2018-11-30

Social Media Big Data Integration: a New Approach Based on Calibration

Dalla Valle, Luciana

10.1016/j.eswa.2017.12.044
Expert Systems with Applications
Elsevier

Social Media Big Data Integration: A New Approach Based on Calibration

Luciana Dalla Valle[1] (University of Plymouth)
luciana.dallavalle@plymouth.ac.uk

Ron Kenett (KPA Ltd, Neaman Institute, Technion and University of Turin)
ron@kpa-group.com

Abstract

In recent years, the growing availability of huge amounts of information, generated in every sector at high speed and in a wide variety of forms and formats, is unprecedented. The ability to harness big data is an opportunity to obtain more accurate analyses and to improve decision-making in industry, government and many other organizations. However, handling big data may be challenging and proper data integration is a key dimension in achieving high information quality. In this paper, we propose a novel approach to data integration that calibrates online generated big data with interview based customer survey data. A common issue of customer surveys is that responses are often overly positive, making it difficult to identify areas of weaknesses in organizations. On the other hand, online reviews are often overly negative, hampering an accurate evaluation of areas of excellence. The proposed methodology calibrates the levels of unbalanced responses in different data sources via resampling and performs data integration using Bayesian Networks to propagate the new re-balanced information. In this paper we show, with a case study example, how the novel data integration approach allows businesses and organizations to get a bias corrected appraisal of the level of satisfaction of their customers. The application is based on the integration of online data of review blogs and customer satisfaction surveys from the San Francisco airport. We illustrate how this integration enhances the information quality of the data analytic work in four of InfoQ dimensions, namely, Data Structure, Data Integration, Temporal Relevance and Chronology of Data and Goal.

Keywords: Bayesian Networks, Calibration, Data Integration, Social Media, Information Quality (InfoQ), Resampling Techniques.

1. Introduction

The growing availability of abundant masses of data in every sector, including business, government and health care, is posing new analytic and statistical challenges. This data may come from different sources such as posts in social media sites, digital pictures and videos, cell phone GPS, purchase transaction records and signal sensors used to gather climate information, to name a few. This is called *Big Data* and is characterized by high volume, variety and gathering velocity. Large quantities of information, mostly unstructured, are generated by social media, every minute. On the web, billions of individuals around the globe simultaneously produce, share and consume content generated by the user themselves. Through social media people express their opinions and sentiments towards specific topics, products and services, and the analysis of this information (called social media mining or sentiment analysis) may be key to organizations and businesses to monitor the satisfaction of their customers or to plan business initiatives or design new products and services.

[1] *Corresponding author*. School of Computing, Electronics and Mathematics, Drake Circus, PL4 8AA, Plymouth (UK), Tel: +44 (0)1752 586319

In recent years, advances in the literature of big data analysis have been significant. Amongst recent contributions to sentiment analysis, Stander et al. (2016a) and Stander et al. (2016b) extracted Facebook data to analyze sentiment scores and voting patterns about the June 2016 EU referendum in the UK. Zhang et al. (2011) used sentiment analysis techniques to predict stock market indicators using Twitter data. Asur and Huberman (2010) predicted box-office movie revenues, performing an analysis of sentiments from comments posted on social media.

However, big data analysis and social media mining may be challenging. The main issues are related to the quality of data collected and reported and to the integration of multiple datasets. The quality of information generated from big data is dependent on the quality of data collected and the robustness of the measures or indicators used. The lack of standardized quality measures and indicators can make comparisons difficult. Moreover, the quality of big data is often compromised by the presence of biased information, which may include fake data and fabricated news stories (BBC News, 2017). In particular, social media big data often contain biased information, especially online blogs describing opinions and sentiments about specific products and services. Indeed, online reviews generally include overly negative comments and feedback, since users tend to feel more free to express their dissatisfaction online, rather than in other contexts. On the other hand, traditional reviews generally include overly positive comments, since people tend not to feel comfortable to voice their opinions in surveys and may not be completely honest about their discontent. In both cases, the levels of the variables expressing customers' views are (sometimes strongly) unbalanced, preventing a correct evaluation of customer satisfaction.

In handling these challenges, data integration is key, especially where data come in both structured and unstructured formats and need to be integrated from disparate sources stored in systems managed by different departments. In most cases, the efficient aggregation and correlation of multiple datasets of considerable dimensions may be very complex (Daniel, 2015). Effective data integration is crucial for analysts and decision makers, since it can provide a broader picture of the problem at hand, avoiding biased results and misleading conclusions. For example, while the analysis of polls data failed to predict the election of Donald Trump in November 2016, data extracted from Facebook correctly predicted the winner (The Economist, 2016).

Dalla Valle and Kenett (2015) show how nonparametric Bayesian Networks (BNs) can be successfully used to integrate data coming from different sources, including official statistics, and to enhance information quality (Kenett and Shmueli, 2016). The aim of this paper is to propose a novel methodology to integrate customer satisfaction surveys and online review data, based on resampling techniques and BNs. Our methodology calibrates the sentiments of online uses with customer surveys using resampling to re-balance variable levels in the data. BNs are then used to propagate calibrated information and perform data integration. This approach allows businesses and organizations to correctly analyze the sentiments of online users on social media, facilitating an accurate evaluation of the satisfaction of their customers. We will illustrate that the proposed big data integration methodology enhances the information quality of the study in four dimensions, namely, Data Structure, Data Integration, Temporal Relevance and Chronology of Data and Goal.

The remainder of this paper is organized as follows: Section 2 is an overview of the literature of big data, information quality, social media mining and data integration; in Section 3 we introduce BNs; Section 4 illustrates the novel big data integration methodology; Section 5 presents the airport passengers' datasets used in our case-study; Section 6 shows the application of our methodology to the passengers' data; concluding remarks are given in Section 7.

2. Big Data and social media mining

Big data consists of data sets of extremely huge size and moving extremely fast, thus exceeding the processing capacity of conventional database systems (Manyika et al., 2011). The opportunities for gaining valuable new insights analyzing and harnessing big data are vast. In order to successfully exploit big data, many organizations are developing new analytics methods to make informed decisions about their strategic and operational directions. The term *analytics* includes a wide variety of mathematical, statistical and computational tools that can turn complex big data into meaningful patterns and value. As stated by Peter Sondergaard, Senior Vice president at Gartner Research, "information is the oil of the 21st century, and analytics is the combustion engine". However, the development of suitable analytics methods to harness big data is challenging. The changing nature of the information available to most organizations leads to complications in managing the volumes and analysis of data. While in the past most organizations handled exclusively structured data, currently 80% of the data (as estimated by IBM) generated are unstructured and come in a variety of formats such as text, video, audio, diagrams and images (Schneider, 2016). The characteristics of this new type of information being generated led to the introduction of proper definitions for the term *big data*. Douglas (2001) in the Gartner's report proposed a threefold definition of big data encompassing the *3Vs*:

a. *Volume*, indicating the increasing size of data, in the order of terabytes and beyond (e.g. the number of tweets created each day by social media users, the annual water meter readings of the households of a specific region);
b. *Velocity*, relating to the growing rate at which information is produced within an organization (e.g. the trade events monitored each day by a financial organization, the daily call detail records in real-time regarding customers' churn);
c. *Variety*, referring to data in diverse range of formats, both structured and unstructured (e.g. live video feeds from surveillance cameras, images and documents uploaded daily on social media platforms).

Later, the definition of big data was expanded by IBM into the *4Vs*, which includes *Veracity* as an additional complementary characteristic of big data, referring to the biases, noise, abnormality, quality issues and uncertainty in the data (e.g. opinion spam on review sharing websites, false illness trends on social network webpages). More recently, a fifth *V* was added, leading to the *5Vs* big data definition, which adds *Value* to the previous *4Vs* (Chen and Zhang, 2014), denoting the ability to generate benefits and value through insights gained by analytics (e.g. the millions of dollars saved by aircraft engine manufacturers using analytics to predict engine events that lead to costly airline disruptions).

The information quality dimensions proposed by Kenett and Shmueli (2014) provide a more general framework than the 5Vs in a wider context. Specifically, information quality (InfoQ) is defined as the potential of a dataset to achieve a specific (scientific or practical) goal using a given empirical analysis method. InfoQ is different from data quality and analysis quality, but is dependent on these components and on the relationship between them. Formally, the definition is

$$InfoQ = U(X, f \mid g),$$

where X is the data, f the analysis method, g the goal and U the utility function.
A key requirement for determining InfoQ is therefore the nature of the study goal. In particular, we distinguish between explanatory, predictive and descriptive goals. An explanatory goal is one that is based on causal hypotheses or seeks causal answers ("does higher income improve satisfaction?"). A predictive goal is aimed at predicting future or new individual observations ("predict the satisfaction level for 100 people, given their income"). A descriptive goal is aimed at quantifying an observed effect using a statistical or other approximation ("how do income levels and satisfaction correlate?"). To assess the level of InfoQ in a particular study, Kenett and Shmueli (2016) propose, with many examples, 8 dimensions of InfoQ:

a. *Data Resolution*: The measurement scale and level of aggregation of the data relative to the task at hand must be adequate for the purpose of the study.

b. *Data Structure*: The data can combine structured quantitative data with unstructured, semantic based data.
c. *Data Integration*: Data is often spread out across multiple data sources. Hence, properly identifying the different relevant sources, collecting the relevant data, and integrating the data, directly affect information quality. In this work, we focus on big data integration.
d. *Temporal Relevance*: A data set contains information collected during a certain time window. The degree of relevance of the data in that time window to the current goal at hand must be assessed.
e. *Chronology of Data and Goal*: Depending on the nature of the goal, the chronology of the data can support the goal to different degrees.
f. *Generalizability*: Two types of generalizability are statistical and scientific generalizability. Statistical generalizability refers to inferring from a sample to a target population. Scientific generalizability refers to applying a model based on a particular target population to other populations.
g. *Operationalization*: Observable data are a construct operationalization of underlying concepts. Action operationalization is about deriving concrete actions from the information provided by a study.
h. *Communication*: If the information does not reach the right person at the right time in a clear and understandable way, then the quality of information becomes poor.

There is some overlap between the 5Vs and the 8 InfoQ dimensions. Volume is related to Data Resolution, Variety is exactly Data Structure, Velocity is part of Chronology of Data and Goal and Value is determined by the Utility, one of the InfoQ components. As one can see the InfoQ framework is wider. In this work, we propose a methodology for Data Integration in the context of big data, and we focus on information produced and communicated by social media.

