figure 1.1: Time series and attributes for a men's Rolex wristwatch auction. Notice that bids are arranged in descending order by bid amount. This order, however, does not reflect the arrival of the bids. Rather, it reflects the current auction high bid.

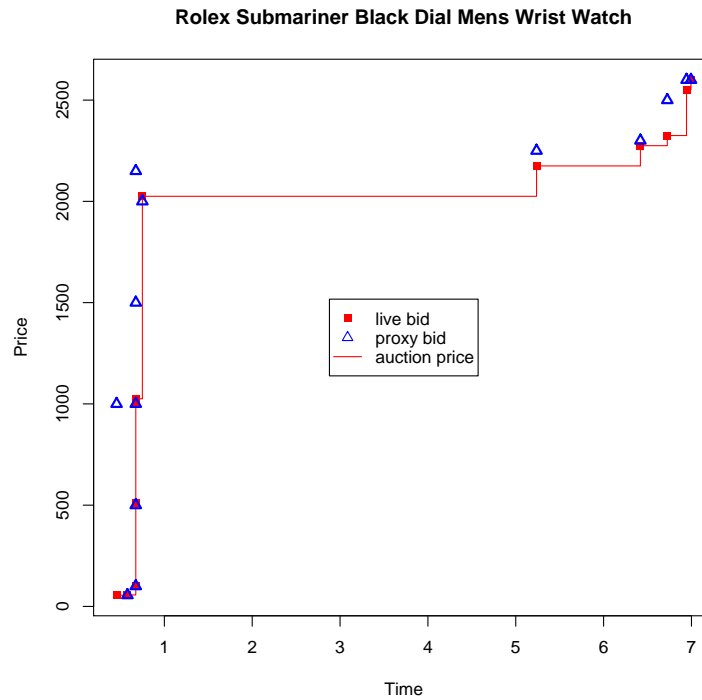


Figure 1.2: Proxy bids (triangles) and live bids (squares) for a men’s Rolex wrist-watch auction. The line connecting the live bids represents current auction price.

highest bid, but s/he pays the second highest price plus an increment. In our auction, Bidder 8 placed the highest bid but paid only Bidder 9’s bid. eBay does not disclose the highest bid (here, by Bidder 8). Furthermore, eBay uses a so-called “proxy bidding” system where bidders place the highest value that they are willing to pay, and then eBay bids on their behalf by increasing the current price by only an increment. For further details see <http://pages.ebay.com/help/buy/proxy-bidding.html>. During the auction, the “current high bid” displayed is actually the second highest bid at the time plus an increment. Figure 1.2 shows this for the auction in Figure 1.1. We call the current price “live bid.” We see that Bidder 9 placed a proxy bid of \$2600 at 19:35:08. Bidder 8 placed an unknown higher bid at 18:22:24 so is credited with the highest bid and wins the auction.

1.2.1 eBay Bid Level Data Sets

We use predominantly three eBay.com bid level data sets throughout this research. All of the data sets contain the auction ID, bids, bid dates and times, opening price, start date and time, closing price, and end date and time. In some cases, we have additional auction attributes such as unique bidder ID and rating, seller ID and rating, and item description.

Our data are bid histories of closed auctions for luxury wristwatches (Rolex and Cartier), Palm Pilot M515 personal handheld organizers, and Xbox game consoles. The luxury wristwatch data set includes a wide variety of items across a wide range of prices. We use this data set when the goal is to capture very different price behavior. The Palm Pilot and Xbox data sets are for the same or similar items. These data sets are used for illustration in cases where the focus is on a single item such as competing concurrent auctions. A full description of each data set as well as descriptive statistics is given in Appendix ???. In some cases, we break down the descriptive statistics by auction attributes since that information is later used in the analysis.

1.3 Online Auction Literature

The popularity of online auctions as well as the rich data sets has encouraged many empirical research studies. Although empirical research has flourished, the great majority of work is by nonstatisticians; therefore, the statistical methods used tend to be standard “off-the-shelf” tools such as regression models. We describe

some of the major contributions to the online auction literature to date. The article by Bajari and Hortascu (2004) provides an extensive overview of the literature by describing various auction attributes and documenting what has been found to date. It is a first-read for anyone interested in studying online auctions.

Lucking-Reiley et al. (2005) conduct one of the first eBay empirical research studies (draft available in 1999). They examine, via regression analysis, the determinants of the final price of auctions where the item, in this case coins, have a known book value. They find that a seller's rating has an effect on the final price; however, a negative rating has a more significant effect than a positive rating (i.e., a negative seller rating will decrease the final selling price much more than a positive rating will increase it). They also find that the length of an auction plays an important role in the final price: longer auctions (7- or 10-days) tend to fetch a higher closing price. Finally, they have mixed results for setting a reserve price (the minimum amount a seller is willing to part with the item for). Although a reserve price significantly increases the final price of the auction, they are quick to note that a reserve, especially a secret reserve, decreases the likelihood of the auction transacting.

Bajari and Hortascu (2003) investigate the determinants of bidder and seller behavior. In terms of bidders, they find that last-moment bidding, often referred to as 'sniping', occurs quite frequently in online auctions and usually results in the winning bid. They notice that sellers tend to set opening bid values much less than the market price of the item, which is thought to encourage bidder entry. Further, reserve prices are usually limited to high-end products.

Roth and Ockenfels (2002) examine bidding behavior based on the rules for ending an auction. For example, eBay has a hard close where the auction has a known fixed end while Amazon auctions may have an unknown end. An Amazon auction remains open (after the specified close) as long as a bid has been placed in the previous 10 minutes. They find that the hard close gives eBay bidders an incentive to bid late, and it is the experienced bidders that tend to bid late. On Amazon, there is still a lot of late bidding, but this tends to be done by inexperienced bidders.

Bapna et al. (2004a) examine bidder-level (rather than auction-level) data to learn about bidders' preferences, behaviors, and economic welfare. Using explanatory variables: time of entry, time of exit, and number of bids (which is a proxy for a bidder's time valuation), they use K-means clustering and find five distinct bidder classes. Evaluators place only one bid in the beginning or middle of the auction, perhaps their willingness to pay. Opportunists bid late in the auction and are most similar to snipers. Sip-and-dippers place a bid in the beginning of the auction (which establishes time priority) and revise their bid at the end of the auction sequence. Participators bid frequently, start bidding early, and place their final bid late, which is the profile of a typical "gamer". Finally, agent bidders use automatic bidding agents to place their bid for them. They are similar to participators except that the agent monitors bidding so they do not have to. Bapna et al. (2004a) find that agent bidders and then participators have the highest level of surplus while opportunists and sip-and-dippers have the highest likelihood of winning the auction. They also find that while experienced bidders are usually the ones who snipe, inexperienced

bidders who do not understand eBay's proxy bidding mechanism tend to bid high early in the auction.

Kauffman and Wood (2005) examine how the fee structure on eBay may motivate bid shilling, the act of bidding in one's own auction. There are two types of shilling: competitive and reserve price. Competitive shilling is where a seller bids in his own auction to increase the price that other bidders need to pay to win the item. Reserve price shilling is where the seller sets a low opening price but bids his reserve in order to avoid the fee associated with setting a secret reserve price on eBay. They are able to detect when shilling has occurred (for example, if a bidder typically bids high in the same seller's auctions and drops out early) and predict which auctions are likely to be shilled.

Recently, Dellarocas and Wood (2007) examined the reporting bias that occurs in the eBay (or any two-party) feedback mechanism. They find that people are more likely to leave positive feedback when they are satisfied than negative feedback when they are dissatisfied. This is likely due to fear of retaliation that would blemish their own ratings. They argue that not leaving feedback carries important information and should also be reported on a trader's profile.

More recently, statisticians have become involved in online auction research. Notably, there is a special issue of *Statistical Science* (May 2006) devoted entirely to "statistical challenges and opportunities in electronic commerce research." Research by statisticians has focused on statistical challenges and proposes suitable probabilistic and statistical approaches. For example, Shmueli et al. (2004) model bid arrivals with a nonhomogenous Poisson process. They find that there is some

bidding activity at the auction start (first day), followed by a period of very little activity, culminating in a surge of bidding at the very end of the auction.

There has also been a trend in studying the price formation and price dynamics rather than just the closing price of an auction. Jank and Shmueli (2005) explore price dynamics by representing the price curve as a smooth continuous function. They use curve clustering to group similarly shaped price curves and compare the price dynamics of the clusters. They then look at auction attributes such as opening bid, seller rating, etc. within each cluster to gain insight into the relationship between price dynamics and auction attributes. The first cluster is thought of as experienced buyer/seller auctions. They have a moderately high opening price and low competition but still achieve a high closing price. The second cluster is formed by greedy sellers, who post a high opening price and thus attract little competition which results in a low closing price. Finally, there are Bazaar auctions, which are characterized by a low opening price, high competition, and the highest closing prices.

Several recent papers in the eCommerce literature have emphasized the importance of the price formation process and dynamics in online auctions (Jank and Shmueli, 2006; Góes et al., 2007). For example, the price-velocity indicates how quickly the price is changing at every point during the auction. This knowledge can be used to build powerful dynamic forecasting models for price (Wang et al., 2007a). It can also be used to visually “mine” a database of auctions for the same or similar products (Shmueli and Jank, 2005; Hyde et al., 2006; Shmueli et al., 2006; Jank et al., 2007).

Hyde et al. (2004) use functional principal components analysis (PCA) to examine the variability in price over the duration of an auction. They find that price varies between auctions mostly in the middle portion of the auction and that a high opening bid leads to little variability in the middle part of the auction, most likely due to a market value for the item. Auctions vary next most in the distance (range) between opening and closing price.

This list is by no means exhaustive; however, these early results motivated some of our research questions. Our contribution is described next in Section 1.4.

1.4 Contributions of this Dissertation

The wide and growing popularity of online auctions creates enormous amounts of publicly available auction bid data providing an interesting and important topic for research. These data also pose unique statistical challenges because of their special structures. Although various statistical methods have been used thus far in the online auction context, they are usually used in a very basic manner, not always accounting for the special data structures that arise. For that reason, there has mostly been a neglect of the entire bid history with a concentration more on the end (final price) of the auction. This dissertation examines some of the special data structures and develops or adapts statistical methods for representing, visualizing, and modeling such data.

1.4.1 Semi-Continuous Transformation

Buyers and sellers participate in online auctions for a variety of reasons. The success of online auctions is due in large part to the ability to bring buyers and sellers together in a manner that may not have previously existed, especially when there are geographic limitations. This is the case in particular for unique or hard-to-find items. The online marketplace can provide a good approximation of the true market value; however, due to imperfect information, there is the opportunity for sellers to make an usually large profit or bidders to get a “steal”.

eBay auctions are second-price, meaning that the winner is the bidder who places the highest bid but pays the second highest price plus a bid increment. Bapna et al. (2005) quantify the overall consumer surplus, the difference between the price paid for an item and the amount that would have been paid, that eBay produces.

Consumer surplus is inherently nonnegative and continuous. However, the data set that Bapna et al. (2005) use is semi-continuous in that there are numerous “too-frequent” values. This is easily seen in a histogram of the consumer surplus data. Bapna et al. (2005) remedy this problem by imposing a ceiling transform, where the surplus values are rounded to the next integer value, reasoning that surpluses within the same \$1 unit carry the same or similar information. This may or may not be true, but it would be better to approximate the underlying continuous distribution that would exist if a discretization contamination mechanism, like the tendency to bid “round” values, did not exist. Further, only the “too-frequent” values should be transformed since they are the only values contaminated.

Our first contribution is a new set of two transformations for semi-continuous data that “iron-out” the too-frequent values and produce completely continuous data. We show that the transformed data maintain the properties of the original data but are suitable for standard analysis. We also assess the performance of the transformed data. This research, discussed in Chapter 2, is also under review (Shmueli et al., 2007).

1.4.2 Visualizing Concurrency

A great majority of the literature that analyzes online auction data assumes *independence* across auctions. This assumption is typically made for reasons of simplicity and convenience, while in reality auctions for the same item, competing items, or even related (substitute) items will influence each other especially if they take place within a close time frame. On eBay.com, an identical product is often sold in numerous simultaneous auctions. While each auction contains a replicate of the product, the resulting sales, prices, and even the number and level of bids during the ongoing auctions are clearly not independent of each other. In such situations, it is likely that there is dependence between the auctions since buyers have the option to select which of the competing items to bid on, and sellers have the option to decide when to post their item for sale using information on similar previously sold items.

We approach the study of concurrency from a visual point of view and propose a series of visualizations that are suitable for the special structure of bid data.

Our major contribution is the *rug plot* which graphically displays the entire price-evolution curves or dynamics curves over calendar time. Other displays that we use are time-grouped box plots and moving statistics plots. We find price trends and relationships between prices in concurrent auctions and raise new research questions. While our main contribution is the rug plot, this research is one of the few to date that considers the dependence among online auctions. Although our analysis takes place in the online auction context, it may be used whenever concurrency of events is prevalent. This research is discussed in Chapter 4 and has been published (Hyde et al., 2006).

1.4.3 Growth Models

In order to analyze an auction's price process, it must be captured with a statistical structure or model. This is a challenging task because we start with a nonstandard time series: bids during an online auction arrive at unequally-spaced discrete time points with sometimes very sparse and/or dense bidding periods. Our goal is to achieve a continuous smooth representation of the auction's price process that is estimated from the discrete bid data.

Previous work treating bid histories as functional objects has focused on non-parametric smoothing methods. Jank et al. (2007) use penalized polynomial smoothing splines due to their computational efficiency and good balance between fit and smoothness. Unfortunately, splines do not necessarily yield a monotone nondecreasing function, as the auction price process demands. They also suffer from extra

volatility at the start and end of the auction, which are of special interest, particularly in forecasting scenarios. In comparison, monotone smoothing used by Hyde et al. (2006) achieves a monotone price process, but it is computationally intensive which might not be useful for large auction data sets. Both methods require estimating a large number of coefficients and for numerous parameters to be set in advance by the user. Finally, both methods are data-driven and thus do not offer a theoretical explanatory model of the process captured. This motivated us to explore the use of meaningful parametric representations.

An important contribution of this research is to introduce a parametric family of growth models that can capture a variety of possible underlying price processes. These are exponential growth, logarithmic growth, logistic growth, and reflected-logistic growth, where “reflected” refers to a mapping of the original model over the line $y = x$. We also develop an automated model selection procedure that uses a specialized proximity metric that measures the distance between bids and the fitted curve in the two dimensions of price and time. The parametric approach is also computationally fast, more parsimonious, and generally more insightful. This research, discussed in Chapter 5, is also under review (Hyde et al., 2007).

This class of parametric growth models can be used not only for data representation and visualization, but also for formal modeling of the auction processes. Statistical methods such as functional linear regression models, functional PCA, and curve clustering can be applied to the parametric objects in order to test hypotheses and model relationships between the price process (and its dynamics) and factors such as auction design and concurrency.

1.4.4 Summary of Dissertation Contributions

To summarize, the contributions of this dissertation are to:

1. Propose a transformation for semi-continuous data found in online auction surplus data (Chapter 2).
2. Investigate concurrency in online auctions through visualization and the invention of the *rug plot* (Chapter 4).
3. Represent unevenly-spaced time series data as a continuous curve (function) and compare various previously used nonparametric smoothing techniques with a parametric family of growth models (Chapter 5).

Chapter 2

Transformations for Semi-Continuous Data

2.1 Introduction and Motivation

Standard statistical methods can be divided into methods for continuous data and those for discrete data. However, there are situations in which observed data do not fall in either category. In particular, we consider data that are inherently continuous but get contaminated by inhomogeneous discretizing mechanisms. Such data lose their basic continuous structure and instead are spotted with a set of “too-frequent” values. We call these semi-continuous data.

Semi-continuous data arise in various settings. Reasons range from human tendencies to enter round numbers or to report values that conform to given specifications (e.g., reporting quality levels that conform to specifications), to contamination mechanisms related to data entry, processing, storage, or any other operation that introduces a discrete element into continuous data. Examples are numerous and span various applications. In accounting, for example, a method for detecting fraud in financial reports is to search for figures that are “too common”, such as ending with 99 or being “too round”. In quality control sometimes data are manipulated in order to achieve or meet certain criteria. For example, Bzik (2005) describes practices of data handling in the semiconductor industry that deteriorate the performance of statistical monitoring. One of these is replacing data with “phys-

ical limits”: in reporting contamination levels negative values tend to be replaced with zeros and values above 100% are replaced with 100%. However, negative and >100% values are feasible due to measurement error, flawed calibration, etc. Another questionable practice is “rounding” actual measurements to values that are considered ideal in terms of specifications. This results in data that include multiple repetitions of one or more values. We use the term “too-frequent” to describe such values.

We encountered two particular studies where semi-continuous data are present. The first is a research project by IBM on customer wallet estimation, where the marketing experts who derived the estimates tended to report round estimates thereby leading to data with several too-frequent values (Perlich and Rosset, 2006). A second study, which motivated this work, studies consumer surplus in the online marketplace eBay. Here, the observed surplus values had several too-frequent values, most likely due to the discrete set of bid increments that eBay uses and the tendency of users to place integer bids. The top panels in Figure 2.1 show the frequencies of the values in samples from each of these two data sets. In the eBay surplus data (left panel) the value \$0 accounts for 6.35% of the values, and values such as \$0.01, \$0.50, \$1.00, \$1.01, \$1.50, \$2.00, \$2.01, \$3.00 are much more frequent than their neighboring values. In the IBM customer wallet estimates, too-frequent values are 0, 100, 200, 250, 300, and 400.

We use the surplus data throughout this chapter to illustrate and motivate the proposed transformations. We therefore describe a few more details about the mechanism that generates the data. Consumer surplus, used by economists to mea-

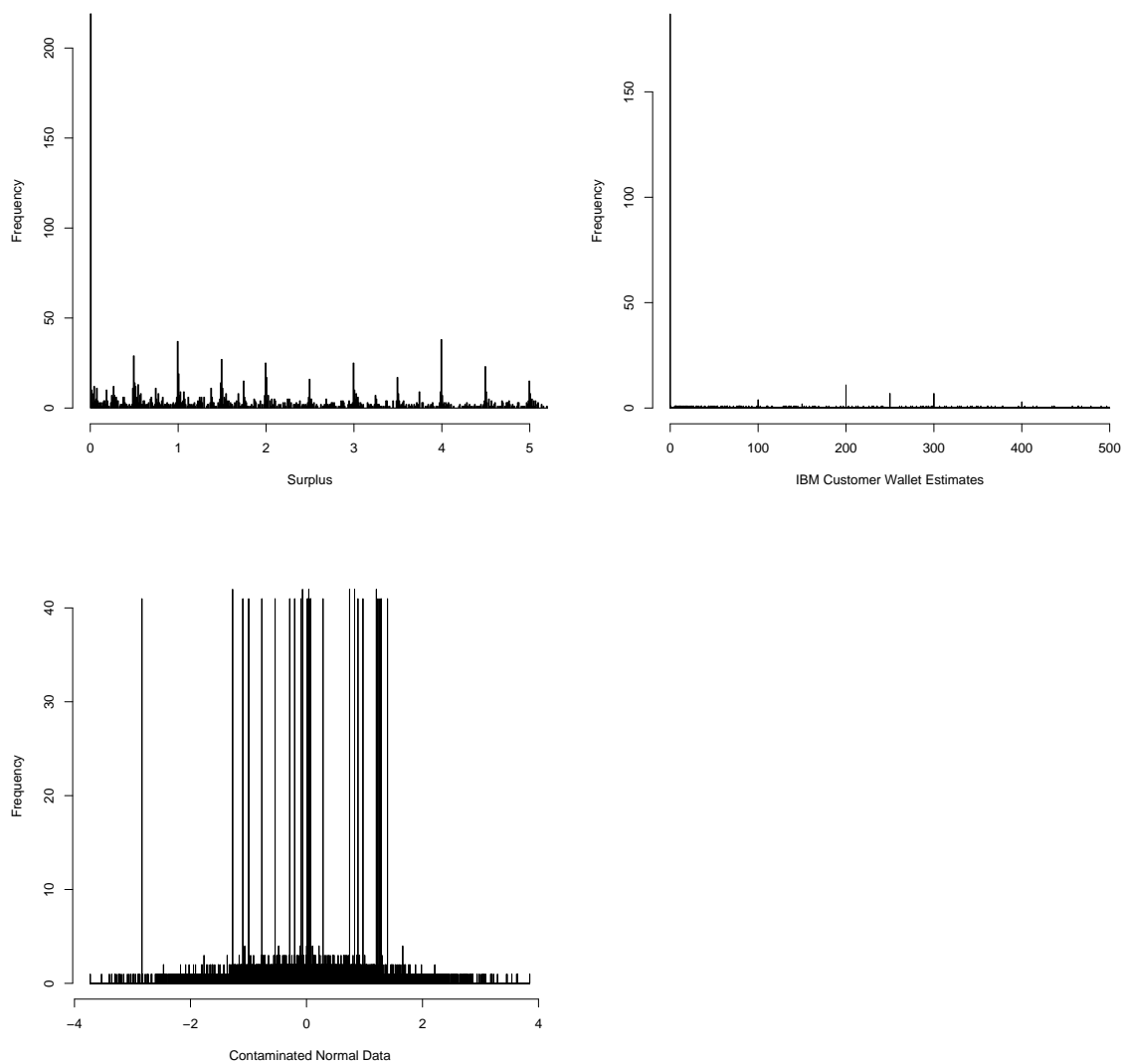


Figure 2.1: Frequencies of values for three data sets: eBay surplus data (top), IBM customer wallet estimates (middle), and contaminated normal simulated data (bottom).

sure consumer welfare, is the difference between the price paid and the amount that consumers are willing to pay for a good or service. In second-price auctions, such as those on the famous online marketplace eBay, the winner is the highest bidder; however, s/he pays a price equal to the second highest bid. Surplus in a second-price auction is defined (under some conditions) as the difference between the price paid and the highest bid. Bapna et al. (2005), who investigate consumer surplus in eBay, find that although surplus is inherently continuous, observed surplus data are contaminated by too-frequent values, as shown in the top left panel of Figure 2.1.

Using the term “contamination” in this case is somewhat artificial because it reflects human behavior. However, the result is similar to the contamination of a physical process by machine error if we assume that humans, acting rationally, would bid their true valuation which may or may not be a round number. For example, a person may value a product for \$5.00 but bid \$5.03 corresponding to a May 3 birthday, which would not be rational behavior. This type of “contamination”, although quite possibly present, is difficult to identify in the data set. In this sense, the surplus data is a “contaminated” version of the continuous bidding process.

Semi-continuous data can appear as if they come from a mixture of discrete and continuous populations. Graphs such as scatter plots and frequency plots might indicate segregated areas or clouds. Such challenges arise in the surplus data described above. Figure 2.2 displays two probability plots for $\log(\textit{surplus} + 1)$: the first is a lognormal fit and the second is a Weibull fit. It is obvious that neither of these two models approximate the data well (other distributions fit even worse) because there are too many 0 values in the data. Furthermore, using $\log(\textit{surplus} + 1)$ as

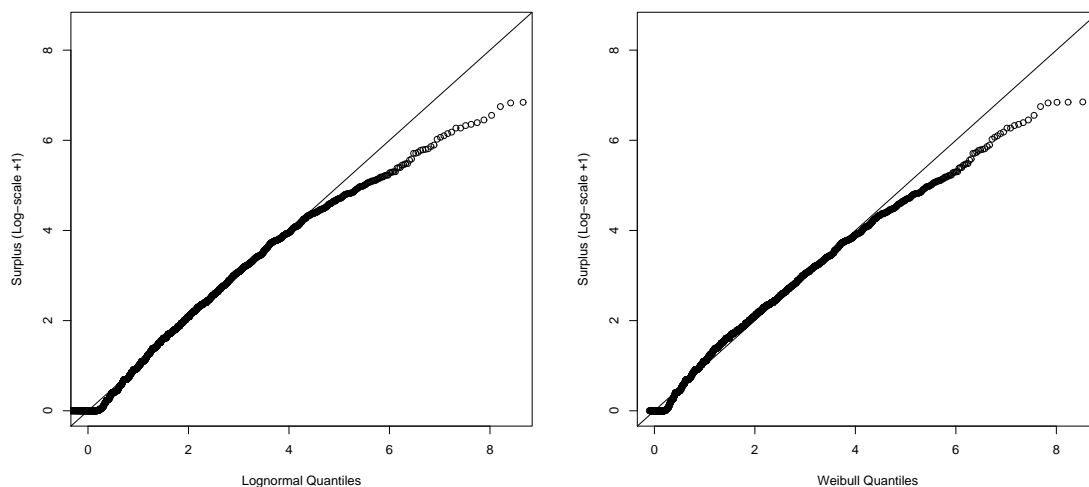


Figure 2.2: Probability Plots for $\log(\textit{surplus} + 1)$: lognormal fit (left) and Weibull fit (right).

the response in a linear regression model yields residual plots that exhibit anomalies that suggest a mixture of populations. Figure 2.3 shows two residual plots exhibiting such behavior. These plots indicate clear violation of the assumptions of a linear regression model.

The main problem with semi-continuous data is that they tend to be unfit for use with many standard statistical analysis methods. One possible solution is to separate the data into continuous and discrete parts, model each part separately, and then integrate the models. For example, in a data set that has too many zeros (zero-inflated) but otherwise positive, continuous values, we might create a classification model for zero/nonzero and then a prediction model for the positive data. This approach has two practical limitations: first, partitioning the data leads to loss of statistical power, and second, it requires the too-frequent values to be concentrated in a limited area or in some meaningful locations. To solve the first

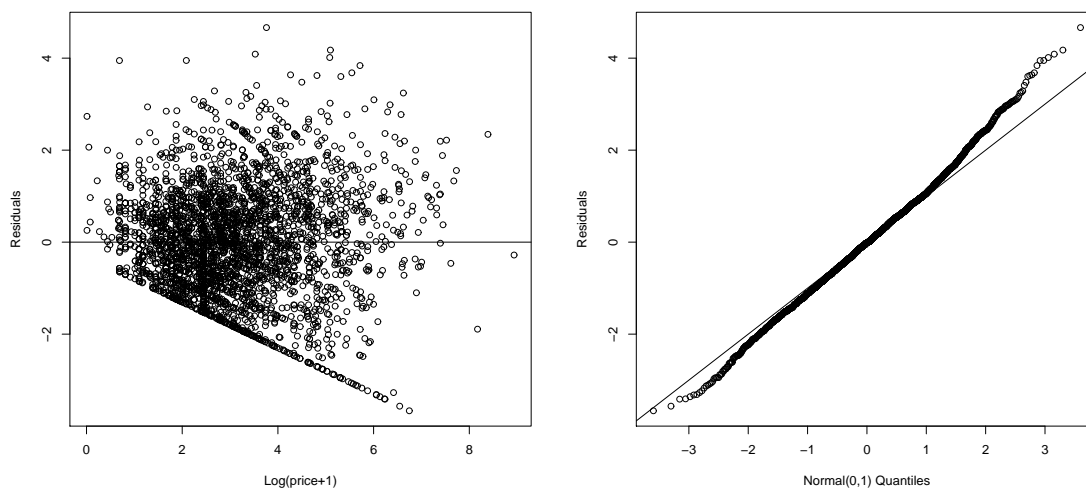


Figure 2.3: Residuals from a linear regression model of $\log(\text{surplus}+1)$ on $\log(\text{price}+1)$ (left) and normal probability plot of residuals (right).

issue one might argue for a mixture model. Although there exists a plethora of such models for mixed continuous populations or for mixed discrete populations (e.g., “zero-inflated” models, Lambert, 1992), we have not encountered models for mixtures of continuous and discrete data. Moreover, there is a major conceptual distinction between mixture and semi-continuous data: unlike mixture data, semi-continuous data are inherently generated from a single process. Therefore, treating them as a mixture of populations is artificial. The ideal solution, of course, would be to find the source of discretization in the data generation mechanism and eliminate it or account for it. However, in many cases this is impossible, very costly, or very complicated. We therefore strive for a method that “unites” the apparent “subpopulations” so that the data can be integrated into a single model and treated with ordinary models for continuous data.

Our proposed solution is a set of two transformations which yield continuous

data. We distinguish between two cases: one, where the goal is to obtain data that fit a particular continuous distribution (e.g., a normal distribution, for fitting a linear regression model); and two, where there is no particular parametric distribution in mind, but the data are still required to be continuous. The first approach, suitable when the data should fit a particular distribution, is based on binning the data in a way that is meaningful with respect to the assumed underlying distribution and then replacing the too-frequent values with randomly generated observations in that bin. We call this the *local regeneration transform*. The second approach, when the goal is simply to obtain continuous data, is based on jittering. As in graphical displays, where jittering is used to better see duplicate observations, our data transformation adds a random perturbation to each too-frequent value, thereby ironing-out the anomalous high frequencies. We call this the *jittering transform*.

Jittering is used not only in graphical displays, but also in data privacy protection. It is a common method for disguising sensitive continuous data while retaining the distributional properties of the sample that are required for statistical inference. The difference between this application of jittering and our proposed jittering transform is that unlike in data privacy protection, we do not jitter each observation but rather only the too-frequent values.

Local regeneration is related to smoothing histograms and rebinning of histograms; however, there are two fundamental differences. First, the assumption about the data origin is different. In histogram smoothing and binning, unless we are dealing with a mixture model, the underlying assumption is that extreme peaks and dips in the histogram result from sampling error. Therefore, increasing the

sample size should alleviate these phenomena, and in the population such peaks would not appear at all (Good and Gaskins, 1980). In contrast, in semi-continuous data the cause of the peaks is not sampling error but rather a structural distortion created by the data generating mechanism. This means that even very large samples will exhibit the same semi-continuity. The second difference is the goal: whereas histogram smoothing and binning are used mainly for density estimation (without attempting to modify the data), the purpose of local regeneration is to transform the semi-continuous data into data that fit a particular continuous parametric distribution, similar to the famous Box-Cox transformation.

Local regeneration and binning both attempt to find the best representation of the data in histogram form. The literature on optimal bin sizes focuses on minimizing some function of the difference between the histogram $\hat{f}(x)$ and the real density $f(x)$ (such as the mean integrated squared error, or MISE). In practice, $f(x)$ is assumed to be a particular density function (e.g., Scott (1979)) or else it is estimated by using an estimator that is known to have a smaller MISE (e.g., a kernel estimator). Methods range from rules of thumb to (asymptotically) theoretically optimal bin width formulas. The methods also vary widely with respect to computational complexity and performance. Simple bin-width formulas such as Sturges' rule and Doane's modifications, which are used in many software packages (e.g., Minitab and S-Plus), have been criticized as leading to oversmoothing (Wand, 1997). On the other hand, methods that have better asymptotic optimality features tend to be computationally intensive and less straightforward to implement in practice. In the semi-continuous data case, we assume that it is impossible to obtain a set of data

that is not distorted. Our target function is therefore a parametric distribution.

