Statistical Science
2019, Vol. 34, No. 4, 669–691
https://doi.org/10.1214/19-STS754

Statistical Theory Powering Data Science

Junhui Cai, Avishai Mandelbaum, Chaitra H. Nagaraja, Haipeng Shen and Linda Zhao

Dedicated to Lawrence D. Brown

Abstract. Statisticians are finding their place in the emerging field of data science. However, many issues considered "new" in data science have long histories in statistics. Examples of using statistical thinking are illustrated, which range from exploratory data analysis to measuring uncertainty to accommodating nonrandom samples. These examples are then applied to service networks, baseball predictions and official statistics.

Key words and phrases: Service networks, queueing theory, empirical Bayes, nonparametric estimation, sports statistics, decennial census, house price index.

1. INTRODUCTION.

In 2009, Hal Varian, the Chief Economist at Google quipped, "I keep saying the sexy job in the next 10 years will be statisticians (Varian, 2009)." This statement was quoted often, much to the delight of statisticians (based on anecdotal evidence). By 2012, *Harvard Business Review* ran an article titled: "Data Scientist: The Sexiest Job of the 21st Century (Davenport and Patil, 2012)."

Junhui Cai is Ph.D. candidate, Department of Statistics, The Wharton School, University of Pennsylvania, 400 Jon M. Huntsman Hall, 3730 Walnut Street, Philadelphia, Pennsylvania 19104, USA (e-mail: junhui@wharton.upenn.edu). Avishai Mandelbaum is Professor Emeritus, Faculty of Industrial Engineering and Management, Technion—Israel Institute of Technology, Room 518, Bloomfield Building, Technion City, Haifa 3200003, Israel. Chaitra H. Nagaraja is Associate Professor of Statistics, Strategy and Statistics Area, Gabelli School of Business, Fordham University, Martino Hall, 45 Columbus Avenue, New York, New York 10023, USA. Haipeng Shen is Patrick S C Poon Professor in Analytics and Innovation, Faculty of Business and Economics, The University of Hong Kong, Room 815, K. K. Leung Building, Pok Fu Lam Road, Hong Kong. Linda Zhao is Professor of Statistics, Department of Statistics, The Wharton School, University of Pennsylvania, 400 Jon M. Huntsman Hall, 3730 Walnut Street, Philadelphia, Pennsylvania 19104, USA.

Statistics had suddenly been eclipsed by the newly-minted "data science." This was not simply an issue of nomenclature—a la the Gershwins', "You like tomato and I like tomatoe. . . (Gershwin and Gershwin, 1937)." Rather, from the statisticians' perspective, it was a question of identity and confidence. At times, it almost felt as if people were implying you could be a "data scientist" without ever needing "statistics."

This question of identity was pervasive enough that the American Statistical Association felt the need to issue an official statement on the subject in 2015, written by a panel of prominent statisticians from both academia and industry (van Dyk et al., 2015). "Framing questions statistically allows us to leverage data resources to extract knowledge and obtain better answers," they wrote. The panel offered examples from the concept of randomness to correlation versus causation to quantifying uncertainty. They also noted that these are all ideas debated in data science today that statistics—and statisticians—have thought deeply about for centuries.

In 2018, *Statistics and Probability Letters* published a special issue titled: "The Role of Statistics in the Era of Big Data." The articles within reiterated the idea that what seemed like new issues in data science were already part of the field of statistics. Moreover, the various authors argued that statistical thinking provided a suitable framework for solving these problems (Sangalli, 2018).

It was—and still is—a time of reckoning for statisticians. Some of data science falls under statistics, some of it does not. That much is clear. It is not our goal in this paper to litigate how much prominence statistics should have in the field of data science or to provide specific recommendations to advocate for our subject. Rather, we hope to highlight how current debates in data science can be cast in terms of statistical thinking. That is, we hope to show the importance of statistics by example.

Lawrence D. Brown, to whom this paper is dedicated, was primarily known for his work on theoretical topics. Emphasizing his theoretical contributions, however, shortchanges a statistician who had an extraordinarily wide set of interests. (Two coauthors of this paper even found that people were surprised that Brown was their thesis advisor when presenting their applied dissertation work!) For example, his work in decision theory and sense of geometry brought a fresh perspective to applications from call centers to baseball to the decennial census.

In this paper, we use those research interests as examples to illustrate the importance of statistical thinking in data science. We begin in Section 2, where a complete framework is proposed for approaching service network problems. The framework demonstrates what data science is about: it is to begin with transaction-level operational data, collected at service systems such as call centers and hospitals; continue with data management, data cleaning, visualization and analysis (in particular Exploratory Data Analysis) through statistical and graphical engines; and culminate with a data-based simulation of the service networks, jointly with validated operational models (theoretical, robust) that support design and management of the service systems that originated the data. This framework has its roots in the seminal work of Brown et al. (2005), which presented the first thorough statistical analysis of transaction-by-transaction service data from a banking call center, and applied those findings to support some and dispute other popular operational models of call centers and their building blocks. One aspect of further development of the framework includes providing valid statistical inference after model selection, based on the pioneering work by Brown's group at the University of Pennsylvania (Berk et al., 2013).

In Section 3, we review and expand Brown (2008) on predicting batting averages for a baseball player. Brown applied several well-known classical parametric methods along with a nonparametric empirical Bayes addition. These methods, however, generate only point estimates without uncertainty quantification, a key issue in data science. Cai and Zhao (2019) proposed an alternative nonparametric empirical Bayes procedure that provides accurate predictions and quantifies uncertainty using predictive intervals.

We take a more historical perspective in Section 4 with a focus on official statistics in the United States. Brown served on numerous committees during his life in an effort to improve statistics produced by federal agencies. Using examples from the history of the U.S. federal statistical system, we will discuss three issues in data science: the effects of repurposing data, data confidentiality and privacy, and working with nonrandom samples. We conclude briefly in Section 5.

2. DATA-BASED SERVICE NETWORKS (SERVNETS): MODELS IN SUPPORT OF INFERENCE, ANALYSIS, DESIGN, CONTROL, PREDICTION AND SIMULATION OF SERVICE SYSTEMS.

Brown et al. (2005) laid the foundation for a large body of research within operations research and operations management, as well as in statistics and more broadly in data science. Its OR and OM impact can be appreciated through the survey by Gans, Koole and Mandelbaum (2003), and over 2400 papers that refer to both Brown et al. (2005) and to that survey.[1] Here, we describe briefly some of its Statistical siblings, as part of the future research horizon that Brown et al. (2005) revealed.

To this end, we start with a description of our world of practice: service networks. Then we proceed with laying out a modeling territory for operational models (empirical, mathematical and simulation), with special attention given to two models, Erlang-A (that motivated Brown et al., 2005) and Erlang-R (a natural sequel). Each of these two models mathematically highlights an operationally-significant phenomenon (Erlang-A: hanging-up while waiting for a phone-service (Garnett, Mandelbaum and Reiman, 2002), and Erlang-R: service-cycling while being treated in an emergency department (Yom-Tov and Mandelbaum, 2014)); and being tractable mathematical models of real service systems, they are necessarily "simple models at the service of complex realities." Now it turns out that both Erlang models have been successful "servants" of these realities, and below we provide (a very)

[1] According to Google Scholar (https://scholar.google.com) as of October 27, 2019.

preliminary support of the hope that this success is generalizable.

We conclude with a paradigm for a future research horizon: it starts with data from a service system and its exploratory-data-analysis (EDA); it continues with a platform for modeling, visualization (of data and models), and performance- and predictive-analysis; and it culminates with validated valuable models that support the design/planning and control/management of the service system from which the data originated.

There are two prerequisites for the success of our paradigm. The first is a multidisciplinary research partnership between Operations Research and Statistics—which is precisely the theme of Brown et al. (2005). The second is operational service-data at the resolution of the individual customer-server transaction—Technion's SEELab (Laboratory for Service Enterpriser Engineering) is a home for such data, which has originated in call centers, hospitals, courts, banks and more. For example, Figures 2 and its corresponding animation are based on such transaction-data from a hospital.

2.1 Service Networks (ServNets).

Our relevant world-of-practice includes telephone call centers (business and marketing, emergency, help desks), hospitals (emergency rooms, outpatient clinics, operating rooms, wards), public service centers (municipal, government), banks (front and back office), airports, supermarkets, field-services, transportation systems and more. Our focus is on operational characteristics, namely customer or server flows, and on system congestion as manifested by its queues (with operational characteristics serving also as surrogates for financial, psychological or clinical performance). In such realities, as will be now demonstrated, a network-view is natural. This leads to our operational network models of service systems, which will be referred to as *Service Networks*, or **ServNets** for short.

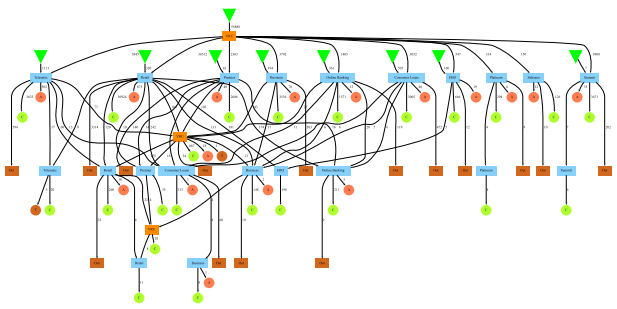
2.1.1 *Static and dynamic depictions of ServNets* (*call centers and hospitals*). To be (visually) concrete, Figure 1 displays two snapshots of call-center ServNets, both generated from SEELab data. The left one represents overall customer flow—through the answering machine, waiting in a queue, being served or abandoning. The right figure zooms in on skills-needs matching—flow from queues of customers (needs, in green) to pools of agents (skills, in orange).

Figure 2, and its corresponding animation (https://youtu.be/H7Td6q-uI7w), are based on SEELab data from Boston's Dana Farber Cancer Institute (DFCI). This real data covers jointly about 1000 *patients* daily, who are treated by around 350 medical personnel (in particular *doctors* and *nurses*) in 100 (20) exam (consultation) *rooms* and over 150 infusion positions. The data/animations capture the precise location/status of *all* these resources, over several years; it has been gathered continuously and automatically through a Real-Time Location System (RTLS), that collects data via 900 sensors scattered over 8 clinical floors.

Figure 3 focuses on the care-path—61 hospital visits—of a single DFCI patient over a 6-month period: a typical single full visit consists of a blood test, then a doctor's exam and finally an infusion treatment; the figure displays also the queues prior to these three activities. Figure 2 in Mandelbaum et al. (2017), with its animation (https://youtu.be/e1qHeYg7hfw), complement Figures 2–3 by displaying all patient locations at a single infusion unit (28 chairs and beds).

2.1.2 *Our examples of ServNets*: *QNets*, *SimNets*, *FNets*, *DNets*. Queueing theory is ideally suited to capture the operational tradeoff that is at the core of any

(a) **Call-Center Flow, US Bank (2/4/01)**



(b) **Skills-Needs Matching, IL Telecom (10/2/08)**

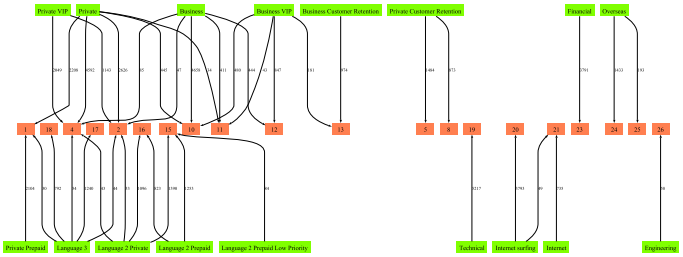


FIG. 1. *Call-Center ServNets* (*SEEGraphs*). *Magnifying the above reveals futher details. Related animations can be viewed on https://www.youtube.com/watch?v=1A6-jzS_scI&t=65s; in particular, "Skills-Needs Matching" starts at minute* 1:40 *of the video.*

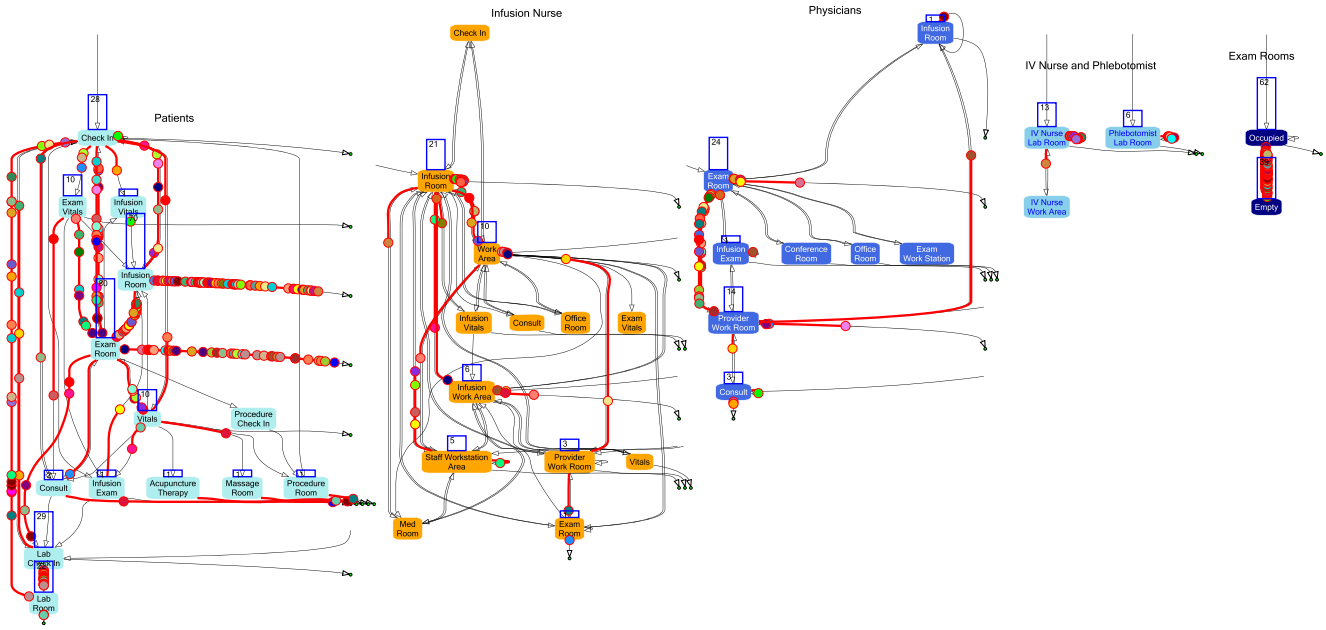


FIG. 2. *ServNets of Resources. From left to right*: *patients*, *infusion nurses*, *doctors*, *blood-draw nurses*, *rooms*. *Data animation can be viewed on https://youtu.be/H7Td6q-uI7w.*

