Expanded Research Statement

Methodologies in Database Development and Causal Inference

Answering new research questions often demand developing novel databases that may take years of efforts, as well as proper analysis of these databases. I started this line of work at my time of pursuing the master's degree in statistics under the supervision of Prof. Donald Rubin. While compiling and analyzing a variety of large databases, I had to confront data limitations due to *endogeneity, selection bias* and *completely missing key variables.* These issues are common in the context of intellectual property, counterfeiting and innovation, data-based brand management and marketing, as well as big-data business analytics. This led me to apply and invent econometric and statistical methodologies to address these common and important issues in database development and data analytics in order to draw valid causal inferences. A central theme of this work is to unify different kinds of problems under the framework of incomplete-data theory by viewing available data as incomplete forms of ideal data, and to develop robust, automated, or scalable methods that (a) develop databases that approach ideal data and (b) reduce or detect the dependence of empirical findings on model assumptions. Consequently, models or methods developed to address a specific kind of problem can have a broader range of applications in quantitative economics, database marketing, and business analytics. My research in this stream spans three intertwined areas: (1) treatment effect evaluation methods (for handling *endogeneity*); (2) missing-data methods for data-based brand management and marketing (for handling *selection bias*); and (3) methods for handling *completely missing key variables*. Given the importance of database development and causal inference in empirical work, this line of work has the potential to impact a range of fields.

a Treatment Effect Evaluation

In this line of work I strive to apply rigorous treatment effect evaluation methods that can handle missing counterfactual outcomes, and to build on identification techniques such as propensity score methods, instrumental variables, and differences-in-differences (DID). When needed, I develop new modeling approaches that permit a more refined analysis of treatment effects that are heterogeneous in various dimensions [20].[1] A fundamental issue in treatment effect evaluation is the inability to observe all counterfactuals, as presented in Figure 1, where X represents a set of covariates, Y_0 (Y_1) denotes the potential outcome of interest under the control (treatment) condition, and shaded (white) areas represent observed (missing) data. For any unit, only one of the two counterfactual outcomes is observed. In nonrandomized observational studies, the observed counterfactual outcome values are often not representative of those from the entire sample. Consequently, the simple comparison of the observed counterfactual outcomes between treatment and control samples is a biased treatment effect estimate. An approach to mimic the gold-standard randomized experiments to form comparable samples for unbiased treatment effect estimation is the use of propensity score methods, which can overcome "the curse of dimensionality" problem in balancing a large number of covariates in X and reduce the sensitivity of empirical results to the functional form extrapolation of simple regression analysis.

I have worked on designing and applying suitable propensity score matched sampling procedures (combined with the panel DID methods) to establish comparable control samples for proper comparison to evaluate the effects of patent protection on pharmaceutical innovations in cross-country panel data analysis from 1978 to 2002 [1], and to evaluate the multichannel spillover effects of opening a factory store using comprehensive panel consumer transactional data from 1994 to 2007 [21]. Matched

[1] Please refer to my curriculum vitae for a numbered reference list of my papers.

sampling in combination with panel DID analyses can tease out the endogenous part of the treatment variable by appropriately controlling for the set of observable confounding covariates and unobserved unit-specific heterogeneity, hence arriving at a reasonable causal inference (Rubin and Thomas 2000). I started the pursuit of causal inference methodology in my very first paper, “Do National Patent Laws Stimulate Domestic Innovation in a Global Patenting Environment?”[1], where I employ a two-stage Mahalanobis matching method to establish comparable country pairs in the face of missing data. I accomplish this by matching in two passes (Appendix III in the paper). Comparing the t-statistics of the covariates before and after matching clearly shows that the covariates are much more balanced after matching. When the countries in the treatment and control groups are comparable in all other characteristics except national patent laws, the differences in their innovation outcomes can be more comfortably attributed to the patent implementation treatment. Having the two control groups is a particularly useful check in that one would expect any remaining bias to go in opposite directions when the two control groups are used as comparators: up, when the control is the never-had-patents group; down, when it is the always-had-patents group. A referee commends it as “a very nice job of applying cutting-edge econometric methodology to a very interesting and important question.”

