A Systematic Approach for Business Data Analytics with a Real Case Study

Kaibo Liu, Department of Industrial and Systems Engineering, University of Wisconsin Madison, Madison, WI, USA

Jianjun Shi, H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, USA

ABSTRACT

Business data analytics is a process of utilizing analytic techniques for resolving business issues based on business performance data. While the avalanche of business data creates unprecedented opportunity, it also poses three fundamental challenges for analytics: (1) Business data often encounters quality issues and needs substantial cleaning efforts; (2) Business data is large in overall size but cannot be fully shared due to the concern of data security; and (3) Business data often needs to be cross-referenced with public databases to reveal more information and knowledge. Due to these challenges, the leading obstacle at many organizations is the lack of a systematic approach to understanding how to leverage the business data analytics techniques to transfer from data-rich into decision-smart. To answer this question, this article proposes a systematic step-by-step procedure for business data analytics. This proposed framework is illustrated and validated by a real case study that involves choosing an optimal location for opening of a new retail site.

Keywords: *Business Data Analytics, Business of Big Data, Data Analytics Software, Gravity Model, New Retail Site Location Selection*

INTRODUCTION

Business operations generate huge amounts of valuable data, such as consumer information, transactions data, shipment information, and service record. As the massive data become available, there has been an urgent need to effectively acquire, assess, analyze, and visualize these big data to gain powerful insights and enhance decision makings for business improvement. *Business analytics*, the application of analytic techniques to resolve business issues, has recently become a popular buzzword (Năstase & Stoica, 2010). For example, according to the IBM Tech Trends Report (2011), business analytics has been identified as one of the four major technology trends in the 2010s. As there is tremendous value embedded in the business data, the correct use of business analytics will create competitive advantages, such as incremental revenue, decreased cost, fast response in the supply chain, and improved customer experience and engagement.

DOI: 10.4018/IJBAN.2015100102

While the avalanche of business data creates unprecedented opportunity for deeper understandings of the business process, the market environment, the customer behavior, and the competitive strategy, it also poses new challenges in handling the massive volume of data, extracting the useful information, and transferring from data-rich into decision-smart (Abdelhafez, 2014). As predicted by McKinsey Global Institute, by 2018, the United States alone will face a shortage of over 140,000 people with deep analytical skills, as well as a shortfall of 1.5 million managers without knowledge on how to leverage the business of big data to make effective decisions (Manyika et al., 2011).

Generally speaking, there are three fundamental challenges when applying the analytic techniques to the business data: (1) *Business data often encounters quality issues and needs substantial cleaning efforts before making meaningful analysis.* Business data is not error-free and often contains noise in the form of inaccuracies, inconsistences, and missing data (Tavana, Trevisani, & Kennedy, 2014). These problems occur when the business data is collected from multiple sources, recorded by different people who interpret the data with different terminologies due to diverse training backgrounds, and entered into the database in a variety of formats, coding standards, and aggregation strategies due to the heterogeneity in software used in each department and company. As a result, there is often a lack of a standard approach from loading to cleaning to processing the data before it can be used for decision making. (2) *Business data is large in overall size but cannot be fully shared due to the concern of data security at the company.* Business data is most favorable for decision making when all information of each company can be shared in a common database. However, as there is invaluable confidential information embedded in the business data, prevention of data leakage is often a primary concern at the company. Many companies and markets operate in a highly competitive environment; thus, any privacy failure can potentially lead to a loss of market share, affect customer retention, and cause significant damage to a company's reputation (Schläfke, Silvi, & Möller, 2013). In other words, due to the concern of data security, each company is often only enriched in its own proprietary business data and lacks the detailed transaction data of its competitors. Although the business data has a large size in the overall network, the limited access to the proprietary transaction data at each company essentially hinders the analytics process. (3) *Business data often needs to be cross-referenced with public databases to reveal more information and knowledge*. In addition to the proprietary transaction data owned by the company, there are enormous public databases reported regularly by the government and consortiums. These additional public data sources, covering a variety of information such as economics, population and geography, create another opportunity for the company to cross-reference the data to maximize the advantages of data analytics. However, there are two challenging questions that remain to be solved: (i) where to find and select the relevant public database in the big data environment and (ii) how to bridge the available data (both proprietary and public) under reasonable assumptions to enhance decision making.

These three aforementioned challenging issues often limit the use of business data analytics in practice. As a result, although new technologies have enabled data collection to be easier and faster than ever before, many companies are still looking for a systematic way to obtain the maximum value from their data and compete in the marketplace. The main goal of this article is to outline a systematic step-by-step procedure to transform from data-rich into decision-smart for business data analytics. An overview of the proposed strategy is elaborated in Figure 1 in the Appendix. Detailed explanations and discussions of each step are provided then. This proposed framework is illustrated step-by-step based on a real case study that involves choosing an optimal location for the opening of a new retail site. Finally, the last section draws a conclusion and discusses future work.

PROPOSED SYSTEMATIC APPROACH TO BUSINESS DATA ANALYTICS

Building an effective business data analytics requires the involvement of at least three different types of experts with strong communication and interactive skills: (i) Business experts, who set the objectives and make the decision; (ii) Information technology experts, who manage the database; and (iii) Data analysis experts, who understand data mining and statistical methods for analyzing the data. These three types of experts need to collaborate and work closely with each other in the proposed analytics process that is explained step-by-step below.

Step 1: Business Strategy

As the big data can bring both more-panoramic and more-granular views of business environment, many organizations strive on gathering data, but pay less attention to understand the potential uses of the data. In fact, without appropriate business strategy, the value of data that is blindly collected and analyzed remains to be seen. Actions taken, if any in such case, might not be the most valuable one for overall business strategy (LaValle, Lesser, Shockley, Hopkins, & Kruschwitz, 2011). Thus, instead of simply fetishizing the data itself, organizations should start the business data analytics from identifying the current priority challenge such as improving customer experience or opening of a new retail site, which if successfully solved, can bring the highest profit to the company. Once such a business strategy is defined, the next step is to transform it into a mathematical formulation, which is often expressed as an optimization function of the decision variables. In this way, any results achieved will be well aligned with the business strategy and actions can be delivered at the right time.

Step 2: Integration of Model Identification and Data Acquisition

Once the business strategy is identified, the second step is to identify the appropriate model and acquire the right data to achieve the business strategy. These two elements, model identification and data acquisition, are needed to be seamlessly integrated together to optimize the data analytics process.

The most effective approach often originates from searching for the related literature to specify an appropriate model for optimizing the identified business strategy (Barton, 2012). Understanding how the model is originally developed and evolved and what assumptions are made in the model is essential. In addition, the model will recognize what types of data need to be acquired as an input to the analytics process. However, due to the existences of these three challenges of business data as mentioned in the introduction section, it often requires much more effort and time in just processing the data before analytics takes place. For example, the data quality may be a minor issue and can be manually corrected when the size of the data is small, but the problem can be significantly exacerbated in the big data environment. Thus, to deal with the quality issues (e.g., missing, incomplete, inconsistent) of business data, dedicated efforts are needed to study the patterns and symptoms of the data first. Organizations may start by focusing on a subset of data to identify a set of systematic rules for rescuing the poor data quality. By implementing these developed rules on the business of big data, additional patterns of errors that are not previously realized in analyzing the subset of data can then be spotted. Consequently, this kind of efforts should be done iteratively before any analytical model could finally play a role. It is possible that there are still some data entries or outliers that need special treatments after the systematic cleaning efforts. However, as the size of these cases is generally small now, manual intervention, interpretation, and cleaning may become feasible.