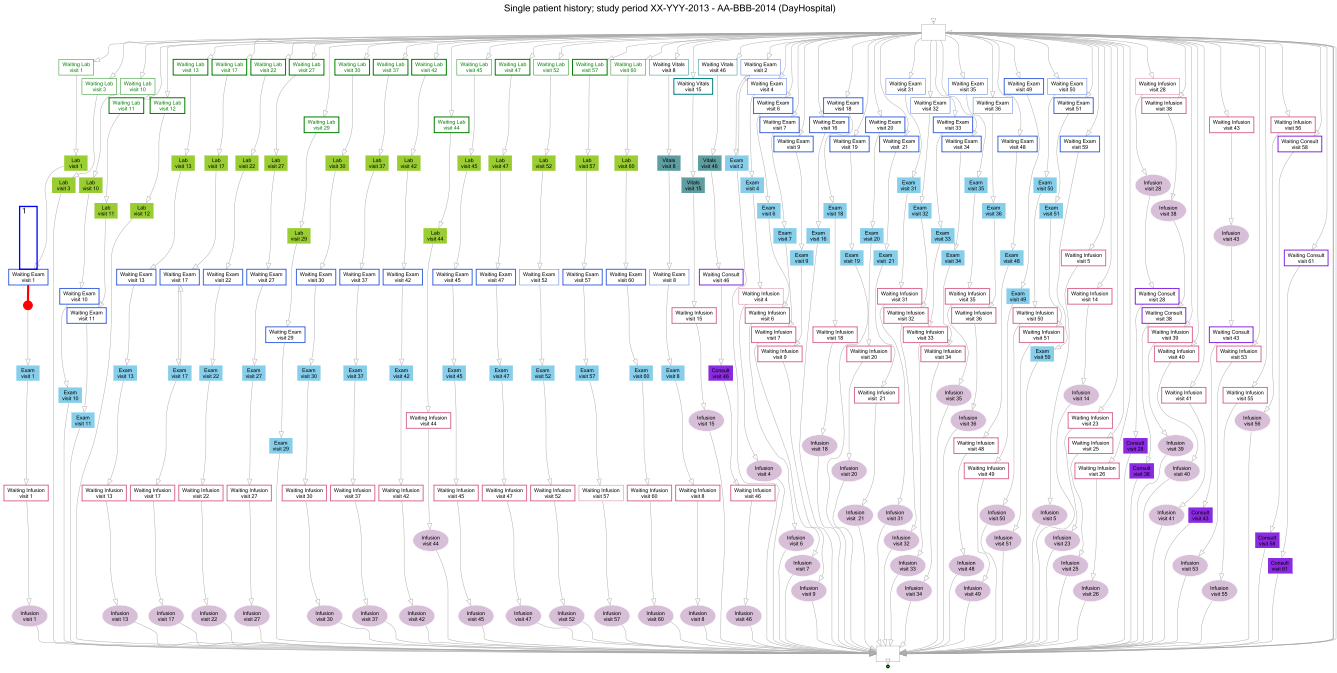


FIG. 3. *Complex care-path at DFCI* = 61 *hospital visits by a single patient* (*the red dot*) *over* 6 *months. The visits are sorted by visit-type* (*as opposed to chronologically*). *Data animation can be viewed on https://youtu.be/e1qHeYg7hfw.*

service, namely quality versus efficiency, and it is flexible enough to provide insights beyond the operational. However, and as already suggested, classical queueing models are unable to tractably accommodate many service characteristics (e.g., transience or finite-horizon, heterogeneity, fork-join or precedence and synchronization constraints, fairness) and service resolutions (from mass customization to flow aggregates). One solution is then to resort to either simulation or approximations. Simulation models, or SimNets, play an important role in the modeling world due to their generality and proximity to reality—but these advantages come at the cost of *complexity*, which renders SimNets difficult to develop, maintain/update and analyze. Below we shall outline an approach/vision that will attempt to overcome some or most of these shortcomings. Toward that end, one must enrich the family of tractable models, which is the role of approximations.

Our framework for approximations (which needless to say is biased by our own research) is based on *asymptotic queueing* theory (Whitt, 2002a, Chen and Yao, 2001, Robert, 2003), specifically fluid (thermodynamic, macroscopic) limits, and their diffusion (mesoscopic) refinements. The former gives rise to Fluid Network models of service systems (**FNets**, derived via strong laws of large numbers for stochastic processes), and the latter to Diffusion Networks (**DNets**, via functional central limit theorems). Growing out of asymptotic laws, FNets and DNets strip their originating QNets off their inessential characteristics. As such, they enjoy dramatically reduced complexity, and hence increased robustness, which makes them potentially valuable for *inference, analysis, design, control, prediction and simulation* of their originating service systems.

2.1.3 *QED ServNets*, *ideally.* Depending on their strategic goals, service systems typically thrive to be Quality-Driven (short delays of customers to servers) versus Efficiency-Driven (short delays of servers to customers). However, service systems that are not "too small" could circumvent this trade-off by being both Quality- and Efficiency-Driven, or QED for short. The QED operational regime entails carefully matching service capacity with customer demand, and asymptotic ServNets provide the recipe for this to happen.

Quantitatively, the recipe for QED performance is the *square-root staffing rule* (Erlang, 1948, Whitt, 1992, Garnett, Mandelbaum and Reiman, 2002, Gans, Koole and Mandelbaum, 2003, Brown et al., 2005), which reads as follows (using the terminology of call centers, though only for concreteness): Suppose that N statistically identical agents cater to a single queue of customers that await telephone service; suppose also that the offered-load demanded by customers is R; here, R is the average amount of work, measured in units of time, that customers demand per time-unit. (For example, an arrival rate of 25 customers per minute, each demanding an average service of duration 4 minutes, gives rise to an offered-load of $R = 25 \times 4 = 100$ minutes-of-work per minute-time.) The QED regime is then guaranteed by a staffing level of $N = R + \beta \times \sqrt{R} + o(\sqrt{R})$ agents, where β is a relatively small constant (e.g., within $[-1, 2]$, as in Plot 5.3); the exact value of β is determined by further refined specifications, for example, quantifying the importance of service quality relative to system efficiency (Borst, Mandelbaum and Reiman, 2004, Mandelbaum and Zeltyn, 2009) (e.g., cost of an abandonment, or its forgone profit, relative to the salary cost of agents).

2.1.4 *Two concrete examples*: *Erlang-A* (Garnett, Mandelbaum and Reiman, 2002) *and Erlang-R* (Yom-Tov and Mandelbaum, 2014). As mentioned, Erlang-A motivated Brown et al. (2005), and has consequently become the basic call center model; Erlang-R has been found useful in healthcare settings (e.g., for modeling emergency departments).

Consider first Erlang-A: Customers arrive in a queue according to a Poisson process with rate λ; they "enjoy" patience of exponential duration with mean $1/\theta$ (they abandon the queue once their wait reaches their patience); customers are served on a first-come-first-serve basis by N independent and statistically identical servers, who provide services of exponential durations with mean $1/\mu$ (rate μ). Defining its state to be the total count of customers (waiting or in-service), Erlang-A is then a 1-dimensional birth-death process, which is characterized by the 4 parameters $(\lambda, \mu, \theta, N)$. Its steady-state could successfully capture/model the operation of say a call-center during a short period such as one hour: see Figure 8 in Brown et al. (2005).

As it turns out (e.g., Brown et al., 2005, Reich, 2011), however, *none* of the underlying assumptions of Erlang-A (Poisson arrivals, exponential service and patience, independence among all its building blocks) prevails in practice. Moreover, being a model of a single queue of impatient customers, who are served by a single pool of agents, Erlang-A is comparable to *merely* the *right-most* arc of ServNet 1.2 in Figure 1.

Still this very simple Erlang-A, as well as its corresponding FNet and DNet, have been found highly valuable for both their theoretical insights into and practical support of call center operations (Garnett, Mandelbaum and Reiman, 2002, Gans, Koole and Mandelbaum, 2003, Brown et al., 2005)—such an operation could be as complex as the model of the basic call center in Figure 1 of Garnett, Mandelbaum and Reiman (2002)), or even ServNet 1.1 in Figure 1.

Moving on from call centers to hospitals, a 2-dimensional time-varying Markov jump process (network) was *needed* for capturing nurse operations in an emergency department, or physicians during a mass-casualty event. That need traces to a combination of two factors: arrival rates are time-varying (as opposed to the steady-state Erlang-A) and patients undergo a service process that consists of Recurrent services (e.g., nurse service, then a "pause" for an x-ray, then nurse service again, etc.), hence Erlang-R. Formally (Figure 1 in Yom-Tov and Mandelbaum (2014)), Erlang-R has arrivals that are nonhomogeneous Poisson; and upon service completion, each customer either completes the service process with probability $1-p$, or alternatively "goes into orbit" where the customer spends an exponential(δ) time (mean $\frac{1}{\delta}$), after which the customer rejoins the queue for an additional service.

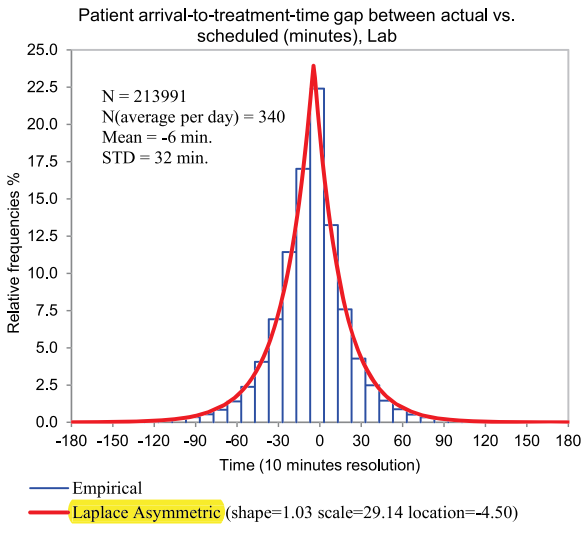
In Yom-Tov and Mandelbaum (2014), it is demonstrated, based on Erlang-R or rather its time-varying FNet model and DNet refinements, that an appropriate time-varying square-root staffing, in fact, stabilizes the operational performance of nurses or physicians. This was validated against two simulation models (SimNets), specifically a (stylized) time-varying Erlang-R and a (realistic) simulation model of an emergency department. For future reference, readers should note that all four ServNets (QNets, FNets, DNets, SimNets) were used in the above-described analysis of Erlang-R (Yom-Tov and Mandelbaum, 2014).

2.1.5 *Why be optimistic?* Because "simple models can valuably portray complex realities." Consider first the basic building blocks of ServNets—*arrival processes* and *service durations*. Unscheduled arrivals (e.g., to call centers or emergency departments) often fit Poisson processes and their relatives (more on this below); appointment-driven arrivals, on the other hand, are determined by system appointment-time and customer-*punctuality*: Figure 4(a) shows that the latter fits an asymmetric Laplace distribution. As for service-durations, for example, telephone services (Brown et al., 2005) and doctor-exams (Figure 4(b))—these are often found to be LogNormally distributed.

Interestingly, there are of yet no theoretical explanations for the above excellent parametric fits, which does not take away from their value (e.g., these distributions serving as building-blocks of SimNets). We have thus shown that simple (parametric) models could fit (the building blocks of) a complex ServNet well.