The remainder of this chapter is organized as follows. In Section 2.2 we describe a method for identifying semi-continuous data. Section 2.3 describes the jittering and local regeneration transforms in detail. We illustrate the performance of these methods using simulated data in Section 2.4 and apply them to the on-line auction surplus data in Section 2.5. Further issues and future directions are discussed in Section 2.6.

2.2 Identifying Semi-Continuous Data

To identify whether a continuous data set suffers from semi-continuity, we examine each of the unique values and see whether there is one that is *too-frequent*. One way to check this is to examine a one-way pivot table, with counts for each value in the data, for very large counts. Sorting the counts in descending order can alleviate the search process. However, since such a table loses the ordering of the values, too-frequent values that appear in sparser areas might remain undetected. An alternative, which enhances the spotting of too-frequent values while preserving the interval scale of the data is a visualization that is a hybrid between a bar chart and a histogram: the *max-bin histogram*. The max-bin histogram is essentially a histogram with bin widths equal to the smallest unit in the data. It is therefore equivalent to a bar chart with as many nonzero bars as there are unique values in the data, except that its x-axis has a continuous meaning rather than labels. On the max-bin histogram of the raw data, where frequencies are represented by

bars, frequencies that are outstandingly high compared to their neighbors indicate values suspected of being too frequent for a continuous scheme. Figure 2.1 displays three max-bin histograms. The top two are for the eBay surplus data and the IBM customer wallet estimates. As mentioned in the previous section, too-frequent values are very visible in each of these plots. The bottom panel shows a simulated data set of size 10,000, where 9,000 observations were generated from a standard normal distribution, retaining 4 decimal digits, and the remaining 1,000 observations were obtained by duplicating 25 of the generated values 40 times. The 25 too-frequent values are clearly visible in the max-bin histogram.

Using plots for continuous data can also assist in identifying semi-continuous data. For instance probability plots might display “steps” or contain other indications of mixed populations. Scatter plots of the suspected variable vs. other variables of interest can also reveal multiple clouds of data. However, depending on the locations and prevalence of the too-frequent values, standard plots might not enhance the detection of such values (e.g., if the too-frequent values are in high frequency areas of the distribution). The max-bin histogram is therefore a powerful and unique display that is ideal for this purpose.

2.3 Transforming Semi-Continuous Data

Given a set of semi-continuous data, the influence of “bouts” of high frequency values can be “ironed-out” (a term coined by Good and Gaskins (1980)) in one of the following ways, depending on the goal of the analysis. The first approach is to jitter

each of the too-frequent values, which means that we add a random perturbation to each such observation. The second approach defines local neighborhoods by binning the data and then replaces the too-frequent values with randomly generated observations within their respective neighborhoods. The two methods are similar in several respects. First, they both assume that the too-frequent values are distortions of other, nearby values. Second, both methods define a local neighborhood for the too-frequent values, and the transformed data are actually randomly generated observations from within this neighborhood. Third, in both cases only the too-frequent values are replaced while the other observations remain in their original form. And finally, in both cases the definition of a local neighborhood must be determined. We describe each of the two methods in detail next.

2.3.1 The Jittering Transform

Although many statistical methods assume a parametric distribution of the data, there are many nonparametric methods that only assume a continuous nature. This is also true of various graphical displays, which are suitable for many data structures, as long as they are continuous (e.g., scatterplots). After identifying the too-frequent values, the transform operates by perturbing each of them by adding random noise. If we denote the i th original observation by X_i and the transformed observation by \tilde{X}_i , then the jittering transformation is

$$\tilde{X}_i = \begin{cases} X_i + \epsilon & \text{if } X_i \text{ is a too-frequent value} \\ X_i & \text{else} \end{cases}, \quad (2.1)$$

where ϵ is a random variable with mean 0 and standard deviation σ_ϵ . The choice of distribution for ϵ depends on the underlying process that the too-frequent values are most likely a result of. If there is no reason to believe that the too-frequent values are a result of an asymmetric processes, then a symmetric distribution (e.g., normal or uniform) is adequate. If there is some information on a directional distortion that leads to the too-frequent values, then that should be considered in the choice of the perturbing distribution. In general, the choice of distribution is similar to that in kernel estimation, where the choice of kernel is based on domain knowledge, trial-and-error, and robustness.

The second choice is the value of σ_ϵ , which should depend on the scale of the data and its practical meaning within the particular application. Specifically, domain knowledge should guide the maximal distance from a too-frequent value that can still be considered reasonable. For example, if the data are measurements of product length in inches, where the target value is 10 inches and a deviation of more than an inch is considered “large”, then σ_ϵ should be chosen such that the jittering will be less than one inch. This, in turn, defines σ_{max} , the upper bound on σ_ϵ . A series of increasing values of σ_ϵ is then chosen in the range $(0, \sigma_{max}]$, and the performance of the transformation is evaluated at each of these values in order to determine the adequate level.

Let 2δ be the width of the jittering neighborhood (symmetric around the too-frequent value). The mapping between δ and σ_ϵ depends on the jittering distribution. If the perturbation distribution for ϵ is normal, then σ_ϵ should be chosen such that $3\sigma_\epsilon = \delta$ since almost 100% of the data should fall within 3 standard deviations of the

mean. If the underlying distribution is $\text{Uniform}(a, b)$, then $2\delta = b - a$ and therefore $\sqrt{3}\sigma_\epsilon = \delta$.

Since jittering is performed in a way that is supposed to (inversely) mimic the data mechanism that generated the too-frequent observations, the transformed data should be similar to the unobserved, continuous data in the sense that they are both samples from the same distribution. In fact, since we only transform the few too-frequent values, most of the observations in the original and transformed data are identical.

2.3.2 The Local Regeneration Transform

Many standard statistical methods assume a particular parametric distribution that underlies the data. In that case the jittering transform can still be used, but it might require additional tweaking (in the choice of the perturbation distribution and of σ_ϵ) in order to achieve a satisfactory parametric fit. The reason is that the jittering operation is anchored around the too-frequent values, and the locations of these values is independent of the unobservable, underlying continuous distribution.

A more direct approach is therefore to anchor the transformation not to the too-frequent values, but rather to anchors that depend on the parametric distribution of interest. We achieve this by first binning the observed data into bins that correspond to percentiles of the distribution. In particular, we create k bins that have upper bounds at the $\frac{100}{k}, 2\frac{100}{k}, \dots, k\frac{100}{k}$ percentiles of the distribution. For example, for $k = 10$ we create 10 bins, each as wide as the corresponding spacing

between the two distribution deciles. Each bin now defines a local neighborhood in the sense that values within a bin have similar densities. This will create narrow bins in areas of the distribution that are high density and much wider bins in sparse areas such as tails. The varying neighborhood width differentiates this transformation from the jittering transform, which has constant sized neighborhoods. This means that the main difference between the two transforms will be in areas of low density that include too-frequent values.

To estimate the percentile of the distribution we use the original data. If the continuous distribution belongs to the exponential family, we can use the sufficient statistics computed from the original data for estimating the distribution parameters and then compute percentiles. Because too-frequent values are assumed to be close to their unobservable counterparts, summary statistics based on the observed data should be sufficiently accurate for initial estimation. In the second step, each too-frequent value is replaced with a randomly generated observation within its bin. As in the jittering transform, the choice of the generating distribution is guided by domain knowledge about the nature of the mechanism that generates the too-frequent values (symmetry, reasonable distance of a too-frequent value from the unobservable value, etc.).

The local regeneration transform for the i th observation can be written as

$$\tilde{X}_i = \begin{cases} l_j + \epsilon & \text{if } X_i \text{ is a too-frequent value} \\ X_i & \text{else} \end{cases}, \quad (2.2)$$

where l_j is the j th percentile closest to X_i from below (i.e., the lower bound of the bin), $j = 1, 2, \dots, k$, and ϵ is a random variable with support $[l_j, l_{j+1})$.

The two settings that need to be determined are therefore the parametric distribution and the number of bins (k). For the distribution, one might try transforming the data according to a few popular distributions, and choose the distribution that best fits the data (according to measures of goodness-of-fit). The number of bins can be set in a similar way: trying a few configurations and choosing the one that yields the best fit. In both cases, of course, practical considerations and domain knowledge must be integrated into the decision process.

2.3.3 Goodness of Fit and Deviation

Our goal is to find a level of jittering or binning that sufficiently “irons-out” the discrete bursts in the data and creates a data set that is continuous and perhaps fits a parametric distribution. In the parametric fitting (using the local regeneration transform), we can measure how well the data fit prespecified distributions at each level of binning. For example, the three most popular distributions used for fitting data and which serve as the basis for many statistical analyses are the normal, lognormal, and Weibull distributions. We can evaluate how well data are approximated by each of these distributions by using graphical tools, such as probability plots, and goodness-of-fit measures, such as the Anderson-Darling statistic, Kolmogorov-Smirnov statistic, Chi-squared statistic, and the correlation based on a probability plot.

In the nonparametric case where the goal is to achieve continuous data without a specified parametric distribution, we use the notion of local smoothness. Treating

the max-bin histogram as an estimate of a continuous density function, we require frequencies of neighboring bins to be relatively similar. One approach is to look for the presence of an “abnormality” in the max-bin histogram by comparing each frequency to its neighbors. Shekhar et al. (2003) define an abnormal point as one that is extreme relative to its neighbors (rather than relative to the entire data set). They construct the following measure of deviation, α_i :

$$\alpha_i = p_i - E_{j \in N_i}(p_j), \quad (2.3)$$

where $p_i = f_i / \sum_j f_j$ is the observed relative frequency at value i , N_i is the set of neighbors of p_i , and $E_{j \in N_i}(p_j)$ is the average proportion of the neighboring values. They then assume an a priori distribution of the data or their histogram, which allow them to determine what is considered an abnormal value for α_i . In our context, the underlying distribution of the data is assumed to be unknown. We therefore use an ad-hoc threshold: values i that have α_i larger than 3 standard deviations of α_i are considered abnormal (assuming that deviations are approximately normal around 0).

An alternative approach is to measure smoothness of the max-bin histogram by looking at every pair of consecutive frequencies ($f_i - f_{i+1}$). Various distance metrics can be devised based on these pairwise deviations, such as the sum-of-absolute-deviations or the sum-of-squared-deviations. We use one such measure here, the “sum of absolute deviations between neighboring frequencies”:

$$SADBNF = \sum_i |f_i - f_{i+1}|. \quad (2.4)$$

A third approach is fitting a nonparametric curve to the max-bin histogram

and measuring the smoothness of the resulting curve or the deviation of the bar heights from the curve (e.g. Efromovich (1997)).

In addition to measuring the fit to a distribution, we also measure deviation from the original data by computing summary statistics at each level of binning or jittering and comparing them to those from the original data. We examine the mean, standard deviation, and median; however, robust statistics may be considered for certain distributions and/or different contamination mechanisms.

2.4 Performance Evaluation Via Simulation

In order to evaluate the performance of the jittering and local regeneration transformations in practice, we simulate data from known distributions and then contaminate them to resemble observable semi-continuous real world data. In particular, we choose parameters and contaminations that mimic the surplus data. We apply the transformations to the data and choose the transformation parameters (δ in jittering and k in local regeneration) that best achieve the goal of approximating the underlying (unobserved) distribution or at least obtaining a continuous distribution (which should coincide with the underlying generating distribution).

2.4.1 Data Simulation

We simulate “unobservable” data from three continuous distributions: log-normal, Weibull, and normal. We then contaminate each data set in a way that resembles the surplus data. The three data sets are called Contaminated Log-

normal (CLNorm), Contaminated Weibull (CWeibull), and Contaminated Normal (CNorm). The steps for all three simulations are similar except that each simulation starts with a different underlying distribution, different too-frequent values, and different contamination spreads. The initial data, whether lognormal, Weibull, or normal, are thought to be the unobservable data that come about naturally. However, a mechanism contaminates the observed data, thereby introducing a few values with high frequencies. Some of the original characteristics of the data are still present, but the contamination makes it difficult to work with the observed data in their present form. In the following we describe the details and use the notation for the CLNorm data. Equivalent steps are taken for the CWeibull and CNorm simulated data.

We generate 3000 observations ($Y_i, i = 1, \dots, 3000$) from a lognormal distribution with parameters $\mu = 0, \sigma = 2$. We choose the too-frequent values $\{s_j\} = \{0, 0.25, 0.50, 0.75, 1.00, 1.50, 2.00, 3.00, 5.00, 10.00\}$ and contaminate 750 of the 3000 observations by replacing values that fall within $\nu = 0.10$ of a too-frequent value by that frequent value (in other words, ν is the width of the contamination neighborhood). The contaminated data are therefore obtained by the following operation:

$$X_i = \begin{cases} s_j & \text{if } Y_i \in [s_j - \nu, s_j + \nu], \text{ and } i = 1, \dots, 750 \\ Y_i & \text{else} \end{cases} . \quad (2.5)$$

The underlying distribution for the CWeibull data is *Weibull*($\gamma = \textit{shape} = 0.5, \beta = \textit{scale} = 10$) with too-frequent values $\{s_j\} = \{0, 0.25, 0.50, 0.75, 1.00, 1.50, 2.00, 3.00, 5.00, 10.00\}$ and a contamination neighborhood of $\nu = 0.12$. The underlying distribution for the CNorm data is $N(\mu = 4, \sigma = 1)$ with too-frequent

values $\{s_j\}=\{2.00, 3.00, 3.50, 3.75, 4.00, 4.25, 4.50, 5.00, 6.00\}$ and a contamination neighborhood of $\nu = 0.05$.

The top two panels of Figures 2.4, 2.5, and 2.6 show max-bin histograms of the unobservable and contaminated data. The lognormal and Weibull plots show a zoomed-in region in order to better see the areas where most of the data are located (i.e., not in the tail(s)). The theoretical density is also drawn on as a solid grey line. We see that the contaminated data have the shape of the original distributions with peaks at the too-frequent values. For the lognormal and Weibull data, only the too-frequent values with high density *overall* are discernable. These are 0, 0.25, 0.50, 0.75, 1.00, 1.50, 2.00, and 3.00 for the CLNorm data and 0, 0.25, 0.50, 0.75, 1.00, and 1.50 for the CWeibull data. For the CNorm data, the too-frequent values 2.00, 3.00, 3.50, 3.75, 4.00, 4.25, and 5.00 stand out. The values 1.80 and 5.12 also stand out, but this is the result of random noise.

Table 2.1 provides the sample mean, standard deviation, and median for each data set. The summary statistics are very close for the original and contaminated samples. This supports the use of summary statistics from the contaminated data for parameter estimation. These statistics serve as benchmarks for comparing the transformed statistics, to make sure that the chosen transformation preserves the main characteristics of the distribution.

Table 2.1: Descriptive statistics for simulated lognormal (top), Weibull (middle), and normal (bottom) data. Summaries are for the original data, the contaminated data, jitter-transformed data (with best parameters), and local regeneration-transformed data (with best parameters).

| Lognormal($\mu=0, \sigma=2$) | Sample Mean (Standard Deviation) | Sample Median |
|------------------------------------|----------------------------------|---------------|
| Original | 8.15 (63.86) | 1.04 |
| Contaminated | 8.15 (63.86) | 1.01 |
| Jittered (Uniform, $\delta=0.07$) | 8.15 (63.86) | 1.03 |
| Local Regeneration (k=5) | 8.15 (63.86) | 1.05 |
| Weibull($\gamma=0.5, \beta=10$) | Sample Mean (Standard Deviation) | Sample Median |
| Original | 19.73 (44.60) | 4.47 |
| Contaminated | 19.73 (44.60) | 4.47 |
| Jittered (Uniform, $\delta=0.02$) | 19.73 (44.60) | 4.47 |
| Local Regeneration (k=16) | 19.74 (44.59) | 4.47 |
| Normal($\mu=4, \sigma=1$) | Sample Mean (Standard Deviation) | Sample Median |
| Original | 3.98 (1.02) | 3.99 |
| Contaminated | 3.98 (1.02) | 4.00 |
| Jittered (Uniform, $\delta=0.07$) | 3.98 (1.02) | 3.98 |
| Local Regeneration (k=20) | 3.98 (1.02) | 3.99 |

2.4.2 Transforming the Contaminated Data: Jittering

We start by choosing a range of values for δ that defines the neighborhood for jittering. We choose six values $6\sigma = 0.01, 0.02, 0.05, 0.08, 0.010$ and 0.12 and compare the uniform and normal perturbation distributions for each of these values. Note that we use the *observed* too-frequent values for the transformations to better mimic a realistic implementation.

Max-bin histograms are produced for each δ and perturbation distribution. To be concise, only the best jitter level/distribution (as determined in Section 2.4.4) is shown for each underlying distribution in the third panel of Figures 2.4, 2.5, and 2.6. The complete set of max-bin histograms can be found at www.smith.umd.edu/ceme/statistics/papers. We see that indeed jittering irons out the too-frequent values and the distribution of the transformed data more closely resembles that of the original data. In addition, for all three distributions the jittering transformation hardly affects the overall summary statistics, as can be seen in Table 2.1.

2.4.3 Transforming the Contaminated Data: Local Regeneration

We start by choosing the levels of binning, the regeneration distribution, and the parametric distributions of interest. For the number of bins we choose seven values: $k=4, 5, 8, 10, 16, 20$ and 25 . Each of these values divide 100 easily. For the CLNorm data, we bin the observed too-frequent values corresponding to the percentiles of the $LNorm(\hat{\mu}, \hat{\sigma})$ distribution where $\hat{\mu}$ and $\hat{\sigma}$ are estimated from the

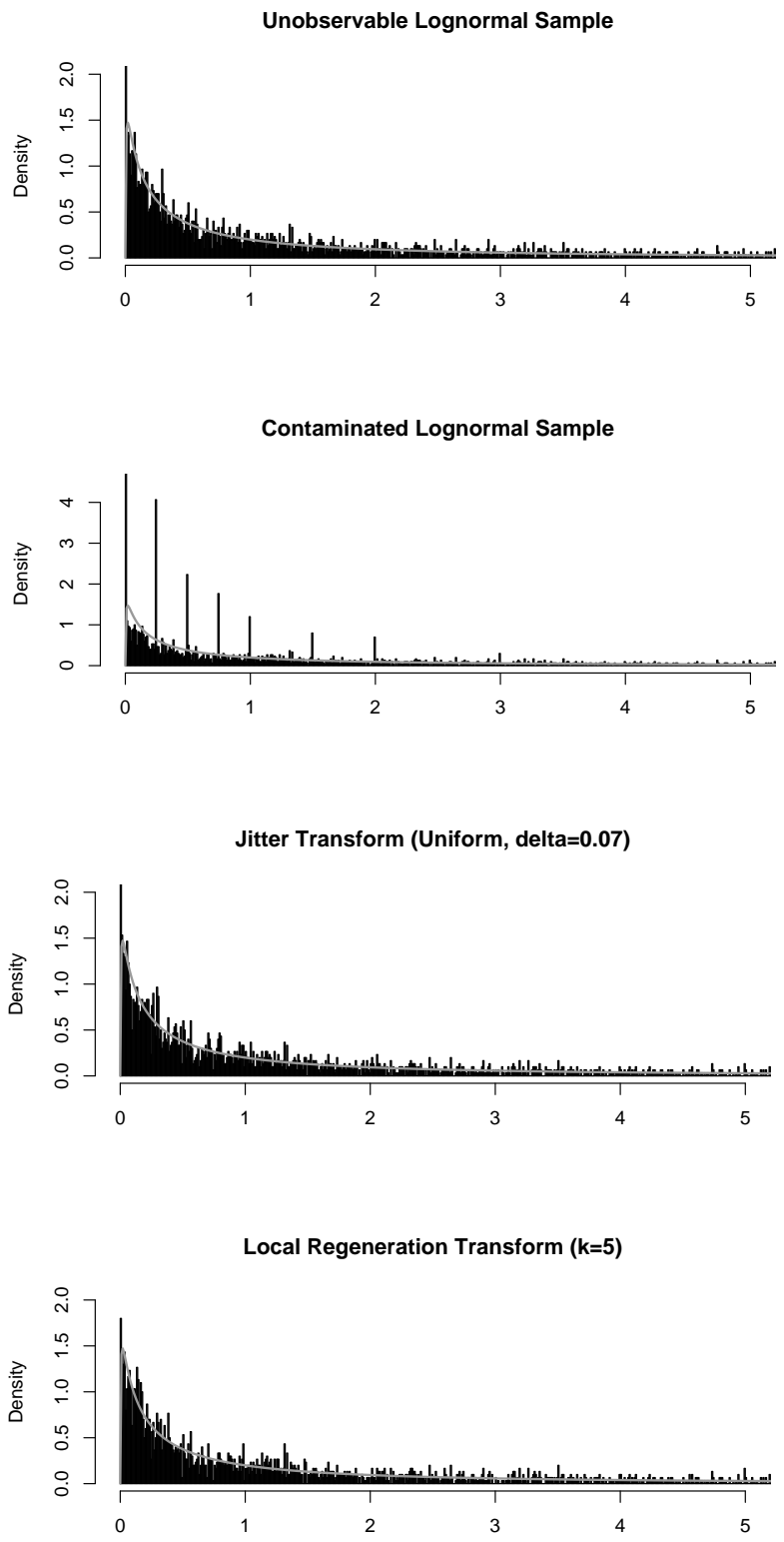


Figure 2.4: Histograms of the original, contaminated, jittered (uniform, $\delta = 0.07$) and locally regenerated ($k=20$) data for the simulated lognormal data. The solid grey line is the lognormal density. Note that the y-axis scale is different for the contaminated data.

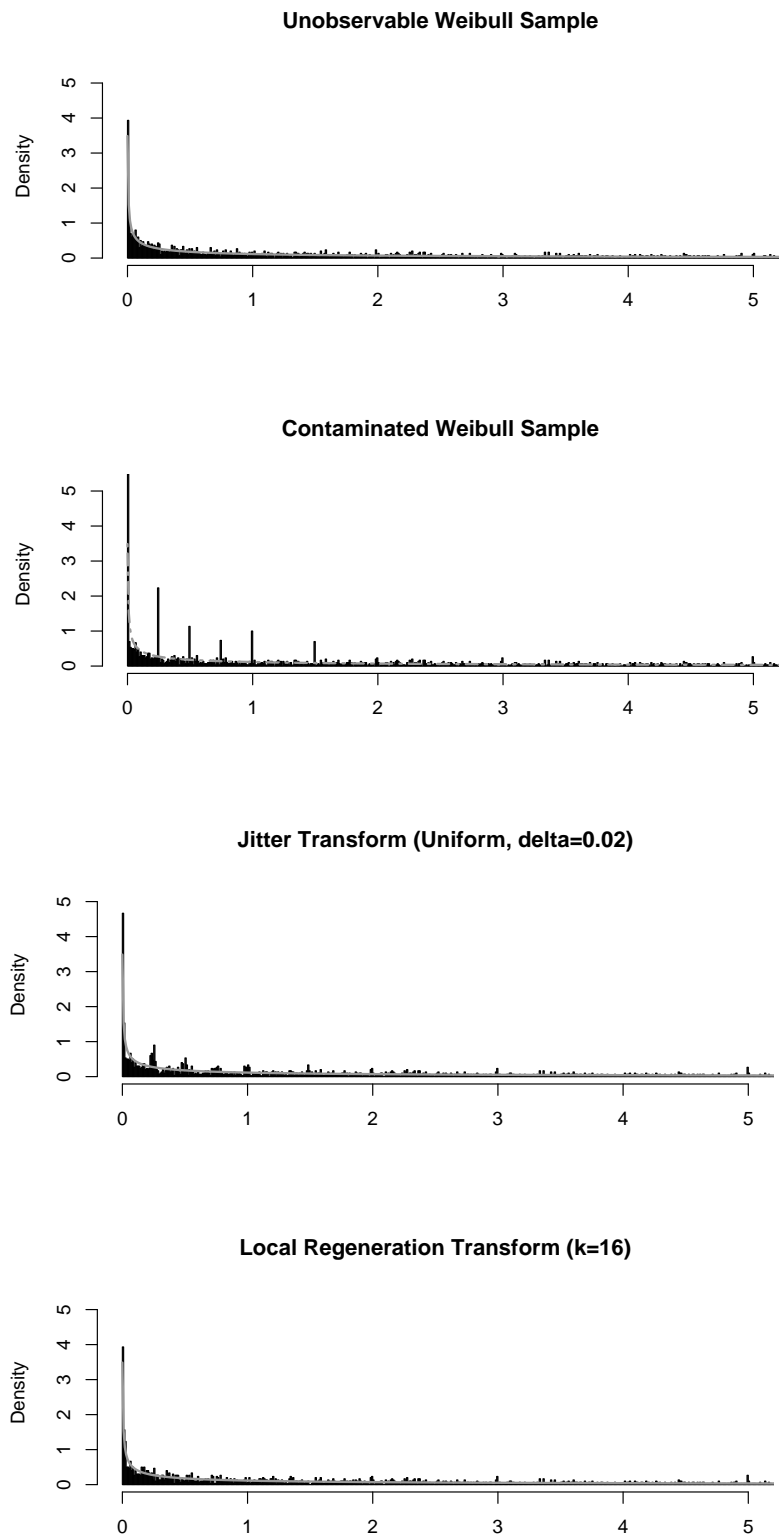


Figure 2.5: Histograms of the original, contaminated, jittered (uniform, $\delta = 0.02$) and locally regenerated ($k=5$) data for the simulated Weibull data. The solid grey line is the Weibull density.

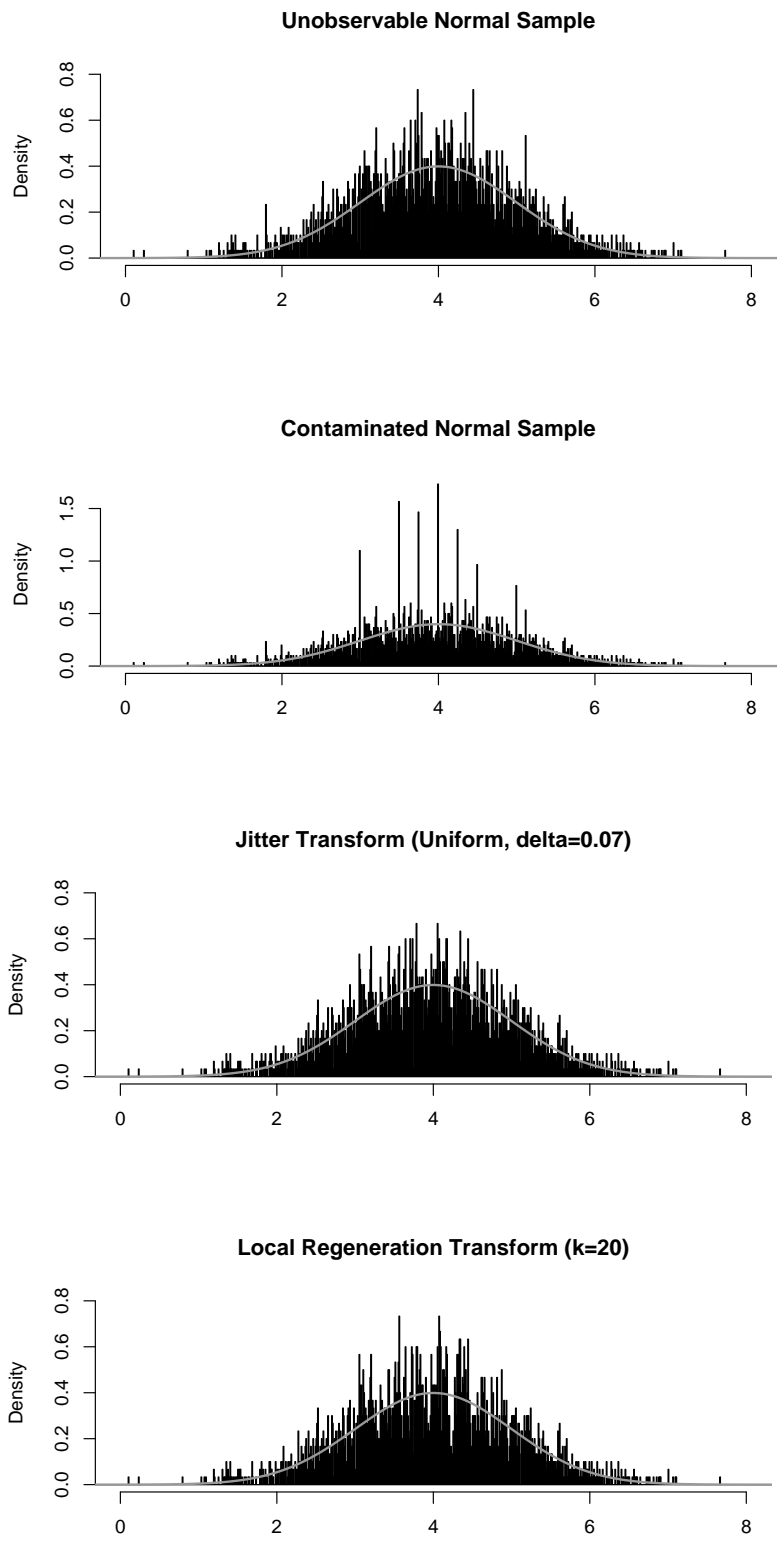


Figure 2.6: Histograms of the original, contaminated, jittered (uniform, $\delta = 0.07$) and locally regenerated ($k=10$) data for the simulated normal data. The solid grey line is the normal density. Note that the y-axis scale is different for the contaminated data.

CLNorm data, and then we “iron-out” the too-frequent values within their corresponding percentile bin. The CWeibull and CNorm data are also “ironed-out” using parameter estimates from the contaminated data.

Max-bin histograms are produced for each k , but only the best k (as determined in Section 2.4.4) is shown for each underlying distribution in the bottom panel of Figures 2.4, 2.5, and 2.6. The complete collection of max-bin histograms can be found at www.smith.umd.edu/ceme/statistics/papers. From the max-bin histograms, we see that the local regeneration transformation yield data that closely resemble the original data. In the Weibull case, it also appears to perform better than the jittering, probably because the too-frequent values are located in low-density areas of the distribution. In such cases, the local regeneration spreads out the too-frequent values over a wider neighborhood (because the percentiles of the distribution are farther away from each other). As with the jittering transform, the summary statistics for the transformed data, for all three distributions, are extremely close to those of the original data, as shown in Table 2.1.