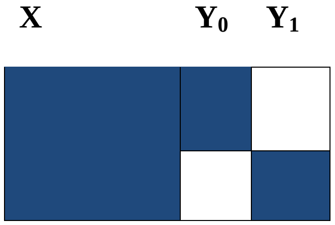

Fig. 1: Treatment Effect Evaluation

An alternative method to draw causal inference from observational studies is to bring in exogenous shocks that give rise to treatment variation by finding appropriate instrumental variables. I provide a set of valid instruments for different aspects of the entry effects of counterfeiting in “Impacts of Entry by Counterfeiters” [2]. In particular, I exploit the plausibly exogenous loosening of government enforcement on footwear trademarks and the different relationship between each brand and the government to identify entry effects of counterfeits. This resembles a difference-in-difference idea in that brands that have a close relationship with the government were less adversely affected by the unexpected reallocation of government enforcement resources and hence faced fewer counterfeit threats after the policy change.

In the next study, “Investigating the Dynamic Effects of Counterfeits with a Hierarchical Random Changepoint Simultaneous Equation Model” [20], I develop a new modeling approach that allows the treatment effect to have *discontinuous* changes at firm-specific unknown change-points (structural breaks), to be *heterogeneous* across firms, and to have *regime-switching* moderating effects. Unlike the standard econometric techniques employed [2], which addressed the research question of the average treatment effects of counterfeit entry, the new modeling approach permits dynamic and heterogeneous treatment effects that are crucial for identifying drivers of inter-firm differences in their response behaviors. The results enhance understanding of firms’ responses to counterfeit entry. Analysis demonstrates that the new model matches the underlying process better than the traditional hierarchical Bayesian (HB) models that do not model these random change-points. Compared with traditional HB models, the new model provides larger estimates for both short-term and long-term effects of counterfeit entry. It considerably improves the power to detect moderating effects, some of which cannot be detected in traditional HB models. The new hierarchical dynamic-effect analysis reveals new findings that (1) pre-entry product quality *moderates* the short-term price competition effects of counterfeit entry and helps alleviate the harmful impact of counterfeit entry; (2) brand popularity *moderates* the long-term price increase effect, and firms that were less popular pre-entry tend to have more price increases, consistent with the hypothesis that counterfeits can serve as free advertising for their authentic counterparts; and (3) firms with more innovation, less diversification from infringed markets, or more human capital were faster in responding to and differentiating from counterfeits.

A common problem underlying observational data is unobserving important relevant variables. Omitted variable bias constitutes one of the most serious threats to causal inference. Researchers often add information by either gathering more exogenous data or imposing structural assumptions. In "Private Label Pricing: Estimating Demand with Data and Structure" [24], co-authors and I use data collected from a randomized field experiment to test the ability of the BLP model to predict counterfactual demand outcomes in the context of retailing. In addition to evaluating structural model performance against the "gold standard" of randomized treatment conditions, we attempt to bridge the structural modeling and natural experiment paradigms in marketing applications.

b Missing Data in Data-based Brand Management and Marketing

With the increasing popularity of database marketing, companies face a greater need to provide data-driven brand management and marketing, such as optimized pricing and revenue management, as well as customized marketing solutions based on consistent and precise elasticity estimates of marketing mix variables. In reality, however, marketing databases are often plagued with missing-data issues, which, as noted in Blattberg et al. (2008), have become "a fact of life for DBM [Database Marketing] applications." These issues have become even more salient with new kinds of data made available by innovations in digital technology. Although these new data-collection methods can collect large amounts of data that were previously unavailable and can more accurately reflect individual attitudes and behaviors in real life, missing-data issues can be especially severe because of the way data are collected. For example, a vast marketing literature studies brand choices and consumption behaviors of individual consumers. When consumer discrete choice models are calibrated using electronic scanner panel data, the prices and coupon values for nonpurchased brands are often missing. Such issues lead to the data pattern in Figure 2. Likewise, for intensive electronic diary data collected through the Internet or from real-time data-capture instruments such as handheld computers, smartphones, or other personal mobile devices, the panel outcomes such as consumer satisfaction ratings, lifestyle outcomes (e.g., smoking behaviors), and consumption behavior data are typically subject to missingness due to attrition and intermittent missingness with complex missing-data patterns (Fig. 3) and reasons. The simple complete-case (CC) analysis can lead to strong self-selection bias and substantial loss of estimation efficiency. In some cases, there may even be no complete case for high-dimensional data, as illustrated in Figures 2 and 3.

