CYBER-INSURANCE PRICING MODELS

James Basil Bardopoulos

Submitted under STA500W (University of Cape Town, Science, Masters) and SA0 (Institute and Faculty of Actuaries). Supervisors: Professor Thiart and Chris Olsen (CAS)

"Every great improvement… has come after repeated failures… Virtually nothing comes out right the first time. Failures, repeated failures, are finger posts on the road to achievement. "

Kettering (1876–1958), quoted by Boyd & Sloan (2002: 40)

DECLARATION

I know the meaning of plagiarism and declare that all of the work in this dissertation, save for which is properly acknowledged, is my own: this is the result of my own work and includes nothing that is the outcome of work done in collaboration except where specifically indicated in the text. The views and opinions expressed in this work are those of the author alone, and do not necessarily reflect those of others.

ABSTRACT

In the present technological age, where *cyber-risk* ranks alongside natural and man-made disasters and catastrophes – in terms of global economic loss – businesses and insurers alike are grappling with fundamental risk management issues concerning the quantification of *cyber-risk*, and the dilemma as to how best to mitigate this risk.

To this end, the present research deals with data, analysis, and models with the aim of quantifying and understanding *cyber-risk* – often described as 'holy grail' territory in the realm of cyber-insurance and IT security. In this dissertation, nonparametric severity models associated with cyber-related loss data – identified from several competing sources – and accompanying parametric large-loss components, are determined, and examined. Ultimately, in the context of analogous cyber-coverage, *cyber-risk* is quantified through various types and levels of risk adjustment for (pure-risk) Increased Limit Factors, based on applications of actuarially founded aggregate loss models in the presence of various forms of correlation.

By doing so, insight is gained into the nature and distribution of volatile severity risk, correlated aggregate loss, and associated pure-risk limit factors. Original contributions include:

- Application of versatile loss models with empirical support and the development of practical model selection techniques
- Derivation of model confidence sets for large cyber-losses
- Applications of existing techniques and models that, according to models identified as part of a systematic review and, to the best knowledge of the author, have not featured in cyber related academia

---

Keywords: cyber-insurance; cyber-risk; Increased Limit Factors; coverage; risk adjustment; correlation; severity distributions; aggregate loss models.

ACKNOWLEDGEMENTS

This thesis is dedicated to my son Alex; special thanks to my wife, mother, and rest of family for their unwavering support, perseverance, and encouragement; and finally, to my professor, for her dedication, guidance, and patience.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS AND ACRONYMS

ACTIVITY BASED COST: ABC	3.4
ADDITIONAL EXPENSE: AE	5.33
AGGREGATE LOSS DISTRIBUTION: ALD	4.1
AKAIKE INFORMATION CRITERIA: AIC	4.17
ANDERSON-DARLING: AD	4.20
BAYESIAN INFORMATION CRITERIA: BIC	4.18
BUREAU OF ECONOMIC ANALYSIS: BEA	2.18
CONSTANT RELATIVE RISK AVERSION: CRRA	2.7
DATA BREACH: DB	5.33
DISCRETE FOURIER TRANSFORM: DFT	4.27
DISTRIBUTED DENIAL OF SERVICE: DDOS	2.5
DOMAIN NAME SETTING: DNS	2.11
EXPECTATION MAXIMISATION: EM	4.26
EXTREME VALUE THEORY: EVT	2.2
GROSS DOMESTIC PRODUCT: GDP	2.9
HYPER TEXT TRANSFER PROTOCOL SECURE: HTTPS	2.11
INCREASED LIMIT FACTORS: ILF	1.3
INDIVIDUAL LIFE: IL	4.3
INFORMATION AND COMMUNICATION TECHNOLOGY: ICT	2.19
INSURANCE SERVICE OFFICE: ISO	4.8
INTERNATIONAL COMPUTER SECURITY ASSOCIATION: ICSA	2.10
INTERNET CRIME COMPLAINT CENTRE: IC3	2.16
INTERNET SERVICE PROVIDERS: ISP	1.2
KOLMOGOROV-SMIRNOV: KS	4.19
KULLBACK-LEIBLER: KL	4.41
LIMITED AGGREGATE SEVERITY: LAS	4.7
LIMITED EXPECTED VALUE: LEV	4.6
MAXIMUM LIKELIHOOD: ML	4.17
MAXIMUM LIKELIHOOD ESTIMATION: MLE	4.35
MEAN EXCESS: ME	4.9
MULTIVARIATE NEGATIVE BINOMIAL: MNB	4.32
NATIONAL CRIME VICTIMISATION SURVEYS: NCVS	2.17
OPERATIONAL RISK: OR	2.2
POTENTIAL SUITABILITY SCORE: PSS	2.21
PROPORTIONAL HAZARD: PH	4.12
RANDOM UTILITY MODEL: RUM	2.8
SUSCEPTIBLE-EXPOSED-INFECTED-RECOVERED: SIER	2.9
SUSCEPTIBLE-INFECTED-SUSCEPTIBLE: SIS	2.10
SYSTEM FOR ELECTRONIC RATE AND FORM FILING: SERFF	2.19
VALUE AT RISK: VAR	2.12
WORLD DEVELOPMENT INDICATORS DATABASE: WDID	2.9
WORLD WIDE WEB: WWW	E.3

Chapter 1

Introduction

> *"Cyberspace is real … so are the risks that come with it. It's the great irony of our Information Age … the very technologies that empower us … also empower those who would disrupt and destroy… "*
>
> (Obama, 2009)

1.1 Background

Cyber-risk, an umbrella term for risks associated with technology and information (CRO Forum, 2014: 3), is a significant threat with an estimated cost to the worldwide economy of over $600bn (McAfee & Center for Strategic and International Studies, 2018: 4). It encompasses a wide host of events caused by inadvertent activities (e.g. loss of data by employees, failure to maintain IT security to protect systems against unauthorised access, use, disruption, destruction, etc.), and criminal threats (e.g. phishing, social engineering, etc.) that can lead to various types of loss (e.g. remediation costs, business interruption, etc.), damage (from physical hardware all the way to diminished reputation) and liability (e.g. media, privacy, security, etc.). Notable examples of publicised events range from targeted breaches (e.g. Sony Pictures – Gara & Warzel (2014)) to large scale cyber-attacks such as *WannaCry*, ransomware that holds a computer hostage for bitcoins and ultimately disrupts "*critical and strategic infrastructure across the world*..." (World Economic Forum, 2018: 15).

There are a number of challenges to surmount: unilateral efforts in regard to managing *cyber-risk* are apparently futile (World Economic Forum, 2016: 78), whilst industrywide efforts require a consensus among many who remain divided as how best to contend with *cyber-risk* (e.g. polarised views concerning IT security and cyber-insurance (Böhme & Kataria, 2006: 3)). Uncertainty in the realm of a nascent insurance market has led to conservative underwriting; premiums are perceived to be large in relation to the level of cover – and thus low product penetration (UK Government and Industry, 2015: 22); and restricted coverage (high deductibles, low policy limits) that fails to protect firms against low frequency events with volatile severity (Solomon, 2017: 7). Similarly, cyber-insurers have to contend with a "*reinsurance barrier*" (Baer & Parkinson, 2007) – proposals include risk-linked securities and other forms of alternative risk transfer (CRO Forum, 2014: 39; BNY Mellon, 2016: 13). Many of these obstacles have been attributed to the following characteristics associated with *cyber-risk*:

1. Lack of reliable (frequency, but mainly severity) data for modelling and quantifying *cyber-risk* in an 'actuarial pricing' context (Radcliff, 2001; Cashell et al., 2004; Kesan, Majuca & Yurcik, 2005; Böhme & Schwartz, 2010)
2. The correlated nature of *cyber-risk* (Böhme & Schwartz, 2010; Baldwin et al., 2012; Mukhopadhyay et al., 2013), which has kindled fears of a global cyber-storm (US Department of Homeland Security, 2012: 1) precipitated by: widespread use of the internet, relatively few *Internet Service Providers* (ISPs), and reliance upon common IT software (Böhme, 2005; Böhme & Kataria, 2006; Wang & Kim, 2009; Laszka, Felegyhazi & Buttyan, 2014)
3. Other features associated with *cyber-risk* such as *interdependence* (i.e. degree of 'interconnectedness' between networks and systems) – (Kunreuther & Heal, 2002; Heal & Kunreuther, 2004; Ogut, Raghunathan & Menon, 2005a; Secretariat of the Security and Defence Committee Eteläinen, 2013; Laszka, Felegyhazi & Buttyan, 2014) and *information asymmetry* (Bandyopadhyay, Mookerjee & Rao, 2010; Böhme & Schwartz, 2010)

In academic circles, these factors have evidently influenced the development of *cyber-risk* models in several ways. Due to data related issues, frequency models appear to be more prevalent than severity (i.e. cost) models; aggregate loss models often assume constant severity leading to (possibly mixed) binomial distributions. Overall, the level of empirical

support is egregiously low. *Correlation* and *interdependence* have led to the consideration of copula (H. Herath & T. Herath, 2011), Markov processes (Barracchini & Addessi, 2014), and Bayesian belief nets (Mukhopadhyay et al., 2013). Many of these models, having been developed beyond the framework of economics and computer science, are abstracted from several peculiarities associated with aggregate *cyber-risk* – especially in the context of cyber-insurance and risk quantification:

- Aggregate loss distributions, risk measures (e.g. variance and *value at risk*), tail dependence, and the effects of *correlation* and *interdependence* in terms of different sections of insurance cover (e.g. business interruption, data breach remediation, etc.) have received little attention
- Loss models are generally underdeveloped in the field of cyber-science – applications concerning (much required) risk theory and aggregate loss modelling techniques have been largely neglected
- There is very little evidence in academic cyber related research of *Increased Limit Factors* (*ILF*s) which, in terms of import, are highly relevant given concerns in regard to 'low policy limits' and 'accurate pricing'

Nominal (academic) contributions from the actuarial domain can be found – despite the potential value that can be demonstrated when data is sparse (Solomon, 2017).

1.2 Research problem and objectives

The research problem appertains to key criticisms in respect of the present state of the *cyber-insurance* market, specifically regarding data and pricing issues, coverage limits, and correlated risks. The points at issue are apparent market deficiency in respect of coverage that is predominantly regarded as being inadequate with low policy limits which fail to provide the level of protection firms require, sparse data, and the related pricing concerns. This appears to correspond with a lack of academic loss models for determining *ILF*s in respect of correlated portfolios of *cyber-risk*, and primarily limited empirical support. Accordingly, in the context of *cyber-risk*, the present *research objectives* are as follows:

1. To review relevant sources of information and data, and, based on this, identify sources most suitable for deriving severity and aggregate loss distributions, and determining

implied *ILF*s

2. To model and explore key attributes associated with underlying loss distributions and the effect of correlation on these and associated risk adjustments

1.3 Notational and other conventions

In the matter of this research terms with a specific meaning or definition are generally italicised (e.g. *limit factor*, *ILF*, and *discount factor*). The prefix '*cyber*-' typically serves as a hyphenated modifier that relates the meaning of a word to information (e.g. storage, processing, communication) or technology (e.g. network, computer) – (Secretariat of the Security and Defence Committee Eteläinen, 2013: 12). Examples include *cyber-risk*, *cyber-insurance*; and *cyber-attack*.

In terms of notation, upper case font is typically used for variables and distributions, observations and density functions are in lower case, matrices and vectors are in bold font. The set of integers greater than zero is denoted by $\mathbb{Z}^+$. For random variable X with distribution, F, $X \sim F$ may be used (depending on the context, '$\sim$' may also be used to indicate approximate or rounded calculations); $X \overset{d}{=} Y$ implies random variables X and Y have the same distribution whilst $X \perp Y$ implies they are independently distributed. The indicator $1_{\{A\}} = 1$ if a given event A occurs (failing which, $1_{\{A\}} = 0$).

1.4 Scope and capacity of research

Many of the results produced (including *ILF*s, distributions, etc.) and assumptions made (e.g. loss count parameters, inflation) within the present research have been formulated based on specific interpretations, and purely for academic interests. There is no warranty for the applicability (or reliability) in other situations or contexts – the results depend critically upon the accuracy of the underlying data collated – details of those responsible for the collection and preparation of such data, in the first instance, are described accordingly in the relevant data section. This research has not been commissioned – there is no designation of *any* recipient, or entitlement to rely upon *any* of the results. The views

and opinions expressed in the present research are purely those of the author, and do not necessarily reflect those of any other individuals, professional body or organisation.

1.5 Outline of research

Chapter 2 reviews implemented *cyber-risk* models (and accompanying data, if utilised), in the context of a model taxonomy by field of study and design. Existing methods for incorporating special features associated with *cyber-risk* (e.g. information asymmetry, correlation, etc.) into models are described, followed by respective summaries in terms of relevant underlying variables (loss count or frequency of attacks, severity of cyber-losses and associated aggregate loss variables) and distributions. A wider investigation of data sources is performed for the purpose of identifying a source most suitable for *ILF* related analysis. This data source is taken forward into Chapter 3 for further scrutiny and preparation.

Chapter 3 describes the data identified in Chapter 2 – data fields are defined in terms of underlying cost activities, accompanied by examples of plausible cyber-insurance products that might cover such costs. These are revisited to formulate a hypothetical cyber-policy after considering regulation, specific correlations pertaining to the data, and after describing survey methodology and inflation adjustments. Once collated and prepared, and the data is ready for subsequent *ILF* related analyses, a preliminary exploration is performed, highlighting key statistics, comparing data before and after the application of inflation adjustments, with consideration of the extent of correlation between various cost categories.

Chapter 4 introduces by Risk Theory, followed by *ILF*s and underlying variables that form part of mathematical expressions to reflect risk and inflation. Attention is then turned to severity models, which are described in terms of composite distributions, model selection procedures, and tail behaviour, followed by a description of fundamental tools relating to aggregate loss models, including characteristic functions and related transforms. This precedes the description of the aggregate loss models with special consideration of associated *ILF*s, before closing with a simulation algorithm that is utilised as part of the validation in Chapter 5.

Chapter 5 provides the results of models in Chapter 4, based on ‘empirical’ data from Chapter 3, and is divided into two key sections:

1. Specification of severity distributions
2. Results and analysis, that include derivation of *ILF*s, aggregate loss distributions, and accompanying investigations pertaining to risk adjustments and correlation scenarios, and validations that consider internal and external consistency of results

Chapter 6 evaluates outcomes against initial objectives (§1.2); summarises contributions and limitations; followed by conclusions, recommendations, and, in finality, proffers direction in terms of future research.

Appendix A describes the search strategy utilised in Chapter 2 for the model review; Appendix B includes supporting material for the data in Chapter 3; Appendix C provides supplementary theory that is related to Chapter 4; Appendix D pertains to the models and results in Chapters 4–5; Appendix E provides information regarding *cyber-risk* and cyber-insurance.

Chapter 2

Review of models and data sources

"You have to know the past to understand the present."

(Sagan, 1983: 41)

Distinguishing features of *cyber-risk* are discussed in §2.1, followed by a review of deployed *cyber-risk* models in §2.2–§2.3, and accompanying data sources which, in the closing of this chapter, are considered in §2.4 as part of a data identification exercise and precursor to Chapter 3.

In terms of the review, Figure 2.1 is a chronological taxonomy that depicts *cyber-risk* models under the following four broad headings:

- ***Economic*** – models that consider the decisions and behaviours of individuals and organisations in the context of IT security and cyber-insurance, which are often brought under the lens of *insurance economics* for "*decision making under risk, risk management, and demand for insurance*" (Zweifel & Eisen, 2012). These typically focus on the "*demand-side*" (Böhme & Schwartz, 2010: 2) of *trade-off decisions* (e.g. for allocating resources between insurance and IT security) using *Utility* or *Decision theory*; in a few cases, insurance premiums are modelled as an output as opposed to an input alone (Kesan, Majuca & Yurcik, 2008; Yannacopoulos et al., 2008)
- ***Correlation based*** – models that include *copula* and *regression* techniques, with some models that straddle the *Economic* sphere (Böhme, 2005; Böhme & Kataria, 2006; Liu, Tanaka & Matsuura, 2007)

- ***Operational Risk*** (OR) – models that stem from *OR* quantification techniques such as those used to determine regulatory capital requirements, (European Commission, 2017). These encompass *Extreme Value Theory* (EVT) and *risk theory* (§4.2.1)
- ***Epidemic*** (and related) – models that utilise *Markov processes* and *regression techniques*, and are analogous to *epidemiological* compartmental (van Mieghem, Omic & Kooij, 2009; Parker & Farkas, 2011) or health insurance (Barracchini & Addessi, 2014) models

Furthermore, models and supporting data are subclassified according to the accompanying icon key. In particular, the type (e.g. *aggregate loss*, *frequency* model, etc. – icon shape) and focus area (e.g. *demand-side*, *financial loss*, etc. – icon colour) of models are indicated, as is the nature of any supporting data (icon fill type), which shall be considered in further detail in terms of the following factors:

- *Content*: *frequency* (e.g. cyber-attack or loss count) and *severity* (cost associated with cyber-incident) and *exposure* to risk (e.g. internet revenue per year; number of network connections: Appendix E.7). When summarised in respect of individual units of exposure, such *content* is referred to as being at an *individual level of detail* (otherwise it is regarded as being at an *aggregate* level)
- *Span*: number of years between the earliest and most recent year of data
- *Age*: number of years between the most recent year of data and a given *reference date* which is taken as the publication date of literature in the *model review*

These factors are used in §2.4 (with a modified 'reference' date) to gauge the potential suitability of various data sources; specific definitions (e.g. frequency, severity, loss, etc.) are then considered in the context of the data identified in this way. The scope of the *model review* includes mathematical models and supporting data in peer-reviewed studies (i.e. articles, journals, books, etc.), published between 1st January 2000 and 1st July 2016 (hereafter, *review period*), on the topics of (*cyber-risk*) management (e.g. IT security, insurance, regulatory intervention, etc.) and statistical assessment and modelling (e.g. risk measures, distributions, etc.). Library journals considered for this purpose are specified in Appendix A.1, together with the underlying *search strings* that are used to identify studies for full text review. Select studies that are related to, or serve as, predecessors for subsequent *cyber-risk* models are also included in the *model review*.

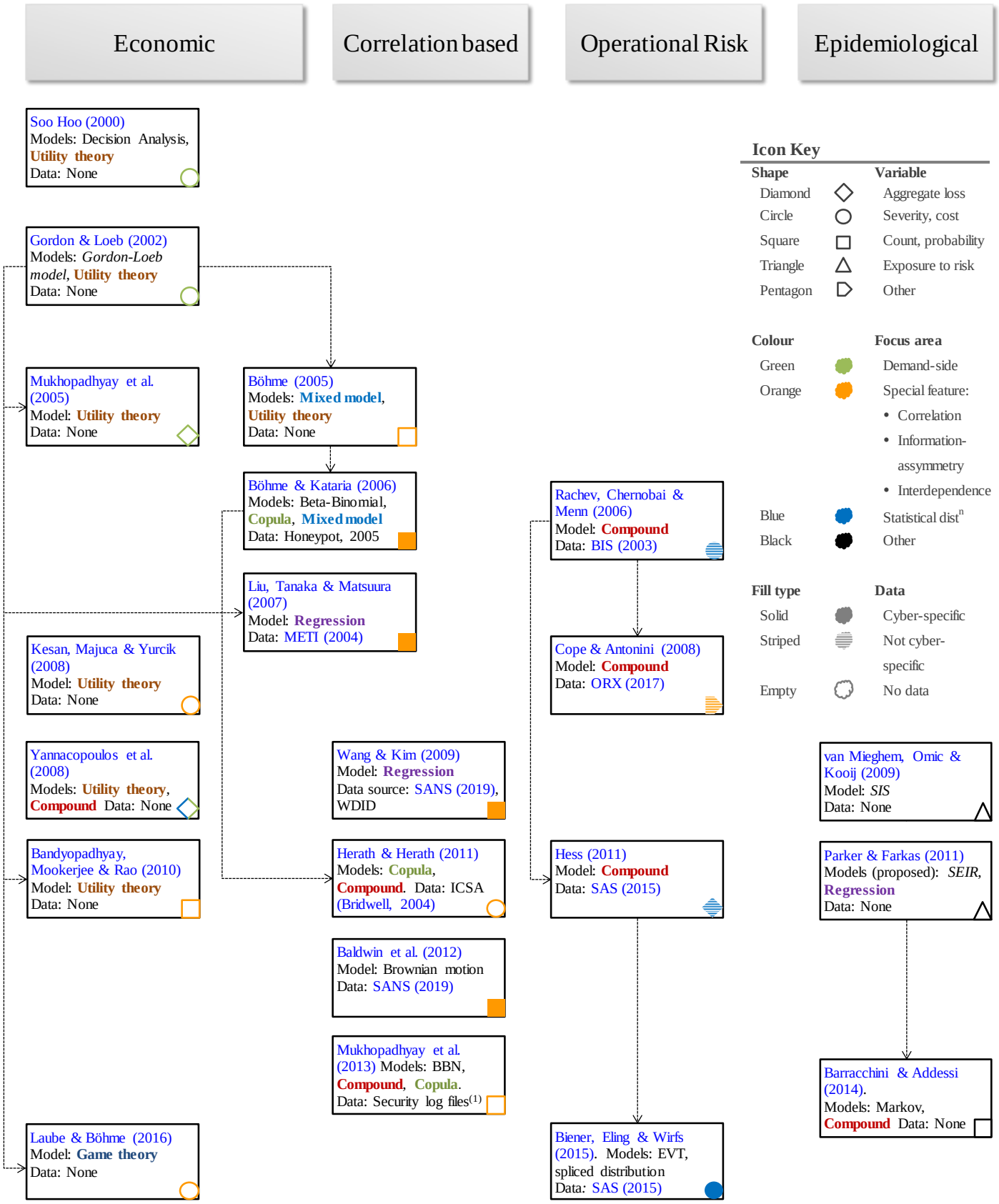


Figure 2.1 Overview of cyber-risk models Text colour: common model types. Abbreviations: Bank for International Settlements [BIS] (2003); Honeypot – Pouget, Dacier & Pham (2005); ICSA: *International Computer Security Association* – Bridwell (2004); Ministry of Economy Trade Industry [METI] (2004); Operational Riskdata eXchange Association [ORX] (2017); SysAdmin, Audit, Admin and Security [SANS] (2019); World Development Indicators Database (WDID): World Bank (2019). SEIR: *Susceptible-Exposed-Infected-Recovered*, SIS: *Susceptible-Infected-Susceptible*. Note (1): undisclosed source.

In this figure, common font colour (unrelated to icon colour) indicates similar techniques (e.g. *copula*, regression, etc.) or theory (e.g. *risk theory*, *utility theory*, etc.); connecting lines show methods that have been incorporated into subsequent models; and years associated with data sources are not necessarily related to underlying data periods (e.g. SysAdmin, Audit, Admin and Security [SANS] (2019) data relates to years 2003–2007). Where available, archived webpages have been referenced in R.1; refer to Table A.2 for data sources and corresponding references.

As can be seen, most *economic* (and all *epidemic*) models have not incorporated empirical data (hence icons with empty fill); in contrast, data has featured in all *OR* models (i.e. solid, striped fill colour), however, only Biener, Eling & Wirfs (2015) focussed on cyber-specific data (i.e. solid). A blue-green icon is used for Yannacopoulos et al. (2008) – (i.e. *demand-side* model with consideration for statistical distributions).

2.1 Special features

This section provides an overview of key *cyber-risk* features which sets the context for the models concerned (Figure 2.1: orange icons; §2.2). These features have been described as being central to cyber-insurance (Romanosky et al., 2017), some of which include causes of "*classic market failures in economics*" (Laszka, Felegyhazi & Buttyan, 2014: 3); whilst others (in isolation or combination) are more unique to cyber-insurance and IT security.

Information asymmetry

Information asymmetry, an imbalance of knowledge or information, stems from economic theory on quality uncertainty (Akerlof, 1970) and forms the basis of several studies in the *model review*. It incorporates phenomena that underpin a number of rudimentary actuarial principles (Allaben et al., 2008: 6); examples, commonplace in insurance and associated cyber-literature, include:

- *Moral hazard*, such as the potential for an insured party to alter its behaviour, upon insurance, in a way that adversely affects the insurer. This tends to increase the probability or severity (or both) of an insured loss. Security, in the context of *cyber-risk*, can have a similar effect (e.g. careless browsing, induced by online protection

software) – (Laszka, Felegyhazi & Buttyan, 2014: 5). Key tools for managing this include risk sharing mechanisms such as insurance deductibles (described further in §4.2.2) and premium 'incentives'. In terms of *cyber-risk*, security measures might be encouraged through partnerships – refer to Gordon, Loeb & Sohail (2003: 83) for examples in this regard

- *Adverse selection* is when material information is not fully accounted for (prior to insurance) due to certain restrictions (e.g. legal, regulatory, etc.) or asymmetric information. This can lead to an imbalance within an exposure group. Risk assessments and differentiated premiums are typically used by insurers to manage this, however, the ability to do so in terms of *cyber-risk* is impeded by the apparent 'under reporting' of security incidents (Gordon, Loeb & Sohail, 2003; Laszka, Felegyhazi & Buttyan, 2014). Whilst regulatory developments regarding disclosure (p3.3) have presumably narrowed the extent of this, they do not appear to have completely resolved the issue

Interdependence

Interdependence associated with *cyber-risk* can manifest in several ways. When claimed to be at the root of IT security (hereafter, *security*) issues (Laszka, Felegyhazi & Buttyan, 2014: 3), it typically refers to the degree of interconnectedness – the situation in which the security of one network (or 'player', in the context of *game theory*) is influenced by that of another. This has been attributed to some of the following:

- Increased probability of an incident (e.g. security breach) leading to increased premiums and demand for insurance (Ogut, Raghunathan & Menon, 2005b: 3–4); and exacerbated effects of *accumulations of risk* (e.g. by vulnerability: *Distributed Denial of Service, DDoS* attacks) – (Romanosky et al., 2017), leading to similar outcomes
- Positive '*externalities*', whereby actions are beneficial to both the enactor *and* others, which reduce firms' incentives to invest in security as a means of self-protection (i.e. '*free-rider*' problem), leading to general under-investment in this regard (Laszka, Felegyhazi & Buttyan, 2014, sec. 2.1)

Interdependence can also relate to organisational structures, for instance, vertically integrated processes and activities. This can have singular implications for *business interruption* coverage in the realm of cyber-insurance (e.g. losses across several firms can erode a common policy limit), which is an area of increasing concern (Marsh, 2015: 10).

Correlation

Various forms of correlation have been considered in the context of *cyber-risk*. Earlier studies have explored the effects of correlated attacks and failures within and across firms with extensions to aggregate correlation within insurance portfolios (Böhme & Kataria, 2006). Other models have focussed on correlation between IT assets (number of computers) and severity of loss (H. Herath & T. Herath, 2011), and the incidence of targeted attacks at the level of individual ports (Baldwin et al., 2012).

Peculiarities associated with correlation pertain to aspects such as self-propagating code, standardised software with common vulnerabilities, and the culture of monopolistic IT markets. Consider the example of antivirus software that can screen for and quarantine viruses before they spread to other computers, preventing damage such as deletion or corruption of files. In this way, the rate of computer failure due to viruses may successfully be reduced; however, this can also lead to accumulations (Böhme & Kataria, 2006: 3):

- The same code can be used to attack computers that are installed with the same version of software, due to common vulnerabilities (i.e. flaws, 'bugs') in that software (news of which often spreads quickly)
- Antivirus updates can usually be downloaded from a common website (e.g. hosted by the software vendor). If this website is compromised it can be used as a host for launching attacks against many users

2.2 Model review

The earliest *cyber-risk* models that fall within the review period can be seen in Figure 2.1 to have originated from the *economic* field:

- Soo Hoo (2000) formulated scenarios in a *decision analysis* with point estimates (i.e. as opposed to '*data*') based on computer security surveys and considered stochastic dominance in respect of various utility curves. Count and severity variables were modelled using simplistic bounded distributions (e.g. uniform, triangular); assumptions were required in respect of initial wealth, utility functions, and the various outcomes and probabilities associated with *decision trees*
- Gordon & Loeb (2002) developed a seminal model for determining the optimal level

of investment security to protect a given set of information technology assets

- Mukhopadhyay et al. (2005: 168) used a *decision tree* approach in a "*utility method backed premium calculation*" where claim frequency and severity were assumed to be stochastic variables (with unspecified distributions)

One of the first (and few) pricing models with empirical support was developed as follows:

- Böhme (2005, sec. 3) proposed a '*supply-side*' model for insurance premiums, based on an aggregate loss distribution (mixed binomial, Example 4.6 later) associated with *Bernoulli* risks with latent correlation and a constant claim severity of one (hereafter, 'unitised' severity). This was combined with a '*demand-side*' perspective based on *Constant Relative Risk Aversion* (CRRA) utility curves, to explore conditions for a market to be feasible. No data was modelled
- Böhme & Kataria (2006) expanded upon this work by exploring correlation within a firm (*beta binomial* failures) and across firms (using a *t-copula*), based on *honeynet* data (Pouget, Dacier & Pham, 2005) – courtesy Leurre.com, Eurecom – in respect of count (attacks) and exposure (total active 'sensors') which spanned less than 3 years (1st February 2003–30th September 2005) but was relatively up to date

This data was collected using *honeypots* (decoy computer systems) which are dedicated online hosts that simulate the activities of vulnerable systems and track network activity. A premium formula was considered in both cases – this incorporated a margin for "*safety capital*" (Böhme, 2005: 7; Böhme & Kataria, 2006: 10) which was based on the opportunity cost of capital (e.g. to protect against a 1 in 200 year event).

Loss distributions were evidently more established in the *OR* field, presumably due to the greater availability of relevant data:

- Rachev, Chernobai & Menn (2006, sec. 6.2) fitted various severity distributions and a homogeneous *Poisson* distribution (i.e. for count) to operational losses that were obtained from an undisclosed "*major European*" data provider. An empirical analysis was also performed in respect of Bank for International Settlements [BIS] (2003) *OR* data that spanned 1 year and was 4 years out of date. Exposure data (e.g. gross income, employees, etc.), which was available in BIS (2003), was not incorporated. Aggregate loss models were considered, however, these were not applied to the data

- Cope & Antonini (2008) measured empirical tail correlation in respect of different 'business lines' (e.g. finance, banking) and event types (e.g. malicious damage, failure), based on Operational Riskdata eXchange Association [ORX] (2017) data which covered 6 years (2002–2007) and was under 1 year old. Count and severity data were not modelled, although an empirical distribution was considered for aggregate loss

BIS (2003) data was based on a loss collection exercise that was carried out in the year 2001 across 89 banking firm members; ORX (2017) represented a collection and mutual exchange of operational loss information.

Concurrent to the formation of this groundwork for cyber-specific *OR* models, progress continued to be made in the *Economic* field where premiums were considered in the context of *demand-side* models based on *utility theory*:

- Kesan, Majuca & Yurcik (2008) constructed an 'asset pricing' model to measure welfare gains associated with cyber-insurance, based on *CRRA* utility; this was used to express the total premium (per dollar cover) a company would be ready to pay as a function of its aversion to risk (i.e. curvature of utility curve) and income level
- Yannacopoulos et al. (2008) proposed an aggregate loss model in a *Collective Risk* framework (§4.2.1) that utilised a *Random Utility Model* (RUM) to reflect subjectivity associated with the 'value' (i.e. severity) of privacy violations in the context of indemnity insurance. Simulation was used to illustrate this in the absence of data

Economic models started to place greater emphasis on correlation with techniques such as regression; unlike *demand-side* models (Kesan, Majuca & Yurcik, 2008; Yannacopoulos et al., 2008) many of these had empirical support:

- Liu, Tanaka & Matsuura (2007), motivated by Gordon & Loeb (2002), used regression to analyse the effect of the number of email accounts on breach probability. This was based on Ministry of Economy Trade Industry [METI] (2004) survey data which was 4 years out of date and reportedly spanned 2 years (but, in actuality, only spanned 1 year: Apr 2002 – Mar 2003). Aggregate count alone was modelled
- Wang & Kim (2009) used regression to describe spatial autocorrelations and the effect of the status and timing of joining an IT security convention. Count data (number of attacks) was sourced from a community-based firewall log system, *DShield* (SANS,

2019). This was between 1–2 years old and spanned 5 years (i.e. 2003–2007). Severity data was not explicitly modelled. Several exposure parameters (e.g. *Gross Domestic Product*, GDP, per capita) were based on *World Development Indicators Database* (WDID) of the World Bank (2019). It is unclear as to what period this data related; however, such indicators are available from as far back as 1960

METI (2004) data (internet archive; in Japanese) concerned the conditions of IT usage for businesses; *DShield* is the data collection engine underlying the so-called '*early warning system*' for the internet, '*Internet Storm Centre*', operated by SANS (2019).

Other progress in this field included an interesting model for depicting *information asymmetry* in the context of insurance and secondary losses:

- Bandyopadhyay, Mookerjee & Rao (2010) used *decision analysis* and *utility theory* to illustrate *information asymmetry* as the propagation of information levels between the parties (first, second) of insurance contracts. No data was modelled, and the severity of loss was assumed to have a uniform distribution. The probability of loss (frequency) was considered, however, no distributions were mentioned in this regard; aggregate loss did not come into question

Within this framework the existence of secondary losses, attributable to the disclosure of cyber-attacks (e.g. loss of stakeholder confidence), was portrayed as having a similar effect on the expected cost of claims (and thus risk premium) as a deductible of equal value. Accordingly, it was argued that *information asymmetry* between the insured and an 'uninformed' insurer (in terms of breach probability and secondary loss) generally leads to overstated premiums. This assumed that (said) nondisclosure was within the "*bounds of accounting norms and…regulatory obligations*" (Bandyopadhyay, Mookerjee & Rao, 2010: 7).

Attempts to model the peculiarities associated with *cyber-risk* started to emanate from the *epidemiological* field:

- Parker & Farkas (2011) depicted *cyber-risk* models as being analogous to compartmental models such as *Susceptible-Exposed-Infected-Recovered* (SEIR). However, concepts and approaches were purely descriptive; no mathematical representations were made. Refer to van Mieghem, Omic & Kooij (2009) for a

variation (*Susceptible-Infected-Susceptible*, SIS) that assumed a *Poisson* arrival process and theoretical underpinnings (e.g. Markov; mean field) in this regard

- Barracchini & Addessi (2014) described an analogous health-insurance model based on Markov processes and Kolmogorov (forward, backward) equations to capture the effect of computer components that transition between different states of operability

Neither of these (Parker & Farkas, 2011; Barracchini & Addessi, 2014) made use of supporting data. Indeed, one of the few examples of the use of 'empirical data' for modelling cyber-insurance premiums can be found in the following ('*correlation*') model:

- H. Herath & T. Herath (2011) estimated "actuarial premiums" for different levels of cover by simulating bivariate outcomes (count of computers infected by viruses; associated costs) with the aid of a copula. Data was based on the *International Computer Security Association* (ICSA) survey, reported by Bridwell (2004) – hereafter, ICSA 2004 – which covered a single year (2003) and was 9 years out of date

A Poisson process was considered for this, however, aggregate loss was not due to simplifying assumptions in respect of insured events (certain) and coverage (single claim per policy period). Data deficiency was highlighted as one of the key limitations – this is evident given that only 15 data points were used to fit distributions (with little information as to how costs were estimated).

Progression towards the development of a cyber-specific *OR* model gained momentum as proponents of *OR* modelling techniques (Rachev, Chernobai & Menn, 2006; Cope & Antonini, 2008) aimed their sites at an operational loss database much larger than previously considered:

- Hess (2011) simulated firm-level aggregate loss with a *compound Poisson* model (§4.2.4) to evaluate the impact of a financial crisis in terms of distributional characteristics (e.g. $VaR_{99.9\%}$). Severity was modelled using a *spliced density* approach (§4.2.3.1) in respect of individual years (2007, 2009) of SAS (2015) *OR* data that was at least 2 years out of date

This data source is purportedly the largest of its kind, featuring disclosed operational losses in excess of $100k, and has been considered (alongside the *spliced-severity* model) in subsequent studies (Biener, Eling & Wirfs, 2015; Eling & Wirfs, 2015, 2019).

At the other end of the diverse pool of *cyber-risk* models, relationships between 'contagious' threats to different security attributes were contemplated:

- Baldwin et al. (2012) modelled temporal relationships in respect of attack count and variations in frequency (i.e. *jumps*) and intensity based on *Brownian motion* with the assumption that *jumps* follow a Hawkes (1971) process

Correlation matrices were used to describe the interrelationship and behaviour between activities of *cyber-crime* on Internet Protocol services (e.g. *Domain Name Setting*: *DNS*; *Hyper Text Transfer Protocol Secure*: HTTPS) and a contagion matrix based on empirical data from SANS (2019), the same source considered by Wang & Kim (2009) which does not appear to feature severity data.

Model complexity continued to increase within the 'correlation based' *Economic* field:

- Mukhopadhyay et al. (2013) modelled the number of failures associated with different forms of security (e.g. firewall, security policy, etc.) as a multivariate normal distribution (i.e. *Gaussian copula*) from which posterior densities were determined, based on expert opinion and security log data (the source of which was not disclosed). Aggregate loss moments were based on a *collective risk* model in respect of a binomial distribution, on account of *unitised* severity

As for earlier *correlation* models (Böhme, 2005; Böhme & Kataria, 2006), premiums incorporated a risk margin; however, in this case, a variance adjustment was applied (risk adjustments are described later, §4.11). Premiums were also modelled in relation to expected utility for various degrees of risk aversion.

Following on from previous work in the *OR* space (Hess, 2011), one of the first analyses of empirical severity data, specific to *cyber-risk*, was performed. In doing so, an attempt to define *cyber-risk* was made and insurability from an Actuarial and Economic perspective was considered:

- Biener, Eling & Wirfs (2015) analysed SAS (2015) operational loss incidents that spanned 38 years (1971–2009), although the most recent year was over 6 years out of date. Whilst severity was explicitly modelled (using *spliced* densities, §4.2.3.1), count, exposure (e.g. revenue, equity), and aggregate loss data were not

This data comprised over 22k losses (in excess of $100k), however, less than 5% of these (~1k) were identified as being cyber-related and were analysed separately. Based on this, Biener, Eling & Wirfs (2015: 139) concluded that cyber and non-cyber severity distributions were "*considerably*" different (in terms of their distributions). In particular, the latter was claimed to be far heavier tailed than the former. However, the veracity of this claim cannot be substantiated due to there being insufficient information (e.g. regarding treatment of inflation; effect of excess; changes in mix with respect to cyber, non-cyber risks; etc.). Indeed, similar analyses based on larger samples of more recent data have indicated a larger *Value at Risk* (VaR_α – inverse of survival function at $\alpha \in (0,1)$, 4.68) at the 95% level for *cyber-risk* (Eling & Wirfs, 2019, sec. 3).

Further, spliced densities appear to have been determined in a fashion that led to sizeable discontinuities (alternative approaches are considered in §4.3.2). It can also be noted that whilst Biener, Eling & Wirfs (2015, n. 41) promised additional information in regard to certain analyses (upon request), this has not been forthcoming due to the proprietary nature of SAS (2015) data which typically requires licensed software (accessibility and other suitability criteria concerning data are considered in §2.4.2).

In the *Economic* field, models were further refined in terms of several *cyber-risk* features:

- Laube & Böhme (2016) explored the effect of disclosure requirements for an economy that comprised two (interdependent) firms ('*agents*') and a regulator ('*principal*'), in terms of an expected cost model. Breach probability was depicted as a function of security investment and incorporated parameters for *interdependence* and propagation of information (based on its effectiveness, firms' compliance, and detection error rate). Costs accounted for breach, disclosure, and security investment; data was not modelled

Concepts that came under consideration included the *principal-agent* problem, *Nash equilibria*, and *social optima* – refer to Laffont & Martimort (2009) for descriptions.

Other models

The following falls outside the *review period* (published by the *Journal of Cybersecurity* in December of 2016), however, it is included here due to its relevance for Chapter 3:

- Edwards, Hofmeyr & Forrest (2016) fitted distributions to Privacy Rights

Clearinghouse [PVC] (2016) data to gain insight into trends, large breaches, and associated costs

This data spanned 10 years (2005–2015) and was less than 1 year out of date (annual updates have since been provided); it includes several records, but neither exposure nor (reliable) severity information. As such, a pre-parameterised log-log regression model (Jacobs, 2014) was utilised to model costs as a function of records.

2.3 Summary of cyber-risk models

Table 2.1 summarises all the models from the *model review* (§2.2), together with that proposed by Jacobs (2014) for completeness, in terms of frequency (i.e. count), severity, and aggregate loss distributions and models (several of which are considered further shortly). Colour coded markers highlight where models or distributions are as follows: unnecessary due to the approach taken (e.g. Markov) or unspecified (grey); likely to misrepresent true underlying distribution (red); 'data dependent' (Klugman, Panjer & Willmot, 2004, sec. 4.2.4), having as many parameters as observations (orange); or plausible or conventional in general insurance practice (green). Also indicated are key outputs (e.g. risk premium, distributional parameters), type of exposure measure, and *special features* (§2.1).

Counting processes and related distributions

As this table shows, a variety of stochastic processes have been considered for count (e.g. number of cyber-related incidents, losses, etc.) and associated interarrival times. The *homogeneous Poisson* process (i.e. constant rate of arrival; independent, exponentially distributed interarrival times) is one common example (van Mieghem, Omic & Kooij, 2009: 2; H. Herath & T. Herath, 2011: 10). Variations (e.g. pareto, lognormal distributed interarrival times) have also been proffered in the context of privacy incidents (Yannacopoulos et al., 2008: 211–212).

The Bernoulli process is another example (Gordon & Loeb, 2002: 441; Böhme, 2005: 6; Böhme & Kataria, 2006: 6). *Non-homogeneous processes* (e.g. *Poisson-gamma* mixture: *negative binomial*: §4.2.5.3, Table D.3 (D.4)) have also been utilised (Edwards, Hofmeyr & Forrest, 2016: 5).

	Distribution, model			Output and related exposure		Cyber-risk feature		
Author(s)	Count (N)	Severity ($X_1,\ldots,X_N$)	Aggregate ($S=X_1+,\ldots,+X_N$)	Output(s)	Exposure-measure	Correlation	Interdependence	Info. asymmetry
Soo Hoo (2000)	Triangular, uniform	Triangular	Not modelled	Expected benefit - security	Customers	×	×	×
Gordon & Loeb (2002)	Bernoulli; probability functions	Given	Not modelled	Optimal investment - security	Not specified	×	×	×
Mukhopadhyay et al. (2005)	Not specified	Not specified	Not modelled [1]	Risk premium - insurance	None	×	×	×
Böhme (2005)	Binomial mixture	Constant (unit cost)	Mixed binomial	Correlation - claims	Risks	✓	×	×
Rachev, Chernobai & Menn (2006)	Poisson [2a]	Variety [2b]	Not modelled [2c]	Parameters - distributions	Loss events	×	×	×
Böhme & Kataria (2006)	Beta-binomial; mixed binomial (EM algorithm) [3]	Constant (unit cost)	Mixed binomial	Correlation, densities	Computers[*]	✓	✓	×
Liu, Tanaka & Matsuura (2007)	Regression - Gordon & Loeb (2002)	Not modelled	Not modelled	Parameters - regression	Companies[*]	✓	×	×
Cope & Antonini (2008)	Not modelled	Not modelled	Empirical	Tail ratios - distributions	Banks[*]	✓	×	×
Kesan, Majuca & Yurcik (2008)	Given	Not modelled	Asset pricing	Premium - insurance Welfare gains - security	Layer of cover	×	×	✓
Yannacopoulos et al. (2008)	Poisson [4a]	Random Utility Model [4b]	Collective risk [4c]	Cost, benefit - insurance	None	×	×	×
Wang & Kim (2009)	Regression	Not modelled	Not modelled	Correlation - residuals	Internet users, GDP	✓	✓	×
van Mieghem, Omic & Kooij (2009)	Poisson	Not modelled	Not modelled	Number, fraction - infected nodes	Not specified	×	×	×
Bandyopadhyay, Mookerjee & Rao (2010)	Not specified	Uniform	Not modelled	Optimal deductible - insurance	Not specified	×	×	✓
Hess (2011)	Poisson	Spliced (exponential, GPD)[5]	Compound Poisson	Parameters - distributions	Companies[*]	×	×	×
Herath et al. (2011)	Poisson	Weibull	Not modelled [6]	Premium - insurance	Computers[*]	✓	×	×
Parker & Farkas (2011)	SEIR [7]	Not modelled	Not modelled	None (descriptive - SEIR)	Systems	×	×	×
Baldwin et al. (2012)	Brownian motion	Not modelled	Not modelled	Correlation - intensity, size	Ports	✓	×	×
Mukhopadhyay et al. (2013)	Gaussian copula; binomial	Constant (unit cost)	Binomial	Risk premium - insurance	Organisation	✓	×	×
Barracchini & Addessi (2014)	Markov	Not modelled	Not modelled	Transition intensity, premium	Computers	×	×	×
Biener, Eling & Wirfs (2015)	Not modelled	Variety [8]	Not modelled	Spliced densities	None	×	×	×
Laube & Böhme (2015)	Gordon & Loeb (2002), probability functions [9a]	Expected cost; parameter-based [9b]	Not modelled	Expected cost, social optima, nash equilibria	None	×	✓	✓
Edwards, Hofmeyr & Forrest (2016)	Negative Binomial, Poisson [10a]	Log-log model; mixed log-skewnormal, lognormal [10b]	Simulation	Records - distributions, Predictions - regression	None	×	×	×

Table 2.1 Extant cyber-risk models Distributions, models − *green* (recognised or plausible in the context of general insurance), *orange* (data dependent), *red* (unrealistic, misrepresentative), *grey* (out-of-scope, not applicable, unspecified). Notes: 1) only moments considered. 2a) Homogeneous; non-homogeneous: lognormal, log-Weibull based functions; b) exponential, lognormal, Weibull, log-Weibull, Pareto αStable (log, symmetric); c) compound processes (e.g. Poisson, Cox) described but not applied. 3) *EM* – Expectation Maximisation. 4a–c) Per simulation example, *RUM* (utility – Pareto, random term – Normal). 5) *GPD* – Generalised Pareto Distribution. 6) Per simulation example (single claim per period, with certainty). 7) *SEIR* – Susceptible-Exposed-Infectious-Recovered. 8) Spliced (exponential, *GPD*), Weibull, gamma, lognormal. 9a) With parameters for *interdependence*, *disseminated information*; b) direct and disclosure costs, security investment. 10a) 'Daily' and 'large' respectively; b) log-log (Jacobs, 2014). Outputs: non-exhaustive examples. Exposure (*): conditional (e.g. given breach). Features: ✓ (considered) × (otherwise).