Figure 5 continues this theme, via four data-stories from call centers ("told" at the SEELab). Figure 5(a) and (b) depict the *congestion law* of *State-Space-Collapse*, which arises in heavy-traffic asymptotics

(a) **Patients punctuality (< 0 = early arrival)**



(b) **Actual vs. planned durations (15, 30, 60 min)**

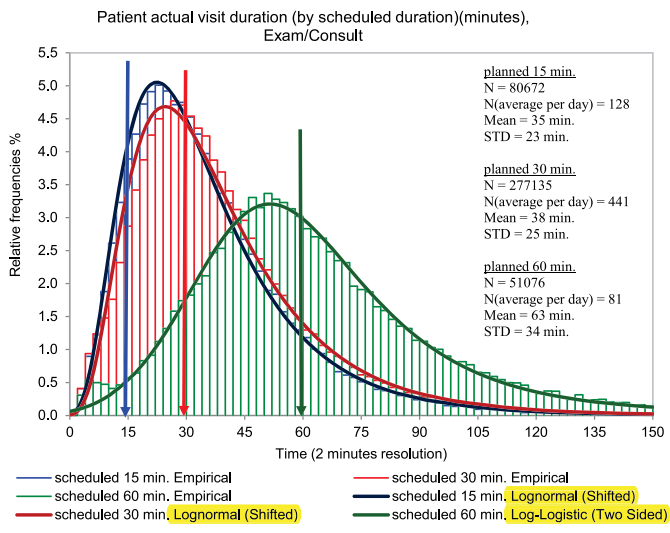
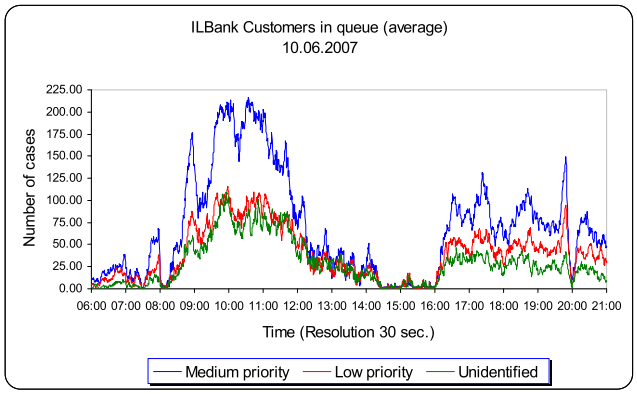
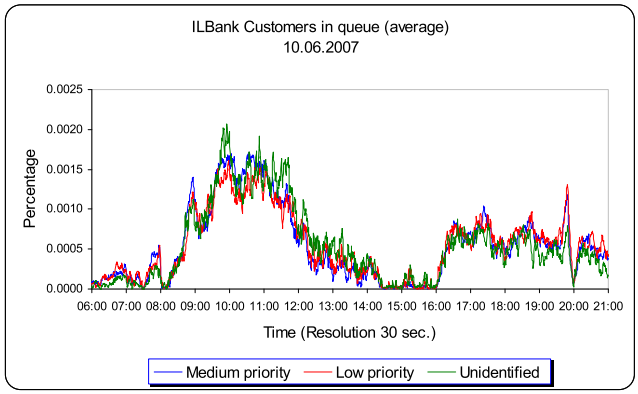


FIG. 4. *Punctuality of patients and durations of doctor exams*: *Simple Models Fit Complex Realities*.

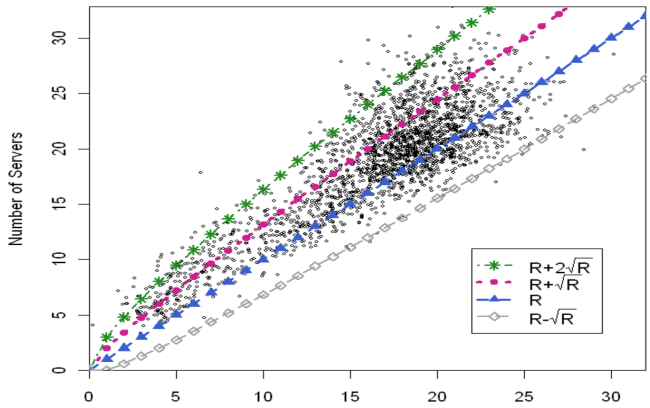
(a) **Three Queues in IL Bank**



(b) **State-Space Collapse (Area = 1)**



(c) **Square-Root Staffing (QED) in Practice**



(d) **Waiting: Announcements & Priorities**

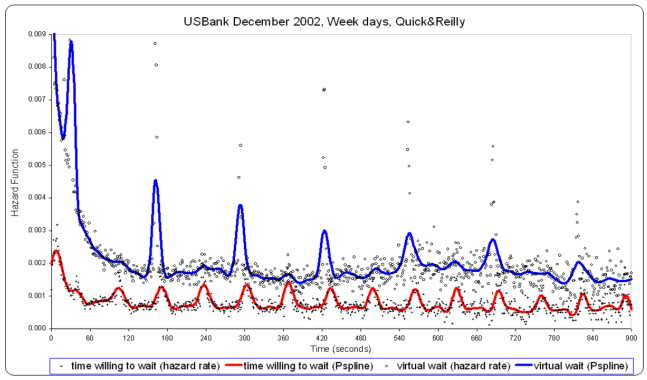


FIG. 5. *Data Stories, as told by SEEStat.*

of DNets (conventional (Bramson, 1998) or many-servers (Gurvich and Whitt, 2009)). According to this law, queue-lengths of different customer classes (Figure 5(a)) are actually proportional (equal in Figure 5(b), after normalization), at the granularity of an individual sample-path and at all times (the time-resolution in Figure 5(a) and (b) is 30 seconds): thus effectively, a low-dimensional model captures (almost surely) the dynamics of a high-dimensional system. Figure 5(c) validates asymptotic theory: it depicts a QED call-center, as introduced earlier, in which $N \approx R + \beta\sqrt{R}$; here N is staffing level, $R = \lambda/\mu$ is offered-load, and β is the constant that determines precise operational performance.

Finally, Figure 5(d) (from Mandelbaum and Zeltyn (2013)) amplifies the complexity of customer waiting, in particular over the phone: the red line corresponds to the time that a customer is *willing* to wait—(im)patience; and the blue line is the time that a customer is *required* to wait—*virtual* wait ("virtual" since this is the waiting time of a customer equipped with unbounded patience; thus the *actual* wait is the minimum between willing and required). Note that, data-wise, patience and virtual-wait right-censor each other (e.g., the patience of a customer who was served after 2 minutes is at least 2 minutes), hence Figure 5(d) displays uncensored graphs. We learn from it that an announcement every minute affects (im)patience (red line)—indeed it triggers abandonment (somewhat surprising (Brown et al., 2005)), and a priority upgrade every 2 minutes affects the required wait (blue line)—it increases the likelihood of being served (which is to be expected).

2.1.6 *Diving deeper into the main building blocks of ServNets, statistically.* Going beyond the above "data-stories," Brown et al. (2005) motivated ample statistical research on ServNet building blocks and their interaction.

Waiting time. There is even more than meets the eye in Figure 5(d): Li, Huang and Shen (2018) studies how

the behavior changes as a function of two timescales—waiting duration, and the time-of-day when a customer calls in. The latter is natural since time-varying environments are prevalent in service systems. Figures 3 and 4 in Li, Huang and Shen (2018) first demonstrate that, conditioning on time-of-day, the effects of announcements and priority upgrades remain similar to Figure 5(d); as for temporal behavior, customers' patience level is quite stable for most of the day, while being the lowest around 5 pm and highest in the evening.

Arrivals. Time-inhomogeneous Poisson processes (with potentially random arrival rates to accommodate overdispersion) have proved to be valuable models for arrivals at call centers and hospitals; see Kim and Whitt (2014), for example. More recently, Chen, Lee and Shen (2018) show that a Poisson process, with a rate that is a simple sum of sinusoids, can adequately describe the reality of call centers and emergency departments; they further introduce a statistical learning technique for estimating the time-varying arrival rate. Empirical research has shown that the rates at which customers arrive are not known with certainty ahead of time, and hence must be forecasted. Statistical models have sought to better characterize the distribution of arrival rates, by time of day, as they evolve; see, for example, Weinberg, Brown and Stroud, 2007, Shen and Huang, 2008a, Shen and Huang, 2008b, Aldor-Noiman, Feigin and Mandelbaum, 2009, Matteson et al., 2011, Taylor, 2012, Ibrahim and L'Ecuyer, 2013, Ibrahim et al., 2016a, Ye, Luedtke and Shen, 2019. Data-Driven forecasting and stochastic programming (SP) framework is proposed by Gans et al. (2015) to cope with arrival-rate uncertainty, integrating statistics and operations management. Intra-day forecast updating and SP with recourse are incorporated so that managers can dynamically adjust agent schedules to make staffing levels more efficient.

Service durations. Another important empirical observation made by Brown et al. (2005) is *service-time heterogeneity*, namely service durations that are agent-dependent (some agents are faster than others). Regression techniques have been developed to systematically understand how time-dependence, learning, shift fatigue, cross-training and other factors affect the distribution of service durations (Shen and Brown, 2006, Gans et al., 2010, Ibrahim et al., 2016b). Mixed-effects models were used to study agent population heterogeneity, which revealed that the rate and ultimate degree of learning can differ significantly across agents. Such statistical findings can guide the development and analysis of operational models, which account for agent heterogeneity in hiring, staffing and retention policies.

Most existing research treats building blocks as being independent. Hence each block is analyzed unilaterally while, in fact, there can exist dependence among building blocks: for example, long patience goes hand in hand with long service (Reich, 2011). Such dependence underscores our network view, in which blocks are connected as parts of a process, and hence analyzed jointly.

2.2 Research Framework for Automatic Creation of Valuable ServNets.

We ask the reader to contemplate the following two questions:

Q1 What did it take to produce the networks in Figures 1, 2 and 3?

Q2 What will it take to create SimNets for such networks?

The two short answers, respectively, are "a lot" and "not much," which suggests that *state-of-the-art is not far from being able to automatically create all ServNets from data*. In support of this vision, we propose the research framework in Figure 6:

1. Start with obtaining data from a service system and convert it to ServData format;
2. Use a statistical engine (SEEStat) and a graphical engine (SEEGraph) to model the building blocks of ServNets;
3. Create a data-based simulation (SimNet) that will serve as a virtual reality of that system;
4. Create a corresponding QNet model;
5. Develop data-based FNets and DNets (approximations) of the QNet. (With enough experience, one could create both directly from data.);
6. Use SimNet to test the accuracy and value of, namely *validate*, all mathematical models (QNet, FNets, DNets);
7. Repeat Steps 2–6 as necessary.

Figure 6 was created after the classical scientific paradigm: experiment, measure, model, validate, analyze, refine, etc. It is traditional and routine in biology, chemistry and physics, and recently has been exercised also in economics and transportation. Here, we propose to adopt it to what can be called *Service Science*; or perhaps the special case of Network Science, where the network is a ServNet. This will yield a novel theory that is born from measurements and experiments.

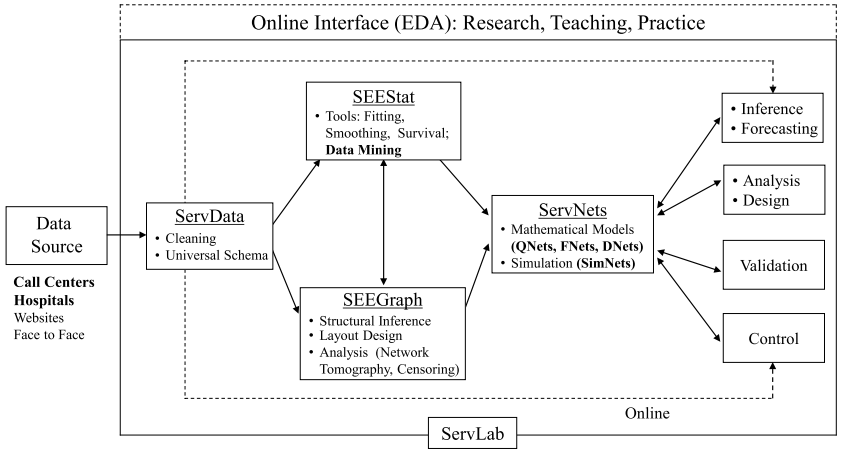


FIG. 6. *Research framework—from data to ServNets and beyond.*

But unlike in most of the above sciences, where experimentalists generate data and conjectures for theoreticians, our new sciences are only now starting to build the infrastructure for collecting, managing, and experimenting with service data and the models it supports (e.g., data from call centers, hospitals and expanding to internet-based shared-economy data).

What have we learned from the above? *That, with the right theoretical (asymptotic) backing, simple QNets, FNets and DNets (though not too simple) can capture specific essentials of a complex service system, for example, total customer count or the dynamics of a subsystem. These simple models can then be validated, theoretically and practically, against a data-based SimNet of that system.* Our lesson has been abstracted to the framework in Figure 6. And the research that will operationalize the framework must do it all: identify, formulate and solve open problems, pave new research directions and explore uncharted territories.

2.3 Further Research.

Significant knowledge gaps must be addressed along several directions: Theory (asymptotics of ServNet idiosyncracies); Design and Staffing (e.g., stabilizing performance over time); Control (time-varying, information-sensitive); Statistical Inference (EDA, structure, missing data); Automatic Creation of ServNets; Validation (value-testing against SimNets). We now expand on some statistics-related ingredients.

2.3.1 *Statistical inference*. Naturally, statistics plays important roles in our data-based research agenda, some leading roles and some facilitating, some existing and many yet to be developed. First and foremost, EDA (Exploratory Data Analysis (Tukey, 1977)) will remain an eye-opener and a research guide, as in Brown et al. (2005) for call centers and Armony et al. (2015) for hospitals. For example, this is how it was discovered that service durations are often LogNormally distributed, which has inspired deep research on many-server queues with general service times (Puhalskii and Reiman, 2000, Mandelbaum and Momčilović, 2012, Kaspi and Ramanan, 2011). Statistics will also help support the modeling and analysis of ServNet building blocks and their interaction—recall Figure 4, then also consider modeling the process of arrivals to service, and calculating offered-loads when there exists dependence between service duration and impatience (Reich, 2011). The roles above mostly fall under traditional Statistics, with some (e.g., validation of models) requiring tools from *Data/Process Mining* (Senderovich, 2016) and very likely/naturally Process Machine Learning.

A theme that goes beyond the traditional is the *interplay of statistics with asymptotics*, which raises novel statistical and mathematical problems. An example of a theory that came out of data is the asymptotic staffing recipes against over-dispersed arrivals (Maman, 2009). Such overdispersion (more stochastic variability than Poisson) prevails in call centers (but, interestingly, not in emergency departments where patient arrivals fit well a nonhomogeneous Poisson process). Conversely, examples of statistical questions that arise from theory, and for which asymptotic models will facilitate, are inference of system structure (Deo and Lin, 2013, Zeltyn et al., 2011), completion of partial information (Zhang and Serban, 2007) and imputation of data that is either missing or costly to get (e.g., service durations of abandoning customers, which are unobserved yet re-

quired for calculating offered-loads (see Reich (2011), but much remains to be done).