Social media are amongst the most prolific generators of big data and allow billions of people all around the world to daily interact, post and share contents and give spontaneous feedback on specific topics. Social media is a group of internet-based applications that allow the creation and exchange of user-generated content (Kaplan and Haenlein, 2010), which can be defined as work published on a publicly accessible website, implying a certain amount of creative effort for its production or adaptation of existing work, and generally created outside of professional routines and practices (OECD, 2007). As opposed to traditional media such as newspapers, books and television, social media is freely accessible, allowing everyone to publish contents and controlling how the information is generated and shared. There are numerous categories of social media: social networking (Facebook, Google+, LinkedIn), microblogging (Twitter), reviews sharing (Amazon, TripAdvisor, Yelp), wiki websites and databases (Wikipedia, GitHub, IMDb), photo sharing (Flickr, Instagram), slides sharing (SlideShare), video sharing (YouTube, Vimeo), livecasting (Periscope) and many others. Social media information is largely unstructured and requires innovative social media mining solutions. Social media mining encompasses the tools to formally represent, measure, model and mine information from large-scale social media data. It includes methodologies from different disciplines, such as data mining, machine learning, sentiment analysis, social network analysis, sociology, statistics, optimization and mathematics (Zafarani et al., 2014). Social media mining allows us to understand complex social phenomena and perform predictions based on *sentiments*, which are expressions of the online opinions, feelings and views of social media users. The process of detecting, extracting, analyzing and classifying the opinions and sentiments of people concerning different topics, as expresses in textual input, is called sentiment analysis (Montoyo et al., 2012). The majority of the contributions in the literature of sentiment analysis are focused on sentiment classification, which is the determination of the orientation of sentiments of a given text in two or more classes (e.g. positive and negative instances or positive, negative and neutral instances). Generally, sentiment classification is implemented using decision trees, support vector machines, neural networks, naïve Bayes and maximum entropy. Another promising, but still underrepresented area of application of sentiment analysis is the measurement of review usefulness, which analyzes online reviews with the purpose of helping customers in making better product or service choices. Ghose and Ipeirotis (2011) identified several features to measure the helpfulness of a review and

observed that subjectivity, informativeness, readability and linguistic correctness in reviews affect sales and perceived usefulness of products. In addition, Krishnamoorthy (2015) developed a predictive model to measure the helpfulness of reviews considering linguistic, readability and subjectivity features.

As pointed out previously, the critical challenges associated to social media mining, such as the lack of effective data integration methodologies, may prevent a broader use of social media data. Foresti et al. (2012) agree that data aggregation from multiple information sources is key to decision-makers and describe a regression-based data integration methodology applied to public and private financial databases. Dalla Valle (2016) illustrates a different approach for blending information from official statistics and organizational data, based on the generalization of Heckman's method where inference is performed according to the Bayesian framework. Dong and Srivastava (2015) describe the big data integration techniques of schema mapping, record linkage and data fusion and identify a range of open problems in this research area. Chakraborty et al. (2015) define a novel approach to integrate diverse data types, such as historic data, survey data, management planning data, expert knowledge and incomplete data, by converting data into Bayesian probability forms. Dalla Valle (2014 and 2017a) and Dalla Valle and Kenett (2015) introduced an innovative approach to integrate survey data with official statistics data based on calibration using copulas and nonparametric BNs. For an overview about copulas and their applications to finance, see Dalla Valle (2017b and 2017c) and references therein. In this paper, we propose a novel methodology that calibrates social media information with specific datasets via resampling and performs integration using BNs. Such an integration, combining different overlapping data sources, enhances the information quality of the data analytic work. The next section introduces Bayesian networks.

3. Bayesian Networks: An Introduction

BNs implement a graphical model structure known as a directed acyclic graph (DAG) that is popular in statistics, machine learning and artificial intelligence. BNs enable an effective representation and computation of the joint probability distribution over a set of random variables (Pearl, 1985). The structure of a DAG is defined by two sets: the set of nodes and the set of directed arcs. The nodes represent random variables and are drawn as circles labelled by the variables names. The arcs represent links among the variables and are represented by arrows between nodes. In particular, an arc from node X_i to node X_j represents a relation between the corresponding variables. Thus, an arrow indicates that a value taken by variable X_j depends on the value taken by variable X_i. Node X_i is then referred to as a 'parent' of X_j and, similarly, X_j is referred to as the 'child' of X_i. This property is used to reduce the number of parameters that are required to characterize the joint probability distribution of the variables. This reduction provides an efficient way to compute the posterior probabilities given the evidence present in the data (Pearl, 2009, Jensen, 2001, Ben Gal, 2007, Koski and Noble, 2009, Pourret et al, 2008). In addition to the DAG structure, which is often considered as the *qualitative* part of the model, a BN includes *quantitative* parameters. These parameters are described by applying the Markov property, where the conditional probability distribution at each node depends only on its parents. For discrete random variables, this conditional probability is represented by a table, listing the local probability that a child node takes on each of the feasible values – for each combination of the values of its parents. The joint distribution of a collection of variables is determined uniquely by these local conditional probability tables.
Formally, a Bayesian Network *B*, is a DAG that represents a joint probability distribution over a set of random variables ***V***. The network is defined by a pair $B = \langle G, \Theta \rangle$, where *G* is the directed acyclic graph whose nodes $X_1, X_2, ..., X_n$ represents random variables, and whose edges represent the direct dependencies between these variables. The graph *G* encodes independence assumptions, by which each variable X_i is independent of its non-descendants given its parents in *G*, denoted generically as π_i . The second component Θ denotes the set of parameters of the

network. This set contains the parameter $\theta_{x_i|\pi_i} = P_B(x_i|\pi_i)$ for each realization x_i of X_i conditioned on π_i, the set of parents of X_i in *G*. Accordingly, *B* defines a unique joint probability distribution over ***V***, namely:

$$P_B(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} P_B(X_i|\pi_i) = \prod_{i=1}^{n} \theta_{X_i|\pi_i}.$$

In learning the network structure, one can include *white lists* of forced causality links imposed by expert opinion and *black lists* of links that are not to be included in the network. For examples of BN applications to study management efficiency, web site usability, operational risks, biotechnology, customer satisfaction surveys, healthcare systems and testing of web services see, respectively, Kenett et al. (2008), Kenett et al. (2009), Kenett and Raanan (2010), Peterson and Kenett (2011), Kenett and Salini (2011), Kenett (2012, 2016, 2017) and Bai et al. (2012). For examples of applications of BNs to education, banking, forensic and official statistics see Pietro et al. (2014), Tarantola et al. (2012), Di Zio et al. (2005), Vicard et al. (2008), Marella and Vicard (2013).

3.1 Parameter Learning

To fully specify a BN, and thus represent joint probability distributions, it is necessary to specify for each node *X* the probability distribution for *X* conditional upon *X*'s parents. The distribution of *X*, conditional upon its parents, may have any form with or without constraints.
These conditional distributions include parameters which are often unknown and must be estimated from data, for example using maximum likelihood. Direct maximization of the likelihood (or of the posterior probability) is usually based on the expectation-maximization (E-M) algorithm which alternates computing expected values of the unobserved variables conditional on observed data, with maximizing the complete likelihood assuming that previously computed expected values are correct. Under mild regularity conditions this process converges to maximum likelihood (or maximum posterior) values of parameters (Heckerman, 1995).

A Bayesian approach treats parameters as additional unobserved variables and computes a full posterior distribution over all nodes conditional upon observed data, and then integrates out the parameters. This, however, can be expensive and lead to large dimension models, and in practice classical parameter-setting approaches are more common (Neapolitan, 2003).

3.2 Structure Learning

BNs can be specified by expert knowledge (using white lists and black lists) or learned from data, or in combinations of both. The parameters of the local distributions are learned from data, priors elicited from experts, or both. Learning the graph structure of a BN requires a scoring function and a search strategy. Common scoring functions include the posterior probability of the structure given the training data, the Bayesian information criterion (BIC) or Akaike information criterion (AIC). When fitting models, adding parameters increases the likelihood, which may result in over-fitting. Both BIC and AIC resolve this problem by introducing a penalty term for the number of parameters in the model, with the penalty term being larger in BIC than in AIC. The time requirement of an exhaustive search, returning back a structure that maximizes the score, is super-exponential in the number of variables. A local search strategy makes incremental changes aimed at improving the score of the structure. A global search algorithm like Markov Chain Monte Carlo (MCMC) can avoid getting trapped in local minima. A partial list of structure learning algorithms includes Hill-Climbing with score functions BIC and AIC, Grow-Shrink, Incremental Association, Fast Incremental Association, Interleaved Incremental Association, hybrid algorithms and Phase Restricted Maximization. For more on BN structure learning see Musella (2013).

3.3 Causality and Bayesian Networks

Causality analysis has been studied from two main different points of view, the *probabilistic* view and the *mechanistic* view. Under the probabilistic view, the causal effect of an intervention is judged by comparing the evolution of the system when the intervention is and when it is not present. The mechanistic point of view focuses on understanding the mechanisms determining how specific effects come about. The interventionist and mechanistic viewpoints are not mutually exclusive. For example, when studying biological systems, scientists carry out experiments where they intervene on the system by adding a substance or by knocking out genes. However, the effect of a drug product on the human body cannot be decided only in the laboratory. A mechanistic understanding based on pharmacometrics models is a preliminary condition for determining if a certain medicinal treatment should be studied in order to elucidate biological mechanisms used to intervene and either prevent or cure a disease. The concept of potential outcomes is present in the work on randomized experiments by Fisher and Neyman in the 1920s and was extended by Rubin in the 1970s to non-randomized studies and different modes of inference (Mealli et al. 2012). In their work, causal effects are viewed as comparisons of potential outcomes, each corresponding to a level of the treatment and each observable, had the treatment taken on the corresponding level with at most one outcome actually observed, the one corresponding to the treatment level realized. In addition, the assignment mechanism needs to be explicitly defined as a probability model for how units receive the different treatment levels. With this perspective, a causal inference problem is viewed as a problem of missing data, where the assignment mechanism is explicitly modelled as a process for revealing the observed data. The assumptions on the assignment mechanism are crucial for identifying and deriving methods to estimate causal effects (Frosini, 2006). Imai et al. (2013) study how to design randomized experiments to identify causal mechanisms. They study designs that are useful in situations where researchers can directly manipulate the intermediate variable that lies on the causal path from the treatment to the outcome. Such a variable is often referred to as a ‘mediator’.

Causal BNs are networks where the effect of any intervention can be defined by a ‘do’ operator that separates intervention from conditioning. The basic idea is that intervention breaks the influence of a confounder so that one can make a true causal assessment. The established counterfactual definitions of direct and indirect effects depend on an ability to manipulate mediators. A BN graphical representation, based on local independence graphs and dynamic path analysis, can be used to provide an overview of dynamic relations (Aalen et al, 2012). In an econometric context, Heckman (2008) develops, as an alternative approach, explicit models of outcomes, where the causes of effects are investigated and the mechanisms governing the choice of treatment are analyzed. In such investigations, counterfactuals, which are possible outcomes in different hypothetical states of the world, are studied. The analysis of causality in studies of economic policies involves: (a) defining counterfactuals, (b) identifying causal models from idealized data of population distributions and (c) identifying causal models from actual data, where sampling variability is an issue.

Pearl developed BNs as the method of choice for reasoning in artificial intelligence and expert systems, replacing earlier ad hoc rule based systems. His extensive work covers topics such as: causal calculus, counterfactuals, Do calculus, transportability, missingness graphs, causal mediation, graph mutilation and external validity (Pearl, 1988). In a heated head to head debate between probabilistic and mechanistic view, Pearl has taken strong standings against the probabilistic view, see for example the paper by Baker (2013) and discussion by Pearl, (2013). The work of Aalen et al. (2012) and Imai et al. (2013) show how these approaches can be used in complementary ways. For more examples of BN applications see Fenton and Neil (2011, 2012, 2014).