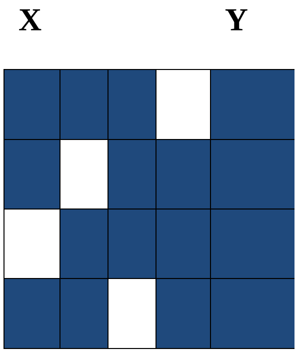

Fig. 2: Missing Covariates

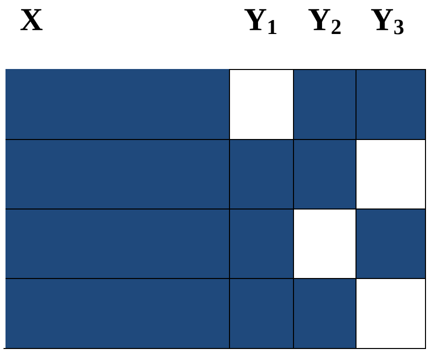

Fig. 3: Missing Outcomes

These new settings raise novel issues as to the choices of general and flexible models and fast and computationally feasible methods, and provide opportunities for developing cutting-edge analytical techniques. Inspired by these interesting and important problems, I have been working on developing a new generation of missing data procedures designed for the rich data environment in the new digital economy. As summarized below, I strive towards developing procedures that better match the underlying process obscured by missing data while still being tractable.

(1) *Methods for handling missing data in covariates.*

In the paper "No Customer Left Behind" [7], I develop a distribution-free Bayesian method to handle missing-covariate problems in marketing applications (Fig. 2). Conventional ad hoc approaches to filling in missing price and coupon values in scanner panel data ignore the dependence between choice outcomes and missing marketing mix variables and can create a strong selection bias in the estimation of brand-choice models. Prior econometric treatment of the problem by Erdem, Keane, and Sun (1999) posited that the covariates (prices and coupon values of different brands) are independent of each other, and each covariate is separately and independently modeled as a parametric polynomial function. The proposed distribution-free Bayesian approach in [7] requires neither the independent assumption nor a search for proper distributional forms. The analysis in both empirical applications and simulated datasets demonstrates the importance of properly modeling both the nonstandard distributional features of marketing mix variables and dependence among them; ignoring either feature can lead to a sizable bias in the estimation of brand-choice models. The implication goes beyond the scanner database application and renders the proposed method applicable to many other applications. One example is a purchase-incidence analysis for consumer targeting and profiling, as illustrated in the second application of the paper, where consumer characteristics used for targeting or profiling are often subject to missingness. In these analyses, I show that the proposed method can correct for strong selection bias due to covariate missingness. It substantially improves model estimates of consumer brand preference, sensitivities to marketing mix variables, and effects of consumer profiles, compared with prior approaches to the problems. Consequently, it can lead to a more accurate assessment of the impact of managerial policy and brand-management strategies and improve the firms' ability to make optimal decisions on pricing and revenue management, consumer targeting, and individualized marketing.

The method is general and can be applied to many other datasets on consumers (e.g., in my study of counterfeit consumption behavior [14], to control for selection bias and improve estimation efficiency when handling missing values in common linkage variables) and datasets on firms or organizations (e.g., [12]). The editor of *Marketing Science* notes, "There is a lot to like about this paper. It is clear that the paper is well-written, attacking an interesting and important marketing problem, and is very competently done." The reviewer team comments, "This research is timely and highly relevant," and adds, "The paper tackles a problem of immense importance, the Bayesian solution is competent, and the method is not difficult to apply in a wide range of empirical settings."