2.4.4 Choosing The Transformation Parameters

While visual inspection of the transformation process (and in particular via max-bin histograms) is valuable in determining the parameters of the transformation (δ and the perturbing distribution in jittering and k in the local regeneration transform), quantitative metrics to assess the data fit can also be helpful, especially for purposes of automation. In order to determine how well the “ironed-out” data

Table 2.2: Goodness-of-fit statistics (and p-values) for the original, contaminated, and jittering transformed lognormal data.

| Lognormal($\mu=0, \sigma=2$) | Anderson-Darling | Cramer von Mises | Kolmogorov | SADBNF |
|--|-------------------|--------------------|---------------------|-------------|
| Original | 1.07(0.25) | 0.147(0.25) | 0.0152(0.25) | 1997 |
| Contaminated | 15.66(0.00) | 0.332(0.11) | 0.0353(0.00) | 2523 |
| Uniform, $\delta = 0.01$ | 7.40(0.00) | 0.300(0.14) | 0.0353(0.00) | 2224 |
| Uniform, $\delta = 0.02$ | 4.15(0.00) | 0.272(0.17) | 0.0316(0.00) | 2264 |
| Uniform, $\delta = 0.05$ | 1.52(0.18) | 0.180(0.25) | 0.0206(0.15) | 2128 |
| Uniform, $\delta = 0.07$ | 1.16(0.25) | 0.154(0.25) | 0.0171(0.25) | 2102 |
| Uniform, $\delta = 0.10$ | 1.12(0.25) | 0.160(0.25) | 0.0185(0.25) | 2131 |
| Uniform, $\delta = 0.12$ | 1.52(0.18) | 0.180(0.25) | 0.0200(0.18) | 2119 |
| Normal, $\delta = 0.01$ | 13.86(0.00) | 0.334(0.12) | 0.0353(0.00) | 2415 |
| Normal, $\delta = 0.02$ | 8.54(0.00) | 0.301(0.14) | 0.0353(0.00) | 2211 |
| Normal, $\delta = 0.05$ | 3.75(0.01) | 0.249(0.20) | 0.0273(0.02) | 2158 |
| Normal, $\delta = 0.07$ | 2.27(0.07) | 0.204(0.25) | 0.0231(0.08) | 2140 |
| Normal, $\delta = 0.10$ | 1.60(0.15) | 0.182(0.25) | 0.0199(0.19) | 2089 |
| Normal, $\delta = 0.12$ | 1.24(0.25) | 0.169(0.25) | 0.0172(0.25) | 2106 |

Table 2.3: Goodness-of-fit statistics (and p-values) for the original, contaminated, and jittering transformed Weibull data.

| Weibull($\gamma=0.5, \beta=10$) | Anderson-Darling | Cramer von Mises | Kolmogorov | SADBNF |
|--|-------------------|--------------------|---------------------|-------------|
| Original | 1.18(0.25) | 0.172(0.25) | 0.0164(0.25) | 3119 |
| Contaminated | 3.92(0.01) | 0.231(0.22) | 0.0344(0.00) | 3442 |
| Uniform, $\delta = 0.01$ | 1.62(0.15) | 0.193(0.25) | 0.0224(0.10) | 3259 |
| Uniform, $\delta = 0.02$ | 1.44(0.20) | 0.187(0.25) | 0.0173(0.25) | 3233 |
| Uniform, $\delta = 0.05$ | 2.21(0.08) | 0.203(0.25) | 0.0280(0.02) | 3149 |
| Uniform, $\delta = 0.07$ | 2.70(0.04) | 0.219(0.24) | 0.0300(0.01) | 3166 |
| Uniform, $\delta = 0.10$ | 3.53(0.02) | 0.258(0.18) | 0.0306(0.01) | 3165 |
| Uniform, $\delta = 0.12$ | 4.29(0.01) | 0.293(0.14) | 0.0328(0.00) | 3125 |
| Normal, $\delta = 0.01$ | 2.81(0.04) | 0.212(0.25) | 0.0274(0.02) | 3357 |
| Normal, $\delta = 0.02$ | 1.68(0.14) | 0.193(0.25) | 0.0220(0.11) | 3209 |
| Normal, $\delta = 0.05$ | 1.57(0.16) | 0.187(0.25) | 0.0230(0.09) | 3324 |
| Normal, $\delta = 0.07$ | 1.67(0.14) | 0.187(0.25) | 0.0246(0.05) | 3183 |
| Normal, $\delta = 0.10$ | 2.08(0.09) | 0.199(0.25) | 0.0280(0.02) | 3149 |
| Normal, $\delta = 0.12$ | 2.35(0.06) | 0.206(0.25) | 0.0290(0.01) | 3180 |

Table 2.4: Goodness-of-fit statistics (and p-values) for the original, contaminated, and jittering transformed normal data.

| Normal($\mu=4, \sigma=1$) | Anderson-Darling | Cramer von Mises | Kolmogorov | SADBNF |
|--|-------------------|--------------------|---------------------|-------------|
| Original | 1.50(0.18) | 0.261(0.18) | 0.0224(0.10) | 1358 |
| Contaminated | 1.52(0.17) | 0.267(0.17) | 0.0224(0.10) | 1710 |
| Uniform, $\delta = 0.01$ | 1.51(0.18) | 0.263(0.18) | 0.0224(0.10) | 1442 |
| Uniform, $\delta = 0.02$ | 1.50(0.18) | 0.261(0.18) | 0.0224(0.10) | 1470 |
| Uniform, $\delta = 0.05$ | 1.51(0.18) | 0.264(0.18) | 0.0224(0.10) | 1462 |
| Uniform, $\delta = 0.07$ | 1.47(0.19) | 0.255(0.19) | 0.0224(0.10) | 1378 |
| Uniform, $\delta = 0.10$ | 1.50(0.18) | 0.261(0.18) | 0.0224(0.10) | 1398 |
| Uniform, $\delta = 0.12$ | 1.59(0.16) | 0.281(0.15) | 0.0224(0.10) | 1422 |
| Normal, $\delta = 0.01$ | 1.52(0.18) | 0.266(0.17) | 0.0224(0.10) | 1592 |
| Normal, $\delta = 0.02$ | 1.51(0.18) | 0.264(0.18) | 0.0224(0.10) | 1408 |
| Normal, $\delta = 0.03$ | 1.50(0.18) | 0.263(0.18) | 0.0224(0.10) | 1410 |
| Normal, $\delta = 0.07$ | 1.50(0.18) | 0.263(0.18) | 0.0224(0.10) | 1406 |
| Normal, $\delta = 0.10$ | 1.49(0.18) | 0.261(0.18) | 0.0224(0.10) | 1428 |
| Normal, $\delta = 0.12$ | 1.51(0.18) | 0.262(0.18) | 0.0224(0.10) | 1420 |

fit an underlying distribution, we use three parametric goodness-of-fit measures: the Anderson-Darling (AD), Cramer von Mises (CvM), and Kolmogorov (K) statistics. These are computed for each combination of distribution and δ for the jittering transform and each k for the local regeneration transform.

We define a “good” parameter as one that captures the underlying (unobserved) distribution, while altering the data as little as possible. In other words, we seek to obtain a reasonable distributional fit while minimizing the differences between the original and transformed too-frequent values.

2.4.4.1 Jittering

Tables 2.2, 2.3, and 2.4 provide the goodness-of-fit test statistics and corresponding p-values for the jittering transformation on the lognormal, Weibull, and normal simulated data sets, respectively. We tried two perturbation distribution (uniform and normal) and six values for δ . Note that when the goal is fitting a parametric distribution we advocate the local regeneration transform over jittering. However, we show the results of the jittering transform for purposes of comparison and also to assess the ability of the jittering transform to approximate the original generating distribution. In order to determine the best jitter level, we plot the tabular results using scree plots for the test statistics, as shown in Figure 2.7. On each scree plot, the lower horizontal line is the test statistic for the original data. The upper horizontal line is the test statistic for the contaminated data. An optimal parameter would have a test statistic close to the lower horizontal line.

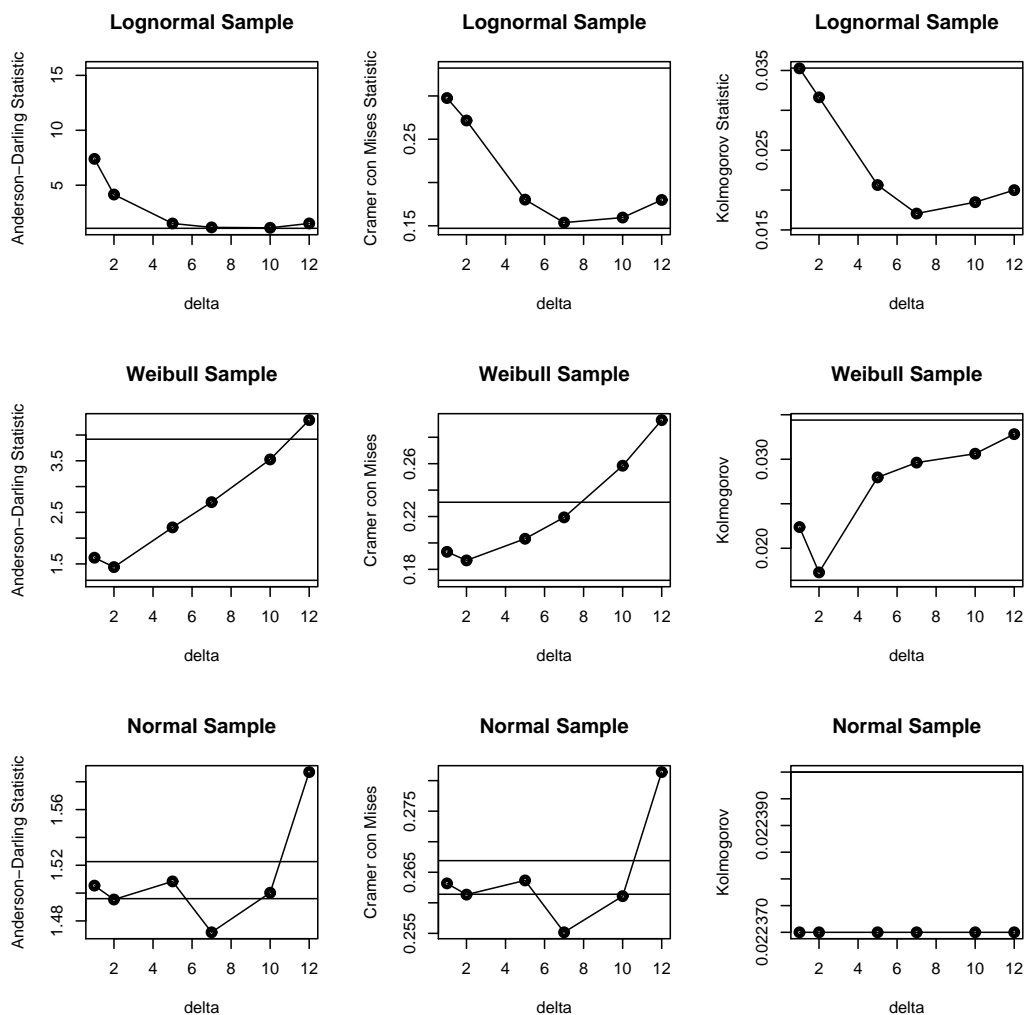


Figure 2.7: Scree plots for the Anderson-Darling (left), Cramer von Mises (middle), and Kolmogorow (right) statistics after applying the uniform jittering transform to the CLNorm (top), CWeibull (center), and CNorm (bottom) data. The upper horizontal line is the test statistic for the contaminated data, and the lower horizontal line is the test statistic for the original data.

For simplicity, we only discuss the results using a uniform perturbation (which provided better results than a normal perturbation). For the CLNorm data, the AD statistic suggests $\delta = 0.10$, the CvM statistic suggests $\delta = 0.07$ (the largest value tested), and the K statistic indicates $\delta = 0.07$. Looking at the p-values in Tables 2.2, 2.3, and 2.4, when $\delta = 0.07$, the p-value is large for all tests, suggesting that these transformed data resembles the underlying distribution. Since we want to manipulate the data as little as possible, we choose $\delta = 0.07$. The CWeibull data has more straightforward results, where the lowest test statistic for each of the three goodness-of-fit statistics is achieved at $\delta = 0.02$. In contrast, results for the CNorm data are very fuzzy, with no clear pattern in the scree plots. Indeed, according to the p-values for these test statistics, even the contaminated data resemble a $N(4, 1)$ distribution quite well! This highlights the robustness of these tests in the normal case, which in our case is a limitation. It is obvious from the max-bin histograms that the contaminated data differ from a reasonable normal sample. We also attempted asymmetric contamination for the CNorm data, but the robustness result remained unchanged.

Finally, to determine the level of jittering using the nonparametric measures, we look at the three proposed methods for evaluating smoothness of the max-bin histogram: kernel smoothers of the max-bin histogram, deviations between pairs of neighboring frequencies (e.g., SADBNF), and looking for local-abnormalities (using the α_i measure). The kernel smoothers are shown in Figure 2.8 for the original data, the contaminated data, and the transformed data. In the lognormal and Weibull cases the kernel smoother for the jittered data shows an improvement over

the contaminated data and closely resembles that of the original data. In contrast, in the normal case we see once again the robustness to deviations from normality, and all kernel smoothers appear identical. We conclude that for nonnormal data, kernel smoothers are a useful tool for the jittering parameter choice. With respect to measures of smoothness, we find these not as helpful. Although the SADBNF measure (shown in the last column of Tables 2.2 2.3, and 2.4), seems to lead to δ values that correspond to those indicated by the parametric goodness-of-fit tests, these drops in SADBNF are more likely an artifact of the particular samples that we generated. Instead, we notice a tendency of this and other local smoothness metrics to always favor larger values of δ . Since our goal is to find the minimal level of smoothing that is able to “iron-out” too-frequent values, one option is to introduce a penalty term into the local smoothness measure. In contrast, using the α_i measure with a three-standard-deviation threshold leads to the opposite conclusion: it favors no jittering at all. The reason for this is that low levels of jittering (with the extreme case being no jittering at all) are associated with a large standard deviation of α and thus no α_i value exceeds this threshold.

2.4.4.2 Local Regeneration

Table 2.5 provides the goodness-of-fit statistics (and p-values) for the local regeneration transformation, and the scree plots are shown in Figure 2.5. In this case we choose seven values for the number of bins k , ranging between 4 and 25. The CLNorm data fits best when $k = 5$ (note that $k = 8$ appears worse than both

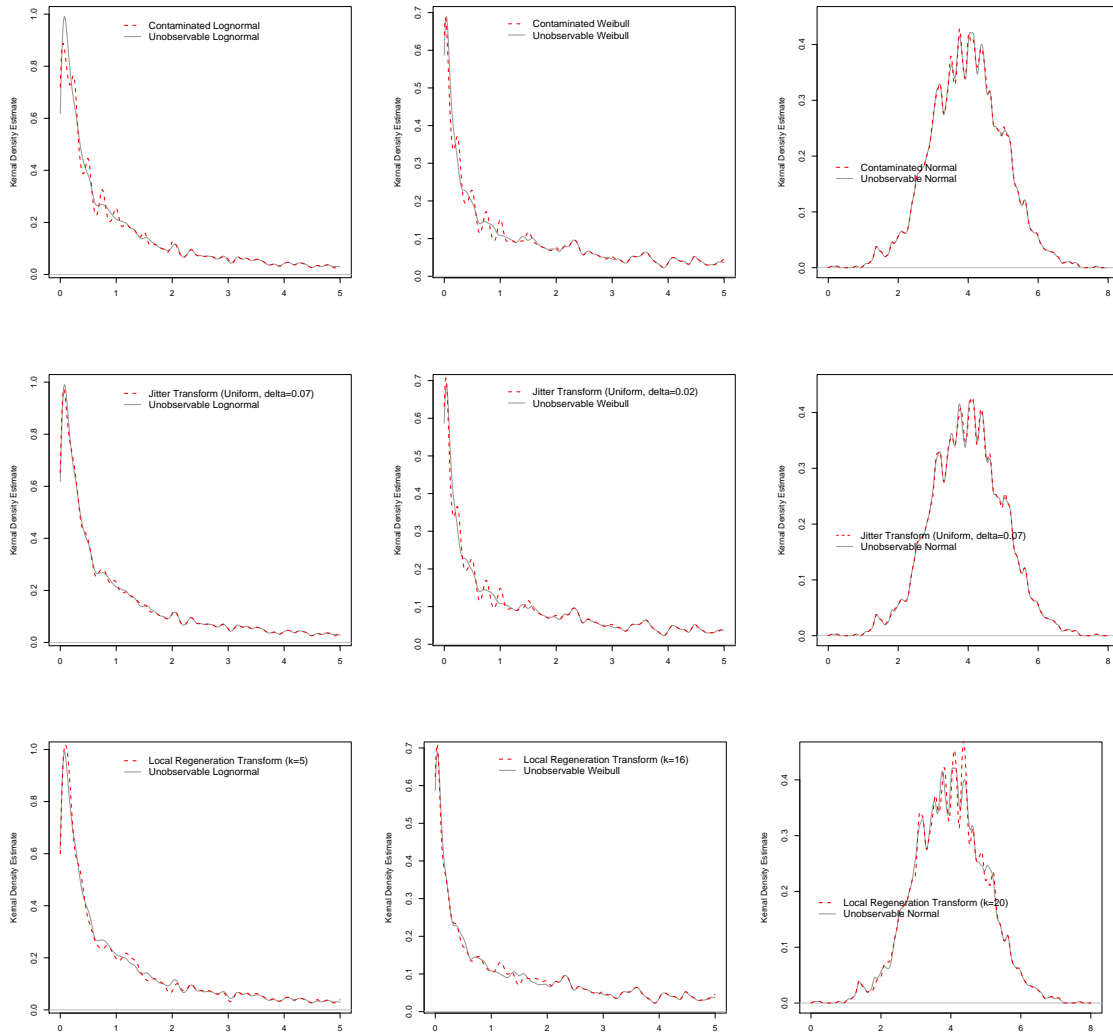


Figure 2.8: Comparing kernel density estimates for the original data (solid line) versus (dashed) the contaminated data (top), jittered data (middle), and locally regenerated data (bottom). The columns correspond to the lognormal (left), Weibull (center), and normal (right) simulated data.

$k = 5$ and $k = 10$, but this is most likely an artifact of these particular data). For the CWeibull data, the best fit is achieved with $k = 16$. Finally, for the CNorm data the picture is again ambiguous, probably due to the strong robustness of the tests to deviation from normality.

The transformed data with these chosen parameters are plotted using max-bin histograms (Figures 2.4-2.6, bottom panels). It can be seen that in each of the three cases, the transformed data no longer exhibit too-frequent values, and they appear very similar to the original “unobservable” data.

2.5 Transforming the Surplus Data

We now apply the local regeneration transform to the auction surplus data, with the goal of obtaining data from a parametric continuous distribution. Since the data are real world, the underlying distribution is unknown. The data best fit a three-parameter Weibull distribution when using the log-scale + 1, as can be seen from the probability plots in Figure 2.2. The data does not fit any familiar distribution without the log transformation.

By inspecting zoomed-in regions of the max-bin histogram (top left panel in Figure 2.1), the too-frequent values appear to be $\{0, 0.01, 0.50, 0.51, 1.00, 1.01, 1.49, 1.50, 1.75, 2.00, 2.01, 2.50, 3.00, 3.50, 4.00, 4.50, 5.00\}$. These same values also appear in a max-bin histogram applied to the log-transformed data (top panel in Figure 2.10). As explained above, we operate on the log-transformed data from here on.

Table 2.5: Goodness-of-fit statistics (and p-values) for the original, contaminated, and local regenerating transformed lognormal (top), Weibull (middle), and normal (bottom) data

| Lognormal($\mu=0, \sigma=2$) | Anderson-Darling | Cramer von Mises | Kolmogorov |
|-----------------------------------|-------------------|--------------------|---------------------|
| Original | 1.08(0.25) | 0.147(0.25) | 0.0152(0.25) |
| Contaminated | 15.66(0.00) | 0.332(0.11) | 0.0353(0.00) |
| k=25 | 4.36(0.01) | 0.286(0.15) | 0.0316(0.00) |
| k=20 | 3.24(0.02) | 0.223(0.23) | 0.0316(0.00) |
| k=16 | 2.83(0.04) | 0.234(0.22) | 0.0280(0.02) |
| k=10 | 1.38(0.21) | 0.117(0.25) | 0.0205(0.16) |
| k=8 | 1.77(0.12) | 0.292(0.14) | 0.0242(0.06) |
| k=5 | 0.90(0.25) | 0.101(0.25) | 0.0169(0.25) |
| k=4 | 2.48(0.05) | 0.375(0.09) | 0.0271(0.02) |
| Weibull($\gamma=0.5, \beta=10$) | Anderson-Darling | Cramer von Mises | Kolmogorov |
| Original | 1.18(0.25) | 0.172(0.25) | 0.0164(0.25) |
| Contaminated | 3.92(0.01) | 0.231(0.22) | 0.0344(0.00) |
| k=25 | 2.03(0.09) | 0.194(0.25) | 0.0224(0.10) |
| k=20 | 1.37(0.22) | 0.175(0.25) | 0.0172(0.25) |
| k=16 | 1.46(0.25) | 0.160(0.25) | 0.0164(0.25) |
| k=10 | 1.55(0.17) | 0.185(0.25) | 0.0216(0.12) |
| k=8 | 1.83(0.12) | 0.180(0.25) | 0.0236(0.07) |
| k=5 | 2.37(0.06) | 0.196(0.25) | 0.0266(0.03) |
| k=4 | 2.33(0.06) | 0.179(0.25) | 0.0246(0.05) |
| Normal($\mu=4, \sigma=1$) | Anderson-Darling | Cramer von Mises | Kolmogorov |
| Original | 1.50(0.18) | 0.261(0.18) | 0.0224(0.10) |
| Contaminated | 1.52(0.17) | 0.267(0.17) | 0.0224(0.10) |
| k=25 | 1.42(0.20) | 0.237(0.21) | 0.0224(0.0985) |
| k=20 | 1.35(0.22) | 0.235(0.22) | 0.0224(0.10) |
| k=16 | 1.45(0.19) | 0.248(0.20) | 0.0224(0.10) |
| k=10 | 1.34(0.23) | 0.215(0.24) | 0.0224(0.10) |
| k=8 | 1.41(0.21) | 0.249(0.20) | 0.0223(0.10) |
| k=5 | 1.60(0.15) | 0.237(0.21) | 0.0254(0.04) |
| k=4 | 1.40(0.21) | 0.212(0.24) | 0.0223(0.10) |

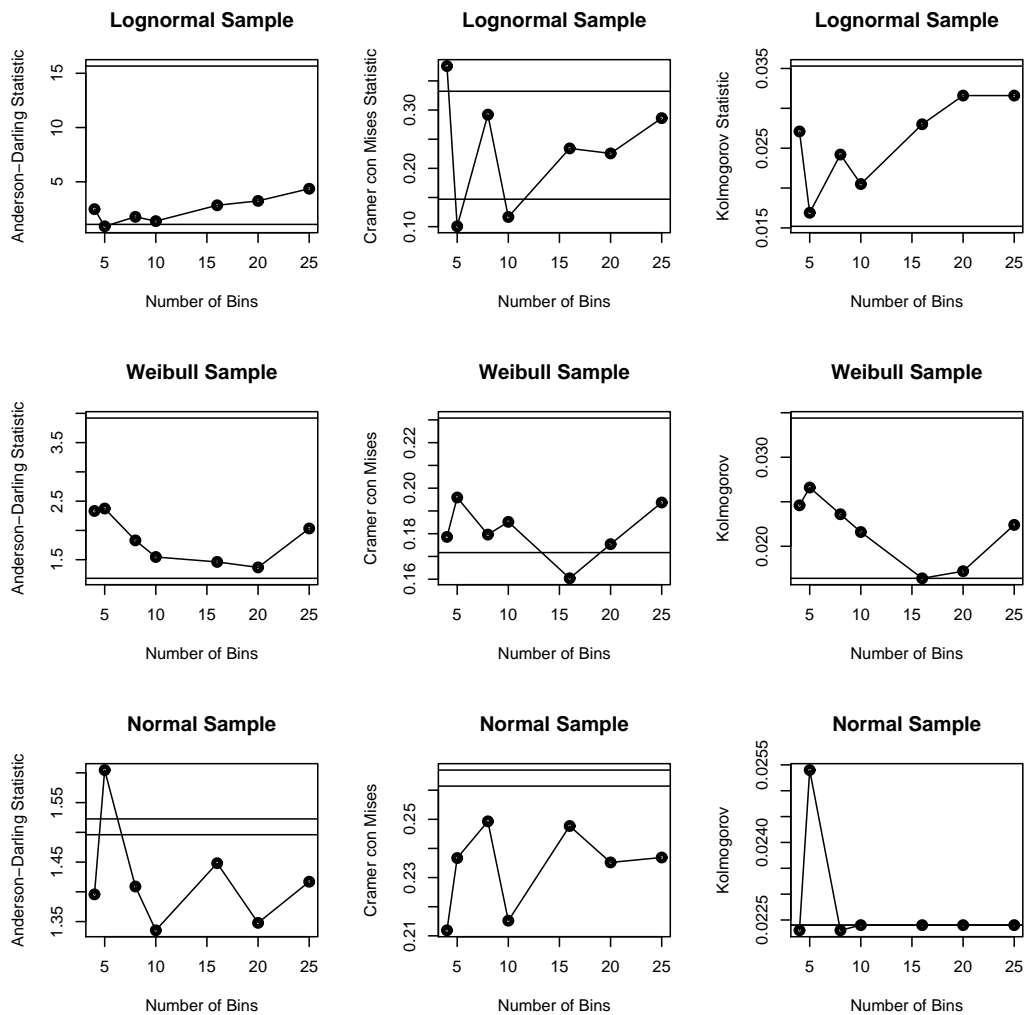


Figure 2.9: Scree plots for the Anderson-Darling (left), Cramer von Mises (middle), and Kolmogorow (right) statistics after applying the local regeneration transform to the CLNorm (top), CWeibull (center), and CNorm (bottom) data. The upper horizontal line is the test statistic for the contaminated data, and the lower horizontal line is the test statistic for the original data.

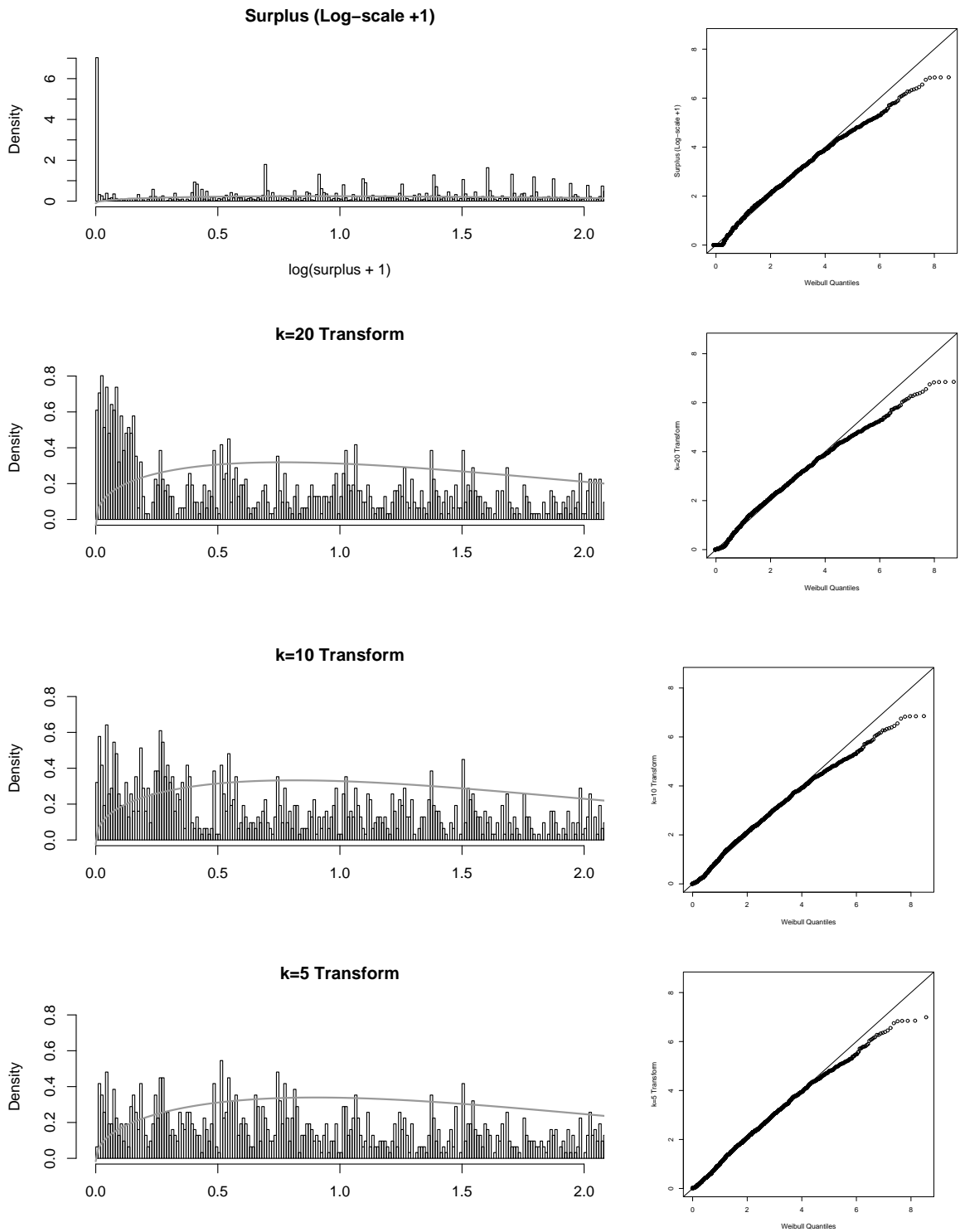


Figure 2.10: Histograms (left) and Weibull probability plots (right) for the original and transformed surplus data. The solid grey line on the histogram is the Weibull density.

The first step is to estimate the three parameters of a Weibull distribution from the observed (contaminated) data. These turn out to be $\gamma = \textit{shape} = 1.403793$, $\beta = \textit{scale} = 2.129899$, and $\tau = \textit{threshold} = -0.10379$. The local regeneration transformation is then applied using k bins from the estimated Weibull distribution, where $k = 4, 5, 8, 10, 16, 20$ and 25 . We show the max-bin histograms of the transformed data for several of these k values in Figure 2.10. The overlaid grey line is the three-parameter Weibull density estimate. Note that the y-axis for the original data is much larger than that for the transformed data. This emphasizes the excess frequency of the zero value. With $k=20$ and even $k=10$ there are still too many zero and near zero values compared to a Weibull distribution. However, $k=5$ appears to “iron-out” the data quite well. This can also be seen in the corresponding probability plots (right column of Figure 2.10), where a small “step” is visible in all plots except for the data transformed with $k=5$. Furthermore, the probability plot with $k = 5$ also appears closest to a straight line, even in the right tail.

To evaluate the deviation between the observed and transformed data, we note that approximately 17% of the original data are considered too-frequent. Comparing the summary statistics before and after the transformation (Table 2.6) shows that they are remarkably similar for all k values tested. This reassures us that the transformation has not altered the main features of the distribution.