4. Data Integration Methodology of Social Media with Survey Data

The methodology proposed in this paper aims at achieving data integration of traditional customer satisfaction survey data with social media data via resampling using BNs, expanding the approach presented in Dalla Valle and Kenett (2015). We perform data integration emphasizing blog-type data, which is a big data environment source. However, our approach is scalable to other social media and big data sources. As mentioned above, properly handling data integration is a key dimension in achieving high information quality (Kenett and Shmueli, 2016).

Self-declared or interview-based surveys are prime research tools in many application areas such as social science research, marketing, service management, risk management and customer satisfaction management. Measuring customer satisfaction is typically based on self-declared or interview-based questionnaires where users or consumers are asked to express opinions on statements, or satisfaction scales, mapping out various interactions with the service provider or product supplier. Customer satisfaction is a key dimension driving business outcomes and performance of processes in service and product organizations (Kenett and Salini, 2011).
BNs are powerful tools for analyzing customer satisfaction surveys, since they provide a visual cause and effect map, or DAG, of the survey variables and show clearly what variable affects customer satisfaction. BNs have several advantages compared to other data modelling techniques, since they can encode and visualize dependencies among all variables, they can be used to learn causal relationships and, since they incorporate both causal and probabilistic semantics, they can combine prior knowledge and data (Heckerman, 1997). BNs can be therefore used effectively to identify the drivers of customer satisfaction, producing knowledge that provides insights to managers and specialists and contributing to decision analysis and decision support systems.
However, interview-based surveys present some drawbacks, that may affect the correct identification of the main determinants of customer satisfaction. One of the main issue related traditional customer surveys is that interviewees are not always honest about their judgements and tend to provide ratings that are biased towards the positive side. Therefore, customers' responses are often unbalanced, with a very low proportion of negative ratings. However, it is key to organizations to correctly identify disappointed customers, to understand the reasons behind their dissatisfaction and to improve their services. On the other hand, social media information, such as online blogs and reviews, often contain a higher proportion of negative comments and feedback, since users tend to feel more free to express their opinions online rather than in other contexts. Sometimes, online reviews are biased towards the negative side, making it difficult to identify the determinants of customers' satisfaction and, hence, areas of excellence within an organization. The integration of traditional surveys with social media data allows us to better model both groups of satisfied and dissatisfied customers, improving our understanding of their motivations by incorporating information that are only present in one of the datasets. The implementation of BNs to integrated data builds a network of causal relationships between variables, which allows organizations to correctly identify the main drivers of customer satisfaction, leading to the improvement of their services and the enhancement of the overall satisfaction of their customers.

The proposed data integration methodology aggregates customer survey data with information extracted from social media, performing calibration of different datasets. The idea is in the same spirit of external benchmarking used in small area estimation (Pfeffermann, 2013). In small area estimation benchmarking robustifies the inference by forcing the model-based predictors to agree with a design-based estimator. Similarly, our methodology is based on qualitative data calibration performed via resampling, where the variables levels are balanced and customer survey estimates are updated to agree with more timely social media data estimates.
Calibration is implemented by altering the class distribution of customers' reviews in one of the datasets to obtain a re-balanced sample, which reflects the distribution of the second dataset. This approach involves the selection of a calibration link variable and the creation of a new artificial data set by suitably resampling the observations belonging to the classes of the

calibration link. In particular, the calibration link variable is resampled by oversampling with replacement the minority class and by undersampling without replacement the majority class.

More formally, the resampling approach can be described as follows. Let us consider the variables denoted by the pairs ($\boldsymbol{x}$, y), where $\boldsymbol{x}$ represents a set of measured characteristics and y is a target (or key) variable. Here, we consider the specific case where $\boldsymbol{x}$ is defined in a d-dimensional space $\mathcal{X}$being the product set between discrete domains, and the target variable y, which is affected by class imbalance, takes values in the categorical domain $\mathcal{Y} = \{\mathcal{Y}_{min}, \mathcal{Y}_{maj}\}$, where $\mathcal{Y}_{min}$ is the minority class and $\mathcal{Y}_{maj}$ is the majority class.
Suppose that a sample $D_n = (\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$, of the pairs ($\boldsymbol{x}$, y), whose generic row is $(\boldsymbol{x}_i, y_i)$, $i = 1, \ldots, n$, is observed on n individuals or objects. The class labels y_i belong to the set $\{\mathcal{Y}_{min}, \mathcal{Y}_{maj}\}$ and $\boldsymbol{x}_i$ are some related attributes supposed to be realizations of a random vector $\boldsymbol{x}$. Let the number of units in class $\mathcal{Y}_j$, $j = min, maj$, be denoted by $n_j < n$ and the corresponding class proportions be denoted by $p_j = n_j/n$. The resampling procedure for generating a new artificially re-balanced dataset, consists of the following steps:

1) Select $y^* = \mathcal{Y}_j$ with probability 1/2.
2) Select $(\boldsymbol{x}_i, y_i) \in D_n$, such that $y_i = y^*$, with probability $1/n_j$.
    a. If $y^* = \mathcal{Y}_{min}$, oversample with replacement by adding $(\boldsymbol{x}_i, y^*)$ to D_n;
    b. If $y^* = \mathcal{Y}_{maj}$, undersample without replacement by removing $(\boldsymbol{x}_i, y^*)$ from D_n.

Repeat steps 1 and 2 until the desired class proportions are achieved or until the minority class reaches the desired size.
This procedure produces a new rebalanced dataset D_m^*, of size m, where the desired proportions of observations belong to the two classes. For more details about the class imbalance problem and resampling techniques see, for example, Chawla (2005) and Menardi and Torelli (2014).
In the present work, the resampling approach described above is applied to interview- and online-based imbalanced datasets to achieve data integration. Following this bias correction, BNs are built to identify the main determinants of customer satisfaction.

The proposed data integration methodology is structured in three phases, represented in Figure 1:

1) *Data structure modelling*. Let D^{SU} denote the interview-based survey dataset and D^{SM} denote the social media dataset. This phase consists in implementing BNs to construct the causal relationships between the variables of both the customer survey, D^{SU}, and social media, D^{SM}, datasets, separately. BNs are chosen amongst other data modelling techniques for their flexibility and ability to encode probabilistic relationships among variables of interest, allowing an easy identification of the determinants of customer satisfaction. However, the presence of unbalanced samples can affect the correct assessment and evaluation of customer satisfaction and may lead to misleading conclusion. Data integration, implemented by rebalancing the unbalanced levels of D^{SU} with the levels of D^{SM} (or viceversa), allows us to accurately analyze customer satisfaction.

2) *Identification of the calibration link*. In the second phase a calibration link, in the form of one or more unbalanced key variables, is identified between customer survey and social media data. Denoting with $(\boldsymbol{x}^{SU}, y^{SU})$ the variables of D^{SU} and $(\boldsymbol{x}^{SM}, y^{SM})$ the variables of D^{SM}, then let y^{SU} be the calibration link of D^{SU} and y^{SM} be the calibration link of D^{SM}. We suppose that calibration links are unbalanced variables, with y^{SU} taking values in the categorical domain $\mathcal{Y}^{SU} = \{\mathcal{Y}_{min}{}^{SU}, \mathcal{Y}_{maj}{}^{SU}\}$, with proportions $p^{SU} = \{p_{min}{}^{SU}, p_{maj}{}^{SU}\}$, and y^{SM} in $\mathcal{Y}^{SM} = \{\mathcal{Y}_{min}{}^{SM}, \mathcal{Y}_{maj}{}^{SM}\}$, with proportions $p^{SM} = \{p_{min}{}^{SM}, p_{maj}{}^{SM}\}$, where $\mathcal{Y}_{min}{}^{SU}$ and $\mathcal{Y}_{min}{}^{SM}$ are the minority classes and $\mathcal{Y}_{maj}{}^{SU}$ and $\mathcal{Y}_{maj}{}^{SM}$ the majority classes of the interview- and blog-based surveys. Calibration links can be target variables expressing overall satisfaction or can be other variables influencing the overall satisfaction.

3) *Performing calibration*. In the last phase calibration is performed by suitably resampling the datasets, based on the distribution of the calibration link variables. In this phase, one of the dataset, for example D^{SU}, is rebalanced following the resampling approach described above, until $p^{SU} \approx p^{SM}$. Therefore, a new rebalanced dataset $D^{SU}*$ with the desired proportions of the calibration link variable will be generated. Similarly, calibration can be performed on D^{SM}, obtaining the new rebalanced dataset $D^{SM}*$. BNs are then updated for the re-balanced datasets $D^{SU}*$ or $D^{SM}*$, allowing the calibrated information to be propagated to achieve data integration. This approach will allow us to properly analyze customer satisfaction surveys and to achieve the goal of accurately identifying pockets of dissatisfaction and areas of excellence within an organization.

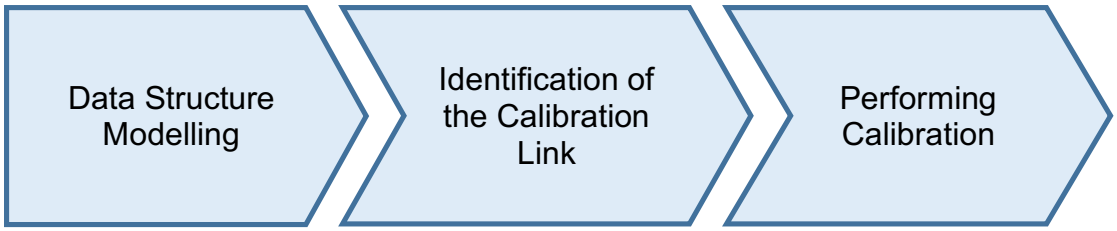

Figure 1: Graphical representation of the big data integration methodology

5. Case-Study: The Airport Passengers Datasets

We illustrate the application of the methodology by integrating airport passengers' data collected via interview-based survey with data extracted from an online review website. The context of this example is an analysis focused on improving the Temporal Relevance of a customer satisfaction survey by linking its results to online reviews that are continuously updated. The data integration methodology described here provides information to decision makers that is both up to date and comprehensive. In this sense, the Data Integration supports proper Chronology of Data and Goal. The example therefore enhances the information quality in four of the InfoQ dimensions: Data Structure, Data Integration, Temporal Relevance and Chronology of Data and Goal.

5.1 San Francisco International Airport Customer Survey

The first dataset we analyze is a subset of the 2016 customer survey administered to the passengers of San Francisco International Airport (SFO). The data are publicly available on the website http://www.flysfo.com/media/customer-survey-data

The passenger dataset contains information pertaining to customer demographics and satisfaction with airport facilities, services, and initiatives. The data was collected in May 2016 through interviews with 3,087 customers in each of SFO's terminals and boarding areas. Customers were asked to rate the airport in several categories, including cleanliness ratings. Additional data collected include customers' income, mode of arrival to the airport, travel style, and various other categories.

The SFO dataset comprises demographic and satisfaction variables, including a variable expressing customers' overall satisfaction, as described in Table 1.