Other counting processes that have been considered (Table 2.1) include the 'self-exciting' *Hawkes* (i.e. 'arrival' rate increases due to previous arrivals), to reflect the effect of correlated attacks directed at systems and ports within a network (Baldwin et al., 2012), and the *Markov*, in relation to the state of damage of a device with internet connection (Barracchini & Addessi, 2014). Count models have also harnessed the *Poisson-gamma* mixture (i.e. *negative binomial*, §4.2.5.3; Table D.3 (D.4)) to reflect serial dependence between loss count distributions, albeit in the context of *OR* as opposed to *cyber-risk* (Cope & Antonini, 2008).

Severity and aggregate loss distributions

Constant severity has often been assumed (Böhme, 2005: 9; Böhme & Kataria, 2006: 16; Mukhopadhyay et al., 2013, sec. 5.2), which has resulted in several impractical aggregate loss models (characterised by binomial distributions, Example 4.6). In the case of Edwards, Hofmeyr & Forrest (2016: 10–11), aggregate loss was estimated using an independent regression model (Jacobs (2014), *log skew-normal* breach size) and a *negative-binomial* distributed breach count variable. Indeed, few severity models have been based on genuine cyber-related loss data – in the case of (Biener, Eling & Wirfs, 2015), this entailed an extensive classification exercise in respect of *OR* data SAS (2015). Other cases (Eling & Wirfs, 2015, 2019) have invariably involved similar (or identical) data and techniques such as *spliced distributions* (§4.2.3, §4.3), *EVT*, and *bootstrap goodness-of-fit tests* (Villaseñor-Alva & González-Estrada, 2009).

Correlation

Copulas have been a popular choice for modelling correlation, for instance:

- *t-copula* (elliptical: tail dependence) – Böhme & Kataria (2006) utilised this to model the dependence structure associated with a multivariate beta-binomial
- *Gumbel* (Archimedean: extreme distributions) – Herath & Herath (2011) modelled a bivariate Weibull distribution with this type of copula
- *Gaussian* (elliptical, multivariate normal: linear correlation) – Mukhopadhyay et al. (2013) used this to combine normal densities in respect of the number of failures associated with different vulnerabilities (e.g. firewall, security policy, etc.)

The present research, however, shall focus on the following areas that appear to have been neglected in terms of cyber-specific models (§2.2, Table 2.1):

1. *ILF*s have yet to be produced or modelled for different types of (correlated) loss
2. *Spliced severity* has yet to be considered in the context of an aggregate loss model
3. *Characteristic functions* (and related transforms) have yet to be utilised as a means of reconstructing aggregate loss distributions (§4.2.4.4: Algorithm 4.1)

2.4 The quest for data

Highly desirable data for the present research includes individual *severities* associated with cyber-incidents, as this can be used to fit severity distributions, model aggregate losses (for given loss count assumption or process), and, ultimately, calculate *ILF*s (at given limits).

Two sources considered thus far (§2.2) appear to have such information (at the desired *level of detail*):

- SAS (2015) – as described, this source is not feasibly accessible
- PVC (2016) – severity data, contained within text descriptions, is unreliable

Other data that may be of use for subsequent aggregate loss models include count and exposure (Klugman, Panjer & Willmot, 2004, sec. 4.6.11). In terms of the data sources depicted in Figure 2.1, such information is generally not sufficiently current (e.g. METI (2004), ICSA, Bridwell (2004)). Further, data that can be used to validate results (e.g. *ILF*s) is of interest. None of the sources considered thus far appear to have such information. This motivates the quest for alternative (hereafter, *'untapped'*) sources.

2.4.1 Untapped data sources

The following sources (*severity*: 1–5, *count*: 6–8, *exposure*: 9–10, validation: 11) are identified through online searches (e.g. industry studies, insurer filings, government and other reports, internet traffic websites, etc.):

1. The *Internet Crime Complaint Centre* (IC3) is a tool that has been in operation since the year 2000 for reporting internet crime complaints to the Federal Bureau of

Investigation [FBI] (2006). Annual reports are published with information pertaining to aggregate *frequency* (i.e. number) and *severity* (i.e. cost) of internet related crimes in the USA, often split geographically by state; some feature exposure information (e.g. number of website visits). The level of detail reported from year to year is not always consistent

2. The Inter-university Consortium for Political and Social Research [ICPSR] (2012) is a large archive of digital social science information and can be used to analyse online data from the *National Crime Victimisation Surveys* (NCVS), in particular, the concatenated files of interviews (claimed to be nationally representative sample sizes of households in the US) conducted between 1992 to 2014. Online queries can be used to retrieve frequency (e.g. number of incidents), severity (e.g. financial loss suffered), and exposure (e.g. number of computers or persons effected) data relating to identity theft crime incidents from mid-2004 to 2007 (year-end)
3. Ponemon Institute [PON] (2019) *cost of data breach* survey reports feature individual (years 2012–2015) and aggregate level severity, representing data breach cost estimates; frequency information (e.g. number of attacks per period, breach probability) is available at an aggregate level (country-year). Exposure is not explicitly reported, however, it appears to be implicitly available at an aggregate level, although various assumptions would be required for its estimation (e.g. dividing *probability of breach* and *customer churn* into *customer lifetime value*, associated with extrapolated *lost-business* costs)
4. NetDiligence [NetD] (2016) regards its analyses of claims data, underlying annual cyber-claim study reports published since 2011, as being the most comprehensive to date. Reports include summary statistics (extrema, mean, median) for aggregate claim payment amounts and numbers, and number of company records, grouped by claim type (e.g. crisis service, regulatory action, legal damages, etc.), type of data (e.g. trade secrets, non-card financial, etc.), and year. In this way, aggregate severity, frequency, and exposure data are available. According to NetD (2016), however, claim payment data has been collected from underwriters of various insurers, and comprises a mixture of open and closed claims that do not appear to have been adjusted to allow for any differences in coverage terms (e.g. policy and retention limits)
5. Lloyd's Market Association [LMA] (2008) provides technical and professional advice to its members, making available to them reports of market statistics and data (e.g. loss ratio triangulations, premium settlement and performance reports). Triangulation

reports are split by *risk codes* that LMA (2008) has mapped to various business and Organisation for Economic Co-operation and Development [OECD] (2018) classes, and includes information such as written premium, rates, and paid and outstanding claims, by quarter, for each *Year of Account* (YOA). *Risk codes* include *CY* (i.e. data and privacy breaches) and *CZ* (i.e. *physical* property damage, which excludes data and other electronic IT assets) in respect of cyber-security. Data for *CY* and *CZ risk codes* in triangulation reports are available for each *Year of Account* since 2013 and 2015 respectively. As such, aggregate severity (i.e. claim amounts) and exposure (i.e. assuming premium is a suitable proxy for this) are available to LMA (2008) members, however, frequency data (e.g. claim count) is apparently unavailable

6. Identity Theft Resource Center [ITRC] (2018) has, since 2005, provided annual data breach information that includes the name, location, and type (e.g. financial services, retail, etc.) of organisation, and the date, type (e.g. payment fraud, inside attack, etc.) and intensity (i.e. records) of breach. Exposure and severity, however, are lacking
7. Verizon Data Breach Incident Response [VER] (2019), since 2008, has produced annual reports of analyses of Information Security incidents based on contributions from a number of private and public enterprises that have included ITRC (2018) and NetD (2016). Interesting infographics are used to depict relationships and patterns between different types of threats and organisations affected, and security controls. However, the level of detail of information usually varies from one year to the next. Whilst aggregate frequency data (i.e. number of attacks and conditional attack probabilities for a given number of incidents) is typically available, unconditional exposure and cost (i.e. severity) are not
8. Digital Attack Map [DIG] (2013) was formed through a collaboration between Jigsaw (Google, 2016), formerly known as *Google Ideas*, and ArborNetworks (2019). DIG (2013) claims to host live visualisations of *DDoS* attacks from around the world that can be viewed from the 2nd of January 2015 onwards, and includes information about individual attacks such as occurrence date, type of attack (e.g. Transmission Control Protocol connection, fragmentation, etc.), their sizes (i.e. bandwidth), and the country source and destination of attacks. Several other websites exist with similar cyber-attack visualisations (Kumar, 2017)
9. *Bureau of Economic Analysis* (BEA) is an agency of the US Department of Commerce (2019), hereafter BEA (2019), that provides relevant exposure data (*GDP* by industry). However, count and severity data do not appear to be available. Annual records from

as far back as 1930 can be found, and an online query tool exists for sourcing more up-to-date, monthly, statistics

10. OECD (2018) has 35 participant member countries and provides country and sector level exposure (e.g. GDP and internet 'value'), as well as other potentially useful measures such as *Product Market Regulation* (PMR) in respect of *Information and Communication Technology* (ICT) which can be queried online. Frequency and severity data do not appear to be available. Depending on the query, data may span from between 1 year (e.g. *ICT* value added for the year 2011) to 57 years (e.g. *GDP* for the years 1960 to 2016, inclusive)
11. The *System for Electronic Rate and Form Filing* (SERFF) was developed by the National Association of Insurance Commissioners [NAIC] (2019) in the mid-1990s, and is regularly updated with new product filings submitted by insurers to regulators with information such as base rates, policy wordings, *ILF*s, rating factors, development factors, and financial indicators. As such, it may be regarded as implicitly consisting of a medley of frequency, severity, aggregate loss, and exposure related information (refer to Romanosky et al. (2017) for a content analysis)

Untapped data sources (blue stripes), together with previously modelled sources (grey stripes) are illustrated in Figure 2.1 according to their *age* and *span* (*reference date*: 31-Dec-16). These factors are used, together with the *content* and *level of detail* of underlying data, to determine the potential suitability (in respect of the present research objectives, §1.2) of each data source. One of the key features illustrated in this figure is the number of years that are spanned. This is factored into account in a more detailed comparison that assigns objectively measurable scores in relation to desirable features (i.e. *content* and *level of detail*, *credibility*: *span*, and *relevance*: *age*). Overall points are then used as a relative indicator of the potential suitability of data sources.

CAS Data Management Educational Materials Working Party (2008) proffers one such illustration which incorporates additional factors based on Dasu & Johnson (2003: 130). Some of these (e.g. completeness, accessibility, accuracy) shall be considered separately (§2.4.3), whilst others (e.g. conformity to business rules and schema) are extraneous to *PSS*s and, therefore, are beyond the parameters of the present research.

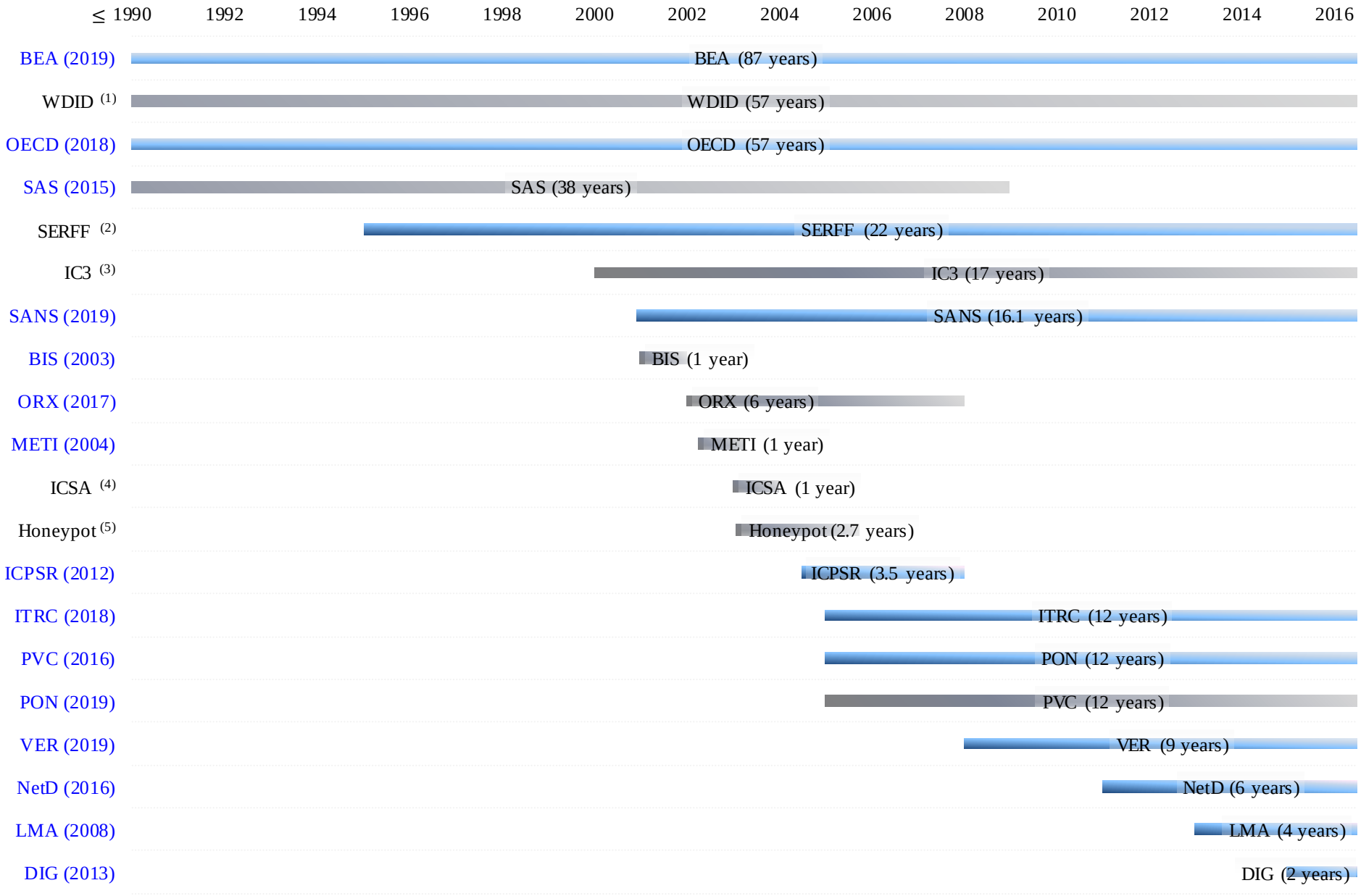


Figure 2.2 Data sources: span and age Years (horizontal, top): full period each data source spans (*reference date* is 31st December 2016). Blue bars: data sources used in literature reviewed (i.e. '*model review*'); grey: '*untapped*' data sources. Notes: (1) WDID of the World Bank (2019). (2) (NAIC, 2019). (3) (FBI, 2006). (4) ICSA reported by Bridwell (2004). (5) (Pouget, Dacier & Pham, 2005).

2.4.2 Measuring Potential Suitability Scores (PSSs)

Relative *PSS*s are derived in respect of each data source by awarding objectively measurable points to the following three factors:

1. Content and level of detail

For a given data source, a single point is awarded if *severity* data is available at an *individual level of detail*; if only available at an *aggregate* level, then half a point is awarded (if unavailable, no points are awarded). The *level of detail* (i.e. *individual* or *aggregate*) is determined on the same basis as before. Points are awarded for *frequency* and *exposure* data in a similar fashion; the total number of points for this factor can, therefore, range between zero and three. Note that if *severity* data is available at an *individual* level of detail, then so too is *frequency* (the opposite, however, is not necessarily true). Thus, sources with *individual severities* automatically score at least two points (one for each of *severity* and of *frequency* content). This is reasonable given that *individual severity* data is deemed to be the most desirable data (and *level of detail*) for the intended purpose.

Thus, this factor contributes up to three points. The *span* and *age* of relevant data underlying each *data source* (Figure 2.1) feed into factors 2 (*credibility*) and 3 (*relevance*).

2. Credibility factor

Generally, more data reduces the volatility associated with estimation errors. Campbell et al. (2006) perform experiments in this regard, and find that, as one could expect, datasets with more historical years of experience produce better estimates (i.e. in terms of accuracy) than those with fewer. At least three historical years of data are generally accepted as the minimum for an 'experience-based' Actuarial pricing exercise; for *ILF*s, which are oftentimes derived in respect of broader, less homogenous risks, more years are generally preferable. Thus, *credibility* points are awarded here by considering the *span* (in years) of relevant *severity*, *frequency*, or *exposure* data associated with each source. In particular, one point is awarded to a given source if relevant underlying available data *spans* at least five years, and half a point for three to five years; if *span* is less than three years, underlying data is deemed as failing admissibility requirements for further analysis in the present research (hereafter, '*inadmissible*').

Thus, this factor contributes up to one point.

3. Relevance factor

Sufficiently current data is typically required for Actuarial analyses (e.g. experience-based rating, risk assessment, etc.). Especially so when dealing with *cyber-risk* given the rapid evolution and, therefore, 'current' nature of technology, and the dynamic nature of business and regulatory environments (Kardoulaki, 2018). Indeed, Actuarial *cyber-risk* assessment has been likened to assessing a moving target (Cullina, 2017). As such, a given data source is deemed *inadmissible* if the *age* of underlying data is over two years; if age is between one and two years, half a point is awarded; if *age* is less than one year, a full point is given.

This factor, therefore, has a similar range of points as factor 2, as can be seen in Figure 2.4.

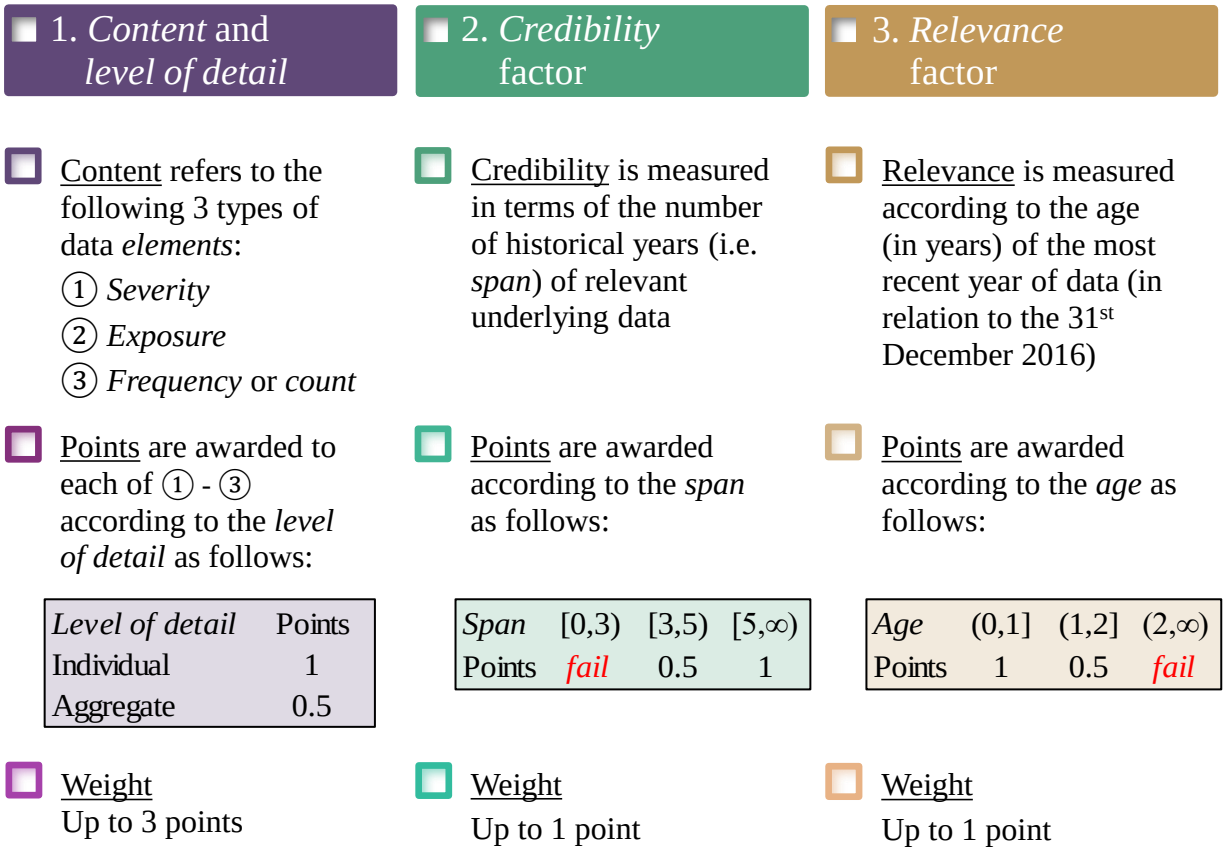


Figure 2.4 Composition of Potential Suitability Scores (*PSSs*) *Span* (for 2) based on number of consecutive historical years of data, subject to a maximum of years between the earliest such historical year, and the reference date (i.e. 31st December 2016).

Overall relative *PSS*

For a given data source, the overall *PSS* is determined by summing up the points awarded to factors 1–3 (Figure 2.4). This provides a useful way to initially screen all data sources using minimal information pertaining to relevant underlying data. Figure 2.5 illustrates relative *PSS*s (y-axis) for all data sources shown in Figure 2.1, by year last modelled (*x*-axis); 'untapped' data sources (§2.4.1) are grouped together as 'previously unmodelled'. The *PSS* for *SERFF* (NAIC, 2019) is calculated by assigning half a point to each type of data (i.e. *severity*, *frequency*, and *exposure*), which, in terms of Figure 2.4, equates to a total of one and a half points for factor 1. The underlying table used to calculate *PSS*s for every *data source* is provided in Table A.1).

The accompanying icon key in Figure 2.5 indicates the following additional information:

- Marker shape (circle, square, triangle, or diamond): for the three types of *data* considered, *severity* (circle) trumps *exposure* (square) which trumps *frequency* (triangle). Thus, any source with *severity* data has a circle marker; sources with *exposure* (or both *exposure* and *frequency*), but no *severity*, have square markers, and so on. A diamond shape can be seen for *SERFF* (NAIC, 2019) indicating it contains information for validation or verifying *ILF* results
- Marker fill-colour (light-blue, grey, or clear): light-blue is used for data sources that have all three data types (regardless of the *level of detail*); grey colour is used for *inadmissible* sources (i.e. fail *credibility* or *relevance*, or both); the default fill-colour is otherwise clear (i.e. white)

Marker outline (and font) colour (multiple): common colours are used for data sources that have comparable underlying data (e.g. purple: BIS (2003), ORX (2017), and SAS (2015) which comprise operational loss data); the default colour is otherwise black

Limitations of *PSS*s

Whilst the *PSS*s in Figure 2.4 provide a simple and practical means to rank data sources, there are several limitations in using this method to identify suitable data. For instance, other factors often associated with *data quality* (e.g. reliability of data field definitions, classification and reporting standards, quality control processes that aim to ensure internal consistency and completeness of data, etc.) would need to be considered separately.

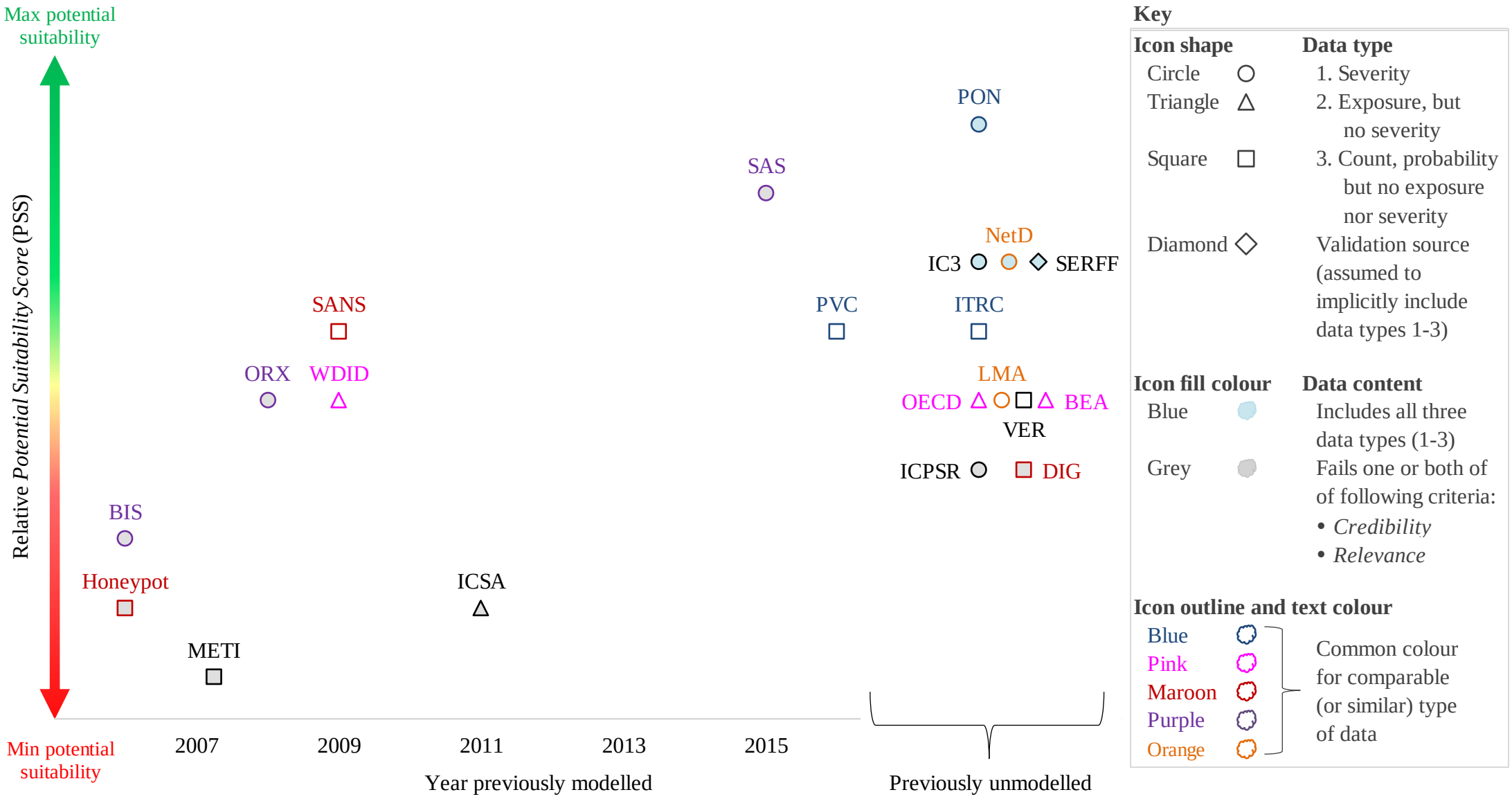


Figure 2.5 Potential Suitability Scores (*PSSs*) Scores (y-axis) represent a relative scale of points that are based upon objectively measurable points associated with the following factors: 1) *content* and *level of* detail; 2) *credibility* (i.e. number of years of data available [*span*]); 3) *relevance* (i.e. *age* of most recent year of data available). The *x-axis* is based on year last modelled. *Span* and *age* are based on potential availability of relevant underlying data (i.e. as opposed to that of actual data modelled) and are calculated in relation to a reference data of 31st December 2016. Previously unmodelled sources refer to those that do not form part of the *model review*.

As such, underlying data for sources with relatively high *PSS*s in this figure may still be 'substandard' in an absolute context. There is also a natural tendency for a positive correlation to exist between year and *PSS*, as subsequent updates to previously modelled data sources may not necessarily have been captured. Results should be interpreted accordingly. Delays between the publication of an academic paper and the most recent period of underlying data are assumed to be negligible (which is unrealistic). Furthermore, the weights ascribed to factors 1–3 are subjectively set and are not based on potentially more accurate scientific measures (e.g. based upon volume of data or underlying probability distributions).

2.4.3 Identifying primary and secondary data sources

None of the data sources considered thus far are necessarily *ideal* for the intended purpose at hand; however, some are preferable to others. This section motivates the selection of a *primary data source* (i.e. for present research objectives, §1.2) and *secondary sources* to consider for potential support and validation of primary data and associated results.

Primary data source

In terms of Figure 2.5, sources with severity data are prioritised over others (i.e. exposure, loss count).

As mentioned, BIS (2003), ORX (2017), and SAS (2015) are ‘comparable’ in the sense that they both feature operational loss data. The latter two sources reportedly have such information at an individual level of detail (i.e. per event), which, as described previously, is highly desirable for the present purpose. Other favourable attributes associated with these sources include: adherence to regulatory minimum reporting standards (relevant to *OR*); well-defined and structured ‘business line’ and ‘event type’ classifications; and quality control processes (Cope & Antonini, 2008). However, as mentioned, SAS (2015) data is not openly available and the underlying data is not necessarily cyber-specific.

Furthermore, *PSS*s indicate that these sources are *inadmissible*: they fail to meet predefined *relevance* criteria, with *ages* that range from 8 to 15 years; BIS (2003) also fails *credibility* requirements, with a *span* of less than one year.

Whilst ICPSR (2012), LMA (2008), IC3 (FBI, 2006), and NetD (2016), do have *cyber-specific* severity data, it is only available at an aggregate level of detail (furthermore, LMA (2008) has restricted access; and ICPSR (2012) has an age of 9 years, and thus fails *relevance* criteria).

PON (2019), on the other hand, scores highly in terms of *PSS* due to underlying (individual level) *severity* data which attracts maximum (relative) points for *credibility* (i.e. *span* of up to 12 years) and *relevance* (age less than one year) factors. Other points in favour of this source include:

- Cyber-specific: a wide range of data is available, such as: data breach probabilities and associated costs; customer churn estimates; number of records breached; and incident to discovery delay patterns
- Relevance to cyber-insurance: data is based on a broad range of cost-activities that can be related to various types of cyber-insurance coverage
- Comparability: data can be compared with other sources (as described shortly)

This source is, therefore, selected as the *primary source* of data for ensuing analyses and models in the present research.

Secondary data sources

Now that the primary source has been decided upon, secondary sources are identified as follows:

- OECD (2018): economic data for setting inflation assumptions (§3.3)
- *SERFF* (NAIC, 2019): for assessing the reasonableness of *ILF* results (§5.3.5)

Other secondary sources could also be identified (e.g. ITRC (2018) and PVC (2016)) to verify the congruency of associated data (e.g. 'records breached'). However, such an exercise lies beyond the scope of the present research.

Chapter 3

Description of data

> *"It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts."*
>
> (Doyle, 1901: 39)

The purpose of this chapter is to describe the data, assimilated from the *primary source* (§2.4.3), and the steps taken to arrive at a consolidated view for analysis in Chapter 5.

3.1 Underlying data

Data is drawn from Ponemon Institute (2012a–i, 2013a–j, 2014a–k, 2015g) global and country-level *cost of data breach* survey reports (hereafter, 2012–2015 years respectively) which form part of the PON (2019) data source (§2.4.1). As described, these reports feature estimated organisational costs in respect of publicly disclosed data breaches (loss or theft of personally identifiable records such as names and account numbers). Permission from the copyright holder, and fair usage considerations, are included in Appendix B.1. The following is a summary of relevant information and basic preparation for subsequent analysis:

- Costs are subdivided into four 'cost centres' – *A*: detection and escalation; *B*: notification; *C*: ex-post response; and *D*: lost business (hereafter, *classes A–D* respectively, with *class E* being the total)

- Years 2012–2014 (country-reports) – organisation-level costs, by class, are collated and US-dollar converted at exchange rates summarised in Table B.1
- Year 2015 (global) – *class E* costs (in US dollars) are depicted in various 'one-way' graphs (e.g. by rank of mean time to discover a breach); *R*-based image-scraping software, *Webplotdigitiser* (Rohatgi, 2013), is used to obtain this data from Ponemon Institute (2015g, fig. 20), before further scrutiny and adjustments (as described shortly)
- Mean and extrema (with respect to costs) are given, by class and year

In terms of the 2015 year, extracted costs appear to resemble the corresponding data points reasonably well (partly due to the ordering represented, which results in volatile and easily identifiable costs). A graphical comparison reveals 8 discrepancies (<2.5% of the data points). These are manually corrected; after doing so, the mean cost falls within 0.2% of the given value and extrema are exact.

Table 3.1 summarises *classes A–D* in terms of underlying activities and reputational damage associated with breaches, alongside examples of *first-party coverage* (i.e. which protect the insured's assets).

Composition of severity data		**Plausible cyber-insurance**
Class	**Associated costs**	
A: Detection and escalation	Detect and report breach (e.g. forensics, crisis management, internal communications, audit and assessment)	• *PortfolioSelect* (CyberEdge, Event Management) – AIG, Illinois National
B: Notification	Notify data subjects (e.g. create contact database, determine regulatory requirements, external experts)	
C: Ex-post response	Assist data subjects in aftermath of privacy event (e.g. help desk, inbound communications, investigations, remediation, legal, product discounts, credit monitoring and identity protection, regulatory fines and penalties)	• *Chubb Cyber Enterprise Risk Management policy* (Cyber Incident Response Fund) – ACE
D: Lost business	Abnormal churn, reputational damage, and diminished goodwill	• *Forefront portfolio* (CyberSecurity, Business Interruption) – Federal
E: Overall	Sum of *class A–D* costs	

Table 3.1 Costs (*classes A–E*) and possible coverage Descriptions for *classes A–E* are based on 'global' cost of data breach reports (Ponemon Institute, 2012d, 2013e, 2014f, 2015g); specimen products are purely illustrative examples of *first-party coverage* in respect of associated costs: AIG – Illinois (Murphy, 2013); ACE – (Cresenzi & Alibrio, 2016); Federal Insurance – (Daigle & Cresenzi, 2018).

There is a wide variety of products on the market: some are offered on a standalone basis which may provide (either one or both) *first-party* and *third-party coverage* (e.g. *AIG* and *ACE*, Table 3.1: *A*–*C*); others, as part of a special package that addresses multiple areas (US Department of Homeland Security, 2012); or through endorsements (e.g. Federal - Table 3.1: *D*) that extend cover under existing arrangements. Coverage and product variations are described further in Appendices E.2–E.3 respectively.

Regulatory environments

Activities associated with regulatory requirements and penalties (Table 3.1: *B*, *C*) can be expected given regulatory and legal developments such as the enactment of data breach laws in US jurisdictions (Greenberg, 2012, 2014, 2015; Kirsch & Greenberg, 2013; Digital Guardian, 2018); and notification requirements under the European Data Protection Directive, which have subsequently been strengthened across all European Union (*EU*) member states by General Data Protection Regulation (*GDPR*) – (European Commission, 2018). The impact of such activities is considered later in terms of severity and aggregate loss distributions (§5.3.2.1).

Class-level correlations

Class A costs are associated with activities that take place prior to public disclosure, whilst those in respect of *classes B–D* occur afterwards (Figure 3.1).

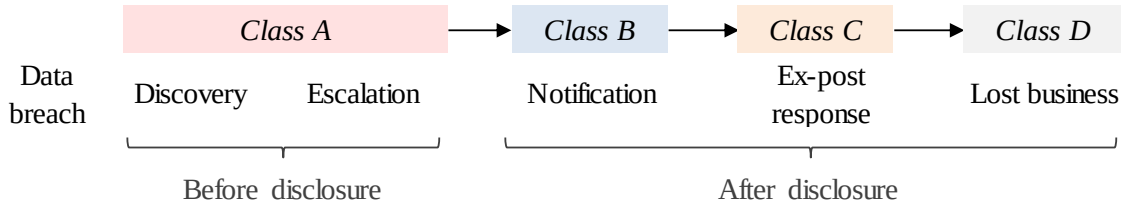


Figure 3.1 Loss generating process (*classes A–E*) Based on country-level cost of data breach reports analysed at a global level in Ponemon Institute (2012d).

This will be relevant for applications concerning interclass correlation in relation to loss distributions and associated characteristic functions in Chapter 5.

The first 10 rows of initial data are shown in Table 3.2. These losses relate to Australian organisations for the 2012 year.

Organisation	*Records*	*Class A*	*Class B*	*Class C*	*Class D*	*Class E*
1	4.7	144.6	19.4	147.6	469.5	781.2
2	65.5	1 005.5	26.0	283.8	2 419.5	3 734.9
3	9.9	460.8	76.3	312.6	17.3	867.0
4	6.5	583.3	120.6	339.3	43.1	1 086.4
5	15.8	550.8	13.5	623.2	153.4	1 340.8
6	26.3	754.3	82.3	255.2	1 105.6	2 197.4
7	33.4	763.5	65.4	237.9	1 155.0	2 221.9
8	2.5	235.0	23.8	202.4	66.2	527.5
9	36.4	419.7	49.5	1 802.9	689.4	2 961.5
10	22.4	787.8	89.9	260.6	1 017.9	2 156.2

Table 3.2 First 10 rows of initial data *Class A–E* costs, for the first 10 rows, are in Australian dollars; costs and records are per thousand (Ponemon Institute, 2012a).

Survey methodology

The same approach to capture and estimate costs appears to have been adopted each year, using *Activity Based Cost* (ABC) methodology which identifies and assigns costs to process-related activities, in respect of *classes A–C*, Table 3.1. For a given class, relative costs for each underlying activity are estimated across a linear scale which represents the monetary range of costs over all such activities. For *class D*, associated costs are extrapolated over the average customer lifetime (of respective firms). Each year represents different, but similar organisations (e.g. geographic presence, workforce size, etc.), which are interviewed over a given period (typically, 10 months). Breaches of less than 1 000 records, and more than 100 000 records are excluded (i.e. costs are *incidentally truncated*; *record-dependent* truncation from above and below is effectively applied).

Inflation

Costs, by class, are inflation-adjusted to make them comparable for analysis, whilst ensuring associated distributions are not overly distorted as a result. Key assumptions, for each survey year, include:

- Costs represent nominal values as at the time of interview
- Uniformity in regard to the timing of interviews (for the given ‘interview period’) and the timing of associated breach incidents (i.e. which occurred during the prior 12 months, (Ponemon Institute, 2012a, n. 8))

- Constant inflation rates, by class, over the entire period of inflation

The inflation period is from the average incident date to 30-06-16, or equivalently, the mid-point of the interview period to 31-12-16 (a convenient reference point for subsequent *ILF*s). Refer to Appendix B.3 for further detail regarding methodology.

3.2 Data limitations

The statistical accuracy of the data used in the present research for respective analyses (and ensuing results) relies upon:

- Survey data, associated methodology (e.g. *ABC* estimation), and the legitimacy of underlying participants' responses
- Data extraction methodology (i.e. 2015-year, *class E*) and severity-trend assumptions

Survey participants are described as constituting a "*representative, non-statistical sample*" of organisations (Ponemon Institute, 2015g: 29), and reports advise against the use of statistical inferences; as such, the data in the present chapter should be regarded as purely heuristic. Whilst analyses may reveal subsurface characteristics that provide insight into special features of *cyber-risk* (e.g. potential impact of correlated classes of *cyber-risk*) and results may align to (or possibly bridge) those of exposure-based approaches (e.g. '*power curve*' *ILF*s, Chapter 5), results cannot necessarily be generalised. In addition to the issue of '*record-dependent*' truncation:

- Samples are believed to be biased towards organisations with more established security measures
- Sampling bias is not measured, and non-participation is not reported

Whilst survey reports (within and across the years) appear to be numerically consistent, representational consistency is sometimes lacking (e.g. depending on the year, tabulated or graphical summaries may be used). For the 2015 year, this inhibits the ability to exact representative and consistently detailed information. Despite having the highest relative *PSS* (Figure 2.5), the data in hand is arguably of questionable veracity as far as an accurate, representative, experienced-based actuarial pricing exercise is concerned.

Uncertainty

It is foreseeable that uncertainty will remain regarding true costs underlying survey reports. There is no intention to reproduce such information, instead, the goal is to transform this information into stylised aggregate loss distributions, capable of reflecting the impact of correlation in terms of risk-adjusted *ILF*s. Uncertainty, in this regard, is communicated in the form of a range of results for various 'scenarios' (in terms of risk, correlation), based on a variety of models.

Homogeneity

To achieve a reasonable balance between sample size and homogeneity, for the present research, severities are grouped by *classes A–E*. Finer groupings (e.g. country, country-year) have not been used as there is insufficient data for intended analyses. Although measures that incorporate variance and proportional hazard transforms are considered later (Chapter 5), data is not explicitly transformed at this stage (e.g. natural logarithm, inverse, square root, etc.) to reduce heteroscedasticity, with respect to variance, within class-level groupings. Homogeneity is considered in further detail in §3.3, with supporting investigations in B.5

3.3 Preliminary exploration

Table 3.3 compares uninflated costs ('raw') with inflated costs in terms of various statistics. There are 15 severities in *class B* that have zero value (i.e. 785 non-zero losses). Classes otherwise have a one to one correspondence between the number of companies and the number of non-zero severities (i.e. loss count). This table appears to be incomplete: the missing values for 2015 (*classes A–D*) do not represent an issue as far as severity and aggregate loss modelling is concerned; the alternative might be to drop the 2015 year altogether, but this would decrease the count for *E* by roughly 30% ($\sim \frac{350}{1150}$).

After applying inflation to raw (uninflated) costs, mean costs appear to be aligned with one another over the years; the exception being *E* (2015 year), which has a relatively larger mean than earlier years (i.e. $3.96m vs. overall average $3.59m).

Mean severity

Class	Data	Year: 2012	2013	2014	2015	All year	Inflation(1)
A	Raw	590.5	608.1	632.2		613.0	
	Inflated	693.2	693.8	695.2		694.2	3%
B	Raw	234.7	211.0	198.5		212.3	
	Inflated	234.7	211.0	198.5		212.3	0%
C	Raw	860.8	818.6	929.0		872.9	
	Inflated	1 064.8	975.1	1 053.8		1 029.4	4%
D	Raw	1 378.2	1 298.6	1 431.8		1 371.7	
	Inflated	1 534.5	1 418.5	1 525.9		1 491.0	2%
E	Raw	3 064.2	2 936.3	3 191.5	3 736.6	3 272.8	
	Inflated	3 527.1	3 298.3	3 473.4	3 957.8	3 588.4	2%
	Entities	209	277	314	350	1 150	

All year

Class	Data	Count	Min	Max	Std dev	Kurtosis	Skewness
A	Raw	800	14.9	4 181.0	668.3	7.31	7.31
	Inflated	800	17.0	4 829.8	757.0	7.44	2.43
B	Raw	785	0.0	3 553.1	346.0	32.52	4.63
	Inflated	785	0.0	3 553.1	346.0	32.52	4.63
C	Raw	800	6.7	8 797.5	1 029.4	13.57	2.94
	Inflated	800	8.3	10 663.7	1 217.7	13.87	2.98
D	Raw	800	3.0	11 869.4	1 874.7	6.45	2.40
	Inflated	800	3.3	12 786.5	2 037.0	6.39	2.39
E	Raw	1 150	75.3	28 290.6	3 231.7	7.26	2.16
	Inflated	1 150	82.6	29 965.1	3 527.6	6.95	2.13

Table 3.3 Summary statistics (uninflated vs. inflated) Raw data (i.e. uninflated costs) source: Ponemon Institute (2012a–i, 2013a–j, 2014a–k, 2015g). Note (1): inflation: annualised compound rate (subject to minimum of 0%) with respect to mean (uninflated) costs, by class, for survey-years 2012 and 2014 (approximately equal to square root of the ratio of 2014 to 2012 mean costs, less one). Inflation period: mean interview date to 31st December 2016. Relevant values in $US '000s. Count refers to non-zero severities.

As can be seen in Table 3.3, uninflated costs for 2014–2015 suggest a rate of inflation of ~17% (i.e. $\frac{3.7}{3.2} - 1$), however, this increase is only partly reflected by corresponding inflated costs, which reveals a shortcoming in assuming constant inflation over the years (Appendix B.3).