In addition to the data quality issues, the concern of data security is another challenge for business data analytics. Each company often only enriches in its proprietary business data but has little information about the detailed transaction performance of its competitors. Most of the analytics models assume all necessary information has been fully acquired; however in practice, some inputs of the model may not be available in the company's database. This requires us seeking additional information from the public domain dataset and then incorporating it into the identified model as a complement. For example, social media is generating huge amount of data in the form of texts, photos and videos (Sasaki, 2014). Moreover, U.S. census regularly provides quality data about the nation's people such as census of population, housing, economics and governments. These data are often cross-referenced with the business performance data of the company. Therefore, by studying these relevant data, customers or competitors' information can then be inferred. It is worth mentioning that analyzing these nontraditional and unstructured data will be effective only if it is related to the identified business strategy and analytics model, as the ratio of the useful information over the total amount of big data is generally small (Pavolotsky, 2012). In addition, identifying the most relevant cross-referenced data is another challenging task and it requires the analytics team to have the capability to think creatively and make reasonable assumptions. In the end, after all the relevant data (both proprietary and public) is acquired, the model identified in the beginning stage often needs to be further modified to ensure it fits the practical problem and dataset well.

Step 3: Visualization

Visualization is the next step that focuses on quickly distilling the potentially huge amounts of data and transferring them into salient results that can be readily understood by everyone in the business data analytics team. Although the concept of visualization is not new, its value and use are often overlooked and poorly implemented in practice during the business decision making process (Stodder, 2013). Moreover, as the size of the business data becomes larger, visualization nowadays is desired to possess advanced features that are beyond conventional reporting functions. Specifically, visualizations are desired to enable (1) quickly dealing with business of big data; (2) multidimensional representations of analysis results via controlling the color, brightness, size, shape and motion of visual objects; and (3) geospatial and location intelligence that depicts physical features and geographically referenced data and relationships by combining geographic and location-related information from a variety of data sources, including aerial maps and consumer demographics (Sallam, Tapadinhas, Parenteau, Yuen, & Hostmann, 2014). For example, dashboards, a style of graphically depicting performance of measures by using such as gauges, sliders, checkboxes and maps have been recently recognized as an important presentation tool for visualization. In addition to these advanced capabilities, visualization in practice is also required to be transparent, easily understood, implemented in an easy-to-use interface, and able to be rapidly shared within the company via mobile devices. Statistical and practical interpretations are also needed based on the visualization results for better understanding of the problem of interest. Please keep in mind that visualization should be always put into the perspective of contextual business strategy such that it enables users to directly take actions based on the observations made.

Another critical concern for visualization is the trade-off between information enrichment and data protection. When the analysis result is prepared for use only within the company, visualization is required to be as enriched and granular as possible. On the contrary, when the analysis result is intended for public use, visualization has to be carefully prepared so that proprietary information can be preserved. This is also a challenging problem faced by many researchers nowadays when they study a practical problem with real datasets, but the results cannot be fully presented in journal publications due to data security. In such case, tables with quantitative business performance measures are not suitable for reporting analytics results anymore. Instead, we should rely on some advanced visualization techniques such as using the palette features (i.e., controlling the contrast, brightness and color) to provide necessary comparisons between different potential business decisions without releasing the exact numerical information.

Step 4: Uncertainty Analysis

A strong visualization should advance users' understanding about their business performance and provide direct answers to the identified business strategy. However, before making the right decision, it is necessary to conduct uncertainty analysis to thoroughly understand the risk of such a potential action. The uncertainty stems from two aspects of the business data analytics: data uncertainty and model uncertainty. Data uncertainty is due to the poor quality of business data and the assumptions made to cross reference the proprietary business performance data with public database as mentioned before. Model uncertainty involves the risks of the model specification and the estimated parameters inside the model. Generally speaking, uncertainty analysis requires assessing the stochastic nature of the decision to be made with some statistical degree of confidence instead of just treating decision making as a deterministic process. For example, Monte Carlo simulation, a method draws repeated samples from probability distributions are commonly used to evaluate data uncertainty. On the other hand, Bayesian statistical modeling approaches can be implemented to study the model uncertainty. A comprehensive review of the methods for uncertainty analysis can be found in (Frey & Patil, 2002; Helton, Johnson, Sallaberry, & Storlie, 2006).

Step 5: Decision Making

Based on the result of uncertainty analysis, robust decision-making process can be achieved accordingly. There are some decisions that need to be made regularly (e.g., determination of the product price). In such case, business data analytics process is required to be incorporated with real-time information and conducted in a timely manner, such that the right decision can be delivered to the right group of customers at the right time. On the other hand, some decisions are like one-step move such as purchasing the competitor's company. For this kind of decision-making process, a set of back-up plans usually needs to be prioritized in case some emergency situations occur. The decision-making process should be well aligned with the identified business strategy in the step 1. Any discrepancies require iterating the whole business data analytics process until a unified scheme is finally achieved.

CASE STUDY: CHOOSING A NEW RETAIL SITE LOCATION BY USING BUSINESS DATA ANALYTICS

In this section, we focus on a popular business analytics problem: the selection of new retail site locations. The goal of this case study is to demonstrate and validate our proposed business data analytics framework to transfer from data-rich into decision-smart when choosing a new retail site location. A detailed step-by-step procedure corresponding to the proposed scheme will be illustrated in the following example by using a real dataset.

Step 1: Business Strategy

To begin with, we first introduce some background information of our studied company. The company of interest in this case study conducts gas station equipment repair and replacement business. Generally speaking, once an order of repair or replacement is received, this company will immediately send new items from its retail site to the designated gas stations. This company currently owns a single retail site located in Georgia. The highest priority strategy of this company right now is to open a new retail site to maximize the potential profits.

Choosing a new retail site location is a crucial question when a company tries to expand its business. As building the retail site can be prohibitively expensive, a retailer often has to live with the site for many years (Buckner, 1998). Thus, the retail site location selection process is critically important, which can make a profitable company suddenly run into debts if the new site fails to meet the investment strategy. In order to rationally decide the optimal location of the new retail site, it is essential to accurately estimate the new market shares of the company over the country if the new retail site is tentatively to be opened at different potential locations (Bruno & Improta, 2008).

Recall that once the business strategy is identified, it needs to be further transferred into a mathematical formulation. Thus, below we will introduce the notation and the general problem formation for choosing the optimal new retail site location. Consider a network:

$$N = \left\{1,2,\ .\ .\ .\ ,n\right\}$$

with n nodes. Let:

$$E = \left\{i_1,\ldots,i_e\right\} \subseteq N$$

be the set of existing e retail sites of our studied company (i.e., $e = 1$ in our example) and $\bar{E} = N - E$ be the set of potential nodes for opening the new retail site. Then, the current total amount of business sales in our studied company can be represented as $\sum_{j \in N} \sum_{i \in E} B_{ij}$, where B_{ij} is the business sales between nodes j and i before opening of the new retail site. The problem of interest is to find an optimal site location $i' \in \bar{E}$ such that the total amount of business sales of the company can be maximized after opening the new retail site:

$$\sum_{j \in N} \left(\sum_{i \in E} B_{ij}^{(new)} + B_{i'j}^{(new)} \right) \tag{1}$$

where $B_{ij}^{(new)}$ is the updated business sales between node j and node i due to opening of the new retail site at location i'. Assume the total demand at each node is constant. Then, the challenge question here is how to estimate $B_{ij}^{(new)}$ and $B_{i'j}^{(new)}$ as its value depends on not only the new retail site location i' but also the market share distribution of competitors in the network.