Table 1: Variables of the SFO airport customer survey

Variable Name	Measurement Levels
• PEAK: type of flight	• 1 = domestic peak (domestic flights departing 8 am to 1 pm) • 2 = domestic off-peak (domestic flights departing before 8 am or after 1 pm) • 3 = international flights
• PURP: what is the main purpose of your trip today?	• 1 = business/work/job interview • 2 = pleasure/vacation/recreation • 3 = visit friends or relatives • 4 = school/school event • 5 = conference/convention • 6 = wedding/funeral/graduation/reunion • 7 = other (specify) • 10 = escorting others (children/elderly)/personal errands/medical purpose • 11 = military • 12 = volunteer/political/religious • 13 = moving/immigration/traveling between homes • 0 = blank/non-response
• AIRTRAIN: rating SFO air train • ART: rating SFO artwork and exhibitions • CLEAN: rating cleanliness of SFO • FOOD: rating SFO restaurants • OVERALL: rating SFO airport as a whole • STORE: rating SFO retail shops and concessions • SIGN: rating SFO signs and directions • SCREENS: rating SFO information on screens/monitors • WALKWAYS: rating SFO escalators/elevators/moving walkways • WIFI: rating SFO accessing and using free Wi-Fi	• 1 = unacceptable • 2 = poor • 3 = satisfactory • 4 = good • 5 = outstanding • 6 = never used • 0 = blank
• SAFE: how safe do you feel at SFO?	• 1 = not safe at all • 2 = unsafe • 3 = neutral • 4 = safe • 5 = extremely safe • 6 = don't know • 0 = blank
• PASSTHRU: passing through security and screening	• 1 = very difficult • 2 = difficult • 3 = average • 4 = easy • 5 = very easy • 6 = don't know • 0 = blank
• COUNTRY: country area of respondent	country name
• AGE: age of respondent	• 1 = under 18 • 2 = 18 - 24

	• 3 = 25 - 34 • 4 = 35 - 44 • 5 = 45 - 54 • 6 = 55 - 64 • 7 = 65 and over • 8 = don't know / refused • 0 = bank/multiple responses
• GENDER: gender of respondent	• 1 = male • 2 = female • 3 = other • 0 = blank/multiple responses
• INCOME: household income	• 1 = under $50,000 • 2 = $50,000 - $100,000 • 3 = $100,001 - $150,000 • 4 = over $150,000 • 5 = other currency (specify) • 0 = blank/multiple responses

As illustrated in Table 1, the satisfaction variables included in the SFO dataset express the passengers' judgements on a five-point scale. For comparison purposes, we transformed the original customers' ratings into dichotomous variables. However, the demographic variables were not transformed. In addition, we removed the observations containing missing data, the maximum percentage frequency being under 2%. The variables were dichotomized following two different schemes. The first of these schemes is called *BOT1+2* and it is constructed by aggregating customers who responded '1' or '2' (corresponding to extreme dissatisfaction and dissatisfaction, respectively). The second scheme is called *TOP5* and identifies customers who responded '5' (corresponding to extremely satisfied) on the five-point scale. BOT1+2 is very effective in identifying pockets of dissatisfaction and areas of improvements, while TOP5 emphasizes areas of excellence. For more on statistical analyses using the two dichotomizing schemes see Kenett and Salini (2011).

5.2 Skytrax Reviews Social Media Data

The second dataset, that we named Skytrax dataset, contains information extracted from the reviews published by passengers of the SFO airport on the website http://www.airlinequality.com. For comparative purposes, only recent reviews of SFO passengers were analyzed.

The dataset includes demographic and satisfaction variables, with judgements on individual characteristics and on the airport as a whole, as described in Table 2.

Table 2: Variables of the SFO airport social media review dataset

Variable Name	**Measurement Levels**
• COUNTRY: country area of passenger	country area name
• EXPERIENCE: airport experience	• 1 = arrival only • 2 = departure only • 3 = arrival and departure • 4 = transit
• TYPE: purpose of flight	• 1 = business • 2 = couple leisure • 3 = family leisure • 4 = solo leisure
• CLEAN: rating SFO cleanliness • FOOD: rating SFO restaurants • QUEUING: rating SFO queuing • SEATING: rating SFO seating • SHOPPING: rating SFO retail shops • SIGNS: rating SFO signs, directions and monitors • STAFF: rating SFO staff friendliness • WIFI: rating SFO accessing and using free Wi-Fi	• 1 = unacceptable • 2 = poor • 3 = satisfactory • 4 = good • 5 = outstanding • 0 = blank
• OVERALL: rating SFO airport as a whole	• 1 = unacceptable • 2 = very poor • 3 = poor • 4 = unremarkable • 5 = average • 6 = satisfactory • 7 = fair • 8 = good • 9 = very good • 10 = outstanding • 0 = blank
• RECOMMEND: would you recommend SFO to a friend?	• 1 = yes • 2 = no

For the sake of comparison, we applied the BOT1+2 and TOP5 dichotomization schemes to the Skytrax satisfaction variables. In order to implement the BOT1+2 scheme, the transformed OVERALL variable was constructed by aggregating the customers' responses '1' to '4'; while for the TOP5 scheme answers '9' and '10' were aggregated. Hence, the 3 phases of the data integration methodology illustrated in Section 4 were applied as to the BOT1+2 as to the TOP5 dichotomized Skytrax datasets.

6. Application of the Big Data Integration Methodology

After transforming the original data, we applied the three phases of the data integration methodology described in Section 4 to the SFO customer survey and to the Skytrax social media datasets using the BOT1+2 as well as the TOP5 dichotomization, as shown in Figure 2. As described in the diagram, initially, from SFO as well as Skytrax, two new datasets were generated according to the BOT1+2 and TOP5 dichotomization schemes. Then, the data integration methodology was applied twice: once to the BOT1+2 datasets and once to the TOP5 datasets, to illustrate the use of different calibration functions. In the first example, data integration was performed rebalancing the levels of the OVERALL variable of the SFO BOT1+2 dataset, while in the second example rebalancing was implemented on the QUEUING variable of the Skytrax TOP5 dataset.

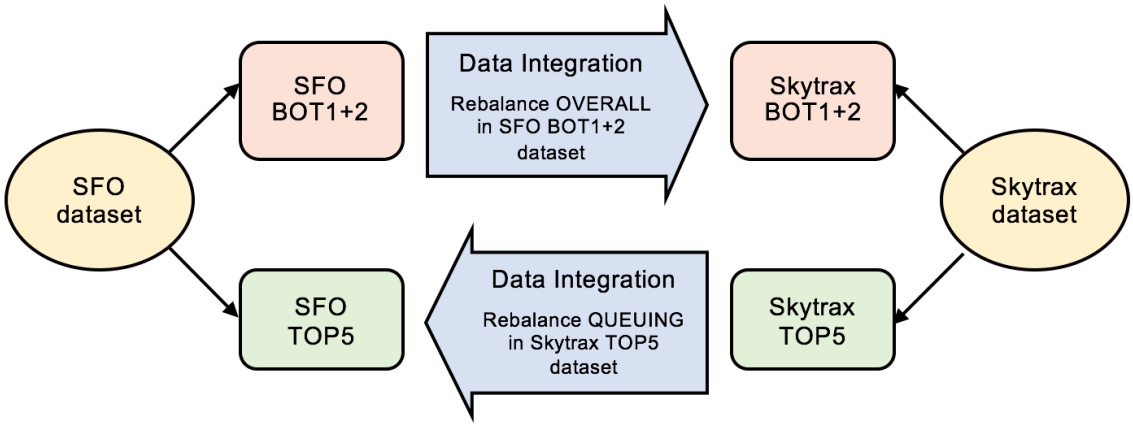

Figure 2: Diagram illustrating the application of the data integration methodology to the SFO and Skytrax datasets.

6.1 Data Integration of BOT1+2 Datasets

Data Structure Modelling

In the first phase of data integration, we analyzed the SFO customer satisfaction survey data with BNs, implemented using the GeNIe software V 2.1 (University of Pittsburgh, Pittsburgh, USA). Before starting the learning procedure, we incorporated prior information by building a blacklist, thus specifying a list of arcs which must be excluded from the network. For example, we constrained all arcs linking OVERALL with the other variables to be directed towards OVERALL. The networks were then built using the Bayesian Search structure learning algorithm implemented by GeNIe, which follows a hill climbing procedure with random restarts, guided by a heuristic scoring.

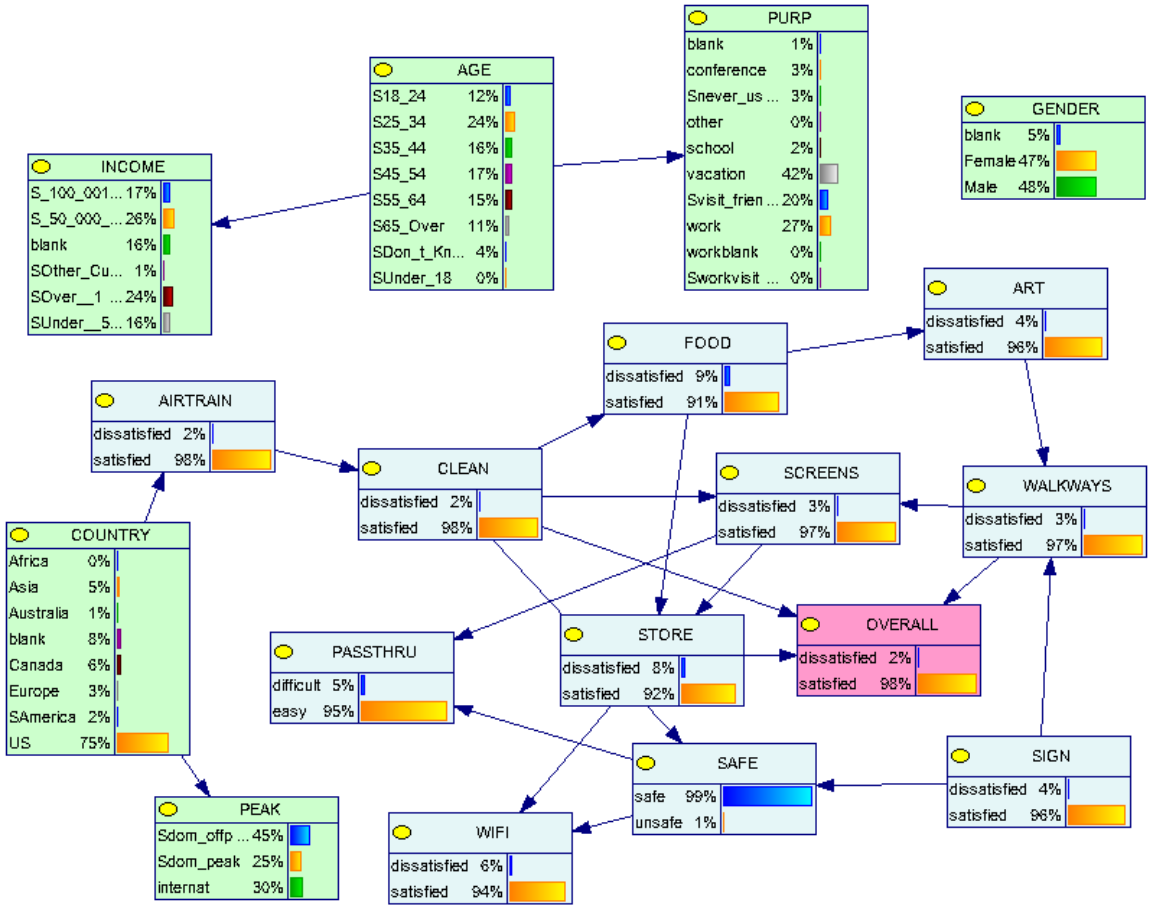

Figure 3: BN of the BOT1+2 SFO customer satisfaction survey dataset

Figure 3 shows the BN created using the dichotomized BOT1+2 SFO survey data. For clarity purposes, the target node, expressing overall satisfaction, is depicted in pink, the demographic information nodes are depicted in green, while the remaining satisfaction nodes are in blue. Most of the demographic nodes are not linked to satisfaction nodes, with the exception of COUNTRY, which influences AIRTRAIN. Hence, the contribution of demographic variables to determine areas of dissatisfaction is limited. The OVERALL node is directly linked to CLEAN, WALKWAYS and STORE. Therefore, customers' dissatisfaction is mainly determined by cleanliness, escalators/elevators/moving walkways and shops available at the airport. The implication is that improvements undertaken in the airport cleanliness, walkways and shops will reduce the proportion of disappointed passengers and increase overall satisfaction.