2.3.2 *Estimating* (*im*)*patience via asymptotics*. An unexplored inference question is that of fitting an asymptotic regime: do our data originate in a QD or ED or QED system? (QED, as already defined, carefully balances service quality and servers' efficiency while, due to economies-of-scales, achieving high levels of both; QD, standing for Quality-Driven, is an operational regime that emphasizes service quality over efficiency, namely customers enjoy very high levels of service but at the cost of high idleness levels—low efficiency—of servers; conversely ED, which corresponds to Efficiency-Driven, thrives for high levels of servers utilization—high efficiency—at the cost of relatively poor service levels.) Why would one wish to know? We now present an example in which knowing the regime simplifies the inference of the originating QNet. In QED systems with impatient customers, theory tells us that the role of the impatience distribution is fully captured by its behavior near the origin, or the value of its density $g(0)$ when positive (Zeltyn and Mandelbaum, 2005, Dai and He, 2010, Mandelbaum and Momčilović, 2012), which we now assume. Thus, it suffices to estimate only $g(0)$ and not worry about the full impatience distribution. (The latter would require survival analysis that accommodates censored data: being served after waiting for 1 minute provides only a lower bound, 1 minute, on that customer's impatience (Brown et al., 2005).) One can estimate $g(0)$ via the congestion law $P\{\text{Abandon}\} \approx g(0) \times E[Wait]$ (Zeltyn and Mandelbaum, 2005); and performance is then calculated via formulae (sometimes tractable and always simpler) of ServNets in which impatience is *exponential* with mean $1/g(0)$ (Zeltyn and Mandelbaum, 2005, Mandelbaum and Momčilović, 2012). This could already be a happy ending: QED analysis reduced the estimation of the full impatience distribution to a straightforward estimation of the slope of a straight line. But the story is getting even more interesting. There is a refined QED regime—after hazard-rate scaling (Reed and Tezcan, 2012)—which does retain the entire abandonment distribution. The theory then tells us that such scaling would lead to more accurate asymptotics when, for example, the magnitude of the derivative of the hazard-rate function at the origin is large. However, *testing the empirical robustness of traditional QED* (equivalently here, testing the need for hazard-rate scaling) calls for estimation of this derivative, which brings back the previously mentioned *need for fitting or testing for a regime*.

2.3.3 *Forecasting*. Forecasting of arrivals has been well surveyed by Ibrahim et al. (2016a) for service systems with one arrival stream. However, ServNets more often than not have multiple arrival streams, that are furthermore dependent, which is a scenario studied in Ibrahim and L'Ecuyer (2013), Ye, Luedtke and Shen (2019). One must also study the impact of interstream dependence on operational decisions, such as agent pooling, in a similar spirit to what Gans et al. (2015) did for single-arrival stream systems. Then Gans et al. (2010) observe that service capacity changes across time due to learning, forgetting and agent attrition, which raises the need to forecast available service capacity (Azriel, Feigin and Mandelbaum, 2014). Waiting time, as the outcome of the interplay between arrivals, services and patience, must be forecasted as well. Delay prediction is presently an active research area; see Ibrahim and Whitt (2011), Senderovich et al. (2015), Dong, Yom-Tov and Yom-Tov (2018), and more is appearing regularly such as Ibrahim (2018). The challenge of prediction can escalate up to the prediction of a whole ServNet. Consider, for example, DFCI: here one seeks to predict complete network performance, given the appointment book of all resources.

2.3.4 *Model selection and post-model selection inference* (*PoSI*). The creation of a suitable ServNet naturally involves model selection, for example, selecting the "right" model for each building block of the ServNet, which shall then be validated and selected again if needed via an iteration process. Classically inference is made assuming that the model selected is correct, which unfortunately rarely is the case. PoSI, recently pioneered by Brown's group at the University of Pennsylvania (Berk et al., 2013), aims at the valid inference that incorporates the effects of model selection. Performing PoSI on ServNets is both an intriguing and challenging research direction.

2.3.5 *Validation*. Both Chen et al. (1988), Adler et al. (1995) are examples of data-based modeling. Their guiding principle prevails here as well: *Search, within a tractable family of ServNets, the one that is most appropriate for its needs*. It remains to specify the *"family"* (in particular, is it a QNet or FNet or DNet) and define *"appropriate."* This requires the development of a *validation framework*, where questions such as the following can be *formally* addressed: How is one to measure or test the *accuracy*, or better yet *robustness or practical value*, of an asymptotic approximation? What are the *sources of uncertainty in a model or its building blocks (*Whitt, 2002b*)?* And "how" to *fit*

a simple model (not too simple) to a complex service reality? There are pitfalls to beware of as mentioned in Whitt (2012) and, generally, this territory is relatively uncharted.

2.3.6 *Performance analysis of ServNets*. Many software tools have been developed for the performance analysis of QNets. As far as we know, only two utilize insights from queueing asymptotics (conventional heavy-traffic): Whitt's (1983) QNA and Dai, Yeh and Zhou's (1997) QNET, both focusing on steady-state. A third numerical tool by Kang and Pang (2013) has been developed for time-varying FNets under many-server scaling. It seems feasible to amalgamate the approaches in Whitt (1983), Dai, Yeh and Zhou (1997), Kang and Pang (2013) to cover QNets, FNets, and DNets. Each of these three could lead to a SimNet. Specifically, QNet-based simulations have been common practice, and many have been customized to hospitals and call centers: an example to follow is the Java suite in Chan and L'Ecuyer. However, *FNet- and DNet-based simulation of service networks* have not yet been attempted. Insights will be gained from the applications to Finance of DNet-based simulations (Glasserman, 2004) and PDEs (Muthuraman and Zha, 2008), as well as from general-purpose simulation theory for stochastic processes (Glynn and Iglehart, 1989).

2.3.7 *Why create ServNets automatically?* There are numerous applications for automatically created ServNets. Demonstrating with SimNets, one could have the following three applications in mind: as already mentioned above, and carried out for Erlang-R (Yom-Tov and Mandelbaum, 2014), SimNets will provide a virtual-reality environment against which other ServNets (QNets, FNets, DNets) can be validated; they will be used to *generate bootstrap ServNet samples* (Cowling, Hall and Phillips, 1996); and they are essential players in *simulation-based staffing and control*, as in Feldman et al. (2008), Feldman and Mandelbaum (2010).

2.3.8 *Broader relevance, in particular to network science*. The framework originating in Brown et al. (2005) and outlined in Section 2 has the potential to significantly affect research and teaching in other areas and disciplines that are central to service systems: marketing, information systems, psychology and human resource management. But the potential goes further, as demonstrated by the emerging field of "Networks" (Newman, 2018, Newman, 2008)—simply recall the ServNets in Figure 2. The mathematical and statistical theory of Networks (Kolaczyk, 2009) has mainly dealt with *static/structural* models of social, biological, information and technical networks, while acknowledging that *dynamic* models are important but their research is yet to mature. Service systems and their ServNets offer a novel area for Network Science and its application, which amplifies the importance of dynamic (*time-varying*) network models and their visualization (Bender-deMoll and McFarland, 2006).

An important research area is network tomography (Vardi, 1996) or inverse problems (Baccelli, Kauffmann and Veitch, 2009), which was referred to previously as completing partial information: for example, *how does one track patient flow within an emergency department, based on entries and intermediate milestones*. A final point is that the asymptotic focus of "Networks" has been large networks with local interactions, but its mathematical framework for limit theorems could prove useful for ServNets as well.

3. NONPARAMETRIC EMPIRICAL BAYES WITH BASEBALL DATA.

Brown has been known as a world-leading statistical decision theorist but he is also a data scientist in today's terminology. His profound understanding of theories powers applications. Brown (2008), published in one of the very early volumes of *The Annals of Applied Statistics*, is one such example. In an attempt to predict batting average for baseball players, he illustrates a golden procedure that a statistician or a data scientist should follow:

1. Identify a meaningful and useful problem with domain knowledge;
2. Start with a suitable working model when needed;
3. Propose methodologies to solve the problem;
4. Validate the entire procedure based on the data;
5. Last but not least make the data used publicly available.

Another focus of the paper lies in Brown's continuous effort to reveal properties of shrinkage methods in normal mean problems. Since Stein's (1956) first discovery of the inadmissibility of classical least squared estimator for multivariate normal means, numerous estimators have been proposed (James and Stein, 1961, Stein, 1962, Lindley, 1962, Strawderman, 1971, Strawderman, 1973, Efron and Morris, 1975, Efron and Morris, 1977). Shrinkage is the key property of these estimators. An empirical Bayes interpretation was developed in Efron and Morris (1973).

The properties of the shrinkage estimators are well studied and all perform similarly for normal mean problems in the homoscedastic case. Nevertheless, the optimal shrinkage scheme for the heteroscedastic setup is unclear to date. With this in mind, Brown provided an insightful study comparing several methods, which include naive standard methods, James–Stein estimator, empirical parametric Bayes, and nonparametric empirical Bayes.

All methods in Brown's paper are preceded by applying the variance-stabilizing transformation. The estimates are evaluated after transformed back into the original scale. Several measures of accuracy are proposed to evaluate the estimates. It is possible, however, to proceed without transforming the data first. Here, we start by reproducing the methods mentioned in Brown's paper. These methods can only provide point estimates. To measure the uncertainty of the estimates, we apply a nonparametric empirical Bayes solution developed in Cai and Zhao (2019) to provide not only point estimates, but also prediction intervals. All the point estimates show comparable accuracy.

In Section 3.1, we highlight Brown's paper. Section 3.2 describes the binomial empirical Bayes method developed in Cai and Zhao (2019). We compare all the methods and results are shown in Section 3.3.

3.1 Highlights from Brown (2008).

Batting ave**R**age (R) is an important metric to evaluate a player's performance. It is the proportion of successful attempts, "Hits" (H), as of the total number of qualifying attempts, "AtBat" (N), that is, $R = H/N$. Prediction of batting average allows the team to develop batting strategies.

3.1.1 *A binomial model*. It is reasonable to assume for each player $H_i \overset{\text{ind}}{\sim} \text{Bin}(N_i, p_i)$, where p_i characterizes the true, batting ability for player i. Assume the players ability is somewhat stable, that is, p_i is a constant, we can then use results from the earlier season to predict the performance in the later one. Using the monthly record of Major League players in 2005, Brown (2008) estimated p_i based on the first half season (from April through June) and then applied the estimate to predict and validate the performance in the second half season (July through September). To be more specific, for player i in season j we have independently

$$H_{ji} \sim \text{Bin}(N_{ji}, p_i), \tag{1}$$

where $j = 1, 2$ and $i = 1, \ldots, P_j$. The first step is to produce a good estimate for p_i.

3.1.2 *Variance stabilization*. When N is not too small and $H \sim \text{Bin}(N, p)$, H/N can be well approximated by a normal distribution with mean p and variance $p(1-p)/N$. Note that the variance involves unknown p but can be transformed using the variance-stabilizing transformation so that it only depends on the observed N. Considering a family of variance-stabilizing transformations,

$$Y^{(c)} = \arcsin\sqrt{\frac{H+c}{N+2c}} \tag{2}$$

for some constant c, $Y^{(c)}$ follows approximately a normal distribution with a stabilized variance

$$\text{Var}\big(Y^{(c)}\big) = \frac{1}{4N} + O\bigg(\frac{1}{N^2}\bigg). \tag{3}$$

The choice of $c = 1/4$ minimizes the approximation error of the mean, that is,

$$E\big[Y^{(c)}\big] = \arcsin\sqrt{p} + O\bigg(\frac{1}{N^2}\bigg) \tag{4}$$

and thus

$$\sin^2\big(E\big[Y^{(c)}\big]\big) = p + O\bigg(\frac{1}{N^2}\bigg). \tag{5}$$

Combining (2)–(4), we have

$$Y^{(1/4)} \sim \mathcal{N}\bigg(\arcsin\sqrt{p}, \frac{1}{4N}\bigg). \tag{6}$$

The above layout allows us to use $Y^{(1/4)}$ to estimate $\arcsin\sqrt{p}$ as an unknown normal mean after which the estimator will be transformed back through (5).

3.1.3 *Methods considered in Brown (2008)*. Following the setup above, we apply the variance-stabilizing transformation to the performance for the first-half of the season to estimate p_i. Let X_{1i} be the transformed batting average for player i in the first-half of the season, that is,

$$X_{1i} = \arcsin\sqrt{\frac{H_{1i}+1/4}{N_{1i}+1/2}}.$$

By (5), X_{1i} can be accurately treated as independent random variables with the distribution

$$X_{1i} \sim \mathcal{N}\big(\theta_i, \sigma_{1i}^2\big), \tag{7}$$

where $\theta_i = \arcsin\sqrt{p_i}$ and $\sigma_{1i}^2 = 1/(4N_{1i})$.

We now describe four approaches considered in Brown (2008), and an alternative using maximum likelihood nonparametric empirical Bayes in Section 3.2.

1. *Naive estimator*: The baseline is a naive estimator that uses the first-half performance to predict the second-half performance.

2. *James–Stein estimator*: The following extended James–Stein estimator is used to accommodate the heteroscedastic setting.

$$\hat{\theta}_i = \hat{\mu} + \left(1 - \frac{P_1 - 3}{\sum (X_{1i} - \hat{\mu})^2/\sigma_{1i}^2}\right)_+ (X_{1i} - \hat{\mu}), \tag{8}$$

where $\hat{\mu} = (\sum X_{1i}/\sigma_{1i}^2)/(\sum 1/\sigma_{1i}^2)$.

