Journal of Statistical Software
February 2011, Volume 39, Book Review 1. http://www.jstatsoft.org/

Reviewer: Nicolas Christou
University of California, Los Angeles

Statistical Methods in e-Commerce Research

Wolfgang Jank, Galit Shmueli
John Wiley & Sons, Hoboken, NJ, 2008.
ISBN 978-0-470-1202-5. 430 pp. USD 105.00.
http://www.wiley.com/WileyCDA/WileyTitle/productCd-0470120126.html

Electronic commerce (e-commerce) has become part of our everyday lives. Using the internet (clicking, buying, selling, rating) generates a vast amount of data. These data can be recorded and stored and can provide important information that it is not available in the offline world. The field of statistics can play a major role in the development of methods for empirical research related to electronic commerce.

The idea for writing this book started at a conference held at the University of Maryland in May 2005. The title of the conference was "Statistical Challenges and Opportunities in Electronic Commerce Research". In this book more than 30 researchers present very interesting ideas and challenges that arise from the substantial amount of data collected from today's online buying, selling, advertising, investing, etc. Today, there is a need for statisticians to collect, clean, and make sense of the massive amount of data generated on the internet. This book is unique in that it attempts to narrow the gap between statistics and e-commerce. To achieve this, the authors present the challenges that arise in e-commerce data, propose novel statistical techniques to tackle the problem, and suggest future research in this field. This book does an excellent job in addressing these goals and highlighting the ability of statistics to facilitate e-commerce research. Additionally, it can be used not only by researchers, but also by instructors for advanced undergraduate and graduate courses. I believe this book can be an invaluable tool in educating statisticians and computer scientists in the many areas of research that e-commerce has to offer.

The book consists of three sections. Five articles from the first section provide an overview of the e-commerce research challenges. One can understand the importance of statistics in e-commerce applications by understanding the offline world when using data from the online world. By using online data we can understand consumer behavior by answering questions on economic, marketing, and information systems issues. The digital divide in the United States (and around the world) presents another potential research problem. Despite the prevalent use of internet there are still low-income, rural, and small-town communities that have not joined the online world, and therefore the study on how the geographical location affect e-commerce is important. The privacy of data gathered online is another issue that requires

new computational and statistical technologies to protect online data from the invasion of databases by commercial and other queries. The last chapter of the first section discusses the enormous data and research problems created on Wikipedia.

The second section of the book examines some e-commerce applications. The topics covered include price dynamics using online auction data, web usability, internet firms' performance and survival, changing of regression coefficients over time by examining pooled cross-sectional data, and optimization of search engine marketing bidding strategies.

The last section presents new methods for e-commerce data. The methods described are clustering e-commerce data into groups with data points in each group similar to each other; functional data analysis using online auction data; price process in online auctions using growth models such as exponential, logarithmic, logistic, and reflected-logistic models, and models describing bidders arrival and departure in online auctions. In particular, I find the discussion of spatial models in e-commerce to be interesting. Despite the belief that online markets are geography-independent, spatial models can dramatically increase the performance of models that explore the spatial correlation exhibited in the data. The authors use data from online mortgage leads and compare the performance of three models: a non-spatial model, a static spatial model, and a dynamic spatial model (the static model uses all the data at once compared to the dynamic model which updates its parameters as consumers join an online business over time). The spatial models easily outperform the classical regression model, with the dynamic spatial model giving the best predictions. The last three chapters of this section deal with differential equations to model the price dynamics in online auctions, understanding the purchasing power of consumers (also called wallet estimation), and use of randomized response in e-commerce.

There is no doubt that internet-based data applications will continue to grow and cover many areas of the science spectrum. For example, during the 2008 swine flu epidemic, Ginsberg *et al.* (2009) at Google Inc. were able to forecast on real time the percentage of influenza-like illness symptoms by collecting Google search results for "flu symptoms, influenza, flu treatments, etc." Google's forecast was two weeks ahead compared to actual data from the Center for Disease Control (CDC).

Interesting articles about e-commerce applications also appear on various websites such as `http://radar.oreilly.com/`, e.g., Loukides (2010). Today's data scientists must be able to use statistics and computer science to make e-commerce data tell their story. UC Berkeley's professor Hal Varian's quote "The ability to take data – to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it – that's going to be a hugely important skill in the next decades," (see McKinsey & Company 2009) summarizes nicely the importance for the development of new tools to tackle the problem of internet-based data.

In closing, I believe this book is a great reference for statisticians who are interested in analyzing data from online transactions and it should be consulted for its useful ideas and proposed research topics.

References

Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L (2009). "Detecting Influenza Epidemics Using Search Engine Query Data." *Nature*, **457**, 1012–1014.

Loukides M (2010). "What is Data Science?" *O'Reilly Radar*. URL http://radar.oreilly.com/2010/06/what-is-data-science.html.

McKinsey & Company (2009). "Hal Varian on How the Web Challenges Managers." *The McKinsey Quarterly*. URL http://www.mckinseyquarterly.com/Hal_Varian_on_how_the_Web_challenges_managers_2286.

Reviewer:

Nicolas Christou
University of California, Los Angeles
Department of Statistics
8125 Math Sciences Bldg.
Los Angeles, CA 90095, United States of America
E-mail: nchristo@stat.ucla.edu
http://www.stat.ucla.edu/~nchristo/

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 39, Book Review 1 *Published:* 2011-03-01
February 2011