Finally, examining at the three goodness-of-fit statistics (Table 2.7) and the corresponding scree-plots (Figure 2.11) indicates that $k=5$ leads to the best fit to an underlying three parameter Weibull distribution. These indications, coupled with visual evidence from the max-bin histograms and probability plots, lead us to

Table 2.6: Summary statistics for the original and transformed surplus data.

| Weibull Fit | Mean (Standard Deviation) | Median |
|-------------|---------------------------|--------|
| surplus | 19.19 (64.77) | 4.49 |
| k=25 | 19.20 (64.77) | 4.48 |
| k=20 | 19.19 (64.77) | 4.40 |
| k=16 | 19.20 (64.77) | 4.45 |
| k=10 | 19.21 (64.77) | 4.49 |
| k=8 | 19.21 (64.77) | 4.52 |
| k=5 | 19.22 (64.76) | 4.49 |
| k=4 | 19.27 (64.75) | 4.54 |

Table 2.7: Goodness-of-fit statistics (and p-values) for surplus original and transformed data (log-scale + 1) fit to a Weibull distribution.

| Weibull Fit | Anderson-Darling | Cramer von Mises | Kolmogorov |
|-------------|-------------------|-------------------|---------------------|
| surplus | 23.47(0.00) | 2.45(0.00) | 0.0567(0.00) |
| k=25 | 22.41(0.00) | 2.72(0.00) | 0.0589(0.00) |
| k=20 | 20.92(0.00) | 2.68(0.00) | 0.0591(0.00) |
| k=16 | 18.30(0.00) | 2.41(0.00) | 0.0536(0.00) |
| k=10 | 12.25(0.00) | 1.82(0.00) | 0.0474(0.00) |
| k=8 | 9.78(0.00) | 1.57(0.00) | 0.0466(0.00) |
| k=5 | 5.29(0.00) | 0.84(0.01) | 0.0352(0.00) |
| k=4 | 6.55(0.00) | 1.09(0.00) | 0.0402(0.00) |

conclude that the (log) surplus data are best approximated by a three parameter Weibull distribution. The transformed data can then be used in further statistical analyses. For example, Figures 2.12 and 2.13 compare the residuals from a linear regression model of surplus on price (both in log form) using the raw data (left panel) vs. the transformed data (right panel). Transforming the raw data removes some of the unwanted pattern from the residuals, thereby making it a more appropriate model for consumer surplus in online auctions.

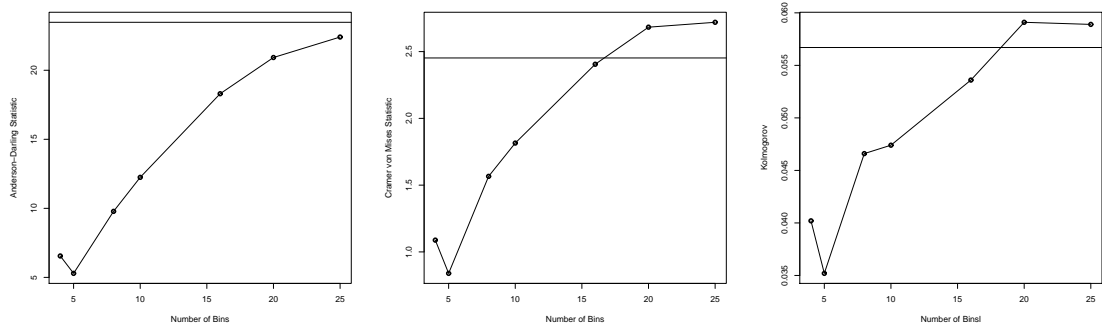


Figure 2.11: Scree plots for Anderson-Darling (left), Cramer von Mises (middle), and Kolmogorow (right) statistics for the transformed surplus data, for an increasing number of bins.

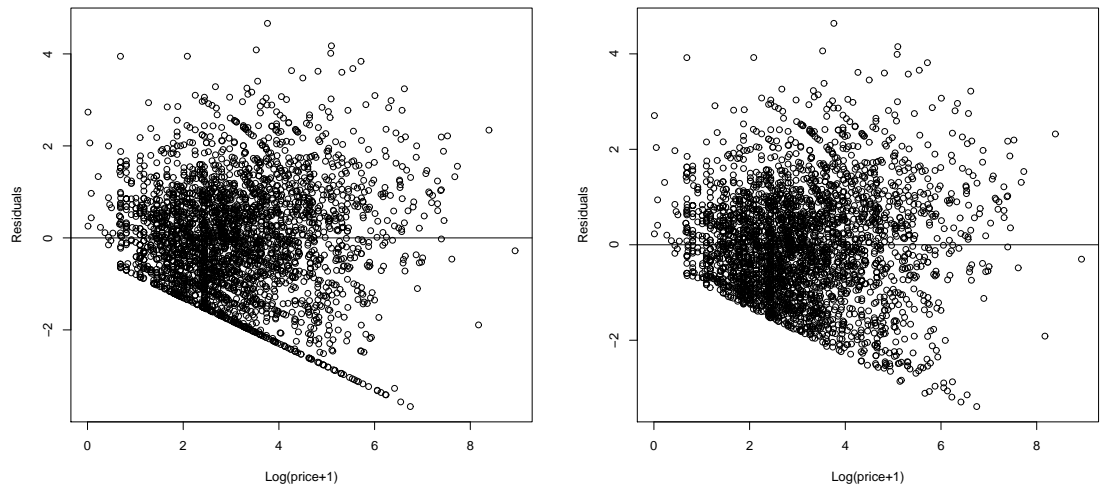


Figure 2.12: Residuals from a linear regression model of $\log(\textit{surplus} + 1)$ on $\log(\textit{price} + 1)$ using the original data (left) vs. the transformed data (right).

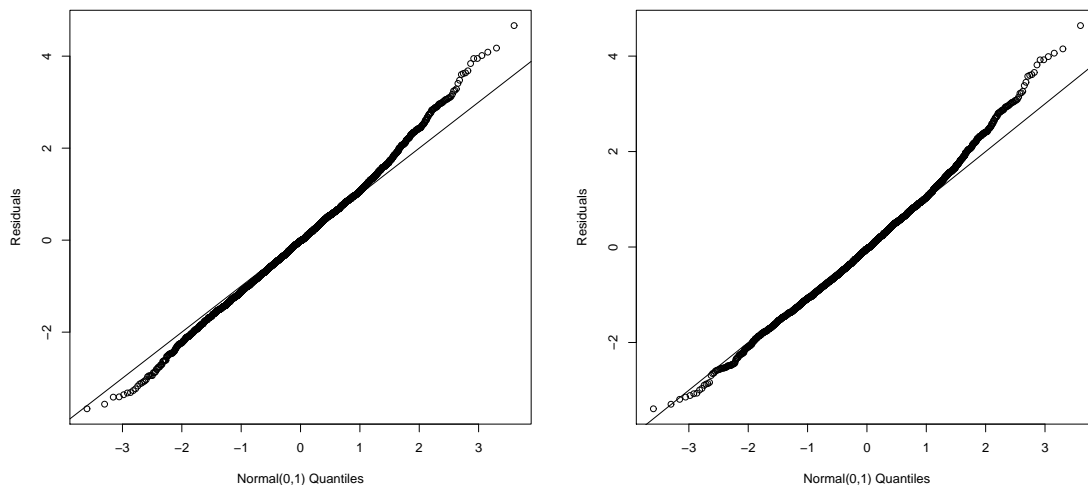


Figure 2.13: Normal probability plots for regression residuals using the original surplus data (left) vs. the transformed data (right).

2.6 Conclusions

We introduce two transformations, jittering and local regeneration, for semi-continuous data that are aimed at yielding data that are continuous. One method is directly aimed at fitting a parametric continuous distribution, while the other is nonparametric and leads to continuous data of an unspecified parametric form. The idea behind both transformations is to replace too-frequent values with values randomly generated within their neighborhood. The difference between the two transformations is with respect to the definition of a neighborhood. While jittering defines a fixed size neighborhood that is anchored around the too-frequent values, local regeneration uses percentiles of the fitted parametric distribution to define neighborhoods. In the latter case, the size of the neighborhood depends on the shape of the fitted distribution, with wider neighborhoods in tail or other low-density areas. The transformed data from the two transformations therefore differ

when too-frequent values are located in low-frequency areas of the data.

The proposed transforms, and in particular the local regeneration transformation, are similar in flavor to the well known and widely used Box-Cox transformation. In both cases the goal is to transform data into a form that can be fed into standard statistical methods. Like the Box-Cox transformation, the process of finding the best transformation level (λ in the Box-Cox case, δ in jittering, and k in local regeneration) is iterative.

To evaluate the performance of the transformations we use simulated data from three common parametric distributions, which are “contaminated”. We find that both transformations perform well, yielding transformed data that are close to the original simulated data. We use both graphical aids and test statistics to evaluate goodness-of-fit. However, for some distributions (e.g., the normal) statistical tests of goodness-of-fit appear to be insensitive to the too-frequent values. The same occurs when comparing transformed data using different parameter values. We therefore advocate the use of max-bin histograms for identifying too-frequent values, for choosing parameters, and for assessing goodness-of-fit.

When the goal is to achieve continuous data without a specific parametric fit, we find that kernel smoothers provide a useful visual aid in determining the level of required jittering, with the exception of the normal distribution. We also note that simple measures of smoothness of the max-bin histogram such as SADBNT or α_i tend to favor no jittering or excessive jittering, and therefore there is room for developing measures that can better assist in determining the right level of jittering.

In conclusion, we believe that further studies with real semi-continuous data

will show the importance and usefulness of these transformations.

Chapter 3

Functional Data Analysis

3.1 What is Functional Data Analysis?

Functional data analysis (FDA) has become popular by the seminal works of Ramsay and Silverman (2005, 2002). At its core, functional data analysis deals with curves, shapes, or objects as the observation, instead of discrete data points as in classical statistics. For example, the temperature at a weather station over a period of one day may be considered a functional observation (curve) since it arises from a continuous process: the temperature at any moment in time over that day. The curves representing daily temperature at several different weather stations would be a set of functional observations.

Functional data analysis generalizes classical statistics to the functional context. For example, data summaries include measures of central tendency, variability, skewness, etc., and traditionally, summary statistics are presented in numerical form. In the functional setting, each summary statistic is itself a functional object, such as the mean function or standard deviation function. FDA has also changed the way that we look at regression. Functional linear models are similar to ordinary linear models except that the explanatory variables are now in functional form. The response may be scalar, multivariate, or another function. See Ramsay and Silverman (2005, Chapters 10 and 11) for details and examples.

One of the main advantages of using functional data is that it often allows for the estimation of derivatives. The first derivative of temperature over one day is also a functional object, and it represents the temperature-velocity: the rate of change in temperature over time. Under normal weather conditions (i.e., no fronts or passing storms), the temperature-velocity is positive approaching 2pm (the warmest part of the day) since the temperature is increasing. After 2pm, the velocity is negative since the temperature is decreasing. Similarly, the temperature-acceleration (second derivative) measures the speed at which the temperature is changing. After the sun sets, the temperature may be decreasing at an increasing rate which corresponds to negative acceleration (or deceleration).

3.2 Representing Auction Price Evolution as a Continuous Curve

It is often difficult or impossible to measure a continuous process. For example, the temperature at a weather station may only be recorded once an hour. Similarly, bids in online auctions are placed at varying time points. The resulting bid histories are therefore time series that are unevenly spaced with sometimes very sparse and other times very dense areas. Instead of considering the discrete set of bids in an auction as a vector, we use them to estimate the complete continuous price evolution that takes place during the auction. While we could simply “connect the dots” to obtain the price of the auction at any given time, this would overfit the data (i.e., model the noise), thereby not providing a good representation of the underlying continuous price process. An alternative is to represent the price as a continuous

smooth curve. This type of curve representation is prevalent in FDA (Ramsay and Silverman, 2005). The first step is therefore to represent/estimate the continuous price function from the discrete bid data.

Let y_j be the recording of an observation at time $t_j, j = 1, \dots, n$ in an auction with n bids. We convert the raw data into a continuous function, $f(t)$, that allows for the evaluation of the price at any point t during the auction. As with any measurement, there is error, so we have

$$y_j = f(t_j) + e(t_j) \tag{3.1}$$

where $e(t_j)$ is considered white noise ($e(t_j) \sim (0, \sigma^2)$). Different nonparametric smoothing methods exist to approximate the price function $f(t)$ and will be discussed in Section 3.3 below. Section 5.3 discusses the smoothing techniques applied to auction bid data thus far. In Chapter 5, we propose a parametric alternative which also produces continuous price curves.

The advantage of the curve representation is that it treats price evolution as a single continuous entity. It captures the complete price evolution in a more compact and easier to visualize way than raw bid data. The price process can then be described by a few coefficients. Furthermore, an appealing feature of smooth curves is that we can gauge their derivatives (the first derivative is the price-velocity and the second derivative is the price-acceleration) in order to learn how price dynamics behave during the auction.

3.3 Representing a Functional Object Nonparametrically

In this section, we discuss various smoothing techniques to estimate a functional object from discrete measurements. Ramsay and Dalzell (1991) represent finite observations in infinite dimensional space since functional data analysis often involves infinite dimensional processes and/or data. For ease of notation and to coincide with our research, we discuss estimating a functional object in a finite dimensional space. Except where noted, the formulas contained in this section are from Ramsay and Silverman (2005); however, the discussion surrounding the formulation is our own.

3.3.1 Basis Series Expansion

Define a linear smoother

$$\hat{f}_t = \sum_{j=1}^n S_j(t)y_j \tag{3.2}$$

that estimates the function $f(t)$. The most popular smoothing procedure is that of representing the function as a linear combination of K known independent basis functions ϕ_k so that

$$f(t) = \sum_{k=1}^K c_k \phi_k(t). \tag{3.3}$$

An example of a basis function would be $\phi_k(t) = t$ which would represent a linear model or $\phi_k(t) = \log(t)$ which is simply a nonlinear transformation of the inputs (Hastie et al., 2001). For n observations, let $\Phi = \{\phi_k(t_j)\}$ be a full-rank $n \times K$ matrix of basis functions at the observation points. An exact representation may be

possible when $n = K$ since we could choose the coefficients c_k such that $x(t_j) = y_j$ for each j . The amount of smoothing that takes place depends on the number K of basis functions. Ideally, we would like a small value of K for computational efficiency; however, we want to choose K large enough so that the main features of the data are captured but small enough to avoid fitting the noise. The best way to determine the weights $S_j(t)$ is to determine the coefficients c_k by minimizing the least squares criterion

$$SMSSE(y|c) = \sum_{j=1}^n [y_j - \sum_{k=1}^K c_k \phi_k(t_j)]^2. \quad (3.4)$$

In matrix terms, this is

$$\mathbf{SMSSE}(\mathbf{y}|\mathbf{c}) = (\mathbf{y} - \mathbf{\Phi}\mathbf{c})^T(\mathbf{y} - \mathbf{\Phi}\mathbf{c}) = \|\mathbf{y} - \mathbf{\Phi}\mathbf{c}\|^2. \quad (3.5)$$

Therefore, the linear smoother to multiply observation points, \mathbf{y}_j (in vector form), by is the weight:

$$\mathbf{S} = \mathbf{\Phi}(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T. \quad (3.6)$$

There are many different choices of basis functions. Examples are Fourier series, polynomials, regression splines, and wavelets. Fourier series are used when the underlying function is cyclical. Wavelet bases deal with discontinuities and rapid changes in behavior, which is a direct contradiction to Fourier series. Polynomials fit well in the center of the data but are often wiggly and inaccurate at the tails.

The most popular bases are regression splines which are functions constructed by joining polynomials together smoothly at special values called *knots*. The researcher determines the number and placement of the knots. More knots are chosen

in areas where the curve should fit the data most accurately. For online auctions, this occurs at the end of the auction when there is the most bidding activity. Let $K_1 + 1$ be the number of knots. Then there are 2 exterior knots and $K_1 - 1$ interior knots. In between any two adjacent knots, a polynomial spline of a fixed degree, K_2 , is fit in such a manner that the value of consecutive polynomials are the same at each interior knot. A polynomial of degree 0 is a step function discontinuous at the knots; degree 1 is a polynomial or piecewise linear function; degree 2 is a piecewise quadratic with continuous first derivative; degree 3 is continuous at the first two derivatives of the cubic spline meaning that the function is virtually smooth, etc. The number of basis functions will depend on the number of internal knots, the number of parameters per region, and the number of constraints per knot. Numerous examples and illustrations of piecewise polynomials are provided in Hastie et al. (2001).

A basis function must be carefully chosen to have features matching those of the function to estimate. Another criterion for choosing bases is that a certain number of derivatives (intrinsic to the particular data problem) behave well. For example, if we want to represent the price of an auction by a function, its first point should be the opening bid, the last point should be the closing bid, and the function should be monotonically nondecreasing since the price of an auction never decreases. We require that two derivatives, representing the price-velocity and price-acceleration, be accurate. In the online auction context, Jank et al. (2007) use polynomial smoothing splines (discussed in Section 5.3.1), and Hyde et al. (2006) use monotone smoothing splines (discussed in Section 5.3.2).

3.3.2 Kernel Smoothers

Another option for smoothing is the use of a kernel smoother. The idea behind kernel smoothing is that the value of a function at any point t must be most influenced by the observations near t . The uniform kernel is

$$Kern(u) = 0.5, |u| \leq 1,.$$

The quadratic kernel is

$$Kern(u) = 0.75(1 - u^2), |u| \leq 1,.$$

The Gaussian kernel is

$$Kern(u) = (2\pi)^{-1/2} \exp(-u^2/2).$$

The weights

$$w_j(t) = Kern[(t_j - t)/h]$$

are concentrated for t_j near t , and the degree of concentration is controlled by h , the bandwidth parameter. Small values of h mean that only observations close to t receive substantial weight, and large values of h use a wider average. The kernel estimator makes use of local weights $S_j(t)$ such that

$$S_j(t) = \frac{Kern[(t_j - t)/h]}{\sum_r Kern[(t_r - t)/h]}$$

for the Nadaraya-Watson kernel estimator.

3.3.3 Roughness Penalty

The above smoothers tend to fit the data too closely (and thus model the noise); therefore, a *roughness penalty* approach is commonly employed. This method

takes into account the trade-off between a curve that fits the data well and smoothness (so not overfitting to the noise). For example, we would like to find a function $x(t_j)$ that minimizes the residual sum of squares $\sum_j [y_j - x(t_j)]^2$ to provide a good fit to the data. However, fitting the data too closely could yield a function that is too oscillatory, or locally variable, that is also modeling the noise in the data. A popular measure of wiggleness (or roughness) of a function is its integrated squared m^{th} derivative

$$PEN_m(x) = \int [D^m x(t)]^2 dt. \quad (3.7)$$

This measures the degree of departure from a straight line. A highly variable function will yield a high value of $PEN_m(x)$. The researcher should choose the highest order of the derivative that she is interested in investigating and then add 2 to ensure that the highest derivative is also a smooth function. For online auctions, we are interested in up to the second derivative, the price-acceleration, so $m = 4$. Define the *penalized* residual sum of squares to be

$$PENSSSE_\lambda = \sum_j [y_j - x(t_j)]^2 + \lambda PEN_m(x), \quad (3.8)$$

where λ is the *smoothing parameter*. When $\lambda = 0$, the penalized squared error drops out, and the function fits the data closely. Indeed, if there were as many knots as there were data points for a linear (degree 1) polynomial, we would have linear interpolation, which would provide too good of a fit to the data. Larger values of λ penalize the function for being curvy. As $\lambda \rightarrow \infty$, the fitted curves approach the standard linear regression line. Ramsay and Dalzell (1991) and Ramsay (1998) suggest the smoothing parameter λ can often be chosen by inspection of the smooths

or through an automated procedure such as generalized cross-validation (GCV).

3.4 FDA Methodology

There has been a lot of efforts in developing models and methods in the functional domain, many of which generalize traditional statistical approaches. We review some of these efforts next; however, Ramsay and Silverman (2005) provide a more extensive list.

Functional summary statistics can be created based on functional observations rather than discrete data points. Clustering discrete points in n -dimensions is performed using a proximity metric, such as K-means. In the functional context, functions can be grouped by the parameters that describe each function. Jank and Shmueli (2007) use functional clustering to examine different groups of price processes in online auctions. Functional linear models is similar to ordinary linear models; however, the explanatory variables are now functional, and the response may be scalar, multivariate, or another function.

Smooth curves also allow for the estimation of derivatives. Understanding a function's dynamics is useful because the dynamics are what drive the process. Differential equations extend quite nicely in the functional context. Ramsay and Silverman (2002) model handwriting and juggling with a differential equation model. Wang et al. (2007b) show that a second order differential equation approximates the price dynamics that take place during an eBay auction well.

FDA also has many applicable uses in the efforts of visualization and ex-

ploratory data analysis (EDA). Unfortunately, visualization has mostly been ignored in the functional literature, which mainly focuses on the derivation of mathematical models. However, there are some noteworthy exceptions. Functional summary statistics such as the mean or median function can be displayed along with the functional observations. Phase plane plots are useful in understanding the interplay of dynamics. Graphing functional principal components, which is the most widely used of the visualization techniques, enables the variability within a function to be studied. Functional PCA has been applied to many different problems: Canadian weather data (Ramsay and Dalzell, 1991; Ramsay and Silverman, 2005), criminal activity (Ramsay and Silverman, 2002), adolescent growth (Ramsay and Silverman, 2002), lip motion (Ramsay et al., 1996), and online auction price (Hyde et al., 2004; Hyde, 2006), to name a few.

Recently, Jank et al. (2007) explored functional data in the online auction context exclusively via visualization. While some standard functional visualization tools are implemented, a key contribution is the cutting-edge interactive time series visualization tool *Timesearcher*. It allows users to see long time series, multivariate time series, zoom-in to various rectangular blocks of time, and search for a selected pattern.

Data visualization, especially in the functional context, provides a deeper understanding of data than summary measures alone. The field of data visualization should continue to expand with increased computer capabilities. We expand the literature in Chapter 4 by presenting a novel graphical tool which displays functional data over calendar time.

Chapter 4

Investigating Concurrency in Online Auctions Through Visualization

4.1 Introduction and Motivation

Concurrency of events is omnipresent in today's commerce, both within the electronic commerce world and even more so if we consider the online and offline domains together. For example, a consumer purchasing a luxury item such as a digital camera can gather information about price from a variety of different sources: by reading a store's sales circular, by checking the price at online retailers, by using comparison shopping websites such as bizrate.com that search across many online retailers, by reading newspaper/online sales postings for used cameras by individuals (e.g., craigslist.com), by checking online auction sites such as eBay.com, and more. The availability of items from multichannel sources gives consumers the power to compare prices and other related information (such as shipping time and cost, trustworthiness of the seller, travel time, and physical inspection). This also means that sources compete with one another and therefore, more than likely, also influence each other. There has been a growing literature investigating how these different channels affect each other, their sales, their prices, etc. (e.g., Etzion et al. (2004); Gallien (2002); Vakrat and Seidmann (1999)).

In this research we focus on concurrency within a single channel, namely, online auctions on eBay.com. Online auctions are different from traditional brick-

and-mortar auctions in that they occur simultaneously or within close temporal proximity. Online auctions tend to be longer, ranging over several days rather than minutes. Further, online auctions are not limited by the geographic barriers of traditional auctions. Sellers and bidders can be located in different parts of the country, or even the world, and still conduct business. The wide and growing popularity of online auctions makes the problem of concurrency an interesting and important topic for research.

There are typically numerous auctions for the same or similar products occurring simultaneously. Figure 4.1 shows a snapshot of an eBay webpage that results from a search on “Palm Pilot M515.” It displays all open auctions closing within the next two days that meet the search criteria. Notice that the soonest closing auction (in 6 hours and 55 minutes) has only received 3 bids, and the highest bid is \$41. This is surprising since two other auctions closing the following day already received more bids and also a higher current price. One reason for this could be the lack of a picture on the display page which attracts more attention. Another reason could be the less detailed product description.

A great majority of the literature that analyzes online auction data assumes *independence* across auctions. This assumption is typically made for reasons of simplicity and convenience, while in reality auctions for the same item, competing items, or even related (substitute) items will influence each other especially if they take place within a close time frame. On eBay.com, an identical product is often sold in numerous simultaneous auctions. While each auction contains a replicate of the product, the resulting sales, prices, and even the number and level of bids during the

eBay.com Palm Pilot M515 Auctions Closing in the Next Two Days

10 items found for

Palm Pilot M515

List View [Picture Gallery](#) Sort by: Time: ending soonest [Customize Display](#)

| <input type="checkbox"/> |  Compare | Item Title | PayPal | Price | Bids | Time Left |
|--------------------------|---|---|---|----------|------|------------|
| <input type="checkbox"/> |  | LEATHER BELT CLIP PALM PILOT V Vx ~ M515 M505 PDA CASE |  | \$7.95 | - | 4h 30m |
| <input type="checkbox"/> |  | Palm Pilot M515 |  | \$41.00 | 3 | 6h 55m |
| <input type="checkbox"/> |  | LEATHER BELT CLIP PALM PILOT M500 M505 M515 PDA CASE!!! |  | \$7.95 | - | 20h 30m |
| <input type="checkbox"/> |  | Palm Pilot M515 PDA with Hard Case Mint Condition |  | \$100.00 | 14 | 1d 02h 25m |
| <input type="checkbox"/> |  | Palm M515 Color Handheld PDA Palm Pilot NR |  | \$51.00 | 8 | 1d 18h 52m |

Figure 4.1: List of Palm Pilot M515 auctions closing within the next two days. Note that some of the auctions are for accessories, not the actual Palm M515 product.

ongoing auctions are clearly not independent of each other. Empirical research has mostly been based on auctions for the same item (e.g., mint condition Indian-head pennies (Lucking-Reiley et al., 2005), coins (Bajari and Hortascu, 2003), personal digital assistants (Jank and Shmueli, 2005; Shmueli and Jank, 2005; Ghani and Simmons, 2004), rare coin auctions (Kauffman and Wood, 2005)), that have closed during a certain time period (typically within a few of months). In such situations it is likely that there is dependence between the auctions since buyers have the option to select which of the competing items to bid on, and sellers have the option to decide when to post their item for sale using information on similar previously sold items.

Auction theory is mainly concerned with a single auction, and there has not

been much theoretical research on concurrent auctions, even in the offline context. One paper that does consider simultaneous auctions is that by Guerre et al. (2000), who examine the underlying distribution of a bidder's private value in sealed bid first-price offline auctions.

Prior empirical research concerned with the interplay of auctions has focused mainly on *sequential* classical (offline) auctions. That is, auctions that occur one after the other, but not simultaneously. Sequential auction research is concerned with price and quality of selection as one auction succeeds another auction.

Allen and Swisher (2000) find that auctions occurring later in an auction sequence tend to fetch a higher price than those in the beginning probably due to the limited supply at the end of a sequential auction. Deltas (1999) examines sequential ordering across numerous auctions for cattle, where it is common that higher value lots are sold at the beginning of an auction. Deltas (1999) hence finds that prices decline throughout an auction but also finds a different rate of decline for auctions of different size: prices decline faster in auctions where only a small number of lots are sold.

While the insight gleaned from these analyses is useful in the traditional auction setting, it is little applicable to the online context. Online auctions present new and challenging questions because of the prevalence of simultaneous events. Even within this particular universe, concurrency is prevalent, and it is challenging to evaluate its magnitude and effects. There have been only few attempts at quantifying concurrency in online auctions. Snir (2007) shows that the expected selling price in S sequential auctions is equivalent to the S th order statistic. While that

work is primarily concerned with the final price, we are also interested in the effect of concurrency on the *process* of bidding. Zeithammer (2006) examines sequential auctions on eBay, and finds that bidders deflate their bids based on the future expected surplus. In particular, when bidders expect more auctions in the future, bids decrease. While the work does not examine price trends over time, it examines how information about the near future affects the bidding equilibrium. Specifically, it investigates whether it is better to bid high today and lose potential surplus tomorrow or bid cautiously today in hopes of surplus tomorrow. Finally, the recent paper by Anwar et al. (2006) finds that bidders on eBay typically place bids in multiple competitive auctions.

We are interested in the effect of concurrency not only on the final price of an auction but also on the relationship between the current bid levels and high bids in simultaneous ongoing auctions. Moreover, very few visualization methods have been developed for the special data challenges arising in online auctions. Exceptions are the work of Jank et al. (2007); Shmueli and Jank (2005); Aris et al. (2005) Although we concentrate on online auctions, our proposed methods can be generalized to other applications where one is interested in studying the effect of concurrency on the final result of an event and also on the *event-evolution* itself. Our work adds to this line of research by specifically focusing on the aspect of concurrency and its impact in online auctions.

The goals of this research are twofold: to study concurrent events via statistical exploratory methods (mainly visualization), and second, to develop new research questions based on these visualizations. In Section 4.2 we propose several new vi-

sualizations and adaptations of classic graphical displays for capturing concurrency. We conclude in Section 4.3.

4.2 Visualizing Concurrent Auctions

We use the Palm Pilot M515 data set (described in detail in Appendix A.2) in this section to investigate concurrency since all auctions are for the same exact good, and they occur in close temporal proximity.

4.2.1 Rug Plots

For studying concurrency of prices in competing auctions, we sometimes consider as the observation the complete price evolution that takes place during each auction (as described in Chapter 3). In this research, we use monotone smoothing splines (described in detail in Section 5.3.2) for their flexibility and suitability in the auction context: since the bidding process is necessarily nondecreasing, the monotone smoothing splines guarantee monotone nondecreasing price curves. Our choice of knots is governed by bidding frequencies, such that more knots are present during the “active period” at the very end of the auction and much fewer during the “quiet period” in the middle of the auction (for further details on the choice of knots and smoothing parameter see Jank and Shmueli (2005)). Figure 4.2 shows the results of monotone smoothing for three sample auctions. Even though the bid patterns in each of the auctions is very different, the same family of monotone splines (using

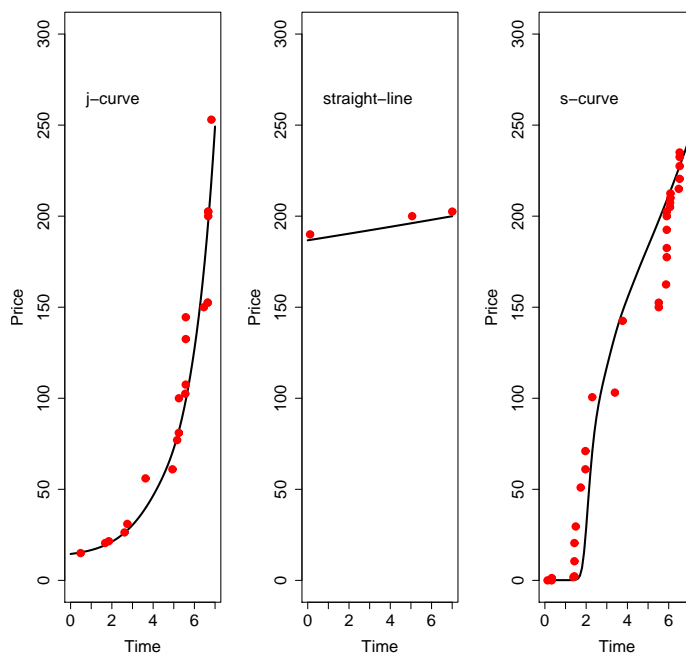


Figure 4.2: Bids and fitted price curve for three different auctions. Knots are placed at day (0, 1, 2, 3, 4, 5, 6, 6.25, 6.50, 6.75, 6.8125, 6.8750, 6.9375, 7). The smoothing parameter is $\lambda=0.01$, and the polynomial order is 5.

the same knots, smoothing parameter, and polynomial order) leads to reasonable approximating curves in each case and avoids adding an extra source of variability across curves.