Uninflated (orange) and inflated (blue) cost distributions in Figure 3.2 appear to agree with one another in broad terms, with greater alignment between lower quantiles (e.g. below median) and greater deviation between upper quantiles (i.e. above median). *Classes A* and *C* appear to be similar with *class B* having the shortest tail, and *D* the longest.

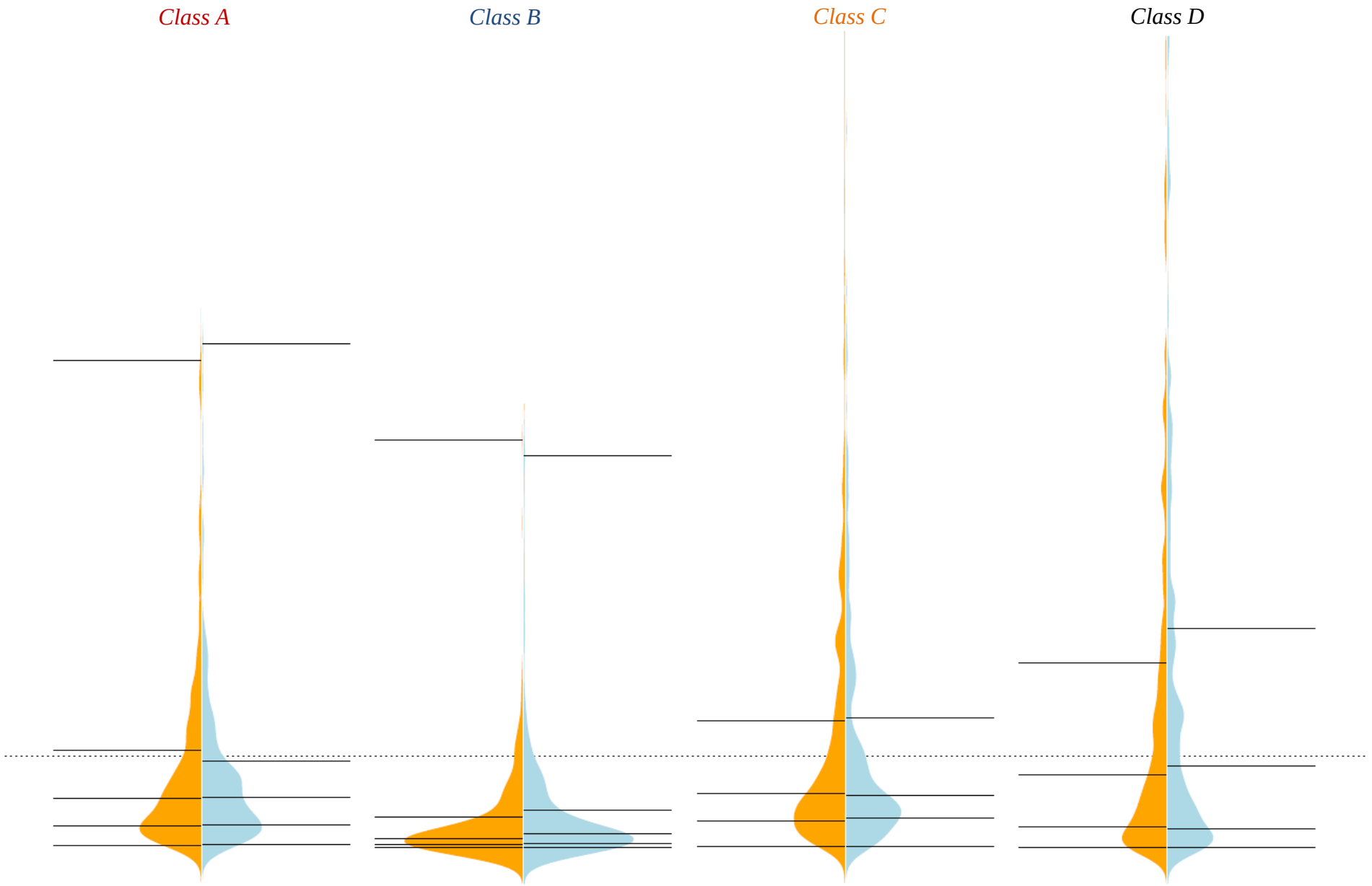


Figure 3.2 Bean plots (*classes A–D*, uninflated vs. inflated) Beans: orange (uninflated cost); blue (inflated); solid lines: class-specific quantiles; dotted line: overall mean cost (costs are depicted on the y-axis). Uninflated costs: Ponemon Institute (2012a–i, 2013a–j, 2014a–k). Software: R, package: '*Beanplot*' v1.2 (Kampstra, 2008).

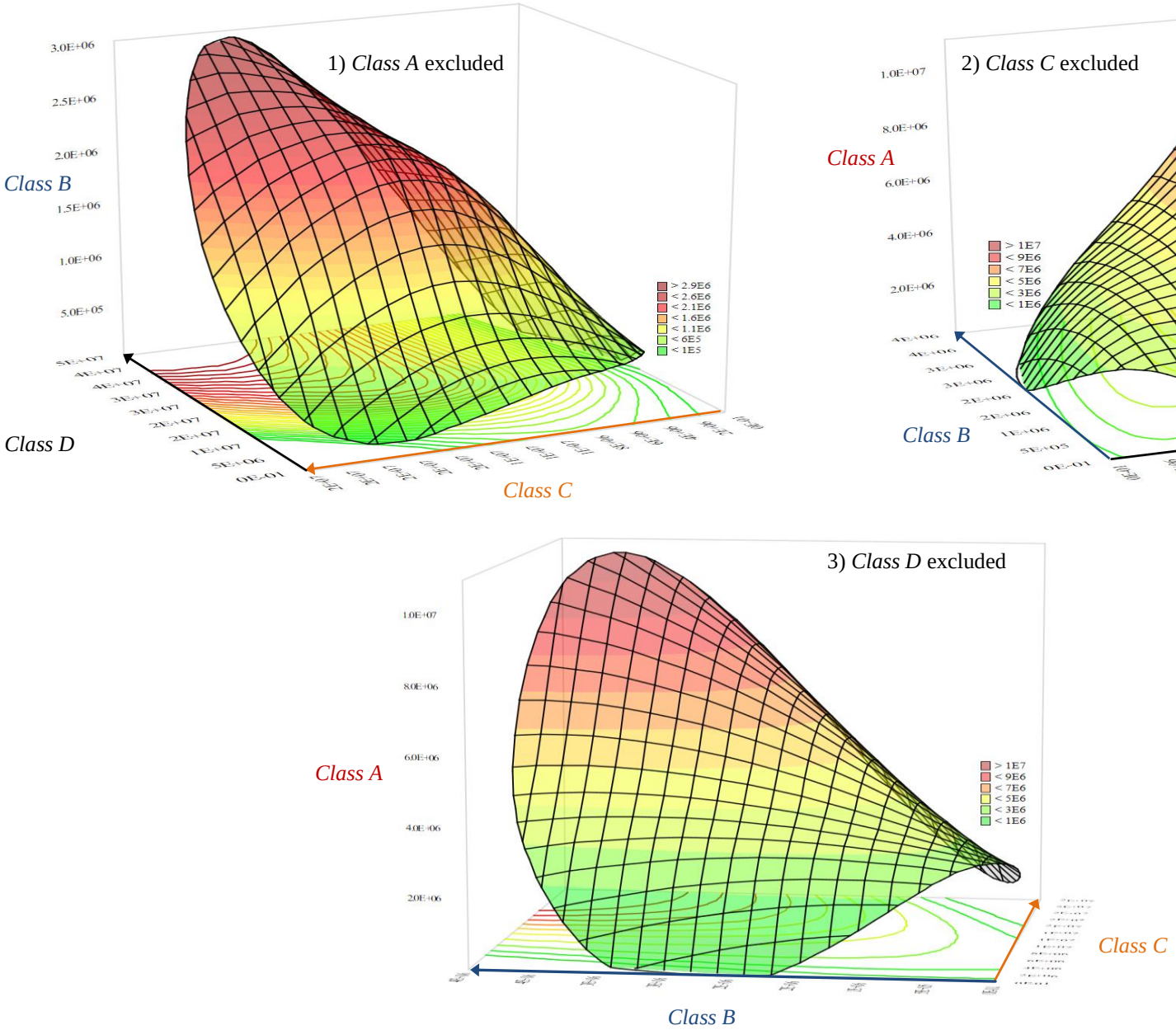


Figure 3.3 Surface plots (inflated costs) Data: costs from Ponemon Institute (2012a–i, 2013a–j, 2014a–k), inflated to end of 2016 year. Ascending order of costs ($US) indicated by arrows (axes); surface-plot (spline fitting method) coloration relates to vertical axes (green – low, red – high costs). Software: Statistica v13.2 (Statsoft, 2016).

As can be seen in Figure 3.3, correlation between *A*–*D* can be positive or negative over different ranges (however, Pearson's correlation coefficient confirms an overall positive correlation between these classes). For instance, red contour lines (i.e. large *B*) in (1) indicate *C* and *D* may either be negatively (lower *C*) or positively (larger *C*) correlated (although the latter form of correlation appears to dominate, with a coefficient of 0.46).

Tail dependence is now examined. Define the *tail ratio*, ω, in respect of n observed pairs (x_i, y_i), $i \in 1,2,..,n$, as follows:

$$\omega(z) = \frac{\sum_{i=1}^{n} \mathbf{1}_{\{x_i, y_i > z\}}}{\sum_{i=1}^{n} \mathbf{1}_{\{x_i > z\}}}, \qquad z \leq \max_{i=1,..,n}\{x_i\} \tag{3.1}$$

where indicators such as $\mathbf{1}_{\{x_i \geq z\}}$ are defined as previously in §1.3 (Cope & Antonini, 2008, sec. 5.1; Parodi, 2014, sec. 28.3.2). Figure 3.4 illustrates ω for different pairs of classes.

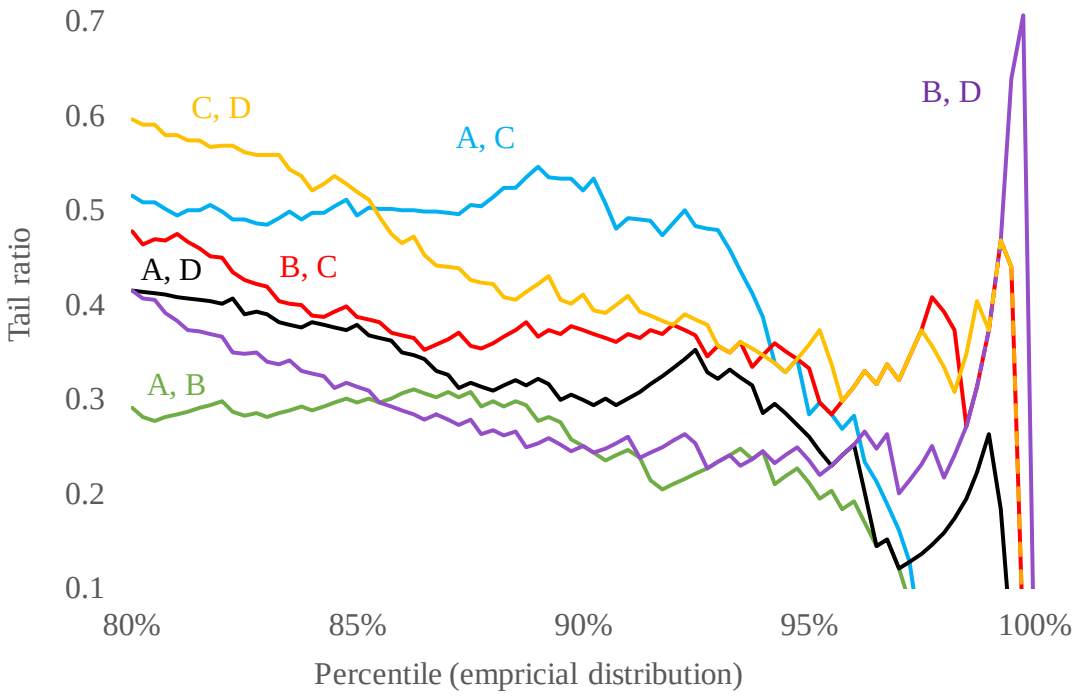


Figure 3.4 Tail dependence ratios Percentile corresponds to empirical quantile above which tail ratios are determined. Underlying costs: Ponemon Institute (2012a–i, 2013a–j, 2014a–k), inflated to end of 2016 year.

As the largest loss in each class corresponds to a different organisation, tail ratios in Figure 3.4 inevitably decline to 0 as the (empirical) percentile increases to 1 (this, and volatility at high percentiles, are typical shortcomings associated with this estimate). Prior to this,

however, there is some evidence of tail dependence: compared to those with *A* (excluding *A*, *D*), pairings with *class D* appear to have stronger tail dependence at high percentiles:

- *Class D*: tail ratios for (*A*, *D*), (*B*, *D*), and (*C*, *D*) exhibit an upward trend as the percentile approaches ~99%; (*B*, *C*) is somewhat similar to (*C*, *D*) in this regard
- *Class A*: tail ratios for (*A*, *B*) and (*A*, *C*) start declining after ~90%; (*A*, *C*), however, does maintain the highest tail ratio over the range 85%-95%

Before closing this chapter, year-on-year homogeneity is considered in terms of the following:

- Country composition of years 2012–2014, according to the number of organisations associated with each country-year combination (proportional to width of adjoining lines, or 'edges') in Figure 3.5
- Comparison with Jacobs (2014) log-log regression model in respect of costs, X, and records breached, R: $\ln X = 7.68 + 0.7584 \ln R$, as studied by Edwards, Hofmeyr & Forrest (2016: 10); as a means of independent validation, by year (2012–2014), and for different country groupings in Table 3.4 and Figure 3.6

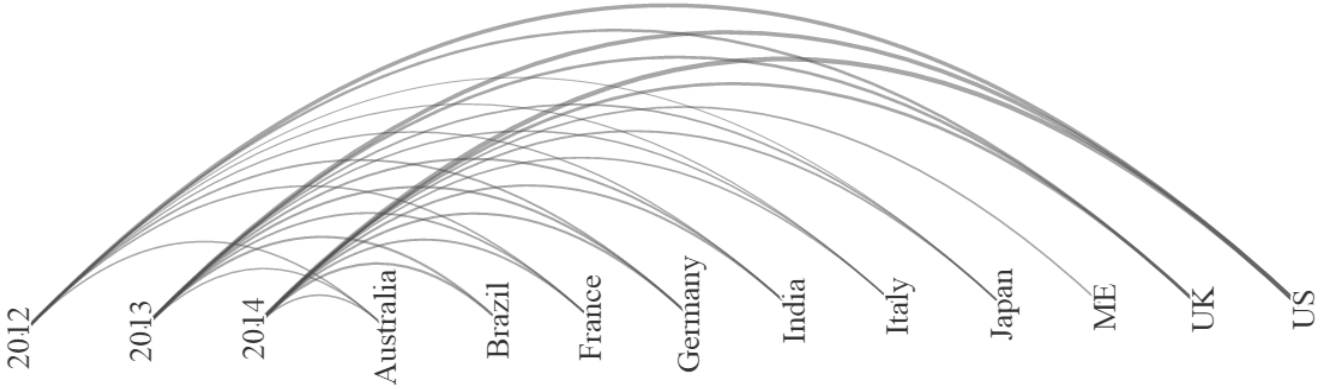


Figure 3.5 Country-year mappings Line width is proportional to number of organisations. Underlying costs: Ponemon Institute (2012a–i, 2013a–j, 2014a–k); *ME*: Middle East (Saudi Arabia, United Arab Emirates). Software: Microstrategy (2016).

In most cases, Figure 3.5 indicates that each country is represented in each year (e.g. Australia has three edges adjoining years 2012–2014; the same can be said for France, Germany, etc.). Brazil and the Middle East (i.e. *ME*, representing Saudi Arabia and United Arab Emirates) are the only two exceptions. These countries were introduced in 2013 and 2014 respectively. Edges, for a given country, generally appear to be consistent with one another (i.e. in terms of width). Typically, only discernible differences can be noted in this way, however, this is confirmed by the actual mix of countries, by year, in Table B.4.

Attention is now turned to Jacobs (2014) log-log model in Table 3.4.

	USA ($\ln X = a \ln R + b$)			***Non-US***			***Global***
Year	*a*	*b*	r^2 (coeff)	*a*	*b*	r^2 (coeff)	r^2 (coeff)
2012	0.801	7.252	0.574	0.884	5.710	0.615	0.593
2013	0.766	7.562	0.523	0.992	4.655	0.617	0.539
2014	0.750	7.800	0.503	1.002	4.492	0.611	0.572
2013 - 2014	0.758	7.680	0.512	0.957	4.980	0.611	0.562
Jacobs (2014)	*0.758*	*7.680*	*0.512*				

Table 3.4 Log-log model by year and country group Costs (*class E*), *X*, and records, *R*, based on Ponemon Institute (2012a–i, 2013a–j, 2014a–k).

Log-log model regression slope and intercept parameters (Table 3.4: USA, 2013–2014) correspond exactly with those determined by Jacobs (2014), as can be expected given the same underlying data. Perhaps of greater interest is the effect of different country and year groupings, as illustrated in Figure 3.6.

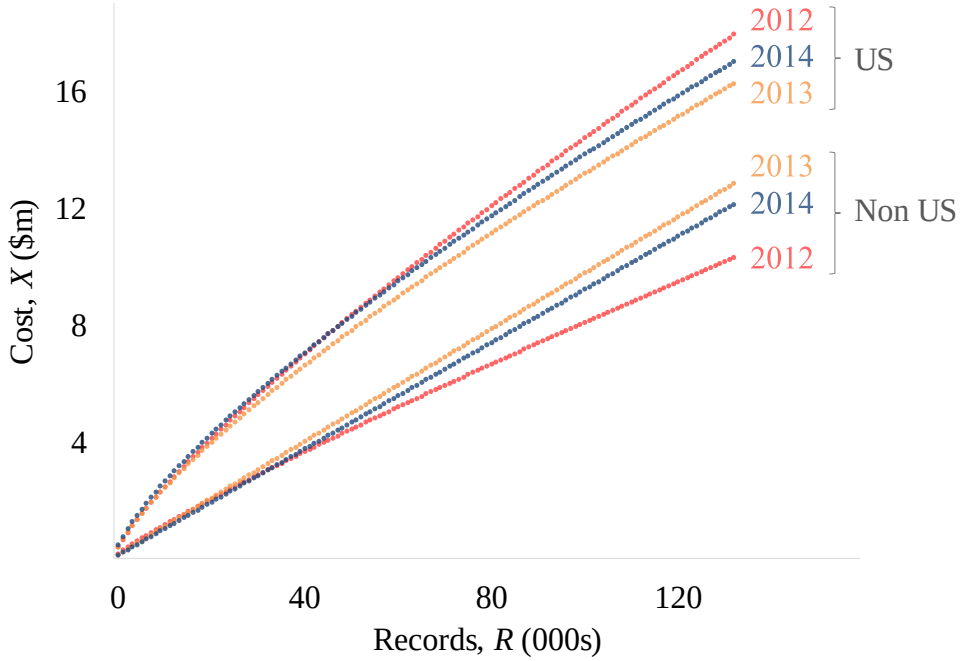


Figure 3.6 Log-log model (year, country groupings) Underlying costs: Ponemon Institute (2012a–i, 2013a–j, 2014a–k).

Whilst greater 'homogeneity' in regard to log-log regression could be achieved with country and year groupings, as depicted in Figure 3.6, resulting sample sizes would not be conducive for subsequent *ILF*s that involve fitting large-loss distributions to even smaller subsets (Table 5.1). Further, using Levene's (1960) test, the null hypothesis that costs (years 2012–2015, inflated to 2016) exhibit homoscedasticity (with respect to variance) cannot be rejected (up to 25% significance). Refer to Appendix B.5 for further support.

Chapter 4

Loss models and underlying theory

"All models are wrong, but some are useful"

(Box, 1979: 202)

4.1 Overview

The aphorism *"all models are wrong, but some are useful"* (Box, 1979: 202) springs to mind when deciding upon a suitable model construct. The purpose of this chapter is to describe models of aggregate losses (total amount of loss that occurs in a defined period in respect of a group of homogeneous risks) and describe the methods that are used to determine the distribution of these (*Aggregate Loss Distribution*, ALD). These models are stylised representations of possible outcomes in respect of data breaches, the cost and number of which are uncertain. In particular, applications of these models for determining *ILF*s in Chapter 5 are based upon the culmination of subjective interpretations pertaining to the underlying data (Chapter 3: inflated costs), and approximations that attempt to balance (apparent) realism with simplicity.

There are essentially two key parts of this chapter: theory and models. The former is divided into the following four sections (introduced by risk theory, §4.2.1) which support the latter as depicted in Figure 4.1:

1. *ILF*s (§4.2.2): this covers mathematical foundations of limit factors; basic definitions (e.g. different bases for limits); key concepts (e.g. consistency properties); related functions (e.g. Mean Excess); and various types of adjustments pertaining to risk, deductibles, and inflation (determined in respect of models)
2. *Composite* (*severity*) *models* (§4.2.3): this concerns spliced densities which are determined later in respect of the severity data from Chapter 3 (underpinning every model), and considers model selection in terms of information criteria and *goodness-of-fit* measures
3. *Aggregate loss models* (§4.2.4–§4.2.5): basic tools for working with and determining distributions are described together with key algorithms that form the basis of Models 4.3–4.6
4. *Simulation* (§4.2.6): this focusses on Monte Carlo simulation and related functions (e.g. quantile, Value at Risk) which are utilised to verify and investigate output relating to Models 4.3–4.6

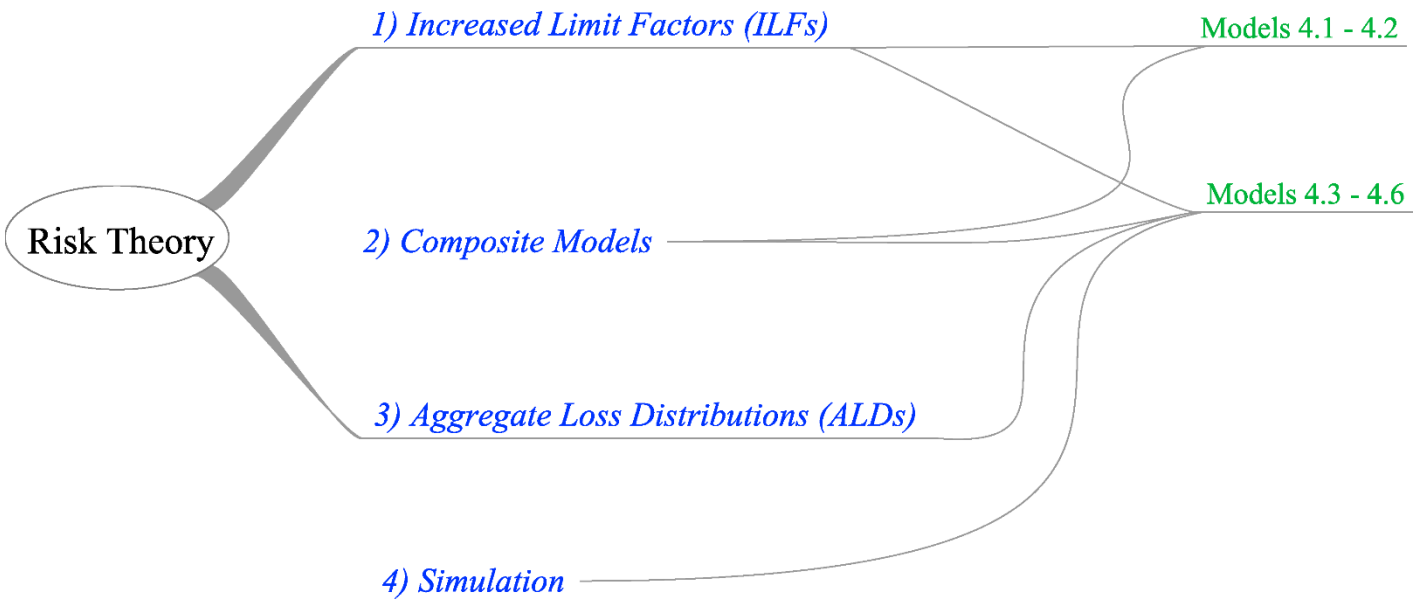


Figure 4.1 Outline of theory and model links Theory 1–4 (blue, in addition to risk theory which introduces 1 and 3); Models 4.1–4.6 (green; all models rely upon 1 and 2; 3 and 4 are only utilised in support of Models 4.3–4.6). Generated using *Freemind* (Müller et al., 2004).

4.2 Background theory

4.2.1 Risk Theory

Aggregate loss, S, represents the total amount for a given period and group of risks,

$$S = X_1 + X_2 +, \ldots, + X_N, \tag{4.2}$$

where N and X_is can be defined from two perspectives of *risk theory*, namely:

- *Collective Risk* (CR): *loss count*, N, and (non-negative) *severities*, $X_1, ..., X_N$, are random variables with *independence assumptions* as follows: N does not depend on the severity of loss; for N given, X_is are *i.i.d.*, independently with respect to count
- *Individual Risk* (IR): here, N denotes a fixed number of risks with respective losses, X_is, that are independently distributed (as opposed to *i.i.d.*) random variables with *mixed cdf*s that may have mass at point zero (i.e. for the probability of no loss)

Following the notation in Klugman, Panjer & Willmot (2004: 142), the first three moments μ'_{S1}, μ_{S2}, and μ_{S3} of S (4.2) for the *CR model* are as follows:

$$\begin{aligned} \mathrm{E}S &= \mu'_{S1} = \mu'_{N1}\mu'_{X1} = (\mathrm{E}N)(\mathrm{E}X) \\ \mathrm{Var}S &= \mu_{S2} = \mu'_{N1}\mu_{X2} + \mu_{N2}\mu'^{\,2}_{X1} = \mu'_{N1}\mu'_{X2} + (\mu_{N2} - \mu'_{N1})\mu'^{\,2}_{X1} \\ \mathrm{E}(S - \mathrm{E}S)^3 &= \mu_{S3} = \mu'_{N1}\mu_{X3} + 3\mu_{N2}\mu'_{X1}\mu_{X2} + \mu_{N3}\mu'^{\,3}_{X1} \end{aligned} \tag{4.3}$$

In terms of μ_{S2}, if $N \sim$ Poisson (i.e. $\mu'_{N1} = \mu_{N2}$), then $\mu_{S2} = \mu'_{N1}\mu'_{X2}$: this represents the '*minimum variance*' (Miccolis, 1978: 43) and is considered later in the context of risk adjustments (§4.2.2.2). Derivations of 4.3, based on *mgf*s and compound Poisson models, can be found in Mildenhall (2005, sec. 3.2). For insurance risks, S (4.2) may represent the total amount paid on claims, in relation to coverage, in a given period, under a defined group of policies (Klugman, Panjer & Willmot, 2004: 135; Liu & Wang, 2017: 362). In this context, *IR* models are a natural construct for a health policy, group life, or pension fund (Boutin-Dufresne, 2003, chap. 1). The *Individual Life* (IL) model is a special type of *IR* model where any risk can only have a loss count of 0 or 1 (i.e. no multiple losses),

(Vernic & Sundt, 2009: 5). However, such models are often too restrictive for describing general insurance risks, which are typically framed in terms of a *CR* model (Parodi, 2014, sec. 6.1.2). As Burnecki, Janczura & Weron (2011: 294) remarked, non-insurance risks (e.g. *OR*, credit) have also been modelled using this framework.

Conventionally, different groups of homogenous risks are modelled separately. When these groups comprise a portfolio of risks, combinations of *IR* and *CR* models can be particularly useful. For instance, model aggregate losses in respect of a portfolio that comprises several sub-portfolios, aggregate losses for each sub-portfolio could be modelled using a *CR* framework; and the aggregation of these could be based on the *IR* framework (in line with underlying *independence assumptions*, 4.2, *CR*). A similar set-up is utilised in §4.4 to model aggregate losses in respect of correlated *classes A–D* (Chapter 3) using a *CR* framework, before combining with an *IR* framework. A special type of *CR* model, which reduces to an *IR* model, is also considered.

Convolutions for compound distributions

Let F_S be the *ALD* for aggregate loss, S, with *CR independence assumptions* (4.2) – this is a compound *cdf* of the following form:

$$F_S(s) = \sum_{n=0}^{\infty} p_N(n) \Pr(S \le s \mid N = n) = E_N F_X^{*(n)}(s), \qquad 4.4$$

where $p_N(n) = \Pr(N = n)$ and $F_X^{*(n)}$ is the *n-fold convolution* of *cdf*, F_X, defined by:

$$F_X^{*(n)}(x) = \begin{cases} \int_0^x f_X(y) dF_X^{*(n-1)}(x-y)dy & n = 2,3,... \\ F_X(x) & n = 1 \end{cases} \qquad 4.5$$

(Klugman, Panjer & Willmot, 2004: 141).

4.2.2 Increased Limit Factors

An *ILF* is a multiplicative factor that is applied to the premium at a *basic limit* to determine the premium at an *increased limit*. *Basic limits* typically refer to the lowest levels of coverage provided, (Werner & Modlin, 2010: 192). However, in principle, any non-

negative limit can be contemplated for this purpose (hereafter, the term *base limit* is used instead of *basic limit*). As a precursor to *ILF* derivations, limits are first described in further detail, followed by some practical considerations.

Limit definitions

A policy limit refers to the maximum amount payable under an insurance policy, either overall, or in respect of a particular section of a policy (Lloyd's, 2019), hereafter, '*coverage section*'. This may be expressed on several bases, for instance:

- *Per-occurrence*: the limit restricts the amount payable in respect of all losses caused by a common occurrence (e.g. IT security failure, data breach, etc.). In this regard, the definition of an '*occurrence*' is crucial, for example, continuous, repeated, or related acts may be deemed as a single occurrence
- Aggregate: the maximum payout in respect of all covered losses is restricted. Such limits can apply to each of, and across all, the coverage-sections of a cyber-policy (Hiscox, 2017)

Limits that restrict the level of payout in respect of an aspect of a coverage section are often referred to as sub limits. For instance, \$5k: data recreation – Munich Re (2015); \$500k–\$2.5m: regulatory defence expenses for small firms (< \$100m turnover) – Deloitte cited by Jensen & Rosenthal (2015: 18). Data breach notification limits can also be equated with or defined in terms of number of persons affected (e.g. Illinois National – Murphy (2013); *National Liability and Fire* (NLF) – Selleck (2015)). As such, cyber-policies may have a multitude of different types of single limits (i.e. maximum amounts payable in respect of individual claims) and compound limits that apply more than one limit to the covered losses (e.g. split limits, ACE – Cresenzi & Alibrio (2016)). Limits may be eroded by legal defence costs incurred by the insurer, in defending the insured against liability claims (i.e. defence inside or within limits). Similar bases may be used to define deductibles (claims in excess of which, subject to limits, are covered).

As depicted previously (Figure 3.1), losses for *classes A–D* are associated with a common occurrence; for demonstrative purposes, limits are assumed to apply as follows:

- *Per-loss*: applies to individual costs (i.e. *classes A–D*)
- *Per-occurrence*: applies to total cost (i.e. *class E*)

Deductibles, where these apply, can be defined on similar bases. The interplay between these, in terms of *ALD*s and *ILF*s, is explored further in Chapter 5.

Practical considerations

Solomon (2017: 7) argued in favour of high *per-occurrence* deductibles and limits, on the premise that frequency (of cyber-related losses) is predictable and, therefore, manageable by the insured, whilst severity is volatile and better managed by insurers through risk-pooling mechanisms. Further, events that trigger losses on multiple policies are described as being rare, mainly relating to cloud-based risks that can be managed separately (i.e. negating the purpose of *aggregate deductibles*). According to a Betterley market report (2006, cited by Baer & Parkinson (2007: 52)), in respect of several major cyber-insurers, limits (presumably on an *aggregate basis*) can be as high as $7.5m–$25m and $7.5m–$50m for different *first-party* and *third-party coverage-sections* respectively. However, it is not difficult to find numerous examples of losses which have far exceeded such levels, such as the Anthem security breach (Osborne, 2015). Whilst there are reports that London-Market insurers have capacity available for $100m limits (Arthur J. Gallagher, 2017), it has been speculated that $1bn limits may be necessary to provide the insured with adequate protection (Chon, 2015) for certain types of *cyber-risk*.

With this overview in mind, relevant variables, functions, and adjustments, associated with *ILF*s, are now formalised.

Limited random variable

The *limited random variable* $X^{(b)}$ is defined as follows:

$$X^{(b)} = \min\left(X, b\right) \tag{4.6}$$

where X is a random variable and $\{b : b > 0\}$ is some limit.

Limited moments

The *Limited Expected Value* (LEV) is the first-order (raw) moment of a *limited* severity random variable. More generally, consider the *limited variable* $X^{(b)}$ (4.6), and suppose X has a *cdf* and *pdf* denoted by F and f respectively; the *limited* k^{th}-order moment of X, when limit b applies, can then be expressed in terms of the *Riemann-Stieltjes* integral:

$$\mathrm{E}X^{(b)k} = \mathrm{E}\min(X,b)^k = \int_0^b x^k dF(x)\,dx + b^k(1-F(b)) = \int_0^b kx^{k-1}S_X(x)\,dx \qquad 4.7$$

where $S_X = 1-F$ and $k \in \mathbb{Z}^+$ ($k=1$ yields the *LEV*). The reader is referred to Lee (1988: 52) for a graphical illustration of 4.7 and Klugman, Panjer & Willmot (2004: 32) for a proof based on *integration by parts*; Riemann-Stieltjes (Stieltjes, 1995) and related integrals are covered in basic calculus, not here. It can be noted that for some *cdf*s, the k^{th}-order moment may not necessarily exist $\forall\ k \in \mathbb{Z}^+$ (e.g. *Pareto*).

Consider $j > 1$ ordered limits, $0 < b_1 < b_2 < ... < b_j$; to approximate the *LEV*, $\mathrm{E}X^{(b_j)}$, from the survival function, S_X, of the severity variable, X (4.7), the product of $S_X(b_{u+1}) + S_X(b_u)$ and $0.5(b_{u+1} - b_u)$ can be summed over $u = 1,..., j-1$. Following on from 4.7 (with $k=1$):

$$\begin{aligned} &\lim_{b\to\infty} \mathrm{E}X^{(b)} = \mathrm{E}X \\ &\frac{d\mathrm{E}X^{(b)}}{db} = 1 - F(b) = S_X(b) \\ &\frac{d^2\mathrm{E}X^{(b)}}{db^2} = -f(b) \end{aligned} \qquad 4.8$$

Refer to Bahnemann (2015: 46) for a proof of these and other characteristics such as the following: $\mathrm{E}X^{(a)} \le \mathrm{E}X^{(b)}\ \forall\ 0 \le a \le b$; $\mathrm{E}X^{(b)}$ is continuous on $X \ge 0$; $\mathrm{E}X^{(b)}$ is *concave-down* on $b \ge 0$; and $\mathrm{E}(aX+c)^{(b)} = b + a\mathrm{E}X^{\left(\frac{b-c}{a}\right)}$ for constants $a > 0$ and c (relevant for inflation adjustments, described in §4.2.2.2). These properties are revisited shortly in the context of desirable features of *limit factors*.

Limited Aggregate Severity (LAS)

Let the aggregate loss in respect of limited severities (hereafter, *LAS*) be $S(b)$ defined by:

$$S(b) = \sum_{i=1}^{N} X_i^{(b)} \qquad 4.9$$

where $b>0$ is a given limit, and X_is are severities, and N is the loss count, as for the aggregate loss in 4.2. This gives rise to the *limit factor*, γ, for a given *base limit*, a, defined by:

$$\gamma(b) := \gamma(b;a) = \frac{\mathrm{E}S(b)}{\mathrm{E}S(a)}, \quad \mathrm{E}S(a), \mathrm{E}S(b) > 0 \qquad 4.10$$

where $a,b>0$. The term *limit factor*, for the purpose of the present research, refers to both *discount factors* and *ILF*s, defined as follows:

- *Discount factor*: $(a>b>0) \Rightarrow \gamma(b;a) \in (0,1)$; in this case, a could represent the highest limit of coverage, or, in the context of coverage *without-limits*, $a \to \infty$
- *ILF*: $(0<a \le b) \Rightarrow \gamma(b;a) \ge 1$; the conventional definition of an *ILF*, where a and b represent '*basic*' and *increased* limits respectively

There are several approaches to derive *limit factors*. Examples includes *top slicing* (Michaelides et al., 1997: 433) which follows the *spliced-severity* model (4.69); *mixed Exponential methodology* which models *ILF*s as weighted *Exponential cdf*s (utilised by *Insurance Service Office*, ISO, of Verisk Analytics (2017)); and various forms of *transformations* (e.g. *power curves* and *PH transforms*, §4.2.2.2).

In terms of 4.10, *CR* independence assumptions lead to the following for *limit factors*:

$$\gamma(b) = \frac{\mathrm{E}X^{(b)}}{\mathrm{E}X^{(a)}} \qquad 4.11$$

where X is the severity variable, and $a,b>0$ are given limits. From 4.8, this implies:

$$\begin{aligned} F(b) &= 1 - \frac{d\gamma(b)}{db}\mathrm{E}X^{(a)} = 1 - \gamma'(b)\,\mathrm{E}X^{(a)} \\ f(b) &= -\frac{d^2\gamma(b)}{db^2}\mathrm{E}X^{(a)} = -\gamma''(b)\,\mathrm{E}X^{(a)} \end{aligned} \qquad 4.12$$

where $\gamma(b)$, X, a, and b are defined as previously (4.11); and F, f represent the *cdf*, *pdf* for X respectively. Thus, given a scale of *limit factors*, 4.12 can be used as a basis for approximating the underlying severity *cdf* (demonstrated later, §5.3.4.1).

Given these relationships, the following properties, for *limit factors*, are apparent.

Properties 4.1 ***Consistent limit factors***

Limit factors, γ, for a given range of limits, are described as 'consistent' if they satisfy the following (notation based on 4.12):

1. Asymptotically constant: $\lim_{b\to\infty}\gamma'(b)=\lim_{b\to\infty}\frac{1-F(b)}{\mathrm{E}X^{(a)}}=0$; $\lim_{b\to\infty}\gamma(b)=c$ for constant $c>0$
2. Monotonically decreasing and non-negative gradient: $F(b)$ is monotonically increasing, therefore $\gamma'(b)=\frac{1-F(b)}{\mathrm{E}X^{(a)}}$ must be monotonically decreasing – any point of inflection in $F(b)$ will also correspond to an inflection point in $\gamma'(b)$ (the converse is also true); $F(b)\in[0,1]$; $EX^{(a)}>0 \Rightarrow \gamma'(b)\geq 0$
3. Concave down: $\gamma''(b)=-\frac{f(b)}{\mathrm{E}X^{(a)}}\leq 0$ – any mode in $f(b)=\frac{dF(b)}{db}$ will correspond to an inflection point in both $F(b)$ and $\gamma'(b)$, Miccolis (1978)

Hereafter, these are referred to as *consistency properties*. In terms of the first of these, it is acceptable for γ to remain constant above some finite limit, y, provided the probability that severity exceeds this is zero (i.e. $F(y)=1$). This point is considered later in the context of empirical *cdf*s (§5.3.1). When testing the second property, in respect of a given set of *limit factors*, $\gamma_1,\gamma_2,\ldots$, corresponding to ordered (positive) limits, $k_1<k_2<,\ldots$, first-order derivatives can be approximated using the well-known divided difference (Milne-Thomson, 2000, chap. 1):

$$\left.\frac{d\gamma_x}{dk_x}\right|_{x=u}\approx\frac{\gamma_{u+1}-\gamma_u}{k_{u+1}-k_u}\quad,u=2,3,\ldots \qquad 4.13$$

4.2.2.1 *Mean Excess*

The *ME* function (also known as *mean residual life function*), which is closely related to the *LEV* (4.7 with $k=1$) is a widely used tool with applications concerning *EVT* in the study of Actuarial Science, Environmental Science, Hydrology, and several other fields. It can be used to signal the potential distribution of the underlying data, for instance, *Paretianity* (Cirillo, 2013), as considered in §5.2.1.

The distribution of the excess over a *threshold* $b > 0$, $F_{(b)}$, in respect of a random variable X with distribution F and density f, is defined by:

$$F_{(b)}(x) = \Pr(X - b \le x \mid X > b), \qquad 4.14$$

(Ghosh & Resnick, 2010: 1492), and the corresponding *ME* function, $\mathrm{e}(b)$, is defined by:

$$\mathrm{e}(b) = E[X - b \mid X > b] = \frac{\mathrm{E}X - \mathrm{E}X^{(b)}}{1 - F(b)} \qquad 4.15$$

where $X^{(b)}$ is as before (4.6), (List & Lohner, 1998: 310; Klugman, Panjer & Willmot, 2004: 29). This clearly defines the relationship between the *LEV* (4.7, $k = 1$) and the *ME* function. The empirical analogue to 4.15 is given by:

$$\hat{\mathrm{e}}(b) = \frac{\sum_{i=1}^{n} \max(x_i - b, 0)}{\sum_{i=1}^{n} 1_{\{x_i > b\}}} \qquad 4.16$$

where $1_{\{x_i > b\}}$ is the indicator defined as previously (§1.3) and $x_1, \ldots, x_n$ are n observed severities. The exponential *cdf* has a horizontal *ME*, as illustrated in Figure 4.2.

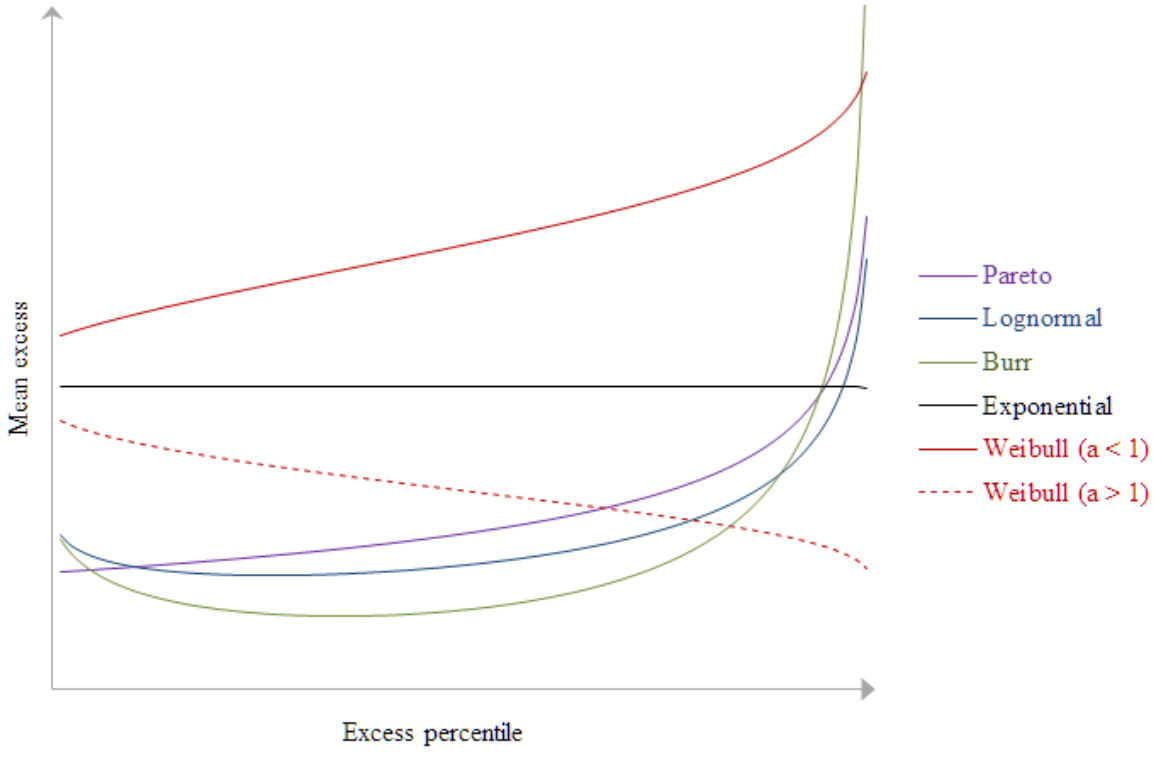


Figure 4.2 ME plots Increasing *ME*s for *heavy-tailed cdf*s: *Pareto*, *lognormal*, *Burr*, and *Weibull* (shape parameter: a < 1); horizontal for *exponential cdf*; and decreasing for *light-tailed Weibull cdf* (i.e. a >1).

As can be seen in Figure 4.2, *heavy-tailed cdf*s (formerly defined later, §4.2.3.3) have increasing *ME*s (e.g. *Pareto*, *lognormal*, heavy-tailed *Weibull*). *Concave-up* patterns in the case of *Burr* and *lognormal cdf*s (depending on parameters) can also be seen in this figure; whilst the *Weibull cdf* (shape parameter greater than one) and other such light-tailed *cdf*s have decreasing *ME*s. Refer to Cirillo (2013) for strengths and weaknesses of *ME* plots when used as a tool to identify Pareto *cdf*s, and Ghosh & Resnick (2010) for further theoretical and practical considerations.