Step 2: Integration of Model Identification and Data Acquisition

Introduction to the General Form of Gravity Model

For retail location analysis, one of the most popular methods is the gravity model. The gravity models have been developed through the adaptation of Newton's law of gravitation to the economic cases (Young, 1975). In theory, there are three important factors in the gravity models: activity, attraction, and friction of distance. The activity factor is associated with the demand of customers and is often measured by the income or the population of customers at a particular area (Grosche, Rothlauf, & Heinzl, 2007). The attraction factor is associated with the facility and the environment of the retail site and is often measured by the square footage of selling space (Huff, 1963; Ozuduru, 2013). Last, the friction of distance factor is used to represent the phenomenon that a customer is less likely to choose a retail site as the distance from the site increases due to the travel time and cost spent in the trip (Huff & Jenks, 1968; Rodrigue, 2012).

The use of the gravity model is based on the assumption that the amount of business sales that the retail site at node i draws from the customer at node j is proportional to the activity factor of node j and the attraction factor of node i, but inversely proportional to the friction of distance factor between nodes j and i. Typically, this gravity model can be represented by the following general form (Bruno & Improta, 2008):

$$B_{ij} = c * P_j * A_i * \left(D_{ij}\right)^{-\gamma} \tag{2}$$

where B_{ij} is the amount of business sales between the customer at node j and the retail site at node i; P_j is the activity factor of node j; A_i is the attraction factor of node i; D_{ij} is the distance between nodes i and j; c is a constant number; and γ is called the distance-decay parameter. The original model in (Reilly, 1931) suggests $\gamma = 2$ according to the Newton's law of gravitation. However, numerous empirical studies (Young, 1975; Drezner & Drezner, 2002) have shown that using a predetermined distance-decay parameter may lead to an inaccurate conclusion. Thus, here we consider γ as a calibration parameter.

The gravity model was originally proposed by Reilly (1931), who used the gravity model to identify the breaking points of retail influence between two competing cities. On the basis of the efforts in Reilly (1931), Huff (1963) developed an alternative gravity model by switching the focus from the destination to the customer to predict shopping probabilities from multiple demand points. Nakanishi and Cooper (1974) further extended the Huff's approach by a multiplicative competitive interaction model which considers a product of variables as the attraction factor for the retail site. Plastria (1997) first considered both the choice of location and the attraction factor simultaneously in the competitive location model. Review of gravity modeling and its impacts on location analysis can be found in (Drezner, 1995; Plastria, 2001; Joseph & Kuby, 2011).

The gravity model has been used for a variety of site location analysis such as identifying the best location for an airline hub (Drezner & Drezner, 2001), university (Bruno & Improta, 2008), hospital (Lowe & Sen, 1996) and shopping center (Drezner & Drezner, 2002). However, none of them have explicitly considered the aforementioned three challenges of business data analytics in the big data environment as mentioned in the introduction section. The existing literature on gravity model analysis often assumes that all the necessary dataset has been perfectly acquired (e.g., the competitor's information) and ignores the privacy issues of the business data. Without information sharing in the marketplace, it is a very challenging task to analyze the impact of new retail site location on the current market shares. To address this issue, it is important to leverage the company's proprietary transaction data with the cross-referenced public database to discover more information and knowledge to enhance the analytics capability. Therefore, there is a pressing need to understand what opportunities and challenges that the business of big data carries and brings to the gravity model nowadays. For example, lots of related information such as the number of gas stations and the number of population at each node can be useful for better implementation of the gravity model. However, where to find these data and what is the impact of data uncertainty for the decision making by using the gravity model are of great interest and will be investigated in this case study.

Proprietary Business Data Acquisition and Processing

In this paper, detailed business transaction data have been collected every day from our studied company. A dataset is exported from the QuickBooks® software, which contains a total of more than one million detailed business transaction data with a size about 8 GB over the last 5-year period. Each transaction data includes a variety of information, such as the order date, the name and the address of the customer (i.e., gas station), the ordered item information, and the sale price.

To implement the gravity model, the first step is to divide the study area of interest into distinct communities, such as cities, states, and regions with same zip code (Drezner & Drezner, 2002). Each community consists of its resident customers. Specifically, the company is currently interested in determining the optimal *state* location for opening of the new retail site, as the gas consumption and price vary a lot among different states (GasBuddy/OpenStore LLC, 2014).

In our example, there are more than one million business transactions; furthermore, different formats are observed in the data entries due to different personal preference when the employee enters the data into database. For instance, as shown in Table 1 in the Appendix, some customers' location information (e.g., customer 3) only contains the zip code information while others (e.g., customers 2 and 4) may contain the detailed street name. Some location information (e.g., customer 1) is written by the abbreviation of the state whereas others may be written with the full name (e.g., customer 2). And some customer's information is partially or even completely missing (e.g., customers 3 and 4) either due to a data loss or were not entered by employees. Therefore, significant efforts are needed to systematically clean and process the business data before the gravity model can be implemented. Based on the proposed approach in the step 2 of the previous Section, we identify the following systematic steps to clean the data: 1) To deal with the unstructured location information, we first implement the text mining method and search for all state names (either full or abbreviation) to project the location information of customers into the state level. 2) As there are some customers (e.g., customer 3) only recorded with the zip code information, additional efforts were made to cross reference the zip code data with state name. Such relationship can be found in the public database (Zip Code Database, 2014). 3) After these two steps, there are still 65 customers without state information due to either missing information or data errors existed in the original state name and zip code. Since these types of data issues are heterogeneous and random, we have to manually correct their location information by googling the customer's name in the internet. Fortunately, as the number of entries is significantly reduced now, manually processing these special cases becomes feasible.

Figure 2 in the Appendix shows the average sale distribution within the last 5-year period before opening of the new retail site after the initial data acquisition and processing efforts. According to Figure 2, Georgia, Ohio, Tennessee, and Texas have the largest amount of business with our studied company. In addition, most of the business is concentrated in the east side of United States and is centered on the Georgia state (current retail site location).

Gravity Model Development Based on Company's Proprietary and Public Databases

In this sub-section, our focus is to demonstrate where to find the relevant public databases and how to bridge the information from both proprietary and public datasets to enhance analytics process. As the existing literature on gravity models has not considered the abovementioned three challenges of business data analytics, a new approach for the gravity model that is tailored to this problem in the business of big data environment needs to be developed.

Recall that when customers choose a retail site for repair or replacement, there are two important factors to consider: (1) the attraction factor that related to the characteristics of each retail site; and (2) the friction of distance factor that related to the shipping time. For example, the square footage of selling space is often used to measure the attraction factor (Huff, 1963; Ozuduru, 2013) as this variable directly relates to many important characteristics of the retail site, such as the production power, the number of personnel and the amount of investment. On the other hand, the friction of distance factor can be measured by the distance between states, if we assume the shipping time is proportional to the shipping distance. In our example, customers need to wait for the repaired or replaced equipment shipped from the retail site. Thus, this scenario is a little bit different from the existing studies, in which customers are required to travel to the retail site, such as a shopping center (Drezner & Drezner, 2002). However, the concepts of gravity (i.e. a longer distance indicates more travelling time) can still be applied here. According to Figure 2, it clearly shows that the amount of business generally decreases as the distance from the retail site in the Georgia state increases. In the gravity models, the center of the community is often used to present the location information of customers that belong to the corresponding community. Such distance information between the centers of any two states can be obtained from the public websites (e.g., http://geography.about.com/library/weekly/aa120699a.htm/). As for the shipping distance within each state (e.g., customers in Georgia make an order from the current retail site), we estimate its value by using the half radius of a circle whose area is equal to the area of the corresponding state.

