Insurance: Mathematics and Economics 75 (2017) 126–136

Contents lists available at ScienceDirect

Insurance: Mathematics and Economics

journal homepage: www.elsevier.com/locate/ime

Data breaches: Goodness of fit, pricing, and risk measurement

Martin Eling [a,*], Nicola Loperfido [b]

[a] Institute of Insurance Economics, University of St. Gallen, Kirchlistrasse 2, 9010 St. Gallen, Switzerland
[b] Dipartimento di Economia, Società e Politica – DESP, University of Urbino, Via Saffi 42, Urbino (PU) 61029, Italy

ARTICLE INFO

Article history:
Received December 2016
Received in revised form May 2017
Accepted 20 May 2017
Available online 26 May 2017

JEL classification:
G22
G31

Keywords:
Cyber risk
Risk measurement
Multidimensional scaling
Goodness of fit
Skew-normal distribution

ABSTRACT

Some research on cyber risk has been conducted in the field of information technology, but virtually no research exists in the actuarial domain. As a first step toward a more profound actuarial discussion, we use multidimensional scaling and goodness-of-fit tests to analyze the distribution of data breach information. Our results show that different types of data breaches need to be modeled as distinct risk categories. For severity modeling, the log-skew-normal distribution provides promising results. The findings add to the recent discussion on the use of skewed distributions in actuarial modeling (Vernic, 2006; Bolancé et al., 2008; Eling, 2012). Moreover, they provide useful insights for actuaries working on the implementation of cyber insurance policies. We illustrate the usefulness of our results in two applications on risk measurement and pricing.

1. Introduction

Cyber risks are operational risks to information and technology assets that have consequences for the confidentiality, availability, and integrity of information and information systems (see Cebula and Young, 2010). Although cyber risks, such as hacking attacks or unintended disclosures, are reported in the media every day and rank high in the business agenda of every Chief Financial Officer and Chief Risk Officer, to our knowledge no research on the topic has been done in the actuarial domain. This is surprising, given the high economic importance (global losses for cyber risk are estimated to surpass US$400 billion per year; see McAfee, 2014) and the increasing efforts of many insurance companies to further develop a market for cyber risk insurance (see Biener et al., 2015).

One reason for the lack of research in the actuarial domain is the lack of data. Recently, however, this situation has changed, especially with the establishment of first data breach databases. In the US, reporting requirements for data breaches have been introduced in many states since 2002 (National Conference of State Legislatures, 2016), and data breach databases are becoming increasingly available. This paper analyzes such data using both exploratory (multidimensional scaling, multiple factor analysis for contingency tables) and confirmatory approaches (goodness of fit).

The literature on cyber risk and information security is mainly limited to the field of information technology, but very little work has been done in business, finance, and economics. Our paper is closest to the data breach analyses of Maillart and Sornette (2010), Edwards et al. (2015), and Wheatley et al. (2016).[1] The intention of this paper is to link what has been done in those three papers with the current discussion on goodness of fit, pricing, and risk measurement in the actuarial domain (see Vernic, 2006; Bolancé et al., 2008; Eling, 2012; Miljkovic and Grün, 2016, among others).

Multidimensional scaling shows that different types of data breaches need to be modeled as distinct risk categories, given their different statistical nature—a result that has not been the focus of existing data breach analyses. For the severity model, it turns out that either the log-normal or the log-skew-normal distribution provides promising results. This is a relevant result, considering the

[1] Maillart and Sornette's (2010) study of the statistical properties of data breaches between 2000 and 2008 reveals the existence of two distinct phases for the breach frequency (explosive growth up to about July 2006 and a stable rate thereafter). Breach size follows a heavy-tailed power-law distribution, remains stable over time and does not depend on the organization's type or size. Edwards et al. (2015) analyze time trends for the size and frequency of malicious and negligent data breaches and show that neither size nor frequency of breaches has increased in recent years; breach size is distributed log-normally and the frequency follows a negative binomial distribution. Wheatley et al. (2016) extend Maillart and Sornette's (2010) work by enlarging the dataset and focusing on the tails of the distribution (i.e., incidents with more than 50,000 records breached). They show that the frequency of large events is independent of time for the US and is increasing over time for non-US firms.

* Corresponding author.
E-mail addresses: martin.eling@unisg.ch (M. Eling), nicola.loperfido@uniurb.it (N. Loperfido).

http://dx.doi.org/10.1016/j.insmatheco.2017.05.008
0167-6687/

recent discussion on the use of skewed distributions in actuarial science (e.g., Vernic, 2006; Eling, 2012). Our results offer important information for insurance companies and regulators seeking to better understand the potential risk exposure when selling cyber insurance policies. Moreover, we also hope to encourage more research on the topic in the risk and insurance domain.

The remainder of this paper is organized as follows. In Section 2, we briefly introduce the methods employed in the paper. Section 3 presents the data and descriptive statistics. The results are given in Section 4, including two applications on risk measurement and pricing. Section 5 concludes the paper.

2. Methodology

We analyze data breach information using both exploratory and confirmatory approaches (Tukey, 1977). Exploratory data analyses aim to uncover previously unanticipated data features; we implement such analyses via multidimensional scaling (MDS) and multiple factor analysis for contingency tables (MFACT). Confirmatory data analyses aim to test statistical hypotheses; we implement such analyses by testing the goodness of fit for several well-known distributions, especially in the tails. For the sake of brevity, we only briefly describe the main methodological approaches used in the paper and refer to the literature for all details.