Beyond the specific applications illustrated in the paper, the proposed method contributes to the broad literature that develops appropriate analytical techniques for use with new kinds of data or new approaches to quantitative marketing and economics. In the new digital economy, the collected data are often big because of a large number of units, variables, collection times (i.e., intensive panel data), or modeling choices. The methods developed above have several merits that are important for meeting the big-data challenge in data-based marketing and brand management. Our Bayesian estimation approach overcomes an important limitation of Chen (2004) in handling high-dimensional missing-covariate problems. The approach of Chen (2004) requires evaluating likelihood, which can contain an exploded number of terms, and its computational workload increases exponentially with the number of missing variables. This constraint limits the number, types, and sizes of analyses that can be performed and makes it infeasible for many marketing applications. The Bayesian estimation approach that I developed in [7] addresses this important problem. It does not require evaluating the model likelihood and reduces the computational workload from an exponential rate to a linear rate, thereby making it feasible for many high-dimensional missing-data problems and/or complex models commonly encountered in business applications.[2] Some other key benefits of the method to the marketing field are (1) its distribution-free feature, (2) its flexibility, and (3) its simplicity in modeling and computation. First, unlike many other advanced missing-covariate methods aiming to correct for selection bias in complete-case analyses, the Bayesian procedure developed in [7] avoids model specifications on a variable-by-variable fashion and can automatically generate suitable distributions for covariates with missing values. It thus overcomes the thorny problem of properly modeling a large number of incompletely observed covariates in the big-data context. Second, the flexibility allows for complex dependence among covariates and for incorporating any useful information for covariate distributions, despite the full freedom from distributional assumptions. Finally, the simplicity substantially reduces the workload associated with carefully modeling covariates compared with alternative parametric approaches that often require a substantial amount of extra effort to search for proper distributional forms and to evaluate intractable high-dimensional integrations with respect to missing covariates.

(2) *Methods for handling missing data in outcomes.*

Researchers in marketing, economics, and other social science fields often analyze panel data collected from household panel surveys, longitudinal experimental studies, and panel scanner databases. Despite their many benefits, panel studies are often plagued by the issue of nonresponse due to both attrition and intermittent missingness. It is important to evaluate the reliability of empirical findings in the presence of panel nonresponse. This problem is well recognized. Particularly relevant and important is how to handle the case in which the nonrespondents are suspected to be different from the respondents on (possibly time-varying) characteristics *unobservable* to researchers, after controlling for all the observed characteristics (i.e., nonignorable nonresponse).

Despite this importance, formally quantifying the potential selection bias caused by nonignorable nonmonotone nonresponse is not trivial. It often requires fitting alternative complex joint nonignorable selection models and involves repeatedly evaluating a large number of intractable high-dimensional integrations with respect to nonmonotone missing data. The computational workload to evaluate these integrals increases exponentially with the amount of missingness per study unit. The problem can

[2] The new Bayesian estimation approach proposed in [7] has implications in a wider range of applications. For example, when I sent the working paper version of [7] to Dr. Chen, who first proposed the odds-ratio models that are used in [7], he commended that “this work solves a problem that troubled me for a long time. Developing a proper Bayesian estimation algorithm to the nonparametric model is nontrivial. You found an elegant solution to this hard problem.” We later worked together on adapting the estimation approach to develop a novel nonparametric multiple imputation method [6]. Unlike the direct Bayesian estimation method in [7], the imputed databases can be useful for database marketing when the intended analysis is unknown beforehand.

quickly become computationally prohibitive with new types of intensive panel data increasingly available in marketing due to technology advancement. These intensive panel data can have more complex missingness reasons and patterns, more collection times (thus more nonresponses per study unit), and more variables collected. Additionally, there can be many more complex models to consider in alternative analyses.