Our first goal is to visualize a set of auctions by describing their entire price-evolution for the purpose of exploring similarities and patterns in price evolutions of concurrent (or partially concurrent) auctions. One step in this direction is the *Calendar of Auctions* plot (Shmueli and Jank, 2005) for displaying a sample of auctions over the data collection period. Although it preserves chronological information, it does not preserve the bid history information. We propose a graphical method that expands upon the Calendar of Auctions, which displays the entire price-evolution curves or dynamics curves over calendar time. Because of its rug like appearance we

name these displays *rug plots*. A rug plot is a functional data analysis visualization tool which displays numerous overlapping continuous processes over calendar time. Figure 4.3 displays the price curves and Figure 4.4 displays the price-velocity curves. Since the end of an auction is of special interest, we emphasize the endpoint of each curve by a darkened dot. Viewing rug plots at multiple resolutions allows seeing an overview but also local details. For instance, the bottom panels of Figures 4.3 and 4.4 display a subset of all curves, zoomed-in between March 29, 2004 and April 8, 2004.

The price-velocity curve, which is the first derivative of the price-evolution curve, is always non-negative (because the price curve is monotone). A zero velocity occurs when the price does not increase, representing a period of bidding inactivity. A small positive velocity means that the price slowly increases at that instant. Similarly, high velocity corresponds to rapid price increases.

The next step is to see what types of concurrency can be gleaned from rug plots. The rug plot display maintains the temporal information of each auction (or event, in general) in order to compare global trends to more local ones. Concurrency can affect the shape of a price curve or the shape of the dynamics curve if we assume that different curve shapes capture different bidding patterns (e.g., gradual vs. bursty bidding or early vs. late bidding). Taking an exploratory approach, we examine the price and velocity curves in the rug plot for certain prominent types and look for patterns of temporal proximity between types of curves.

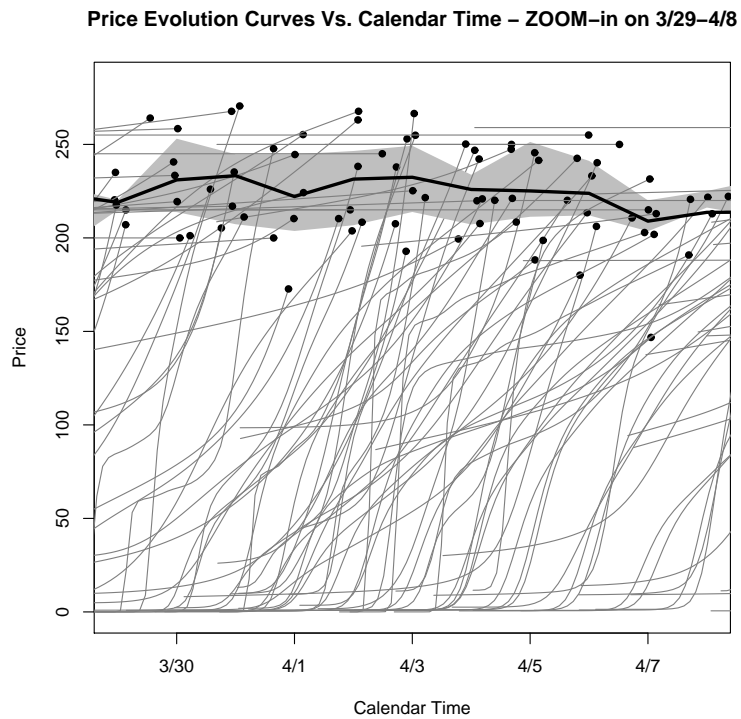
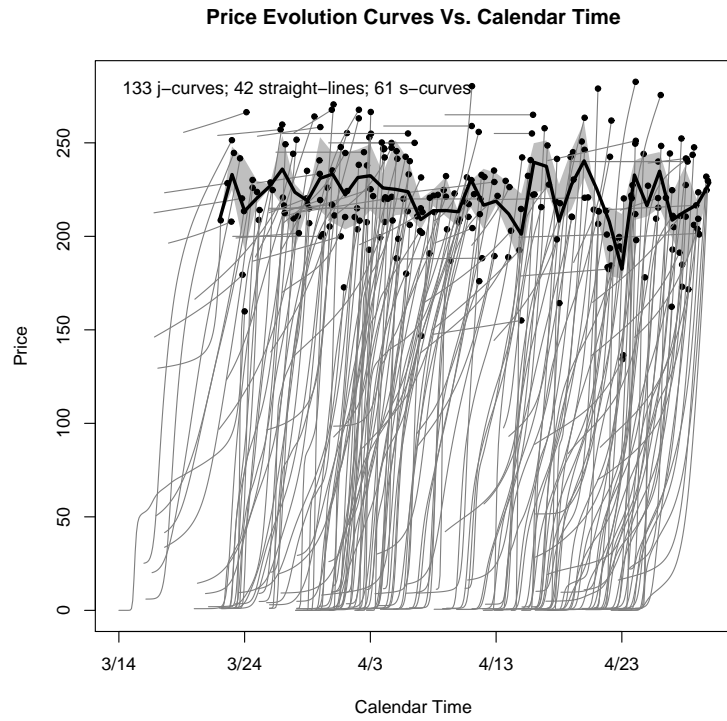


Figure 4.3: Price evolution over calendar time (top) and zoomed-in to March 29 - April 8 region (bottom). Thick curve and grey band are the daily median and IQR closing price.

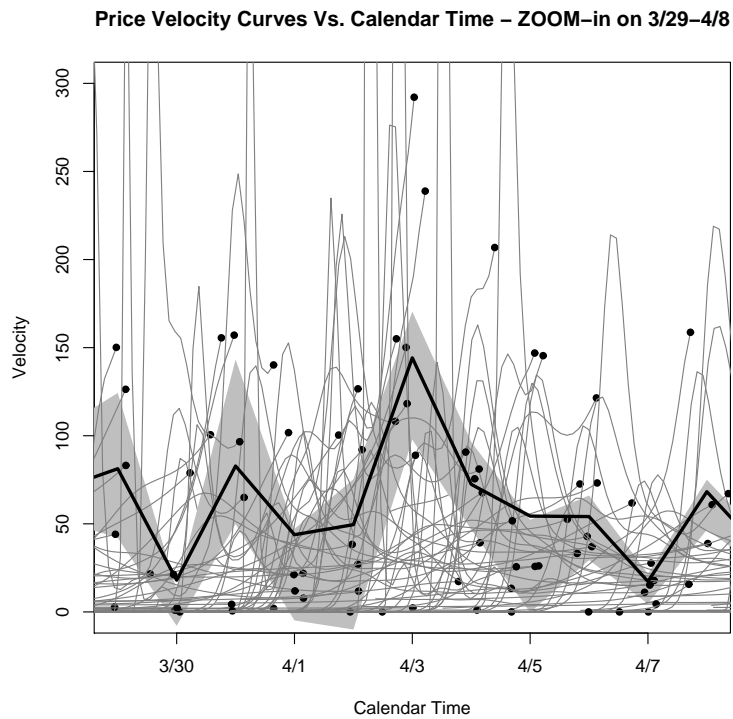
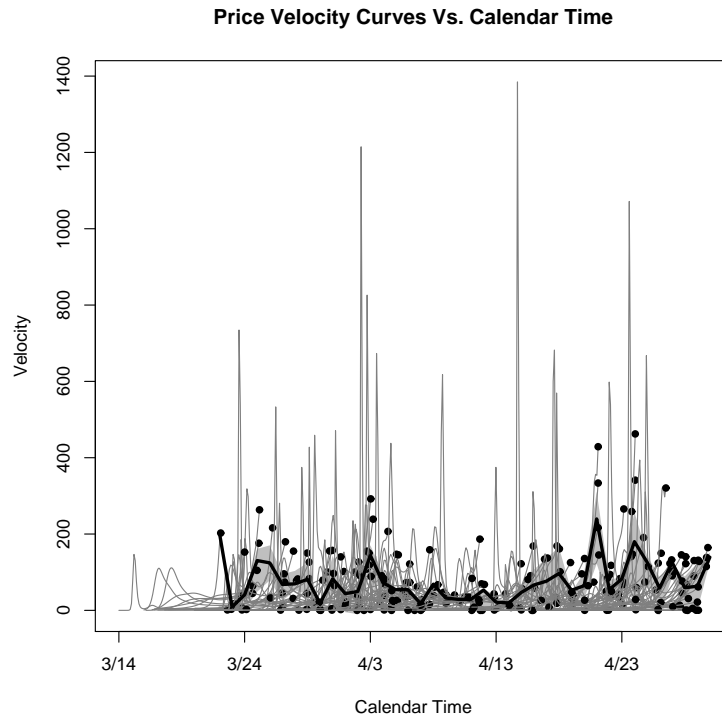


Figure 4.4: Price velocity over calendar time (top) and zoomed-in to March 29 - April 8 region (bottom). Thick curve and grey band are the daily median and IQR velocity.

4.2.1.1 Curve Shapes

The empirical online auction literature describes several bidding patterns. One of these is late bidding (also called “sniping”), which results in very high bidding frequency at the end of an auction. We therefore expect auctions with sniping to have price-velocity curves that spike towards the auction end. Bidding frequency should also manifest itself in the resulting curve: gradual bidding would lead to more gradual curves (and slow dynamics), whereas jump bidding would lead to bursty velocity curves. Relating different types of curves with bidding behavior can be useful in interpreting the rug plot results.

Examining the price curves in Figure 4.3 reveals three main shapes of price curves corresponding to the three shapes seen in Figure 4.2. The most frequent (133 auctions) is a concave-up “j”-shaped price curve which represents auctions with gradual price increases until midauction and a price jump towards the auction end, perhaps indicative of sniping. The corresponding velocity curve slowly increases and culminates in a peak. Further inspection reveals that most of the “j”-shaped auctions start at the lowest opening bid of \$0.01.

A second type of price curve that we identify is the straight line where the slope depends on the ratio of the opening and closing prices. The corresponding velocity curve is constant. These curves represent auctions with little to no bidding in the middle of the auction. Since all of our auctions transacted, a flat price curve represents auctions with a high opening bid (set by the seller), and therefore, very little bidding. In our data, the 42 straight line curve auctions vary widely in their

opening and closing bids, but they all open above \$150.

The third typical shape in our data (61 auctions) is a stretched-out “s”-shaped curve which reflects auctions where the price increases slowly, then jumps up during midauction, and slowly increases to the close. The corresponding velocity curve spikes at the start of the price jump and then decreases when the price slows down again, resulting in a single “hump” shape. This can be indicative of jump bidding that occurs during midauction, where a single bidder raises the price drastically.

4.2.1.2 Temporal Groupings of Curve Types

When searching for auctions for a particular item on eBay, users are presented with information on all open auctions for that item as well as for auctions that closed in the last 60 days. A typical user would examine each of the open auctions (their current price, number of bids, and closing time) and perhaps some closed auctions to learn about closing prices. All these factors can influence how this bidder will place a bid. In addition, sellers can also use this information to schedule their auction, to set their opening bid, etc. We therefore expect to see temporal patterns in bidding (and posting) behavior in concurrent auctions. These patterns can manifest themselves as similarities in bidding behavior, for example when bidders mimic a sniping behavior. Or they can show dissimilarities, for instance, if low closing price auctions lead new sellers to change strategy.

Our rug plots (Figures 4.3-4.4) show temporal effects such as clustering of shapes in certain periods or the lack of such clustering in others. One global obser-

vation is that the straight line price curves appear to be scattered throughout the collection period. As mentioned earlier, many of these auctions have a high opening bid (sometimes higher than the 75th percentile of daily closing price), perhaps because there is something special about the product, like an included accessory.

In general, during periods with many auctions, it is useful to zoom-in on the x-axis to better separate curves and make their shape more visible. For this reason we concentrate on three periods: the calendar start, the zoomed-in period of March 29 - April 8 (more visible in the bottom panels of Figures 4.3-4.4), and the calendar end. During the beginning of the data period (before March 24), we see mostly “s”-shaped price curves and a few straight line price curves. The main distinction between the two appears to be the opening bid, with the “s”-shaped price curves starting at lower prices. This is more visible in the velocity rug plot where the first group of auctions has a positive velocity “bump” at midauction (corresponding to the “s”-shaped price curves), followed by a period with nearly flat velocity curves. This raises questions about the bidding behavior during this period. What is going on during this flat price velocity period? Why was there no jump bidding during this period?

The second period (March 29 - April 8) contains many concurrent auctions (see bottom panels of Figures 4.3-4.4). First, we see a global peak in velocity around April 3, as indicated by the median daily acceleration, indicating fast price increases during this period. Second, during this entire period, we see two dominant types of auctions going on: a few “j”-shaped price curves (more easily identified in the zoomed-in velocity rug plot, as curves that end high), followed by a large group of

“s”-shaped price curves (i.e., single-humped velocity curves). The price curves of both types appear to intersect mainly around \$100-\$150 perhaps suggesting that bidders who see later closing auctions reaching a price in this range will tend to jump bid to this amount in auctions soon to close, reasoning that the price will most likely reach (at least) this amount.

Finally, during the last period of the calendar, there appear to be more and more auctions with high opening bids. One possible explanation is that the Palm M515 has been auctioned long enough by this time that an opening bid of \$50 or \$100 seems reasonable to bidders and sellers . The resulting price curves are mostly “j”-shaped curves, and many of them tend to close lower than the 25th percentile of daily price.

4.2.1.3 Other Groupings of Curves

The rug plot allows one to explore groupings not only of curve types but also of other relevant factors such as the number of bidders, the seller rating, and the auction duration. This can be performed by coding each curve according to the level or category of the variable of interest. Such an extension allows investigating hypothesized concurrency effects such as: how does a high rated seller’s auction affect the bidding progress in a concurrent low rated seller’s auction? Or, do sellers choose the auction duration according to the durations of other open auctions for the similar item?

To show how this can reveal interesting effects, consider Figure 4.5, which

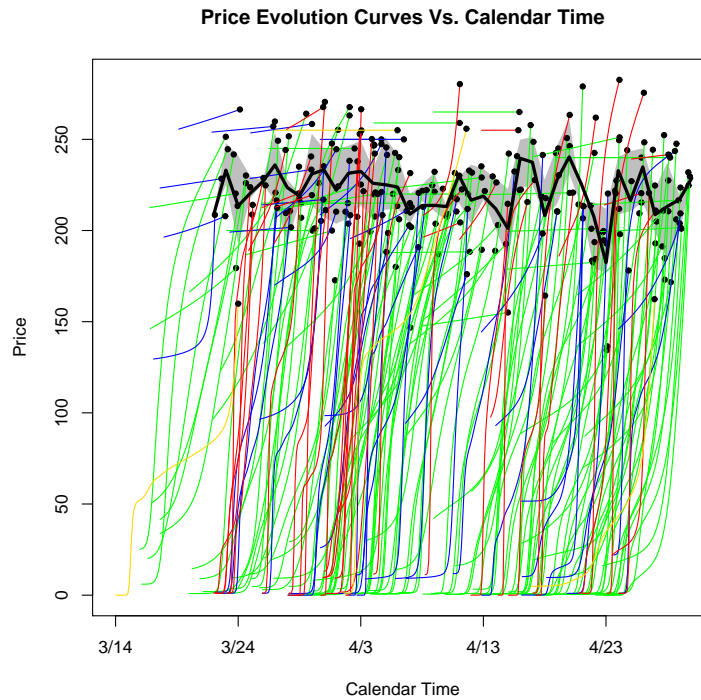


Figure 4.5: Price evolution by auction duration (green = 7-day, blue = 5-day, red = 3-day, gold = 10-day.)

displays the price curve rug plot from Figure 4.3 coded by auction duration. Besides the popularity of 7-day auctions, we see that some auction durations seem to group temporally while others do not exhibit such a pattern. There are two periods when 3-day auctions were very popular and a number of periods where 5-day auctions are prevalent. In contrast, 10-day auctions are relatively rare and do not seem to group together.

4.2.2 Time Grouped Box Plots

In order to study the relationship between closing prices of nearby auctions as well as trends over time, we group each auction's closing price into categories based on the calendar date and ending time of the auction. Using multi scale box

plots for visualizing closing prices in a series of online auctions was proposed by Shmueli and Jank (2005). In particular, they propose an interactive computation of box plots according to temporal scales of interest (“STAT-zoom”). To incorporate the volume of auctions within each temporal window, the widths of the box plots are proportional to the sample size, or alternatively, the box plots are coupled with histograms.

To study the relationship between selling prices of concurrent or nearly concurrent auctions we apply the same method. The levels of grouping we examine are weekly, daily, and half-daily. Figure 4.6 displays time grouped box plots. The width of the box is proportional to the square root of the number of auctions (McGill et al., 1978) to show the volume of auctions closing in a particular period. We see that the median of weekly closing prices is relatively constant over time whereas median prices are highly variable on the daily and half-daily scale. This is comparable to the stock market where prices are highly volatile within a day but less so when an entire week is considered. The within day price fluctuations could be attributed to different time zones that bidders live in or to day/night differences in bidding activity.

While the width of the box plot is visible in the weekly displays it is not easily discernible at the daily and half-daily scales; therefore, we enhance the box plots with histograms of auction closings. Figure 4.7 displays box plots that are coupled with histograms that contain the number of observations in each group. For the daily displays (top of Figure 4.7), the number of auctions closing on each day is between 1 and 10. In some cases where there are many auctions, the median closing

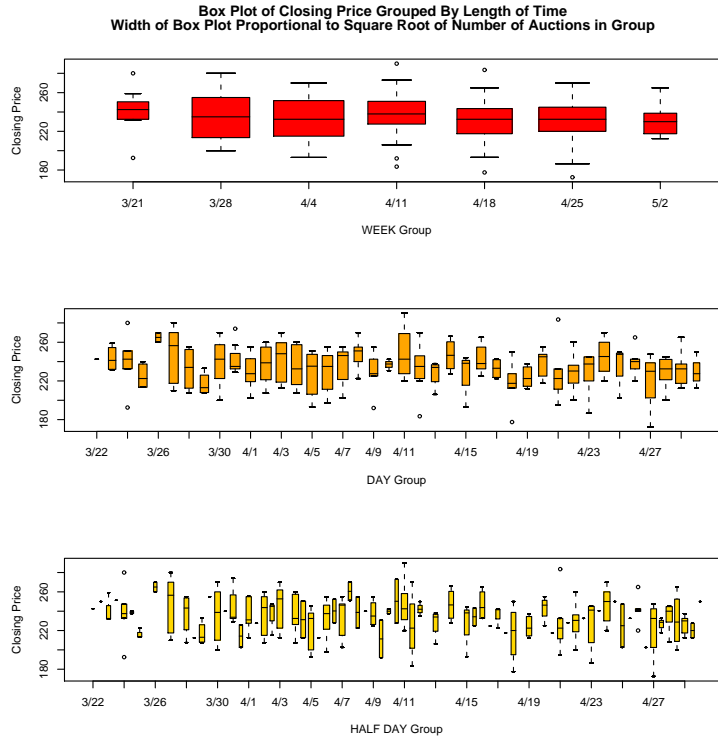
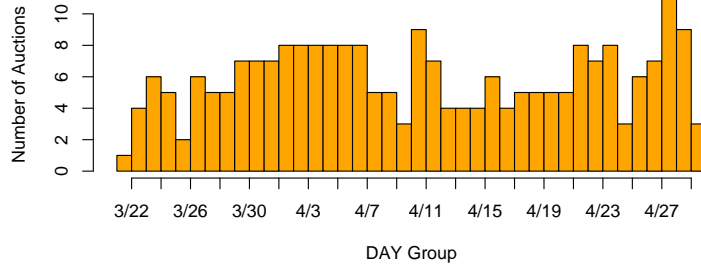
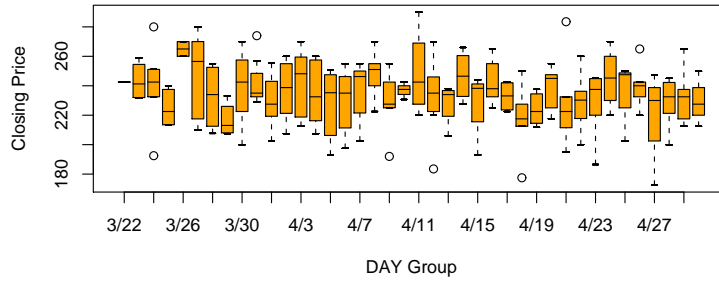


Figure 4.6: Grouped box plots of closing price with width proportional to the square root of the number of auctions.

price is small as would be expected by economic theory since as supply increases, price decreases. However, this is not always the case due to factors such as seller rating, shipping costs, experience of the bidders, etc. that may also influence price. For instance, quite a few auctions end around April 3 and result in a high median price, much higher than the median price on the days before or after where the auction volume is comparable.

The histograms provided in the half-daily displays (bottom of Figure 4.7) show that most auctions close only during one half of the day, probably due to the fact that people sleep during the other half! However, there is also a significant number of auctions that close at night (e.g., around April 27) which may be attributed to time differences or simply to unfortunate and sloppy seller choices.

**Box Plot of Closing Price Grouped By DAY
and Histogram of Number of Auctions Closing Each DAY**



**Box Plot of Closing Price Grouped By HALF DAY
and Histogram of Number of Auctions Closing Each HALF DAY**

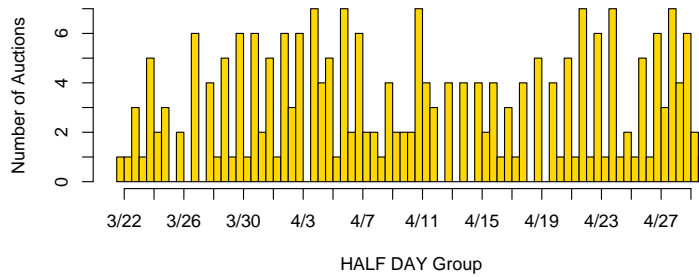
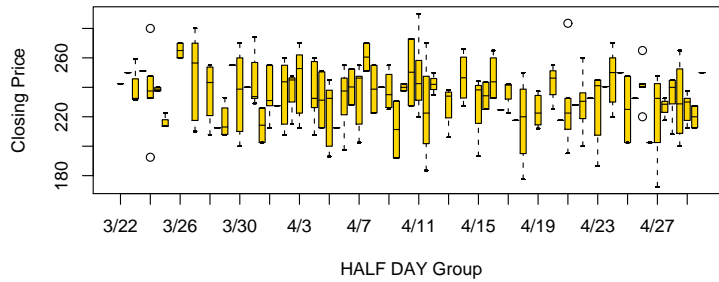


Figure 4.7: Closing price box plots and auction frequency for day (top) and half day (bottom) groups.

4.2.2.1 Comparing Medians of Adjacent Auctions

To better differentiate between adjacent medians, we use the test by Chambers et al. (1983) that compares medians of box plots. The histogram function in R performs this test with the option “notch.” The notches extend to $\pm 1.58IQR/\sqrt{n}$ where IQR is the interquartile range. If the notches overlap, then the medians are not significantly different with 95% confidence. This test assumes asymptotic normality of the median and roughly equal sample sizes. Both assumptions seem reasonable with the auction data.

There are a total of 41 daily box plots between March 22 - April 30 in Figure 4.7. In order to declutter the graphical display, we partition the 41 plots into 3 groups (March 21-April 10, April 1-April 20, and April 11-April 30) and enhance them with notches described above. The result can be seen in Figure 4.8. For instance, we see that the notches for March 26 and March 27 overlap; median prices for these two days are *not* significantly different. In contrast, the median prices of March 25 and March 26 are significantly different. In general, while the median daily closing prices are variable, the differences are not statistically significant in most instances. There also does not appear to be any day of the week effects: the median daily closing price for any particular day of the week is not consistently significantly different than the other days of the week.

Figure 4.8 also shows the notched box plots for the half-daily displays. In this display, we find more statistical differences between adjacent medians. This suggests that the time of day can have an advantageous effect for the bidder or the seller.

This finding also coincides with Snir (2007) who finds lower bids now (i.e., at night) if more auctions are expected in the future (i.e., during the day).

4.2.3 Moving Statistics Plots

Another way of visualizing time trends in prices (or any other measure of interest) is by applying *moving statistics plots*, that is, via graphing the moving average or median. Moving statistics plots are a form of smoothing the data. The particular statistic is computed over subsets of the data. We use the observation index for the purpose of grouping (the earliest closing auction is denoted “1”, the next closing auction is denoted by “2”, etc.) For example, using a moving window of 3 auctions for our 236 auctions yields a series of 234 statistics.

Note that this method does not take into account the temporal distance between auctions. This means that moving statistics do not differentiate between auctions that are close in time and those that are far apart. One possibility is to use weights that reflect these temporal distances.

A moving statistic plot helps in understanding time trends in the data, and unlike side-by-side box plots, they present a smoother transition over time. Figure 4.9 shows moving average plots for time-windows of 3, 5, and 10 auctions. The dark solid line shows the moving average, and the lighter lines correspond to upper and lower 95% confidence bands. The straight line through the center of the data shows the overall average closing price.

The increased variability in average price for the 3- and 5-auction time windows

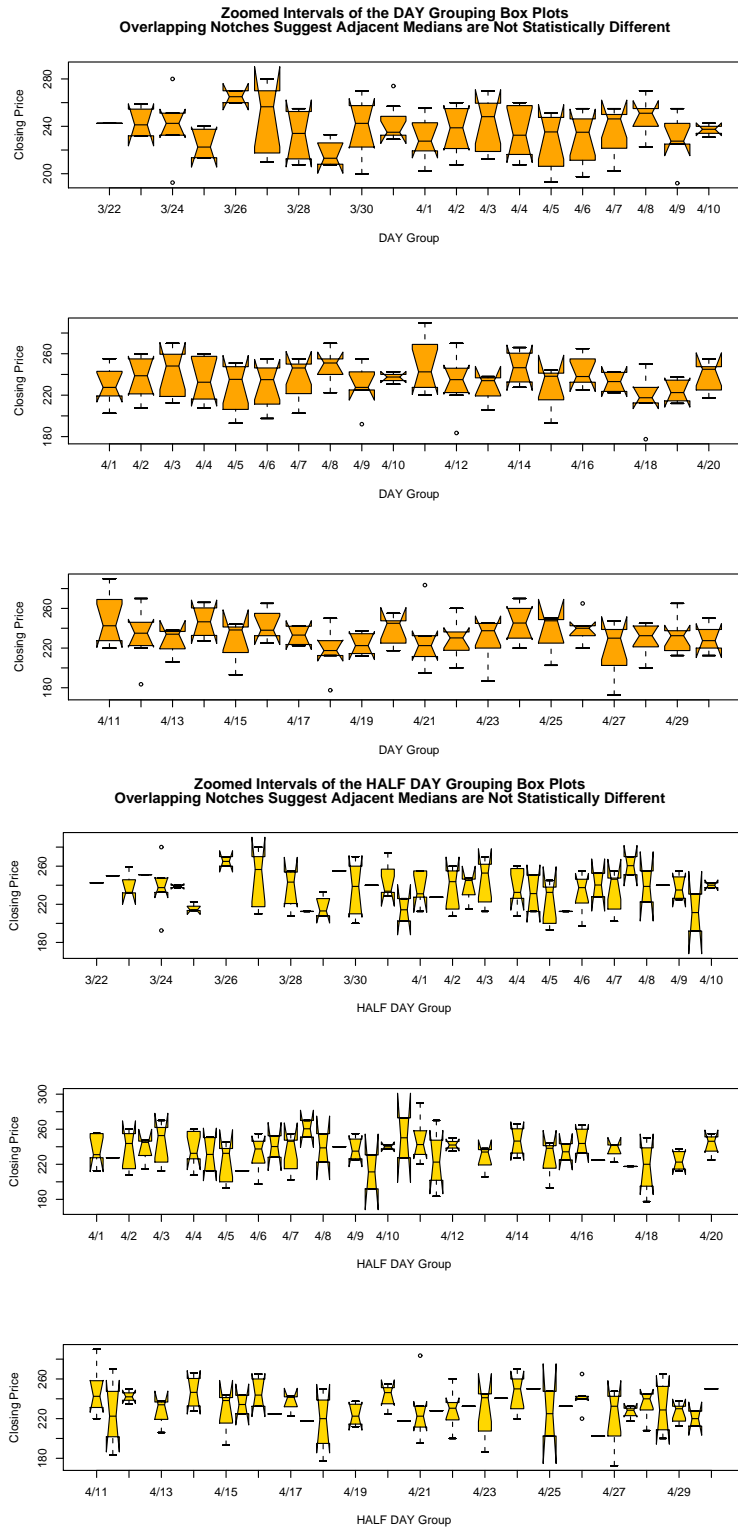


Figure 4.8: Closing price box plots zoomed at different overlapping periods of calendar time for day groups (top three) and half day groups (bottom three).

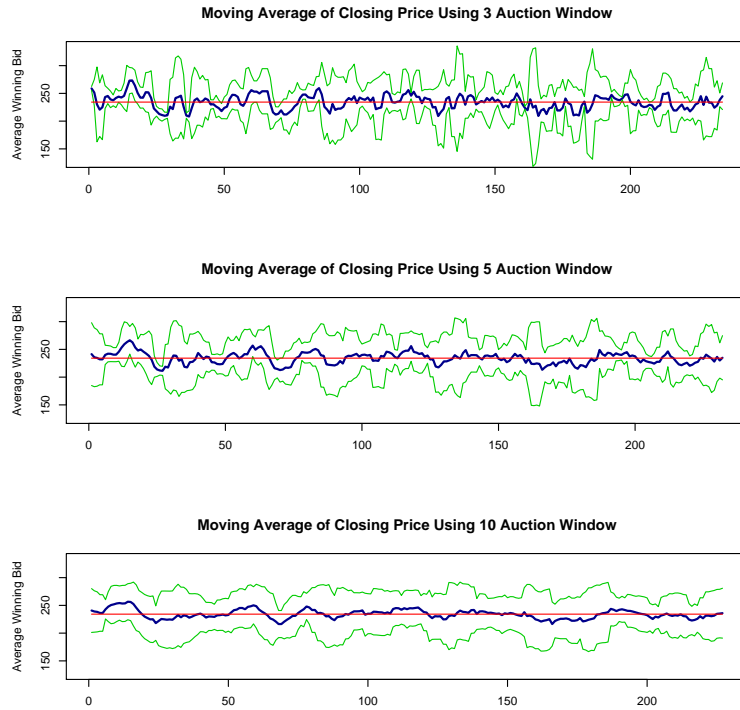


Figure 4.9: Closing price moving average based on different numbers of auctions.

compared to the relatively constant 10-day window suggests, again, that auctions are similar to the stock market: while there exists some variability in the short run, prices tend to converge to the market value in the long run. This also suggests that the smart (or lucky) bidder can take advantage of the variability to obtain a price that is lower than the market value.