4.2.2.2 *Adjustments and transformations*

This section describes limit factor adjustments in regard to risk, inflation, and deductibles.

Risk adjustment

As suggested by Feldblum (1993: 1), actuaries have proposed several methods for determining risk loads to compensate insurers for the level of risk they accept when writing business (e.g. measures based on the loss distribution, *utility theory*, and modern portfolio theory). In the context of stochastic loss modelling, there are three main sources of risk:

- *Process risk*: the inherent variability associated with the stochastic nature of frequency and severity of losses
- *Parameter risk*: the uncertainty in estimating the expected loss due to, for example, the occurrence of catastrophes, inflation, changes in the volume and mix of business (for a line of insurance), inadequate data (Freifelder 1976, cited by Miccolis (1978: 41)), errors in estimating parameters for frequency and severity distributions (Miccolis, 1978, n. 12) or the application of knowledge that is based on incomplete information (Allaben et al., 2008: 12)
- *Model risk*: the use of an imperfect model, or one that fails to accurately represent the situation (Allaben et al., 2008: 12) due to, for instance, modelling errors that produce inaccurate outputs or incorrect usage of the model (Aggarwal et al., 2015: 233)

Parameter risk considered in the context of *mixture models* (§4.2.5) and an algorithm is used to simulate *model error*. The current section concerns *process risk* (in particular, its quantification). In the context of *limit factors*, the mean *LAS* (4.10) will fail to reflect *process risk* otherwise described by the *ALD*. As such, higher order moments (e.g. standard

deviation and variance) may be required. Alternatively, transformation such as the *Proportional Hazard* (PH) or *power curve* can be used; these are considered shortly.

Example 4.1 ***Variance principle risk adjustment***

The *variance principle* has the following desirable properties: it satisfies basic ratemaking axioms, as set out by Freifelder (1979: 520), has theoretical backing (Bühlmann, 1985), and enables direct use of the severity distribution (assuming independence with frequency, losses in different layers, etc.). According to Feldblum (1993: 167), however, whilst variance (or standard deviation) may be mathematically tractable there is often no '*a priori*' reason for equating risk to such measures. Further, *consistency* (Properties 4.1) may not be preserved (Wang, 1995, sec. 10).

Formerly, let $\pi_{\text{var}}(S;w)$ be the *variance-adjusted* (pure-risk) premium in respect of the aggregate loss amount, S, and a risk parameter, $w>0$, be defined by:

$$\pi_{\text{var}}(S;w)=\text{E}S+w\text{Var}S\,, \tag{4.17}$$

then the variance-adjusted *limit factor*, γ_S, can be defined as:

$$\gamma_S(b;a,w)=\frac{\pi_{\text{var}}(S(b);w)}{\pi_{\text{var}}(S(a);w)}\,, \tag{4.18}$$

where $S(a)$ and $S(b)$ are *LAS*s (4.9) with limits $a,b>0$ respectively. Now suppose the underlying loss count variable is N and respective *limited severity* variables are $Y^{(a)}$ and $Y^{(b)}$. *Independence assumptions* (4.2) concerning loss count and *i.i.d.* severity (i.e. N, Y respectively), in conjunction with a *Poisson cdf* for N, collapse the risk-adjusted *limit factor*, γ_S (4.18), to the following:

$$\begin{aligned}\gamma_Y(b;a,w)&=\frac{\text{E}N\text{E}Y^{(b)}+w\text{E}N\text{E}Y^{(b)2}}{\text{E}N\text{E}Y^{(a)}+w\text{E}N\text{E}Y^{(a)2}}\\&=\frac{\pi_{\text{var}}(Y^{(b)};w)+w(\text{E}Y^{(b)})^2}{\pi_{\text{var}}(Y^{(a)};w)+w(\text{E}Y^{(a)})^2}\\&=\frac{\pi^*_{\text{var}}(Y^{(b)};w)}{\pi^*_{\text{var}}(Y^{(a)};w)}\,,\qquad \pi^*_{\text{var}}(Y^{(b)};w)=\pi_{\text{var}}(Y^{(b)};w)+w(\text{E}Y^{(b)})^2\end{aligned} \tag{4.19}$$

In addition to Properties 4.1, Wang (1995, sec. 10) described desirable qualities for risk-adjusted *ILF*s that include the following:

- Relative to the mean, risk loadings should increase with the size of the limit
- The same price should be produced regardless of how layers of cover are subdivided

In terms of the *variance principle*, the first of these should be satisfied in most practical circumstances, however, it is possible to violate the second by dividing cover into sufficiently many layers (however, the present research generally considers upper-limits for ground-up losses, as opposed to layers of insurance).

Example 4.2 ***Proportional-Hazard (PH) transform***

The *PH transform* is a member of a class of functions that preserves *stochastic dominance* and exhibits *comonotonic additivity*, and is defined as the mapping: $S_Y(x) \coloneqq S_X(x)^r$ where S_X and S_Y are survival functions in respect of the severity random variable X and the transformed variable, Y, respectively, and $0 < r \leq 1$ is a given risk parameter.

Let π_{PH} be the mean in respect of the *PH transform* defined by:

$$\pi_{PH}(Y^{(b)}; b, w) = \int_0^b S_Y(x)^{\frac{1}{w}} dx, \qquad 4.20$$

where b is a given non-negative limit, and, and $w \geq 1$ (Wang, 1995: 44, 1999b: 943).

This transform satisfies *limit factor consistency* properties since $\pi_{PH}'(b) = S_Y(b)^{\frac{1}{w}}$ is a monotonically increasing function (i.e. with decreasing marginal rate of increase). One approach to modelling aggregate loss distributions, described by Wang (1999b: 955), is to apply the transform to the associated severity and frequency *cdf*s (as demonstrated later, §5.3.3). Further, application of this transform to certain *cdf*s (e.g. *Pareto*, *Weibull*, *Burr*) results in the same type of *cdf*, with altered parameters; in other cases (e.g. lognormal, Poisson) numerical integration or analysis techniques are required. Refer to Wang (1995: 45) for examples of the former and a description of the properties of this type of transform.

The *power transform* is another type of transformation that several insurers such as *ACE* (Cresenzi & Alibrio, 2016) and *NLF* (Selleck, 2015) have utilised for cyber-liability *ILF*s.

Example 4.3 ***Riebesell curves (power transform)***

Power curves are commonly used in the London Market for insurance and reinsurance excess of loss pricing, and are also known as *power* or *alpha curves*, the *German method*, or *power curves*, named after their founder, Riebesell, as described by Mack & Fackler (2003: 231).

The following rule is assumed: $\gamma(2^k a, a) = (1+r)^k$, where γ is the *ILF* in respect of an *increased limit* and *base limit*, in this case, $2^k a$ and a respectively, with $a > 0$, $r \in (0,1)$, and $k > 1$. Substituting $b = 2^k a$ yields $k = \log_2(\frac{b}{a})$, and, therefore, one has the following for the *power curve limit factor*:

$$\gamma(b; a, w) = (1+r)^{\log_2(ba^{-1})} = (ba^{-1})^{\log_2(1+r)} = (ba^{-1})^w \qquad 4.21$$

where $w = \log_2(1+r)$. Refer to Halliwell (2013) for further detail on *power curves* (and *exponential transforms*); for an evaluation of the performance of several principles (e.g. *PH*, square root, logarithmic, quadratic, etc.), see Wang (1996: 85). In contrast to the *variance principle*, *power curves* and *PH transforms* are scale invariant (i.e. *limit factors* are unaffected by scale transforms such as currency and inflation adjustments). In terms of inflation invariance, this can lead to inconsistent results when compared to *experience-based* calculations that consider the experience of the risk in question.

Inflation adjustment

Constant inflation trend in underlying loss-severity can have a disproportionate effect on the *LAS* at higher limits. In the case of excess of loss covers, this is commonly referred to as a *leveraging* effect of inflation. The effect of such inflation, in terms of *limit factors*, is now considered. Let Y and X be two random variables with *cdf*s F_Y and F_X respectively, where $Y = vX$ for constant $v > 1$. From 4.7 ($k = 1$), the *LEV* for Y with limit $b > 0$ is:

$$E[Y^{(b)}; v] = \int_0^b S_Y(y)dy = v\int_0^{\frac{b}{v}} S_X(x)dx = vEX^{(\frac{b}{v})}, \qquad 4.22$$

through substitution, $x = yv^{-1}$.

As such, *limit factors* γ_X and γ_Y, corresponding to severity variables X and $Y = vX$ respectively, are related as follows:

$$\gamma_Y(b;v) = \gamma_X(\tfrac{b}{v}), \tag{4.23}$$

where b and v are as before (4.22). Variations of inflation adjustments include application of two inflation trends: the first, in relation to the *LEV* at the *basic limit*; the second, in relation to the average severity in excess of that limit.

Related methods exist for updating *ILF*s to reflect inflation, for instance, applying one trend function to average severity for *basic* limits and a separate function to the average severity for layers in excess of the *basic* limit, as described by Miccolis (1978).

Thus far, risk and inflation adjustments have been considered in isolation; their combined effect is now considered.

Proposition 4.1 ***Limit factors: inflation and variance principle risk adjustments***

From 4.7 $(k = 2)$, the second-order moment for limited variable Y, and limit b, is:

$$\begin{aligned} E[(Y^{(b)})^2; v] &= \int_0^b y^2 dF_Y(y) + b^2(1 - F_Y(b)) \\ &= \int_0^{\frac{b}{v}} (xv)^2 dF_Y(xv) + b^2(1 - F_X(\tfrac{b}{v})) \\ &= v^2\left[\int_0^{\frac{b}{v}} x^2 dF_X(x) + \left(\tfrac{b}{v}\right)^2 (1 - F_X(\tfrac{b}{v}))\right] \\ &= v^2 E[(X^{(\frac{b}{v})})^2], \end{aligned} \tag{4.24}$$

where F_X and F_Y denote *cdf*s for $Y = vX$ and X respectively ($v > 1$, as for 4.22). Here, the third equality follows from the previous since $dF_Y(y) = dF_X(\frac{y}{v})$. For a *CR* model with *independence assumptions* (4.2), and *Poisson* loss count, the variance-adjusted *limit factor*, γ_Y (4.19), with parameter w (as before), becomes:

$$\gamma_Y(b;a,w,v) = \frac{\pi^*_{\text{var}}(Y^{(b)};w)}{\pi^*_{\text{var}}(Y^{(a)};w)} = \frac{\pi^*_{\text{var}}(X^{(\frac{b}{v})};vw)}{\pi^*_{\text{var}}(X^{(\frac{a}{v})};vw)} = \gamma_X(\tfrac{b}{v};\tfrac{a}{v},vw) \tag{4.25}$$

The following is an extension of 4.25 that recognises deductibles:

Example 4.4 ***Excess losses with inflation and variance principle risk adjustments***

If $Y = \max(0, vX^{(\frac{b}{v})} - d)$ for non-negative variables X, Y; and constants $v > 1$, d, and b, s.t. $0 \le d < b$; then $Y = \max(0, (vX)^{(b)} - d)$, and Y can be expressed as follows:

$$Y = \begin{cases} 0 & vX < d \\ vX - d & d \le vX \le b \\ b - d & vX > b \end{cases} \qquad 4.26$$

The first two (raw) moments of Y are:

$$\begin{aligned} E[Y; b, d, v] &= v(EX^{(\frac{b}{v})} - EX^{(\frac{d}{v})}) \\ E[Y^2; b, d, v] &= v^2(EX^{(\frac{b}{v})2} - EX^{(\frac{d}{v})2} - \frac{2d}{v}(EX^{(\frac{b}{v})} - EX^{(\frac{d}{v})})), \end{aligned} \qquad 4.27$$

(Klugman, Panjer & Willmot, 2004: 127). For a *compound-Poisson* 'excess' *LAS*, $S = \sum_{i=1}^{N} Y_i$, where $Y_i = \max(0, vX_i^{(\frac{b}{v})} - d)$, $i = 1, 2, ..., N$ (i.e. $N \sim Poisson$), under *CR independence assumptions* (4.2), limits $a, b > 0$; and deductible d, s.t. $0 \le d < \min(a, b)$, the variance-adjusted *limit factor*, γ_Y (i.e. 4.25, based on 4.19), with parameter w (as before), becomes:

$$\gamma_Y(b; a, d, w, v) = \frac{\pi^*_{\text{var}}(X^{(\frac{b}{v})}, vw) - \pi^*_{\text{var}}(X^{(\frac{d}{v})}, vw) - 2dw(EX^{(\frac{b}{v})} - EX^{(\frac{d}{v})})}{\pi^*_{\text{var}}(X^{(\frac{a}{v})}, vw) - \pi^*_{\text{var}}(X^{(\frac{d}{v})}, vw) - 2dw(EX^{(\frac{a}{v})} - EX^{(\frac{d}{v})})} \qquad 4.28$$

where π^*_{var} is defined as previously.

4.2.3 Modelling severity

This section is relevant for spliced-severity *cdf*s, model selection, and tail behaviour (§4.3).

4.2.3.1 *Composite models*

The composite model considered here is an *m-component spliced* density, a convex combination of other densities with given weights.

Definition 4.1 ***m-component spliced density***

An *m-component spliced* density, h, given m densities, h_i, $i=1,...,m$, is defined as:

$$h(x)=\begin{Bmatrix} p_1h_1(x) & b_1<x<b_2 \\ p_2h_2(x) & b_2<x<b_3 \\ ... & ... \\ p_mh_m(x) & b_m<x<b_{m+1} \end{Bmatrix}, \qquad 4.29$$

where h_i is valid on (b_i,b_{i+1}), $i=1,..,m$, and $p_i\in(0,1)$ s.t. $\sum_{i=1}^{m}p_i=1$.

Precise model specification is required to ensure h (4.29) is continuous (Klugman, Panjer & Willmot, 2004, sec. 4.4.7); the same goes for differentiability (e.g. at $b_2,...,b_m$).

Several proponents of the *Maximum Likelihood* (ML) technique, for parameterising two component *spliced-densities*, have incorporated such restrictions to reduce the number of unknown parameters and associated equations. *ML* and other such techniques are considered later as part of a threshold determination exercise for a spliced-severity model (§4.3). The k^{th} order moment for a variable, Y, with density h, can be expressed in terms of the k^{th} order moments of the variables $X_1,\ldots,X_m$ (provided these exist), if each X_i has density $h_i\ \forall\ i=1,...,m$ as follows:

$$\mathrm{E}Y^k=\sum_{i=1}^{m}p_i\mathrm{E}X_i^k \qquad 4.30$$

where p_is are defined as previously. Applications of 4.30 range from mixed compound *Poisson* (Halliwell, 2009), §5.3.3, to spliced-severity *cdf*s.

4.2.3.2 *Model selection*

Information Criteria (IC)

Akaike Information Criterion (AIC) is a form of penalised likelihood criteria for model selection, measured as follows (Akaike, 1998):

$$AIC=2k-2l^* \qquad 4.31$$

where $k \in \mathbb{Z}^+$ is the number of parameter estimates, and l^* is the maximised log-likelihood.

The model with the lowest *AIC* is preferred; however, this is asymptotically valid in relation to sample size (Burnham & Anderson, 2002: 353). For small samples (in relation to k), a 'second-order' bias correction term may be introduced, which gives rise to the *corrected Akaike information criterion*, AIC^C, defined by:

$$AIC^c = AIC + \frac{2k(k+1)}{n-k-1}; n \neq k+1 \qquad 4.32$$

where n is the sample size.

According to Motulsky & Christopoulos (2004, chap. 23), the probability of making an incorrect selection, in respect of two models with an AIC^C difference of d, is $(\exp(0.5d)+1)^{-1}$.

Bayesian Information Criteria (BIC) is another measure that is based on a different underlying perspective to the *AIC*:

$$BIC = k\ln(n) - 2l^* \qquad 4.33$$

where k, n, and l^* are defined as previously. Key points relating to these criteria include:

- In contrast to the *BIC*, *AIC* is based on the Kullback & Leibler (1951) distance between two models (AIC^C simply enforces greater parsimony in terms of number of parameters)
- *AIC* does not assume that the true model is one of the candidates (Burnham & Anderson, 2002: 211–212)
- Whereas *AIC* aims for parsimonious selection, *BIC* (asymptotically) aims to determine the dimension of the true model
- In either case, *AIC*s and *BIC*s cannot be compared for models of different data

Such criteria depend heavily on relative measures for a given set of candidate models. As such, the absolute 'quality' of a selected model depends heavily on the suitability of candidate models.

In particular, 4.31–4.33 only reflect *goodness of fit* through likelihood term (l^*); as such, measure based on the Kolmogorov-Smirnov and Anderson-Darling tests are now introduced.

Kolmogorov-Smirnov test

Let F_n be the empirical distribution for n (*i.i.d.*) observations $x_1,\ldots,x_n$ defined by:

$$F_n(x)=\frac{1}{n}\sum_{i=1}^{n}1_{\{x_i\le x\}} \qquad 4.34$$

where indicator, $1_{\{x_i<x\}}$, is defined as previously (§1.3). For some continuous distribution, G, the *Kolmogorov-Smirnov* (KS) statistic, d_n, is defined as:

$$d_n=\sup_x\left|F_n(x)-G(x)\right|. \qquad 4.35$$

Now consider the following hypotheses:

$$H_0:F_n=G,\quad H_1:F_n\neq G \qquad 4.36$$

- $H_0:\ \lim_{n\to\infty}(d_n)=0$ *almost surely* (*Glivenko-Cantelli* – van der Vaart (1998: 266)), rejected at the $\alpha\in(0,1)$ level in favour of H_1 if $\sqrt{n}d_n>\mathcal{K}^{-1}(1-\alpha)$, where $\mathcal{K}$ represents the *Kolmogorov distribution*
- Critical values $k(\alpha)=\mathcal{K}^{-1}(1-\alpha)$ for this two-sided test can be determined using:

$$k(\alpha)=\left(-\frac{1}{2}\ln(\frac{\alpha}{2})\right)^{\frac{1}{2}} \qquad 4.37$$

- To allow for discontinuities (i.e. *jumps*) in the empirical *cdf*, F_n (4.34), for n ordered observations, $x_1\le x_2\le,\ldots,\le x_n$, d_n (4.35) can be derived as:

$$d_n=\max_i\left[\max\left(|G(x_i)-F_n(x_i)|,|G(x_i)-F_n(x_{i-1})|\right)\right] \qquad 4.38$$

where G is defined as previously in 4.35–4.36 (Klugman, Panjer & Willmot, 2004, sec. 13.4.1). As this relies on the comparison of the test statistic, $\sqrt{n}d_n$, and the critical value,

$k(\alpha)$ (given significance, α), the ratio of these terms represents a compact metric for testing 4.36 and 'score-based' comparisons that account for *goodness of fit* (§4.3.1).

Proposition 4.2 *KS-ratio*

Using definitions for n (4.34), d_n (4.35 or 4.38), $k(\alpha)$ and α (4.37), the *KS-ratio*, $r(n,\alpha)$, is proposed as follows:

$$r(n,\alpha) = \frac{\sqrt{n}d_n}{k(\alpha)},\ \alpha \in (0,1);\ \ k(\alpha) > 0 \qquad 4.39$$

where $r(n,\alpha) > 1$ implies there is insufficient evidence to reject H_0 (4.36).

In a similar way, an alternative ratio can be defined in terms of the *Anderson-Darling* (AD) measure which is related to 4.35 but recognises differences (weighted, squared) between the empirical and proposed (i.e. model) *cdf*s.

Anderson-Darling test

The *AD* test statistic, A^2, in respect of model (F) and empirical (F_n) *cdf*s, is defined by:

$$\frac{A^2}{n} = -F_{k+1} + \sum_{j=0}^{k-1}\left(S_{n,j}{}^2 \ln\frac{S_j}{S_{j+1}} + F_{n,j+1}{}^2 \ln\frac{F_{j+2}}{F_{j+1}}\right) + S_{n,k}{}^2 \ln S_k \qquad 4.40$$

where $F_j = 1 - S_j = F(x_j)$, $F_{n,j} = 1 - S_{n,j} = F_n(x_j)$, and F_n is based on $n > 1$ observations that span the (ordered, unique, uncensored) set of $k+2 \le n$ observations in question, $x_0 < x_1 <, ..., < x_{k+1}$, (Klugman, Panjer & Willmot, 2004, sec. 13.4.2). The '*AD-ratio*' analogue to 4.39 can, therefore, be defined in terms A^2 (4.40) divided by the desired critical value and related hypotheses (4.36, $G := F$) can be tested as before (i.e. reject the null hypothesis, thus the proposed model, if this ratio is less than one).

Preference may be given to *AD* test over the *KS* in 'standard' applications that require greater emphasis to be placed on *goodness of tail fit* (as opposed to in the 'body' of the *cdf*). Later, a similar (but moderated) result, in terms of *goodness of fit*, is considered using an average score that incorporates, as one of its components, the *KS-ratio* (§4.3.2). This avoids shortcomings associated with *AD* critical values (e.g. specificity in relation to the

cdf being tested), and, as alluded to previously, provides greater control over fluctuations in the tail.

4.2.3.3 *Tail behaviour*

Tail behaviour can be studied in several ways, one of which involves the *ME* as described previously (4.14). Here, a *limiting ratio* introduces the concept of *relative tail weight*, followed by the *absolute* concept which defines a class of distributions with a particular property. Consider two *cdf*s, G and H, with respective *pdf*s, g and h: should G have a *heavier tail* than H, then the *limiting ratio*, λ, diverges to infinity as follows:

$$\lim_{x\to\infty}\lambda(x)=\lim_{x\to\infty}\frac{1-G(x)}{1-H(x)}=\lim_{x\to\infty}\frac{g(x)}{h(x)}=\infty \qquad 4.41$$

Now that *relative tail weight* has been considered, attention is turned to the absolute concept. A class of *sub-exponential* 'heavy-tail' distributions, $\mathcal{S}$, is considered followed by the subclass of *long-tailed cdf*s.

Definition 4.2 ***Heavy- and light- tailed cdfs (absolute context)***

For *cdf* F and associated *pdf* f, let $S(x)=1-F(x)$ be the survival (i.e. tail) function and $S^{*(n)}=1-F^{*(n)}$ be the tail function where $F^{*(n)}$ is the *n-fold convolution* of F (4.5), $n\in\mathbb{Z}^+$. The tail of F depends on the asymptotic properties of $\lambda^{(n)}(x)=\frac{S^{*(n)}(x)}{S(x)}$, for instance:

- $\lim_{x\to\infty}\lambda^{(n)}(x)=n\ \Rightarrow F$ is *heavy tailed* and belongs to $\mathcal{S}$; examples include *Weibull* (with shape parameter, $\alpha<1$), *Burr*, *Lognormal*, and *Pareto cdf*s
- $\lim_{x\to\infty}\lambda^{(n)}(x)=\infty\Rightarrow F$ is *light tailed* (e.g. *Binomial*, *Poisson*, and *Negative Binomial*), Panjer & Willmot, 1992, cited by Wang (1998: 29)

*Long-tailed cdf*s are a subclass of *heavy-tailed* distributions (i.e. if F is *long tailed* then it is also the case that $F\subset\mathcal{S}$, Definition 4.2).

Definition 4.3 ***Long-tailed distributions***

The *cdf* F is defined as *long tailed* if $\lim_{x\to\infty} S(x+y) = \lim_{x\to\infty} S(x)\ \forall\ y>0$. For m long-tailed *cdf*s, $F_1,\ldots,F_m$, the following asymptotic result holds:

$$\lim_{x\to\infty} \frac{(S_1 * S_2 * \ldots * S_m)(x)}{S_1(x)+\ldots+S_m(x)} \geq 1 \qquad 4.42$$

Now that the all preliminary theory for this chapter has been covered, attention is turned to models that describe the severity of loss.

4.2.4 Aggregate loss distributions and transforms

In terms of 4.1 (*IR, CR* models), determining the *ALD* is one of the classical problems in the realm of risk theory. As there is generally no closed-form solution alternative techniques are often required (Shevchenko, 2010, sec. 1).

In the context of an insurance portfolio, the *ALD* for the *IR* model becomes a convolution of *ALD*s in respect of the individual risks that comprise the portfolio (Vernic & Sundt, 2009: 5); generally, the exact distribution can only be obtained in this way (i.e. convolution), although De Pril recursion can be used provided the portfolio follows a certain set up (Tse, 2009: 86). It may also be possible to identify (or, with *FFT*, reconstruct) the density with the aid of transforms (e.g. *cf*; *mgf*, *pgf* – provided these exist) – (Kaas et al., 2008, sec. 2.1). Alternatively, the *ALD* might be approximated (e.g. Normal, translated gamma (ibid., sec. 2.5); compound Poisson, Klugman, Panjer & Willmot (2004, sec. 6.11.3)).

For *CR*, the *ALD* can be treated as a compound *cdf*, with primary (loss count) and secondary (severity) component *cdf*s (the heavier of which determines the shape of the *ALD* tail (ibid., secs 6.2–6.3)). A special case, for instance, is when the primary *cdf* is Poisson which is closed under convolution and, therefore, results in a (mixed) Poisson distributed *ALD* (ibid., 98) which facilitates recursion (e.g. Panjer) and *FFT* methods (e.g. as a means of approximating the *IR* model).

There are relative advantages and disadvantages associated with each of these methods (e.g. *FFT* can be quicker than Panjer recursions when modelling severities with high *per-loss*

limits, although the opposite may be true at lower limits (ibid., sec 6.10); exact computation can be laborious, etc.). The main methods used to determine *ALD*s in this chapter are those relating to *transforms* such as *FFT* (§4.2.4.2) and *mixture models* (§4.2.5); *MC* simulation is also utilised (§4.2.6). This allows *ALD*s based on different techniques to be compared and checked against one another and provides a means to deal with some of the following areas that come under scrutiny in Chapter 5:

- Determining *ALD*s in respect of correlated classes (*FFT*)
- Quantifying *process risk* for risk-adjusted *limit factors* (mixture models)
- Varying the *per-loss* limit for individual severity distributions that incorporate empirical and statistical losses (*MC simulation*)

Basic concepts concerning *transforms* (§4.2.4.1) are now covered; these 'tools' will be key for several subsequent algorithms and models in the present chapter.

4.2.4.1 *Characteristic functions and related transforms*

Transforms are defined here in univariate and multivariate settings, followed by two illustrative examples for independent and correlated risks.

Univariate and multivariate transforms

In a univariate setting, respective definitions for the *pgf*, *mgf*, and *cf*, relating to a non-negative random variable, X, are given by:

$$P_X[t] = \mathrm{E}t^X,\ M_X[t] = \mathrm{E}\exp(tX),\ C_X[t] = \mathrm{E}\exp(itx) \tag{4.43}$$

where $i \in \mathbb{C}$. This assumes the *pgf* and *mgf* exist (hereafter, this goes without saying); the *cf* (of a real valued argument), however, always exists. Useful properties and relationships include:

$$\begin{aligned} &P_X[\exp(it)] = M_X[it] = C_X[t];\ \ P_X[t] = M_X[\ln(t)] \\ &\mathrm{E}X^k = \frac{\partial^k P_X[1]}{\partial t^k} = \frac{\partial^k M_X[1]}{\partial t^k} = (-i)^k \frac{\partial^k C_X[0]}{\partial t^n},\ k \in \mathbb{Z}^+ \\ &P_X[1] = C_X[0] = M_X[0] = 1, \end{aligned} \tag{4.44}$$

assuming k^{th}-order derivatives exist, (Wang, 1998; Mildenhall, 2005; Shevchenko, 2010).

Further, P_X, M_X, and C_X uniquely characterise the *cdf* of X (i.e. random variables with the same *pgf*, *mgf*, or *cf* are identically distributed). For the multivariate case, with slight abuse of vector notation, respective versions of joint *pgf*, *mgf*, and *cf* of $\mathbf{X} = [X_1, \ldots, X_n]$, $n \in \mathbb{Z}^+$, are defined as:

$$\begin{aligned} P_{\mathbf{X}}[\mathbf{t}] &:= P_{X_1,\ldots,X_n}[t_1,\ldots,t_n] = \mathrm{E}\, t_1^{X_1} \ldots t_n^{X_n} \\ M_{\mathbf{X}}[\mathbf{t}] &:= M_{X_1,\ldots,X_n}[t_1,\ldots,t_n] = \mathrm{E}\exp(\mathbf{t}\cdot\mathbf{X}') \\ C_{\mathbf{X}}[\mathbf{t}] &:= C_{X_1,\ldots,X_n}[t_1,\ldots,t_n] = \mathrm{E}\exp(i\mathbf{t}\cdot\mathbf{X}'), \quad \mathbf{t} = [t_1,\ldots,t_n] \end{aligned} \tag{4.45}$$

Transforms for ALDs

Let P_S, M_S, and C_S be the respective *pgf*, *mgf*, and *cf* of the aggregate loss, $S = X_1 +, \ldots, + X_n$ (X_is and n, as for 4.45). If $X_i s$ are independently distributed, with *pgf*, *mgf*, and *cf* given by P_i, M_i, and C_i respectively, then:

$$P_S[t] = \prod_{i=1}^{n} P_i[t],\ M_S[t] = \prod_{i=1}^{n} M_i[t],\ C_S[t] = \prod_{i=1}^{n} C_i[t], \tag{4.46}$$

(Klugman, Panjer & Willmot, 2004, sec. 3.3). On the other hand, for correlated or independent $X_i s$, it can easily be shown that:

$$P_S[t] = P_{\mathbf{X}}[\mathbf{t}],\ M_S[t] = M_{\mathbf{X}}[\mathbf{t}],\ C_S[t] = C_{\mathbf{X}}[\mathbf{t}] \tag{4.47}$$

where $\mathbf{t} = [t, ..., t]$ is *n* dimensional (see, for instance, Wang (1998, sec. 4.3)). In the context of aggregate loss models, X_is could represent aggregate losses in respect of n independent (4.46) or correlated (4.47) risk portfolios. In either case, S would resemble an *IR* framework. For a *CR* model, where X_is are defined accordingly, with common *pgf*, *mgf*, and *cf* given by P_X, M_X, and C_X, respectively; and N, the loss count variable, has *pgf* P_N. Transforms for the aggregate loss are now given by:

$$P_S[t] = P_N[P_{X|N}[t]],\ M_S[t] = P_N[M_{X|N}[\mathrm{t}]],\ C_S[t] = P_N[C_{X|N}[t]] \tag{4.48}$$

where $P_{X|N}$ is the conditional *pgf* for X (i.e., given N).

Here, S is treated as a *compound cdf*, with respect to N, and $P_{X|N}$ represents the *pgf* of the *secondary severity cdf* (as mentioned in §4.2.3); the *compound Poisson cdf* is considered in Klugman, Panjer & Willmot (2004: 94). Examples for independent and correlated risks are now considered: the first is elementary; the second, which builds upon this, is somewhat more elaborate.

Example 4.5 ***Aggregation of independent random events***

Consider a random variable U with *pdf* $\Pr(U=u)=\{p:u=1; q=1-p:\text{elsewhere}\}$ for $p\in[0,1]$. The *mgf* of U is, therefore, $M_U[t]=q+pe^t$ (i.e. *Bernoulli* Table D.3, eqn. D.2 with $n=1$), and aggregation over $n>1$ *i.i.d.* of such variables relates to the *Binomial cdf* (i.e. parameters n, p).

Example 4.6 ***Aggregation of correlated random events***

This example revisits the correlation model, proposed by Böhme (2005) and Böhme & Kataria (2006), §2.2, where the individual losses are correlated with a latent variable, based on *Pearson correlation coefficient* (defined shortly, §4.2.5.1). For this , consider m random variable loss events, $U_1,\ldots,U_m$, and assume these are correlated with the random variable $R\sim Bin(1,p_1)$, s.t. U_is are conditionally *i.i.d.* with $U_i\,|\,(R=r)\sim Binomial(1,p_{1|r})$ $\forall\ i=1,...,m$; $p_{1|r}=\Pr(U_i=1|\,R=r)$; and joint *pgf*, $P_{\mathbf{U}}[\mathbf{t}\,|\,r]:=P_{U_1,...,U_m}[t_1,...,t_m\,|\,R=r]$. Then, from 4.47 with $X_i=U_i$ and $S=U_1+,\ldots,+U_m$,

$$\begin{aligned} P_{\mathbf{U}}[\mathbf{t};r] &= P_S[t;r] \\ &= E[t^{U_1}...\,t^{U_m}\,|\,R=r] \\ &= (p_{1|r}+q_{1|r}t)^m, \end{aligned} \qquad 4.49$$

which is the *pgf* for a $Binomial(m,p_{1|r})$ *cdf*. The unconditional *pgf* of S is given by:

$$\begin{aligned} P_S[t] &= E_R E[t^{U_1+\ldots+U_m} \mid R=r] \\ &= E_R[(p_{1|R}+q_{1|R}t)^m] \\ &= p_1(p_{1|1}+q_{1|1}t)^m+(1-p_1)(p_{1|0}+q_{1|0}t)^m \\ &= p_1P_S[t\,|\,1]+(1-p_1)P_S[t\,|\,0]\,, \end{aligned} \qquad 4.50$$

and this resembles the weighted average of two *Binomial pgf*s: $P_S[t\,|\,1]$ and $P_S[t\,|\,0]$, with weights p_1 and $(1-p_1)$ respectively. This is consistent with the aggregate claim count *cdf* arrived at by Böhme (2005: 9) and Böhme & Kataria (2006: 17).

In terms of parameter estimation concerning 4.50 (i.e. $p_{1|1}$, $p_{1|0}$, p_1) Böhme & Kataria (2006) utilised *Expectation Maximisation*, EM (an iterative method to find the maximum likelihood of parameter estimates), and measured *parameter risk* using a *beta binomial* model (i.e. *mixture model*). Refer to Hisakado, Kitsukawa & Mori (2006, secs 2–3) for further detail regarding equations and solutions for $(p_{1|1},\ p_{1|0})$ in the context of correlated binomial distributions. Other types of *mixture models* are considered in §4.2.5. For other considerations and methods (e.g. Panjer recursion) concerning correlation in terms of aggregate loss, refer to Sundt (1999).

4.2.4.2 *Fourier transform*

The *Fourier* transform $\hat{f}$ of an integrable function f is a mapping $f:\mathbb{R}\to\mathbb{C}$ defined by:

$$\hat{f}(z)=\int_{-\infty}^{\infty}\exp(ixz)\,f(x)dx \qquad 4.51$$

where $z\in\mathbb{R}$ and $i\in\mathbb{C}$ s.t. $\sqrt{-1}=i$.

The original function f can be recovered using the *inverse Fourier transform* (i.e. of the *Fourier* transform) which can be represented as:

$$f(x)=\frac{1}{2\pi}\int_{-\infty}^{\infty}\exp(-izx)\,\hat{f}(z)dz,\ x\in\mathbb{R};i\in\mathbb{C}\,, \qquad 4.52$$

(Klugman, Panjer & Willmot, 2004, sec. 6.9.1). Several properties are now described.

Properties 4.2 ***Fourier transforms***

Let $f(x)$ and $g(y)$ be integrable functions defined on $x, y \in \mathbb{R}$ s.t. $x \perp y$, with (*Fourier*) transforms $\mathcal{T}[f] := \hat{f}$ and $\mathcal{T}[g] := \hat{g}$ respectively; the following can then be shown:

1. The transform $\hat{f}(0)$ in respect of function f is, according to 4.51, the following integral:

$$\hat{f}(0) = \int_{-\infty}^{\infty} f(x)dx \qquad 4.53$$

2. The transform of $f * g$ (i.e. convolution) is the product of their transforms (this follows from 4.46):

$$\mathcal{T}[f * g] = \mathcal{T}[f]\mathcal{T}[g] \qquad 4.54$$

3. When f is a *pdf*, its transform is the *cf* of f (*cf* in 4.43 is same as 4.51)

4.2.4.3 *Discrete Fourier Transform (DFT)*

Let f_x be a function defined on all discrete integers, such that f_x has a period of length n (i.e. $f_{x+n} = f_x$). The *Discrete Fourier Transform* (DFT) that applies to a vector $f^{(n)} = [f_0, f_1, \ldots, f_{n-1}]$ generates $\hat{f}_z$ defined as:

$$\hat{f}_z = \sum_{x=0}^{n-1} \exp\left(\frac{2\pi i z x}{n}\right) f_x, \quad z = 0,1,..,n-1 \qquad 4.55$$

The original functions in $f^{(n)}$ can then be recovered using the inverse of the *DFT* that is applied to the *DFT* of these functions, with the following result:

$$f_x = \sum_{x=0}^{n-1} \exp\left(\frac{-2\pi i z x}{n}\right) \hat{f}_z, \; x \in \mathbb{Z}, \qquad 4.56$$

(Klugman, Panjer & Willmot, 2004, sec. 6.9.1).

4.2.4.4 *Fast Fourier Transform (FFT)*

To generate n values of $\hat{f}_z$ (4.55), n vectors of f, each having n functions of the form f_x (4.56) are required. As such, the number of terms for evaluation is of the order n^2; the *FFT* algorithm reduces this to the order of $n\log_2 n$. To illustrate, it begins with a vector $f^{(n)} = [f_0, f_1, \ldots, f_{n-1}]$ of length $n = 2^r$, for some $r \in \mathbb{Z}^+$, and subdivides this into two subvectors of equal length ($m = \frac{n}{2}$), functions f_{2x} (even indices) are assigned to one sub vector whilst functions f_{2x+1} (odd indices) are assigned to the other, $x = 0,1,..,m-1$:

$$\begin{aligned}
\hat{f}_z &= \sum_{x=0}^{n-1} \exp\left(\frac{2\pi i z x}{n}\right) f_x \\
&= \sum_{x=0}^{m-1} \exp\left(\frac{2\pi i z (2x)}{n}\right) f_{2x} + \sum_{x=0}^{m-1} \exp\left(\frac{2\pi i z (2x+1)}{n}\right) f_{2x+1} \quad \ldots \quad m = \frac{n}{2} \\
&= \sum_{x=0}^{m-1} \exp\left(\frac{2\pi i z x}{m}\right) f_{2x} + \exp\left(\frac{\pi i z}{m}\right) \sum_{x=0}^{m-1} \exp\left(\frac{2\pi i z x}{m}\right) f_{2x+1} \\
&= \hat{f}_{z1} + \exp\left(\frac{\pi i z}{m}\right) \hat{f}_{z2} \ ,
\end{aligned} \tag{4.57}$$

(Klugman, Panjer & Willmot, 2004: 186). In turn, each of the transforms $\hat{f}_{z1}$ and $\hat{f}_{z2}$ are subdivided into two further subvectors, each of equal length, and so on until each vector is comprised of only one function. *This entire procedure is hereafter referred to as FFT.*

For n, m,... (i.e. the lengths of each subdivided vector of functions) to be integer valued, the original vector, $f^{(n)}$, must have n functions such that $n = 2^r$ (i.e. $r \in \mathbb{Z}^+$ is the number of bisections required). If there are fewer than 2^r functions the original vector $f^{(n)}$ can be *padded* with zeros to make up for the shortfall. Should f_x be continuous (i.e. $x \in \mathbb{R}$ as opposed to $x \in \mathbb{Z}$) or not *periodic*, then *discretisation*, the mathematical process by which continuous functions (or models, equations, etc.) are relayed to discrete counterparts, is required before *FFT* (4.57) can be applied. Appendix C.1 describes the *mass-dispersal discretisation* (*rounding* method), which is utilised in Chapter 5. Refer to Wang (1999a: 862) and Klugman, Panjer & Willmot (2004: 655–656) for further information pertaining to this and the *mean-preserving* method.

The following steps define a general purpose *FFT* algorithm in terms of 4.55–4.57. These form the basis of several key models (§4.4).

Algorithm 4.1 ***General FFT steps for reconstructing ALDs***

1. Perform discretisation (Appendix C.1) in respect of a given severity *cdf* to produce the vector $f^{(n)} = [f_0, \ldots, f_{n-1}]$ (where $n = 2^r$, $r \in \mathbb{Z}^+$)
2. Apply *FFT* to the vector $f^{(n)}$, and obtain $\hat{f}_z$, $z = 0,1,\ldots,n-1$ (as in 4.55)
3. Apply the *cf* (step 2) within the *pgf* of the loss count cdf, or raise to the power of the given number of risks (i.e. for *CR* and *IR ALD*s respectively, 4.2) – based on relevant application of 4.46–4.48
4. Apply the inverse *Fourier* transform to the *cf* (step 3), and obtain the *ALD* as a discretised pdf vector with dimension $n = 2^r$, as in 4.56

The *ALD* (step 4, Algorithm 4.1) can be '*undiscretised*' as necessary; further details in this regard can be found in Klugman, Panjer & Willmot (2004, sec. E.3).

4.2.5 Correlated *ALD*s

This section describes *cf*s for correlated aggregate loss and count, based on pioneering contributions by Wang (1998, 1999a) and conventional techniques for mixture models (Klugman, Panjer & Willmot, 2004; Mildenhall, 2005). To begin with, prerequisite definitions regarding correlation are first covered.

4.2.5.1 *Correlation and covariance coefficients*

Definition 4.4 ***Pearson's correlation coefficient***

Pearson's correlation coefficient between two random variables, X_i and X_j, with standard deviations σ_i, σ_j respectively, is ρ_{ij} defined by:

$$\rho_{ij} = \frac{\mu_{ij} - \mu_i \mu_j}{\sigma_i \sigma_j} \qquad 4.58$$

where $-1 \le \rho_{ij} \le 1$ and $\mu_{ij} = EX_iX_j$ (Pearson cited by Lawrence & Lin (1989)).

Definition 4.5 ***Covariance coefficient***

For random variables X_i and X_j, with Pearson correlation coefficient ρ_{ij}, means and standard deviations as before (4.58), the *covariance coefficient* κ_{ij} is given by:

$$\kappa_{ij} = \frac{\mathrm{Cov}(X_i, X_j)}{\mu_i \mu_j} = \frac{\rho_{ij}\sigma_i\sigma_j}{\mu_i\mu_j} \qquad 4.59$$

The range of κ_{ij} (4.59) depends on the shape of marginal distributions for X_i and X_j, as described later (§4.2.5.1) in terms of *tail behaviour*.

4.2.5.2 *Cfs for correlated aggregate losses*

Define the joint *cf*, $C_{\mathbf{S}} := C_{S_1,\ldots,S_m}$, for $m \in \mathbb{Z}^+$ random variables, $\mathbf{S} = [S_1, \ldots, S_m]$, by:

$$C_{\mathbf{S}}[\mathbf{t}] = \left(1 + \sum_{i<j} \kappa_{ij} \left(1 - C_i[t_i]\right)\left(1 - C_j[t_j]\right)\right) \prod_{i=1}^{m} C_i[t_i], \qquad 4.60$$

where $S_i, S_j \in \mathbf{S}$ have respective *cf*s C_i, C_j, and *covariance coefficient* κ_{ij}, $1 \le i < j \le m$; and $\mathbf{t} = [t_1, \ldots, t_m]$, Wang (1998, pt. IV).

The joint *pdf*, $f_{\mathbf{S}} := f_{S_1,\ldots,S_m}$, can be represented as follows:

$$f_{\mathbf{S}}(\mathbf{s}) = \left(1 + \sum_{i<j} \kappa_{ij} \left(1 - \frac{f_i^{*(2)}(s_i)}{f_i(s_i)}\right)\left(1 - \frac{f_j^{*(2)}(s_j)}{f_j(s_j)}\right)\right) \prod_{i=1}^{m} f_i(s_i), \qquad 4.61$$

where $f_i := f_{S_i}$ represents the marginal *pdf* of S_i, with *two-fold convolution* (4.5) $f_i^{*(2)} := f_{S_i}^{*(2)}$; and $i = 1, \ldots, m$. Key features of $C_{\mathbf{S}}$ (4.60), and representation of $f_{\mathbf{S}}$ (4.61), include:

- *Covariance coefficients* are incorporated by utilising the entire marginals (where these are given)
- For valid $f_{\mathbf{S}}$ (i.e. non-negative), κ_{ij}s must fall within a *permissible range*, which can be defined for *heavy-tailed* marginals (Definition 4.2) as *limiting ratios* ($\frac{f^{*(2)}}{f}$, 4.61)

that are bounded from above; whilst *light-tailed cdf*s lead to negative probabilities

- The density of $S_1+,\ldots,+S_m$ can be reconstructed from $C_{S_1+,\ldots,+S_m}$ using *FFT* (Algorithm 4.1) which, interestingly, may not necessarily be invalid in the case of *light-tailed* marginals

The univariate density, referred to in the final point, is now considered further.