To address these issues, in “Measuring the Impact of Nonignorability in Panel Data with Nonmonotone Nonresponse” [8], we develop a fast index method to quantify the potential selection bias caused by nonignorable dropout and intermittent missingness. We derive the index for selection bias using a Markov transitional multinomial logit model for modeling dropout and intermittent missingness that allows nonresponse behaviors to depend on unobserved outcomes. The methodology is evaluated through extensive simulation studies and illustrated in the context of a longitudinal cohort study of 10-year smoking trends for young adult Americans. Estimating these smoking-rate changes and evaluating the validity of these estimates with respect to nonignorable missing smoking outcomes are of both public policy and managerial importance: (1) the National Institutes of Health sets goals to achieve reduction in the adolescent smoking rate, and (2) these smoking-rate changes can help firms and organizations estimate the dynamics of their future consumer bases for targeting purposes. The method can also be used to evaluate the validity of elasticity estimates of marketing activity variables. This research also tackles two high-priority research areas recommended by the Division of Behavioral and Social Sciences and Education of the National Research Council, “(1) methods for sensitivity analysis and principled decision making based on the results from sensitivity analyses” and “(2) analysis of data where the missingness pattern is non-monotone.” As a result, the method can potentially have an impact on a range of fields requiring missing-data methods.

Our analysis also demonstrates several features of the method that are important for new types of intensive panel data. The method removes obstacles encountered by alternative approaches to assess the validity of empirical findings in the presence of nonresponse. It completely avoids estimating complicated nonignorable models and does not require evaluating those high-dimensional integrals in the likelihood of nonignorable selection models. The gain in computational efficiency is large. As shown in [8], the computational time is reduced from *several hours* to only *a few seconds* for problems of moderate size. The proposed method can handle large problems without difficulty when alternative approaches become computationally prohibitive. Furthermore, it can handle a larger number of robustness analyses, which becomes computationally prohibitive for alternative approaches. In our application, because the type of missingness (dropout versus intermittent missingness) is unknown for some missing observations in the dataset, we conduct a large number of robustness analyses and use a bound approach to summarize the results. To the best of our knowledge, this is the first work in the literature to systematically address the problem of not observing the missingness reason and type in panel studies.

c Methods for Handling Completely Missing Key Variables

In a wide variety of databases and empirical studies including some of my own, researchers often find available datasets *completely missing key variables*. In some cases, a single-source comprehensive dataset that includes all essential variables is unavailable, prohibitively expensive to collect, or subject to bias due to survey context effects. In other cases, creating or sharing such comprehensive databases is constrained or even *prohibited* by data privacy laws or public concerns when sensitive respondent data are involved (e.g., see Winer 2001). I contribute to the literature by developing three kinds of methods to tackle the problem and to exploit the full potential of available datasets.

(1) In situations where key variables are collected in different samples, I develop effective and theoretically sound nonparametric data fusion methods to link key data elements collected from these different samples by using a set of linkage variables common to these independent datasets. The idea is to match nonoverlapping data items from *similar* units when matching these data from *same* units is impossible. Figure 4 illustrates the data pattern in data fusion problems where X denotes common linkage variables and Y_A (Y_B) denotes the unique variables collected only in Dataset A (B). At first glance, data fusion problems have the identical data pattern as in treatment-effect evaluation (i.e., Figure 1). However, unlike in Figure 1, Y_A and Y_B are typically not counterfactual outcomes but variables of a different nature (e.g., media exposure and buying behavior). Instead of estimating treatment effects such as $E(Y_1)$-$E(Y_0)$, data fusion typically estimates the association between Y_A and Y_B, e.g., Cor(Y_A, Y_B). These important differences call for effective methods tailored for the data fusion purpose.