4.2.4 Price Autocorrelation

We now examine the correlation between adjacent auctions. If the closing price in an auction affects the closing price in nearby auctions, we expect to find some degree of autocorrelation across the series of prices, especially if the groups are small enough. One is less likely to find autocorrelation in weekly prices because

of the large price variability in auctions every week and also possibly because the most popular auction duration is one week long. Narrowing the window to the day or half-day window may provide a better indication of autocorrelation. Ideally, we would like to look at hourly autocorrelation; unfortunately, this data set is too sparse to do so. Since most auctions do not close in the middle of the night, it would be hard to find a data set rich enough for this type of investigation. Therefore, a method for dealing with this type of missing values is necessary.

Even in the half-daily series we encounter missing data. Our solution is to impute missing values with the average of prices from one time period before and after the missing value. This “nearest neighbor” approach can be extended to multiple neighbors if it is reasonable to assume that the price is stable within a wider window of auctions. Another possibility is to impute the missing value using only information prior to the missing value, to reflect that fact that bidders do not have information about future prices.

Figure 4.10 shows price autocorrelation based on the series of moving averages and moving medians for several window widths. It appears that there is no autocorrelation in the means or medians at any window. We suspect that the groups are too large and do not differentiate well between auctions that closed hours apart and those that closed minutes apart. Perhaps a richer data set would reveal more or a moving statistic that would account for the temporal distances between auctions. The general question is how to represent time lags in unequally spaced observations (e.g., Jank and Shmueli (2006)).

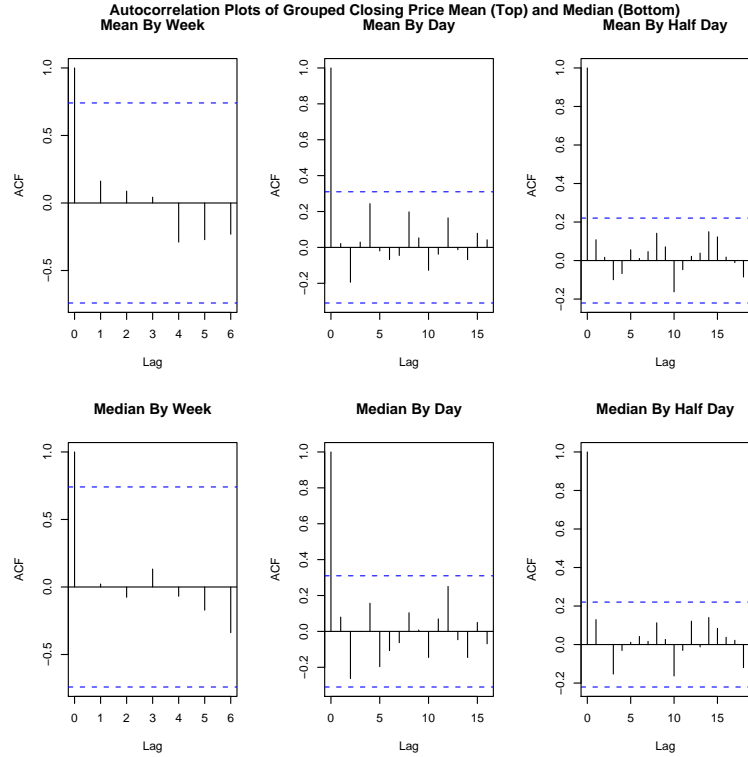


Figure 4.10: Autocorrelation for week (left), day (middle), and half day (right) groups for mean (top) and median (bottom).

4.3 Conclusions

Concurrency is prevalent in the world of commerce. Whether it is stocks, sales promotions of competing products, or mortgage rates offered by lenders, action often results in reaction. Concurrency is particularly prevalent in the online environment since the openness of the Internet allows players to observe each other's moves in real time. In this work, we consider concurrency of online auctions on eBay. We approach the study of concurrency from a visual point of view and propose a series of visualizations that are suitable for the special structure of bid data. Considering the entire price evolution during an auction and its dynamics, we find several patterns and raise new research questions. We also find that online auctions tend to resemble

the stock market in that most auctions tend to close around their market value but still contain variability that allows bidders to purchase products significantly below market value. While some of this variability could be attributed to other factors such as seller rating and the winner's experience, it creates opportunities for an eBay bidder. We also find a difference in median closing prices at different times of the day. A bidder, for instance, may be able to obtain a lower value by taking part in auctions that close later in the evening when fewer people are taking part in an auction. These observations are all based on the exploratory analysis performed here and can serve as a basis for further studies of concurrency in online auctions.

Chapter 5

A Family of Growth Models for Representing the Price Process in

Online Auctions

5.1 Introduction and Motivation

The wide and growing popularity of online auctions creates enormous amounts of publicly available auction bid data providing an interesting and important topic for research. These data pose special statistical challenges because of their special structures. This research focuses on uncovering the entire continuous price process during an auction.

Bids during an online auction arrive at unevenly-spaced discrete time points chosen by bidders; however, we are interested in obtaining a continuous approximation of the underlying continuous price evolution: the price at any time during the auction. From a visualization point of view, a series of discrete unevenly-spaced bids (e.g. presented as a scatterplot of bid vs. time) loses the conceptual as well as continuous nature of the price process. Furthermore, such plots do not scale to multiple auctions. A better representation is a continuous function with only a few parameters. Such a representation is conceptually more appealing, is more parsimonious, and can be further used for analyses such as clustering or regression models. Finally, the underlying continuous process may depend on derivatives of

the function. A smooth price curve allows the calculation of derivatives (i.e., the first derivative is price-velocity and the second derivative is price-acceleration).

In FDA, nonparametric smoothing techniques are used to estimate a functional object from discrete measurements (as described in Chapter 3). Examples are kernel smoothers, polynomial splines, and monotone splines (Ramsay and Silverman, 2005; Ramsay, 1998). Smoothing should be done in such a way that the resulting underlying object adequately represents the continuous process. In the auction setting, we know that the price is always positive and monotonically nondecreasing, so our smoothed price curve must reflect this nature.

To date, all of the smoothing methods applied to online auctions have proven to involve the specification of many parameters such as the number and position of knots, a roughness penalty parameter, and the polynomial order. Resulting curves capture the price curves reasonably well; however, the fit often requires manual visual inspection (for parameter selection). Smoothing splines, which suffer from edge fitting, often result in a deteriorated fit at the start and end of the price curve and produce curves that are not necessarily monotone. Monotone smoothing splines produce monotone curves; however, they are computationally intensive and require lengthy run times for even a moderate number of auctions. Lastly, nonparametric smoothing does not provide an explanatory model for the price process. This motivated us to explore meaningful parametric representations of the price process. A parametric approach would be more elegant in the sense that it would provide a theoretical explanation of the process, it would potentially be computationally fast, and it would provide a more parsimonious representation.

For example, in Chapter 4, monotone smoothing is employed to estimate the price curves of a set of Palm M515 auctions. We find three distinct shapes of price curves: “j”, stretched-out “s”, and straight line curves. The most frequent shape in the Palm Pilot data set is a concave-up “j”-shaped price curve which represents auctions with gradual price increases until mid-to-late auction and then a price jump towards the end; the second most popular shape is the straight line, where the slope depends on the ratio of the opening and closing prices; and the third typical shape is a stretched-out “s”-shaped curve which reflects auctions where the price increases slowly, jumps up during mid-auction, and slowly increases to the close. These price curve shapes are the result of several bidding styles documented by Bapna et al. (2004b). Determining the number of each type of curve requires visual inspection of every single price curve. Clearly, a better method for grouping curves is needed beyond physical examination.

The goal of this paper is to introduce, within an FDA framework, a new *parametric* family of growth functions that describe the price processes in online auctions. This chapter is organized as follows: Section 5.2 discusses the data we use in this research. In Section 5.3, we discuss two nonparametric smoothing methods that have been used to describe auction price evolution and their limitations. In Section 5.4, we describe two popular growth models (logistic and exponential) and two additional useful growth modes (logarithmic and reflected-logistic) and show how our parametric family of growth models is used for representing auction price growth. Section 5.5 introduces an automated selection procedure to choose the best growth model. Section 5.6 compares the quality of curves obtained through non-

parametric methods with those from our parametric family of curves. In Section 5.7, we suggest several further uses of growth models. We conclude in Section 5.8 with final remarks and future research.

5.2 Wristwatch Data

We use the luxury wristwatch data described in Section A.1 to test our family of growth models because of its variability. This sample includes a variety of items in terms of make and model, new and used, and closing price. We know from the literature (Bapna et al., 2004b; Wang et al., 2007a) that auction attributes such as opening price, seller experience, number of bids, etc. affect not only the closing price of an auction but also the entire price process. Indeed, our sample is varied in all of those attributes.

Figure 5.1 shows the distribution of the number of bids in our sample for each day in the auction. Typically, there is some bidding activity at the auction start (first day), followed by a period of very little activity, culminating in a surge of bidding at the very end of the auction (Shmueli et al., 2004). This last moment bidding is often referred to as “sniping” (Bajari and Hortascu, 2003; Roth and Ockenfels, 2002). Clearly, when modeling the price process, the start and end of the auction are of particular importance.

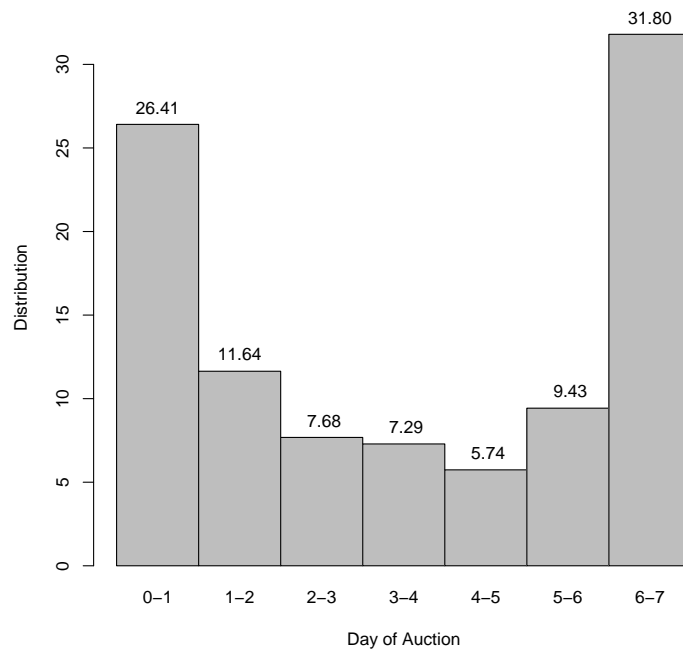


Figure 5.1: Distribution of daily bids over the course of 472 luxury wristwatch auctions.

5.3 Representing Auction Price Nonparametrically

There have been predominantly two approaches for representing the price process in online auctions. Jank et al. (2007) use penalized polynomial smoothing splines (p-splines), and Hyde et al. (2006) use penalized monotone splines. A comparison of the two approaches for auction data is given in Alford and Urimi (2004). We now describe each of the two methods and their properties.

5.3.1 Smoothing Splines

Each auction has bids placed at different times. Rather than applying smoothers to the raw data directly, we apply them to a derived data set that are sampled on

the same set of time points for all auctions as follows. Consider the observed price during an auction, represented by the step function in Figure 1.2, where there is a new step every time a bid is placed. We use this step function to obtain our sampled data by selecting a set of knots at times $\tau_1, \tau_2, \dots, \tau_L$ and determining the corresponding auction prices at those knots, $y_1^*, y_2^*, \dots, y_L^*$. The noisy bid data are then discarded and replaced with the observations (τ_i, y_i^*) . This transformation then allows us to use the same method to estimate the price process in different auctions using the same smoother method.

The polynomial spline (Green and Silverman, 1994) of order p is given by

$$f(t) = \beta_0 + \beta_1 t + \beta_2 t^2 + \dots + \beta_p t^p + \sum_{l=1}^L \beta_{pl} (t - \tau_l)_+^p \quad (5.1)$$

where $u_+ = uI(u \geq 0)$ is the positive part of the function u . Many smoothers of this type tend to fit the data too closely (and thus model the noise); therefore, a *roughness penalty* approach is commonly employed. This method takes into account the trade off between data fit (i.e., minimizing $f(t) = \sum_j (y_j^* - f(t_j))^2$) and function smoothness. A popular measure of roughness, which measures degree of departure from a straight line, is of the form

$$PEN_m(t) = \int [D^m f(t)]^2 dt \quad (5.2)$$

where $D^m f, m = 1, 2, 3, \dots$ denotes the m^{th} derivative of the function f . A highly variable function will yield a high value of $PEN_m(x)$. If the highest derivative of interest is m , then using $m + 2$ as the polynomial order will assure m continuous derivatives. For online auctions, where the first and second derivatives have been

shown to be of interest, we use polynomials of order $m = 4$. The penalized smoothing spline f minimizes the penalized squared error

$$PENSSSE_{\lambda,m} = \int (y(t) - f(t))^2 + \lambda PEN_m(t). \quad (5.3)$$

When the roughness parameter is set to $\lambda = 0$, the penalized squared error drops out, and the function fits the data. Larger values of λ penalize the function for being curvy, and as $\lambda \rightarrow \infty$, the fitted curves approach a linear regression. Ramsay and Dalzell (1991) and Ramsay (1998) suggest that the smoothing parameter λ can often be chosen by inspection of the smooths or through optimizing metrics such as the generalized cross-validation (GCV).

We have encountered a number of challenges using penalized smoothing splines in the online auction context. First, and most detrimental, is that the created functions are not always monotone nondecreasing, as auction price necessarily must be. Second, the functions are often very wiggly, especially at the ends. This is particularly egregious in the online auction context where the start and end of the auction are of major importance. Third, there tends to be a large number of coefficients to estimate (due to adequate choices of knots and the polynomial order). Finally, there are multiple decisions about smoothing parameters that must be made in advance: the number and position of knots, the polynomial order, and the roughness penalty parameter λ , in order to provide a reasonable fit to the entire set of auctions. An alternative, where each auction is fit separately using a different set of parameters, will lead to confounding in the analysis results (see Jank et al. (2007)).

5.3.2 Monotone Splines

Since the bidding process by nature is nondecreasing, Hyde et al. (2006) use monotone smoothing splines to represent the price process. The idea behind monotone smoothing (Ramsay, 1998) is that monotone increasing functions have a positive first derivative. The exponential function has this property and can be described by the differential equation $f'(t) = w(t)f(t)$. This means that the rate of change of the function is proportional to its size. Consider the linear differential equation

$$D^2 f(t) = w(t)Df(t). \quad (5.4)$$

Here, $w(t) = \frac{D^2 f(t)}{Df(t)}$ which is the ratio of the acceleration and velocity. It is also the derivative of the logarithm of velocity which always exists (because we define velocity to be positive by the equation $Df(t) = e^{w(t)}$). The differential equation has the following solution:

$$f(t) = \beta_0 + \beta_1 \int_{t_0}^t \exp\left(\int_{t_0}^v w(v)dv\right) du \quad (5.5)$$

where t_0 is the lower boundary over which we are smoothing. After some substitutions (see Ramsay and Silverman (2005)), we can write

$$f(t) = \beta_0 + \beta_1 e^{wt}. \quad (5.6)$$

and estimate β_0 , β_1 , and $w(t)$ from the data. Since $w(t)$ has no constraints, as $f(t)$ does in the form of the differential equation, it may be defined as a linear combination of K known basis functions (i.e., $w(t) = \sum_k c_k \phi_k(t)$). Examples of a basis functions are $\phi_k(t) = t$ which represents a linear model or $\phi_k(t) = \log(t)$ which

is a nonlinear transformation of the inputs. The penalized least squares criterion is thus

$$PENSSSE_{\lambda} = \sum_i [y_i - f(t)]^2 + \lambda \int_0^T [w^2(t)]^2 dt. \quad (5.7)$$

While monotone smoothing solves the wiggly problem of the penalized smoothing splines, some of the same challenges remain and new ones arise. First, monotone smoothing is computationally intensive. The more bids, the longer the fitting process. Second, we still cannot fit the original raw data but rather the derived data (τ_i, y_i^*) (obtained from the step function). Finally, as with smoothing splines, the researcher must determine the number and location of knots and the roughness parameter λ that provide a reasonable fit to the entire set of auctions.

5.4 Representing Price Evolution Parametrically

We now introduce a parametric family of four growth models that are able to capture the price evolution in many types of auctions. These are exponential growth, logarithmic growth, logistic growth, and reflected-logistic growth functions. Exponential and logistic growth functions have long been used to model population growth, dissemination of information, spread of disease, and more.

Our parametric approach is elegant, computationally fast, and parsimonious. It allows automated fitting, and there is no need to specify any parameters in advance. We choose models that are theoretically relevant in terms of monotonicity (to accurately reflect the price process) and that provide insight into the price process in online auctions.

Our approach, although motivated by online auctions, actually provides an alternative to the nonparametric smoothing methods that are customary in the field of Functional Data Analysis (FDA). This research opens the door for parametric curves to be the basis of FDA.

In the following, we describe each of the four models. Their functional form, derivative form, and parameters are summarized in Table 5.1.

5.4.1 Exponential Growth

5.4.1.1 Exponential Model

Exponential growth has been used for describing a variety of natural phenomena including the dissemination of information, the spread of disease, and the multiplication of cells in a petrie dish. In finance, the exponential equation is used to calculate the value of interest-bearing accounts compounded continuously. The idea behind exponential growth is that the rate of growth is proportional to the function's current size; that is, growth follows the differential equation

$$Y'(t) = rY(t), \tag{5.8}$$

or the equivalent equation

$$Y(t) = Ae^{rt}, \tag{5.9}$$

where t is time, and $r > 0$ is the growth constant. Equivalently, exponential decay, when $r < 0$, can model phenomena such as the half-life of an organic event.

From a theoretical standpoint, we hypothesize that exponential growth can

Table 5.1: Price-evolution, -velocity, and -acceleration functions for growth models.

| Growth Model | Price Evolution | Price Velocity | Price Acceleration | Parameters |
|--------------------|--|--|--|------------|
| Exponential | $Y(t) = Ae^{rt}$ | $Y'(t) = Are^{rt}$ | $Y''(t) = Ar^2e^{rt}$ | A, r |
| Logarithmic | $Y(t) = \frac{\ln(\frac{t}{A})}{r}$ | $Y'(t) = \frac{1}{rt}$ | $Y''(t) = \frac{-1}{rt^2}$ | A, r |
| Logistic | $Y(t) = \frac{L}{1+Ce^{rt}}$ | $Y'(t) = \frac{-LCre^{rt}}{(1+Ce^{rt})^2}$ | $Y''(t) = \frac{-LCr^2e^{rt}(1-Ce^{rt})}{(1+Ce^{rt})^3}$ | C, r |
| Reflected-Logistic | $Y(t) = \frac{\ln(\frac{L}{t}-1)-\ln(C)}{r}$ | $Y'(t) = \frac{-L}{rt^2(\frac{L}{t}-1)}$ | $Y''(t) = \frac{L(L-2t)}{rt^4(\frac{L}{t}-1)}$ | C, r |

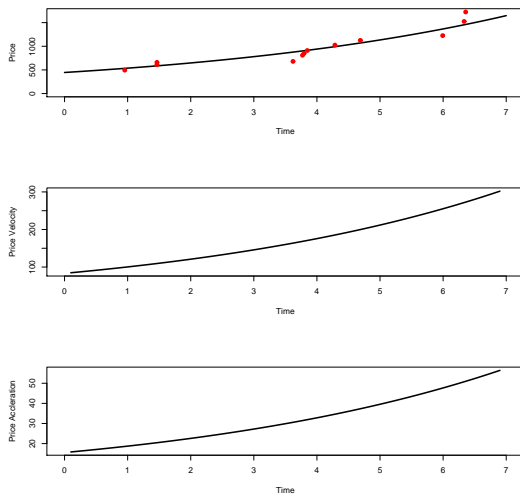


Figure 5.2: Price (top), velocity (middle), and acceleration (bottom) for an auction fit with exponential growth.

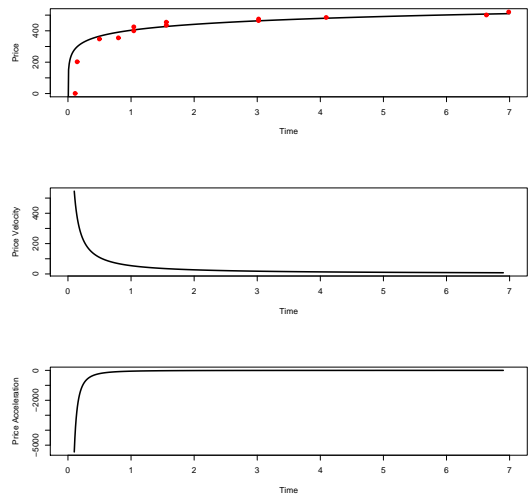


Figure 5.3: Price (top), velocity (middle), and acceleration (bottom) for an auction fit with logarithmic growth.

describe a price process for auctions where there are gradual price increases until mid-to-late auction and a price jump towards the end, perhaps due to sniping (Bajari and Hortascu, 2003; Roth and Ockenfels, 2002). This is reminiscent of the “j”-shaped price curves that Hyde et al. (2006) find. An example of an auction that is captured well with exponential growth is shown in Figure 5.2 (top). The corresponding velocity curve (middle) and acceleration curve (bottom) are proportional to the price curve, as the differential equation implies. The price-velocity and -acceleration are zero or small during most of the auction before spiking at the end.

5.4.1.2 Logarithmic Model

We have learned from Bapna et al. (2004a) that inexperienced bidders who do not understand the proxy bidding mechanism on eBay tend to bid high early,

which increases the price early in the auction. Since the item for sale in an auction can often be purchased in other venues, such as online and brick-and-mortar stores, there is a “market value” that caps price. The existence of a market value for the product necessitates that the price process flattens out for the remainder of the auction. This type of price behavior tends to be rare, as most bidders do not wish to reveal their bids early in the auction. The inverse of the exponential function,

$$Y(t) = \frac{\ln\left(\frac{t}{A}\right)}{r}, \quad (5.10)$$

which is called logarithmic growth, approximates this price process well. We choose a form of the logarithmic model that is the mapping of the original exponential model over the line $y = x$. An example of an auction whose price process can be described well by a logarithmic price process is as shown in Figure 5.3). In this model, the velocity starts at its maximum and then decays to little or zero velocity as the auction progresses. The acceleration is always negative, since price increases more slowly throughout the auction, and approaches zero at the end of the auction (where very little change in price is occurring).

5.4.2 Logistic Growth

5.4.2.1 Logistic Model

While exponential growth often makes sense over a fixed period of time, in many cases, growth can not continue indefinitely. For example, there are only a finite number of people to spread information or disease; the petrie dish can only hold a maximal number of cells. A typical application of the logistic equation is in

population growth. In the beginning, there are seemingly unlimited resources and population grows increasingly fast. At some point, competition for food, water, land, and other resources slows down the growth; however, population is still growing. Finally, overcrowding, lack of food, and susceptibility to disease limit the population to some maximal carrying capacity.

In auctions, it is possible that growth starts out exponentially with a big price increase in the middle of the auction, perhaps due to an inexperienced bidder. In the presence of a market value (or “carrying capacity”), the increased price slows competition, as smart bidders will make sure not to overpay for the item, and it is necessary for the price to flatten out near the end of the auction. The closing price that we witness is analogous to the carrying capacity L in the logistic growth function.

The logistic model is given by

$$Y(t) = \frac{L}{1 + Ce^{rt}}, \quad (5.11)$$

and the differential equation is

$$Y'(t) = rY(t) \left(\frac{Y(t)}{L} - 1 \right), \quad (5.12)$$

where L is the carrying capacity, t is time, r is the growth rate, and C is a constant. Logistic growth forms a stretched-out “s”-shaped curve, discussed by Hyde et al. (2006), where the price increases slowly, then jumps up during mid-auction, and finally levels off through the end of the auction. This price process and dynamics can be seen in Figure 5.4. The velocity is small or zero at the beginning and end of the auction, where there is little change in price, with a peak mid-auction corresponding

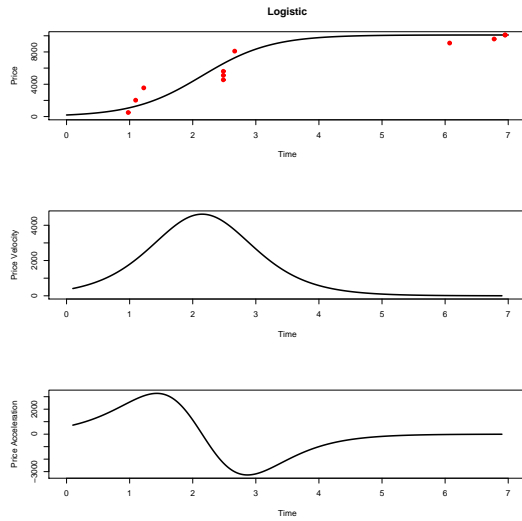


Figure 5.4: Price (top), velocity (middle), and acceleration (bottom) for an auction fit with logistic growth.

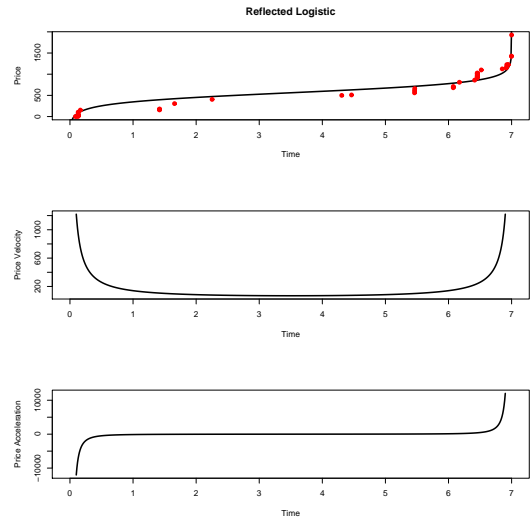


Figure 5.5: Price (top), velocity (middle), and acceleration (bottom) for an auction fit with reflected-logistic growth.

to the price increase. The acceleration is zero for most of the beginning and end of the auction. It peaks during the first part of the growth spike, where price is increasing at an increasing rate, followed by a valley during the second part of the growth spike, where price is increasing at a decreasing rate.

5.4.2.2 Reflected-Logistic Model

Shmueli et al. (2004) found that there is some bidding activity at the auction start (first day), followed by a period of very little activity, culminating in a surge of bidding at the very end of the auction. The early price increases is indicative of early bidding by inexperienced bidders, and the price spike at the end may be caused by sniping. We hypothesize that the early bidding will increase price, followed by

a period of no price increase when little or no bidding occurs, and then a large price increase with the surge of bidding towards the very end of the auction. This type of price process may be described well by the inverse of logistic growth, or reflected-logistic growth, given by the function

$$Y(t) = \frac{\ln(\frac{L}{t} - 1) - \ln(C)}{r}. \quad (5.13)$$

An example of reflected-logistic growth is shown in Figure 5.5. The velocity peaks at the start and end of the auction where there are jumps in price and little or zero velocity during the middle of the auction where price does not change significantly. The acceleration curve is similar in shape to the price curve; however, it is negative in the beginning (where price increases at a decreasing rate) and positive at the end (where price increases at an increasing rate).

5.4.3 Fitting Growth Models

Unlike the nonparametric smoothers, we fit the growth models directly to the live bids. A simple and computationally efficient method for fitting each of the growth models is by linearizing the function and then performing least squares. Since we are especially interested in obtaining an accurate fit at the beginning and end of the auction, it is usually necessary to add two additional points representing the price at the start and close of the auction: $(t = 0, y = \min(y_j))$ and $(t = l, y = \max(y_j))$ where l is the duration of the auction, and y_j is the value of bid j . Note that $\min(y_j) = \text{opening price}$ and $\max(y_j) = \text{closing price}$. It is not necessary to add these extra points for logistic growth, where the maximum price is already incorporated

into the function by defining $L = \max(y_j)$. Further, empirical evidence shows that auctions whose underlying price process is logistic tend to start close to zero, where logistic growth must start.

5.4.3.1 Fitting Exponential Growth

The exponential growth model from equation (5.9) can be linearized as

$$\ln Y = \ln A + rt. \quad (5.14)$$

We fit this model directly to the live bids with the two additional points in order to constrain the price at the start and end of the auction. In this case, two parameters, A and r are estimated. Although we can fix A as the opening price since $Y(t = 0) = Ae^{0r} = A$, we chose to estimate both parameters for two reasons: first, empirical evidence shows that the two-parameter estimation allows a better fit at the end of the auction, and second, the other three growth models require estimating two parameters. Thus, it is easier to compare model fit.

The straight line curves that are discussed in Hyde et al. (2006) are a special case of exponential growth. When $r = 0$, we obtain a horizontal straight line, and when r is close to 0, the price curve resembles a straight line with positive slope.

5.4.3.2 Fitting Logarithmic Growth

The logarithmic growth model is given in equation (5.10). As with exponential growth, the two extra points are added to the live bids ensure a good fit at the start and end of the auction. Notice that we can not reduce this function to one parameter

because when $t = 0$, $\ln(0)$ does not exist. Also, we can not linearize this function. Therefore, rather than using optimization methods for parameter estimation where guesses of the initial value are necessary, we fit $T(y) = Ae^{ry}$ and linearize as in the exponential growth case. This time, least squares minimizes over time instead of price.

5.4.3.3 Fitting Logistic Growth

The logistic growth model from equation (5.11), where L is the distribution's asymptote, can be linearized as

$$\ln\left(\frac{L}{y} - 1\right) = \ln(C) + rt. \quad (5.15)$$

We know that $\lim_{t \rightarrow l} = L$ (since for logistic growth $r < 0$ and $L > 0$). Define $L = \max(\text{price}) + \delta$, where $\delta = 0.01$ is needed so that the left hand side (LHS) is defined over all bids y . In this case, there is no need to add the start and closing points to the live bids because defining the asymptote takes care of the fit at the end. Auctions whose underlying price process can be described by logistic growth tend to start out low, so there is also no need to set the start value.

5.4.3.4 Fitting Reflected-Logistic Growth

The reflected-logistic function is given in equation (5.13). As with logarithmic growth, we can not linearize this function. We instead fit $T(y) = \frac{L}{1+Ce^{ry}}$, where $L = l + \epsilon$ ($\epsilon = 0.00001$), to obtain the coefficients for C and r . We need $\epsilon \ll \delta$ since the time range is much smaller than the price range. As with logarithmic growth,

least squares minimizes over time instead of price. Note that here, the extra points are $(t = 0.000001, y = \min(y_j))$ and $(t = l, y = \max(y_j))$ so that the LHS is defined over all bid times.