Example 4.7 ***ALD for sum of correlated aggregate losses***

Following on 4.60, let the univariate *cf* of $S = S_1+,\ldots,+S_m$ be C_S; from 4.47 (with $\mathbf{S} = \mathbf{X}$), it follows that:

$$C_S[t] = \left(1 + \sum_{i<j} \kappa_{ij} \left(1 - C_i[t]\right)\left(1 - C_j[t]\right)\right) \prod_{k=1}^{m} C_k[t], \qquad 4.62$$

where κ_{ij}s and C_is are defined as previously. The mean and variance of aggregate loss, S, is then:

$$\begin{aligned} \mu &:= \mathrm{E}S = \mathrm{E}[S_1+,\ldots,+S_m], \\ \mathrm{Var}S &= \sigma^2 + 2\sum_{i<j} \kappa_{ij} \mathrm{E}S_i \mathrm{E}S_j \end{aligned} \qquad 4.63$$

where $\sigma^2 = \sum_{j=1}^{m} \mathrm{Var}S_j$, Wang (1998: 31). As alluded to previously, the univariate *cf* (4.62) is apparently less restrictive, in terms of *covariance coefficients* (for valid *pdf*), than is the case for the *joint cf* (4.60).

4.2.5.3 *Cfs for correlated loss count (mixture models)*

Often, the extent, existence, or number of insurance claims are influenced by common external loss generating mechanisms, for instance: a hurricane or motor-vehicle accident may result in multiple claims including bodily injury and property damage; claim costs may be affected by the common regulation or economic climate (e.g. inflation). In the context of stochastic modelling, this is a cause of uncertainty referred to previously as *parameter risk* (§4.2.2.2).

To reflect such uncertainty, a secondary *mixture cdf* can be incorporated within the model. In this section, a joint *pgf* for correlated aggregate loss count variables is built up using

Poisson mixtures. Refer to Klugman, Panjer & Willmot (2004, sec. 4.6.10) for examples of various other mixtures with theoretical underpinnings.

Poisson mixture models

Let $\mathbf{N}=[N_1,\ldots,N_m]$ be a vector of m discrete random variables with joint *pgf* given by $P_{\mathbf{N}} := P_{N_1,\ldots,N_n}$ (see 4.45 for similar notation) and assume there exists a random variable θ with *mgf* M_θ such that $(N_j \mid \theta=\omega) \sim Poisson(\lambda_j\omega)$ (i.e. Appendix D.1, D.3) where $EN_j(\theta=\omega)=\omega\lambda_j,\ j=1,...,m.$ The marginal *pgf* of $N_j \mid (\theta=\omega)$ is then $P_{N_j|\theta=\omega}[t_j]=e^{w\lambda_j(t_j-1)}$, which leads to the following joint *pgf* for $\mathbf{N}$:

$$P_{\mathbf{N}}[\mathbf{t}]=E_\theta E t_1^{N_1}\ldots t_m^{N_m} \mid \theta = E_\theta \exp(\theta\boldsymbol{\lambda}\cdot(\mathbf{t}'-\mathbf{1}_m'))=\mathrm{M}_\theta[\boldsymbol{\lambda}\cdot(\mathbf{t}'-\mathbf{1}_m')], \qquad 4.64$$

where $\boldsymbol{\lambda}=[\lambda_1,\ldots,\lambda_m]$, $\mathbf{t}=[t_1,\ldots,t_m]$, and $\mathbf{1}_m$ is a (row) vector with m ones.

Example 4.8 ***Gamma-mixed Poisson model***

Suppose $\theta \sim Gamma(\alpha,1)$, for some $a>0$, has *mgf* $M_\theta[t]=(1-t)^{-\alpha}$, then the joint *pgf* in 4.64 becomes $P_{\mathbf{N}}[\mathbf{t}]=(1-\boldsymbol{\lambda}\cdot(\mathbf{t}'-\mathbf{1}_m'))^{-\alpha}$. This specifies a form of *multivariate negative binomial cdf*, where marginals, $N_j \sim NB(\alpha,\lambda_j)$ (D.4), have respective *pgf*s, P_{N_j}, $j=1,...,m,$ defined by:

$$P_{N_j}[t_j]=(1-\lambda_j(t_j-1))^{-\alpha}, \qquad 4.65$$

(Wang, 1999a: 803). Refer to Mildenhall (2005: 120) for *mgf*s with alternative parameterisations, and Reshetar (2008) for practical application in the context of *OR* (Chapter 2).

Example 4.9 ***Multivariate Negative Binomial*** **(MNB)** ***distribution***

From Example 4.8, let $N_j \sim NB(a_j,\lambda_j)$ – the *joint pgf*, $P_{\mathbf{N}}$, is now:

$$P_{\mathbf{N}}[\mathbf{t}]=(\mathbf{1}_m\cdot\mathbf{k}'-m+1)^{-\frac{1}{w}}, \qquad 4.66$$

where $\mathbf{t} = [t_1, \ldots, t_m]$, $\mathbf{1}_m$ is a row vector of m ones; $\mathbf{k} = [k_1, \ldots, k_m]$ with $k_j = (1 - \lambda_j (t_j - 1))^{\alpha_j w}$, $j = 1, ..., m$; and $w \neq 0$.

This specifies a family of *MNB cdf*s, with marginals $N_j \sim NB(\alpha_j, \lambda_j)$, in either of the following cases:

1. $0 < w < \min_{j \in [1,m]} \{\alpha_j^{-1}\}$
2. $w < 0$ s.t. $\mathrm{P}_{\mathbf{N}}[\mathbf{0}_m] > 0$ and $-\frac{1}{w} \in \mathbb{Z}^+$

where $\mathbf{0}_m$ is a row vector of m zeros, (Wang, 1998: 47). Here, the random vector $\mathbf{N}$ follows an *MNB* distribution, denoted by $\mathbf{N} \sim \mathrm{MNB}(\boldsymbol{\alpha}, \boldsymbol{\lambda}, w)$ with vector parameters $\alpha = [\alpha_1, \ldots, \alpha_m]$ and $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_m]$. Suppose $S_1, \ldots, S_m$ represent $m \in \mathbb{Z}^+$ *CR* loss models (4.2) that are specified by their severities and loss count variables, (X_i, N_i), $i = 1, ..., m$, and only correlated through $\mathbf{N} = [N_1, \ldots, N_m] \sim \mathrm{MNB}(\boldsymbol{\alpha}, \boldsymbol{\lambda}, w)$ (Example 4.9).

Recall the relationship between aggregate loss *cf*, loss count *pgf*, and severity *cf* (4.48); accordingly, the *cf* for the overall aggregate loss, $C_S := C_{S_1 + \ldots + S_m}$, is defined by:

$$C_S[t] = (\mathbf{1}_m \cdot \mathbf{y}' - m + 1)^{-\frac{1}{w}}, \qquad 4.67$$

where $\mathbf{1}_m$ is a row vector of m ones, $\mathbf{y} = [y_1, \ldots, y_m]$ with $y_j = (1 - \lambda_j (C_j - 1))^{\alpha_j w}$, C_j is the *cf* of X_j $j = 1, ..., m$ (Meyers & Heckman, 1984: 36; Wang, 1998: 27). As such, *FFT* reconstructs the *cdf* of $S = S_1 + ... + S_m$, from transforms C_S (4.67). The mean and variance of S can be determined using 4.63 (substituting κ_{ij}s with w, the correlation parameter in 4.67).

4.2.6 Monte Carlo (MC) simulation

Monte Carlo simulation refers to a broad class of algorithms that repeatedly sample from a process, to assimilate results for a process that typically exhibits some form of variability.

To illustrate this, the *inverse probability transformation* (as considered previously for the purpose of *FFT* in Algorithm 4.1) and the *quantile function* are defined.

Definition 4.6 ***Probability integral transformation***

For random variable, X, with continuous distribution, F, the transformation, $U = F(X)$, yields random variable, $U \sim Uniform(0,1)$.

Definition 4.7 ***Inverse probability integral transformation***

The inverse for Definition 4.6: if $U \sim Uniform(0,1)$ and X has a distribution F, then random variable $F(U^{-1})$ has the same distribution as X. Thus, simulating $F^{-1}(U)$ is equivalent to simulating random variable X (Definition 4.7), however, it may be the case that F does not have a unique inverse (e.g. $F(b) = F(a)$ s.t. $a < b$, $F(a) > 0$, and $F(b) < 1$). In such instances, it is useful to define the inverse F^{-1} in terms of a non-decreasing *quantile function* Q as follows:

$$Q(u) = F^{-1}(u) = \inf\{x : F(x) \geq u\} \qquad 4.68$$

where $u \in (0,1)$, (Devroye, 1986: 28). In terms of a given time horizon and loss X with *cdf* F, the $\alpha \in (0,1)$ quantile, $Q(\alpha) \equiv VaR_{\alpha}(X)$, where VaR is the *Value at Risk* – that is, the value of loss s.t. the probability of a larger loss is less than $1-\alpha$.

Algorithm 4.2 ***Monte Carlo (MC) simulation of a random variable***

To simulate random variable X with distribution F, first simulate $U = u$ from $U \sim Uniform(0,1)$, then calculate $Q(u)$ using 4.68, (Wang, 1999a: 880). This represents one iteration of the *MC* simulation (increasing the number of iterations generally reduces associated *simulation error*).

4.3 Severity model

The *spliced* (severity) *model*, considered in this section, assumes individual losses are generated by processes that differ according to the severity of loss. In particular, define a two component *spliced model* in terms of n observed *severities*, ordered as

$x_1 < x_2 <, \ldots, < x_n$. Losses in the interval $[0, \tau]$, for a given non-negative *threshold*, τ (i.e. '*splicing point*'), are assumed to follow a *small loss cdf* (in this case, estimated by the empirical *cdf*, F_n). To cover the interval (τ, ∞), a parametric distribution, G, is estimated using (observed) losses greater than τ. Following on from 4.29 ($m = 2$), let H be the *spliced* distribution in question:

$$1 - H(x) = \begin{cases} 1 - F_n(x) & x \le \tau \\ \left(1 - F_n(\tau)\right)\left(1 - \dfrac{G(x) - G(\tau)}{1 - G(\tau)}\right) & x > \tau \end{cases}$$
$$= \begin{cases} 1 - F_n(x) & x \le \tau \\ \left(1 - F_n(\tau)\right)\left(\dfrac{1 - G(x)}{1 - G(\tau)}\right) & x > \tau \end{cases} \qquad 4.69$$

where the *first component cdf*, $\frac{F_n(x)}{F_n(\tau)}$ (for $x \le \tau$) and *second component cdf*, $\frac{G(x) - G(\tau)}{1 - G(\tau)}$ (for $x > \tau$), are spliced with weights $F_n(\tau)$ (i.e. p_1, 4.29) and $1 - F_n(\tau)$ ($p_m = p_2$, 4.29) respectively. There is no notation for the true underlying distributions, which are "*unknown and unknowable*" (Klugman, Panjer & Willmot, 2004: 421).

Approaches to identify a *model* and *threshold* (i.e. G and τ respectively, 4.69) are now described in further detail.

4.3.1 Selection (large-loss model)

The following steps are used to select a *large-loss cdf*, from a set of $k \in \mathbb{Z}^+$ candidate models (e.g. *Burr, Weibull, Pareto, etc.*) and identify a suitable *threshold* for application of the *spliced model* in 4.69 (i.e. given $\mathbf{x}_n = [x_1, \ldots, x_n]$):

Step 1 — Fit $m > 1$ *cdf*s, $G_{i1}, \ldots, G_{im}$, to the largest $n - i + 1$ *severities*, for some $i = 2, 3, \ldots, n - k - 1$, where $k \le n - 2$ is the minimum number parameter estimates for each *cdf* (e.g., based on *Maximum Likelihood Estimation*, MLE)

Step 2 — Let $G_i^* = \min_j \{c_j\}$, where c_j is the AIC^c for G_{ij}, $j = 1, \ldots, m$

Step 3 Calculate B_i^*, the *KS-ratio* (4.39) for G_i^*

Steps 1–3 have the following outputs: the *large-loss* distribution, G_i^*, empirical *threshold*, x_i, and *KS-ratio*, B_i^* (valid scores require $i = 2,...,n-k-1$, as in step 1). In terms of the *spliced model*, H (4.69), $G_i^*(x) = \frac{G(x)-G(\tau)}{1-G(\tau)}, x > \tau$ and $x_i \le \tau < x_{i+1}$ – if $\tau < x_2$ or $\tau > x_{n-k-1}$, then the unconditional *cdf*s, G and F_n respectively, might be used (in parallel to similar set-ups, such as Ralucavernic (2009: 86), where an *ML* approach is utilised).

The threshold itself can be expressed in terms of the empirical rank as follows:

$$j = nF_n(\tau) \qquad 4.70$$

where $j = 1,...,n$; F_n, τ, and x_1 are defined as previously (4.69).

4.3.2 Threshold determination

Threshold determination is a common challenge when dealing with *spliced models* such as H (4.69), and several techniques exist in this regard. In the statistical branch of *EVT*, for instance, these include graphical (e.g. *ME* 4.2.2.1, *Hill* plot) and analytical methods (e.g. *ML*, Scollnik & Sun, (2012); square error; etc.). These and other methods include *EM* algorithms (Reynkens et al., 2016); techniques pertaining to *GPD* models (Gharib et al., 2017); and *tail-fit* optimisation (Buch-Kromann, 2009).

However, elements of subjectivity are introduced (e.g. choice of weights, w_is in 4.72, as described shortly); and there may be loss of predictive power, with results that are highly dependent on data (Michaelides et al., 1997).

The present chapter adopts a score based approach (Klugman, Panjer & Willmot, 2004, sec. 13.5.3), using a similar set-up adopted for *ML* approaches, but with greater emphasis being placed on *tail fit* and *limit-factor consistency*. Differentiability and continuity requirements, (Cerchiara & Acri, 2016: 3), are not explicitly allowed for, however, model selection incorporates the *Kullback-Leibler* distance estimate (AIC^C, §4.2.3.2). This provides a practical and simplified means to identify both *parametric cdf* and *threshold* (additional considerations pertain to *limit-factor consistency* and *ME plots*, Chapter 5).

Criteria 4.1 ***Splicing point***

Ordinarily, criteria for determining a threshold in the context of *EVT* depict a variance-bias trade-off associated with *GPD* parameter estimates. As there is no such 'bias' in the present case, an artificial index is created using the underlying empirical *cdf*, which is combined with the *KS-ratio* (§4.2.3.2). An alternative approach would be to use the *Anderson-Darling goodness-of-fit* measure (Klugman, Panjer & Willmot, 2004, sec. 13.4.2) – as this places greater emphasis on *tail fit* (i.e. at larger values). However, to illustrate concepts such as weighted scores and balancing trade-offs, the following criteria are contemplated for determining threshold, τ, in terms of output from steps 1–3 (§4.3.1).

1. τ, with the greatest rank, i
2. τ, with the lowest *KS-ratio,* B_i^*

In this way, larger thresholds are favoured through the first criterion, whilst the second attempts to optimise *tail fit*. As described in §5.2.1, upper bounds are established for thresholds by considering *ME plots*. Recall τ was defined as being equivalent to x_i, $i = 2,...,n-k-1$, as before (step 1, §4.3.1) – this is relevant for the following section.

Normalising scores

According to Criteria 4.1, preference is given to higher and lower values of x_is and B_i^*s respectively (i.e. steps 1–3, §4.3.1). Equivalently, higher values of α_i and β_i, defined as follows, are favoured over lower values:

$$\alpha_i = \frac{x_i}{x_{n-k-1}}, \quad \beta_i = \frac{\min_{i\in[1,n]}\{B_i^*\}}{B_i^*} \;\forall\; B_i^* > 0 \qquad 4.71$$

With i defined as previously (step 1, §4.3.1); thus $\alpha_i, \beta_i \in (0,1]$ are on the same scale.

Combining scores

The concept of combining different measures or estimates is not uncommon. For instance, in insurance practice, Actuaries may determine premiums as the credibility weighted average of exposure-based and experience-based estimates (Boor, 1997: 2; Werner &

Modlin, 2010, chap. 12). Some of the many other applications include constructing *cubic splines* where competing objectives relating to *smoothness* (measured using second derivatives) and *goodness of fit* (based on the sum of least squares) are combined, (Klugman, Panjer & Willmot, 2004: 485).

The weighted average score, z_i, with respect to measures α_i and β_i (4.71), is determined as follows:

$$z_i = w_i\alpha_i + (1-w_i)\beta_i, \qquad 4.72$$

where $w_i \in (0,1)$ represents the weight associated with α_i, $i = 2,3,...,n-k-1$. To determine an optimal splicing point, z_i (4.72) can be maximised over $i = 2,3,...$ given $\mathbf{x} = [x_2,...,x_{n-k-1}]$, B_i^*, and associated weights (i.e. $w_2, w_3, ...$). The algorithm outlined shortly utilises *steps 1–4* (§4.2.4.4), transformations (§4.2.2.2), and *tail-fit* scores (§4.2.3.2).

The choice of weight, w_i, $i = 2,...,n$, for α_i (or equivalently, $1-w_i$ for β_i) in 4.72 is indeed a subjective one. Three options are considered here, the first of which, $w_i^{(1)}$, represents a naïve approach that assumes equal weights for α_i and β_i:

$$w_i^{(1)} = 50\% \qquad 4.73$$

This is the most straightforward option, however, as the *threshold*, τ (or equivalently, i) increases, the reliability of α_i reduces (since fewer observed severities are used to parameterise the *cdf*, G_{i*}^*, Algorithm 4.3, upon which α_i is based). Therefore, $w_i^{(2)}$, which reduces as τ increases, can be defined as follows:

$$w_i^{(2)} = \frac{n-i}{n} \qquad 4.74$$

Likewise, as τ reduces $w_i^{(2)}$ increases, which also appears to be acceptable if this implies G_{i*}^*, and thus α_i, is more reliable due to parameterisation in respect of a larger number of observed severities. However, the suitability of G_{i*}^* itself depends on the suitability of candidate severity *cdf*s considered (i.e. $G_{i1}, G_{i2}, ...$, §4.3.1 step 1), which is independent of

the threshold. As described earlier (§4.2.3.2), 'absolute' quality of a model relies on that of the candidate models. If unsuitable *cdf*s are considered in the first instance then β_i should reflect this, however, using $w_i^{(2)}$ will mask this at lower thresholds. The third and final weighting option, $w_i^{(3)}$, attempts to address this potential issue by forcing the ratio $w_i^{(2)}:(1-w_i^{(2)})$ to remain constant across all $i=2,...,n$:

$$w_i^{(3)} = \frac{n-i}{2n-i} \qquad 4.75$$

This weight results from the division of $1+w_i^{(2)}$ into $w_i^{(2)}$. As is the case for $w_i^{(2)}$, $w_i^{(3)}$ reduces as τ increases, however, it allows greater weight to be placed on β_i at lower τ. Table 4.1 illustrates $w_i^{(1)}$, $w_i^{(2)}$, and $w_i^{(3)}$ for a sample of low to high percentiles. In terms of the various options for weights depicted in this table:

- Option 1 appears to be inferior to Options 2–3, for reasons already provided
- Option 3 is unnecessary here, due to a variety of suitable candidate *cdf*s considered in Chapter 5 (Appendix D.1), where the sensitivity of these options, in terms of Algorithm 4.3 outputs, is also considered

Option 2 is, therefore, selected for use in Chapter 5.

	Option 1		Option 2		Option 3	
Percentile $\left(\frac{i}{n}\right)$	$w_i^{(1)}$	$1-w_i^{(1)}$	$w_i^{(2)}$	$1-w_i^{(2)}$	$w_i^{(3)}$	$1-w_i^{(3)}$
0.25%	50.0%	50.0%	99.8%	0.3%	49.9%	50.1%
65.5%	50.0%	50.0%	34.5%	65.5%	25.7%	74.3%
66.5%	50.0%	50.0%	33.5%	66.5%	25.1%	74.9%
80.5%	50.0%	50.0%	32.5%	67.5%	24.5%	75.5%
81.5%	50.0%	50.0%	31.5%	68.5%	24.0%	76.0%
82.5%	50.0%	50.0%	30.5%	69.5%	23.4%	76.6%
97.5%	50.0%	50.0%	2.5%	97.5%	2.4%	97.6%
98.5%	50.0%	50.0%	1.5%	98.5%	1.5%	98.5%
99.5%	50.0%	50.0%	0.5%	99.5%	0.5%	99.5%

Table 4.1 Scale of weights for scores Considered for determining the weighted average where α_i and β_i are transformed AIC^c and *KS-ratio* measures respectively for different percentiles, $\frac{i}{n}$, in respect of n severities, $i=1,2,...,n$. Option 1: 50%; 2–3: weights ($w_i^{(2)}=\frac{n-i}{n}$ and $w_i^{(3)}=\frac{n-i}{2n-i}$ respectively) that reduce as the percentile increases.

Algorithm 4.3 ***Optimal threshold and large-loss cdf***

For a given group (i.e. class) of n ordered, homogeneous, and independent severities, $x_1,\ldots,x_n$, with empirical *cdf* F_n (4.34); *steps 1- 3* (p. 4.35) are run for each $i \in [2,n]$ to produce the following input vectors for this algorithm:

- $\mathbf{G} = [G_2^*, G_3^*, \ldots, G_{n-k-1}^*]$ (i.e. selected *large-loss* distributions from *step 3*)
- $\mathbf{x} = [x_2^*, \ldots, x_{n-k-1}^*]$ (i.e. vector of 'thresholds')
- $\mathbf{B} = [B_2^*, B_3^*, \ldots, B_{n-k-1}^*]$ (i.e. associated vector of *KS-ratios*)

Next, 4.71 is applied to $\mathbf{x}$ and $\mathbf{B}$ (element by element) to obtain the vector of scores $\boldsymbol{\alpha} = [\alpha_2, \ldots, \alpha_n]$ and $\boldsymbol{\beta} = [\beta_2, \ldots, \beta_n]$ respectively.

For a given vector of weights $\mathbf{w} = [w_2, \ldots, w_n]$, where $w_i \in (0,1)\ \forall\ i = 2,3,...,n$, the vector of (calculated) weighted scores, $\mathbf{z} = [z_2, \ldots, z_n]$, is determined using 4.72. The *optimal threshold*, τ^*, is x_{i^*}, where $i^* \in \{2,3,...,n\}$ is the *optimal index* value that yields the solution to the following:

$$z_{i^*} = \max\{z_i : i = 2,3,...,n\} \qquad 4.76$$

The corresponding (parameterised) *optimal distribution* is then $G^*_{nF_n(\tau^*)} = G^*_{i^*}$ (which follows from 4.70 with $j := i^*$). Thus, the outputs of this algorithm are the *optimal threshold*, *optimal index value*, and *optimal distribution* (i.e. τ^*, i^*, and $G^*_{i^*}$ respectively).

Algorithm 4.4 ***Model confidence sets – Kullback-Leibler***

This algorithm follows the bootstrap approach of Burnham & Anderson 2002 (sec. 4.5), which is based on essential Kullback & Leibler (1951) theory associated with *AIC* and other such information criteria. For each candidate *cdf* (i.e. parametric family), G_i, and bootstrap sample indexed $i = 1,...,m$ and $j = 1,...,M$ respectively, m, $M > 2$, determine *Akaike* differences, δ_{ij}, in relation to the minimum AIC^C, $A_j^* = \min_{i=1,..,m}\{A_{ij}\}$, and associated *Akaike weights*, w_{ij} (that sum to one for each sample) as follows:

$$\delta_{ij} = A_{ij} - A_j^* \qquad w_{ij} = \frac{\exp(-0.5\delta_{ij})}{\sum_{i=1}^{m}\exp(-0.5\delta_{ij})} \qquad 4.77$$

where A_{ij} is the AIC^C score for *cdf* G_i, parameterised (e.g. using *MLE*) in respect of data for sample $j \in \{1,...,M\}$.

Differences and weights accompanying the M samples can provide insight into model (in this case, *cdf*) selection uncertainty. For instance, in terms of the following 'model confidence set' and selection probability estimates:

- The α100% 'Kullback-Leibler' (*KL*) confidence set, for specified *cdf* with (common) index $s \in \{1,...,m\}$, comprises the set of candidate *cdf*s with corresponding *Akaike* differences below the α100% empirical quantile, $q^{(\alpha)}$, of *Akaike* differences for the specified *cdf*; the probability that *cdf* indexed $i = 1,...,m$ is in such a confidence set, $c_i^{(\alpha)}$, can be estimated from the samples as follows: $\hat{c}_i^{(\alpha)} = M^{-1}\sum_{j=1}^{M} 1_{\{A_{ij}-A_j^* \le q^{(\alpha)}\}}$ (indicator, $1_{\{.\}}$, defined as previously in §1.3)
- Correspondence between the average weight, $\hat{w}_i = M^{-1}\sum_{j=1}^{M} w_{ij}$, for a given *cdf* with index $i = 1,...,m$, and the proportion of (M) minimum Akaike scores that correspond to the *cdf* in question, $\hat{\pi}_i = M^{-1}\sum_{j=1}^{M} 1_{\{\delta_{ij}=0\}}$, attests to the veracity of the (aforementioned) *KL* confidence set, and associated model inference uncertainty

4.4 Limit factor and aggregate loss models

This section describes and formulates various *limit factor* and aggregate loss models, which are grouped in Figure 4.3 according to whether correlation (between aggregate losses for *classes A–E*, Chapter 3) is recognised, and how loss count, N, is modelled:

- *IR* framework: $N = n$ is given
- *CR* framework: N is a random variable with a given *pdf*

In this way, *IR* represents a special type of *CR*, where N has a degenerate distribution such that $\Pr(N = n) = 1$, as contemplated by Klugman, Panjer & Willmot (2004, sec. 6.1).

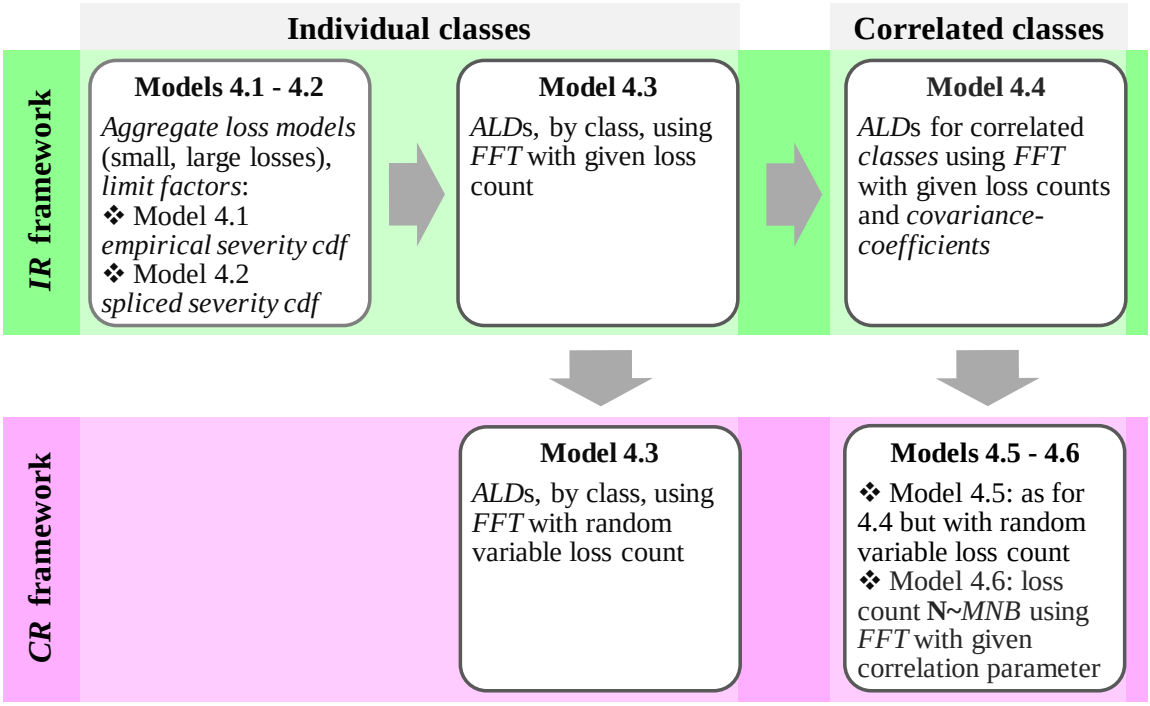


Figure 4.3 Flow chart for Models 4.1–4.6 Models 4.4–4.5 and Model 4.6 assume correlated aggregate loss amounts and counts (*classes A–D*) respectively. Adjustments (e.g. inflation, risk) may apply to limit factors based on any of these models.

Models 4.1–4.2 Limit factors for independent, individual classes (IR model)

The following is an overview of Models 4.1–4.6, as depicted in this figure

- Models 4.1– 4.2 model aggregate losses in respect of *small* and *large* severities, using empirical *cdf*s and the *spliced-severity* model (§4.3); relevant *limited moments* (4.7) are used to determine the risk-adjusted *LAS* (4.17) and *limit factors* (4.18) in an *IR* framework with consideration for possible application in a *CR* framework
- Model 4.3 (*IR* and *CR*) derives *ALD*s in respect of *classes A–D* (subject to *per-loss* limits), and *class E* (subject to a *per-occurrence* limit) from which *limit factors* are determined in respect of *ground-up* or *excess* losses; inflation and risk adjustments (4.28)
- Models 4.4–4.5 rely on given *covariance coefficients* between aggregate losses in *classes A–D* (4.62)
- Model 4.6 applies 4.67 with relevant parameters for the (correlated) marginal loss count *cdf*s (*NB*, Table D.3, eqn. D.4)

Models 4.1–4.2 are formerly defined in this section; Models 4.3–4.6 are more descriptive in nature and are framed in the context of tailored *FFT* steps, with compound-Poisson and negative binomial applications for Models 4.5–4.6.

Assumptions for ILFs

The method used to determine *ILF*s in this section is based on 'top slicing' and relies on the following assumptions:

- Severities, by class, are homogenous, independent, and independent of loss count
- Non-risk elements (e.g. expenses) are negligible
- There is no anti-selection (e.g. by size of limit)

Variables and definitions

Define the following for a given class with n observed severities:

- F_n and τ : empirical *cdf* and *splicing point* respectively
- $x_1 \leq,\ldots,\leq x_u \leq \tau$: the smallest, ordered, u (*i.i.d.*) severities with *LAS*, *LEV*, and '*limited*' *variance* denoted by $Z_S(b) = \sum_{i=1}^{u} x_i^{(b)}$, $\mu_{S;b} = \mathrm{E}X_S^{(b)} = \frac{1}{u}\sum_{i=1}^{u} x_i^{(b)}$, and $\sigma_{S;b}^2 = \mathrm{Var}X_S^{(b)} = \frac{1}{u}\sum_{i=1}^{u}(x_i^{(b)} - \mu_{S;b})^2$ respectively, where $b > 0$ is a *single limit* that applies to severity (§4.2.2); $u = nF_n(\tau) \in \{0,1,...,n\}$; $X_S \in \{x_1,...,x_u\}$ is the *small* severity random variable where $x_i \overset{d}{\sim} X_S$, $i = 1,...,u$, and $X_S \sim F_n$
- $X_1,\ldots,X_{n-u}$: $n-u$ random variable '*large'* severities with *LAS*, *LEV*, and *limited variance* $Z_L(b) = \sum_{i=1}^{u} X_i^{(b)}$, $\mu_{L;b} = \mathrm{E}X_L^{(b)}$, and $\sigma_{L;b}^2 = \mathrm{Var}X_L^{(b)}$ respectively, where X_is are *i.i.d.* such that $X_i \overset{d}{\sim} X_L$, $i = 1,...,n-u$; $X_L \sim G$, where X_L and G are the *large* severity random variable and *cdf* (unconditional with respect to τ), respectively; $X_L \perp X_S$; b is the limit as before

Thus, $\mu_{L;b} = \int_0^b S_X(x)dx$ and $\sigma_{L;b}^2 = 2\int_0^b xS_X(x)dx - {\mu_{L;b}}^2$, which follows from 4.30 with $k=1$ and $k=2$ respectively, and $S_X = 1 - F_X$ where $F_X(x) = \frac{G(x)-G(\tau)}{1-G(\tau)}$, $x > \tau$ $(F_X(x) = 0,\ x \leq \tau)$. The overall aggregate loss, Z , its mean, μ_Z , variance, σ_Z^2 , and

associated (*variance principle*) risk-adjusted *LAS*, $\pi_Z := \pi_{\text{var}}$ (4.17, $S = Z$), and *limit factor*, $\gamma_Z := \gamma_S$ (4.18, $S = Z$), are defined by Models 4.1–4.2, in an *IR* framework, as follows:

$$Z(b) = Z_S(b) + Z_L(b) = \sum_{i=1}^{u} x_i^{(b)} + \sum_{i=1}^{n-u} X_i^{(b)}$$

$$\text{E}Z(b) = \mu_{Z;b} = u\mu_{S;b} + (n-u)\mu_{L;b} \;; \quad \text{Var}Z(b) = \sigma^2_{Z;b} = u\sigma^2_{S;b} + (n-u)\sigma^2_{L;b} \qquad 4.78$$

$$\pi_{Z;b} = \mu_{Z;b} + w\sigma^2_{Z;b} \;; \qquad \gamma_{Z;a,b} = \frac{\pi_{Z;b}}{\pi_{Z;a}}$$

where $a, b > 0$; (u, n, b, Z_S , Z_L, $\mu_{S;b}$, $\mu_{L;b}$, $\sigma^2_{L;b}$) as before; and $\text{Cov}X_S X_L = 0$.

Models 4.1–4.2 can now be distinguished from one another as follows:

- Model 4.1 – by setting $u = n$ (or equivalently, $\tau \geq x_n$, the maximum observed severity), s X_L and associated terms in 4.78 become redundant and Z , μ_Z , σ^2_Z, . π_Z, . and γ_Z are expressed solely in terms of x_i , $i = 1,...,n$) and calculated numerically
- Model 4.2 – this relies on the *spliced-severity* model (and associated algorithms) developed in §4.3, by setting τ , u , and G to the optimal outputs from Algorithm 4.3 (i.e. *threshold* τ^* , *index* i^* , and *large-loss cdf* , $G^*_{i^*}$ respectively); analytical solutions, developed in respect of large-loss *limited moments*, are checked using *Model Risk* by Vose (2019), risk analysis software and simulation

*ILF*s and associated measures for Models 4.1–4.2 can then be determined for a range of different splicing points and associated (small and large) severity *cdf*s.

CR modifications – Model 4.2

Limited moments, risk-adjusted *LAS*s, and *limit factors* (4.78) can be modified for *CR* applications. For instance, suppose N is the random variable loss count with mean n , and all other relevant assumptions underlying Models 4.1–4.2 remain unaltered; then $\mu_{Z;b}$ does not change; if $N \sim Poisson(n)$, then replacing $\sigma^2_{S;b}$ and $\sigma^2_{L;b}$, in the expression for $\sigma^2_{Z;b}$ (4.78), with $\text{E}X_S^{(b)2}$ and $\text{E}X_L^{(b)2}$ respectively, yields the *CR* equivalent for the variance of

$Z(b)$ (associated *limit factors* follow suit). Attention is now turned to Models 4.3–4.6, which rely on *FFT* (Algorithm 4.1), with steps summarised in Table 4.2.

- *Steps 1–2:* limited severity *cdf*s that reflect *per-loss* (classes *A–D*) and *per-occurrence* (*E*) limits (§4.2.2) are *discretised* for application of standard (Excel) *FFT* routine (Chapter 5), based on specified *spans*, *ranges*, and *limits*
- *Step 3a*: varies according to whether the model belongs to the *IR* or *CR framework* (as defined earlier); *Step 3b* combines *cf*s in respect of *ALD*s with correlated aggregate severity (i.e. Models 4.4–4.5) or count (i.e. Model 4.6)
- *Step 4*: reconstructs the density in question by applying inverse *Fourier transform* (4.56) to respective *cf*s from the previous step

Model:	Steps 1 - 2	Step 3a	Step 3b	Step 4
	1) *Discretise* (limited, *spliced*) severity *cdf*s in respect of *classes A-E*; 2) Apply *FFT* (element by element) to obtain their *cf*s	*Cf*s (*step 2*): raise to power of *n* (i.e. given loss count), or apply within *pgf* of *N* (i.e. random variable *loss count*) to obtain aggregate loss *cf*s	Combine *cf*s (*step 3a*) to obtain overall aggregate loss *cf* using given *covariance coefficients* or *Multi-NegBin model*	Reconstruct *ALD*(s) from *cf*(s) in penultimate step (i.e. *step 3a* or *step 3b*), using *inverse FFT*
Model 4.3 (*IR*)	✓	Raise *cf*s to power of *n*	x	*Inverse FFT* (*cf*s: *step 3a*)
Model 4.3 (*CR*)	✓	Apply *cf*s in *pgf* of *N*	x	*Inverse FFT* (*cf*s: *step 3a*)
Model 4.4	✓	Raise *cf*s to power of *n*	Combine using *cov coeff*	*Inverse FFT* (*cf*: *step 3b*)
Model 4.5	✓	Apply *cf*s in *pgf* of *N*	Combine using *cov coeff*	*Inverse FFT* (*cf*: *step 3b*)
Model 4.6	✓	Apply *cf*s in NegBin *pgf*s	Combine with *MNB*	*Inverse FFT* (*cf*: *step 3b*)

Table 4.2 FFT steps for ALDs (Models 4.3–4.6) Check mark (✓) if step is relevant, cross (x) otherwise. Common font colour (i.e. red or blue) for common procedures within a step; '*cov coeff*'- given covariance coefficient parameters.

As mentioned, *covariance coefficient* parameters (Step 3b, Models 4.4–4.5) are investigated and formulated as part of a sensitivity analysis (Chapter 5).

Model 4.3 ALD for independent classes (IR, CR models)

Of the Models 4.3–4.6, this model represents the most straightforward application of *FFT* (Algorithm 4.1). In terms of *steps 1- 4* in that algorithm, consider a class with n observed severities. Model 4.3 (*IR*) proceeds with *step 1* by discretising the *spliced-severity*

distribution (of limited severities) using the *rounding method* (Appendix C.1). The corresponding vector of *cf*s (determined in *step 2*) are raised to the power of n (element by element) to obtain *cf*s in respect of *ALD*s (*step 3a*), which are yielded using the *inverse Fourier transform* (*step 4*) – *undiscretisation* of these *ALD*s is unnecessary for the intended purpose, and, therefore, not performed. Model 4.3 (*CR*) is very similar except, instead of raising severity *cf*s to the power of n (*step 3a*), the *pgf* of an assumed loss count *cdf* (in this case, *Poisson*) is incorporated (*steps 1*, *2*, and *4* remain otherwise unchanged). A simulation algorithm (presented shortly) is used to verify Model 4.3. Models 4.4–4.6 are distinguished from one another in terms of *steps 3–4* (Table 4.2), as is now described.

Model 4.4 ALD for correlated aggregate losses (IR model)

Step 3a is relevant for Model 4.4 as this is based on the *IR* framework which assumes each class has a (deterministic) loss count, n. The *cf* for each of the *classes A–D* is thus raised to the power of n (element by element) to obtain corresponding (class-level) *cf*s in respect of their marginal *ALD*s. *Step 3b* combines these using 4.62 (with $m = 4$, and assumed *covariance coefficients*, $\kappa_{ij} = \kappa$), before taking the *inverse Fourier transform* (4.56) in *step 4* to yield the aggregate loss *cdf* (i.e. joint *cdf* for correlated marginal *ALD*s with respect to *classes A–D*).

Model 4.5 ALD for correlated aggregate losses (CR model)

Model 4.5 is the *CR* analogue to Model 4.4. Instead of raising the *cf*s in each of the classes *A–D* to the power of a deterministic count parameter, n, as is the case for Model 4.4 in *step 3a*, the *pgf* of an assumed *loss count* variable is incorporated within the *cf* (element by element). Following on from 4.48, this yields the *cf*s in respect of the (marginal) *ALD*s for each of the *classes A–D*. *Step 3b* (i.e. application of 4.62 with given marginals and *covariance coefficients*) and *step 4* (i.e. *inverse Fourier transform*) used in this model are otherwise identical to those used for Model 4.4. For *variance principle* adjustments regarding *limit factors*, 4.63 is utilised later (§5.3.3).

Model 4.6 ALD for correlated loss count (CR model)

Model 4.6 utilises a (multivariate) mixture model, as considered for Example 4.9. In particular, *step 3a* assumes that the class has random variable *loss count*, N_j, with

$NB(a_j, \lambda_j)$ *cdf* and specified parameters a_j, λ_j, $j = 1,2,3,4$ (Table D.3, D.4). The associated *pgf* is thus incorporated (element by element) within *cf*s in *step 2* to produce (class-level) vectors of *cf*s (*step 3a*) for respective *ALD*s. These are then combined using 4.67 (with $m = 4$, and assumed correlation parameter, w) in *step 3b*, before using the *inverse Fourier transform* (4.56) to yield the aggregate loss *cdf* in *step 4* (i.e. joint *cdf* in respect of classes with correlated aggregate *NB* loss count variables). Relationships between Models 4.5–4.6, with respect to *LAS* moments, are considered later (Chapter 5).

Applications for Models 4.5–4.6

In terms of Model 4.5, aggregate loss, S (4.63) with constant covariance, $\kappa_{ij} = \kappa_r \ \forall\, i < j$, is assumed later for some scenario $r = 1,2,3$ ($\perp i, j$) – to this end, let $C_V := 2\sum_{i<j} \mathrm{E}S_i \mathrm{E}S_j$. The variance-adjusted *LAS*, π_r (4.17) for scenario r, with *covariance-coefficient*, κ_r, and (common) risk-adjustment parameter, $w \geq 0$, can then be expressed as:

$$\pi_r = \mu + w(\sigma^2 + \kappa_r C_v) = \pi_1 + w\kappa_r C_v \qquad 4.79$$

For compound-Poisson S_i (4.63), having (primary) *Poisson* parameter, $\lambda > 0$, and (secondary) survival function, S_{X_i}, in respect of severity variable X_i, $i = 1,...,m$, it can be shown that $S = S_1 + ,...,+ S_m$, in the independent case (i.e. $\kappa_r = 0$), is also *compound-Poisson*, having primary (Poisson) parameter $m\lambda$ and secondary (mixed) survival function $m^{-1}\sum_{i=1}^{m} S_{X_i}$. Refer to Klugman, Panjer & Willmot (2004: 99) for a general case proof. In this way C_V, μ, and σ^2 can be related to the moments of a random variable Y with (mixed) survival $S_Y = m^{-1}\sum_{i=1}^{m} S_{X_i}$, as follows:

$$\begin{aligned} C_V \lesssim \tilde{C}_V &= m(m-1)(\lambda \mathrm{E}Y^{(b)})^2, \quad \frac{d\tilde{C}_V}{db} = 2m(m-1)\lambda^2 S_Y(b)\mathrm{E}Y^{(b)} \\ \mu &= m\lambda \mathrm{E}Y^{(b)}, \quad \frac{d\mu}{db} = m\lambda S_Y(b) \\ \sigma^2 &= m\lambda \mathrm{E}Y^{(b)2}, \quad \frac{d\sigma^2}{db} = 2mb\lambda S_Y(b) \end{aligned} \qquad 4.80$$

where $b > 0$ is a given limit; m and λ as before; and $\tilde{C}_V$ denoting an upper bound (or, when the variance between $\mathrm{E}X_1^{(b)},...,\mathrm{E}X_m^{(b)}$ is small, an approximation) for C_V. In fact, $\tilde{C}_V - C_V$ is directly proportional to the variance between $\mathrm{E}X_1^{(b)},...,\mathrm{E}X_m^{(b)}$ (i.e. akin to 'variance in hypothetical means'), and thus, at sufficiently low limits, $C_V \lesssim \tilde{C}_V$ (with equality when $\mathrm{E}X_i^{(b)} = b \;\forall\; i = 1,...,m$). To see this, define $x_i := \mathrm{E}X_i^{(b)}$, $i = 1,...,m$; then $\mathrm{E}Y^{(b)} = \frac{1}{m}\sum_{i=1}^{m} x_i$, $\frac{\tilde{C}_V}{\lambda^2} = \frac{m-1}{m}(\sum_{i=1}^{m} x_i)^2$, and $\frac{C_V}{\lambda^2} = \sum_{i \neq j} x_i x_j = (\sum_{i=1}^{m} x_i)^2 - \sum_{i=1}^{m} x_i^2$. The difference, $\frac{1}{\lambda^2}(\tilde{C}_V - C_V)$, after rearranging, yields $\tilde{C}_V - C_V \propto \sum_{i=1}^{m} x_i^2 - \frac{1}{m}(\sum_{i=1}^{m} x_i)^2$ (i.e. $\mathrm{Var}\{\mathrm{E}X_1^{(b)},\ldots,\mathrm{E}X_m^{(b)}\}$, *q.e.d.*). The extent to which this difference, as a percentage of C_V (i.e. $\frac{\tilde{C}_V - C_V}{C_V}$), increases with the size of the limit, b, depends on the nature of the underlying severity *cdf*s (i.e. $1 - S_{X_1}(b), 1 - S_{X_2}(b), ...$).