3. *Parametric empirical Bayes*: Various parametric Bayes estimators have been investigated. The classical approach imposes a normal prior $\theta_i \sim \mathcal{N}(\mu, \tau^2)$ and estimates the hyperparameters by the method of moments or maximum likelihood, which yields the parametric empirical Bayes estimator,

$$\hat{\theta}_i = \hat{\mu} + \frac{\hat{\tau}^2}{\hat{\tau}^2 + \sigma_{1i}^2}(X_{1i} - \hat{\mu}).$$

Other methods have been proposed after Brown (2008). Xie, Kou and Brown (2012) developed a SURE shrinkage estimator for the heteroscedastic case and Weinstein et al. (2018) proposed a group-linear Bayes estimator that exploits the dependence between true means and sampling variances by grouping observations with similar variances. It is also possible to directly work on the original scale. For instance, Muralidharan (2010) proposed a binomial mixture model using a Beta prior.

4. *Nonparametric empirical Bayes*: Parametric priors are mathematically tractable but may not best approximate the distribution of θ. Nonparametric priors provide flexibility to capture the heterogeneity in the unknown parameters of interest. We focus on a few variants of the nonparametric Bayes method here.

Let G denote an unknown distribution and assume $\theta_i \sim G$ independently, then the posterior mean is

$$E(\theta_i|X) = \frac{\int \theta_i \phi(\frac{X_i - \theta}{\sigma_i}) G(d\theta)}{\int \phi(\frac{X_i - \theta}{\sigma_i}) G(d\theta)},$$

where ϕ denotes the standard normal density. Using Formula 1.2.2 in Brown (1971), we can indirectly obtain the posterior mean

$$\theta_G(X)_i = X_i + \sigma_i^2 \frac{\frac{\partial}{\partial X_i} g_i(X)}{g_i(X)}, \tag{9}$$

where $g_i(X) = \int \phi(\frac{X_i - \theta}{\sigma_i}) G(d\theta)$ is the marginal density of X and can be estimated by kernel estimators. Brown and Greenshtein (2009) proposed an estimator that adapts well to both sparse and nonsparse cases based on this formulation.

Along this line, Raykar and Zhao (2010) incorporated a sparsity-adaptive mixture prior. Jiang and Zhang (2009), Jiang and Zhang (2010) developed a general maximum likelihood empirical Bayes method for the homoscedastic case that directly estimates the prior G and extended it to the heteroscedastic case with known variance. Koenker and Mizera (2014), Gu and Koenker (2017) recast the nonparametric maximum likelihood problem as a convex optimization problem and solved by interior point method, which gained considerable speed, and further improved the estimate by incorporating covariates and longitudinal data. Dicker and Feng (2016) further bounded the approximation error. All the methods above focus on point estimation.

3.2 The Binomial Empirical Bayes Approach.

All the above mentioned methods use standard variance-stabilizing transformations which transform a binomial distribution into a normal distribution with known variance, and at the same time achieve the best control of bias asymptotically.

Brown (2008) focuses on batters having more than 10 at-bats since the transformed data is well-approximated by a normal distribution when N is larger than 10. Since batters with more than 20 at-bats account for a relatively large population (91%), the binomial data can be approximated by the normal distribution reasonably well.

In particular, we model the batting average R_i as follows:

$$R_i = \frac{H_i}{N_i} \sim \mathcal{N}(p_i, \sigma_i^2), \quad \text{independent.} \tag{10}$$

We directly estimate the variance

$$\hat{\sigma}_i = \sqrt{p_i(1 - p_i)/N_i}.$$

The variance can also be estimated using a more robust plug-in estimator such as the mean absolute deviation.

Again, assume a nonparametric prior on $p_i \sim G$ independently, where G denotes an unknown distribution. The unknown prior can be estimated using the data. We then compute the posterior distribution of p_i using Bayes' theorem. Point estimates, such as the posterior mean, credible intervals and predictive intervals are readily followed.

Alternatively, we can adopt the independent binomial model $H_i \sim \text{Bin}(N_i, p_i)$ and follow the same recipe to estimate p_i.

3.2.1 *Nonparametric empirical Bayes via maximum likelihood*. In this section, we will describe our procedure based on normal approximation (10). Since a fully nonparametric prior is infinite-dimensional, we estimate the nonparametric prior G of batting probability by discretizing its domain into a fine equal-length mesh with M grid points $\{\tau_j\}_{j=1}^M$ supported on the range of the observed data. We will estimate $\pi_j = G(\tau_j)$.

To be more specific, we approximate the nonparametric prior as a multinomial distribution,

$$\pi(p_i \mid \boldsymbol{\pi}) \sim \text{Multinomial}(\boldsymbol{\pi}), \quad \text{independent},$$

where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_M)$ and $\sum \pi_j = 1$. Consider $\boldsymbol{\pi}$ as the hyperparameter estimated by maximizing the conditional marginal log-likelihood of $\{R_i\}$ given $\boldsymbol{\pi}$,

$$\hat{\boldsymbol{\pi}} = \operatorname*{argmax}_{\boldsymbol{\pi}} \sum_{i=1}^{n} \log\left(\sum_{j=1}^{M} \varphi\big((R_i - p_j)/\sigma_i\big) \cdot \pi_j \right), \tag{11}$$

where φ is the standard normal density. The maximizer can be obtained by a modified EM algorithm originally proposed in Jiang and Zhang (2009). We adopted early stopping to avoid overfitting.

The posterior distribution can be obtained through

$$\pi\big(p \mid R_i, \hat{\pi}, \sigma_i^2\big) = \frac{\varphi((R_i - p)/\sigma_i)\hat{\pi}(p)}{\sum_{j=1}^{M} \varphi((R_i - \tau_j)/\sigma_i)\hat{\pi}(\tau_j)}, \tag{12}$$

where $\hat{\pi} = \pi(p|\hat{\boldsymbol{\pi}})$. The posterior mean

$$\hat{p}_i(R_i, \sigma_i) = \sum_j \tau_j \pi\big(\tau_j | R_i, \hat{\pi}, \sigma_i^2\big) \tag{13}$$

is used as the estimator of p and $(1-\alpha)$ credible intervals are readily followed.

We further obtain the predictive distribution and construct a predictive interval for a future R_i^*,

$$\begin{aligned} &\pi_{\text{pred}}\big(R_i^* \mid R_i, \hat{\pi}, \sigma_i^2\big) \\ &\quad = \sum_j \varphi\big((R_i^* - \tau_j)/\sigma_i\big)\pi\big(\tau_j | R_i, \hat{\pi}, \sigma_i^2\big). \end{aligned} \tag{14}$$

As a remark, the procedure to obtain point estimates based on the binomial model is similar but instead uses a different likelihood. For details, see Cai and Zhao (2019).

3.3 Prediction with 2005 Baseball Data.

Using the monthly batting records for all Major League baseball players in the 2005 season as Brown (2008), we predict the second-half season performance using the first-half. Note that we only use batters with at-bats more than 10 in the first-half to estimate and batters with at-bats more than 10 in both the first- and second-half for validation. We compare the above-mentioned nonparametric empirical Bayes estimators against the naive estimator and the extended James–Stein estimator. Specifically:

1. Naive: $\delta_0 = R_{1i}$;
2. James–Stein: the extended James–Stein estimator (8);
3. NPEB: Brown's (2008) nonparametric empirical Bayes method via variance-stabilizing transformation (7);
4. NPEB_Normal_Approx: the nonparametric empirical Bayes method described in Section 3.2 with direct normal approximation (10) using (9) to estimate the posterior mean;
5. NPEBML_Normal: the nonparametric empirical Bayes via maximum likelihood (11);
6. NPEBML_Binomial: the nonparametric empirical Bayes via maximum likelihood with the binomial model (1).

The accuracy of the point-estimates is measured by the normalized total squared error as follows. Denote $S_j = \{N_{ji} \geq 11\}$ where $j = 1, 2$, The Total Squared Error (TSE) for $\hat{R}$ is defined as

$$\widehat{TSE}_R[\hat{R}] = \sum_{i \in S_1 \cap S_2} \left[(R_{2i} - \hat{R}_{2i})^2 - \frac{R_{2i}(1 - R_{2i})}{N_{2i}} \right].$$

TSE is an unbiased estimator of the predictive risk. To compare different estimators, we use the naive estimator δ_0 as a baseline and normalize each total squared error by the total squared error of δ_0, that is,

$$\widehat{TSE}_R^*[\hat{R}] = \frac{\widehat{TSE}_R[\hat{R}]}{\widehat{TSE}_R[\delta_0]}.$$

Table 1 reports the normalized total squared error of the six methods. The first block reports the normalized total squared error of the naive estimator and the James-Stein estimator; the second block reports NPEB and NPEB_ Normal_Approx, the nonparametric empirical Bayes estimators using (9); the third block reports NPEBML_Normal that uses the normal approximation (10) and NPEBML_Binomial that uses the binomial model (1).

All nonparametric empirical Bayes estimators achieve comparable performance in terms of TSE. Note that the estimate of NPEB is transformed back using $\hat{R} = \sin^2(\hat{\theta})$ while others use the data on the original scale. The comparable performance justifies the direct normal approximation of the binomial distribution.

TABLE 1
$\widehat{TSE}_R^*$ *of half-season predictions for all batters*

	Naive	James–Stein	NPEB	NPEB_Normal_Approx	NPEBML_Normal	NPEBML_Binomial
$\widehat{TSE}_R^*$	1	0.540	0.51	0.536	0.599	0.645

3.3.1 *Credible and predictive intervals*. In addition to a point estimate of p_i, the nonparametric empirical Bayes method can also provide both a credible interval for p_i and predictive interval for $R_{2,i}^*$ using the equal quantiles on both sides of the posterior distribution (12) and the predictive distribution (14), respectively.

Figure 7(a) shows the 95% predictive intervals based on the normal approximation model (10) for 50 randomly selected players. We check the empirical coverage of the known R_{2i}'s. The green solid lines are the predictive intervals that cover the observed R_2 (green dots); and the red dashed lines are those that do not cover the observed R_2 (red cross). The average coverage is 93% and the average width of the predictive intervals is 0.19.

Figure 7(b) displays the 95% credible intervals of p_i's for 50 randomly selected players. We will not be able to check the coverage probability for the unknown true p_i's here.

4. OFFICIAL STATISTICS.

In 1790, James Madison, one of the founding fathers of the United States, said, "... in order to accommodate our laws to the real situation of our constituents, we ought to be acquainted with that situation" (Madison, 1790). At this point, the United States was a nascent country, embarking on its first census. Madison's opinion was that the country could be effectively governed only if the administration knew about its people and their activities.

The federal statistical system has grown markedly since then. In modern times, it can tell us that the midwestern state of Ohio had 11,658,609 residents in 2017, according to the Population Estimates Program U.S. Census Bureau (2017a). From just the 2017 American Community Survey, we can also discover (with margins of error corresponding to 90% confidence):

- The median age in Ohio was 39.8 ± 0.1 years (U.S. Census Bureau, 2017b);
- 8,830,185 ± 11,095 people were U.S. citizens of voting-age (U.S. Census Bureau, 2017c);
- Among those with a Bachelor's degree and at least 25 years old, 30.7% ± 0.5% have earned a degree in science or engineering (U.S. Census Bureau, 2017d);
- 2.1% ± 0.1% of the working population 16 years and older walked to work (U.S. Census Bureau, 2017e); and

FIG. 7. 95% *predictive intervals of* R_{2i}^* *and credible intervals of* p_i *for randomly selected* 50 *players*.

- 3.6% ± 0.3% of adult civilians in the state were World War II veterans (U.S. Census Bureau, 2017f).

In addition, this system yielded maps outlining geographies from states to school districts (see Figure 10(a) for an example) and unique labels for those areas through Federal Information Processing Standard (FIPS) codes. These innovations standardized information processing and reporting. Reams of paper (now bytes) are devoted to tables and descriptions and explanations of this data. All of it is available to the public for free, regarded as a public good (Citro, 2016, Eberstadt et al., March 2017).

These are official statistics, numbers produced by governments (and other large organizations). With such a large population and a wide range of surveys, official statistics are probably one of the first examples of big data and the use of technology for tabulation and analysis.

The history of the U.S. federal statistical system can offer guidance on how to think about three key issues in data science today: (a) using data for purposes beyond originally intended, (b) protecting data privacy and confidentiality, and (c) effectively using data which is not, by its very nature, a random sample. (See Citro (2016), Groves (2011), Pfeffermann (2015) for more about the U.S. federal statistical system.)

Brown was an integral part of this system through his testimony to the U.S. Congress in 1997 and 1998 about the 2000 U.S. Census and his participation on numerous National Research Council panels (e.g., Panel to Evaluate National Statistics on Research and Development; Panel on Coverage Evaluation and Correlation Bias in the 2010 Census; Chair of the Committee on National Statistics). This service spanned over 30 years.

4.1 Repurposing Data.

As noted, the first census in the United States took place in 1790. It was carried out by many men riding around on horseback, knocking on doors, and counting the people (and slaves) inside. The logistics were complicated even then, although the information collected was not.

Required by the U.S. Constitution (Article 1, Section 2) every 10 years (and so, "decennial census"), the enumeration of every resident was needed primarily to determine how many representatives each state sent to the House of Representatives, part of the U.S. Congress. This process is called apportionment. (The next round of enumeration will occur in 2020.)

Schedule of the whole number of persons within the division allotted to A. B.

Names of heads of families.	Free white males of 16 years and upwards, including heads of families.	Free white males under 16 years.	Free white females, including heads of families.	All other free persons.	Slaves.

FIG. 8. *Example of a schedule for the first census in* 1790. *Native Americans who were not taxed by the government were excluded from the count*. (*Source*: *U.S. Congressional Record* (*U.S. Census Bureau*, 1907).)

The first census, however, served two other functions: (a) to obtain the number of potential army recruits and (b) to determine how to tax the population to pay for the American Revolutionary War (Nagaraja, 2019). Both impacted the census in different ways.

Figure 8 shows the type of information enumerators collected during the 1790 census. There are two columns for free white males: one for those sixteen and older and another for those younger than sixteen. Males in the former group were eligible to join the army. In this case, the census schedule was altered to incorporate data collection beyond what was required for apportionment calculations.