Note that there is a strong imbalance among the categories of the target variable OVERALL, since the 'dissatisfied' category (comprising extreme dissatisfaction and dissatisfaction in the BOT1+2 dichotomization), represents only the 2% of the interviewed customers. This situation is common in customer satisfaction surveys, since people tend to avoid expressing strong negative opinions. Several other variables in the dataset are also showing severe class imbalance, including the most influential determinants to the overall satisfaction. In particular, the percentage of passengers who are dissatisfied with cleanliness is 2%, those dissatisfied with walkways is 3% and those dissatisfied with shops is 8%. In this case, it is difficult to determine the motivations of dissatisfaction and identify areas where improvements are needed. In this paper, we propose a new data integration methodology, which addresses this issue, calibrating satisfaction information with online reviews.

In the first phase of the data integration methodology, we also implemented the data structure modelling using BNs to analyze the Skytrax BOT1+2 dichotomized dataset. We used the Greedy Thick Thinning structure learning algorithm, which consists in two steps: the thickening and the thinning step. The thickening step starts with an empty graph and then arcs that maximally increase the marginal likelihood are repeatedly added until no arc addition will result in a positive increase. Then, in the thinning step arcs are repeatedly removed until no arc deletion will result in a positive increase in the marginal likelihood. BNs were implemented using the GeNIe software.

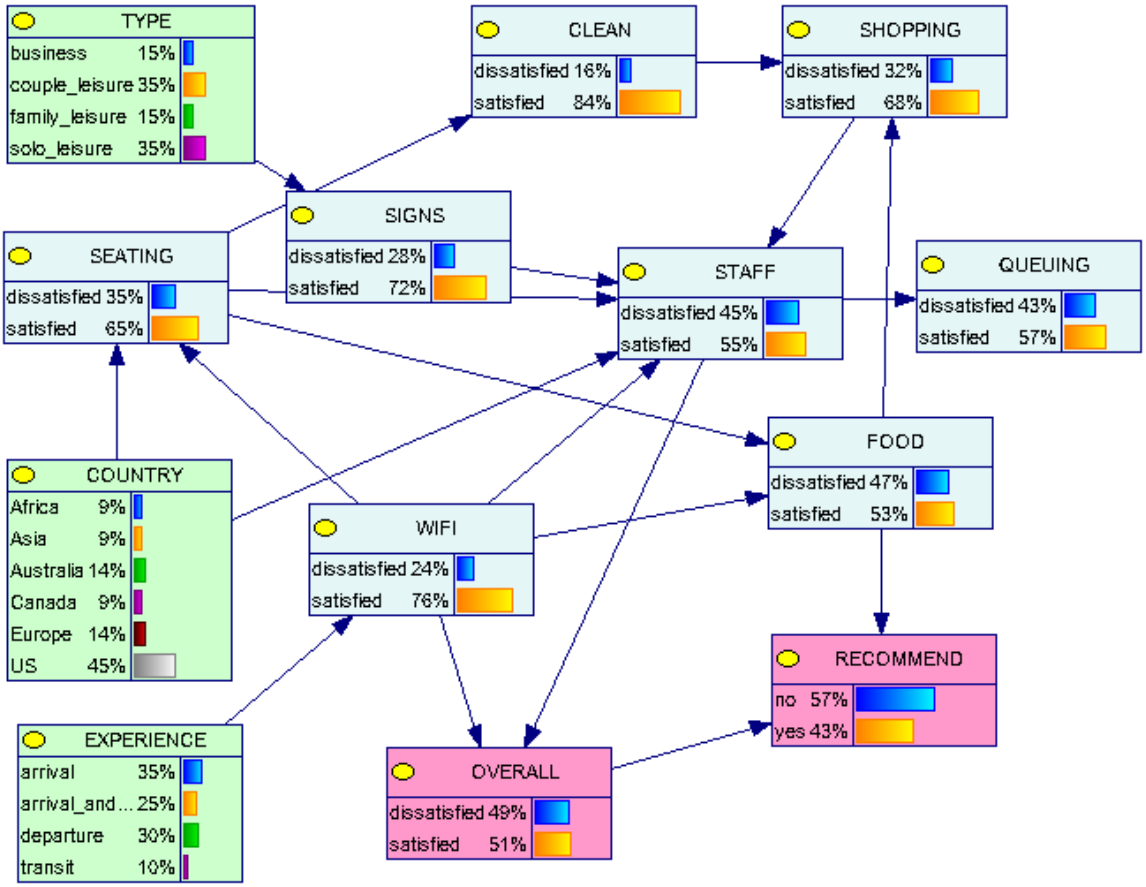

Figure 4: BN of the BOT1+2 Skytrax reviews social media dataset

The BN obtained from the BOT1+2 Skytrax social media review dataset is displayed in Figure 4, where we adopted the same color code used for the SFO customer satisfaction survey analysis. Since the role of the variable RECOMMEND is similar to OVERALL, the relevant node was depicted in pink. In the Skytrax dataset, the distribution of the OVERALL variable is well-balanced, as opposed to the same distribution of the SFO customer satisfaction survey dataset. In particular, the proportion of overall dissatisfied customers of the online review dataset is much higher than the same proportion in the survey dataset. In addition, the majority of passengers will not recommend SFO airport to a friend. A high number of negative feedback is frequent in online blogs and social media pages, since reviewers feel more free to express their opinions online rather than via traditional surveys. The availability of information on dissatisfied customers is key to organizations in order to identify their weaknesses and to improve their services. Therefore, the integration of traditional surveys with online reviews is fundamental to correctly analyze customer satisfaction.

Among the rating variables, the main determinants of passengers' overall dissatisfaction are STAFF and WIFI. The percentage of dissatisfaction with staff is 45% and the percentage of

dissatisfaction with accessing and using Wi-Fi is 24%. Therefore, the primary areas of weakness in the airport are related to staff and Wi-Fi and interventions in these areas will sensibly reduce customers' overall disappointment with the airport services.

The demographic variables are affecting passengers' overall dissatisfaction via the rating given to staff friendliness, accessing and using the free Wi-Fi service and signs, directions and monitors. Therefore, particular attention needs to be given to specific groups of passengers, who might be more sensitive than others to unsatisfactory airport services.

Identification of the Calibration Link

The calibration link for the BOT1+2 dichotomized datasets is the OVERALL variable. The percentage of dissatisfied passengers in the SFO survey dataset is only 2%, while the same percentage in the Skytrax online dataset is almost 50%. Therefore, the levels of OVERALL in the SFO survey dataset need to be re-balanced by resampling, to make the distribution similar to that of the Skytrax online dataset.

Performing Calibration

The SFO customer survey dataset was resampled, as explained in Section 4, using the R package ROSE (Lunardon et al., 2014). The BN was updated via parameter learning and hence calibrated to reflect the information contained in the online reviews. Figure 5 illustrates the BN of the BOT1+2 SFO customer satisfaction survey dataset, after calibration of the OVERALL node via resampling. The distribution of the overall satisfaction is now balanced, with a higher proportion of dissatisfied customers, as appears in online reviews. This calibrated BN shows that the percentages of passengers who are dissatisfied with cleanliness, walkways, shopping areas and the free Wi-Fi are 19%, 23%, 33% and 14%, respectively. These results highlight, much more clearly than those based on the original unbalanced dataset, the weaknesses and corresponding areas of improvement of the airport.

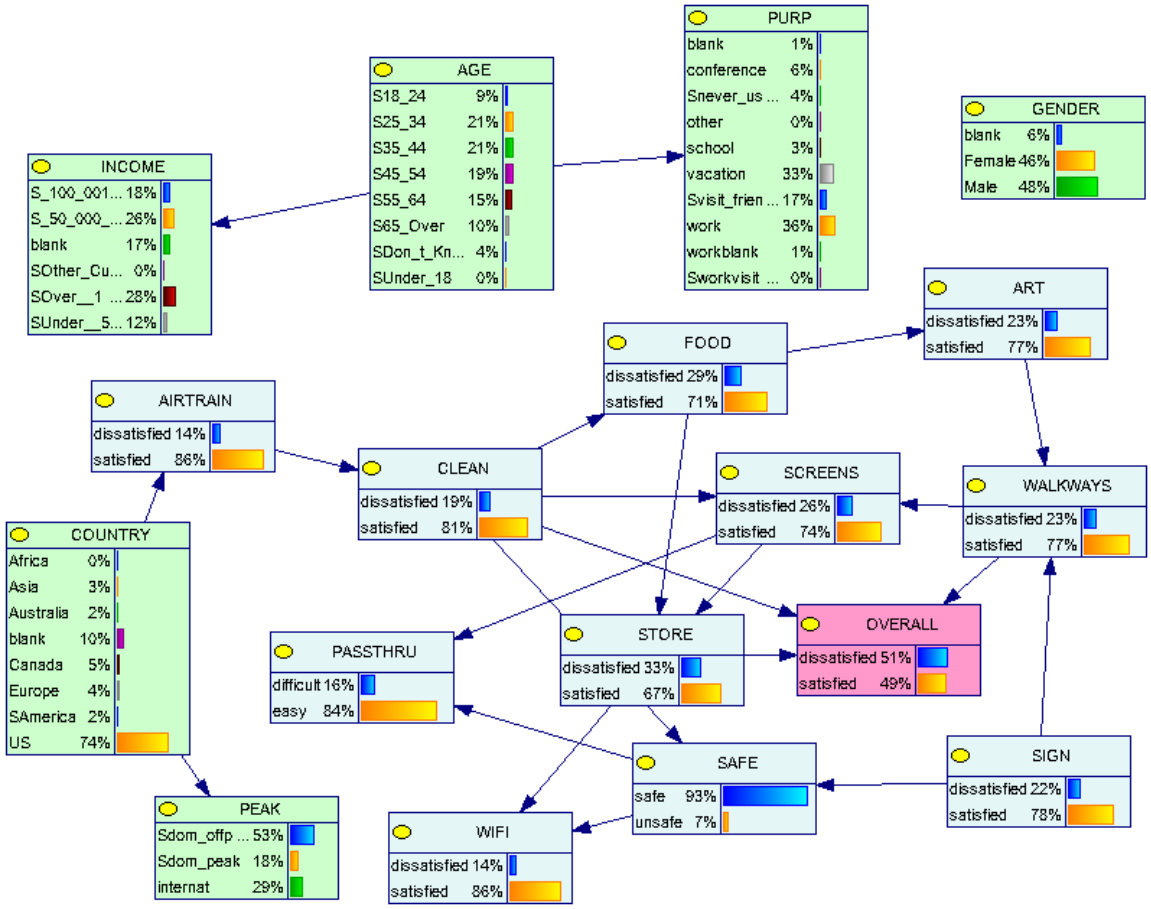

Figure 5: BN of the BOT1+2 SFO customer satisfaction survey dataset, after calibration of the OVERALL node via resampling.