For Model 4.6, let $S^* = S_1^* + \ldots + S_m^*$ denote the *LAS* for this model, where S_i^* s have compound *cdf*s with (common) marginal $NB(\lambda, c)$ loss count variables and secondary severity *cdf*s identical to those of Model 4.5. Then the mean and variance of S^* can be expressed in terms of μ, σ^2, and C_V (defined for S in 4.80) as follows:

$$\mathrm{E}S^* = \frac{\lambda^*}{\lambda}\mu\,; \qquad \begin{aligned} \mathrm{Var}S^* &= \sum_{i=1}^{m} \mathrm{Var}S_i^* + \kappa^* C_V \frac{\lambda^{*2}}{\lambda^2} \\ &= \frac{\lambda^*}{\lambda}\sigma^2 + \lambda^* c \sum_{i=1}^{m} (\mathrm{E}X_i^{(b)})^2 + \kappa^* C_V \frac{\lambda^{*2}}{\lambda^2} \\ &= \sigma^2 + c\frac{\mu^2 - C_V}{\lambda} + \kappa^* C_V \quad \text{.... for } \lambda = \lambda^* \\ &= \sigma^2 + \frac{\mu^2 c}{\lambda} - C_V \frac{c - \kappa^* \lambda}{\lambda} \end{aligned} \qquad 4.81$$

The final line can also be written as $\mathrm{Var}S + \lambda c \sum_{i=1}^{m} (\mathrm{E}X_i^{(b)})^2 \geq \mathrm{Var}S$, where $\mathrm{Var}S$ represents the variance for S in Model 4.5 (i.e. $\sigma^2 + \kappa C_v$) with covariance coefficient $\kappa = \kappa^*$ (in which case Model 4.6 would have a larger risk-adjusted *LAS* than Model 4.5 with this covariance coefficient). The algorithm used to simulate *LAS cdf*s, to verify compound *Poisson ALD*s based on Model 4.3, is now described.

Algorithm 4.5 ***Monte Carlo simulation of class-level ALDs (CR framework)***

Let S have a compound-Poisson *cdf* with primary *Poisson* loss-count (and constant parameter) and secondary spliced-severity *cdf* (given q splicing percentile; and large- and small- loss *cdf*s):

1. Realise $N = n$ from the primary *cdf*
2. Realise $N_L = n_L$ severities from the secondary *cdf* (large-loss) *cdf*, where N_L denotes large-loss count with a conditional Binomial *cdf*, parameters (n, q)
3. Realise $n - n_L$ small severities from the secondary (small-loss) *cdf*, and sum

Summing large (2) and small (3) severities and repeating 1–3 will simulate S (Homer & Rosengarten, 2011). Refer to Sundt (1999, sec. 2) for a generalised set-up for various applications.

Model choice

The choice of model depends on several factors (e.g. suitability for desired purpose, validity of underlying assumptions, etc.). As summarised in Table 4.3, Models 4.1–4.2 provide an effective way to calculate *limit factors* over a wide range of limits using risk analysis software or analytical solutions (where these exist) to derive underlying *limited moments*. Models 4.3–4.6 allow for determination of the entire *ALD*. However, if the purpose is to derive *limit factors*, then an analytical approach, provided solutions exist, would be far more effective since application of *FFT* would otherwise be required for each and every (single) limit (the same goes for simulation based approaches, 4.12).

As for Models 4.1–4.2, analytical complexity depends on factors such as severity (e.g. *spliced*) and loss count (e.g. *Poisson*) *cdf*s, and risk-adjustment method (e.g. *variance principle*, *PH*, etc.). The usual quirks associated with *FFT* (e.g. *wrap around*) extend to Models 4.3–4.6. The choice between Models 4.4–4.6 depends on respective loss count and correlation assumptions. For instance, Models 4.4–4.5 allow for correlation between class-level (i.e. marginal) *LAS*s by utilising given covariance coefficients and marginal *ALD*s with respect to different classes, whilst Model 4.6 utilises a mixed-model in relation to correlated aggregate count variables.

Model	Advantage	Disadvantages
4.1 - 4.2	1.1(a) Efficient method for calculating *limit factors* or the exact mean *LAS* over a wide range of (single) policy limits	1.1(b) Based on expected value which disregards other features of the *ALD* (e.g. skewness, kurtosis, etc.)
	1.2(a) *ILF*s can be expressed in terms *LEV*s alone, which simplifies calculations	1.2(b) Assumes frequency and severity are independent
	1.3(a) Easily extended to other applications such as testing *ILF* consistency for different splicing points and performing simulations	1.3(b) Requires simulation or other extension to cater for the *CR* framework, non-independence, correlation, or compound limits
4.3 - 4.6	2.1(a) Makes use of *FFT*, an efficient algorithm for approximating the *ALD*	2.1(b) Needs to be run for each *per-loss/occurrence* limit
	2.2(a) Quick determination of pure-risk premium for different aggregate limits	2.2(b) Requires specification of a suitable span and truncation point
	2.3(a) *FFT* routines can be implemented using widely available software	2.3(b) Some frequency *pgf*s may be difficult/impossible to explicitly formulate (e.g. negative hypergeometric)
	2.4(a) Flexible: easily modified to model *IR* or *CR* frameworks, univariate or multivariate severity and frequency, and various forms of correlations	2.4(b) Potential for distortions and inaccuracies (e.g. discretisation error, wrap-around and over/under-flow issues)
	2.5(a) Strong theoretical backing with consideration of characteristic, moment and probability generating functions	

Table 4.3 Advantages and disadvantages of different models

Model 4.5 (Table 4.3) provides greater flexibility than Model 4.6 in that the latter requires *NB* loss count *cdf*s and a common covariance-coefficient, whilst the former allows for different cdfs and coefficients in this regard. Whereas both models are utilised, in Chapter 5, to combine limited severity cdfs for *classes A–D* with per-loss limits (and relevant correlation assumptions), Model 4.3 (when applied to *class E*) implicitly combines *classes A–D* without-limits (allowing for any empirical correlation that may exist) before imposing a per-occurrence limit.

Chapter 5

Results and Discussions

"Again, you can't connect the dots looking forward, you can only connect them looking backwards… you have to trust that the dots will somehow connect in your future"

(Jobs, 2010)

5.1 Overview

Various models and algorithms (Chapter 4) are put to the test and analysed in this chapter using data from Chapter 3.

The objective of this chapter is to determine and compare aggregate loss *cdf*s and *limit factors* (§4.2.2) for a range of limits, modelling perspectives (i.e. *IR*, *CR*), types of interclass correlations, and *risk adjustments*.

Definitions of various terminologies have been introduced (e.g. *ALD*s; *IR*, *CR* §4.2.1; *LAS* 4.9 and related measures; *FFT* §4.2.4.4; and *MC* §4.2.6), and Models 4.1–4.6 (§4.4), based on *spliced cdf*s (§4.3), have been described. Depending on the context, *ALD* refers to a *cdf* or *pdf*; and *empirical* or *observed*, and unless stated otherwise, refer to (inflated severity) *data* (Chapter 3).

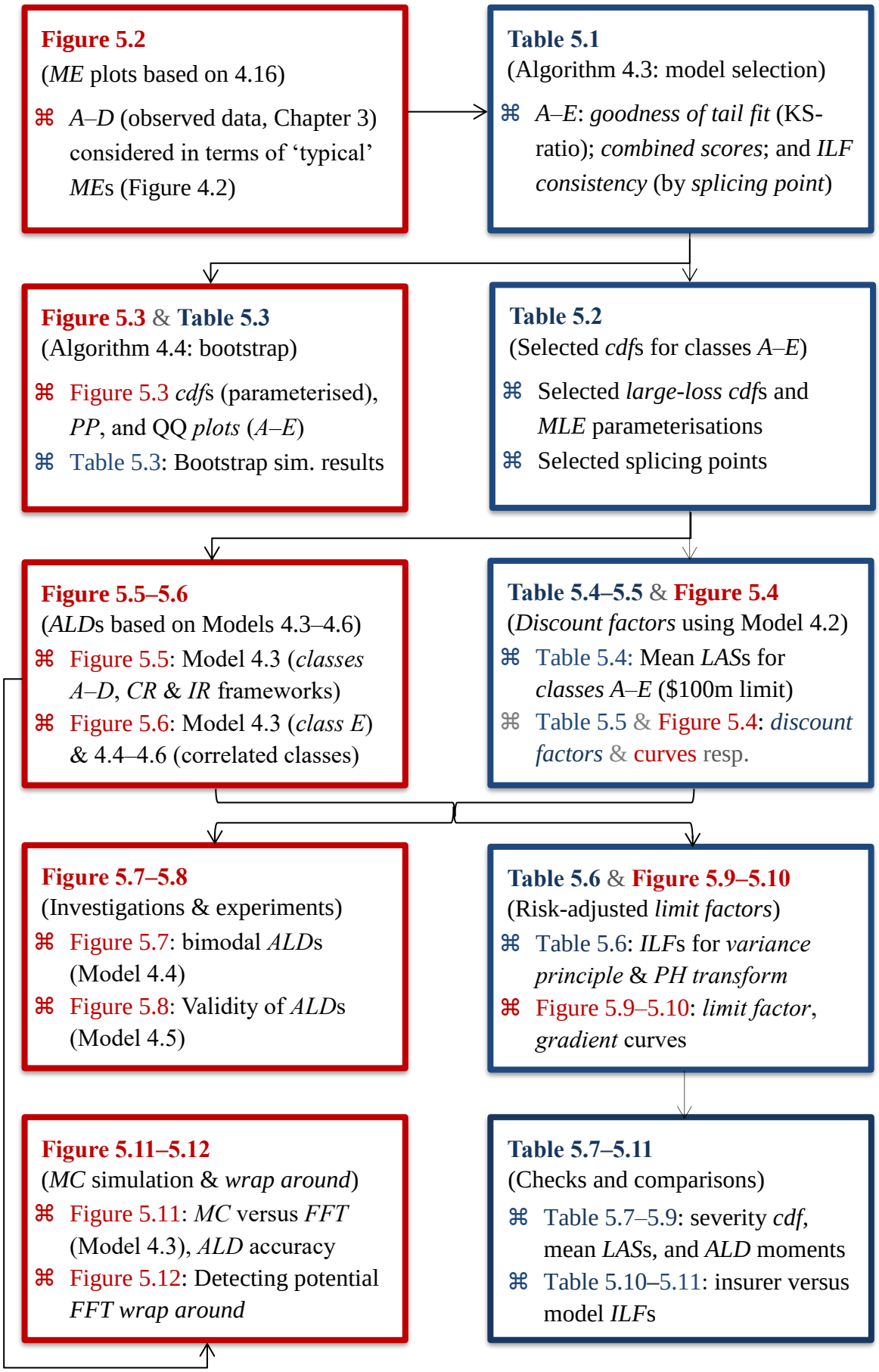


Figure 5.1 Flow of results between figures and tables

The tables and figures in Figure 5.1 can essentially be divided into two main parts:

1. *Severity cdf*s (§5.2): results from §4.3 are used to specify and assess *large-loss* severity *cdf*s underlying *spliced cdf*s of the form 4.69
2. Model *results* (§5.3): feature *discount factors* (Model 4.2), *ALD*s (Models 4.3–4.6), and risk-adjusted *ILF*s (Models 4.3, 4.5–4.6), and concludes with validations and additional investigations (§5.3.4)

1) *Severity cdfs* (§5.2)

Here, data is represented by *ME plots* (Figure 5.2), followed by results of Algorithm 4.3 (Table 5.1) where *severity cdf*s are selected (Table 5.2); *goodness of fit* (i.e. *QQ*- and *PP*-plots: Figure 5.3); and model confidence sets (Table 5.3) based on Algorithm 4.4 are considered.

2) *Model results* (§5.3)

This section is further subdivided into *discount factors* (§5.3.1), *ALD*s (§5.3.2), risk-adjusted *ILF*s (§5.3.3), and validations (§5.3.4) based on Models 4.1–4.6, as follows:

- *Discount factors* (Models 4.1–4.2): these are based on mean *LAS*s at the $100m limit mark (Table 5.4), with a tabulated summary (Table 5.5) of factors that underpin *limit factor curves* in Figure 5.4
- *ALD*s (Models 4.3–4.6): Figure 5.5–Figure 5.6 illustrate *ALD*s for Models 4.3–4.6; correlation is considered in terms of the effect in the tail of distributions (Figure 5.7), and the validity of resulting *ALD*s (Figure 5.8)
- Risk adjustments (Models 4.3, 4.5–4.6): graphical illustrations that consider the impact of correlation in terms of *consistency* and different risk 'environments' (Figure 5.9) are followed by a summary of *variance principle* and *PH transform* risk-adjusted *limit factors* (with ranges indicating different intensity levels, Table 5.6), and gradient curves as part of a 'stress test' (Figure 5.10)
- Validations are made in terms of: severity *cdf*s (*FFT* vs. *limit factors*: Table 5.7); accuracy of mean *LAS*s (Table 5.8); *ALD*s (Model 4.3 vs. Algorithm 4.5: Figure 5.11, Table 5.9); *FFT aliasing error* (Figure 5.12); and reasonableness (insurer *ILF*s: Table 5.10–Table 5.11)

5.2 Spliced severity

Recall that in Algorithm 4.3 that steps 1–3 (p. 4.35) are repeated over a range of *splicing points* to determine input vectors **G**, **x** , and **B**. To this end, a number of candidate *cdf*s are considered in step 1 (Appendix D.2, Table D.1), and maximum *splicing points* (i.e. minimum number of large losses) are set in relation to *ME plots*.

5.2.1 Mean Excess plots

ME plots (§4.2.2.1) for the *data* are illustrated in Figure 5.2. Markers that indicate the apparent onset of volatility, or other such irregularity due to having too few data points (percentiles correspond to maximum permissible thresholds for use in Algorithm 4.3).

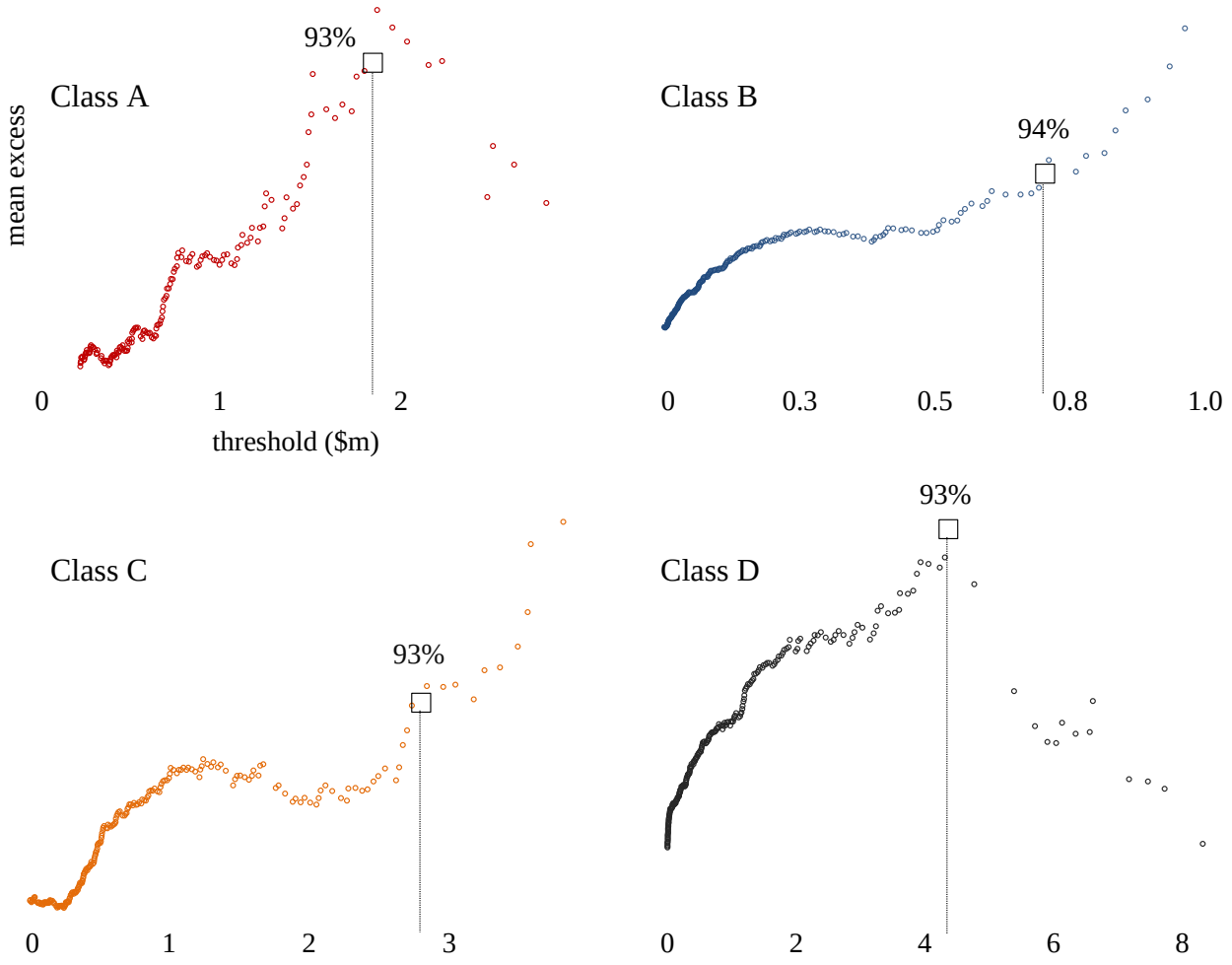


Figure 5.2 Empirical ME plots Axes: *x* (*threshold*, $m), *y* (mean excess, values omitted as they are unnecessary for this exercise). Data: costs sourced from Ponemon Institute (2012a–i, 2013a–j, 2014a–k), inflated to 2016. Square markers (i.e. 94th, 96th, 93rd, and 92nd percentiles: *A*–*D* respectively) indicate the onset volatile or irregular trends (used as maximum percentiles for Algorithm 4.3).

The *ME*s for *classes B* and *C* (Figure 5.2) initially decrease before assuming upward concavity (possibly indicating a *Burr* type *cdf*), and ultimately, continue to increase beyond the indicated percentiles (i.e. 94%, 93% respectively). This could also be indicative of a heavy-tailed *Weibull*, possibly a *Pareto*. In contrast, *ME*s for *class A* and *D* reduce after the threshold of 93% (sharply so, in *class D*), which undermines a *cdf* such as the *Pareto*, and may even imply a short-tailed *cdf* for *D*, as will be explored in further detail. For completeness, the empirical *ME* for *E* (Figure D.1) and 'shifted' *ME*s by percentile for *A*–*E* (Figure D.2) are included in Appendix D.1.

5.2.2 Selecting *large-loss cdf*s (Algorithm 4.3)

Vectors **G** , **x** , and **B** (Algorithm 4.3), representing large-loss *cdf*s (based on AIC^C Kullback-Leibler criterion; *threshold* values, and scaled, inverted *KS-ratios*, §4.3.1, respectively), and overall *combined scores*, **z** (4.76), are summarised by class *A*–*E* and percentile in Table 5.1.

In the first column are threshold percentiles for which large-loss *cdf*s and associated scores have been determined as part of two runs of Algorithm 4.3: the 1st run identifies 'optimal' *splicing points* in relation to the set of percentiles presented (i.e. with increments of one – 65.5%, 66.5%, …, 98.5%, 99.5%); the 2nd run considers percentiles with finer increments (based on underlying empirical *cdf*s) that fall within 4 per cent of the thresholds identified in the initial run. In terms of the *large-loss models*, 'Weibull3' refers a 3-parameter (shifted) *Weibull*; Burr *cdf*s are of a type 3 (i.e. *Dagum, inverse Burr*) with 4 parameters. Colour-coded bars represent the empirical data (quantiles for severity divided by respective maxima, *A*–*E*). *Goodness-of-fit* scores, for each class, are evaluated relative to the lowest *KS-ratio* achieved in the 2nd run. Colour-coded bars associated with these (and *overall combined*) scores illustrate the relative magnitude of the scores.

Final selections (i.e. percentiles with colour-coded font; *cdf*s within boxes: *A*–*E*) correspond to the largest *combined scores* – similar selections (in terms of percentile; type and tail behaviour of *cdf*s) could be made using the second and third largest scores (these are produced in the 2nd run and are, therefore, not shown in this table). Green check marks indicate where percentiles fall within acceptable ranges (based on *ME* plots: Figure 5.2; Figure D.1) and where resulting *spliced cdf*s produce *consistent ILF*s (across all limits considered, Table D.4); red-crosses are used otherwise.

	Large-loss models (*AIC*^{c} selections; final selections are boxed)					Goodness-of-fit score (KS critical: 5%; minimum *KS-ratio by* class, divided by *KS-ratio*)					Overall scores and consistency check (are limit factors consistent over all limits considered?)				
Threshold percentile	***Class A***	***Class B***	***Class C***	***Class D***	***Class E***	***Class A***	***Class B***	***Class C***	***Class D***	***Class E***	***Class A***	***Class B***	***Class C***	***Class D***	***Class E***
0.25%	Weibull	Weibull	Pearson	Weibull	Weibull*	0.63	0.36	0.32	0.71	0.39	✓ 0.31	✓ 0.19	✓ 0.16	✓ 0.35	✓ 0.20
65.5%	Weibull	Weibull	Weibull	Weibull	Weibull	0.73	0.62	0.41	0.79	0.56	✓ 0.71	✓ 0.63	✓ 0.47	✓ 0.76	✓ 0.58
66.5%	Weibull	Weibull	Weibull	Weibull	Weibull	0.73	0.48	0.56	0.75	0.62	✓ 0.72	✓ 0.53	✓ 0.59	✓ 0.73	✓ 0.63
67.5%	Weibull	Weibull	Weibull	Weibull	Weibull	0.80	0.51	0.49	0.76	0.53	✓ 0.77	✓ 0.55	✓ 0.53	✓ 0.74	✓ 0.55
68.5%	Weibull	Weibull	Weibull	Weibull	Weibull*	0.73	0.39	0.52	0.77	0.77	✓ 0.72	✓ 0.46	✓ 0.56	✓ 0.75	✓ 0.76
69.5%	Weibull	Weibull	Weibull	Weibull	Weibull*	0.60	0.49	0.50	0.79	0.73	✓ 0.62	✓ 0.54	✓ 0.54	✓ 0.77	✓ 0.72
70.5%	Weibull	Burr	Weibull	Weibull	Weibull*	0.47	0.25	0.54	0.67	0.74	✓ 0.53	✓ 0.36	✓ 0.58	✓ 0.68	✓ 0.73
71.5%	Weibull	Burr	Weibull	Weibull	Weibull	0.50	0.28	0.53	0.65	0.53	✓ 0.55	✓ 0.37	✓ 0.57	✓ 0.66	✓ 0.56
72.5%	Weibull	Burr	Weibull	Weibull	Weibull*	0.43	0.36	0.40	0.69	0.59	✓ 0.49	✓ 0.44	✓ 0.47	✓ 0.70	✓ 0.62
73.5%	Weibull	Burr	Weibull	Weibull	Weibull	0.43	0.40	0.37	0.59	0.50	✓ 0.49	✓ 0.47	✓ 0.45	✓ 0.62	✓ 0.55
74.5%	Weibull	Burr	Weibull	Weibull	Weibull	0.43	0.36	0.44	0.54	0.56	✓ 0.49	✓ 0.44	✓ 0.50	✓ 0.58	✓ 0.60
75.5%	Weibull	Burr	Weibull	Weibull	Weibull	0.47	0.52	0.38	0.49	0.57	✓ 0.53	✓ 0.57	✓ 0.45	✓ 0.55	✓ 0.61
76.5%	Weibull	Burr	Weibull	Weibull	Weibull	0.50	0.27	0.36	0.59	0.48	✓ 0.55	✓ 0.37	✓ 0.43	✓ 0.62	✓ 0.54
B: 77.50%	Weibull	Burr	Weibull	Weibull	Weibull	0.68	**1.00**	0.35	0.53	0.51	✓ 0.70	✓ **0.96**	✓ 0.43	✓ 0.58	✓ 0.57
78.5%	Weibull	Burr	Weibull	Weibull	Weibull*	0.64	0.80	0.29	0.55	0.65	✓ 0.66	✓ 0.80	✓ 0.37	✓ 0.59	✓ 0.68
79.5%	Weibull	Burr	Weibull	Weibull	Weibull*	0.63	0.31	0.29	0.49	0.75	✓ 0.66	✓ 0.39	✓ 0.37	✓ 0.54	✓ 0.76
C: 81.00%	Weibull	Burr	Burr	Weibull	Weibull*	0.56	0.34	**1.00**	0.62	0.73	✓ 0.60	✓ 0.42	✓ **0.97**	✓ 0.65	✓ 0.75
81.5%	Weibull	Burr	Burr	Weibull	Weibull*	0.53	0.51	0.28	0.59	0.85	✓ 0.58	✓ 0.56	✓ 0.36	✓ 0.63	✓ 0.84
82.5%	Weibull	Burr	Weibull	Weibull	Weibull*	0.52	0.51	0.33	0.54	0.84	✓ 0.56	✓ 0.56	✓ 0.41	✓ 0.58	✓ 0.83
E: 83.91%	Weibull	Weibull	Weibull	Weibull	Weibull*	0.49	0.28	0.42	0.53	**1.00**	✓ 0.54	✓ 0.36	✓ 0.48	✓ 0.57	✓ **0.98**
84.5%	Weibull	Burr	Weibull	Weibull	Weibull	0.57	0.36	0.41	0.54	0.66	✓ 0.61	✓ 0.42	✓ 0.47	✓ 0.58	✓ 0.71
85.5%	Weibull	Burr	Weibull	Weibull	Weibull	0.69	0.40	0.35	0.65	0.58	✓ 0.71	✓ 0.45	✓ 0.41	✓ 0.68	×
86.5%	Weibull	Weibull	Weibull	Weibull	Weibull	0.82	0.48	0.41	0.68	0.58	✓ 0.82	✓ 0.52	✓ 0.47	✓ 0.70	×
A: 87.25%	Weibull	Weibull	Weibull	Weibull	Weibull	**1.00**	0.46	0.39	0.75	0.66	✓ **0.99**	✓ 0.50	✓ 0.44	✓ 0.76	×
88.5%	Weibull	Weibull	Weibull	Weibull	Weibull	0.80	0.48	0.39	0.63	0.56	✓ 0.81	✓ 0.52	✓ 0.44	✓ 0.66	×
89.5%	Weibull	Weibull	Weibull	Weibull	Weibull	0.91	0.34	0.48	0.49	0.52	✓ 0.91	✓ 0.39	✓ 0.52	✓ 0.53	×
90.5%	Weibull	Burr	Weibull	Weibull	Weibull	0.75	0.63	0.41	0.51	0.58	✓ 0.76	✓ 0.65	✓ 0.45	✓ 0.54	×
91.5%	Weibull	Weibull	Weibull	Weibull*	Weibull	0.88	0.28	0.49	0.65	0.49	✓ 0.88	✓ 0.33	✓ 0.52	✓ 0.67	×
D: 92.12%	Weibull	Burr	Weibull	Weibull*	Weibull	0.67	0.30	0.31	**1.00**	0.49	✓ 0.68	✓ 0.35	✓ 0.35	✓ **0.99**	×
93.5%	Weibull	Burr	Burr	Weibull	Burr	0.52	0.26	0.25	0.67	0.28	✓ 0.55	×	×	×	×
94.5%	Weibull	Burr	Weibull	Weibull	Burr	0.44	0.30	0.48	0.77	0.26	×	×	×	×	×
95.5%	Weibull	Burr	Burr	Weibull	Weibull	0.41	0.28	0.27	0.64	0.30	×	×	×	×	×
96.5%	Fatigue	Burr	Burr	Fatigue	Burr	0.32	0.30	0.20	0.21	0.29	×	×	×	×	×
97.5%	Fatigue	Fatigue	Fatigue	Fatigue	Weibull	0.24	0.19	0.26	0.30	0.61	×	×	×	×	×
98.5%	Fatigue	Fatigue	Fatigue	Fatigue	Fatigue	0.39	0.20	0.25	0.67	0.15	×	×	×	×	×
99.5%	Fatigue	Fatigue	Fatigue	Fatigue	Fatigue	0.47	0.31	0.35	0.71	0.34	×	×	×	×	×

Table 5.1 Large-loss cdfs and scores Final selections (percentiles: coloured font, *A*–*E*; *cdf*s: boxed) correspond to maximum *overall scores* (boxed). *Weibull* (shifted; asterisked: light-tailed), *Burr* (type III: *Dagum*), and *Pearson*: 3, 4, and 6 parameter *cdf*s respectively. Coloured bars: models – quantile divided by maximum (empirical severity); scores – relative magnitude. Criteria for ✓ (failing which, ×): percentile deemed to be acceptable (in terms of *ME* plots); *spliced cdf* yields *consistent ILF*s over a given set of limits ($10k, $100m). Underlying costs: Ponemon Institute (2012a–i, 2013a–j, 2014a–k, 2015g), inflated to 2016.

The *cdf*s identified by AIC^c (Table 5.1: *A*–*E*) generally agree with *ME plots* (Figure 5.2: *A*–*D*; Figure D.1: *E*) for respective classes (e.g. *A*, *C*: heavy-tailed *Weibull*, *Burr*; *B*: *Burr*; *D*: light-tailed *Weibull*, 92% threshold; *E*: light- and heavy- tailed *Weibull cdf*s).

As mentioned, final selections (Table 5.2) are reinforced by the fact that the top 3 largest *combined scores* (2nd run) yield similar results in terms of distributions and percentiles (Table D.2).

	A	*B*	*C*	*D*	*E*
Threshold ($m)	1.40	0.29	1.67	4.14	6.50
Percentile	87.3%	77.5%	81.0%	92.1%	83.9%
Distribution	Weibull	Burr	Burr	Weibull	Weibull
Shape	0.76	2.12, 0.53	2.13, 0.57	1.56	1.11
Scale ($m)	0.82	0.38	1.34	3.37	3.81
Location ($m)	1.40	0.29	1.67	4.14	6.50

Table 5.2 Selected large-loss cdfs and splicing points *Threshold*: dollar value of splicing point; *Burr* represents inverse *Burr* (i.e. *Dagum cdf*); *cdf*s fit using *MLE* to severities from Ponemon Institute (2012a–i, 2013a–j, 2014a–k, 2015g), inflated to 2016.

5.2.3 Model confidence sets (Algorithm 4.4)

Table 5.4 shows results for Algorithm 4.4 (10k bootstrap samples), by class. In the first column are the top four models (selected *cdf*s, Table 5.2, are colour coded), according to how frequently they were selected on the basis of AIC^C (i.e. % selected, 2nd column). For each such model, the average AIC^C weight, *KS*, and *AD*-ratios are shown together with the rate (per 100) for which resulting *ILF*s were consistent (over the range of limits in Table 5.5). The proportion of samples for which a given model falls within the 90% confidence set (based on *Akaike* differences, as described previously) is reported under the heading '*Confidence set %*'.

- *Selected %* ($\hat{\pi}$, following 4.77), *AIC weight* ($\hat{w}$), *KS* and *AD* ratios (based on 4.39 and 4.40) are in agreement; $\hat{\pi}$ and $\hat{w}$ are highest for *selected cdfs*, except for *C* (*Weibull*, the highest, fails the *AD*-test, 5% critical; also, the selected *Burr cdf* has a similar 90% *confidence set* success rate, $\hat{c}^{(90\%)}$)
- Light-tailed *cdf* selections are confirmed for *D*, *E* (with average shape $\alpha > 1$)

- Lowest and highest $\hat{c}^{(90\%)}$ can be seen for *D* (due to high, 92.5% truncation, Table 5.2) and *E* (due to additional 350 observations, year 2015, Chapter 3) respectively
- Selected *cdf*s appear to strike an appropriate balance between $\hat{\pi}$, $\hat{c}^{(90\%)}$, and *tail-fit* ratios

	Model	*Selected %*	*AIC weight*	*Confidence set %*	*KS-ratio*	*AD-ratio*	*Consistent ILFs*
A	Weibull3 (α=0.76)	79%	77%	68%	0.5	0.6	99.8
	Burr	17%	18%	25%	1.1	1.3	54.4
	Fatigue	3%	3%	2%	2.5	17.8	24.5
	LogLaplace	1%	1%	2%	3.4	67.8	99.5
B	Burr	74%	67%	48%	0.7	0.5	99.9
	Weibull3	22%	22%	23%	0.7	1.1	99.9
	LogGamma	3%	7%	17%	0.4	0.2	100
	GEV	1%	2%	8%	0.4	0.2	100
C	Weibull3	55%	52%	26%	0.7	1.0	99.9
	Burr	29%	29%	25%	0.8	0.6	98.0
	LogGamma	14%	12%	17%	0.4	0.2	100
	LogLaplace	1%	2%	8%	1.0	0.3	99.9
D	Weibull3 (α=1.20)	49%	34%	11%	0.6	0.6	99.8
	Fatigue	21%	21%	4%	2.2	14.0	41.9
	Burr	9%	9%	10%	1.5	1.5	52.1
	Pearson5	8%	6%	8%	0.5	0.2	100
E	Weibull3 (α=1.04)	85%	80%	74%	0.5	0.6	100
	Burr	8%	10%	13%	0.9	0.3	95.6
	LogGamma	7%	8%	12%	0.4	0.2	100
	LogLaplace	0%	1%	1%	1.7	0.7	99.3

Table 5.3 Bootstrap results 10k samples; *selected %* achieving minimum AIC^C; 90% confidence sets based on Kullback-Leibler distance estimate for selected *cdf* (colour coded font, *A*–*E* - average shape parameter for *Weibull cdf* selections). *Tail-fit* ratios (*KS*, *AD* - 5% critical); *consistent ILF*s (rate per 100). Underlying costs based on Ponemon Institute (2012a–i, 2013a–j, 2014a–k, 2015g) inflated to 2016.

5.2.4 *Cdf*, *QQ*, and *PP* plots

The *cdf*, *QQ*, and *PP* plots in Figure 5.3 (first, second, and third column, respectively) illustrate *goodness of fit*, by class (i.e. row), in relation to the *large-loss cdf*s (Table 5.2).

In terms of *PP plots*, these *cdf*s appear to resemble the empirical *cdf*s reasonably well; as for the QQ plots, deviations occur in the extremities of the data as might be expected (e.g. maximum: *A*; largest three values: *B*, *C*, and *E*); distributions are otherwise reasonably well aligned with the empirical data.

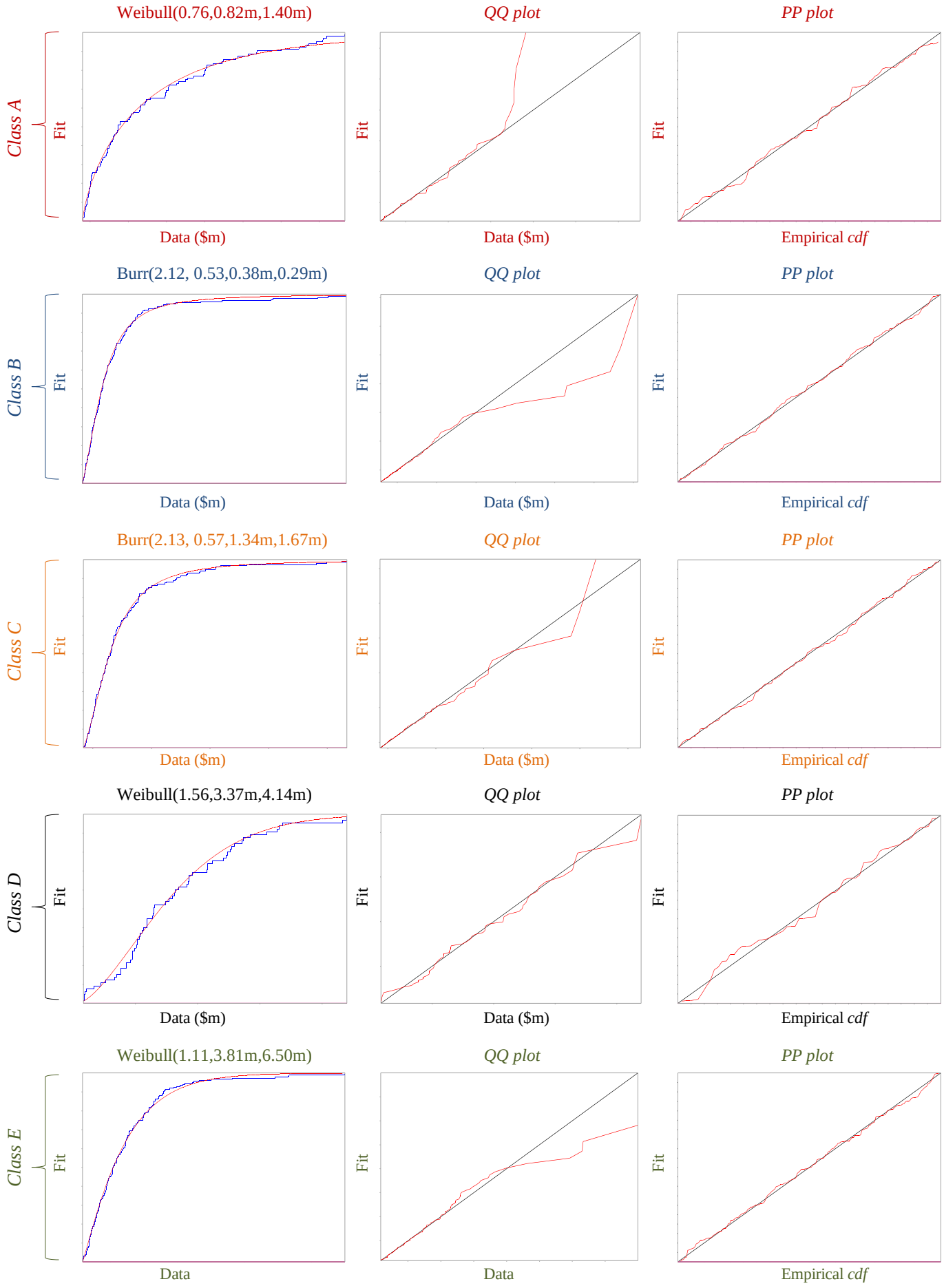


Figure 5.3 Cdfs, QQ, and PP plots for large losses *Rows* correspond to different classes: columns (1–3) correspond to different types of plots: 1) empirical (blue line) vs. model (red line) *cdfs*; 2–3) quantile-quantile (*qq*) and probability-probability plots, respectively: data (red line) vs. fitted (black line). Costs based on Ponemon Institute (2012a–i, 2013a–j, 2014a–k, 2015g), inflated to 2016.

5.3 Limit factors and ALDs

5.3.1 Discount factors (Models 4.1–4.2)

The *discount factor* at a given limit is derived by dividing the mean *LAS* at that limit by the mean *LAS* at the $100m limit. Table 5.4 reports mean *LAS*s at the $100m limit for severities based on the *empirical data* (alone) and *spliced models* (§4.3), for each of *classes A–E*.

The *cdf*s and *thresholds* used to construct *spliced cdf*s were presented in Table 5.2 – *limit factors* based on these are compared to those for the empirical data (hereafter, '*spliced*' and '*empirical*' *limit factors* respectively; *LAS*s are referred to in a similar fashion).

Severity *cdf*	*A*	*B*	*C*	*D*	*E*
Empirical	555.4	169.8	823.5	1 192.8	4 126.7
Spliced	564.7	169.9	832.3	1 189.3	4 123.4

Table 5.4 Empirical vs. spliced mean LASs (Model 5.2) $m; apply to mean *LAS*s at $100m limit. Underlying costs based on Ponemon Institute (2012a–i, 2013a–j, 2014a–k, 2015g), inflated to end of 2016 year.

*Spliced LAS*s at $100m are greater than *empirical* counterparts (Table 5.4, *A–E*), due to *spliced cdf*s having heavier tails (Definition 4.3), which is very similar to the effect previously described for *QQ* plots previously (Figure 5.3, *A–E*).

Attention is now turned to the *discount factors* that apply to these mean *LAS*s (Table 5.5, *base limit* $100m), based on Models 4.1–4.2, and in particular, the extent to which these satisfy *consistency tests* (Properties 4.1).

Number formats in this table are as follows: percentages (to the nearest two decimals) represent limit factors less than one, whilst numeric '1' is equivalent to one (100.00% refers to a number less than one). As such, *spliced discount factors* continue to increase as limits increase towards $100m – Figure 5.4 illustrates the effect this has on *limit factors*, at high limits.

	Empirical discount factors					*Spliced discount factors*				
Limit ($)	*A*	*B*	*C*	*D*	*E*	*A*	*B*	*C*	*D*	*E*
10 000		4.40%	0.97%	0.66%			4.40%	0.96%	0.66%	
20 000	2.88%	8.35%	1.94%	1.28%		2.83%	8.35%	1.91%	1.28%	
30 000	4.32%	11.93%	2.89%	1.88%		4.24%	11.92%	2.86%	1.89%	
40 000	5.73%	15.20%	3.83%	2.48%		5.64%	15.18%	3.79%	2.49%	
50 000	7.13%	18.26%	4.76%	3.06%		7.01%	18.24%	4.71%	3.07%	
60 000	8.50%	21.17%	5.69%	3.64%		8.36%	21.15%	5.63%	3.65%	
70 000	9.85%	23.87%	6.61%	4.22%		9.69%	23.85%	6.54%	4.23%	
80 000	11.18%	26.38%	7.52%	4.79%		10.99%	26.36%	7.45%	4.80%	
90 000	12.48%	28.75%	8.43%	5.35%	2.51%	12.28%	28.72%	8.34%	5.36%	2.51%
100 000	13.77%	31.01%	9.33%	5.91%	2.79%	13.55%	30.98%	9.23%	5.92%	2.79%
250 000	30.75%	55.17%	21.79%	13.72%	6.92%	30.25%	55.13%	21.56%	13.76%	6.93%
400 000	43.89%	70.16%	32.27%	20.53%	10.91%	43.17%	70.01%	31.93%	20.59%	10.92%
550 000	54.39%	79.91%	40.67%	26.51%	14.79%	53.50%	79.64%	40.24%	26.59%	14.80%
700 000	62.76%	85.83%	47.59%	31.85%	18.51%	61.72%	85.42%	47.09%	31.94%	18.52%
850 000	69.10%	89.50%	53.53%	36.69%	22.02%	67.96%	88.95%	52.96%	36.80%	22.03%
1 000 000	74.31%	91.68%	58.65%	41.14%	25.38%	73.09%	91.24%	58.03%	41.26%	25.40%
1 150 000	78.61%	92.97%	63.08%	45.21%	28.61%	77.32%	92.80%	62.41%	45.34%	28.63%
1 300 000	82.06%	93.98%	67.02%	48.81%	31.70%	80.71%	93.93%	66.31%	48.96%	31.72%
1 500 000	85.73%	94.91%	71.66%	53.08%	35.57%	84.27%	95.00%	70.91%	53.23%	35.59%
1 700 000	88.38%	95.74%	75.66%	56.93%	39.19%	86.86%	95.78%	74.85%	57.09%	39.22%
1 900 000	90.46%	96.47%	79.18%	60.39%	42.62%	88.89%	96.35%	78.30%	60.57%	42.66%
2 100 000	92.12%	97.13%	82.25%	63.53%	45.83%	90.54%	96.80%	81.33%	63.72%	45.87%
2 300 000	93.52%	97.72%	84.83%	66.42%	48.86%	91.89%	97.16%	83.91%	66.61%	48.90%
2 500 000	94.78%	98.31%	87.02%	69.04%	51.72%	93.02%	97.44%	86.06%	69.24%	51.76%
3 000 000	97.07%	99.36%	90.88%	74.73%	58.10%	95.12%	97.97%	89.96%	74.95%	58.14%
3 500 000	98.41%	99.97%	93.40%	79.26%	63.46%	96.53%	98.33%	92.40%	79.49%	63.51%
4 000 000	99.30%	1	94.99%	82.78%	67.94%	97.50%	98.58%	93.99%	83.03%	67.99%
4 500 000	99.91%	1	96.16%	85.55%	71.95%	98.18%	98.78%	95.08%	85.82%	72.01%
5 000 000	1	1	97.11%	87.99%	75.47%	98.66%	98.93%	95.87%	88.29%	75.53%
5 500 000	1	1	97.77%	90.32%	78.46%	99.01%	99.04%	96.46%	90.51%	78.52%
6 000 000	1	1	98.27%	92.42%	81.11%	99.26%	99.14%	96.91%	92.44%	81.17%
6 500 000	1	1	98.60%	94.06%	83.47%	99.44%	99.22%	97.27%	94.08%	83.54%
7 000 000	1	1	98.84%	95.35%	85.64%	99.58%	99.29%	97.57%	95.43%	85.67%
7 500 000	1	1	99.06%	96.46%	87.57%	99.68%	99.35%	97.81%	96.53%	87.58%
8 000 000	1	1	99.24%	97.35%	89.29%	99.76%	99.40%	98.01%	97.41%	89.26%
8 500 000	1	1	99.42%	98.10%	90.78%	99.81%	99.44%	98.18%	98.09%	90.73%
9 000 000	1	1	99.60%	98.70%	92.09%	99.86%	99.48%	98.32%	98.61%	92.02%
9 500 000	1	1	99.78%	99.10%	93.23%	99.89%	99.51%	98.45%	99.00%	93.14%
10 000 000	1	1	99.92%	99.36%	94.21%	99.92%	99.54%	98.56%	99.29%	94.11%
11 000 000	1	1	1	99.71%	95.78%	99.95%	99.60%	98.74%	99.66%	95.68%
12 000 000	1	1	1	99.88%	96.93%	99.97%	99.64%	98.89%	99.84%	96.84%
13 000 000	1	1	1	1	97.76%	99.98%	99.67%	99.01%	99.93%	97.70%
14 000 000	1	1	1	1	98.27%	99.99%	99.70%	99.11%	99.97%	98.34%
15 000 000	1	1	1	1	98.60%	99.99%	99.73%	99.19%	99.99%	98.80%
16 000 000	1	1	1	1	98.86%	100.00%	99.75%	99.26%	100.00%	99.14%
17 000 000	1	1	1	1	99.04%	100.00%	99.77%	99.32%	100.00%	99.38%
18 000 000	1	1	1	1	99.19%	100.00%	99.79%	99.38%	100.00%	99.56%
19 000 000	1	1	1	1	99.31%	100.00%	99.80%	99.42%	100.00%	99.68%
20 000 000	1	1	1	1	99.43%	100.00%	99.81%	99.47%	100.00%	99.77%
35 000 000	1	1	1	1	1	100.00%	99.92%	99.78%	100.00%	100.00%
40 000 000	1	1	1	1	1	100.00%	99.94%	99.82%	100.00%	100.00%
45 000 000	1	1	1	1	1	100.00%	99.95%	99.86%	100.00%	100.00%
50 000 000	1	1	1	1	1	100.00%	99.96%	99.88%	100.00%	100.00%
55 000 000	1	1	1	1	1	100.00%	99.97%	99.91%	100.00%	100.00%
60 000 000	1	1	1	1	1	100.00%	99.97%	99.92%	100.00%	100.00%
65 000 000	1	1	1	1	1	100.00%	99.98%	99.94%	100.00%	100.00%
70 000 000	1	1	1	1	1	100.00%	99.98%	99.95%	100.00%	100.00%
75 000 000	1	1	1	1	1	100.00%	99.99%	99.96%	100.00%	100.00%
80 000 000	1	1	1	1	1	100.00%	99.99%	99.97%	100.00%	100.00%
85 000 000	1	1	1	1	1	100.00%	99.99%	99.98%	100.00%	100.00%
90 000 000	1	1	1	1	1	100.00%	100.00%	99.99%	100.00%	100.00%
95 000 000	1	1	1	1	1	100.00%	100.00%	99.99%	100.00%	100.00%
100 000 000	1	1	1	1	1	1	1	1	1	1

Table 5.5 Empirical versus spliced discount factors (*classes A–E*) Based on Model 4.2. Factors apply to mean *LAS* at $100m limit. Costs based on Ponemon Institute (2012a–i, 2013a–j, 2014a–k, 2015g), inflated to 2016.