In addition to the attraction and the friction of distance factors, the activity factor of the state also determines the amount of business between customer and retail site according to Equation (2). The activity factor is associated with the demand of gas stations; however, this variable is difficult to be measured directly. Thus, we choose to use other factors as a substitute under reasonable assumptions. For example, one possible choice is the number of gas stations, which can be considered to be proportional to the demand of gas stations if we assume each gas station has the same failure rate. Another possible choice is to use the population data. According to the U.S. Census Bureau press release (2008), there is a gas station for approximately every 2,500 people in the United States. Figure 3 in the Appendix further validates this positive linear relationship between the population and the number of gas stations in each state (the p value is less than 0.0001). The number of gas stations and the population data in each state can be obtained from U.S. Census Bureau public database (https://www.census.gov/).

Now, all the three factors in the gravity model have been acquired. To tackle the issue that the competitor's information is unknown, we further propose the following two-step procedures for location analysis:

In step one, we utilize the gravity model from the retailer viewpoint to understand how each retail site in E attracts business from customers in N. To begin with, we take a log transformation on both sides of the Equation (2):

$$\log\left(B_{ij}\right)=\log\left(P_j\right)-\gamma\log\left(D_{ij}\right)+\log\left(A_i\right)+log\left(c\right) \tag{3}$$

denote:

$$Y_{ij}=log\left(B_{ij}\right)-\log\left(P_j\right)$$

and:

$$X_{ij}=-\log\left(D_{ij}\right)$$

Then, we can use the least squares method to estimate the parameters c, γ and A_i. The detailed procedure is as follows. Define:

$$Y=\left[Y_{i_1 1},\ldots,Y_{i_1 n},Y_{i_2 1},\ldots,Y_{i_2 n},\ldots,Y_{i_e 1},\ldots,Y_{i_e n}\right]'$$

by the vector with size $\left(e\cdot n\right)*1$ and:

$$X=\left[V_{i_1};V_{i_2};\ldots;V_{i_e}\right]$$

by the matrix with size $\left(e\cdot n\right)*\left(e+2\right)$, where:

$$V_{i_k}=\left[1,v_{i_k},X_{i_k 1};1,v_{i_k},X_{i_k 2};\ldots;1,v_{i_k},X_{i_k n}\right]$$

is a matrix with size $n*\left(e+2\right)$ and v_{i_k} is a unit row vector with the k-th $\left(k=1,\ldots,e\right)$ element equal to 1. Assume $Y=X\beta+\varepsilon$, where $\varepsilon\sim N\left(0,\sigma^2 I\right)$ and:

$$\beta = \left[\log\left(c\right), \log\left(A_{i_1}\right), \log\left(A_{i_2}\right), \ldots, \log\left(A_{i_e}\right), \gamma\right]'$$

Then, the least squares method estimates β with the equation:

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T Y$$

In the case that there is only one retail site in the company (i.e., $e = 1$ in our example), the variable A_i will be alias with the variable c, and the least squares method will estimate their joint effects.

2. In step two, we utilize the gravity model from the customer viewpoint to understand how each customer in N chooses business from retail sites in E. One of the challenges here is that when customers choose retail sites, they will not only consider the retail sites of our studied company in E, but also consider the retail sites of other competitors in N. However, due to the data privacy, the detailed transaction data of different competitors is not shared with each other. Let $A_{i'}$ be the attraction factor of the new retail site opened at state $i' \in \overline{E}$. To address this issue, we first estimate the potential new business $\hat{B}_{i'j}$ between the new retail site at state i' and other state $j \in N$ without considering the current market share in the network:

$$\hat{B}_{i'j} = \hat{c} * P_j * A_{i'} * \left(D_{i'j}\right)^{-\hat{\gamma}} \tag{4}$$

Denote the current market share of our studied company in the United States as M. Then, the estimated annual cost of equipment repair or replacement spent at each gas station is:

$$\frac{\sum_{i \in E} \sum_{j \in N} B_{ij}}{M * \sum_{j \in N} G_j}$$

where G_j is the number of gas station at each state j. As a result, we can estimate the current market share of our studied company at each node $j \in N$ by:

$$\hat{M}_j = \frac{M * \sum_{j \in N} G_j}{\sum_{i \in E} \sum_{j \in N} B_{ij}} * \frac{\sum_{i \in E} B_{ij}}{G_j}$$

In this way, the increased market share $M_{i'j}$ of our studied company due to opening of the new retail site at state i' is:

$$\widehat{M}_{i'j} = \left(\frac{\widehat{B}_{i'j}}{\sum_{i \in E} B_{ij}}\right) * \widehat{M}_j$$

However, as the total demand at each node is assumed to be constant, the estimated $\hat{B}_{i'j}$ in Equation (4) must be scaled according to the current market share of our studied company and the total demand constraint:

$$\widehat{B}_{i'j}^{(new)} = \frac{\widehat{B}_{i'j}}{\widehat{M}_{i'j} + 100\%} \tag{5}$$

Similarly, due to the opening of new retail site at node i', the expected business $\widehat{B}_{ij}^{(new)}$ between the customers at state $j \in N$ and the current retail site at state $i \in E$ of our studied company will be changed to:

$$\frac{B_{ij}}{\widehat{M}_{i'j} + 100\%}$$

In other words, the total amount of business sales after the opening of the new retail site will become:

$$\sum_{j\in N}\left(\sum_{i\in E}\hat{B}_{ij}^{(new)} + \hat{B}_{i'j}^{(new)}\right)$$

Then we can search for all possible nodes in $\bar{E}$ and the optimal site location i' will be the one that maximizes the total amount of business sales:

$$argmax_{i'\in\bar{E}}\sum_{j\in N}\left(\sum_{i\in E}\widehat{B}_{ij}^{(new)} + \widehat{B}_{i'j}^{(new)}\right)$$

Step 3: Visualization

Selection of Business Data Analytics Software

As the size of the dataset increases, how to make the data easily understood via rapid and meaningful visualization has become increasingly important. Good data visualization can quickly convey the analytical results and yield actionable insights. As mentioned in the step 3 of the previous section, here visualization faces a trade-off problem. On one hand, visualization is required to be transparent, information enriched, and easily understood. On the other hand, propriety information needs to be preserved when visualization appears in publication. In addition, it is desired that the visualization can be conducted in an easy-to-use interface and information can be rapid shared within companies via mobile devices.

In this case study, we choose to use the Tableau software as a platform to demonstrate our business analytical results and all figures presented in this paper are generated from the Tableau software. One of the advantages for Tableau software is that it is easy to use (known as “the democratization of analytics within organizations”). The key technology underpinning the Tableau Software is the VizQL (visual query language) that does not require the complexity of a traditional OLAP cube (Cosentino, 2012) and allows the user to explore the data by dragging and dropping variables on the screen for rapid visualization. Another advantage for Tableau software is that it can provide a unique interface for map-based geospatial analyses and much flexibility in quickly generating different types of visuals even with big datasets. As a result, we can elaborate our analytics results comprehensively by using the palette feature (e.g., see Figure 2). This unique feature not only provides us with sufficient information for decision making but also preserves the exact proprietary information of the company.