This practice of adding extra questions to the census continued throughout much of U.S. history, culminating in the "long form" of the census, sent to a small percentage of households. (Everyone else received the usual "short" census form.) In 2005, however, the American Community Survey (ACS) and its Puerto Rican equivalent replaced the long form. These surveys are administered using a rolling sample design, allowing the U.S. Census Bureau to publish more current data about the U.S. population. Now, every year, roughly 3.5 million households are surveyed through the ACS (Torrieri et al., 2014).

As shown in Figure 9, the first census in 1790 recorded around 3.9 million free and enslaved people in the United States; the twenty-third census in 2010 counted more than 308 million (free) people (U.S. Office of the Secretary of State, 1793). At the time of the first census, however, the government felt that the 3.9 million count seemed too low. They even suspected that households pretended to have smaller families to avoid the taxes associated with paying for the recent war (Wright and Hunt, 1900).

While it was eventually decided that it was the colonial-era population estimates which were too high instead of the census count being too low, this high-

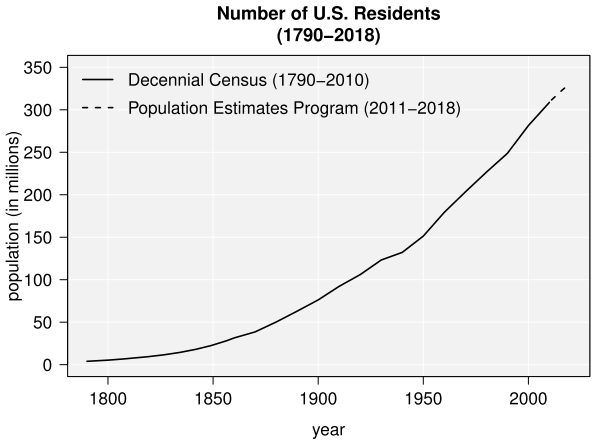


FIG. 9. *Resident population of the U.S. based on the decennial census starting in* 1790 *and the Population Estimates Program for the recent, intercensal years*, 2011–2018. *The* 1790 *census counted people from the* 13 *states and the territorial areas which eventually became states*. (*Sources*: *Decennial census of population and housing*; *U.S. Census Bureau*, 2017a).

lights an important point. The primary purpose of the census was (and still is) for apportionment. However, at that time, people worried that a secondary purpose—a repurposing—dampened the final counts. If that had happened, then an auxiliary objective would have degraded the constitutionally mandated one. Consequently, it is possible for repurposing to be damaging to the original goal.

4.2 Data Confidentiality.

During the 2010 census, the U.S. Census Bureau launched a $133 million campaign to encourage the public to complete their census forms. Reasons for failing to respond ranged from apathy to hostility toward government data collection (U.S. Census Bureau, 2010). The latter has always been a concern.

Charles Pidgin, Chief Clerk of the Massachusetts Bureau of the Statistics of Labor, wrote in 1888, "The American citizen is jealous of his individual rights and opposed, on principle, to inquisitorial inquiries by the government. He is not so much opposed to giving information of a private nature, but he is very solicitous as regards the use to be made of the information. He will give a statistical office individual facts but he wishes, naturally, to be 'covered up' in the print (Pidgin, 1888, p. 3)." He goes on to recommend that the questionnaire schedules themselves should explain the legal basis for collecting data. (U.S. Census Bureau forms do include such explanations.)

Pidgin address two issues here. First, published data should not reveal any individual's information. The U.S. Census Bureau handles this problem using disclosure avoidance techniques: data swapping, synthetic data, top-coding and so forth (Lauger, Wisniewski and McKenna, 2014).

Pidgin hints at the second issue when speaking about the use of that information. Governments are made up of many agencies and departments and bureaus. Not only does the public want to be assured that their personal information will not be printed, but they also do not want it used against them by the government.

The only way to formally ensure this is through the law. The U.S. Census Bureau uses Title 13 and Title 26 of the United States Code to protect respondent confidentiality. Title 13 makes it a federal crime to disclose individual (or business) information to the public and to other government agencies or courts (U.S. Census Bureau, 2017g). Title 26 is more specific and covers the use of Internal Revenue Service (IRS) data (i.e., tax information) by the U.S. Census Bureau to produce official statistics (U.S. Census Bureau, 2017h). Again, strict privacy protections are in place.

Terrible things can happen when confidentiality is violated. For example, it has been shown that census records were used to identify and intern Japanese residents (many of them U.S. citizens) during World War II. This could occur because privacy protections were temporarily suspended by then-President Franklin D. Roosevelt, and only reinstated after the war (Anderson, 2015).

Privacy and confidentiality are becoming increasingly important as it becomes easier to collect data about a person's daily activities through cell phones, internet activity, cameras, etc. This data is then sold and disparate data sets are connected, allowing a business (or a government) to track a person across multiple devices through time and space (another statistical idea pops up here: record linkage). Relying on organizations to self-regulate with regards to privacy may be too much to ask; the law may be the only way forward. Europe, for example, has taken the lead with rules on the "right to be forgotten" and the General Data Protection Regulation (GDPR).

4.3 Nonrandom Samples.

Much of "big data" is opportunistic in that it is collected, for example, as a byproduct of doing business (e.g., scanner data, Google searches, Uber trips, Netflix

shows watched). Groves (2011) offers a useful term for this type of data: "organic data."

The upside of organic data is that there is a lot of it. One downside is that it is very messy. This contributes to another critical disadvantage: it is often unclear what part of the population the data represents. These are not random samples nor are they censuses, rendering traditional statistical techniques for inference ineffectual.

That said, statistical thinking can still provide guidance. For instance, the difference between the sample and population mean is a product of data quality, data quantity, and problem difficulty (i.e., standard deviation) (Meng, 2018). The data quality element is the component that, in part, allows one to understand the representativeness of the sample.

Official statistics offer an accepted use of a nonrandom sample: house price indices. These indices are intended to measure the state of the housing market. As Brown (2015) in his comment to Pfeffermann's (2015) Morris Hansen Lecture observes, house price indices are most commonly constructed using sale prices. In any given time period (e.g., month, quarter) only a small fraction of homes in a region are sold. An example in Nagaraja (2019) shows that for roughly 733,600 single-family homes in the San Diego metropolitan area, only 2.8% were sold in 2011. Therefore, any price index based on sale prices utilizes a very small, nonrandom subset of all single-family homes (excluding apartments, rental properties, mobile homes, etc.).

The first house price indices were hedonic indices that regressed price against housing characteristics (i.e., hedonics) such as the number of bedrooms, bathrooms, lot size and so forth (Nagaraja and Brown, 2013, Nagaraja, Brown and Wachter, 2014). This type of index was eventually rejected as hedonic variables were difficult to collect consistently over long periods of time and across large geographic areas. (Perhaps they may return in the current, web-scraping age.)

Bailey, Muth and Nourse (1963) proposed the repeat sales index as a replacement for the hedonic index. Their idea was that the difference in price between two sales of the same house could be used to construct an index. ("Two sales of the same house" was shortened to "repeat sales.") Using repeat sales would sidestep the issue of hedonics because, in theory, the same house—with the same hedonic characteristics—would be compared.

Using the notation from Nagaraja and Brown (2013), the Bailey, Muth and Nourse (1963) model can be expressed as follows:

$$y_{it'} - y_{it} = \beta_{t'} - \beta_t + u_{itt'}, \tag{15}$$

where y_{it} is the log sale price of house i at time t, β_t is the log price index at time t, $u_{itt'}$ is the Gaussian error term, and $t' > t$. The resulting $\hat{\beta}_t$ values are converted to a price index with $t = 1$ being the base period: 1, $e^{\hat{\beta}_2 - \hat{\beta}_1}$, $e^{\hat{\beta}_3 - \hat{\beta}_1}$, …, $e^{\hat{\beta}_T - \hat{\beta}_1}$.

More than two decades later, the model in (15) resurfaced in Case and Shiller (1987, 1989) with an additional component. Case and Shiller (1987, 1989) argued that the gap time between sales of the same house was important. Specifically, the wider the gap (e.g., 10 years between sales vs. 5 years between sales), the less valuable the older sale price. This information was incorporated into (15) by introducing a heteroscedastic error term. A modified weighted-least squares method was used to fit the model where wider gap times resulted in lower observation weights (Nagaraja, Brown and Wachter, 2014).

This repeat sales model became very popular and is used in two prominent indices today: the S&P CoreLogic Case–Shiller Home Price Indices and the Federal Housing and Finance Agency (FHFA) House Price Index. The weights used in the FHFA index vary slightly from the Case and Shiller (1987, 1989) method and the index incorporates an adjustment for the depreciating effect of age on a house. Both indices make adjustments for seasonal effects (Calhoun, 1996).

Figure 10(b) shows an example of the quarterly and seasonally adjusted FHFA index for five Ohio Core Based Statistical Areas (metropolitan/micropolitan areas): Akron, Cincinnati, Cleveland-Elyra, Columbus and Dayton. All five indices have the same base, the first quarter (Q1) of 1991. We can see that prices in Dayton rose more slowly than the other four regions, which had comparable growths. Prices fell in all five regions after the 2007 housing bubble burst (see the vertical line). After that point, the indices diverge with the capital, Columbus, having the fastest relative change in prices and Dayton remaining as the area with the slowest relative change in prices.

A repeat sales index is a reasonable solution to a problem of limited data. However, there are a few issues, even though they have been adopted for widespread use. Primarily, the limited data issue is worsened with repeat sales methods as all homes which sold only once in the sample period are discarded.

For example, in a study of 20 U.S. cities with sales covering a nearly 20-year period, between 53% and 79% of homes in the sample period were only sold once (Nagaraja, Brown and Zhao, 2011). Those homes

(a) **CBSAs in Ohio**

(b) **FHFA Seasonally Adjusted Quarterly House Price Indices (Q1 1991 - Q3 2018; base period = Q1 1991)**

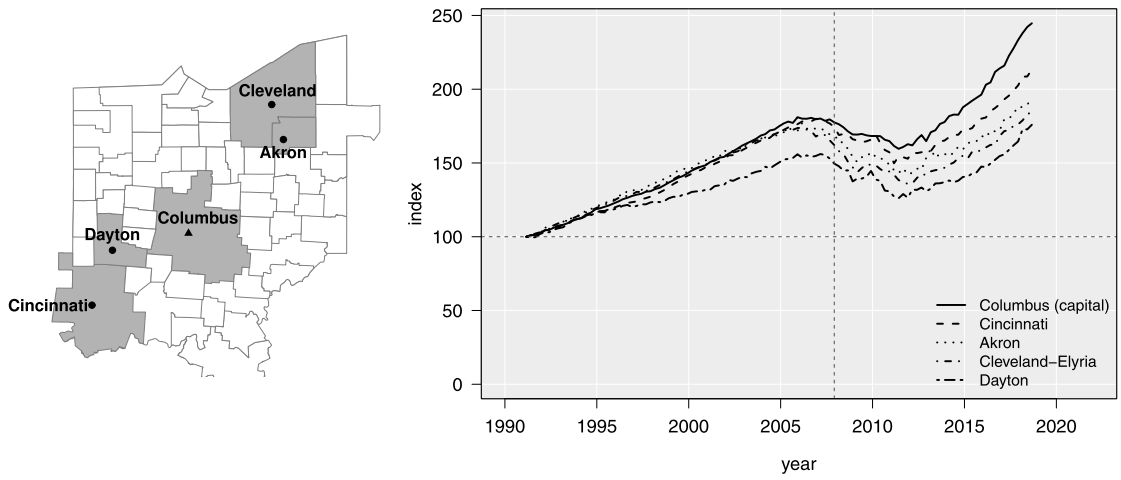


FIG. 10. *Graph* (*a*) *is a map of the state of Ohio demarcated by Core Based Statistical Areas* (*CBSAs*) *with major cities marked* (*Columbus is the state capital*). *Some CBSAs*, *like Cincinnati*, *cross state boundaries*. *Graph* (*b*) *shows the Federal Housing and Finance Agency House Price Indices for these CBSAs*. *The base year is the first quarter in* 1991 *where the index is* 100; *the repeat sales index is quarterly and has been seasonally adjusted*. *The burst of the housing bubble in* 2007 *is marked by the vertical dotted line*. (*Sources*: *Federal Housing and Finance Agency and Geography Division*, *U.S. Census Bureau* (*U.S. Census Bureau*, 2009, *Federal Housing and Finance Agency*, 2019)).

would not appear in the index until they had been sold a second time. Among those single sales, some would be new home sales. That means, in theory, two cities could have the same house price index even though one was growing far more rapidly and constructing new homes at a feverish pace. A second issue is that there is some evidence that homes sold many times may be cheaper than those sold less frequently, further biasing the sample (Nagaraja, Brown and Wachter, 2014).

To address these issues, an alternate, hybrid index was proposed in Nagaraja, Brown and Zhao (2011). This index has the benefit of being able to incorporate both single and repeat sales; furthermore, the effect of gap times are included through the use of a decaying autoregressive parameter and in the variance of the error term.

To borrow notation from Nagaraja, Brown and Zhao (2011), let $1, \ldots, T$ denote the (discrete) time periods (e.g., months, quarters, years) in the sample period. Further, let $y_{i,1,z}, y_{i,2,z}, \ldots, y_{i,j,z}, \ldots$ be the log sale price of the jth sale of the ith house in area z (e.g., ZIP code, census tract) within the sample period. Then let $\gamma(i,j,z) = t(i,j,z) - t(i,j-1,z)$ denote the gap time between two consecutive sales where $t(i,j,z)$ is the time period when the jth sale of the ith house in area z occurred.