In order to evaluate the performance of the proposed data integration methodology, we compared the results of the uncalibrated and calibrated BNs, estimating the following measures:

- Absolute Bias = $| p_{min} - p_{min}^* |$ (1)
- Relative Bias = $| (p_{min} - p_{min}^*) / p_{min} | \times 100$ (2)
- Percentage Bias = $| (p_{min} - p_{min}^*) / \text{Mean}[p_{min}, p_{min}^*] | \times 100$ (3)

where p_{min} is the minority class proportion (i.e. the proportion of dissatisfied customers) of the uncalibrated SFO BOT1+2 variables and p_{min}^* is the corresponding class proportion of the calibrated SFO BOT1+2 variables.
Table 3 compares the uncalibrated and calibrated SFO BOT1+2 datasets, listing the proportions of dissatisfied passengers for the calibration link variable and the most influential determinants of the overall dissatisfaction. The estimated bias measures clearly show that the uncalibrated results are largely underestimating the extent of customer dissatisfaction. The bias measures reach their maximum with the calibration link variable OVERALL, while among the other variables, STORE has the highest absolute bias and CLEAN has the highest relative and percentage bias. These results show that there are pockets of dissatisfaction with airport cleanliness and shops, that would be hidden and ignored with a simple analysis of the uncalibrated results. The exclusive study of interview-based data, with their extremely low proportions of dissatisfaction, may lead to the erroneous conclusion that there are no areas on improvement in the airport, which could be

dangerous for the future of the organization. On the contrary, the proposed approach, integrating survey with social media reviews, provides a more accurate picture of customer satisfaction, pointing out the existence of pockets of dissatisfaction, which are crucial for improving the service supplied by the organization.

Table 3: Comparison between the results of the uncalibrated and calibrated SFO BOT1+2 data

		Uncalibrated Dissatisfied Proportion	Calibrated Dissatisfied Proportion	Absolute Bias	Relative Bias	Percentage Bias
SFO BOT1+2 Variables	OVERALL	0.02	0.51	0.49	2450.00	184.91
	CLEAN	0.02	0.19	0.17	850.00	161.90
	STORE	0.08	0.33	0.25	312.50	121.95
	WALKWAYS	0.03	0.23	0.20	666.67	153.85
	WIFI	0.06	0.14	0.08	133.33	80.00

6.2 Data Integration of TOP5 Datasets

Data Structure Modelling

The BNs of the SFO and Skytrax TOP5 datasets were constructed using the GeNIe software following the same procedure adopted for the BOT1+2 datasets described in Section 6.1.

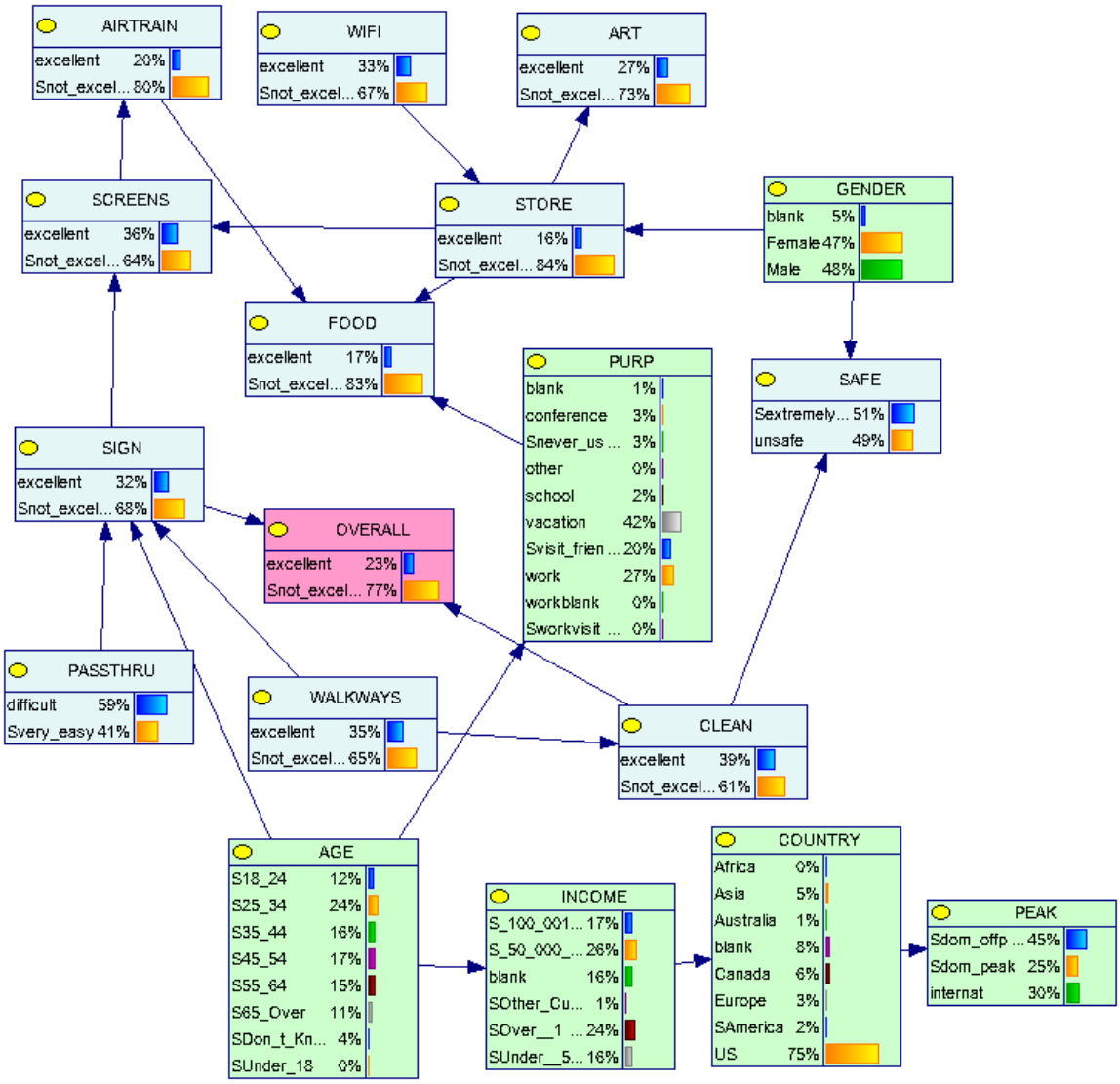

Figure 6: BN of the TOP5 SFO customer satisfaction survey dataset

The BN of the SFO customer satisfaction survey dataset dichotomized according to the TOP5 scheme is shown in Figure 6, where nodes follow the same color code adopted in the previous Figures. Again, only some of the demographic variables (namely AGE, GENDER and PURP) affect customers' satisfaction. The most influential variables to customers' overall satisfaction are CLEAN, SIGN and WALKWAYS, suggesting that a high satisfaction with airport cleanliness, signs, directions and walkways will enhance customers' overall satisfaction. The implication is that if the airport increases the percentage of customers with top-level satisfaction from cleanliness, signs and walkways, overall satisfaction levels will reach their maximum.

Note that the SFO TOP5 dataset is not affected by strong class imbalance. For example, the 41% of the interviewees states that passing through security and screening is very easy (variable PASSTHRU). Also, the highest level of satisfaction from cleanliness (percentage of '5') is 39%, from signs is 32% and from walkways is 35%.

Figure 7 shows the BN built using the Skytrax dataset, dichotomized according to the TOP5 scheme. The node colors follow the coding adopted in the previous Figures. Differently from the BOT1+2 Skytrax BN, the demographic variables are not related to passengers' satisfaction. Since the variable OVERALL depends on FOOD, QUEUING, SHOPPING and SEATING, the airport's

areas of excellence are identified by the quality of restaurants, the queueing system, the presence of a variety of shops and the availability of seating spaces. Note that the levels of several variables dichotomized according to the TOP5 scheme, such as QUEUING, are imbalanced, with a percentage of '5' equal to 24%. Also, the highest level of satisfaction from restaurants is 31%, from shops is 24% and from seating is 11%. This implies that customers' overall satisfaction will be maximized by an increased top-level satisfaction with food, queuing, shops and seating spaces. However, class imbalance makes it difficult to clearly identify areas of excellence within the organization.

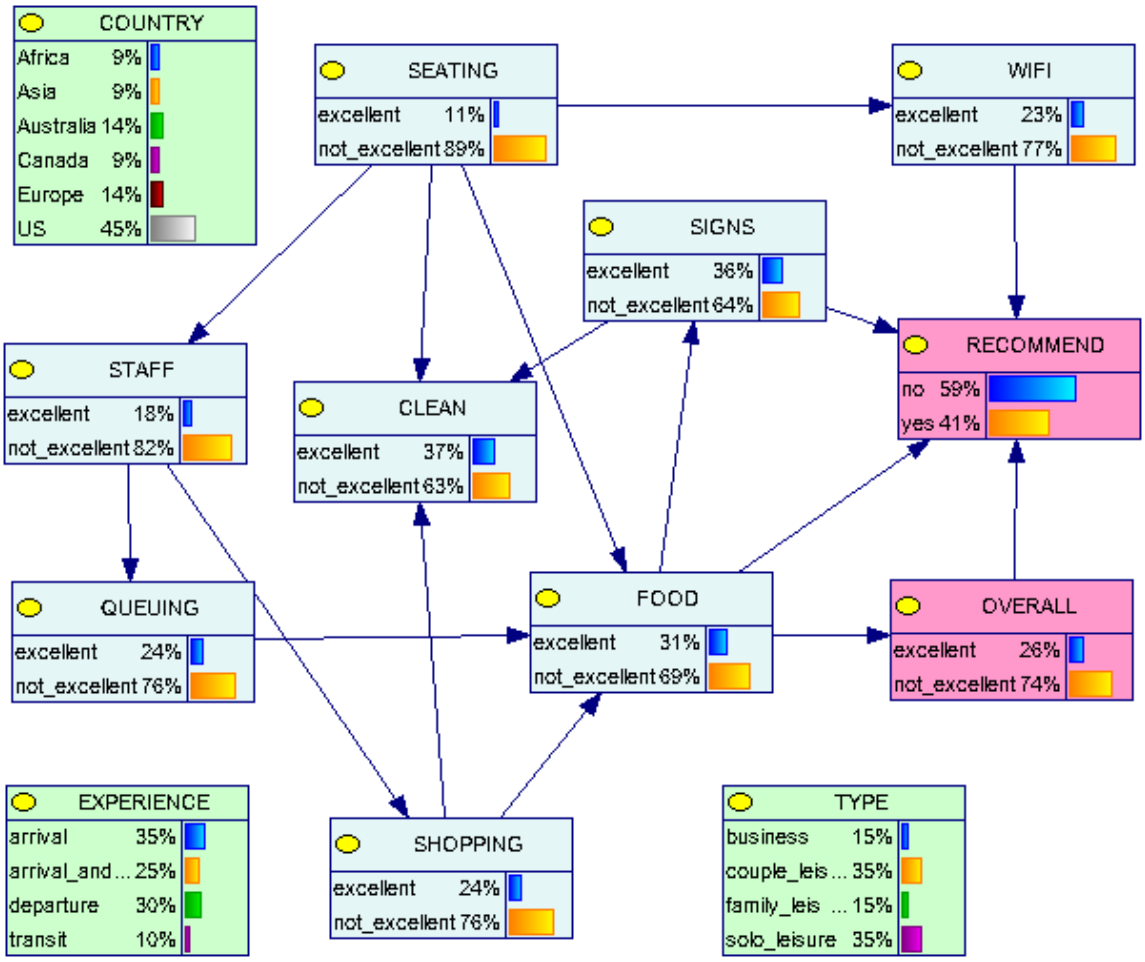

Figure 7: BN of the TOP5 Skytrax reviews social media dataset

Note that the determinants of passengers' overall satisfaction in the BOT1+2 and TOP5 schemes are different. In particular, while staff friendliness and the availability of Wi-Fi can be identified as areas of improvement, the quality of restaurants, the queueing system, the presence of a variety of shops and the availability of seating spaces can be identified as areas of excellence in the SFO airport.