Visualization for Retail Site Location Selection

Recall this study includes more than one million transaction data within the last 5-year period. Significant efforts have been made for cleaning and processing the data and the average sale amount information has been obtained at each state as shown in Figure 2. In addition, the distance information between any two states have been acquired and documented for use in the gravity model. The population data is chosen as a quantitative measure for the activity factor of each state. As our studied company is planning to open a new retail site which has a similar square footage of selling space as the current one in Georgia, the attraction factor of the new retail site is assumed to be the same as the current one. Furthermore, this company has already recruited a certified research firm to estimate its current market share, M in the Unites States. However, due to the concern of data security, the value of M is not given here. We can then estimate the current market share of our studied company at each state based on the approach developed in the step 2 of this section. Figure 4 in the Appendix shows the estimated current market share of the company in the Unites States (due to data privacy, the exact number has already been removed).

By using our proposed method, Figure 5 in the Appendix further shows the comparisons of the estimated total business sales if we tentatively open the new retail site at a different state each time and try all possible state locations. The darker the color of the state, the higher the estimated total amount of business will be if the new retail site opens at the corresponding state. This visualization has many advantages. Not only can any two candidate states be easily compared by referring to the brightness of these two states, but also this figure clearly shows the geographic information of the state at which the new retail site should be opened. In addition, this visualization only highlights the relative differences in the estimated total business sales if opening of the new retail site at different locations. Thus, it can prevent releasing the exact amount of business sales that are concerned as the confidential information in the company. According to Figure 5, we can see that:

1. The best state for opening the new retail site is Pennsylvania, and followed by Maryland, Virginia, and West Virginia. The worst state for opening the new retail site is Washington, and followed by Montana, Oregon, and Idaho. This result is a little surprised if we consider the current amount of business sales before opening of the new retail site is relatively small at the west side of the United States in Figure 2. Thus, intuitively the new retail site should be opened at the west side to increase the total business. However, this does not consider the potential business sales and the current market share at each state. Figure 6 in the Appendix illustrates the number of gas stations at each state in 2007 according to the U.S. Census Bureau Public Database. It clearly shows that the number of gas stations is relatively small at the west side of United States, which indicates that the total potential business at the west side of Unites States is low, and thus opening of the new retail site at the this region may not be able to maximize the total business sales in the overall country;
2. The top four states for opening the new retail site are all located in the middle of east side of the United States and the estimated total amount of business gradually decreases as the state gets farther from this region. One possible explanation is that the geographic size of the states in this region is relatively small and these states are geographically close to each other; thus, opening the retail site at this compact district requires shorter shipping distances but can increase much business. For example, although there is much business potential in the California as indicated in Figure 6, it is too far from other states that are on the east side of United States. Thus, opening the new retail site in the California may not be able to maximize the total business sales in the overall country.

Step 4: Uncertainty Analysis for Retail Site Location Selection

As discussed in the step 4 of previous section, uncertainty analysis is the next essential step that provides a comprehensive risk analysis before making the final decision. Uncertainty analysis for the retail site location selection has been studied in the literature. For example, Plastria and Vanhaverbeke (2008) incorporated future changes in market conditions into the location model, and the optimal location is determined with the anticipation of the future competitor's entry. Drezner (2009) considered the uncertainty in the activity and attraction factors when choosing the optimal location of retail sites.

Unlike most of the existing studies, here we focus on the uncertainty of the transaction data on the optimal retail site location analysis. Specifically, we employ the bootstrap sampling technique (Hastie, Tibshirani, & Friedman, 2009; Liu, Jain, & Shi, 2013) that randomly draws data points with replacement from the company's proprietary database and each bootstrap dataset has the same size as the original transaction data. For each bootstrap sample, we conduct the optimal retail site location analysis by using the proposed method. We repeat the experiments 100 times, and Figure 7 in the Appendix shows the variation in the expected total business sales of the company if tentatively opening of the new retail site at a different state each time and for all possible state locations. Each dot represents the expected total amount of business sales in one bootstrap sample and the red lines represent the three quartiles according to the variations of the expected total amount of business sales in the 100 bootstrap samples.

Based on the visualization in Figure 7, we can see that there are some overlapping areas between the three quartiles of the top four states (Pennsylvania, Maryland, Virginia, and West Virginia) for opening the new retail site. This indicates that there is no significant difference if opening of the new retail site at any one of these four states. However, there is significant difference if the new retail site is opened at Pennsylvania or New Jersey. In addition, the rank of the states based on the second quartile (i.e., median) in Figure 7 is generally consistent with the results in Figure 5. In this way, not only a set of back-up plans for opening of the new retail site is provided in Figure 7, but also we can thoroughly compare the risks in the estimated total amount of business if the new retail site is opened at any one of the two states by looking at the percentage of overlapped bar length. This uncertainty study indicates that the proposed retail site selection method is relatively robust to the data uncertainties. It also shows that the impact of data uncertainty on the estimated total amount of business is not the same if the retail site is opened at different states.

CONCLUSION

With the massive growth of "big data" in business operations, it becomes vitally important to take advantage of the business analytics tools to effectively assess, analyze and visualize the data to transform from data-rich into decision-smart. Three fundamental challenges in business data analytics are highlighted in this paper: (1) Business data often encounters quality issues and needs substantial cleaning efforts before making any of sense; (2) Business data is large in overall size but cannot be fully shared due to the concern of data security at the company; and (3) Business data often needs to be cross-referenced with public databases to reveal more information and knowledge. These three aforementioned challenging issues often limit the effectiveness of business data analytics. The main contribution of this article is to outline a systematic step-by-step procedure to transforming from data-rich into decision-smart for business data analytics. Specifically, a real case study involving choosing an optimal location for the opening of a new retail site is selected to illustrate and validate the proposed framework. The dataset considered in this study includes the company's proprietary transaction data and a variety of public databases such as population, number of gas stations, and zip code information in different states, in which the public databases are acquired explicitly for better use of the gravity model. A detailed illustration on how to obtain these relevant public databases is also given. Based on the identified problem and dataset for choosing a new retail site location, a new approach for implementing the gravity model is developed that bridges the proprietary and public databases together to enhance the data analytics process. Next, detailed analysis results are presented for data visualization and decision making. Finally, an uncertainty analysis is provided to understand the impacts of data uncertainty on the optimal retail site location selection.

This research establishes a generic guideline on leveraging data analytics tools for resolving business issues when dealing with business of big data. Such research efforts have rarely been done in the literature before. There are several important topics for future research that are related to this work. First, the current study projects the customer's location information into the state level. Once the best state location is determined, the next step is to further specify the location of the new retail site with a more refined scale. Second, the current location analysis focuses on the average of the amount of business sales within the last 5-year period. Further studies can be done to develop a “dynamic” gravity model to investigate the changes in the optimal retail site location at different time points. This may also lay a foundation for us to incorporate the future trends of market conditions into the location selection model.

REFERENCES

Abdelhafez, H. A. (2014). Big data technologies and analytics: A review of emerging solutions. *International Journal of Business Analytics*, *1*(2), 1–17. doi:10.4018/ijban.2014040101

Barton, D. (2012). Making advanced analytics work for you. *Harvard Business Review*, *90*(10), 78–83. PMID:23074867

Bruno, G., & Improta, G. (2008). Using gravity models for the evaluation of new university site locations: A case study. *Computers & Operations Research*, *35*(2), 436–444. doi:10.1016/j.cor.2006.03.008

Buckner, R. W. (1998). *Site selection: New advancements in methods and technology*. New York, NY: Chain Store Publishing Corp.