5.5 Selecting the Best Growth Model

We develop an automated model selection procedure to choose for each auction the best fitting growth model among the four models. The procedure uses a specialized proximity metric that measures the distance between bids and the fitted curve in the two dimensions of time and price. This metric is reminiscent of the Mahalanobis distance. Most model selection criteria only measure the residual distance in the y (price, in this case) dimension; however, we are interested in capturing the best fit in both the price and time dimensions because bid times are informative and are random variables. Furthermore, the fit for the logarithmic growth and reflected-logistic growth models are minimized over the x (time) dimension. If we were to choose between models based simply on the price dimension, we would tend to choose the exponential growth and logistic growth models, even though the reflected models may provide a better representation of the price process (as can be visually seen). While our model selection criteria is primarily aimed at choosing among growth models, the metric may also be used to choose among other methods: growth models, p-splines, monotone smoothing, etc.

5.5.1 Model Selection Metrics

For auction i , let $\{\mathbf{t}_i, \mathbf{y}_i\}$ be the vectors of live bids (t_{ij}, y_{ij}) where bid y_{ij} is placed at time t_{ij} , and the number of bids in the auction is n_i . Define a new vector with two additional price points ($n_i^* = n_i + 2$) that also includes the open and close price of the auction as $\{\mathbf{t}_i^*, \mathbf{y}_i^*\} = \{(t = 0, y = \min(y_{ij})), \{\mathbf{t}_i, \mathbf{y}_i\}, (t = l, y = \max(y_{ij}))\}$, where l is the length of the auction. It is important that we examine the fit at the start and end of the auction because that is where most of the bid activity takes place, they are conceptually important, and also because that is where modeling falls short.

We propose two measures of fit: the weighted sum-of-squares standardized by the range (WSSER) and the weighted sum-of-squares standardized by the variance (WSSEV). Both metrics are weighted averages of fit in the y -direction and fit in the x -direction, using weights w_y and w_x , such that $w_y + w_x = 1$. The WSSER for auction i is defined as

$$WSSER_i = \frac{w_y \sum_{j=1}^{n_i^*} (y_{ij}^* - \hat{y}_{ij}^*)^2}{(\max_j(y_{ij}^*) - \min_j(y_{ij}^*))^2} + \frac{w_x \sum_{j=1}^{n_i^*} (x_{ij}^* - \hat{x}_{ij}^*)^2}{(\max_j(x_{ij}^*) - \min_j(x_{ij}^*))^2}. \quad (5.16)$$

Notice that the denominator is the squared price range in the y dimension and the squared time range (in our case the auction length l) in the x dimension. The WSSEV for auction i is defined as

$$WSSEV_i = \frac{w_y \sum_{j=1}^{n_i^*} (y_{ij}^* - \hat{y}_{ij}^*)^2}{\text{variance}_j(y_{ij}^*)} + \frac{w_x \sum_{j=1}^{n_i^*} (x_{ij}^* - \hat{x}_{ij}^*)^2}{\text{variance}_j(x_{ij}^*)}. \quad (5.17)$$

5.5.2 Model Selection Procedure

In this section, we describe how to automatically select between different growth models and choose the best one. The model selection procedure is as follows:

1. Select weights, w_y and w_x , representing the importance of fit in the price and time dimensions, respectively.
2. Fit each of the four growth models to the live bids of an auction.
3. Compute model selection metric(s).
4. Choose model with best fit (minimum WSSE).

For our sample of auctions, we choose equal weights $w_y = w_x = \frac{1}{2}$ since we are equally interested in fit in the price and time dimensions. One may overweight time (large w_x) if capturing bid timing is of special interest. One such case is in studying bid shilling, where a seller may cancel the auction or illegally bid in his own auction if the price has not reached a certain level by a certain time (Kauffman and Wood, 2005). A researcher may overweight price (large w_y) when the focus is on the price level itself (e.g., using this information to make more informed bid decisions.) Note that overweighting price tends to favor exponential and logistic growth models, whereas overweighting time leads to favoring logarithmic and reflected-logistic models. Examples of overweighting are shown in Figures 5.6 and 5.7. In Figure 5.6, overweighting in the price dimension leads to exponential growth selection, whereas reflected-logistic growth would have been selected had the weights been equal. In Figure 5.7, overweighting in the time dimension leads to reflected-logistic growth

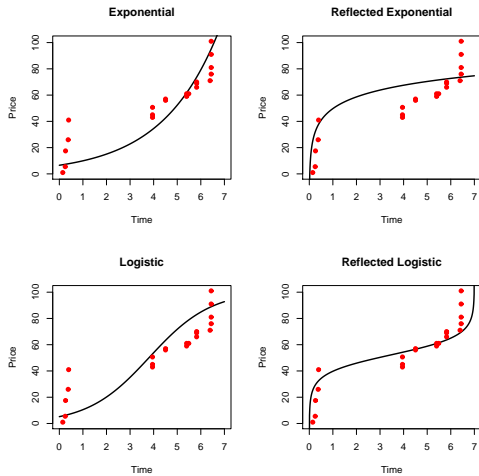


Figure 5.6: Exponential (top left), logarithmic (top right), logistic (bottom left), and reflected-logistic (bottom right) models fit to bids for the same auction. Overweighting in the y-dimension selects exponential growth when reflected-logistic growth should be chosen.

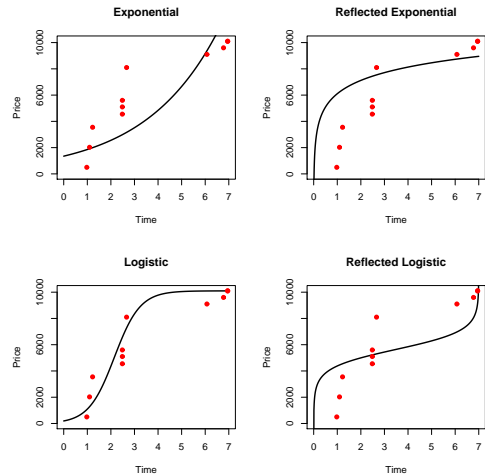


Figure 5.7: Exponential (top left), logarithmic (top right), logistic (bottom left), and reflected-logistic (bottom right) models fit to bids for the same auction. Overweighting in the x-dimension selects reflected-logistic growth when logistic growth should be chosen.

selection, whereas logistic growth would have been selected had the weights been equal. In addition to the task at hand, visual inspection of a subset of the auctions can provide an appropriate weighting scheme.

From empirical evidence, we find that both WSSE measures provide very similar results, matching on 453 out of 472 (or 95.97%) of the auctions. In cases where the results do not match, visual inspection shows that the models selected by each metric provide reasonably good results with a slight preference towards WSSER. We therefore use WSSER in the following.

Figure 5.8 shows live bids and fitted price curves for four different auctions. The top left is best fit with exponential growth, the top right with logarithmic growth, the bottom left with logistic growth, and the bottom right with reflected-

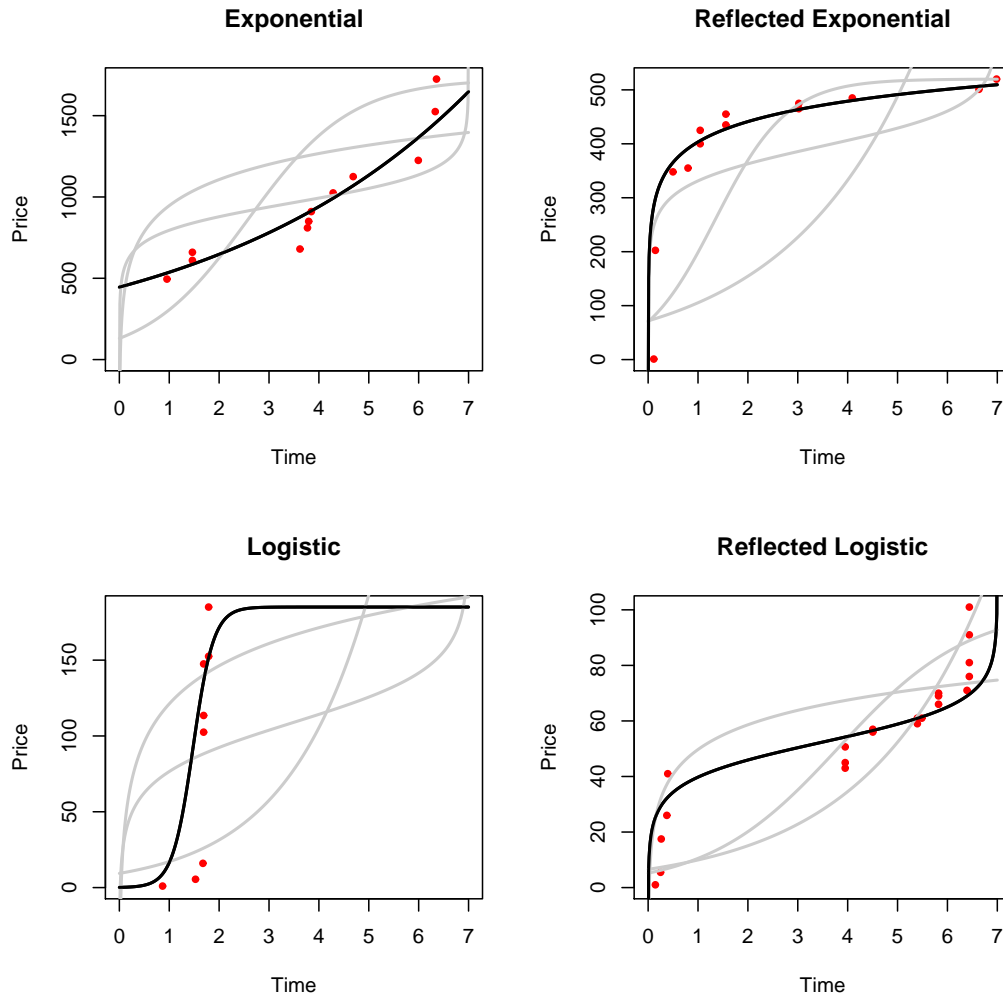


Figure 5.8: Live bids and fitted price curves for four different auctions. The top left is best fit with exponential growth, the top right with logarithmic growth, the bottom left with logistic growth, and the bottom right with reflected-logistic growth.

logistic growth. For model comparison purposes, the selected model is drawn in black and the three models that are not selected (but fitted) are drawn in grey.

Figure 5.9 provides the distribution of auctions across the four models. As expected, and in accordance with previous empirical evidence, exponential growth best fits the majority of the auctions. The next most popular model is reflected-logistic growth, which captures the common phenomena of early and late bidding.

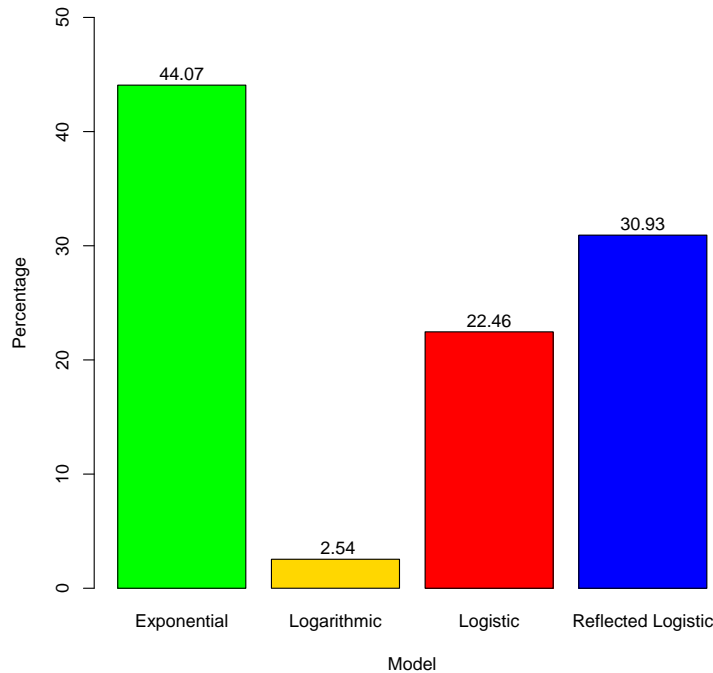


Figure 5.9: Distribution of selected model for 472 luxury wristwatch auctions.

Logistic growth curves are also selected in many cases. In contrast, logarithmic growth is rarely chosen (2.54% of the auctions). This is most likely because in this set of auctions, early high bidders were rare.

5.6 Smoothing Method Comparison

To assess the differences between our proposed growth models and the non-parametric smoothing methods (p-splines and monotone smoothing splines), we compare them on several dimensions:

Nature of fitted curves - How well does the fitted curve capture the main features of the underlying price process? Specifically, are the estimated curves monotone?

Data fitted - Which data are used for fitting the curve? Can the actual bid data be used, or do we have to sample from the “actual price” step function? In addition, what type of auctions, in terms of the number of bids, can be fit?

Overall fit - How well does the curve fit the actual bid data?

Parsimony - The level of complication in terms of number of parameters

Explanation - The level of informativeness of the fitting mechanisms (model driven vs. data driven)

Automation and computational considerations - The level of user input that is needed for curve fitting and the ability to automate the process. In addition, considerations of computational complexity and run time (especially when considering large auction data sets).

Relationship - Is there a linear relationship?

Table 5.2 summarizes the comparison of parametric and nonparametric curve fitting on all these dimensions.

In terms of fitted curves, growth models have the advantages of fitting monotone curves, fitting directly to the “live bids” (in some cases with the addition of the price at the start and end of the auction), and fitting any number of bids, including single-bid auctions (using the additional start and end prices). The resulting curves fit a variety of bid histories and capture the main features of the price dynamics during the auction. In comparison, nonparametric curves are not fit directly to the bid data but rather to the derived step function that conveys the price seen during

Table 5.2: Comparison of parametric growth models and nonparametric smoothing.

| | Growth | P-splines | Monotone |
|------------------|----------------------------------|---------------------------|---------------------------|
| Nature of curves | monotone | nonmonotone | monotone |
| Data fitted | bid data | step function | step function |
| Overall fit | good | good | variable |
| Parsimony | model type + 2 parameters | many parameters | many parameters |
| Explanation | available | unavailable | unavailable |
| Automation | no user input | user specifies parameters | user specifies parameters |
| Computation | fast | fast | slow |
| Relationship | nonlinear | linear | nonlinear |

Table 5.3: Distribution of chosen price model for 472 completed 7-day luxury wrist watch auctions.

| | Growth | P-splines | Monotone |
|----------------|---------------|-----------|----------|
| Percent chosen | 25.42% | 69.49% | 5.08% |
| Percent chosen | 88.35% | NA | 11.65% |

the auction. Although monotone splines produce monotone curves, p-splines do not guarantee such monotonicity. In fact, there is a balance between monotonicity and data fit, such that a large smoothing parameter might create monotone curves but create larger deviations between the curve and the data points and vice versa. The wiggleness of the p-splines can be seen in several of the auctions in Figure 5.10. With respect to the minimal number of bids needed for fitting, monotone splines can only be used to fit auctions with at least two bids. P-splines can be fit to single-bid auctions, but the result will be a wiggly horizontal line.

When comparing goodness-of-fit of the curve to the bid data, growth models appear to provide a good fit without overfitting. To compare goodness-of-fit, we fit each of the three methods (growth models, p-splines, and monotone splines) to each of the 472 auctions in the luxury wristwatch data set. Using the WSSE metric, we find that p-splines provide the best fit roughly 70% of the time, growth models are selected 25% of the time, and monotone smoothing only 5% of the time (Table 5.3). However, nearly 90% of the curves fit by p-splines are not monotone (which is verified by our sample in Figure 5.10). When comparing only monotone smoothing and growth models, we find that growth models are selected 88% of the time.

With respect to parsimony and explanatory power, the parametric growth models have a clear advantage: they include only two parameters, and the family of

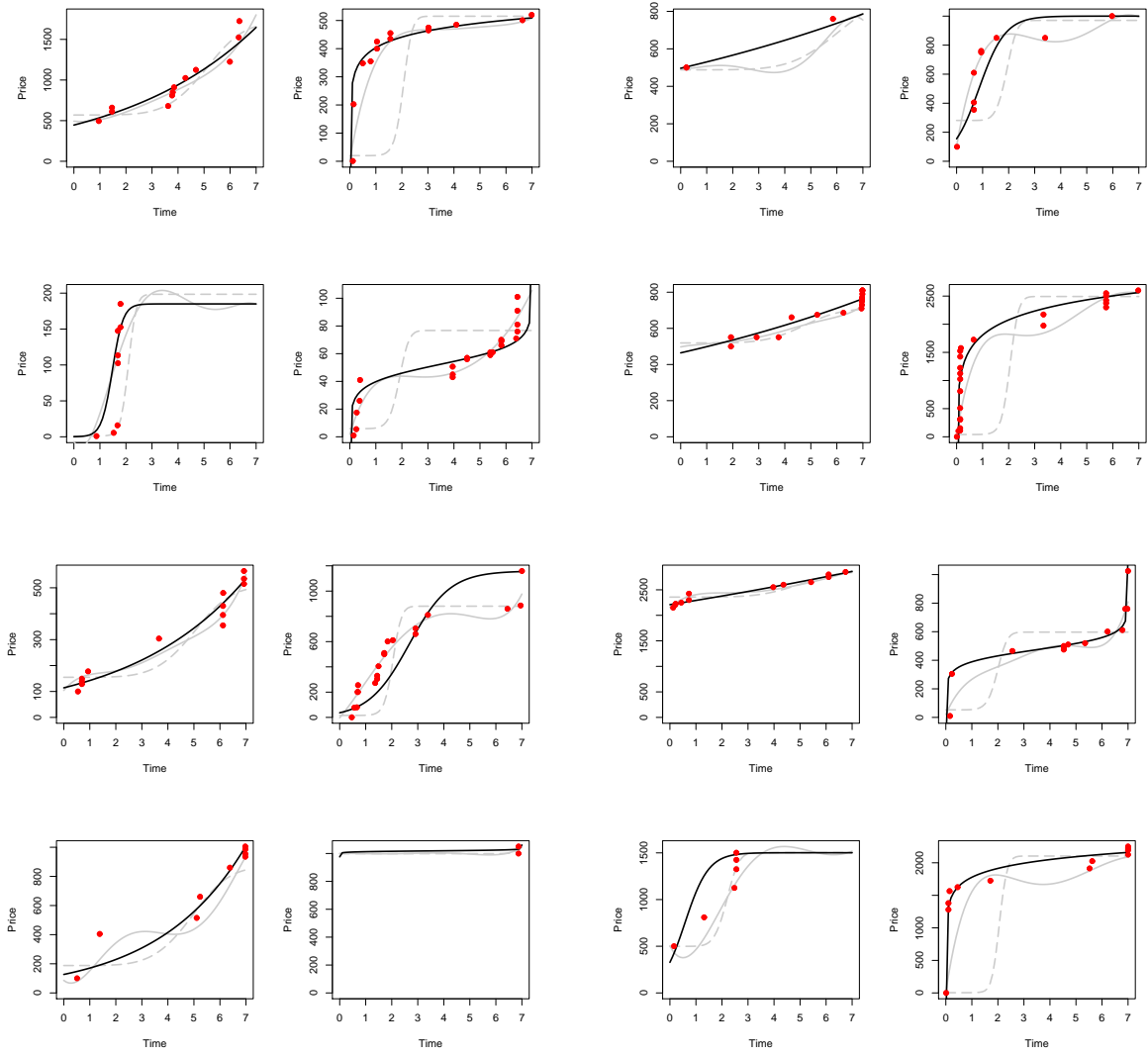


Figure 5.10: Live bids and smoothed price curves for randomly selected 7 day luxury wristwatch auctions. They black line is fit with the growth models, the solid gray line is fit with p-splines, and the gray dashed line is fit with monotone smoothing splines.

four models is able to capture a wide variety of price processes. Furthermore, growth models provide a theoretical basis that describes the price “growth” during the auction and its dynamics. Exponential models are associated with sniping, where the rate of the price increase grows faster and faster. The logistic and reflected-logistic models capture the change in price dynamics associated with early bidding. In contrast, the nonparametric methods are purely data-driven and as such do not provide a theoretical model for price growth. While they do capture the price process and its dynamics, they require a large number of parameters (the polynomial coefficients between each pair of consecutive knots, usually each such polynomial is of order 4, in order to obtain smooth curve derivatives).

From a computational point of view, fitting the growth models to data is very easy to automate and is reasonably fast, even for a large data set of auctions. The fitting can be completely automated, and we find that the results of automated fitting are satisfactory. The combination of easy automation and computation time is a major advantage over nonparametric smoothing. When fitting curves nonparametrically, the user is required to specify several parameters in advance: the number and position of knots, the order of the polynomials, and the roughness parameter. The set of knots and roughness parameter that optimally uncover the price process in one auction may not accurately capture the underlying price process of another auction. However, the same number and position of knots and roughness parameter is often used for all auctions in the data set of interest in order to avoid confounding the curve fitting from other manipulations (see Jank and Shmueli (2007)).

Table 5.4: Elapsed time (in seconds) to fit 472 and a subset of 10 luxury wristwatch auctions by p-splines, monotone splines, and growth models.

| | Growth | P-spline | Monotone |
|--------------|---------------|----------|----------|
| 10 Auctions | 4 | 2 | 75 |
| 472 Auctions | 33 | 6 | 2082 |

To evaluate computation time, we measure the elapsed time for each of the three smoothing methods on the 472 luxury wristwatch auctions as well as the first 10 auctions (Table 5.4). It is clear that for even moderate data sets, monotone splines require very long run times, whereas p-splines and growth models are much faster.

One of the disadvantages of growth models, also present in monotone smoothing, is that the model is not linear in the parameters. From an application standpoint, this means that performing linear operations on the curves (such as computing an average curve or fitting a linear regression model to the curves) must be performed on a grid. By grid, we mean that we slice time into a set of discrete time points and smooth the resulting estimates. In contrast, when using p-splines, which are linear combinations of basis functions, linear operations can be performed directly on the coefficients, and there is no need for a grid.

5.7 Using Growth Curves

One of the main advantages of modeling an auction’s price process through parametric growth models is that an initial distinction between auctions is directly obtained: each auction is represented (or, labeled) by one of exponential, logarithmic,

mic, logistic, or reflected-logistic growth. Knowing the shape of the price curve tells us about the underlying price process. This knowledge is useful in many applications, some of which will be discussed in this section.

5.7.1 Rug Plots

The *rug plot* is a visualization tool, proposed by Hyde et al. (2006), for displaying concurrent processes over a period of time. In the online auction context, the rug plot displays the entire price-evolution of all auctions in the data set over the period of data collection (calendar time). Specifically, the x-axis is calendar time, the y-axis is price, and each auction's price process is plotted as a curve. Rug plots for data sets of Palm Pilot M515 auctions and Xbox auctions (see Appendix ?? for descriptions of these data sets) are shown in the middle panels of Figure 5.11. The final price of each auction is marked with a dot, and the thick black line and grey band are the daily median closing prices and interquartile ranges (IQR), respectively. The rug plot code "GrowthModelRugPlotCode.txt" is available at <http://www.rhsmith.umd.edu/ceme/statistics>.

The rug plot supports visual exploration of temporal groupings of curves. When curves are fit nonparametrically, it is difficult to ascertain types of curves without visual inspection of each curve, which could be a daunting task for even a moderate data set. Growth models offer an easy solution, by using the WSSE measure to choose the best growth model of the four. The model type can then be easily integrated into the rug plot via color coding. To further improve the

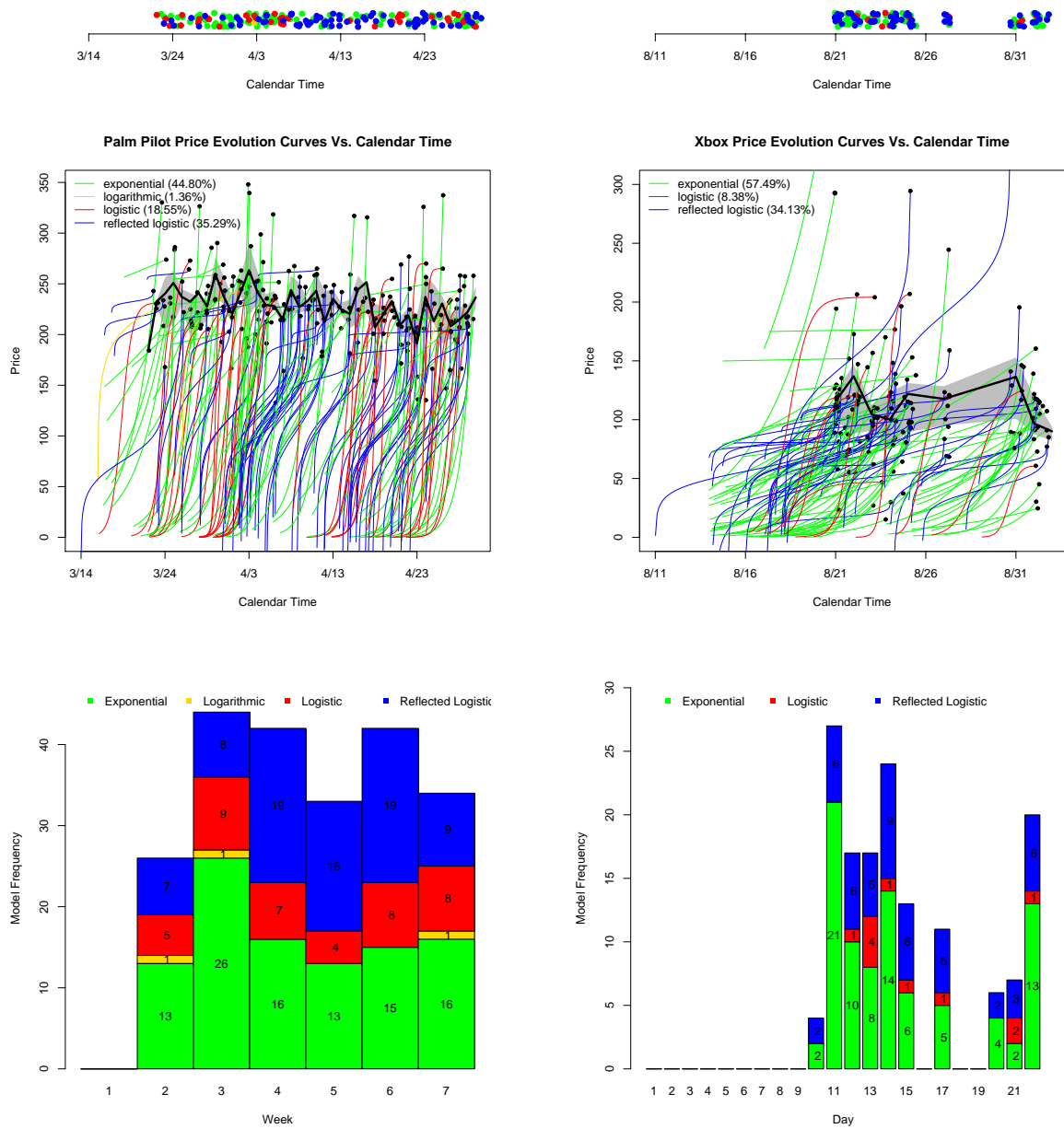


Figure 5.11: Visualizing temporal clustering of price curve types. The left column describes the Palm data; the right column describes the Xbox data. All plots are color-coded by growth model type. The top panels are dot plots (points are jittered for visibility). The middle panels are rug plots. The bottom panels are temporally-aggregated stacked bar charts of auction volume.

information contained in a rug plot, we add color coded dot plots, where a dot represents an auction that closes on that date (top panels in Figure 5.11). The dots are jittered to visualize periods where many auctions close on the same day. In addition, we create time-grouped stacked bar charts for the volume of auction closings during the time period, as can be seen in the bottom panel of Figure 5.11 (week for Palm data, day for Xbox data). Using this display we investigate temporal groupings of price processes. For the Palm data, we see a grouping of reflected-logistic price curves over April 6 - April 13 among very few exponential growth curves, whereas during other times (and especially before April 6) most price curves are exponential. Perhaps the large number of exponential growth curves before April 6 led bidders in later auctions to believe that most auctions close well over \$150, and they therefore bid early in the auction. In the Xbox data, most curves are either exponential or reflected-logistic with no logarithmic price curves. There is also a period with almost no auctions (due to data collection issues). Here we see during the beginning of the period, between August 14-18, a clustering of exponential and reflected-logistic curves, with many of the reflected-logistic auctions opening higher than the exponential auctions.

We further explore the relationship of other information, such as auction duration, to the temporal clustering of different price curves. Figure 5.12 is a set of rug plots (for the Palm data) separated by auction duration (10,7,5, and 3 days). We see a temporal clustering of reflected-logistic auctions between April 6 and April 13 in the 7-day auctions. Another observation is sporadic exponential price curves in the 5- and sometimes 7-day auctions.

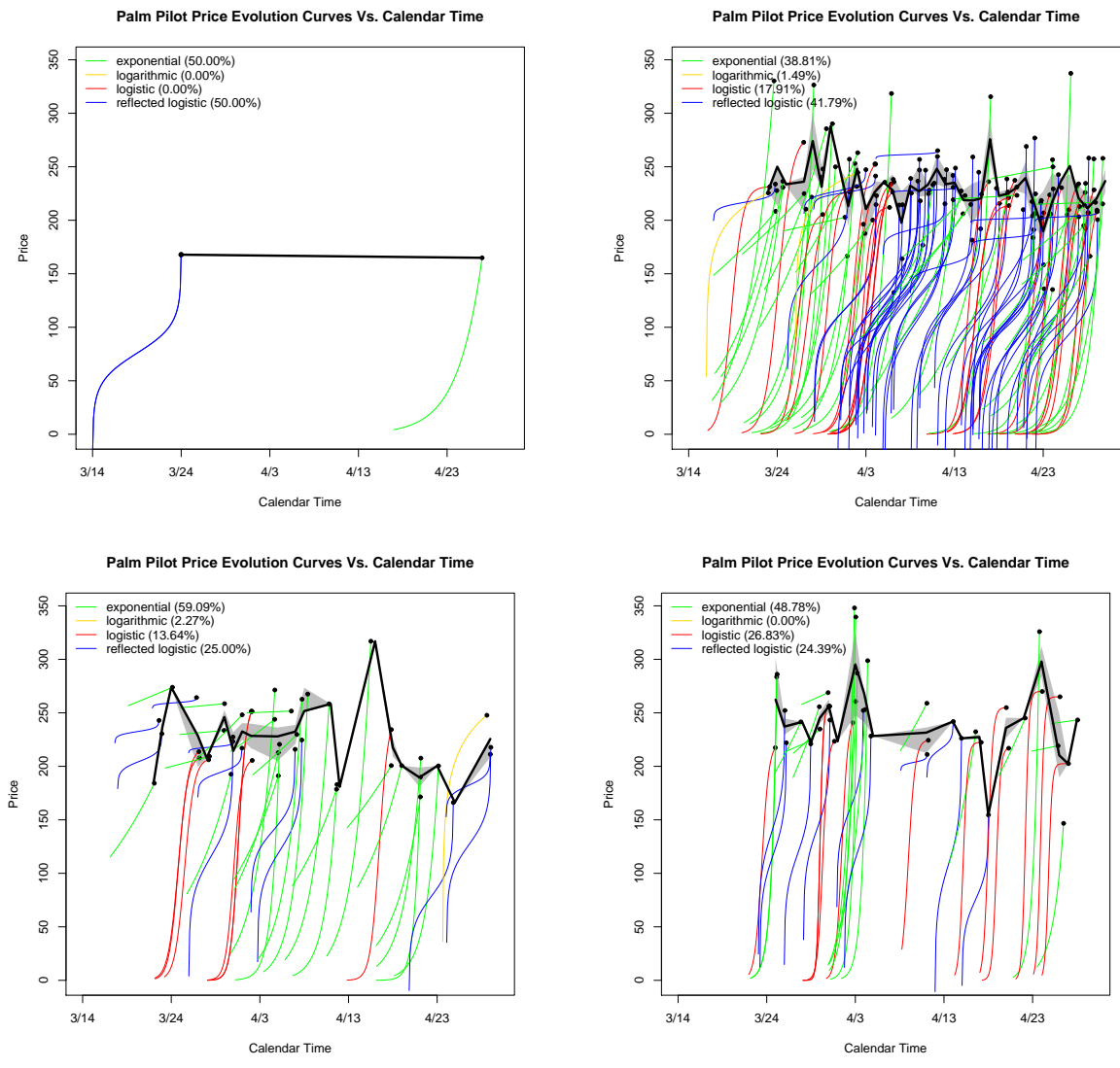


Figure 5.12: Rug plots for Palm Pilot M515 auctions broken down by length: 10-day (top left), 7-day (top right), 5-day (bottom left), and 3-day (bottom right).