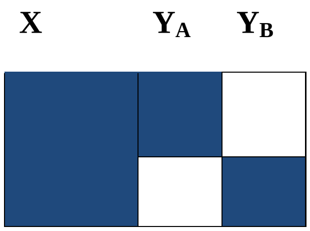

Fig. 4: Data Fusion

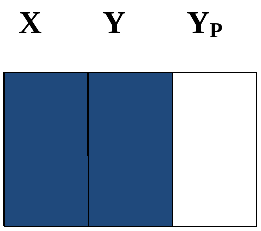

Fig. 5: Privacy Protection

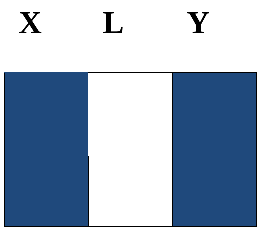

Fig. 6: Changepoint Modeling

Although different data fusion methods have been proposed in the literature, there is a strong need for more efficient and robust data fusion methods. As noted in Kamakura et al. (2006), "current [fusion] methods are strongly dependent on distributional assumptions, and [effective] nonparametric approaches are called for." In particular, the conventional industry-standard nonparametric hot-deck fusion procedure suffers from several important drawbacks noted in the literature. To address these important issues, in "Which Brand Purchasers Are Lost to Counterfeiters? An Application of New Data Fusion Approaches" [14] we develop effective and highly efficient nonparametric data fusion methods tailored for solving data fusion problems. These methods retain the main merits of the hot-deck fusion procedure, such as imposing no distributional assumptions for the variables in the datasets and using a weighted distribution on the set of observed values for fusion. On the other hand, the weights in our methods are determined by principled and coherent rules determined from statistical models with theoretical justifications, and thus our methods overcome major drawbacks of hot-deck fusion procedures. Furthermore, our methods provide closed-form joint predictive distributions of unique variables even if these unique variables contain a mixture of discrete, semicontinuous, and continuous variables, thereby enabling fast and efficient data fusion. These novel features of the proposed methods overcome important limitations of existing methods and can substantially improve the identification of consumer behavior patterns, individual prediction, and targeting. I apply them to study which types of consumers are lost to counterfeiters by combining independent data from different sources. Some findings from combining these consumer databases provide empirical support for the findings in my other work, which uses firm-level data. The study represents the first step to overcoming the important data-limitation issue in the study of underground economics and counterfeit purchase behaviors, and it opens the door to conducting more detailed investigations into counterfeiting phenomena by combining complementary consumer-level datasets from multiple sources. I am working on developing more powerful data integration methods that can examine more complex associations and relax the commonly used conditional independence assumption [30].

(2) In some situations, a comprehensive dataset does exist, but data on some key variables are confidential and cannot be shared with (thus becoming completely missing to) researchers who need them. Data privacy has become an important issue in the new digital economy and a key concern in innovation policy (Goldfarb and Tucker 2012). To solve the problem, in "Drive More Effective Data-Based Innovations: Enhancing the Utility of Secure Databases" [13], co-authors and I propose a new methodology to provide analytically valid data that also protect data privacy. The approach can be viewed as an incomplete-data problem, as shown in Figure 5, where X denotes nonconfidential variables and Y denotes confidential data that cannot be shared because of privacy concerns. The approach is to generate a perturbed version of Y, named Y_P, that can be shared because Y_P protects the sensitive values while also maintaining the statistical properties of the original data.

The current state-of-the-art perturbation methods for providing secure datasets are based on Copula models. Copula modeling has been very popular recently for modeling multivariate distributions [3], and nonparametric Copula-based perturbation methods have proven superior to prior methods. In [13], co-author and I demonstrate a range of important limitations of these nonparametric Copula-based methods, some of which are unknown in this literature. These issues limit the scope of distributional properties, as well as the types of attributes and relationships among the attributes for which these Copula-based methods are applicable, and can affect the ability of the users of secure databases to make optimal managerial and policy decisions.

We develop a new set of nonparametric perturbation methods that *simultaneously* address all these limitations of existing methods while retaining their good properties. Compared with existing methods, the proposed procedures substantially increase the utility of secure databases without increasing disclosure risk. As demonstrated in our application, the proposed methods preserve the important inverted-U relationship between competition and innovation while the existing methods create secure datasets that lead to the conclusion of no relationship between competition and innovation, a difference with very significant managerial and policy implications. Our evaluation also shows that simple strategies, such as rounding to different decimal places, can sufficiently control the computational time without affecting the utility of secure database, thereby making the proposed methods suitable for creating secure large databases. Given the importance of data, the proposed methods can have important impacts on driving data-based innovations in many fields. The reviewer team commended this paper as "an important contribution to the literature for secure analytic data." It continued, "The problem addressed in the manuscript is an interesting and important one for the readers of this journal [*Management Science*]" and "the problem investigated is very important and relevant to the field. The proposed technique is not tied to a specific statistical model/distribution. It can preserve nonlinear and nonmonotonic relationships between attributes."