Cosentino, T. (2012). Tableau thrives in providing visual discovery for business analytics. Retrieved February 20, 2014, from http://www.ventanaresearch.com/blog/commentblog.aspx?id=3521

Drezner, T. (2009). Location of retail facilities under conditions of uncertainty. *Annals of Operations Research*, *167*(1), 107–120. doi:10.1007/s10479-007-0253-6

Drezner, T., & Drezner, Z. (2001). A note on applying the gravity rule to the airline hub problem. *Journal of Regional Science*, *41*(1), 67–72. doi:10.1111/0022-4146.00207

Drezner, T., & Drezner, Z. (2002). Validating the gravity-based competitive location model using inferred attractiveness. *Annals of Operations Research*, *111*(1-4), 227–237. doi:10.1023/A:1020910021280

Drezner, Z. (1995). *Facility location: A survey of applications and methods*. New York, NY: Springer. doi:10.1007/978-1-4612-5355-6

Frey, C. H., & Patil, S. R. (2002). Identification and review of sensitivity analysis methods. *Risk Analysis*, *22*(3), 553–578. doi:10.1111/0272-4332.00039 PMID:12088234

GasBuddy/OpenStore LLC. (2014). USA national gas price heat map. Retrieved February 14, 2014, from http://www.gasbuddy.com/gb_gastemperaturemap.aspx

Grosche, T., Rothlauf, F., & Heinzl, A. (2007). Gravity models for airline passenger volume estimation. *Journal of Air Transport Management*, *13*(4), 175–183. doi:10.1016/j.jairtraman.2007.02.001

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. New York, NY: Springer. doi:10.1007/978-0-387-84858-7

Helton, J. C., Johnson, J. D., Sallaberry, C. J., & Storlie, C. B. (2006). Survey of sampling-based methods for uncertainty and sensitivity analysis. *Reliability Engineering & System Safety*, *91*(10), 1175–1209. doi:10.1016/j.ress.2005.11.017

Huff, D. L. (1963). A probabilistic analysis of shopping center trade areas. *Land Economics*, *39*(1), 81–90. doi:10.2307/3144521

Huff, D. L., & Jenks, G. F. (1968). A graphic interpretation of the friction of distance in gravity models. *Annals of the Association of American Geographers*, *58*(4), 814–824. doi:10.1111/j.1467-8306.1968.tb01670.x

IBM. (2011). The 2011 IBM tech trends report: the clouds are rolling in ... Is your business ready? Retrieved December 18, 2013, from http://ai.arizona.edu/mis510/other/2011IBMTechTrendsReport.pdf

Joseph, L., & Kuby, M. (2011). Gravity modeling and its impacts on location analysis. In H. A. Eiselt & Vladimir Marianov (Ed.), Foundations of location analysis (pp. 423-443). New York, NY: Springer. doi:10.1007/978-1-4419-7572-0_18

LaValle, S., Lesser, E., Shockley, R., Hopkins, M. S., & Kruschwitz, N. (2011). Big data, analytics and the path from insights to value. *MIT Sloan Management Review*, *52*(2), 21–31.

Liu, K., Jain, S., & Shi, J. (2013). Physician performance assessment using a composite quality index. *Statistics in Medicine*, *32*(15), 2661–2680. doi:10.1002/sim.5710 PMID:23280761

Lowe, J. M., & Sen, A. (1996). Gravity model application in health planning: Analysis of an urban hospital market. *Journal of Regional Science*, *36*(3), 437–461. doi:10.1111/j.1467-9787.1996.tb01111.x

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity. *McKinsey Global Institute*. Retrieved December 18, 2013, from http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

Nakanishi, M., & Cooper, L. G. (1974). Parameter estimate for multiplicative interactive choice model: Least squares approach. *JMR, Journal of Marketing Research*, *11*, 303–311. doi:10.2307/3151146

Năstase, P., & Stoica, D. (2010). A new business dimension - business analytics. *Accounting and Management Information Systems*, *9*(4), 603–618.

Ozuduru, B. H. (2013). Assessment of spatial dependence using spatial autoregression models: Empirical analysis of shopping center space supply in ohio. *Journal of Urban Planning and Development*, *139*(1), 12–21. doi:10.1061/(ASCE)UP.1943-5444.0000129

Pavolotsky, J. (2012). *Demystifying big data. Business Law Today*. American Bar Association.

Plastria, F. (1997). Profit maximizing single competitive facility location in the plane. *Studies in Locational Analysis*, *11*, 115–126.

Plastria, F. (2001). Static competitive facility location: An overview of optimization approaches. *European Journal of Operational Research*, *129*(3), 461–470. doi:10.1016/S0377-2217(00)00169-7

Plastria, F., & Vanhaverbeke, L. (2008). Discrete models for competitive location with foresight. *Computers & Operations Research*, *35*(3), 683–700. doi:10.1016/j.cor.2006.05.006

Reilly, W. J. (1931). *The law of retail gravitation*. New York, NY: Knickerbocker Press.

Rodrigue, J. P. (2012). Supply chain management, logistics changes and the concept of friction. In P. V. Hall & M. Hesse (Eds.), *Cities, Regions and Flows*. London: Routledge.

Sallam, R. L., Tapadinhas, J., Parenteau, J., Yuen, D., & Hostmann, B. (2014). *Magic Quadrant for Business Intelligence and Analytics Platforms*. Stamford, CT: Gartner Group.

Sasaki, H. (2014). Time lags related to past and current IT innovations in japan: An analysis of ERP, SCM, CRM, and big data trends. [IJBAN]. *International Journal of Business Analytics*, *1*(1), 29–42. doi:10.4018/ijban.2014010103

Schläfke, M., Silvi, R., & Möller, K. (2013). A framework for business analytics in performance management. *International Journal of Productivity and Performance Management*, *62*(1), 110–122. doi:10.1108/17410401311285327

Stodder, D. (2013). Data visualization and discovery for better business decisions. *TDWI Best Practices Report*, TDWI research.

Tavana, M., Trevisani, D. A., & Kennedy, D. T. (2014). A fuzzy cyber-risk analysis model for assessing attacks on the availability and integrity of the military command and control systems. *International Journal of Business Analytics*, *1*(3), 21–36. doi:10.4018/ijban.2014070102

U.S. Census Bureau Press Release. (2008). A gas station for every 2,500 people. Retrieved December 30, 2013, from http://www.census.gov/newsroom/releases/archives/county_business_patterns/cb08-96.html

Young, W. J. (1975). Distance decay values and shopping center size. *The Professional Geographer*, *27*(3), 304–309. doi:10.1111/j.0033-0124.1975.00304.x

Zip Code Database. (2014). United States zip code database (v1.0). Retrieved February 20, 2014, http://www.populardata.com/zipcode_database.html

Kaibo Liu received the B.S. degree in industrial engineering and engineering management from the Hong Kong University of Science and Technology, Hong Kong, China, in 2009, the M.S. degree in statistics and the Ph.D. degree in industrial engineering from the Georgia Institute of Technology, Atlanta, in 2011 and 2013, respectively. Currently, he is an assistant professor at the department of Industrial and Systems Engineering, University of Wisconsin-Madison, Madison. His research interests are data fusion for process modeling, monitoring, diagnosis and prognostics. Dr. Liu is a member of IEEE, INFORMS, IIE and ASQ.

Jianjun Shi received the B.S. and M.S. degrees in electrical engineering from the Beijing Institute of Technology, Beijing, China, in 1984 and 1987, respectively, and the Ph.D. degree in mechanical engineering from the University of Michigan, Ann Arbor, in 1992. Currently, he is the Carolyn J. Stewart Chair Professor in the H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta. His research interests include the fusion of advanced statistical and domain knowledge to develop methodologies for modeling, monitoring, diagnosis, and control for complex manufacturing systems. Dr. Shi is a Fellow of the IIE, a Fellow of ASME, a Fellow of INFORMS, an academician of the International Academy for Quality, an elected member of the ISI, and a life member of ASA.