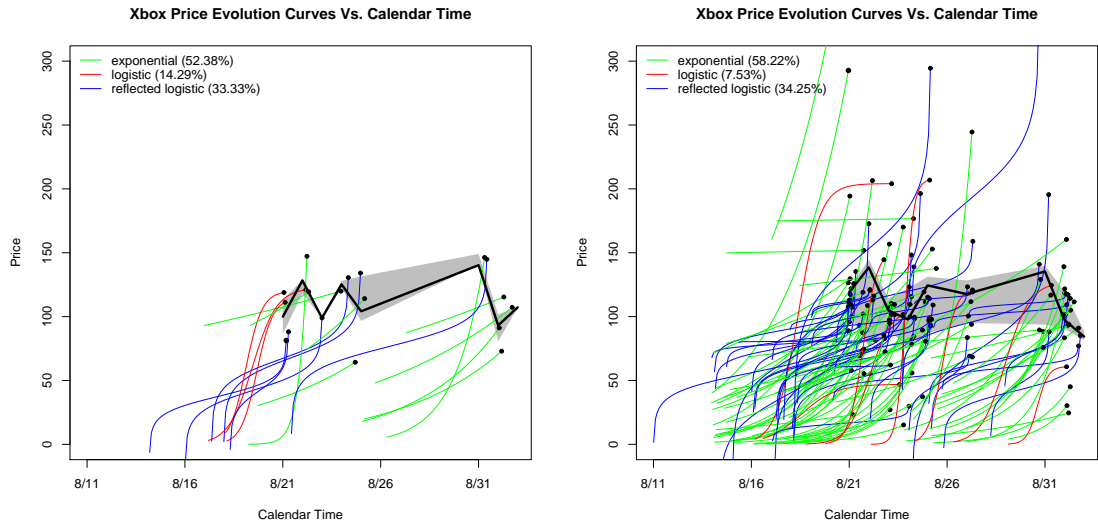


Figure 5.13: Rug plot for Xbox auctions broken down by type: “New” (left) and “Used” (right).

Another such exploration is the comparison of new vs. used Xbox game consoles (Figure 5.13). We see temporal groupings of logistic curves and reflected-logistic curves for new Xbox’s at the beginning of the calendar whereas for used Xbox’s, there does not appear to be such distinct groupings. Because there is still a large volume of auctions in the used data set, perhaps zooming-in on different calendar dates would reveal more temporal groupings.

5.7.2 Integrating Growth Model Parameters Into Analyses

The parametric growth model representation provides a compact representation of the entire price curve in an auction which is simple and parsimonious – the model type and its two estimated parameters alone. We can then integrate this compact representation into analyses by applying the statistical or data mining

method directly to this representation. This is a classic data mining approach where complicated information is summarized and the summaries are used in the analysis.

One possible application is clustering auctions using the growth model representation and perhaps additional auction related information. Another direction is in distance based methods, where the parametric representation can be used for measuring the distance between auctions. A third example is using classification trees, where the model type and the estimated parameters serve either as predictors (for predicting an outcome of interest) or as the outcome variable. In particular, such a tree could be used for predicting the type of price curve of a new auction as a function of information that is given at the auction start (e.g., opening bid, seller rating, presence of a picture, and closing day). Potential bidders can then use the predicted information in order to decide on which auction to bid, their bid timing, and bid amount.

We describe only a few possible applications here, but obviously the approach is general and parametric representation can be used in almost any type of statistical analysis and/or data mining technique.

5.8 Conclusions

This research introduces a family of growth models that describe the underlying continuous price process of online auctions. These are exponential growth, logarithmic growth, logistic growth, and reflected-logistic growth. We also present a metric to choose between models (and more generally, to choose between any type

of fitted curves), which allows automation in the data fitting stage.

Our parametric approach is parsimonious, the models are easily fitted to bid data, and they capture a variety of price process shapes. They also provide an appealing theoretical explanation of the price process rather than being purely data driven. The resulting curve is monotone, as expected for price curves in ascending auctions. Our method is computationally fast and can therefore be applied to large auction data sets. All of these reasons give the parametric approach an advantage over nonparametric smoothing methods.

Our contribution is not limited to the auction setting, but rather proposes the use of parametric functional representations as an alternative to the more popular nonparametric functional objects. We show how the parametric representation provides advantages in data visualization, as well as offers a compact summarization of the price process that can then be used in a variety of statistical analysis and data mining techniques.

One of the limitations of our family of growth models is in the case of auctions with very sparse activity throughout the auction that then changes into very steep price increases at the last moments of the auction. In this case, the exponential growth model, which provides the best fit among the four models, often fails to adequately capture the intense bid activity at the auction end. One solution is to first transform the data (e.g., by moving to log-scale). Another possibility is to heavily weight the data points towards the auction end in the fitting process. And yet another option is to include an additional growth model that describes processes that change little until a peak at the end.

There are many other functions that could potentially be used to model growth, and we provide a few examples. The Chapman-Richards growth function is similar to logarithmic growth but places a limit on growth. The Couttsian growth model is similar to exponential growth except that the growth rate is variable. Different models are popular in different disciplines such as biology, ecology, economics, etc. to describe a variety of phenomena.

We believe that parametric functional representations enhance the field of FDA and provide additional information for statistical analysis. We hope to spur interest in using theoretically relevant parametric models to describe continuous processes.

Chapter 6

Future Research

In the following sections, we describe future research directions for the three topics presented in this dissertation.

6.1 Semi-Continuous Transformation

In Chapter 2, we propose two transformations to “iron-out” the too-frequent values in semi-continuous data in order to obtain a continuous, and sometimes parametric, distribution. The too-frequent values that are transformed are selected based on visual inspection of the max-bin histogram of the data. We select those values that appear to be too-frequent such that they appear unusually large compared to other values in a surrounding small neighborhood. In the simulated data, where we specify the too-frequent values in advance, only those values in high density areas are found to be too-frequent based on visual inspection. We also find values that appear too-frequent even though they are simply a product of the particular sample, not the contamination mechanism itself.

There is a need for an automated method to select too-frequent values without visual inspection. While the method may not be able to account for too-frequent values that are a consequence of the particular sample, it will clearly be more structured and not biased by the researcher’s visual perception of what constitutes a

“large” frequency. The criteria for selecting too-frequent values needs to compare frequencies locally rather than to a global measure of largeness. This way, values in a low density area will be compared to their neighbors rather than frequencies in a high density area. Further, the size of the local comparison neighborhood must be considered and will be data dependent.

Another research contribution is to produce a better notion of what it means to be “continuous” with respect to data that is observed. We present the nonparametric measure SADBNF (Equation 2.4), a global measure of local frequency differences, but perhaps other measures would be more accurate.

Finally, we would like to implement our method on other semi-continuous data sets to illustrate the importance and usefulness of our transformations.

6.2 Visualizing Concurrency

In Chapter 4, we present several visual displays to learn how concurrency affects not only the final price of an auction but also the relationship between the current bid levels and high bids in simultaneous auctions. These displays may be used to study any overlapping continuous events, not just auction prices. Perhaps the most important next step is the need to model concurrency. Traditional ordinary least squares (OLS), which makes the assumption of independence between observations, is clearly is not appropriate in this situation. Perhaps general linear modeling can be employed to deal with the correlated variance-covariance matrix. The model should also take into account the temporal proximity between neighbor-

ing auctions such that the influence of nearby auctions are weighted more heavily than those further away.

Another extension of our research is in the area of interactive rug plots, perhaps as an additional tool for the innovative interactive AuctionExplorer (Shmueli et al., 2006). AuctionExplorer, an extension of TimeSearcher which interactively displays time series data (Aris et al., 2005; Jank et al., 2007), is a suite of tools for exploring databases of online auctions. There are tools for collecting, processing, and interactively exploring auction attributes. Ideally, the interactive rug plot would incorporate many of the functionalities that we have already discussed: the ability to zoom-in on different time domains, color-code price curves based on various static attributes (such as auction duration, seller rating, etc.), and break down the rug plots by different attributes on the same screen for visual comparison. Such an interactive tool would allow users to easily look for trends in the data as well as answer specific research questions.

6.3 Growth Models

In Chapter 5, we introduce a family of growth models for representing the underlying price process of eBay auctions. There are many ways to expand upon this area of research.

First, we find that all four growth models are manifested in the wristwatch and Palm Pilot data, but logarithmic growth is absent from the Xbox data. While it is possible that the lack of logarithmic growth is a product of this sample or

something special about Xbox auctions in general, it is also possible that the price process of online auctions has changed over time. The wristwatch data, from 2001, is the oldest, and logarithmic growth best describes the price process of 2.54% of the auctions. The Palm data, from 2003, best describes the price processes of 1.36% of the auctions with logarithmic growth. Perhaps bidder strategy, which is reflected in an auction's price process, has evolved over time. A logarithmic growth price process is hypothesized to occur when an inexperienced bidder places a high bid early in the auction, and the price levels off approaching the end of the auction due to a market value. Perhaps eBay bidders have learned over time not to reveal their willingness to pay so early in the auction. To test the hypothesis that bidder strategy has evolved over time, we would compare the distribution of price processes for different time periods (preferably holding the item being auctioned constant) using, for example, a chi-square test for independence or a Mantel-Haenszel chi-square test for trend. Similarly, we may test the hypotheses of independence of price process distribution across items, such as those with a known market value versus antiques where there may be imperfect price information.

While one research avenue is to compare the distribution of auction price processes over time, another research experiment would be to examine whether bidders in general change their style over time. Bapna et al. (2004a) profile bidders in 1999 and 2000 and find five distinct types, discussed in Section 1.3; however, the groups change slightly over the two years. In 1999, there are clear "early" and "middle" evaluator groups, whereas in 2000, the groups merge into the single class "evaluator". In 1999, there are no agent bidders, but in 2000, there is a class

of agent bidders. Bidder style could be tracked over time to detect any changes in the distribution of bidder strategies, and possibly, new strategies will emerge. Another possible extension is to examine individual bidders (with bidder level data) to investigate whether they change their bidding strategy over time, and if so, how long it takes the average bidder to transition into a style that yields more auction wins or higher surplus.

Another extension of this research is to search for additional theoretically relevant growth models. This is also essential if we find that bidding strategy changes over time, which affects price processes. We hypothesize that auctions with experienced bidders often have an exponential price process. Perhaps over time there will be so many experienced bidders that the spike in price only occurs in the final moments of the auction. This may require a different growth curve that has little growth during most of the auction then a distinct spike at the end. We suggest in Section 5.8 that perhaps weighting the end of the auction heavily would help. However, additional growth models may be useful. For example, the Couttsian growth model is similar to exponential growth except that the growth rate is variable. For price processes that are best fit exponentially, perhaps the Couttsian model with a small growth rate for most of the auction and a large growth rate in the last moments would provide a more accurate representation of the price process.

In this research, we fit exponential and logistic models in the price dimension but fit logarithmic and reflected-logistic models in the time dimension. For simplicity and computational efficiency, OLS fitting is performed in the dimension that is linearizable. However, both metrics (WSSER and WSSEV) for selecting the best

growth model evaluate fit in both the price and time dimensions simultaneously to avoid overselection of models fit in a certain dimension. We reason that for online auctions, both the time and magnitude of the bid are random variables, so the fit in both dimensions is important. This suggests that our fitting method should also be based on both dimensions simultaneously. Since we can not linearize the growth models in both dimensions, we could employ other optimization techniques, such as steepest ascent or Newton-Raphson, to estimate parameters. One of the reasons we were initially hesitant to employ optimization methods was the simplicity and computational speed of our method. Further, no parameters needed to be set in advance, as is required of the nonparametric smoothing methods. If we employ iterative fitting in both dimensions simultaneously, we may use the estimates obtained via one dimensional OLS as the starting values. This line of research should be expanded and a comparison of parameter estimates as well as computational complexity should be considered.

Finally, we encourage other researchers to adopt parametric families to describe continuous processes in their research. This is not limited to growth models. Rather, parametric models may describe decay or another continuous process. Parametric models are parsimonious, computationally fast, and provide a theoretical explanation. Further, the model that fits the price process best is itself an auction attribute, and it compactly describes the entire price-evolution. We believe that parametric families of models will become a popular method for capturing the underlying continuous process in many applications and will be a solid basis of FDA.

Appendix A

eBay Bid Level Data Sets

A.1 Luxury Wristwatch Data

Our data contain information on 472 completed 7-day luxury wristwatch (375 Rolex and 97 Cartier) auctions on eBay.com that took place between September 15, 2001 and October 27, 2001. Our sample includes a variety of items in terms of make and model, new and used, and closing price. The average selling price for all the auctions is \$2019.00, with a median of \$1300.00, and standard deviation of \$2561.89. The range of opening prices is \$0.01 to \$6,500.00, closing price is \$70.00 to \$24,000.00, number of bids is 2 to 57. The average seller experience is 571.99 with standard deviation of 1505.94, and the average (unique) bidder experience is 64.62 with standard deviation 179.14. Descriptive statistics are provided in Table A.1.

A.2 Palm Pilot M515 Data

Our data contain information on 221 closed auctions for a brand new Palm Pilot M515. The data were collected between March 11, 2003 and April 20, 2003, roughly a year after the Palm M515 was released to the market. At that time, eBay auctions lasted 3, 5, 7, or 10 days, depending on the length set by the seller. More

Table A.1: Descriptive statistics for 472 completed eBay wristwatch auctions.

| Variable | Mean(Std) | Median | Minimum | Maximum |
|--------------------------|-----------------------|-----------|---------|-------------|
| Closing Price | \$2019.00(\$2,561.89) | \$1300.00 | \$70.00 | \$24,000.00 |
| Opening Price | \$509.70(\$896.96) | \$100.00 | \$0.01 | \$6,500.00 |
| Number of Bids | 14.65(9.61) | 13.00 | 2.00 | 57.00 |
| Number of Unique Bidders | 7.38(4.05) | 7.00 | 2.00 | 21.00 |
| Unique Bidders Rating | 64.62(179.14) | 11.00 | -4.00 | 2648.00 |
| Seller Rating | 571.99(1,505.94) | 107.00 | -2.00 | 9055.00 |

recently, 1 day auctions were introduced on eBay.

Even though the Palm Pilot has a known market value (\$250.00 at the time of this analysis), auctions do not always close near this value. Low prices can result, for instance, if the box is already open or if the seller has a questionable reputation. Auctions can close high if something special is offered with the product such as an accessory or free shipping. Prices also vary because bidders often get caught up in the excitement of bidding (“auction fever”) and pay more than would be expected. The average selling price for all Palm Pilots in our data is \$234.00 with a median of \$232.50 and standard deviation of \$20.86. The least expensive Palm Pilot sells for \$172.50, and the most expensive auction closes at \$290.00. Table A.2 provides descriptive statistics for the closing price, opening price, number of bids, number of unique bidders, and unique bidder rating. We also provide descriptive statistics broken down by auction length since we group Palm Pilot auctions by length in Sections 4.2.1 and 5.7. Note that this data set does not contain any seller information.

A.3 Xbox Data

Our data contain information on 167 closed auctions for an Xbox game console. The auctions took place between August 11, 2005 and August 30, 2005 and are of fixed duration: 1, 3, 5, 7, or 10 days. While the Xbox product is the same, it may be “new” or “used”, and the auction may include extras such as additional games and/or controllers. Therefore, bidders will not have the same valuation for each

Table A.2: Descriptive statistics for 221 completed eBay Palm Pilot M515 auctions and broken down by auction length: 3-day (41), 5-day (44), 7-day (134), and 10-day (2) auctions.

| Variable | Duration | Mean(Std) | Median | Minimum | Maximum |
|--------------------------|----------|-------------------|----------|----------|----------|
| Closing Price | 3 Day | \$238.60(\$24.67) | \$232.50 | \$177.50 | \$290.00 |
| | 5 Day | \$233.40(\$23.08) | \$235.00 | \$183.50 | \$280.00 |
| | 7 Day | \$233.60(\$17.78) | \$232.80 | \$186.50 | \$283.50 |
| | 10 Day | \$182.50(\$14.14) | \$182.50 | \$172.50 | \$192.50 |
| | Total | \$234.00(\$20.86) | \$232.50 | \$172.50 | \$290.00 |
| Opening Price | 3 Day | \$63.46(\$94.87) | \$1.00 | \$0.01 | \$259.00 |
| | 5 Day | \$91.09(\$97.36) | \$35.50 | \$0.01 | \$259.00 |
| | 7 Day | \$41.69(\$72.47) | \$1.00 | \$0.01 | \$259.00 |
| | 10 Day | \$2.51(\$3.53) | \$2.51 | \$0.01 | \$5.00 |
| | Total | \$63.70(\$84.05) | \$1.00 | \$0.01 | \$259.00 |
| Number of Bids | 3 Day | 17.51(9.85) | 19.00 | 2.00 | 43.00 |
| | 5 Day | 16.18(9.18) | 17.50 | 2.00 | 36.00 |
| | 7 Day | 20.86(10.25) | 21.00 | 2.00 | 51.00 |
| | 10 Day | 20.50(6.36) | 20.50 | 16.00 | 25.00 |
| | Total | 19.33(10.09) | 19.00 | 2.00 | 51.00 |
| Number of Unique Bidders | 3 Day | 9.22(5.05) | 9.00 | 2.00 | 23.00 |
| | 5 Day | 8.25(4.22) | 9.00 | 2.00 | 19.00 |
| | 7 Day | 10.85(4.71) | 11.50 | 1.00 | 23.00 |
| | 10 Day | 10.50(3.54) | 10.50 | 8.00 | 13.00 |
| | Total | 10.03(4.77) | 10.00 | 1.00 | 23.00 |
| Unique Bidders Rating | 3 Day | 91.14(68.43) | 86.00 | 3.00 | 204.00 |
| | 5 Day | 89.93(69.01) | 84.00 | 1.00 | 217.00 |
| | 7 Day | 84.56(70.29) | 74.00 | 1.00 | 217.00 |
| | 10 Day | 73.81(57.02) | 74.00 | 3.00 | 167.00 |
| | Total | 86.46(69.68) | 74.00 | 1.00 | 217.00 |

auction. Descriptive statistics for all the auctions as well as broken down by item condition (new or used) are shown in Table A.3.

This game console is no longer sold in stores (as it is the predecessor of the Xbox 360); however, Amazon.com's list price was \$179.98 at the time the data were collected. The average selling price in our sample is \$132.40 with a standard deviation of \$62.59, median of \$125.00, minimum of \$28.00, and maximum of \$501.80. The auction that closed at \$28.00 is for a damaged console, and the auction that closed at \$501.80 is used but includes 84 games.

Table A.3: Descriptive statistics for 167 completed eBay Xbox auctions and broken down by condition: “new” (21) and “used” (146).

| Variable | Condition | Mean(Std) | Median | Minimum | Maximum |
|--------------------------|-----------|-------------------|----------|---------|----------|
| Closing Price | New | \$121.00(\$14.21) | \$123.20 | \$85.00 | \$142.50 |
| | Used | \$134.10(\$66.60) | \$125.50 | \$28.00 | \$501.80 |
| | Total | \$132.40(\$62.59) | \$125.00 | \$28.00 | \$501.80 |
| Opening Price | New | \$26.19(\$37.33) | \$1.00 | \$0.01 | \$99.99 |
| | Used | \$40.84(\$43.40) | \$32.49 | \$0.01 | \$290.00 |
| | Total | \$39.00(\$42.86) | \$25.00 | \$0.01 | \$290.00 |
| Number of Bids | New | 20.95(8.81) | 22.00 | 6.00 | 38.00 |
| | Used | 18.64(11.75) | 18.00 | 2.00 | 75.00 |
| | Total | 18.93(11.43) | 18.00 | 2.00 | 75.00 |
| Number of Unique Bidders | New | 9.19(3.16) | 9.00 | 3.00 | 14.00 |
| | Used | 8.18(3.80) | 8.00 | 1.00 | 19.00 |
| | Total | 8.31(3.73) | 8.00 | 1.00 | 19.00 |
| Unique Bidders Rating | New | 234.70(340.79) | 164.00 | 5.00 | 1325.00 |
| | Used | 299.70(837.90) | 36.00 | -1.00 | 5560.00 |
| | Total | 291.50(792.29) | 44.00 | -1.00 | 5560.00 |
| Seller Rating | New | 30.16(69.63) | 5.00 | 0.00 | 605.00 |
| | Used | 42.51(166.44) | 5.00 | -1.00 | 2736.00 |
| | Total | 40.79(156.63) | 5.00 | -1.00 | 2736.00 |

Bibliography

- Alford, B. and Urimi, L. (2004). An analysis of various spline smoothing techniques for online auctions. Term Paper, AMSC Research Interaction Team. Available at <http://www.smith.umd.edu/ceme/statistics/brianlakshmi.pdf>.
- Allen, M. T. and Swisher, J. (2000). An analysis of the price formation process at a hud auction. *Journal of Real Estate Research*, 20:279–298.
- Anwar, A., McMillan, R., and Zheng, M. (2006). Bidding behavior in competing auctions: Evidence from ebay. *European Economic Review*, 50:307–322.
- Aris, A., Shneiderman, B., Plaisant, C., Shmueli, G., and Jank, W. (2005). Representing unevenly-spaced time series data for visualization and interactive exploration. In *Proceedings of the International Conference on Human Computer Interaction (INTERACT), Lecture Notes in Computer Science*, volume 3585, pages 835–846, Rome, Italy.
- Bajari, P. and Hortascu, A. (2003). Winner’s curse, reserve price and endogenous entry: Empirical insights from ebay. *RAND Journal of Economics*, 34:329–355.
- Bajari, P. and Hortascu, A. (2004). Economic insights from internet auctions. *Journal of Economic Literature*, 42:457–486.
- Bapna, R., Goes, P., Gupta, A., and Jin, Y. (2004a). User heterogeneity and its impact on electronic auction market design: An empirical exploration. *MIS Quarterly*, 28(1):21–43.
- Bapna, R., Jank, W., and Shmueli, G. (2004b). Price formation and its dynamics in online auctions. Working Paper, Smith School of Business, University of Maryland. Available at <http://ssrn.com/abstract=902887>.
- Bapna, R., Jank, W., and Shmueli, G. (2005). Consumer surplus in online auctions. Working Paper, Smith School of Business, University of Maryland. Available at <http://ssrn.com/abstract=840264>.
- Bzik, T. J. (2005). Overcoming problems associated with the statistical reporting and analysis of ultratrace data. <http://www.micromanagemagazine.com/archive/05/06/bzik.html>[08 January 2007].
- Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P. A. (1983). *Graphical Methods for Data Analysis*. Wadsworth International Group, Belmont, CA.
- Dellarocas, C. and Wood, C. (2007). The sound of silence in online feedback: Estimating trading risks in the presence of reporting bias. Under review at Management Science.

- Deltas, G. (1999). Auction size and price dynamics in sequential auctions. Working Paper, University of Illinois.
- Efromovich, S. (1997). *Nonparametric Curve Estimation*. Springer-Verlag, New York.
- Etzion, H., Pinker, E., and Seidermann, A. (2004). Analyzing the simultaneous use of auctions and posted prices for online selling. Working Paper No. CIS 03-01, Simon Business School.
- Gallien, J. (2002). Dynamic mechanism design for online commerce. Working Paper, MIT Sloan School of Management.
- Ghani, R. and Simmons, H. (2004). Predicting the end-price of online auctions. In *Proceedings of the International Workshop on Data Mining and Adaptive Modeling and Methods for Economics and Management*.
- Góes, P., Pereira, A., Rocha, L., Mourão, F., Torres, T., and Meira Jr., W. (2007). A hierarchical characterization model for online auctions. In *ICISTM 07 - First International Conference on Information Systems Technology and Management*, New Delhi, India.
- Good, I. J. and Gaskins, R. J. (1980). Density estimation and bump hunting by the penalized maximum likelihood method exemplified by scattering and meteorite data. *Journal of the American Statistical Association*, 75(369):42–56.
- Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Chapman and Hall, London.
- Guerre, E., Perrigne, I., and Vuong, Q. (2000). Optimal nonparametric estimation of first-price auctions. *Econometrica*, 68:525–574.
- Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York.
- Hyde, V. (2006). Visualizing and exploring time series data using functional data analysis (fda). AMSC Candidacy Prospectus. Available at <http://math.umd.edu/~valhyde/research>.
- Hyde, V., Jank, W., and Shmueli, G. (2006). Investigating concurrency in online auctions through visualization. *The American Statistician*, 34(3):241–250.
- Hyde, V., Jank, W., and Shmueli, G. (2007). *Statistical Methods in eCommerce Research*, chapter A Family of Growth Models for Representing the Price Process in Online Auctions. Wiley & Sons. Jank and Shmueli (Eds.). Under review.
- Hyde, V., Moore, E., and Hodge, A. (2004). Functional pca for exploring bidding activity and times for online auctions. Term Paper, AMSC Research Interaction Team. Available at http://www.smith.umd.edu/ceme/statistics/functional_pca.pdf.

- Jank, W. and Shmueli, G. (2005). Dynamic profiling of online auctions using curve clustering. Working Paper, Smith School of Business, University of Maryland. Available at <http://www.smith.umd.edu/ceme/statistics/papers.html>.
- Jank, W. and Shmueli, G. (2006). Functional data analysis in electronic commerce research. *Statistical Science*, 21(2):155–166.
- Jank, W. and Shmueli, G. (2007). *Business Computing*, chapter Studying Heterogeneity of Price Evolution in eBay Auctions via Functional Clustering. Handbook of Information Systems Series. Elsevier. Adomavicius and Gupta (Eds.). Accepted.
- Jank, W., Shmueli, G., Plaisant, C., and Shneiderman, B. (2007). *Handbook on Computational Statistics on Data Visualization*, chapter Visualizing Functional Data with an Application to eBay’s Online Auctions. Springer-Verlag, Heidelberg. Chen, Hardle and Unwin (Eds.). In press.
- Kauffman, R. J. and Wood, C. A. (2005). The effects of shilling on final bid prices in online auctions. *Electronic Commerce Research and Applications*, 4:18–31.
- Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34:1–14.
- Lucking-Reiley, D., Bryan, D., Prasad, N., and Reeves, D. (2005). Pennies from ebay: the determinants of price in online auctions. Working Paper, University of Arizona.
- McGill, R., Tukey, J. W., and Larsen, W. A. (1978). Variations of box plots. *The American Statistician*, 38:12–16.
- Perlich, C. and Rosset, S. (2006). Quantile tress for marketing. In *Proceedings of Data Mining in Business Applications Workshop, International Conference on Knowledge and Data Mining*, Philadelphia, PA.
- Ramsay, J. O. (1998). Estimating smooth monotone functions. *Journal of the Royal Statistical Society B*, 60:365–375.
- Ramsay, J. O. and Dalzell, C. J. (1991). Some tools for functional data analysis. *Journal of the Royal Statistical Society B*, 53:539–572.
- Ramsay, J. O., Munhall, K. G., Gracco, V. L., and Ostry, D. J. (1996). Functional data analysis of lip motion. *Journal of the Acoustical Society of America*, 99:3718–3727.
- Ramsay, J. O. and Silverman, B. W. (2002). *Applied Functional Data Analysis: Methods and Case Studies*. Springer-Verlag, New York.
- Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer-Verlag, New York.

- Roth, A. E. and Ockenfels, A. (2002). Last-minute bidding and the rules for ending second-price auctions: Evidence from ebay and amazon auctions on the internet. *The American Economic Review*, 92.
- Scott, D. W. (1979). On optimal and data-based histograms. *Biometrika*, 66:605–610.
- Shekhar, S., Lu, C. T., and Zhang, P. (2003). Unified approach to spatial outliers detection. *GeoInformatica*, 7(2):139–166.
- Shmueli, G. and Jank, W. (2005). Visualizing online auctions. *Journal of Computational and Graphical Statistics*, 14:299–319.
- Shmueli, G., Jank, W., Aris, A., Plaisant, C., and Shneiderman, B. (2006). Exploring auction databases through interactive visualization. *Decision Support Systems*, 42(3):1521–1538.
- Shmueli, G., Jank, W., and Hyde, V. (2007). Transformations for semi-continuous data. Working Paper, Smith School of Business, University of Maryland. Available at <http://ssrn.com/abstract=956938> Under review for Computational Statistics and Data Analysis.
- Shmueli, G., Russo, R., and Jank, W. (2004). The barista: A model for bid arrivals in online auctions. Working Paper, Smith School of Business, University of Maryland. Available at <http://ssrn.com/abstract=902868>.
- Snir, E. M. (2007). Online auctions enabling the secondary computer market. *Information Technology and Management*.
- Vakrat, Y. and Seidmann, A. (1999). Can online auctions beat online catalogs? In *Proceedings of the 20th International Conference on Information Systems (ICIS)*, Charlotte, NC.
- Wand, M. P. (1997). Data-based choice of histogram bin width. *The American Statistician*, 51(1):59–64.
- Wang, S., Jank, W., and Shmueli, G. (2007a). Dynamic forecasting of online auction price using functional data analysis. *Journal of Business and Economic Statistics*. In press.
- Wang, S., Jank, W., Shmueli, G., and Smith, P. (2007b). Modeling price dynamics in ebay auctions using principal differential analysis. Working Paper, Smith School of Business, University of Maryland.
- Zeithammer, R. (2006). Forward-looking bidding in online auctions. *Journal of Market Research*, 43(3):462–476.