Then the autoregressive model for log price is

$$
\begin{aligned}
&y_{i,1,z} = \mu + \beta_{t(i,1,z)} + \tau_z + \varepsilon_{i,1,z}, \quad \text{(i.e. } j = 1\text{)};\\
&y_{i,j,z} = \mu + \beta_{t(i,j,z)} + \tau_z\\
&\qquad + \phi^{\gamma(i,j,z)}(y_{i,j-1,z} - \mu - \beta_{t(i,j-1,z)} - \tau_z)\\
&\qquad + \varepsilon_{i,j,z}, \quad j > 1\\
&\text{subject to } \sum_{t=1}^{T} n_t \beta_t = 0,
\end{aligned}
\tag{16}
$$

where μ is a constant, n_t is the number of sale at time t, μ is the overall average log price, $\beta_{t(i,j,z)}$ are the fixed effects for time period, ϕ is the autoregressive coefficient where $|\phi| < 1$, $\tau_z \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\tau^2)$ is a random effect, $\varepsilon_{i,1,z} \sim \mathcal{N}(0, \frac{\sigma_\varepsilon^2}{1-\phi^2})$, and for $j > 1$, $\varepsilon_{i,j,z} \sim \mathcal{N}(0, \frac{\sigma_\varepsilon^2(1-\phi^{2\gamma(i,j,z)})}{1-\phi^2})$ with independent $\varepsilon_{i,j,z}$ terms. The fitted $\hat{\beta}_t$ are then converted to an index using the same method as with the Bailey, Muth and Nourse (1963) index.

The first equation in (16) handles single sales and the second harnesses the extra information contained in having repeated sales. However, while the autoregressive index uses more of the sale data, it still is not constructed from a random sample of all homes. Nonetheless, these types of house price indices are an accepted way of describing the state of the housing market.

5. SUMMARY.

There are countless links between statistics and data science. The cases in this paper try to highlight the use of statistical thinking to address a few of these current debates in data science. Moreover, these examples emphasize the need to understand the nature of the data and its connection to the population along with a need to interpret the fitted model. These are two ideas that underpin much of statistical thinking and much of Lawrence D. Brown's work.

ACKNOWLEDGMENTS

We would like to thank the Editor, the Associate Editor and the referee for the opportunity to write this paper and their constructive comments. Mandelbaum's work has been partially supported by the Israeli BSF grant 2014180 and ISF Grants 357/80 and 1955/15. Shen's work is partially supported by the Ministry of Science and Technology Major Project of China 2017YFC1310903, University of Hong Kong Stanley Ho Alumni Challenge Fund, HKU University Research Committee Seed Funding Award 104004215 and BRC Fund. Zhao's work is partially supported by NSF Grant DMS-1512084.

REFERENCES

ADLER, P. S., MANDELBAUM, A., NGUYEN, V. and SCHWERER, E. (1995). From project to process management: An empirically-based framework for analyzing product development time. *Manage. Sci.* **41** 458–484.

ALDOR-NOIMAN, S., FEIGIN, P. D. and MANDELBAUM, A. (2009). Workload forecasting for a call center: Methodology and a case study. *Ann. Appl. Stat.* **3** 1403–1447. MR2752140

ANDERSON, M. (2015). *The American Census: A Social History*, 2nd ed. Yale University Press, New Haven.

ARMONY, M., ISRAELIT, S., MANDELBAUM, A., MARMOR, Y. N., TSEYTLIN, Y. and YOM-TOV, G. B. (2015). On patient flow in hospitals: A data-based queueing-science perspective. *Stoch. Syst.* **5** 146–194. MR3442392

AZRIEL, D., FEIGIN, P. and MANDELBAUM, A. (2014). Erlang-S: A data-based model of servers in queueing networks. *Manage. Sci.* **65** 4607–4635.

BACCELLI, F., KAUFFMANN, B. and VEITCH, D. (2009). Inverse problems in queueing theory and Internet probing. *Queueing Syst.* **63** 59–107. MR2576007

BAILEY, M. J., MUTH, R. F. and NOURSE, H. O. (1963). A regression method for real estate price index construction. *J. Amer. Statist. Assoc.* **58** 933–942.

BENDER-DEMOLL, S. and MCFARLAND, D. A. (2006). The art and science of dynamic network visualization. *J. Soc. Struct.* **7** 1–38.

BERK, R., BROWN, L. D., BUJA, A., ZHANG, K. and ZHAO, L. (2013). Valid post-selection inference. *Ann. Statist.* **41** 802–837. MR3099122

BORST, S., MANDELBAUM, A. and REIMAN, M. I. (2004). Dimensioning large call centers. *Oper. Res.* **52** 17–34. MR2066238

BRAMSON, M. (1998). State space collapse with application to heavy traffic limits for multiclass queueing networks. *Queueing Syst.* **30** 89–148. MR1663763

BROWN, L. D. (1971). Admissible estimators, recurrent diffusions, and insoluble boundary value problems. *Ann. Math. Stat.* **42** 855–903. MR0286209

BROWN, L. D. (2008). In-season prediction of batting averages: A field test of empirical Bayes and Bayes methodologies. *Ann. Appl. Stat.* **2** 113–152. MR2415597

BROWN, L. D. (2015). Comments on "Methodological issues and challenges in the production of official statistics." *J. Surv. Statist. Methodol.* **3** 478–481.

BROWN, L. D. and GREENSHTEIN, E. (2009). Nonparametric empirical Bayes and compound decision approaches to estimation of a high-dimensional vector of normal means. *Ann. Statist.* **37** 1685–1704. MR2533468

BROWN, L., GANS, N., MANDELBAUM, A., SAKOV, A., SHEN, H., ZELTYN, S. and ZHAO, L. (2005). Statistical analysis of a telephone call center: A queueing-science perspective. *J. Amer. Statist. Assoc.* **100** 36–50. MR2166068

CAI, J. and ZHAO, L. (2019). Nonparametric empirical Bayes method for sparse noisy signals. Preprint.

CALHOUN, C. (1996). OFHEO House Price Indices: HPI Technical Description. Available at https://www.fhfa.gov/PolicyProgramsResearch/Research/Pages/HPI-Technical-Description.aspx.

CASE, K. E. and SHILLER, R. J. (1987). Prices of single-family homes since 1970: New indexes for four cities. *N. Engl. Econ. Rev.* **Sept/Oct** 45–56.

CASE, K. E. and SHILLER, R. J. (1989). The efficiency of the market for single family homes. *Am. Econ. Rev.* **79** 125–137.

CHAN, W. and L'ECUYER, P. CCOptim: Call Center Optimization Java Library. Available at http://simul.iro.umontreal.ca/contactcenters/index.html.

CHEN, N., LEE, D. and SHEN, H. (2018). Can Customer Arrival Rates Be Modelled by Sine Waves? Submitted. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3125120

CHEN, H. and YAO, D. D. (2001). *Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization, Stochastic Modelling and Applied Probability. Applications of Mathematics* (*New York*) **46**. Springer, New York. MR1835969

CHEN, H., HARRISON, J. M., MANDELBAUM, A., VAN ACKERE, A. and WEIN, L. (1988). Empirical evaluation of a queueing network model for semiconductor wafer fabrication. *Oper. Res.* **36** 202–215.

CITRO, C. F. (2016). The US federal statistical system's past, present, and future. *Annu. Rev. Stat. Appl.* **3** 347–373.

COWLING, A., HALL, P. and PHILLIPS, M. J. (1996). Bootstrap confidence regions for the intensity of a Poisson point process. *J. Amer. Statist. Assoc.* **91** 1516–1524. MR1439091

DAI, J. G. and HE, S. (2010). Customer abandonment in many-server queues. *Math. Oper. Res.* **35** 347–362. MR2674724

DAI, J. G., YEH, D. H. and ZHOU, C. (1997). The QNet method for re-entrant queueing networks with priority disciplines. *Oper. Res.* **45** 610–623.

DAVENPORT, T. H. and PATIL, D. J. (2012). Data Scientist: The Sexiest Job of the 21st Century. *Harvard Business Review.*

DEO, S. and LIN, W. (2013). The impact of size and occupancy of hospital on the extent of ambulance diversion: Theory and evidence. *Oper. Res*. **61** 544–562.

DICKER, L. H. and ZHAO, S.D. (2016). High-dimensional classification via nonparametric empirical Bayes and maximum likelihood inference. *Biometrika* **103** 21–34.

DONG, J., YOM-TOV, E. and YOM-TOV, G. B. (2018). The impact of delay announcements on hospital network coordination and waiting times. *Manage. Sci*. **65** 1969–1994.

EBERSTADT, N., NUNN, R., SCHANZENBACK, D. W. and STRAIN, M. R. (2017). "In order that they might rest their arguments on facts": The vital role of government-collected data. The Hamilton Project at Brookings and the American Enterprise Institute.

EFRON, B. and MORRIS, C. (1975). The efficiency of logistic regression compared to normal discriminant analysis. *J. Amer. Statist. Assoc*. **70** 311–319.

EFRON, B. and MORRIS, C. (1973). Stein's estimation rule and its competitors—an empirical Bayes approach. *J. Amer. Statist. Assoc*. **68** 117–130. MR0388597

EFRON, B. and MORRIS, C. (1977). Stein's paradox in statistics. *Sci. Am*. **236** 119–127.

ERLANG, A. K. (1948). On the rational determination of the number of circuits. In *The Life and Works of A. K. Erlang* (E. Brockmeyer, H. L. Halstrom and A. Jensen, eds.) 216–221. The Copenhagen Telephone Company, Copenhagen.

FEDERAL HOUSING AND FINANCE AGENCY. House Price Index, Quarterly Purchase-Only Indexes (Estimated Using Sales Price Data), 100 Largest Metropolitan Statistical Areas (Seasonally Adjusted and Unadjusted). Available at https://www.fhfa.gov/DataTools/Downloads/pages/house-price-index.aspx; accessed 30 January 2019.

FELDMAN, Z. and MANDELBAUM, A. (2010). Using simulation-based stochastic approximation to optimize staffing of systems with skills-based-routing. In *Proceedings—Winter Simulation Conference*. 3307–3317.

FELDMAN, Z., MANDELBAUM, A., MASSEY, W. A. and WHITT, W. (2008). Staffing of time-varying queues to achieve time-stable performance. *Manage. Sci*. **54** 324–338.

GANS, N., KOOLE, G. and MANDELBAUM, A. (2003). Telephone call centers: Tutorial, review, and research prospects. *Manuf. Serv. Oper. Manag*. **5** 79–141.

GANS, N., LIU, N., MANDELBAUM, A., SHEN, H. and YE, H. (2010). Service times in call centers: Agent heterogeneity and learning with some operational consequences. In *Borrowing Strength*: *Theory Powering Applications—a Festschrift for Lawrence D. Brown. Inst. Math. Stat.* (*IMS*) *Collect*. **6** 99–123. IMS, Beachwood, OH. MR2798514

GANS, N., SHEN, H., ZHOU, Y. P., KOROLEV, N., MCCORD, A. and RISTOCK, H. (2015). Parametric forecasting and stochastic programming models for call-center workforce scheduling. *Manuf. Serv. Oper. Manag*. **17** 571–588.

GARNETT, O., MANDELBAUM, A. and REIMAN, M. (2002). Designing a call center with impatient customers. *Manuf. Serv. Oper. Manag*. **4** 208–227.

GERSHWIN, G. and GERSHWIN, I. (1937). Let's Call the Whole Thing Off. Shall We Dance?

GLASSERMAN, P. (2004). *Monte Carlo Methods in Financial Engineering*: *Stochastic Modelling and Applied Probability. Applications of Mathematics* (*New York*) **53**. Springer, New York. MR1999614

GLYNN, P. W. and IGLEHART, D. L. (1989). Importance sampling for stochastic simulations. *Manage. Sci*. **35** 1367–1392. MR1024494

GROVES, R. M. (2011). Three eras of survey reseach. *Public Opin. Q*. **75** 861–871.

GU, J. and KOENKER, R. (2017). Empirical Bayesball remixed: Empirical Bayes methods for longitudinal data. *J. Appl. Econometrics* **32** 575–599. MR3631958

GURVICH, I. and WHITT, W. (2009). Queue-and-idleness-ratio controls in many-server service systems. *Math. Oper. Res*. **34** 363–396. MR2554064

IBRAHIM, R. (2018). Sharing delay information in service systems: A literature survey. *Queueing Syst*. **89** 49–79. MR3803957

IBRAHIM, R. and L'ECUYER, P. (2013). Forecasting call center arrivals: Fixed-effects, mixed-effects, and bivariate models. *Manuf. Serv. Oper. Manag*. **15** 72–85.

IBRAHIM, R. and WHITT, W. (2011). Wait-time predictors for customer service systems with time-varying demand and capacity. *Oper. Res*. **59** 1106–1118. MR2864327

IBRAHIM, R., YE, H., L'ECUYER, P. and SHEN, H. (2016a). Modeling and forecasting call center arrivals: A literature survey and a case study. *Int. J. Forecast*. **32** 865–874.

IBRAHIM, R., L'ECUYER, P., SHEN, H. and THIONGANE, M. (2016b). Inter-dependent, heterogeneous, and time-varying service-time distributions in call centers. *European J. Oper. Res*. **250** 480–492. MR3435009

JAMES, W. and STEIN, C. (1961). Estimation with quadratic loss. In *Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. I* 361–379. Univ. California Press, Berkeley, CA. MR0133191

JIANG, W. and ZHANG, C.-H. (2009). General maximum likelihood empirical Bayes estimation of normal means. *Ann. Statist*. **37** 1647–1684. MR2533467

JIANG, W. and ZHANG, C.-H. (2010). Empirical Bayes in-season prediction of baseball batting averages. In *Borrowing Strength*: *Theory Powering Applications—a Festschrift for Lawrence D. Brown. Inst. Math. Stat.* (*IMS*) *Collect*. **6** 263–273. IMS, Beachwood, OH. MR2798524

KANG, W., PANG, G. (2013). Fluid limit of a many-server queueing network with abandonment. Preprint. http://scripts.cac.psu.edu/users/g/u/gup3/fluidnetwork2013r.pdf.