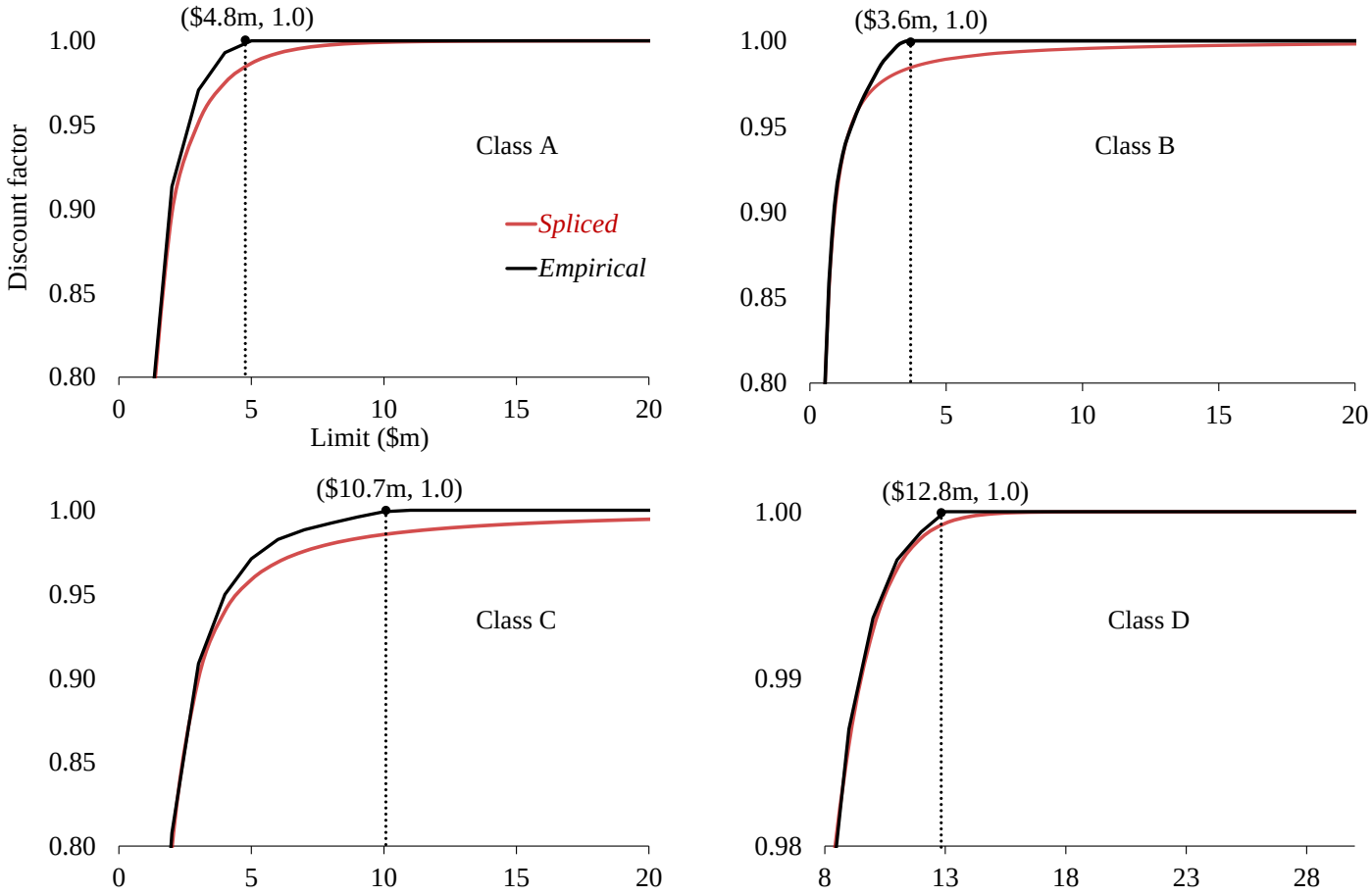


Figure 5.4 Discount factor curves (*classes A–D*, Model 4.2) *Limit factors* expressed as discount factors (to respective *LAS*s at $100m limit for *classes A–D*) - based on costs Ponemon Institute (2012a–i, 2013a–j, 2014a–k), inflated to 2016.

Naturally, *empirical* limit factors reach '1' beyond observed maxima. However, given the nature of *incidental truncation* described previously (§3.1), larger values can be expected for these classes (in general, larger values can always be expected for empirical samples). In the absence of any external restrictions, *limit factors* should be strictly monotonic and increasing over the entire range of limits (in this case, up to $100m).

In this this way, *empirical limit factors* are regarded as undermining the first *consistency property*. In contrast, *spliced limit factors* should adhere to this property, and, provided threshold values are suitably low (and continuity, differentiability, and other Properties 4.1 are in order) *consistency properties* should be satisfied.

5.3.2 Aggregate Loss Distributions (Models 4.3–4.6)

In Figure 5.5 individual *ALD*s for *A–D* (Model 4.3 *IR*, *CR*) are illustrated; combined versions of these (Models 4.4–4.6) are then shown in Figure 5.6, accompanied by Model 4.3 (*IR*, *CR*) in respect of *class E*.

The following assumptions are made for loss count, limits, *discretisation* (for *FFT*, §4.2.4.4), and correlation (for Models 4.4–4.6).

Loss count

This assumption depends upon the framework (i.e. *IR*, *CR*) as follows:

- *IR* – Models 4.3–4.4 assume a deterministic loss count of 10
- *CR* – Models 4.3 and 4.5 assume a *Poisson* loss count with mean 10
  – Model 4.6 assumes a multivariate negative binomial with mean 10 and variance 20 (4.66 with *MNB*(**10**,**1**,0.09))

It can be noted that loss count assumptions are convenient, but otherwise arbitrary (however, assumptions are consistent across models) – modelling empirical severity is of greater interest for the present research.

Severity limits *A–E*

Illustrated in Figures 5.5–5.6 are *per-loss* and *per-occurrence* limits (§4.2.1), defined as:

- *Per-loss* limit: $20m (*A–D*)
- Per-occurrence limit: $80m (*E*)

These are high enough to ensure large-loss *cdf*s make a reasonable contribution towards the *ALD* (Table 5.2).

This allows for suitable *discretisation* (as is described shortly), and, in this particular case, a *per-occurrence* limit four times the *per-loss* limit provides some level of consistency between the limited severities for Model 4.3 (*CR*) and Model 4.5 (Scenarios 1–3): Figure 5.6 (1), and Model 4.3 (*IR*) and Model 4.4 (Scenarios 1–3): Figure 5.6 (2).

Class E (subject to *per-occurrence* limit: \$80m) can be expected to be at least as large as the sum of *A*–*D* (each subject to per-loss limit: \$20m). It can be shown that equality will occur whenever (severities for) *A*–*D* are (simultaneously) less than, greater than, or equal to \$20m with strict inequality diversely (that is, the limited severity *E* will be greater than the sum of limited severities in *A*–*D*).

According to the data, equality occurs in every one of the 800 cases. However, these limits have implications for *ALD*s, as is described shortly.

Discretisation of *spliced-severity cdf*s

ALDs (Figures 5.5–5.6) represent 4 096 discretised points, based on the method of *mass dispersal* (Appendix C.1) as illustrated by Wang (1998: 46–47) and Klugman, Panjer & Willmot (2004, sec. 6.6.5), with the following *truncation points*:

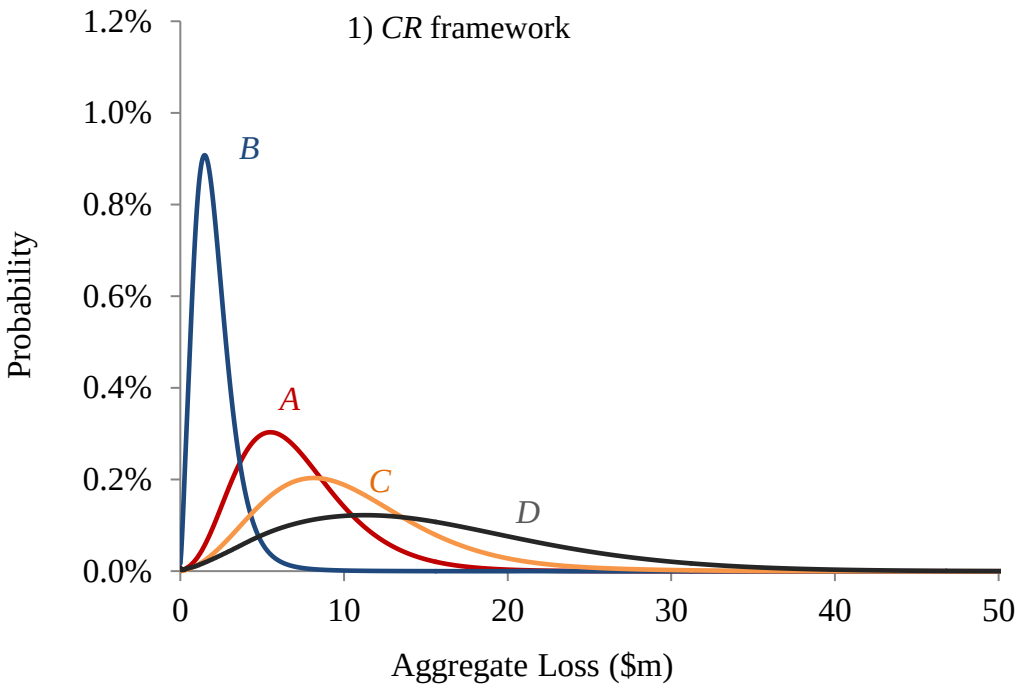
- Figure 5.5 (Model 4.3): *truncation* of \$96.2m, roughly 6.5 times the mean *LAS* for *class D*, which has the largest mean *LAS* compared to other classes. This equates to a *span* of approximately \$23.5k $(\approx \frac{96.2\text{m}}{4096})$
- Figure 5.6 (Models 4.3–4.6): *truncation* of \$287.1m, roughly 8 times the mean *LAS* for *class E*, corresponding to a *span* of \$70.1k $(\approx \frac{287.1\text{m}}{4096})$

In each case, *truncation* points are selected to ensure there is no *wrap-around* issue encountered by the *FFT*, whilst maintaining an acceptably low level of *discretisation error* (*wrap around* and *discretisation* are described in further detail later in this section).

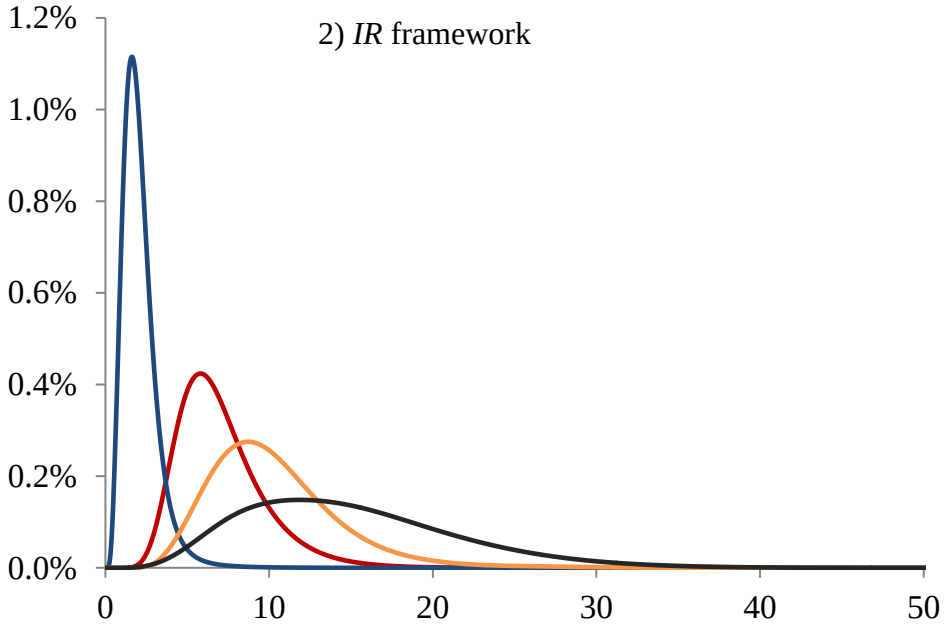
Correlation parameters (Models 4.4–4.6)

Scenario $r = 1, 2, 3$ for Models 4.4–4.5, is based on 4.62 with covariance coefficient $\kappa_{ij} = \kappa_r = 0.05(r-1) \;\forall\; i < j$ (i.e. 0%, 5%, and 10% for scenarios 1–3 respectively).

Model 4.6 represents a 'single' scenario: $\alpha_j = \alpha = 10$ (4.66, $j = 1, \ldots, 4$) with parameter $w = 0.09$ satisfying the first condition: $w \in (0, \alpha^{-1})$.

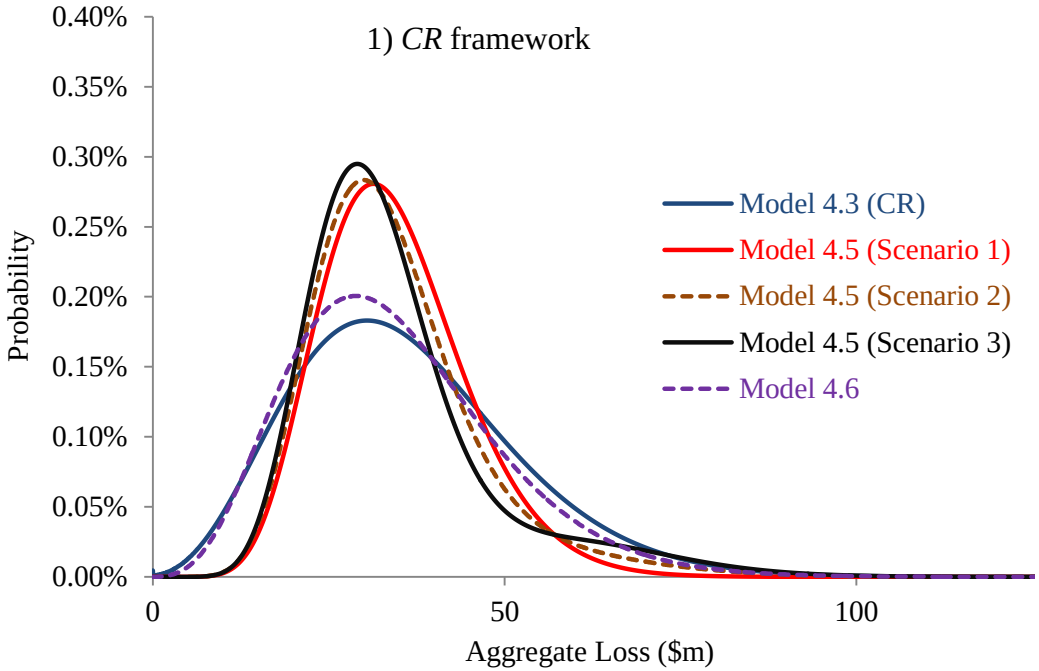


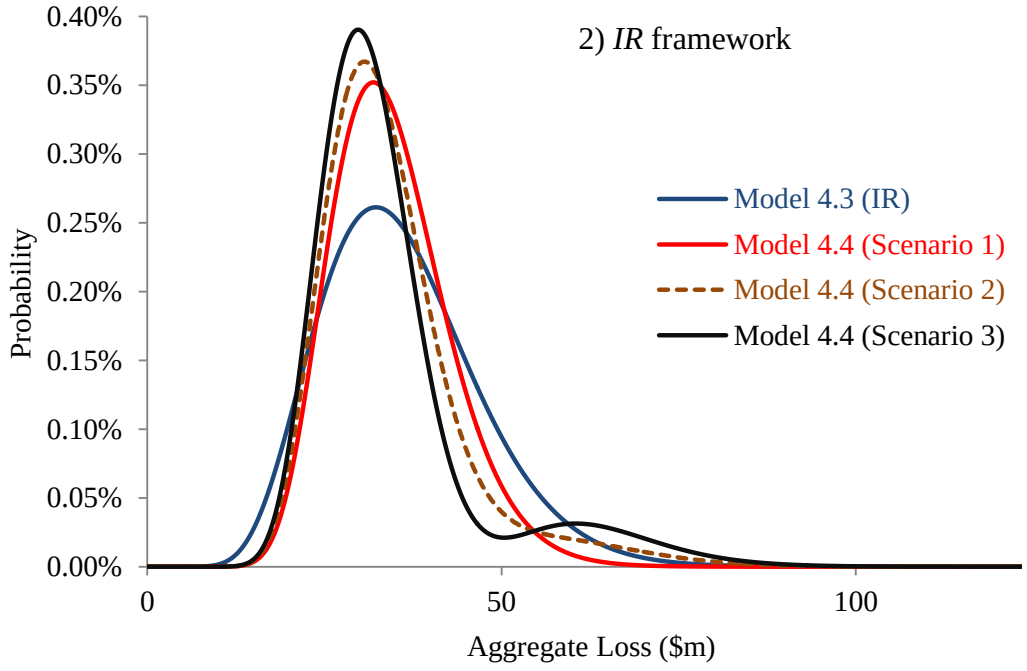
	A	*B*	*C*	*D*
Mean	7.05	2.12	10.35	14.99
Min	0	0	0	0
Std dev	3.48	1.39	5.24	8.08
Skew	1.00	3.03	1.10	0.84
Kurt	4.67	28.60	5.25	3.87
VaR $_{1\%}$	17.58	6.56	26.87	38.38



	A	*B*	*C*	*D*
Mean	7.05	2.12	10.35	14.99
Min	0.54	0	0.61	0.09
Std dev	2.67	1.22	4.10	6.54
Skew	1.23	3.93	1.39	0.77
Kurt	5.74	43.29	6.91	3.68
VaR $_{1\%}$	15.55	6.08	24.36	33.61

Figure 5.5 ALDs: Model 4.3 *CR* is based on Poisson(10) loss count, and *IR* assumes a deterministic loss count of 10. Data based on Ponemon Institute (2012a–i, 2013a–j, 2014a–k), with costs inflated to 2016.





	Model 4.3 (*CR*)	Scenario 1	Scenario 2	Scenario 3	Model 4.6	Model 4.3 (*IR*)	Scenario 1	Scenario 2	Scenario 3
Mean	35.84	34.46	34.46	34.46	34.46	35.84	34.46	34.46	34.46
Min	0	1.75	1.68	1.61	0	3.50	7.22	7.15	7.08
Std dev	15.83	10.33	12.12	13.67	14.75	11.06	8.26	10.41	12.18
Skew	0.66	0.59	1.27	1.52	0.75	0.62	0.61	1.63	1.84
Kurt	3.56	3.51	5.85	6.08	3.79	3.50	3.55	7.08	6.79
VaR $_{1\%}$	79.85	62.88	75.40	81.58	76.44	66.28	57.23	72.28	77.31

Figure 5.6 ALDs: Models 4.3–4.6 $m; Scenarios 1–3 represent constant covariance coefficients of 0%, 5%, and 10% respectively, for use in Models 4.4 (*IR* framework) and 4.5 (*CR* framework). *CR* loss count: *Poisson* with mean 10 (Models 4.3–4.5); *MNB*(**10**,**1**,0.09) for Model 4.6; *IR* loss count: 10 (deterministic). Underlying data based on costs from Ponemon Institute (2012a–i, 2013a–j, 2014a–k, 2015g), inflated to end of 2016 year.

5.3.2.1 *Underlying cost types*

*ALD*s for *A–E* (Figures 5.5–5.6) are now considered in terms of associated cost types (Table 3.1) and underlying large-loss *cdf*s (Table 5.2)

- *B*: this has the lowest mean (Figure 5.5: 1, 2) and largest kurtosis – in keeping with the fact that these costs are not significant drivers of overall loss (e.g. data recreation, expert engagement, possibly customer notification); and the element of 'determining regulatory requirements', suggesting a heavier tail than otherwise (i.e. in support of the *Burr cdf*, Table 5.2)
- *A*, *C*: most similar in terms of *ALD*s and moments – this agrees with underlying cost types which appear to be overlapping in some aspects (e.g. forensic, investigative, communication, assessment costs); however, the nature of other costs in *C* (legal, regulatory fines and penalties, product discounts, and credit monitoring) would explain its relatively larger moments and heavier tail
- *D*: the largest mean and, as implied by the lowest kurtosis and skewness (relative to mean), lightest (severity cdf and *ALD*) – this appears to reflect the nature of the underlying extrapolated cost estimate that has been derived from some other distribution

5.3.2.2 *Impact of correlation*

By increasing the *covariance coefficient*, κ (Figure 5.6, 1), *variance*, *skewness*, and *kurtosis*, for Model 4.5, also increase. This is consistent with Model 4.4, Scenarios 1–2, but not Scenario 3, which has a lower *kurtosis* than Scenario 2 (i.e. 6.79 vs. 6.08), which is due to the formation of a *bimodal ALD*. Bimodal *ALD*s appear to result from 'spikes' in underlying discretised severity *cdf*s. In this case, due to common *per-loss* limits coupled with *FFT* discretisation issues when combining heterogonous *cdf*s (e.g., *B*, *D*). To investigate further, Scenario 3 is compared to Scenario 1 in terms of Model 4.4 (as described shortly, similar comments apply to Model 4.5). Scenario 3 should bear closer resemblance to Scenario 1 if one of *A–D* were to be 'exchanged' with a mutually independent but otherwise identically distributed class (akin to the concept of '*reciprocity*' in the context of reinsurance).

To this end, an experiment is performed using a log-log scale for Model 4.4 (Scenario 3), as Figure 5.6 illustrates, where, in turn, each of one *A*–*D* is assumed to be independent of the others, as follows:

Sensitivity 1: *Class A* is independent of *classes B*, *C*, and *D*

Sensitivity 2: *Class B* is independent of *classes A*, *C*, and *D*

Sensitivity 3: *Class C* is independent of *classes A*, *B*, and *D*

Sensitivity 4: *Class D* is independent of *classes A*, *B*, and *C*

*ALD*s for sensitivities 1–4, as well as for Scenarios 1 and 3 are illustrated in Figure 5.7. The *covariance coefficient* in respect of a class, assumed to be independent of every other class (sensitivities 1–4), is set to zero, and kept at 10% for other classes (in accordance with Scenario 3).

For example, if *C* is being tested as the independent class (i.e. sensitivity 3) then $\kappa = 0\%$ for (*A*,*C*), (*B*,*C*), and (*C*,*D*), and $\kappa = 10\%$ for other pairs: (*A*,*B*), (*A*,*D*), and (*B*,*D*).

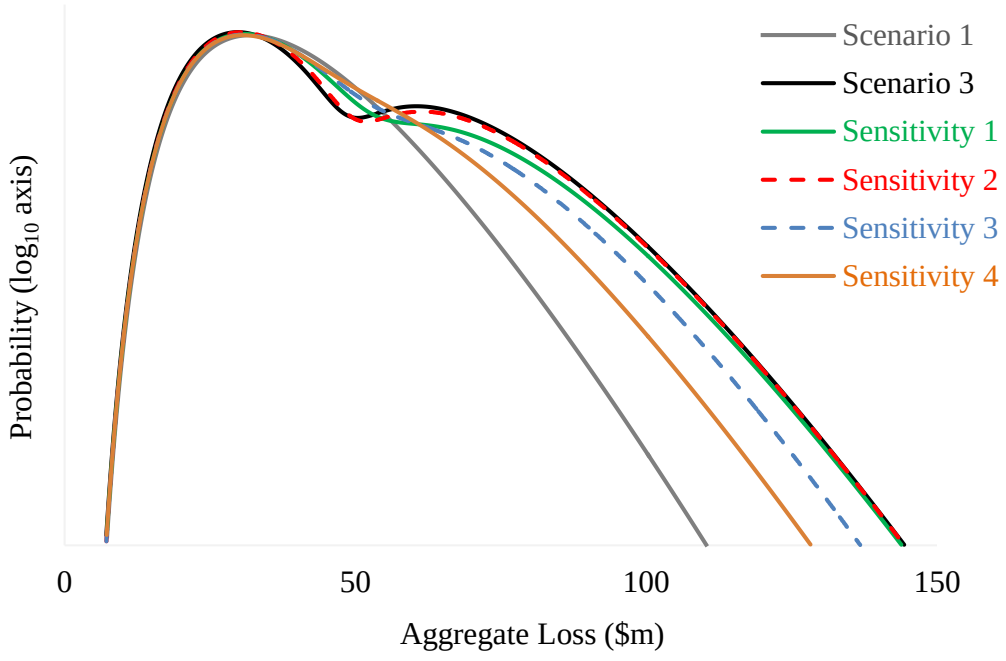


Figure 5.7 Bimodal feature for different sensitivities (Model 4.4) *ALD*s for sensitivities 1–4 as well as the common independent *ALD* (Scenario 1) and Scenario 3 are based on Model 4.4 with (deterministic) loss count of 10. Data based on Ponemon Institute (2012a–i, 2013a–j, 2014a–k), with costs inflated to 2016.

As can be seen, the *ALD* for Sensitivity 4 lies between Scenario 1 and other sensitivities (Figure 5.7); further, it does not appear to have a clear bimodal feature (which can otherwise be seen near the \$50m mark). Therefore, in terms of the extent of the *bimodal* feature (and potential *ALD* invalidity) for Model 4.4, correlation with respect to *D* has the greatest impact compared to other classes. This is consistent with previous observations for relative *tail ratios* (Figure 3.4).

Applying a *PH transform* to the severity *cdf* of *D* could be another way to deal with the bimodal issue (*FFT* discretisation may require updating to prevent aliasing issues). As previously mentioned, similar conclusions apply for Model 4.5, which, as can be seen in Figure 5.8, also leads to bimodal (and eventually invalid) *ALD*, as κ increases.

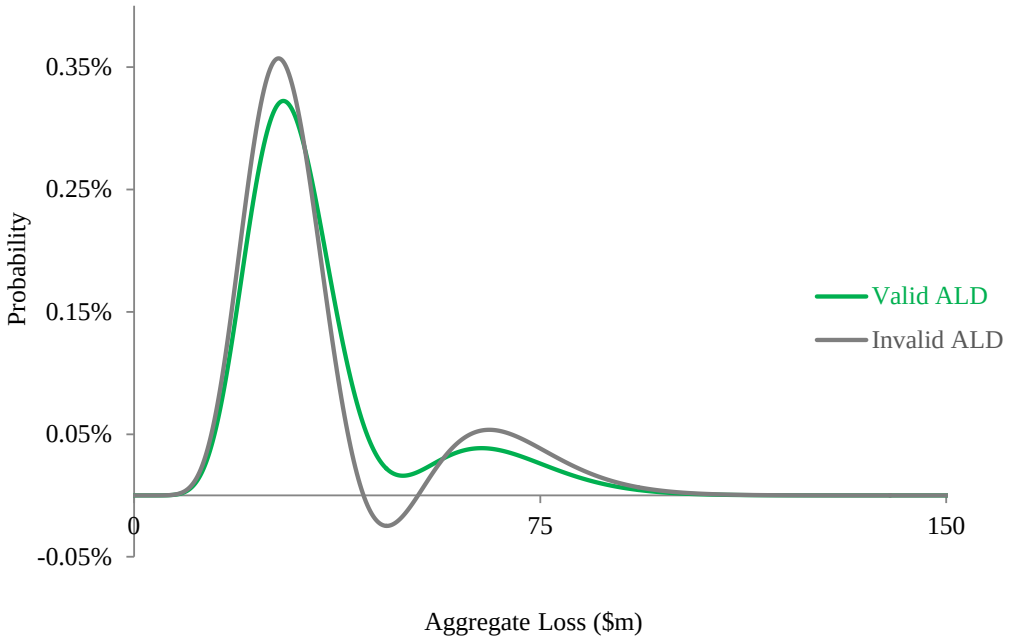


Figure 5.8 Valid and invalid ALDs (Model 4.5) Uniform *covariance coefficients* (i.e. across all classes) of 20 and 30 per cent result in *valid* and *invalid ALD*s respectively (i.e. maximum permissible *covariance coefficient* lies within this range). Data based on Ponemon Institute (2012a–i, 2013a–j, 2014a–k), with costs inflated to 2016.

According to Figure 5.8, the maximum permissible *covariance coefficient* lies between 20 to 30 per cent (corresponding to valid and invalid *ALD*s, respectively). Using trial and error, this maximum is determined as ~ 0.24. Whilst this is true for Models 4.4 and 4.5 considered in this chapter, it may not be the case for other *ALD*s that have been derived in a similar fashion (i.e. using 4.62). Despite this, however, it can be noted that the example *ALD* provided by Wang (1998, sec. 12), based on the same technique, appears to exhibit a similar bimodal feature.

One of the key differences between Model 4.4 and 4.5 is that the 'tail' of the discretised *ALD* for Model 4.5 is less sensitive to changes in the coefficient, κ. For instance, if this parameter is set to 0% for class *D* (i.e. sensitivity 4), then other classes can enjoy a coefficient as high as 85%–90% before the *ALD* becomes invalid. This compares to the maximum permissible value of only 50%–60% in the case of Model 4.4.

5.3.3 Risk-adjusted *limit factors* (Models 4.3, 4.5–4.6)

This section considers risk adjustments (*variance principle*, *PH transform* – §4.2.2.2) for Models 4.3, 4.5–4.6; types of limits (per-loss, per-occurrence), and risk and correlation parameters; in particular:

- Loss count is defined as previously: *Poisson* with mean 10 – Models 4.3 (*CR*) and 4.5; *MNB*(**10**,**1**,0.09) – Model 4.6; *Poisson*(10) is also used for the *PH transform*
- Risk parameter w (4.17, 4.20) is calibrated to achieve a risk-adjusted (to) mean *LAS* ratio of 1.05 (*low*), 1.25 (*medium*), and 1.50 (*high* 'risk environment'), at a *per-occurrence* limit of \$10m (Model 4.3, *PH transform*) and a *per-loss* limit of \$2.5m (Models 4.5–4.6)
- *PH transform* follows 4.20: with $\pi_{PH}(S;b,w) \approx \pi_{PH}(N;\sqrt{w})\ \pi_{PH}(X^{(b)};\sqrt{w})$ where $S, N,$ and X (*Weibull* and *lognormal*, fit to *class E* using *MLE*) denote *LAS*, loss count, and *limited severity* variables respectively (given limit $b>0$ and parameter $w \geq 1$); this assumes equal confidence can be placed on assumed loss count and severity, as described by Wang (1999b: 955)

Variance principle adjustments rely on analytical and computational results for first and second-order moments (4.3, 4.7) of a *spliced* limited severity variable (based on 4.30). *PH transforms* are based on algorithmic integration using Vose (2019) software; for variable X with *Weibull cdf*: $X \sim Weibull(a,b)$ (Table D.3: D.7), $\pi_{PH}(X;b,w) = \mathrm{E}Y^{(b)}$, where $Y \sim Weibull(aw^{-1/b},b)$ is considered. Parameters (for each method, model) are summarised in Appendix D.6.

Figure 5.9 illustrates *discount factor* curves (*base limit*, \$100m) and associated gradients, followed by Table 5.6 which summarises the range of *low*–*high* risk-adjusted *limit factors* (*base limit*, \$1m) for different *per-occurrence* and *per-loss* limits.

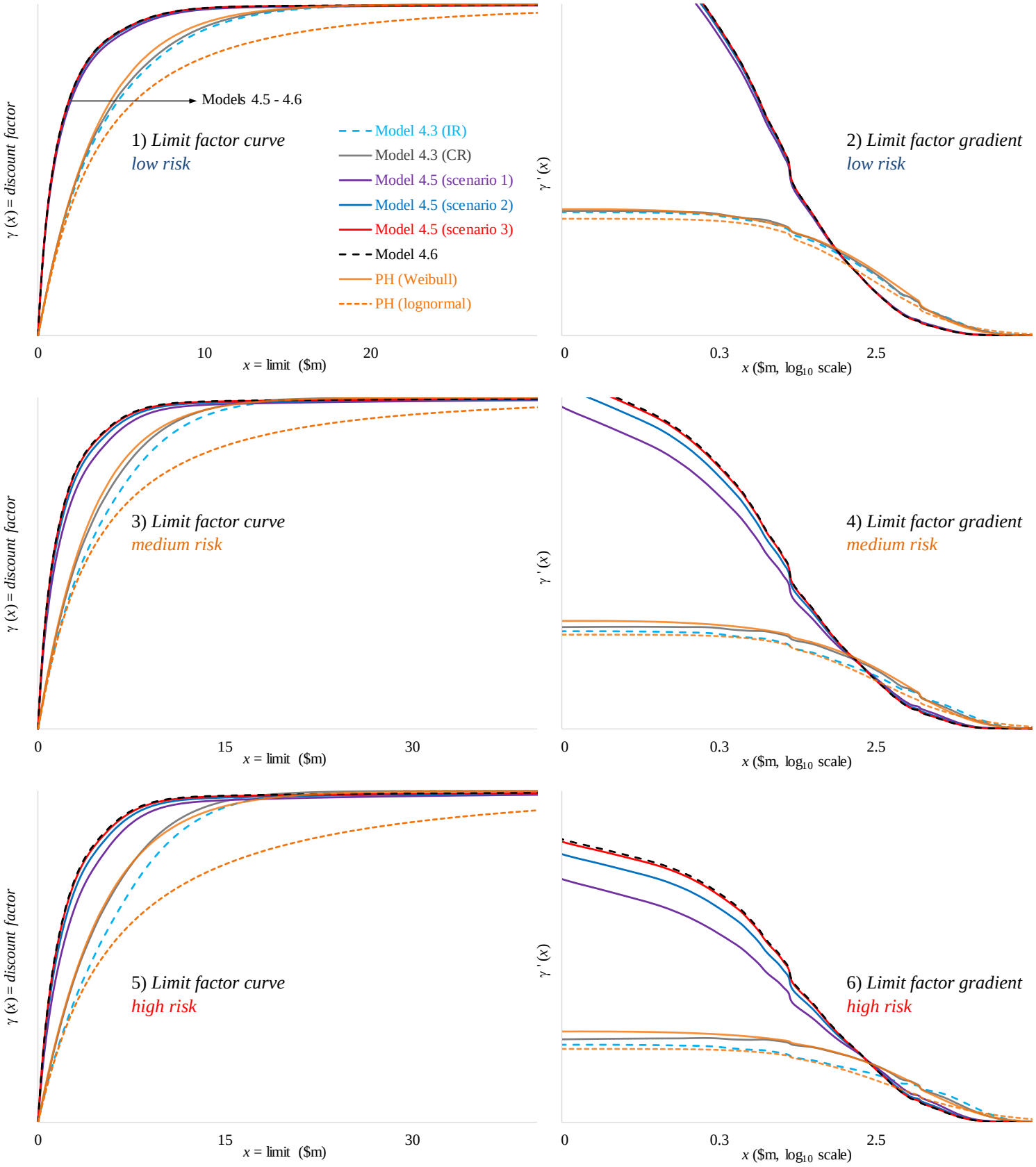


Figure 5.9 Limit factor and gradient curves Base limit: $100m. Model 4.3 (*CR*) in *low* (1–2), *medium* (3–4), and *high* environments achieves a *risk margin* of 5% at $10m, $100k, and $10k limits, respectively, based on *variance principle* which also applies to Models 4.5–4.6. *PH transform* applies to a compound *Poisson-Weibull* and *lognormal* model (fit to costs: Ponemon Institute (2012a–i, 2013a–j, 2014a–k, 2015g), inflated to end of 2016 year). Random variable loss count assumed to follow a *Poisson cdf* with mean 10 for all *CR* models and a deterministic value of 10 for Model 4.3 (*IR*).

		← Limit →						
	Model	*$1m*	*$2m*	*$5m*	*$10m*	*$15m*	*$20m*	*$100m*
Per-occurrence	4.3 (*IR*)	1	1.75 - 1.81	3.04 - 3.59	3.89 - 5.51	4.14 - 6.33	4.20 - 6.58	4.21 - 6.65
	4.3 (*CR*)	1	1.76 - 1.86	3.05 - 3.64	3.86 - 5.13	4.08 - 5.65	4.13 - 5.79	4.14 - 5.83
	Weibull *	1	1.77 - 1.83	3.14 - 3.55	3.91 - 4.88	4.09 - 5.35	4.13 - 5.52	4.15 - 5.61
	Lognormal *	1	1.73 - 1.80	2.93 - 3.35	3.69 - 4.65	4.00 - 5.34	4.15 - 5.75	4.40 - 6.85
Per-loss	4.5 (1)	1	1.37 - 1.53	1.75 - 2.31	1.89 - 2.76	1.91 - 2.82	1.91 - 2.83	1.92 - 2.90
	4.5 (2)	1	1.37 - 1.50	1.73 - 2.15	1.86 - 2.48	1.88 - 2.53	1.88 - 2.54	1.89 - 2.58
	4.5 (3)	1	1.36 - 1.49	1.72 - 2.09	1.85 - 2.37	1.86 - 2.40	1.86 - 2.41	1.87 - 2.44
	4.6	1	1.36 - 1.49	1.72 - 2.08	1.85 - 2.35	1.86 - 2.38	1.86 - 2.39	1.87 - 2.42

Table 5.6 Risk-adjusted limit factors Base limit: $1m. Loss count: all *CR* models (*Poisson*, mean 10); *IR* (deterministic, 10). *Variance principle* (Models 4.3, 4.5–4.6); *PH transform*: *CR* model with same risk parameter for both loss count (i.e. *Poisson*) and *severity* (*orange) *cdf*. Risk parameters (each method, model) calibrated to achieve 5%, 25%, and 50% risk margin (i.e. *low–high*, corresponding to each range of *limit factors*) at $10m (*per-occurrence*) and $2.5m (*per-loss*) limits. Underlying cost data: Ponemon Institute (2012a–i, 2013a–j, 2014a–k, 2015g), inflated to end of 2016 year.

Key observations relating to Figure 5.9 and Table 5.6 include:

- *PH* (*Weibull*) *limit factors* are closely aligned to (variance-adjusted) Model 4.3 (*CR*), as is the case for Models 4.5 (scenario 3) and 4.6; *PH* (*lognormal*) and Model 4.3 *ILF*s crossover at a limit between the $15m-$20m (due to the underlying *cdf*s, Figure D.3)
- *Variance principle* risk-adjusted *limit factors*, in this case, are generally consistent (i.e. positive and decreasing gradients, which is always the case for *PH*), although a subtle initial increase can be seen for Model 4.3 (i.e. closing the gap between *CR* and *PH Weibull* in *medium–high* risk, Figure 5.9: 4, 6)
- Increasing the risk parameter leads to a greater risk adjustment at higher limits than lower limits for a given model (i.e. *discount factor* reduces, whilst *ILF*s increase at limits greater than $1m); a similar effect can be achieved through the correlation parameter in Models 4.5–4.6 (although this is partially offset by equalising risk margins at the $2.5m limit)

Attention is now turned to risk-adjusted *LAS*s and associated gradients for Model 4.5 Scenarios 1–3, as they relate to a *compound Poisson* model with *Poisson* parameter and secondary mixed severity *cdf*.

For this, model 4.5 is specified in terms of 4.63, 4.79–4.80 as follows:

- 4.63, Model 4.5 *LAS*: $S = S_1 +, \ldots, +S_4$; respective *LAS*s for *A*–*D*: $S_1, ..., S_4$ (i.e. $m = 4$): constant *covariance* coefficient, scenario $r = 1, 2, 3$: $\kappa_r = 0.05(r-1)$,
- 4.79, variance-adjusted *LAS*: $\pi_r = \mu + w(\sigma^2 + \kappa_r C_v)$ with $w > 0$; covariance term: $C_V := 2\sum_{i<j} \mathrm{ES}_i \mathrm{ES}_j$
- 4.80, mixed survival: $S_Y(b)$ based on *spliced-severity* variables, $X_1, ..., X_4$, for respective classes (*A*–*D*); per-loss limit: $b > 0$; and *Poisson* parameter (*A*–*D*): $\lambda = 10$

In this case, $\frac{\tilde{C}_V - C_V}{C_V}$ (i.e. percentage difference described in respect of 4.80) grows from 0.06% to 11% between the limits (\$10k, \$10m), but only increases by a further 1% between (\$10m, \$100m). Clearly, $\lim_{w \to 0^+} \pi_r = \mu$, $r = 1, 2, 3$, which is why (*low*-risk) *limit factor* (and *gradient*) *curves* for Scenarios 1–3 (purple, blue, and red respectively) are virtually indistinguishable (Figure 6.13, 1), but appear to deviate in the *medium* (3) and (more so) *high* risk (5). Moreover, in this case, σ^2 and C_V are initially (at low limits) *concave-up* (this is certainly the case $\forall b$ s.t. $\mathrm{E}Y^{(b)} = b$, since $\frac{dC_r}{db}, \frac{d\sigma^2}{db} \propto b$, 4.80: Figure D.4). Given this, and the fact that greater weight is placed on σ^2 and C_V as w increases (4.79), this would explain why gradients (scenarios 1–3) exhibit a decline that transitions from being steep (*low* risk), to being relatively gentle (*high* risk), as the risk parameter increases.

Without digressing too far, this trend (i.e. *upward concavity*, C_V, σ^2) continues until underlying severity *cdf*s (i.e. for *A*–*D*) have gathered sufficient probability mass below the corresponding limit, b (which marks the respective inflection points in C_V, in this case, \$450k, and σ^2, \$1m). Compared to other classes, *B* has a severity *cdf* that accumulates the greatest such mass early on (i.e. due to it having a greater probability of relatively smaller severities). In this way, it is the first class that serves as a countermeasure against the (initial) *concave-up* nature – theoretically analogous to the effect of diversification associated with heterogeneity. Similarly, its *cdf* is also the first to exhibit a diminishing contribution towards the marginal increase in C_V and σ^2 as the limit increases even further (e.g. at $b = \$450\text{k}$, severity *cdf* for *B* is around 80%–90%; compared to 40%–50% for other classes).

Returning to the point at issue, π_1, by definition (4.79), is independent of C_V, but still depends upon μ and σ^2; π_2 and π_3 depend not only upon these terms, but also C_V. Whilst μ may be relatively 'well-behaved' in terms of *consistency properties*, this is not necessarily the case for π_r, $r=1,2,3$, due to dependence upon one (or both) of the terms σ^2 and C_V. In particular, the risk adjustment associated with Scenario r is $\pi_r - \mu = w(\sigma^2 + 2\kappa_r C_V),. r=1,2,3$, all terms defined as previously. As such, the shape of the corresponding *limit factor curves* is influenced by that of σ^2 (i.e. through w) and, in the case of Scenarios 2–3, C_V (i.e. through w and κ_r). Since σ^2 and C_r are initially *concave-up*, there will be a tendency for *limits factors* to exhibit a similar pattern should the variance parameter, w, be large enough. In this regard, w (*high* risk, s.t. risk margin, $m=50\%$, at limit \$2.5m) does not appear to cause any issues, however, 'stress testing' w s.t. $m=500\%$ reveals the effect in Figure 5.10.