Identification of the Calibration Link

The calibration link for the TOP5 dichotomized datasets is the QUEUING variable. This is a key determinant of the overall satisfaction variable in the Skytrax dataset. However, there is an imbalance in its classes, since the percentage of 'excellent' answers is only 24%. The same variable appears to be well-balanced in the SFO survey dataset, where the percentage of 'excellent' is close to 50%. Therefore, the Skytrax dataset needs to be resampled, in order to re-balance the distribution of QUEUING according to the distribution of the SFO survey dataset.

Performing Calibration

In order to re-balance the QUEUING variable, the Skytrax online reviews dataset was resampled, to reflect the distribution of a similar variable (PASSTHRU) in the SFO customer survey dataset. Calibration between the two datasets was performed and the BN of the TOP5 Skytrax dataset was updated via parameter learning. Figure 8 illustrates the BN of the TOP5 Skytrax reviews social media dataset, after calibration of the QUEUING node via resampling. The distribution of passengers' satisfaction with queuing is now balanced, with a higher proportion of extremely satisfied passengers, as appears in the SFO customer survey dataset. This calibrated BN shows that the percentages of passengers who are extremely satisfied with cleanliness, restaurants, shopping and seating areas have increased and are equal to 44%, 37%, 34% and 14%, respectively. In addition, the percentage of very satisfied passengers overall is 34%. These results calibrate the overly negative online reviews and underline the areas of excellence of the airport.

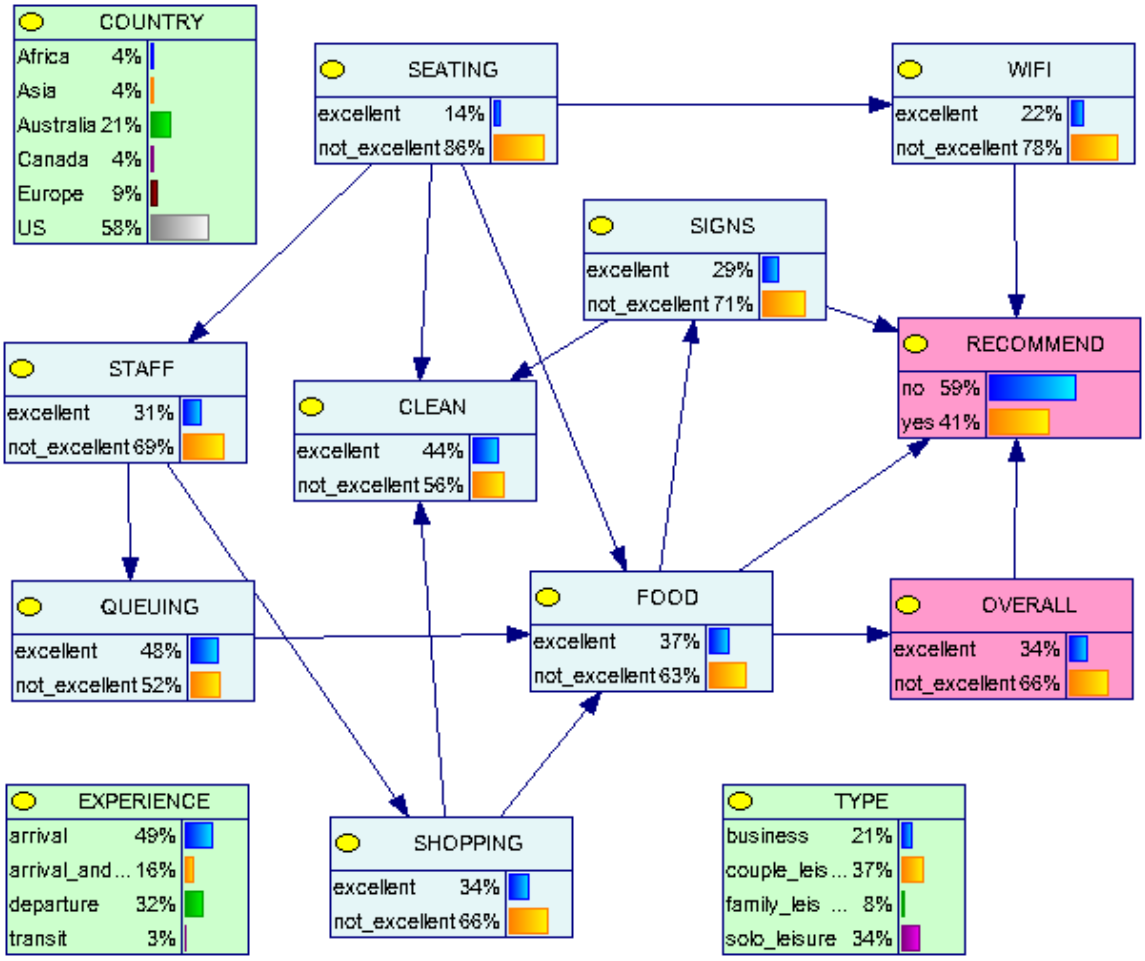

Figure 8: BN of the TOP5 Skytrax reviews social media dataset, after calibration of the QUEUING node via resampling.

Table 4 compares the uncalibrated and calibrated Skytrax TOP5 datasets, listing the proportions of extremely satisfied passengers for the calibration link, the overall satisfaction variable and the most influential determinants of customer satisfaction. We compared uncalibrated and calibrated results using the bias measures (1), (2) and (3) introduced in Section 6.1, where p_{min} here is the proportion of extremely satisfied customers. In Table 4, since class unbalance in the uncalibrated datasets is less severe in the TOP5 than the BOT1+2 datasets, the estimated bias results are generally lower than those listed in Table 3. However, these results show that the underestimation of the proportion of satisfied customers in the uncalibrated analysis is still relevant. The bias measures are particularly high for the calibration link QUEUING and the SHOPPING variables. These results demonstrate that the proposed approach, based on the integration of survey and social media reviews, allows us to highlight more clearly the areas of excellence within an organization. Therefore, the data integration methodology leads to a more accurate customer satisfaction analysis and provides a valuable tool for decision-makers.

Table 4: Comparison between the results of the uncalibrated and calibrated Skytrax TOP5 data

		Uncalibrated Satisfied Proportion	Calibrated Satisfied Proportion	Absolute Bias	Relative Bias	Percentage Bias
Skytrax TOP5 Variables	QUEUING	0.24	0.48	0.24	100.00	66.67
	CLEAN	0.37	0.44	0.07	18.92	17.28
	FOOD	0.31	0.37	0.06	19.35	17.65
	OVERALL	0.26	0.34	0.08	30.77	26.67
	SEATING	0.11	0.14	0.03	27.27	24.00
	SHOPPING	0.24	0.34	0.10	41.67	34.48

7. Discussion and Conclusions

With the growing exploitation of big data, integration of data sources becomes a key capability. Traditional integration methods rely on extract transform and load (ETL) and record linkage techniques (Kenett and Raanan, 2010). In this paper, we propose a novel approach to data integration that combines online big data with a comprehensive survey. The methodology is derived from resampling and modeling the data using BNs, and identifying overlapping links that are used for calibration. We show, with an example, how data integration between online blogs and a customer satisfaction survey supports proper chronology of data and goal. The example demonstrates of such data integration enhances the information quality of a study in four of the InfoQ dimensions: Data Structure, Data Integration, Temporal Relevance and Chronology of Data and Goal.
The approach is applicable in a wide range of domains such as the integration of administrative data with official statistics or combining data from different sensors in a production environment. In particular, with continuous variables, the proposed methodology can be modified by combining nonparametric BNs and Vines (Dalla Valle and Kenett, 2015; Dalla Valle, 2016, 2017a, 2017b and 2017c). Vines are extremely flexible in high-dimensional cases, allowing the specification of various types of non-linear dependencies. Results from the application of vines can be used to determine the causal effects in non-parametric BNs.

This research addresses a growing need in big data analytics and requires follow up, for example considering methods for integration of a very high number of data sources to increase accuracy of results. It is one of relatively few studies which attempt to address the generalizable problem of big data integration. It proposes and demonstrates a methodology designed to increase information quality.

References

1. Aalen O., Røysland K. and Gran JM. (2012). Causality, mediation and time: a dynamic viewpoint. *Journal of the Royal Statistical Society (Series A)*, 175(4), 831-861.
2. Asur, S. and Huberman, B.A. (2010). Predicting the future with social media, Web Intelligence and Intelligent Agent Technology (WI-IAT), *2010 IEEE/WIC/ACM International Conference*, Vol. 1, pp. 492-499.
3. Bai, X., Kenett, R.S. and Yu, W. (2012). Risk Assessment and Adaptive Group Testing of Semantic Web Services. *International Journal of Software Engineering and Knowledge Engineering*, 22 (5), 595-620.