(3) In some other situations, the completely missing key variables can be inferred from data using proper statistical models. In [20], co-author and I develop a third type of method that relies on latent variable models to infer firm-specific response times to counterfeit entry. In our application, the completely missing variables L refer to the latent firm-specific changepoints (structural breaks) in response to counterfeit entry. Our analysis demonstrates that the new model, which models these latent firm-specific changepoints, matches the underlying process better than the traditional hierarchal Bayesian (HB) models, which do not model them, and yields improved estimates of the dynamic and moderating effects of counterfeit entry. The new model also reveals that firms with more innovation,

less diversification from infringed markets, or more human capital are faster in responding to and differentiating from counterfeits.

Overall, with the availability of “big data,” robust, automated, and scalable methods are in high demand to harness these data for understanding economic phenomena and guiding managerial and policy decisions. I have developed a stream of research that formulates various research issues as incomplete data problems, and develop methods that can reduce or detect model independence so that empirical findings are not artifacts driven by unnecessarily imposed assumptions. These automated methods can match more closely with the true underlying processes, make better use of available datasets, and have important modeling and computational merits for scalable data-driven decisions in a big-data environment. They are applied in a wide range of ways for better treatment-effect evaluation in studies of innovation, counterfeits, and intellectual property rights, and for improved data-driven decisions in brand management, pricing and revenue management, targeted marketing and customer relationship management, and data integration and database development.

Ongoing and Future Projects

My work suggests ample opportunities for future research. I am currently developing a novel statistical matching procedure that matches data from independent consumer samples to study the impact of counterfeits on consumer brand awareness [30]. The procedure allows for estimating a consumer-level model of brand awareness as a function of the extent of counterfeit presence and a rich set of consumer demographic and purchasing behavior variables, even if the two key variables (brand awareness and extent of counterfeit presence) are observed only in two independent samples. The procedure allows for examining more general relationships among key variables as compared with existing data fusion methods and relaxes the strong conditional independence assumption commonly employed in the existing data fusion methods. Preliminary results from the new statistical matching analysis show that increased counterfeit exposure raises brand awareness for products that are less known to consumers.

I have also been analyzing the Hertz consumer databases to develop appropriate models for estimating the effects of customer and employee satisfaction ratings on customer retail purchase behaviors, while accounting for self-selection biases associated with informative satisfaction survey nonresponses. There are research opportunities to develop more general and efficient approaches to achieve optimal tradeoffs between data-driven innovations and data privacy concerns. Because some methods that I have developed require relatively sophisticated programming, there is a need to develop user-friendly software for others to use. Such work could be supported through external funding (e.g., NSF) and make these methods more useful.

References

1. Blattberg, R.C., Kim, B.D., and Neslin, S.A. (2008) **Database Marketing: Analyzing and Managing Customers**. Springer, New York.

2. Chen, H. Y. (2004) “Nonparametric and Semiparametric Models for Missing Covariates in Parametric Regression,” Journal of the American Statistical Association, 99, 1176-1189.

3. Erdem, T., Keane, M.P., and Sun, B. (1999) “Missing Price and Coupon Availability Data in Scanner Panels: Correcting for the Self-selection bias in Choice Model Parameters,” Journal of Econometrics, 89:177-196.

4. Goldfarb, A., and Tucker, C. (2012), ``Privacy and Innovation," **Innovation Policy and the Economy,** Vol. 11, Josh Lerner and Scott Stern (Eds), NBER.

5. Kamakura, W.A., Mela, C.F., Ansari, A., Bodapati, A., Fader, P., Iyengar, R., Naik, P., Neslin, S., Sun, B., Verhoef, P.C., Wedel, M., and Wilcox, R. (2005), “Choice Models and Customer Relationship Management,” Marketing Letters, 16:279-291.

6. Rubin, Donald, and Neal Thomas (2000), “Combining Propensity Score Matching with Additional Adjustments for Prognostic Covariates,” **Journal of the American Statistical Association** 95, 573–585.

7. Winer, S Russell (2001), “A Framework for Customer Relationship Management”, **California Management Review**, 43: 89-105.