APPENDIX

Table 1. Illustration of the data quality issues in the customer's location information (some of the information is removed due to data privacy)

Customer	Customer's Name	Location Details 1	Location Details 2	Location Details 3
1	---	--- Street	Atlanta, GA 30309	
2	---	--- Road	Suite E	Montgomery, Texas 77316
3	---	32218		
4	---	--- Street	27409	

Figure 1. An overview of the proposed framework for business data analytics

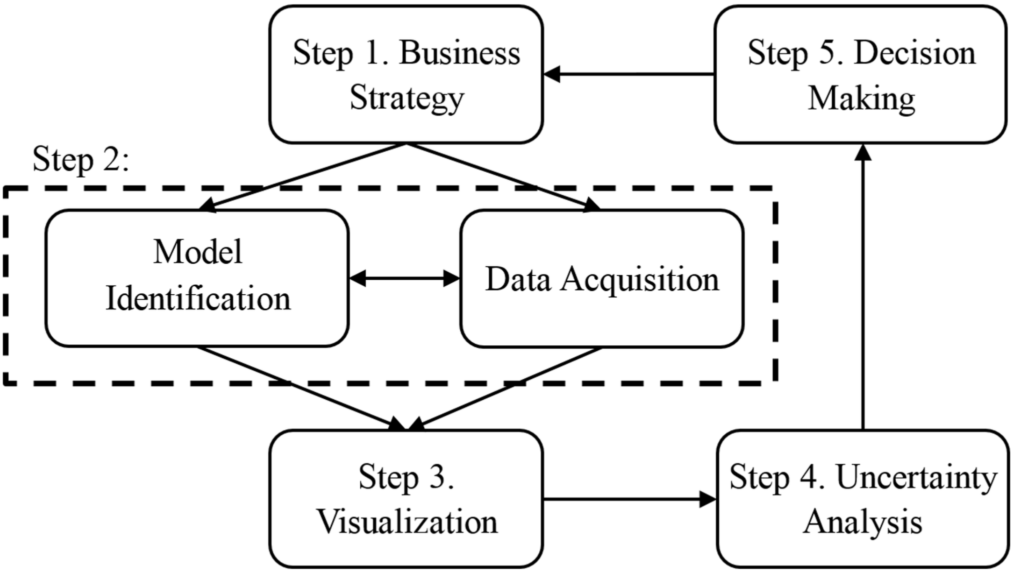


Figure 2. The amount of business of the company before opening of the new retail site (the exact number is removed in the legend due to data privacy)

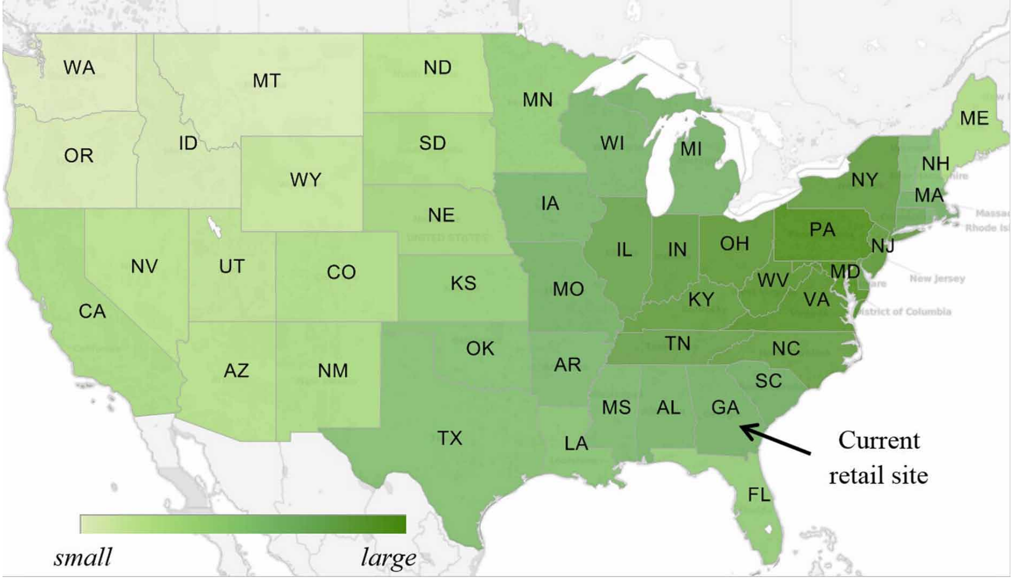


Figure 3. A positive linear relationship between the number of gas stations and the population data in each state

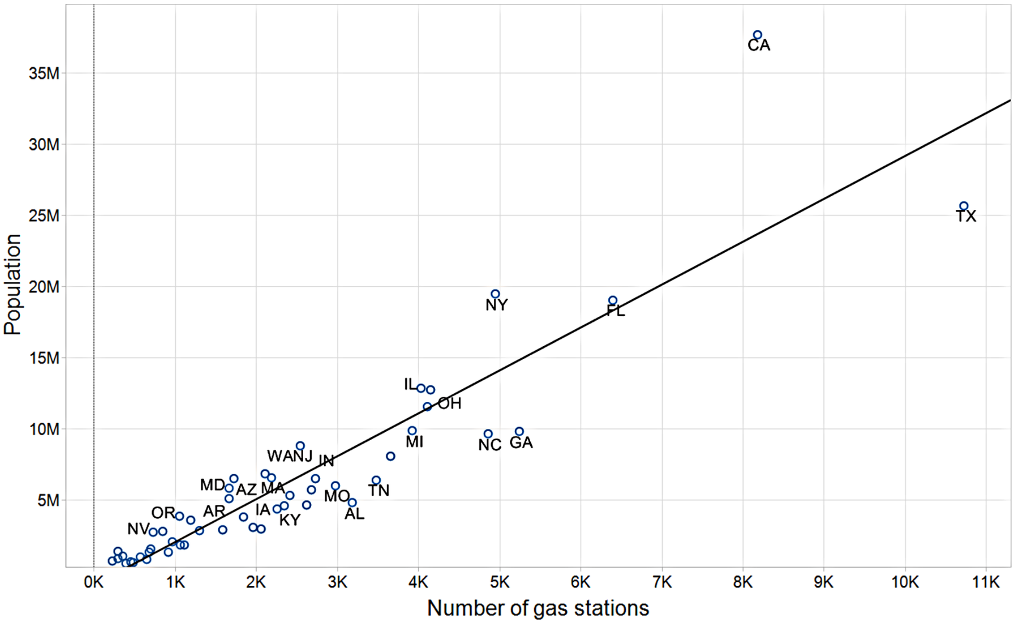


Figure 4. Estimated current market share of the company before opening of the new retail site