MDS is a multivariate statistical technique that is rarely used in the context of insurance (one exception is Brechmann et al., 2013), but is widely used in other fields. It aims to recover the data structure from the distances between data points. MDS approximates interpoint distances with Euclidean distances between numbers, pairs of numbers, or trios of numbers. When distances are Euclidean, MDS and principal component analysis lead to the same results; thus, the former generalizes the latter. A detailed description of MDS, together with its potential applications, is given by Mardia et al. (1979). In this paper, we use MDS to investigate differences between entities that suffer data breaches and differences between various types of attacks. In both cases, we first use the number of data breaches (to measure frequency) and then use the number of lost records (to measure severity). Differences between either entities or types will be evaluated via the chi-squared distance, which is the default choice for a distance when the data matrix is a contingency table.[2,3]

A dynamic approach taking the evolution of data breaches through time into account requires a joint analysis of several contingency tables. Bécue-Bertaut and Pagés (2004, 2008) introduced MFACT, a multivariate statistical method specifically developed for such situations. It has been applied to textual analysis (Bécue-Bertaut, 2014) and implemented in the R package FactoMineR (Husson et al., 2007), which has been thoroughly illustrated by Kostov et al. (2013).

In the confirmatory data analysis, we test several established actuarial models with respect to their goodness of fit. The data breach frequency is modeled by either a Poisson or a negative binomial distribution (see, e.g., Moscadelli, 2004). For the data breach severity, we fit the data to several distributions used in recent actuarial literature (see, e.g., Eling, 2012). Furthermore, we include a non-parametric transformation kernel estimation (see Bolancé et al., 2003, 2008)[4] and implement the peaks-over-threshold (POT) method from extreme value theory (EVT; see, e.g., Chapelle et al., 2008). In the latter approach, losses above a threshold (e.g. the 90% quantile) are modeled by a generalized Pareto distribution (GPD), while losses below the threshold are modeled with another common loss distribution, such as exponential, log-normal, or Weibull. To identify the best models, we apply various goodness-of-fit tests (log-likelihood value, the AIC, Kolmogorov–Smirnov (KS)-test, Anderson–Darling (AD)-test).[5]

All models are implemented in the R packages sn, ghyp, and MASS. We use all packages to derive the best-fitting parameters and compare these distributions. Some of the benchmark distributions are also involved in the risk measurement and pricing procedure, where we compare model results with the empirical results to evaluate the accuracy of different models. More details on skewed distributions can be found in Adcock et al. (2015) and Azzalini (2013); a description of the other benchmark models is given in actuarial textbooks, such as Mack (2002), Kaas et al. (2009), and Panjer (2007). It should also be mentioned that a better fit does not necessarily mean that a model is better, as actuaries need to keep in mind many other aspects, such as the risk of change of the underlying stochastic process.

3. Data and descriptive statistics

The data breach information we consider is taken from the "Chronology of Data Breaches" provided by the Privacy Rights Clearinghouse (PRC). This dataset has not yet been used in the context of actuarial science, but it has been applied in other fields (see Maillart and Sornette, 2010; Edwards et al., 2015; Wheatley et al., 2016). The PRC is a non-profit organization with the mission to engage, educate, and empower individuals to protect their privacy (Privacy Rights Clearinghouse, 2016); their data breach dataset is regularly updated and can be downloaded from the PRC website. The data sample we use here consists of data breaches in the US between January 10, 2005, and December 15, 2015. We follow Edwards et al. (2015) in erasing all observations that do not give information on the number of records; this yields a sample of 2266 observations. The data contain only the number of records affected by data breaches and do not include financial losses.[6]

[2] Information about data breaches is transformed into a contingency table, where each cell contains the number of data breaches of a given type and entity and each row (column) represents an entity (a type). A second contingency table contains the amount of data breached. The chi-squared distance (Izenman, 2008, p. 642) is the measure of discrepancy between rows (columns) and gives a weighted distance between rows (columns) of the table. The chi-squared distance is mostly used in correspondence analysis, a multivariate statistical technique used for exploring the association between cross-tabulated data (Izenman, 2008, Chapter 17).

[3] Chi-squared distance also plays a central role in correspondence analysis (CA), which is particularly apt for analyzing contingency tables. However, risk managers and insurance companies are primarily interested in whether different entities need different insurance contracts against cyber risks. Hence the need to focus on pairwise distances between entities, whose visualization is the primary goal of MDS. CA aims at approximating the total inertia of a contingency table that is a weighted average of squared distances between row profiles and the row centroid (see, for example, Izenman, 2008, p. 642). Proper approximation of pairwise distances between entities is not the primary goal of CA. For this reason, we focus on MDS when analyzing frequencies of differences among data breaches and use CA for validating MDS results.

[4] In the main body of the paper, the standard Silverman's rule smoothing parameter is implemented. In additional tests, available from the authors upon request, we also implement alternative estimation approaches following Alemany et al. (2013), which do not materially change our results; one explanation might be that the data are not too extreme in the tails.

[5] The bootstrap goodness-of-fit test by Villaseñor-Alva and González-Estrada (2009) is used to identify the optimal threshold value for the POT method.

[6] We apply our analyses to the original dataset (i.e., the number of records breached) and the natural logarithm of the number of records breached. An open research question is how to transform the number of records breached into actual loss data; one potential approach is the transformation described by Jacobs (2014); losses are estimated by $\ln(\text{loss}) = 7.68 + 0.76 \cdot \ln(\text{records breached})$. Jacobs (2014) generated this relationship between the number of records breached and the actual losses for the years 2013 and 2014 only, showing no significant differences in the two years. We use this formula to estimate insurance prices in Section 4.3. In additional tests, available upon request, we also present the results for alternative transformations presented by Jacobs (2014). The estimated prices vary substantially, depending on the type of transformation used, illustrating the need for future research on this topic.