KASPI, H. and RAMANAN, K. (2011). Law of large numbers limits for many-server queues. *Ann. Appl. Probab*. **21** 33–114. MR2759196

KIM, S. H. and WHITT, W. (2014). Are call center and hospital arrivals well modeled by nonhomogeneous Poisson processes? *Manuf. Serv. Oper. Manag*. **16** 464–480.

KOENKER, R. and MIZERA, I. (2014). Convex optimization, shape constraints, compound decisions, and empirical Bayes rules. *J. Amer. Statist. Assoc*. **109** 674–685. MR3223742

KOLACZYK, E. D. (2009). *Statistical Analysis of Network Data*: *Methods and Models. Springer Series in Statistics*. Springer, New York. MR2724362

LAUGER, A., WISNIEWSKI, B. and MCKENNA, L. (2014). Disclosure Avoidance Techniques at the U.S. Census Bureau: Current Practices and Research. Research Report Series (Disclosure Avoidance #2014-02).

LI, G., HUANG, J. Z. and SHEN, H. (2018). To wait or not to wait: Two-way functional hazards model for understanding waiting in call centers. *J. Amer. Statist. Assoc*. **113** 1503–1514. MR3902225

LINDLEY, D. V. (1962). Discussion on professor Stein's paper. *J. R. Stat. Soc.* **24** 285–287.

MADISON, J. (1790). Census of the Union. In *Annals of Congress*, *House of Representatives*, 1*st Congress*, 2*nd Session*.

MAMAN, S. (2009). Uncertainty in the demand for service: The case of call centers and emergency departments Ph.D. thesis Technion-Israel Institute of Technology, Faculty of Industrial. http://ie.technion.ac.il/serveng/References/Thesis_Shimrit.pdf.

MANDELBAUM, A. and MOMČILOVIĆ, P. (2012). Queues with many servers and impatient customers. *Math. Oper. Res.* **37** 41–65. MR2891146

MANDELBAUM, A. and ZELTYN, S. (2009). Staffing many-server queues with impatient customers: Constraint satisfaction in call centers. *Oper. Res.* **57** 1189–1205. MR2583510

MANDELBAUM, A. and ZELTYN, S. (2013). Data-stories about (im)patient customers in tele-queues. *Queueing Syst.* **75** 115–146. MR3110635

MANDELBAUM, A., MOMČILOVIĆ, P., TRICHAKIS, N., KADISH, S., LEIB, R. and BUNNELL, C. (2017). Data-driven appointment-scheduling under uncertainty: The case of an infusion unit in a cancer center. Under Revision to Management Science.

MATTESON, D. S., MCLEAN, M. W., WOODARD, D. B. and HENDERSON, S. G. (2011). Forecasting emergency medical service call arrival rates. *Ann. Appl. Stat.* **5** 1379–1406. MR2849778

MENG, X.-L. (2018). Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. *Ann. Appl. Stat.* **12** 685–726. MR3834282

MURALIDHARAN, O. (2010). An empirical Bayes mixture method for effect size and false discovery rate estimation. *Ann. Appl. Stat.* **4** 422–438. MR2758178

MUTHURAMAN, K. and ZHA, H. (2008). Simulation-based portfolio optimization for large portfolios with transaction costs. *Math. Finance* **18** 115–134. MR2380942

NAGARAJA, C. H. (2019). *Measuring Society*. CRC Press, Boca Raton, FL.

NAGARAJA, C. H. and BROWN, L. D. (2013). Constructing and evaluating an autoregressive house price index. In *Topics in Applied Statistics* (M. Hu, Y. Liu and J. Lin, eds.). *Springer Proceedings in Mathematics & Statistics* **55** 3–12. Springer, Berlin.

NAGARAJA, C. H., BROWN, L. D. and WACHTER, S. (2014). Repeat sales house price index methodology. *J. Real Estate Lit.* **22** 23–46.

NAGARAJA, C. H., BROWN, L. D. and ZHAO, L. H. (2011). An autoregressive approach to house price modeling. *Ann. Appl. Stat.* **5** 124–149. MR2810392

NEWMAN, M. E. J. (2008). *The Mathematics of Networks*. *The New Palgrave Encyclopedia of Economics*. Palgrave Macmillan, Basingstoke, UK.

NEWMAN, M. (2018). *Networks*. Oxford Univ. Press, Oxford. MR3838417

PFEFFERMANN, D. (2015). Methdological issues and challenges in the production of official statistics: 24th Annual Morris Hansen Lecture. *J. Surv. Statist. Methodol.* **3** 425–477.

PIDGIN, C. F. (1888). *Practical Statistics*: *A Handbook for the Use of the Statistician at Work*, *Students in Colleges and Academies*, *Agents*, *Census Enumerators*, *Etc*. The W.E. Smythe Company.

PUHALSKII, A. A. and REIMAN, M. I. (2000). The multiclass $GI/PH/N$ queue in the Halfin–Whitt regime. *Adv. in Appl. Probab.* **32** 564–595. MR1778580

RAYKAR, V. and ZHAO, L. (2010). Nonparametric prior for adaptive sparsity. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* 629–636.

REED, J. and TEZCAN, T. (2012). Hazard rate scaling of the abandonment distribution for the $GI/M/n + GI$ queue in heavy traffic. *Oper. Res.* **60** 981–995. MR2979435

REICH, M. (2011). The workload process: Modelling, inference and applications. M. Sc. research proposal.

ROBERT, P. (2003). *Stochastic Networks and Queues*: *Stochastic Modelling and Applied Probability*, French ed. *Applications of Mathematics* (*New York*) **52**. Springer, Berlin. MR1996883

SANGALLI, L. M. (2018). The role of statistics in the era of big data. *Statist. Probab. Lett.* **136** 1–3. MR3806826

SEELAB (Service Enterprise Engineering Laboratory). Available at https://web.iem.technion.ac.il/en/service-enterprise-engineering-see-lab/general-information.html.

SENDEROVICH, A. (2016). Queue Mining: Service Perspectives in Process Mining Ph.D. thesis Technion-Israel Institute of Technology, Faculty of Industrial. http://ie.technion.ac.il/serveng/References/Thesis_Submission_Arik_Senderovich.pdf.

SENDEROVICH, A., WEIDLICH, M., GAL, A. and MANDELBAUM, A. (2015). Queue mining for delay prediction in multi-class service processes. *Inf. Syst.* **53** 278–295.

SHEN, H. and BROWN, L. D. (2006). Non-parametric modelling for time-varying customer service time at a bank call centre. *Appl. Stoch. Models Bus. Ind.* **22** 297–311. MR2275576

SHEN, H. and HUANG, J. Z. (2008a). Interday forecasting and intraday updating of call center arrivals. *Manuf. Serv. Oper. Manag.* **10** 391–410.

SHEN, H. and HUANG, J. Z. (2008b). Forecasting time series of inhomogeneous Poisson processes with application to call center workforce management. *Ann. Appl. Stat.* **2** 601–623. MR2524348

STEIN, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1954–1955, *Vol. I* 197–206. Univ. California Press, Berkeley and Los Angeles. MR0084922

STEIN, C. M. (1962). Confidence sets for the mean of a multivariate normal distribution. *J. Roy. Statist. Soc. Ser. B* **24** 265–296. MR0148184

STRAWDERMAN, W. E. (1971). Proper Bayes minimax estimators of the multivariate normal mean. *Ann. Math. Stat.* **42** 385–388. MR0397939

STRAWDERMAN, W. E. (1973). Proper Bayes minimax estimators of the multivariate normal mean vector for the case of common unknown variances. *Ann. Statist.* **1** 1189–1194. MR0365806

TAYLOR, J. W. (2012). Density forecasting of intraday call center arrivals using models based on exponential smoothing. *Manage. Sci.* **58** 534–549.

TORRIERI, N., ACSO, DSSD and SEHSD PROGRAM STAFF (2014). American Community Survey Design and Methodology. U.S. Census Bureau.

TUKEY, J. W. (1977). *Exploratory Data Analysis*. *Addison-Wesley Series in Behavioral Science*. Addison-Wesley Pub. Co., Reading, MA.

VAN DYK, D., FUENTES, M., JORDAN, M. I., NEWTON, M., RAY, B. R., TEMPLE LANG, D. and WICKHAM, H. (2015). ASA Statement on the Role of Statistics in Data Science. *Amstat news*.

VARDI, Y. (1996). Network tomography: Estimating source-destination traffic intensities from link data. *J. Amer. Statist. Assoc.* **91** 365–377. MR1394093

VARIAN, H. (2009). Hal Varian on how the Web challenges managers. McKinsey & Company.

WEINBERG, J., BROWN, L. D. and STROUD, J. R. (2007). Bayesian forecasting of an inhomogeneous Poisson process with applications to call center data. *J. Amer. Statist. Assoc.* **102** 1185–1198. MR2412542

WEINSTEIN, A., MA, Z., BROWN, L. D. and ZHANG, C.-H. (2018). Group-linear empirical Bayes estimates for a heteroscedastic normal mean. *J. Amer. Statist. Assoc.* **113** 698–710. MR3832220

WHITT, W. (1983). The queueing network analyzer. *Bell Syst. Tech. J.* **62** 2779–2815.

WHITT, W. (1992). Understanding the efficiency of multi-server service systems. *Manage. Sci.* **38** 708–723.

WHITT, W. (2002a). *Stochastic-Process Limits*: *An Introduction to Stochastic-Process Limits and Their Application to Queues*. *Springer Series in Operations Research*. Springer, New York. MR1876437

WHITT, W. (2002b). Stochastic models for the design and management of customer contact centers: Some research directions. Department of Industrial Engineering and Operations Research, Columbia Univ., New York.

WHITT, W. (2012). Fitting birth-and-death queueing models to data. *Statist. Probab. Lett.* **82** 998–1004. MR2910048

WRIGHT, C. D. and HUNT, W. O. (1900). The history and growth of the United States census: Prepared for the Senate Committee on the Census. In 56*th Congress*, 1*st Session*; *Document No*. 194.

XIE, X., KOU, S. C. and BROWN, L. D. (2012). SURE estimates for a heteroscedastic hierarchical model. *J. Amer. Statist. Assoc.* **107** 1465–1479. MR3036408

YE, H., LUEDTKE, J. and SHEN, H. (2019). Call center arrivals: When to jointly forecast multiple streams? *Prod. Oper. Manag.* **28** 27–42.

YOM-TOV, G. and MANDELBAUM, A. (2014). Erlang-R: A time-varying queue with reentrant customers, in support of healthcare staffing. *Manuf. Serv. Oper. Manag.* **16** 283–299.

ZELTYN, S. and MANDELBAUM, A. (2005). Call centers with impatient customers: Many-server asymptotics of the $M/M/n+G$ queue. *Queueing Syst.* **51** 361–402. MR2189598

ZELTYN, S., MARMOR, Y. N., MANDELBAUM, A., CARMELI, B., GREENSHPAN, O., MESIKA, Y., WASSERKRUG, S., VORTMAN, P., SCHWARTZ, D. et al. (2011). Simulation-based models of emergency departments: Real-time control, operations planning and scenario analysis. *ACM Trans. Model. Comput. Simul.* **21** 3.

ZHANG, P. and SERBAN, N. (2007). Discovery, visualization and performance analysis of enterprise workflow. *Comput. Statist. Data Anal.* **51** 2670–2687. MR2338995

U.S. CENSUS BUREAU (1907). Heads of Families at the First Census of the United States Taken in the Year 1790. Government Printing Office, Washington, DC.

U.S. CENSUS BUREAU (2009). TIGER/Line Shapefiles. Available at https://www.census.gov/geo.

U.S. CENSUS BUREAU. Decennial census of population and housing. Available at https://www.census.gov/programs-surveys/decennial-census.html.

U.S. CENSUS BUREAU (2010). Census Bureau Launches 2010 Census Advertising Campaign: Communication Effort Seeks to Boost Nation's Mail-Back Participation Rates. Available at https://www.census.gov/newsroom/releases/archives/2010_census/cb10-cn08.html, January 2010.

U.S. CENSUS BUREAU (2017a). "Annual Estimates of the Resident Population: April 1, 2010 to July 1, 2017—Table PEPANNRES." Population Estimates Program.

U.S. CENSUS BUREAU (2017b). "Geographic Mobility by Selected Characteristics in the United States—Table S0701." American Community Survey 1-Year Estimates. Available at https://factfinder.census.gov.

U.S. CENSUS BUREAU (2017c). "Citizen, Voting-age Population by Age—Table B29001." American Community Survey 1-Year Estimates. Available at https://factfinder.census.gov.

U.S. CENSUS BUREAU (2017d). "Field of Bachelor's Degree for First Major—Table S1502." American Community Survey 1-Year Estimates. Available at https://factfinder.census.gov.

U.S. CENSUS BUREAU (2017e). "Commuting Characteristics by Sex—Table S0801." American Community Survey 1-Year Estimates. Available at https://factfinder.census.gov.

U.S. CENSUS BUREAU (2017f). "Veteran Status—Table S2101." American Community Survey 1-Year Estimates. Available at https://factfinder.census.gov.

U.S. CENSUS BUREAU—Census History Staff (2017g). Title 13, U.S. Code. Available at https://www.census.gov/history/www/reference/privacy_confidentiality/title_13_us_code.html. Last revised: July 18, 2017.

U.S. CENSUS BUREAU—Census History Staff (2017h). Title 26, U.S. Code. Available at https://www.census.gov/history/www/reference/privacy_confidentiality/title_26_us_code_1.html. Last revised July 18, 2017.

U.S. OFFICE OF THE SECRETARY OF STATE (1793). Return of the Whole Number of Persons Within the Several Districts of the United States According to, "An Act Providing for the Enumeration of the Inhabitants of the United States."