4. Baker S. (2013). Causal inference, probability theory, and graphical insights. *Statistics in Medicine*; 2(25), 4319-4330.
5. BBC News (2017). Facebook to roll out fake news tools in Germany. Available online at: http://www.bbc.co.uk/news/business-38631847
6. Ben Gal, I. (2007). Bayesian Networks, in *Encyclopedia of Statistics in Quality and Reliability*, Ruggeri, F., Kenett, R. S. and Faltin, F. (editors in chief), Wiley, UK.
7. Chawla, N.V. (2005) Data mining for imbalanced datasets: An overview. In: *Data mining and knowledge discovery handbook*. Springer US, p. 853-867.
8. Chakraborty, S., Mengersen, K., Fidge, C., Ma, L., and Lassen, D. (2015). Multifaceted modelling of complex business enterprises. *PloS one*, Vol. 10, No.8, e0134052.
9. Chen, C.L.P. and Zhang, C.Y. (2014). Data-intensive applications, challenges, techniques and technologies: a survey on big data, *Information Sciences*, Vol. 275, pp. 314-347.
10. Dalla Valle, L. (2017a) Data Integration. in: Wiley StatsRef: Statistics Reference Online, ed. M. Davidian, B. Everitt, R. Kenett, G. Molenberghs, W. Piegorsch and F. Ruggeri, John Wiley & Sons, pp. 1-6.
11. Dalla Valle, L. (2017b) Copula and Vine Modeling for Finance, in: Wiley StatsRef: Statistics Reference Online, ed. M. Davidian, B. Everitt, R. Kenett, G. Molenberghs, W. Piegorsch and F. Ruggeri, John Wiley & Sons, pp. 1-5.
12. Dalla Valle, L. (2017c) Copulas and Vines, in: Wiley StatsRef: Statistics Reference Online, ed. M. Davidian, B. Everitt, R. Kenett, G. Molenberghs, W. Piegorsch and F. Ruggeri, John Wiley & Sons, pp. 1-5.
13. Dalla Valle, L. (2016) The Use of Official Statistics in Self-Selection Bias Modeling. *Journal of Official Statistics*, Vol. 32, No. 4, pp. 887–905.
14. Dalla Valle, L. (2014). Official statistics data integration using copulas. *Quality Technology and Quantitative Management*; 11(1), 111-131.
15. Dalla Valle, L. and Kenett, R.S. (2015). Official Statistics Data Integration for Enhanced Information Quality, *Quality and Reliability Engineering International*, Vol. 31, No. 7, pp. 1281-1300.
16. Daniel, B. (2015). Big Data and analytics in higher education: opportunities and challenges. *British Journal of Educational Technology*, 46(5), pp. 904-920.
17. Di Zio, M., Sacco, G., Scanu, M., Vicard, P. (2005). Multivariate techniques for imputation based on Bayesian networks. *Neural Network World*, 4, 303–309.
18. Dong, X.L. and Srivastava, D. (2015). *Big Data Integration*. Synthesis Lectures on Data Management. Morgan Claypool Publishers.
19. Douglas, L. (2001). 3D data management: controlling data volume, velocity and variety. *Gartner Report*. Available online at: https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf
20. Fenton, N.E. and Neil, M., (2011). The use of Bayes and causal modelling in decision making, uncertainty and risk, *UPGRADE*, the Journal of CEPIS (Council of European Professional Informatics Societies), 12(5), pp. 10-21.
21. Fenton, N. E. and Neil, M. (2012). Risk Assessment and Decision Analysis with Bayesian Networks, CRC Press, http://www.bayesianrisk.com.
22. Fenton, N. E. and Neil, M. (2014). Decision Support Software for Probabilistic Risk Assessment Using Bayesian Networks". IEEE Software, 31(2), 21–26.
23. Foresti, G., Guelpa, F. and Trenti, S. (2012). Enterprises in a globalized context and public and private statistical setups. *SIS Scientific Meeting*.
24. Frosini B. (2006). Causality and causal models: a conceptual perspective*. International Statistical Review*; 74, 305-334.
25. GeNIe. *Decision Systems Laboratory*. University of Pittsburgh: USA, 2006. http://genie.sis.pitt.edu.
26. Ghose, A. and Ipeirotis, P.G. (2011). Estimating the helpfulness and economic impact of product reviews: mining text and reviewer characteristics, *IEEE Transactions on Knowledge and Data Engineering*, 23(10), pp. 1498-1512.
27. Heckerman, D. (1995). A tutorial on learning with Bayesian networks. Microsoft Research tech. report MSR-TR-95-06. Revised November 1996, from http://research.microsoft.com.

28. Heckerman, D. (1997). Bayesian Networks for Data Mining. Data Mining and Knowledge Discovery, 1(1), 79-119.
29. Heckman, J. (2008). Econometric Causality. *International Statistical Review*, 76, 1-27.
30. Imai K., Tingley D. and Yamamoto T. (2013). Experimental designs for identifying causal mechanisms. *Journal of the Royal Statistical Society (Series A)*; 176(1), 5-51.
31. Jensen, F. V. (2001). *Bayesian Networks and Decision Graphs*, Springer.
32. Kaplan, A.M. and Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of social media, *Business Horizons*, Vol. 53, pp. 59-68.
33. Kenett, R. S. (2016). On Generating High InfoQ with Bayesian Networks, *Quality Technology and Quantitative Management,* 13(3), http://dx.doi.org/10.1080/16843703.2016.11891
34. Kenett, R., De Frenne, A., Tort-Martorell, X and McCollin, C. (2008). The Statistical Efficiency Conjecture, in *Applying Statistical Methods in Business and Industry – the state of the art*, Greenfield, T., Coleman, S. and Montgomery, R. (editors), John Wiley and Sons, Chichester: UK.
35. Kenett, R.S. (2012). Risk Analysis in Drug Manufacturing and Healthcare, in *Statistical Methods in Healthcare*, Faltin, F., Kenett, R.S. and Ruggeri, F. (editors in chief), John Wiley and Sons.
36. Kenett, R.S. and Raanan, Y. (2010). *Operational Risk Management: a practical approach to intelligent data analysis*, John Wiley and Sons, Chichester, UK.
37. Kenett, R.S. and Salini, S. (2011). *Modern Analysis of Customer Satisfaction Surveys: with applications using R*, John Wiley and Sons, Chichester: UK.
38. Kenett, R.S. and Shmueli, G. (2014). On Information Quality, *Journal of the Royal Statistical Society (Series A),* 177(1), 3-38.
39. Kenett, R.S. and Shmueli, G. (2016). Information Quality: The Potential of Data and Analytics to Generate Knowledge, John Wiley and Sons. www.wiley.com/go/information_quality.
40. Kenett, R.S. and Zacks, S. (2014), *Modern Industrial Statistics: with applications using R, MINITAB and JMP, 2nd edition,* John Wiley and Sons, Chichester, UK.
41. Kenett, R.S. (2017) "Bayesian networks: Theory, applications and sensitivity issues", Encyclopedia with Semantic Computing, World Scientific press. DOI: 10.1142/S0000000016300146
42. Koski, T. and Noble, J. (2009). *Bayesian Networks – An Introduction*, John Wiley and Sons, Chichester, UK.
43. Krishnamoorthy, S. (2015) Linguistic features for review helpfulness prediction, Expert Systems with Applications, 42(7), pp. 3751-3759.
44. Li, J., Conradi, R., Slyngstad, O., Torchiano, M., Morisio, M., and Bunse, C. (2008). A State-of-the-Practice Survey of Risk Management in Development with Off-the-Shelf Software Components. IEEE Trans. Software Eng. 34(2): 271-286.
45. Lunardon, N., Menardi, G., and Torelli, N. (2014). ROSE: A Package for Binary Imbalanced Learning. *R Journal*, Vol. 6, No. 1, pp. 82-92.
46. Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. and Hung Byers, A. (2011). *Big data: the next frontier for innovation, competition, and productivity*. Available online at: http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation
47. Marella, D. and Vicard, P. (2013). Object-Oriented Bayesian Networks for Modelling the Respondent Measurement Error, *Communications in Statistics - Theory and Methods*, 42(19), 3463-3477
48. Mealli, F., Pacini, B. and Rubin, D.B. (2012). Statistical inference for causal effects In *Modern Analysis of Customer Satisfaction Surveys: with applications using R* (Kenett, R.S. and Salini, S., editors) John Wiley and Sons, Chichester: UK.
49. Menardi, G., and Torelli, N. (2014). Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery*, Vol. 28, No, 1, pp 92–122.
50. Montoyo, A., Martinez-Barco, P. and Balahur, A. (2012). Subjectivity and sentiment analysis: an overview of the current state of the area and envisaged developments, Decision Support Systems, 53(4), pp. 675-679.
51. Musella F. (2013). A PC algorithm variation for ordinal variables, *Computational Statistics*, 28(6), 2749-2759.
52. Neapolitan, E.R. (2003), *Learning Bayesian Networks*. Prentice Hall.

53. Organization for Economic Co-operation and Development (OECD) (2007). *Working Party on the Information Economy*, Participative Web: User-Created Content.
54. Pearl J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.
55. Pearl J. (2013). Comment on Causal inference, probability theory, and graphical insights (by Stuart G. Baker). UCLA Cognitive Systems Laboratory, *Statistics in Medicine,* 32(25), 4331-4333
56. Pearl, J. (1985). Bayesian Networks: A Model of Self-Activated Memory for Evidential Reasoning" (UCLA Technical Report CSD-850017). *Proceedings of the 7th Conference of the Cognitive Science Society*, University of California, Irvine, CA, 329–334.
57. Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*, 2nd ed., Cambridge University Press, UK.
58. Peterson, J. and Kenett, R.S. (2011). Modelling Opportunities for Statisticians Supporting Quality by Design Efforts for Pharmaceutical Development and Manufacturing, *Biopharmaceutical Report*, ASA Publications, 18(2), 6-16.
59. Pfeffermann, D. (2013). New Important Developments in Small Area Estimation, *Statistical Science*, 28, pp. 40-68.
60. Pietro, L.D., Mugion, R.G., Musella, F., Renzi, M.F., Vicard, P. (2015). Reconciling internal and external performance in a holistic approach: a Bayesian network model in higher education, *Expert Systems with Applications*, 42(5), 2691–2702
61. Pourret, O, Naïm P. and Marcot, B. (2008). *Bayesian Networks: A Practical Guide to Applications*, John Wiley and Sons, Chichester: UK.
62. Salter-Townshend, M., White, A., Gollini, I., Murphy, T. B.: Review of statistical network analysis: models, algorithms, and software. Statistical Analysis and Data Mining, 5(4):243–264 (2012)
63. Schneider, C. (2016) The biggest data challenges that you might not even know you have. *IBM Watson*. Available online at: https://www.ibm.com/blogs/watson/2016/05/biggest-data-challenges-might-not-even-know/
64. Stander, J., Dalla Valle, L. and Cortina Borja, M. (2016a). Sentiments, surnames and so long EU. *Communicator, Special Supplement: Science Communication*, Autumn 2016, pp. 19–23.
65. Stander, J., Dalla Valle, L., Eales, J., Baldino, A. and Cortina Borja, M. (2016b). The EU referendum: extracting insights from Facebook using R, *Significance Magazine*, available on-line at www.statslife.org.uk/significance/2889.
66. Tarantola, C., Vicard, P., and Ntzoufras, I. (2012). Monitoring and improving Greek banking services using Bayesian networks: An analysis of mystery shopping data. *Expert Systems with Applications*, 39(11), 10103-10111.
67. The Economist (2016). *Where polling failed, Facebook prevailed*, available online at: http://www.economist.com/blogs/graphicdetail/2016/11/social-media-and-american-election
68. Vicard, P., Dawid, A. P., Mortera, J., Lauritzen, S. L. (2008). Estimation of mutation rates from paternity casework. *Forensic Science International Genetics*, 2, 9–18.
69. Zafarani, R., Abbasi, M.A. and Liu, H. (2014) *Social media mining: an introduction*, Cambridge University Press.
70. Zhang, X., Fuehres, H. and Gloor, P.A. (2011). Predicting stock market indicators through twitter "I hope it is not as bad as I fear". *Procedia – Social and Behavioral Sciences*, Vol. 26, pp. 55-62.