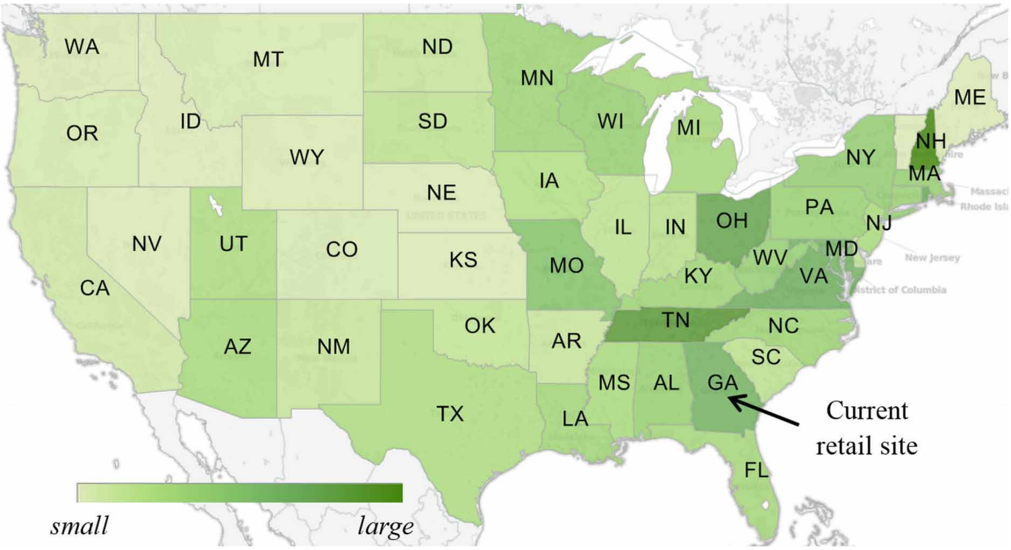


Figure 5. Comparisons of the estimated total amount of business of the company if opening of the new retail site at a different state each time (the exact number is removed due to data privacy)

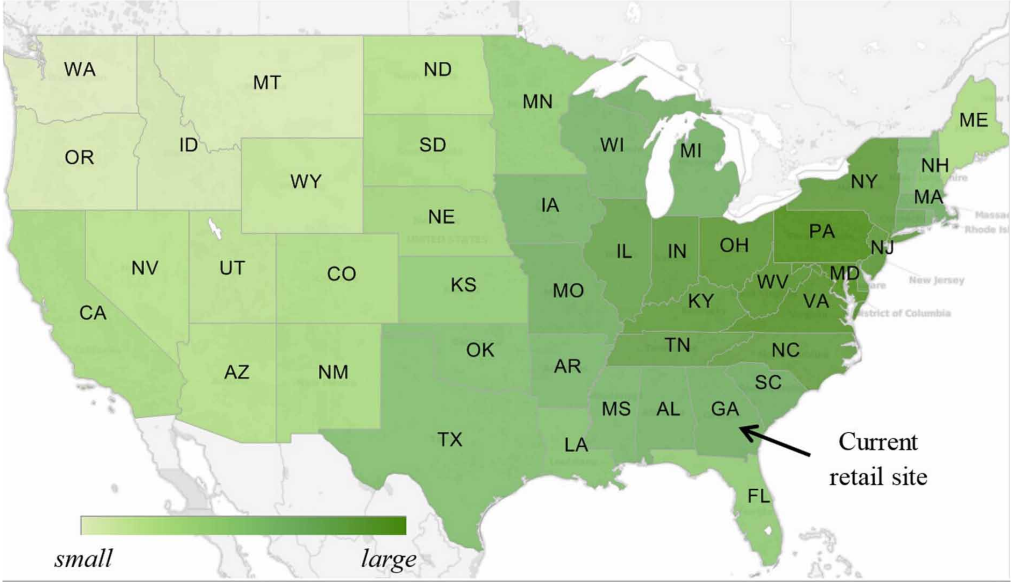


Figure 6. The number of gas stations at each state in 2007

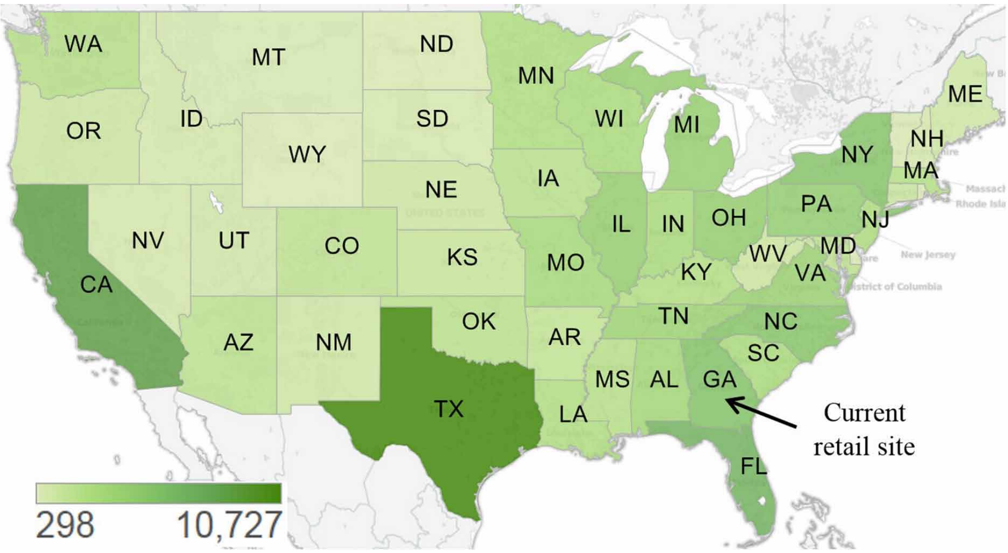
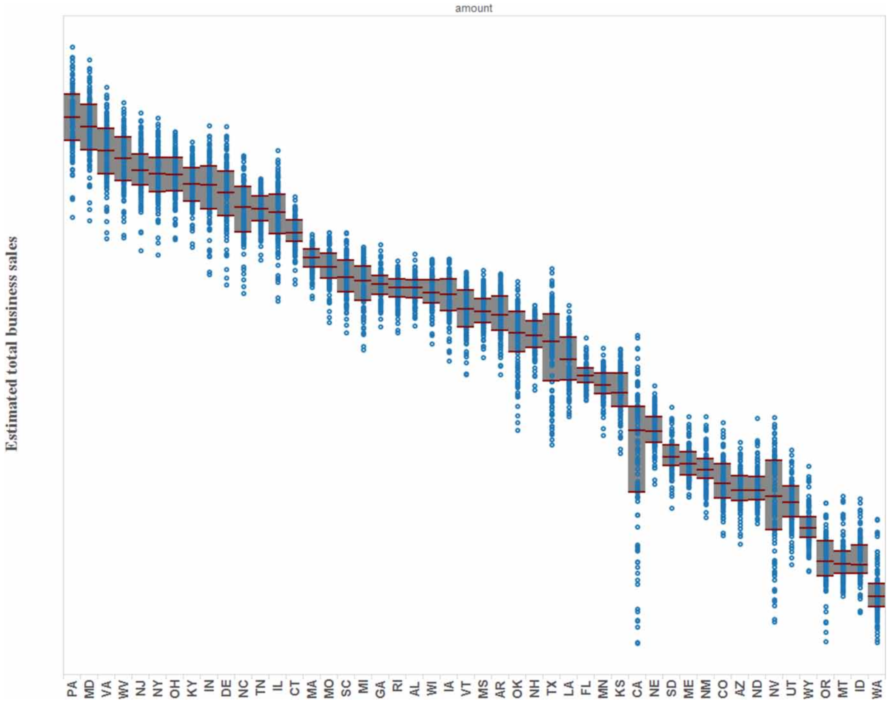


Figure 7. Uncertainty analysis of the estimated total business sales of the company after opening of the new retail site at a different state each time (the exact number is removed due to data privacy)


amount
Estimated total business sales
PA
MD
VA
WV
NJ
NY
OH
KY
IN
DE
NC
TN
IL
CT
MA
MO
SC
MI
GA
RI
AL
WI
IA
VT
MS
AR
OK
NH
TX
LA
FL
MN
KS
CA
NE
SD
ME
NM
CO
AZ
ND
NV
UT
WY
OR
MT
ID
WA