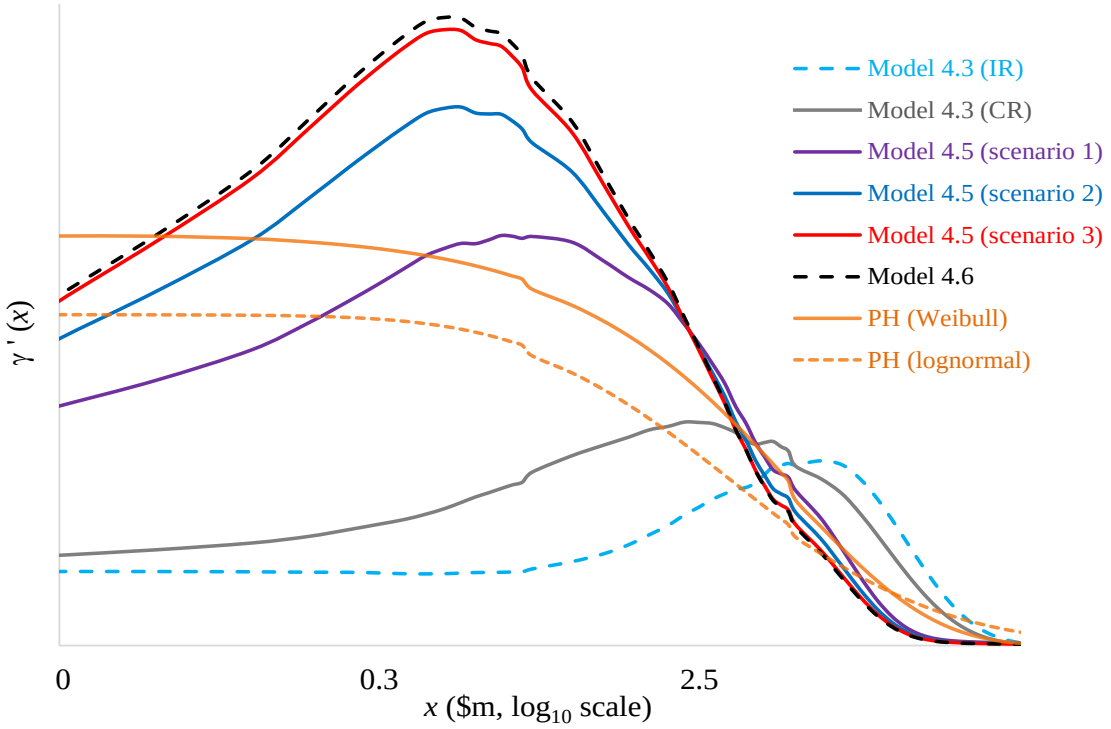


Figure 5.10 Gradient curves (risk parameter stress test) 500% *risk margin* at limits \$10m (Model 4.3 -*variance principle*; *PH transforms*) and \$2.5 (other models - variance principle). \$100m base limit; *PH transform* applies to a compound *Poisson-Weibull* and *lognormal* model (fit to costs: Ponemon Institute (2012a–i, 2013a–j, 2014a–k, 2015g), inflated to end of 2016 year). Loss count = 10 for *IR* (and *Poisson* parameter for *CR*).

Figure 5.10 illustrates the case where w is stressed to the point where gradients for scenarios 1–3 reflect the combined effect of underlying gradients associated with σ^2 and C_V (Appendix D.6). As can be seen, *PH transforms* remain resilient in terms of *consistency properties* (i.e. decreasing gradients); whilst Model 4.3 (*CR*, *IR*) has increasing gradients, which violates these properties (*CR* earlier than *IR*, due to greater variance associated with the former). There is still remarkable similarity between scenario 3 (i.e. $\kappa_3 = 0.1$) and Model 4.6 in terms of gradients.

The mathematical relationship follows 4.81 with $S^* = S_1^* + \ldots + S_4^*$; marginal compound *LAS*s for *A*–*D* with respective primary *negative binomial* parameters $(\lambda, c) = (10,1)$; σ^2, μ based on Model 4.5 scenario 3; and $\kappa^* = 0.09$. In particular, $\mathrm{Var}S^* - \mathrm{Var}S = 0.1\mu^2 - 0.11C_V$ (i.e. $\frac{\mathrm{Var}S^* - \mathrm{Var}S}{\mathrm{Var}S^*} \sim 0$; C_V based on scenario 3 as before).

This supports Figure 5.9 (1, 3, 5), although the effect is rather subtle and difficult to see as *limit factor curves* for Scenario 3 (red) and Model 4.6 (dark grey, dashed) appear to coincide with one another. Refer to Halliwell (2009) for generalised extensions relating to mixed *cdf*s in the context of *CR* models with and without heterogeneity. In Figure 5.9 (2, 4, 6), gradients also appear to be fairly similar to one another (differences can be observed more easily in this case, due to the log-scale used for limits along the *x-axis*, especially in the case of Figure 5.10).

In summary, due to common dependence on terms such as σ^2 (i.e. associated with π_1) and C_V, *limit factor curves* and gradients for Model 4.5 (in particular Scenario 3) and Model 4.6 are similar in shape as illustrated (Figure 5.9: 1–6; Figure 5.10). Previous comments regarding Model 4.5 *consistency* properties (*low*–*high* risk) also apply to Model 4.6.

5.3.4 Validations and investigations

As outlined in §5.1, this section reviews, to a practicable degree, models and results (§5.2–§5.3) in terms of consistency, accuracy, and reasonableness.

5.3.4.1 *Discretisation versus first-order derivative*

As mentioned previously, *spliced-severity cdf*s form the basis of mean *LAS* calculations based on Model 4.2 as well as *ALD*s based on Models 4.3–4.6. *Goodness of fit* (§5.2.3) has already been considered for *large-loss cdf*s at selected thresholds. It is now of interest to assess *spliced-severity cdf*s in terms of Models 4.1–4.6, by comparing *cdf*s implied by *LEV*s (based on Models 4.1–4.2) to *discretised cdf*s used in the *FFT* algorithms that underlie Models 4.3–4.6. In Table 5.7 discretised severity *cdf*s for *A*–*E* (underlying *ALD*s in Figure 5.5 and Figure 5.6) are compared with *cdf*s using the first-order derivative of *LEV*s based on Model 4.2 (i.e. 4.12).

The term *discretisation* (Appendix C.1) in Table 5.7 refers to discretised severity *cdf*s that formed the basis of *FFT* used to determine the *ALD*s represented in Figure 5.5 (i.e. for *classes A*–*D*) and Figure 5.6 (i.e. for *class E*). Recall that a separate *discretisation* was used for each. This table shows that, for each class, there is close correspondence between the moments of the *spliced-severity cdf*, based on discretisation, and moments of *cdf*s that have been derived using the first derivative of *LEV* over a range of limits.

Class	Method	Mean	Std Dev	Skewness	Kurtosis
A	*LEV* derivative	705 820	844 039	3.875	30.407
	Discretisation	704 985	844 115	3.876	30.408
B	*LEV* derivative	212 040	384 233	12.467	407.459
	Discretisation	211 924	384 384	12.443	405.979
C	*LEV* derivative	1 034 841	1 294 057	4.418	42.267
	Discretisation	1 034 591	1 295 137	4.409	42.076
D	*LEV* derivative	1 486 589	2 028 142	2.422	9.755
	Discretisation	1 498 916	2 069 401	2.449	9.819
E	*LEV* derivative	3 585 582	2 656 221	4.443	23.870
	Discretisation	3 583 591	2 655 215	4.456	23.940

Table 5.7 Discretised severity cdf versus first-order derivative of ILF *Discretisation* (*truncation*, *span*) - *A*–*D*: ($96.2*m*, $23.5*k*), *E*: ($287.1*m*, $70.1*k*). Limits: *A*–*D*: $20*m*; *E*: $80*m*. Underlying data based on Ponemon Institute (2012a–i, 2013a–j, 2014a–k, 2015g) costs inflated to end of 2016 year.

The first-order derivative of *LEV*s are approximated over 4096 limits (with loss of one point, as described shortly), based on discrete multiples of the *span* used for *discretisation* (§5.3.2).

From these, survival functions are evaluated at each discretised severity point, and the moments (summarised in this table) are calculated accordingly.

To illustrate these calculations (to one decimal, in $000s), for example, the first three limits (with increments equal to a *span* of $23.5k) are 0, 23.5, and 47.0 respectively; corresponding *LEV*s (underlying Model 4.2, 4.78) for class *A* are calculated as: 0, 23.5, and 46.6, respectively. Divided differences (4.13) approximate first-order *LEV* derivatives (i.e. survival functions) at limits $0 and $23.5 as $1 \sim \frac{23.5}{23.5}$ and $0.98 \sim \frac{46.6-23.5}{47.0-23.5}$ respectively. The probability at the nearest point that is greater than or equal to $20m will be set equal to one less the sum of probabilities up to this point; probabilities greater than this point are set to zero. Moments are then calculated using the probabilities calculated in this fashion at each of the 4095 points (i.e. using 4096 limits).

The correspondence between the moments of the discretised and *ILF-implied* severity *cdf*s in Table 5.7 not only confirms commonality of *spliced-severity cdf*s in different models, but also serves as a check for Model 4.2 in terms of basic underlying *ILF* theory (§4.2.1, 4.8).

5.3.4.2 *Mean LAS comparisons*

Table 5.8 includes the following entries, which are used to compare mean *LAS*s for Model 4.2, by class, to those based on Model 4.3 (Figure 5.5 and Figure 5.6) and various other approximate methods:

(1) Small	Mean *LEV* for *empirical* severities that fall below the *threshold* (Table 5.2), multiplied by the expected number of small losses for a given total loss count of 10 (i.e. 10 times the threshold percentile, also given in Table 5.2), calculated empirically
(2) Large	Mean *LEV* implied by the *large-loss cdf*, multiplied by the expected number of *large losses* for a given total loss count of 10 (i.e. 10 less the number of expected small losses), based on analytical solutions
(3) Combined	Model 4.2 (spliced *cdf*s): regarded as ‘correct’ values
(4) Model 4.3	Model 4.3 (*CR* models, *FFT*): Figure 5.5 (1, *A*–*D*), Figure 5.6 (1, *E*)

(5) Vose software Vose (2019) built-in *FFT* aggregate function

(6) Empirical Model 4.2 (empirical *cdf*s): entry (1) plus observed (large-loss) mean

	Mean *LAS*	*A*	*B*	*C*	*D*	*E*
	(1) Small (Model 4.2)	4 048 433	615 060	4 558 041	9 217 511	19 497 087
	(2) Large (Model 4.2)	3 009 651	1 505 338	5 790 314	5 648 380	16 358 737
	(3) Combined (1+2)	7 058 084	2 120 398	10 348 355	14 865 891	35 855 824
Approximations	(4) Model 4.3	7 049 848	2 119 242	10 345 908	14 989 156	35 835 908
	% diff = [(3) - (4)] / (3)	*0.12%*	*0.05%*	*0.02%*	*-0.83%*	*0.06%*
	(5) Vose Software	7 076 448	2 111 060	10 287 907	15 015 410	35 861 806
	% diff = [(3) - (5)] / (3)	*-0.26%*	*0.44%*	*0.58%*	*-1.01%*	*-0.02%*
	(6) Empirical data	6 942 036	2 122 505	10 294 099	14 909 599	35 884 272
	% diff = [(3) - (6)] / (3)	*1.64%*	*-0.10%*	*0.52%*	*-0.29%*	*-0.08%*

Table 5.8 Accuracy of Model 4.3 and other approximations 1), 6) based on *data*; 2) reflects *large-loss cdf*; 4) *FFT* (*truncation*, *span*) - *A*–*D*: ($96.2*m*, $23.5*k*), *E*: ($287.1*m*, $70.1*k*); 5) mean, aggregate *FFT* Vose (2019) functions. Loss count: Poisson, mean 10. Limits: $20m (*A*–*D*), $80m (*E*). Cost data: Ponemon Institute (2012a–i, 2013a–j, 2014a–k, 2015g), inflated to 2016.

The same severity limits (i.e. *A*–*D*: $20; *E*: $80m) and mean loss count (i.e. 10), used for Model 4.3, are also used for Model 4.2 in Table 5.8. Key points and observations relating to this table include the following:

- Model 4.2 (*entries 3*, *6*) reconciled with previous *discount factors* (Table 5.4, Table 5.5) – for instance, *B* (entry 3): $\$2.12\text{m} \sim (1-0.2\%) \times \frac{\$169.9\text{m}}{800} \times 10$ (i.e. *spliced discount factor* at $20m, Table 5.5: 0.9981%; applicable mean *LAS*, Table 5.5: $169.6m; loss count, observed and assumed mean: 800 and 10 respectively)
- Model 4.3 (*entry 4*): corresponds to within 1% for *D*, and 0.2% for other classes
- Model 4.3, Vose (2019), and empirical (*entries 4–6*): in this case, the built-in *FFT* function is less accurate than Model 4.3 for *A*–*D* (and less accurate than the empirical estimates for *B*–*D*, *entry 6*), but the most accurate approximation for *E*

It may be possible to improve the accuracy of Model 4.3 even further by discretising severity *cdf*s for each class separately (although this will make it much more difficult to combine *ALD*s using Models 4.4–4.6).

Whilst not as simple as the rounding-method (used here), the '*mean-preserving*' method of discretisation should also achieve greater accuracy in this regard (Klugman, Panjer & Willmot, 2004: 168).

5.3.4.3 *Higher order moments (Algorithm 4.5)*

MC simulation (Algorithm 4.5) is used to determine *ALD*s for *A*–*E* (Figure 5.11, left), which are compared with those based on *FFT* (right) using Model 4.3 (*CR*).

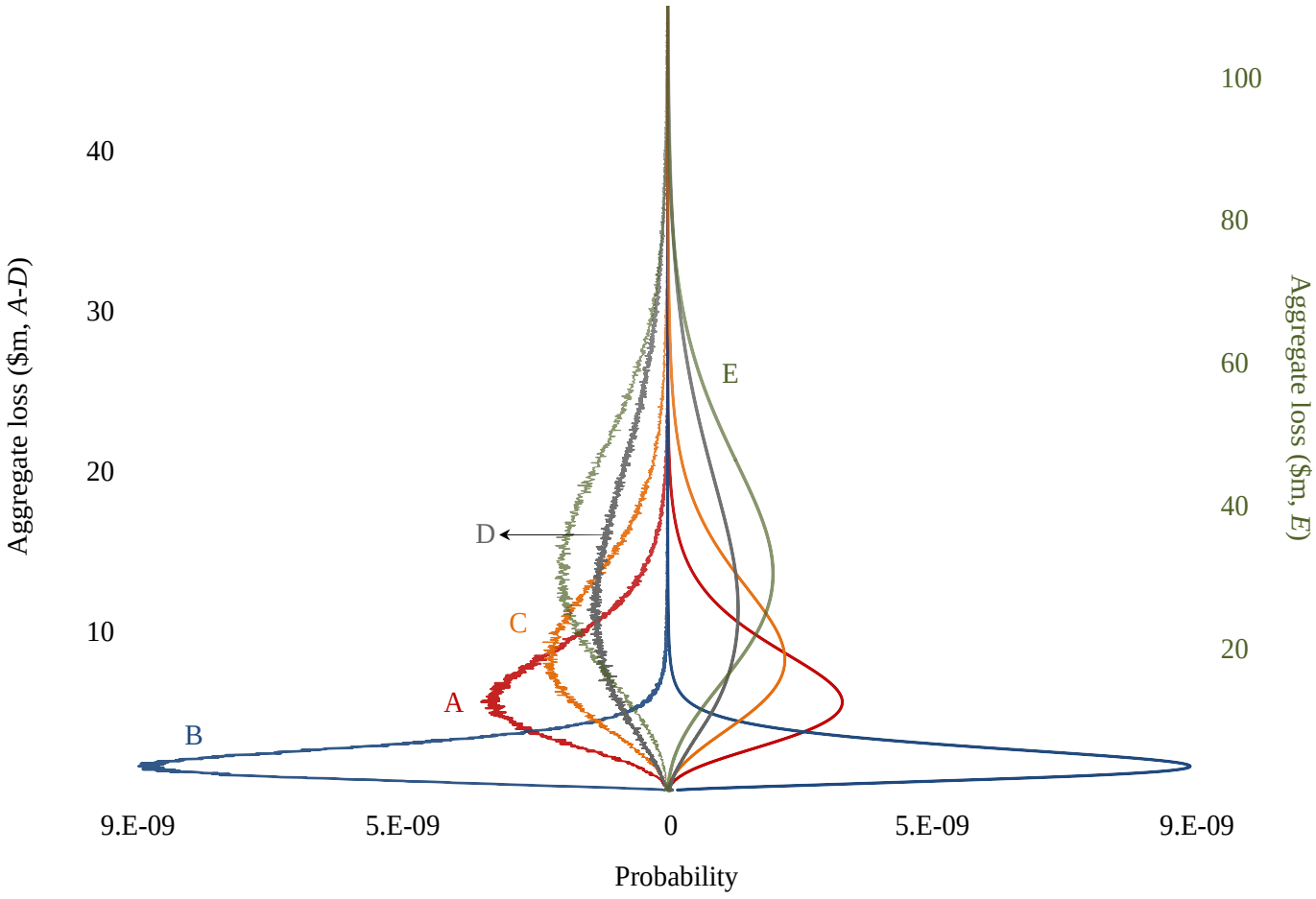


Figure 5.11 ALDs: Monte Carlo versus FFT (Model 4.3, CR) - (*1*) *Left*: *MC* simulation with 500k iterations; (*2*) *Right*: Model 4.3 (*CR*) with *FFT* (*truncation*, *span*) - *A*–*D*: ($96.2m, $23.5k), *E*: ($287.1, $70.1k). Limits: *A*–*D* ($20m), *E* ($80m); *Poisson* loss count with mean 10. Vertical axes - left (*A*–*D*); right (*E*). Underlying data: Ponemon Institute (2012a–i, 2013a–j, 2014a–k, 2015g), costs inflated to year 2016.

*ALD*s for *A–E*, based on *FFT* (Figure 5.11: 2), are copies of previous *ALD*s (Figure 5.5–Figure 5.6), rotated by 90 degrees (for visual convenience, aggregate loss for *E* is placed on a second vertical axis, shown in green on the right).

*Severity cdf*s for *A–D* are discretised separately to *E*, as before (the same limits and *Poisson* loss count assumption are also used, §5.3.2, §5.3.4.2). *ALD*s for *A–E*, based on the *MC* algorithm (Figure 5.11: 1) appear as reflections of Model 4.3 (2) due to their close alignment.

Simulation error associated with the *MC algorithms* is somewhat apparent (1); however, this does not seem to detract from the correspondence that can be seen between means and (standardised) moments (Table 5.9).

Class	Method	Mean	Min	Std dev	Kurt	Skew
A	Monte Carlo	7.080	0	3.485	4.683	0.998
	Model 4.3	7.050	0	3.478	4.666	0.996
B	Monte Carlo	2.137	0	1.397	27.695	2.985
	Model 4.3	2.119	0	1.388	28.599	3.029
C	Monte Carlo	10.362	0	5.238	5.204	1.095
	Model 4.3	10.346	0	5.242	5.250	1.103
D	Monte Carlo	14.895	0	7.961	3.852	0.831
	Model 4.3	14.989	0	8.080	3.875	0.840
E	Monte Carlo	35.806	0	15.710	3.521	0.643
	Model 4.3	35.836	0	15.834	3.556	0.657

Table 5.9 Moments: Monte Carlo versus FFT *MC* simulation with 500k iterations; Model 4.3 (*CR*) with *FFT* (*truncation*, *span*) - *A–D*: ($96.2*m*, $23.5k), *E*: ($287.1, $70.1*k*). Means: $m. Limits: *A–D* ($20*m*), *E* ($80*m*); *Poisson* loss count with mean 10. Underlying data based on Ponemon Institute (2012a–i, 2013a–j, 2014a–k, 2015g), with costs inflated to end of 2016 year.

This is confirmed by the means and moments reported in Table 5.9; it is also reassuring to see similar *CR* features for both approaches (e.g. minimum of 0, as observed previously §5.3.2).

5.3.4.4 *Detecting potential aliasing errors*

In terms of *FFT* used for Models 4.3–4.6, should there be any (non-zero) compound mass at (or beyond) the *truncation* point (in this case, Figure 5.5–Figure 5.6, $96.2m for *A–D* and $287.1m for *E*, §5.3.2) it will simply *wrap around* and reappear (erroneously) at zero, giving rise to what is known as an *aliasing error*.

This has been likened to a year 2000 problem and the 'wagon-wheel' effect and is an issue that can lead to an uplift in the left tail of the *ALD* (more so for heavy-tailed *cdf*s). Techniques to address this include:

- Increasing the *truncation* point (although this must be balanced against associated *discretisation error*)
- Applying a *tilting* operator that commutes with convolutions and increases the tail decay (Shevchenko, 2010, sec. 6.2)

The latter can lead to 'overflow' or 'underflow' – results too large or too small to be represented in computer memory – (Grübel & Hermesmeier, 1999). To detect potential wrap around in the present case, Figure 5.11 is converted to a log-log scale (Mildenhall, 2005: 175) in Figure 5.12 (1–4), and left tails are inspected.

If *wrap-around* errors were an issue for Model 4.3 (*CR*), Figure 5.12 (2, 4), then positive deviations, relative to *MC* simulations (i.e. 1, 3), could be expected in the *left tail* of these *ALD*s. In this case, only *C* exhibits such a positive deviation in Figure 5.12 (2), however, this only occurs at the (left-most) single point, corresponding to the span.

Thus, *wrap around* does not appear to be an issue for any of *A–E* in this case, where limits are suitably low in relation to the truncation point used for discretisation. If this were not the case, then *C* would be especially prone to *wrap around*, followed by *B*, *A*, then *E* (due to relative *tail weight*, §4.2.3.3, associated with selected *cdf*s, Table 5.2).

Whilst subtle in effect (due to the log-scale), the application of this limit results in a small 'spike' in the right tail of the *ALD* for *B* (Figure 5.12: 1, 2).

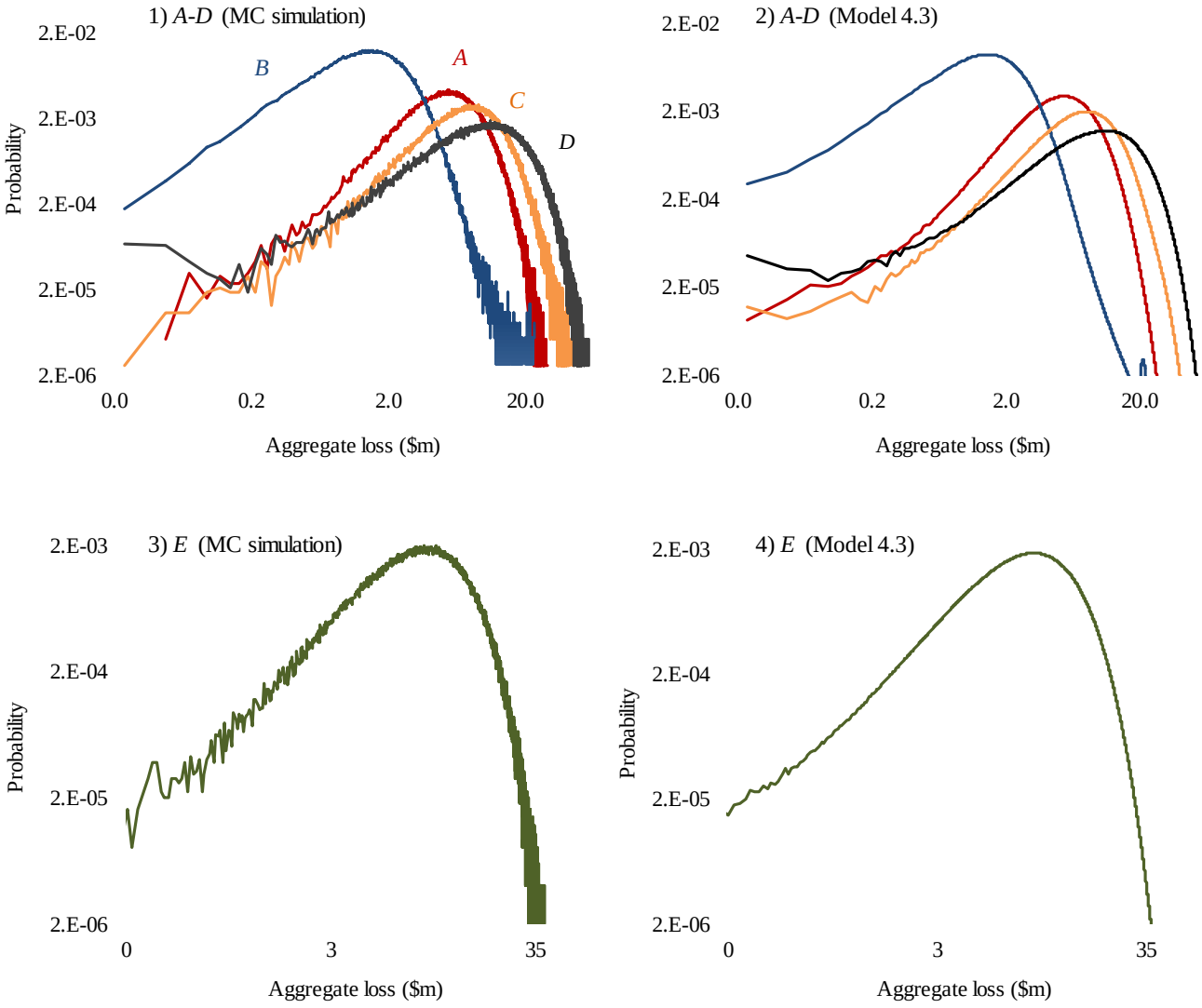


Figure 5.12 ALDs: MC versus FFT (log-log scale) - (*1,3*) *MC* simulation with 500k iterations; (*2,4*) Model 4.3 (*CR*) based on *FFT*. Limits: *A–D* ($20m), *E* ($80m); *Poisson* loss count with mean 10. Costs: Ponemon Institute (2012a–i, 2013a–j, 2014a–k, 2015g), inflated to year 2016.

5.3.5 Reasonability of *limit factors*

To assess reasonableness, *ILF*s in respect of relevant models and classes are compared with insurer *ILF*s (allowing for the effect of inflation and deductibles as required). Examples are provided according to the type of limit that applies.

Per-occurrence limit comparison

The Hanover (2015) filing includes premiums (that vary by size of limit) for *Data Breach* (hereafter, *DB*) and *Additional Expense* (*AE*) coverage for 'services and expenses' which appear to overlap with *A*–*D* (Chapter 3) as follows:

- *A* – forensics, consultations (e.g. audit and assessment)
- *B* – notification, breach restoration, consultations
- *C* – help line, legal, investigations, public relations, identity restoration
- *D* – business interruption (i.e. lost business)

As such, Table 5.10 compares *ILF*s for Models 4.2–4.3 (*CR*, *class E*), with *low*–*high* (*variance principle*) risk adjustments, to *ILF*s based on the Hanover (2015) filing.

	Hannover *ILF* s (by band of annual Turnover, $m)						**Model 4.2**	**Model 4.3**		
Limit ($)	**(0,1]**	**(1 , 2]**	**(2 , 5]**	**(5 , 10]**	**(10 , 20]**	**20+**	***No risk adj.***	***Low risk***	***Medium***	***High***
10 000	1	1	1	1	1	1	1	1	1	1
25 000	2.03	2.05	2.07	2.23	2.48	2.64	2.27	2.50	2.50	2.50
50 000	2.95	3.34	3.75	4.26	4.84	5.36	4.55	5.00	5.01	5.02
100 000	4.91	5.98	7.04	8.15	9.24	10.30	9.09	10.01	10.04	10.08
250 000	9.78	12.58	15.37	18.17	20.96	23.76	22.58	24.89	25.09	25.35
500 000	17.22	22.98	28.73	34.48	40.22	45.99	44.03	48.66	49.47	50.48
750 000	23.48	32.29	41.10	49.90	58.73	67.54	64.19	71.13	72.89	75.08
1 000 000	29.90	41.86	53.80	65.78	77.76	89.70	82.79	91.81	94.79	98.51
1 500 000	37.38	56.04	74.76	93.44	112.13	130.82	116.02	129.15	135.26	142.88
2 000 000	45.14	70.93	96.72	122.54	148.30	174.10	144.28	161.27	171.12	183.42
2 500 000	49.91	83.14	116.39	149.64	182.90	216.15	168.71	189.09	203.08	220.56
5 000 000	75.40	150.80	226.20	301.60	377.00	452.40	246.20	279.61	314.74	358.62

Table 5.10 Insurer ILF comparison (per-occurrence limits) Hanover (2015) *ILF*s based on premiums filed under *Data Breach* coverage (and 40% marginal loading for *Additional Expense*). Underlying costs for modelled *ILF*s: Ponemon Institute (2012a–i, 2013a–j, 2014a–k, 2015g), inflated to year 2016. Font colour indicates Hanover (2015) *ILF*, at a given limit, with closest match to *ILF*s based on Models 4.2 - 4.3.

Hanover (2015) *ILF*s in Table 5.10 are summarised as a matrix of values: different rows and columns represent limits and bands of annual company turnover (in $m, 2015) respectively, and adjacent to this are *ILF*s based on Models 4.2–4.3. The smallest absolute difference between insurer *ILF*s and the *ILF* for each model, by limit, is indicated with common font colour (e.g. 377 at $5m limit, Hanover (2015), is the insurer *ILF* closest to

358.62, Model 4.3 *high-risk* at that limit; likewise 174.10 is closest to 171.12 at the $2m limit; as is 89.70 to 91.81 at the $1m limit, etc.). It can be noted that model *ILF*s are different to those previously compiled in Table 5.6 (due to a different *base limit*).

In accordance with this filing, *ILF*s for *DB* coverage are multiplied by a variable factor (that increases with the limit) to incorporate *AE* coverage. This coverage overlaps with several covers listed previously (i.e. alongside *C*), and certain others that fall outside the scope of *E* (i.e. *A*–*D*). Key assumptions underlying the present comparison can now be stated as follows:

- 'Out of scope' covers comprise 40% of the *AE* loading which otherwise relates to a number of covers listed previously alongside *C*; thus, instead of multiplying by a given *AE* factor of $y \geq 1$, at some limit, $0.6(y-1)+1=0.6y+0.4$ is used
- Filed rates came into effect during the 2016 year; under the premise that insurer and model *ILF*s relate to the same period, no (further) inflation adjustment is made
- *ILF*s relate to ground-up coverage (i.e. $0 excess); expense, profit, and other 'non-risk' adjustments and loadings are ignored

Contrary to the concept of reducing marginal increases, associated with *consistent limit factors*, Hanover (2015) *DB* (implied) *ILF*s (i.e. based on filed 'base premiums'), for example, with or without *AE* adjustment, do not produce *consistent ILF*s across all limits. Other observations include the following:

- Turnover bands appear low in relation to the size of underlying costs (the reason for this is explained shortly)
- In terms of the *variance principle* (as it relates to Model 4.3), the equivalent risk-adjustment parameter, in relation to Hanover (2015) *ILF*s, for a given turnover band, is one that generally increases with the size of the limit
- Model 4.3 appears to produce reasonable *ILF*s, in relation to Hanover (2015) *ILF*s, for the $20m+ turnover band at lower limits (i.e. 250k–$1m) with *no* or *low*-risk adjustment, and at higher limits ($1.5m–$2.5m) with *medium*–*high* risk adjustment (the equivalent *risk margin* at limit $5m is over 100%, double that assumed for *high* risk)

One of the (potentially material) flaws associated with this comparison pertains to the type of risks to which (*DB*, *AE*) premiums are related. In particular, this program (according to its name) relates to religious institutions which are likely to have a different risk profile to many of the organisations associated with underlying data (Chapter 3). This would explain why the turnover bands appear to be low in relation to the costs in *class E*. For instance, the mean *LEV* (Table D.4, Model 4.3, \$5m limit, divided by mean loss count 10) would imply, for institutions with \$20m turnover-year, a 'pure-risk' rate of \$0.16 ($\sim \frac{3.2}{20}$) per \$1 turnover-year. In comparison, premium rates in the order of \$0.01–\$10 per mille turnover-year might be expected for such coverage (indeed, the actual rate filed by Hanover (2015) for the \$20m turnover band was \$0.07 turnover-year). The following comparison is somewhat more consistent in this regard.

Per-loss limit comparison

Table 5.11 compares *ILF*s for several major league insurers to those based on Models 4.5–4.6 (*low*–*high* risk). The portion relating to insurers formed part of a 'competitor comparison' (in relation to 2015-year *ILF*s) filed by Cresenzi & Alibrio (2016) on behalf of *ACE* (Chubb, 2017); *ILF*s at the \$100m limit are extrapolated, as is described shortly. According to Fitch Ratings (2016), *AIG* and *Chubb*, whose cyber-insurance products are included in this comparison, are the two largest writers of cyber-insurance with a market share, based on direct written premium for the year 2016, of 34% (22% and 12% respectively). As for the previous comparison, costs covered by the insurance products underlying this table correspond with *A*–*D* categories (Chapter 3). Since insurer *ILF*s incorporate a *base retention* and *base limit* of \$10k and \$1m respectively, model *ILF*s are derived using 4.28 with $v = 1.025$ (based on inflation used for *E*, year 2015, Chapter 3), $d = \$10\text{k}$, and $a = \$1\text{m}$. It can be noted that with the same risk-adjustment parameter a different risk margin will be generated.

As mentioned, insurer *ILF*s at the \$100m limit, in this table, have been estimated separately (the competitor comparison only goes as far as the \$10m limit). For this, *Riebesell* (i.e. '*power*') *curves*, giving a third (and final) representation of risk-adjusted *ILF*s in the present research, are determined for each insurer. In particular, from 4.21, with $a = \$1\text{m}$, $b = \$100\text{m}$, and $\gamma_{b,a} := \gamma(b; a, w)$: $w = \ln \gamma_{c,a}(\ln c - \ln a)$ where $c = \$10\text{m}$ (i.e. $\gamma_{c,a}$ is given) and w is the insurer specific *Riebesell* parameter.

	←			*Limit*			→
Insurer	***$1m***	***$2m***	***$3m***	***$4m***	***$5m***	***$10m***	***$100m****
Chubb	1	1.29 - 1.50	1.49 - 1.89	1.65 - 2.21	1.77 - 2.50	2.20 - 3.60	4.84 - 12.96
AIG National	1	1.50	1.88	2.14	2.35	3.04	9.24
Travelers	1	1.42	1.62	1.83	1.99	2.73	7.44
Philadelphia	1	1.58	1.98	2.27	2.47	3.15	9.92 - 9.92
ACE	1	1.30 - 1.50	1.50 - 1.89	1.65 - 2.22	1.78 - 2.51	2.21 - 3.62	4.88 - 13.10
Overall range (A , B)	1	1.29 - 1.58	1.49 - 1.98	1.65 - 2.27	1.77 - 2.51	2.20 - 3.62	4.84 - 13.10
Models 4.5 - 4.6							
Low risk	1	1.37	1.56 - 1.57	1.66 - 1.68	1.72 - 1.75	1.84 - 1.89	3.39 - 3.57
(Median - A) / (B - A)	-	*27%*	*15%*	*3%*	*-5%*	*-24%*	*-16%*
Medium risk	1	1.43 - 1.45	1.67 - 1.73	1.81 - 1.90	1.90 - 2.02	2.09 - 2.31	4.37 - 5.32
(Median - A) / (B - A)	-	*52%*	*43%*	*33%*	*26%*	*0%*	*0%*
High risk	1	1.49 - 1.54	1.78 - 1.90	1.96 - 2.13	2.07 - 2.30	2.33 - 2.73	5.44 - 7.48
(Median - A) / (B - A)	-	*77%*	*71%*	*64%*	*57%*	*23%*	*20%*

Table 5.11 Insurer ILF comparison (per-loss limits) Insurer comparison: 2016 ACE *SERFF* filing - *Chubb Enterprise Risk Management Cyber* and *Digitech* products (Cresenzi & Alibrio, 2016), with reference to (2015 year) *SERFF* filings by: *AIG* (*Speciality Risk Protector*) [AGNY-130104025], *Travelers* (*Cyber-Essentials*) [TRVD-130748646], *Philadelphia* (*Cyber-Security Liability*) [PHLX -G128091742], and *ACE* (*MPL Advantage*) [ACEH-125807939]. *$100m: *ILF*s estimated with *Riebesell* curve (implied at $10m limit). *Base limit*: $1m; retention: $10k. Shading: model range within insurer range (*A*:*B*):=(min, max); partial if ranges overlap. 'Median': model *ILF* range. Ponemon Institute (2012a–i, 2013a–j, 2014a–k), inflated to year 2016 (*ILF*s: adjusted to 2015).

The self-same parameters, which range (0.34, 0.56), are then reapplied at the $100m limit to determine insurer *ILF*s at that limit (Table 5.11). It is worth noting that estimated *ILF*s at $100m limits are likely to be overstated, given that there is evidence that implied ws decrease (slightly) as the limit increases (e.g., with $b_1 = \$5\text{m}$, implied ws range 0.35–0.57). A *power curve* with $w \approx 0.14$ would be required to achieve *low*-risk modelled *ILF*s at the $100m limit, in comparison to $w \approx 0.17$ (*medium* risk) and $w \approx 0.19$ (*high* risk). Indicated with green shading are model *ILF*s that fall (entirely) within the range of insurer *ILF*s at a given limit. All models can be seen to achieve this – the *high* risk does this across every (given) limit (i.e. $1m–$10m). Given the data limitations associated with underlying (publicly available) data, manipulations, adjustments, simplifying assumptions, and what appears to be somewhat narrow *insurer ranges*, it is reassuring with regard to any such alignment between *model ranges* and *insurer ranges*.

Chapter 6

Conclusions, Recommendations

> *"That is the way to learn the most, that when you are doing something with such enjoyment that you don't notice the time passes."*
>
> Einstein (1879–1955), cited by Lawson (2004: 12)

This research has explored key issues associated with cyber-risk and related pricing models through empirical analyses and applications of *spliced-severity* and aggregate loss models – the main aim was to investigate different types and levels of risk adjustment and correlation in terms of (pure-risk, cyber-insurance) *Increased Limit Factors*, *ILF*s. This chapter evaluates the primary objectives (§1.2), highlights key contributions, limitations, and conclusions, and, in finality, makes recommendations in regard to future research.

6.1 Evaluation of objectives

For reference, each objective is restated, followed by a summary of what has been considered.

Objective 1 (a-c) "To [a] review relevant sources of information and data, and, based on this, identify sources most suitable for [b] deriving severity and [c] aggregate loss distributions and determining implied *ILF*s"

1a) Data sources: *review*, *identify*

- Twenty sources were compared in terms of practical factors associated with data quality (refer to Table A.2 for quick reference); *primary* (Ponemon Institute, 2019) and associated *secondary* (e.g. *SERFF* (NAIC, 2019), OECD (2018)) sources were identified and data was extracted

- Validity (e.g. consistency, completeness, accuracy) of data (i.e. data breach costs, *primary*) was considered to a practicable degree; the effect of basic inflation adjustments was assessed; key limitations were disclosed; and applicability in terms of analogous cyber-coverage was considered

1b) Severity *cdf*s: *derive*

- *Large* (severity) *cdf*s, selected using the corrected Akaike, AIC^c, in terms of the *Kullback-Leibler* distance estimate (candidate models: Appendix D.1) were subject to a *Kolmogorov-Smirnov*, *KS*, test (5% critical) to determine *splicing points* on the basis of *goodness of tail fit*

- Model (90%) confidence sets were estimated for *cdf*s identified in this way (*Burr*: *B*, *C*; *Weibull*: other classes), left-truncated at selected percentiles (75^{th}–92^{nd}), and based on 10k bootstrap samples – these considered *ILF consistency*

- This reinforced *light-tail* selections in respect of certain classes (i.e. *D*, *E*: *Weibull*) – subject to the degree of uncertainty associated with model selection (greatest for *D*, due to its truncation); indicated alternative models (e.g. *log-gamma*, *fatigue*, *GEV*, *log Laplace*, and *Pearson*); and assessed *KS-test* performance in relation to an equivalent (i.e. 5% critical) *Anderson Darling*, *AD*, test

1c) *ALD*s and *ILF*s: *derive*, *determine*

- Various aggregate loss models were considered: *Collective Risk* , *CR*, models for determining *ALD*s based on *Fast Fourier Transform*, *FFT*, included: *compound Poisson* (Model 4.3 with primary *Poisson* loss count, secondary *spliced cdf*s); *correlated* aggregate loss and count models (Models 4.5–4.6, based on *characteristic functions*, *cf*s, and related transforms); and *deterministic* analogues (i.e. Models 4.2–4.4, crudely dubbed '*Individual Risk*', *IR*)

- These incorporated different types of limits (defined in terms of '*per-loss*': *A–D*; and '*per-occurrence*': *E*), and, in respect of Models 4.4–4.5, different 'correlation scenarios'

- *ILF*s were determined in respect of both severity *cdf*s and *ALD*s; as mentioned, these incorporated different types (i.e. *variance principle*, *Proportional Hazard* – PH – *transforms*, and, for comparisons with insurer *ILF*s, *Riebesell* or *power* curves) and levels of risk adjustments (based on implied risk margins at given limits)

Objective 2 (a-b) "To model and explore [a] key attributes associated with underlying loss distributions and [b] the effect of correlation on these and associated risk adjustments"

2a) Distributions: *key attributes*

Severity cdfs and *ALD*s were considered in terms of the nature of underlying costs:

- *A*, *C*: similar underlying cost types (and, therefore, distributions); although larger moments and a heavier tail for the latter were associated with distinguishing cost types (e.g. legal fees)

- *B*: costs associated with regulatory requirements were speculated to lead to a highly skewed distribution

- *D*: low skewness, kurtosis (relative to mean) and large values were associated with the nature of its underlying extrapolated cost estimates

2b) Correlation and risk adjustments – *effects*

- Bimodal distortions (Model 4.4) in the right tail of the *ALD* were attributed to aggregate correlation associated with *class D* (diversification analogous to retrocession, in the context of insurance, was considered in this regard); permissible ranges for covariance coefficients, required to ensure valid (i.e. non-negative) *ALD*s in respect of Models 4.4–4.5 were derived; aliasing issues (associated with *FFT* methodology) were investigated in the left tail of *ALD*s

- Variance-adjusted *ILF*s (Models 4.3, 4.5–4.6) were considered in terms of *consistency properties*, using the *Proportional Hazard* transform as a benchmark;

the *variance principle* was found to produce consistent *ILF*s provided the risk margin (as a function of risk-adjustment and correlation parameter) was acceptably low (e.g. 50% at $2.5m limit, Model 4.5); stress testing (e.g. 500% margin) revealed issues associated with covariance and variance terms at lower limits

- In addition to model comparisons and *ALD* investigations (e.g. *FFT wrap around*); as part of validation, insight into coverage and pricing issues was gained through insurer *ILF* comparisons and *Riebesell curves*

6.2 Contributions, limitations

The *model review* (§2.2) found cyber-pricing models to be in want of further development and empirical support – particularly derelict aspects included severity and aggregate loss; there was no evidence of *ILF* related models. Empirical support, based on statistically viable severity data, featured only once (Biener, Eling & Wirfs (2015), almost 1 000 cases). Key contributions made by the present research include:

1. Model confidence sets for various severity *cdf*s, derived in relation to key forms of first-party data-breach coverage
2. New insight into aspects associated with correlated *ALD*s and risk-adjusted *ILF*s

This was done in terms of nonparametric models based on empirical data, extracted from data breach survey reports (4×800 : *A–D*; 1 150: *E*). There was no evidence of such applications or findings in the *model review* (or, to the best knowledge of the author, elsewhere in cyber related academia).

Further, several algorithms were developed as a means of demonstrating practical data screening and model selection approaches. However, these contributions are not without limitations:

- Data: non-transparent, non-statistical, survey methodology; consequential (left, right) incidentally truncated data breach costs; in combination with graphical extraction methods (2015 year) are associated with uncertainty and inaccuracy; there is also the issue of the internalised nature of analysis, which only considered the external economic environment when setting inflation assumptions (and technological

environment in a general context)

- Assumptions: homogeneity (by year, country – as explored in Figure 3.6, Figure B.2) and constant underlying exposure, unchanging technological, regulatory, and legal environment are, admittedly, unrealistic (however, provided necessary simplicity for analysis); pure-risk *ILF*s ignored expenses and other 'non-risk' components (although total implied margins, including the risk element, were considered in relation to insurer *ILF*s)
- Results: uncertain – as previously described (although, to some extent, this was communicated through a range of results based on different models, correlation, and risk parameter assumptions, and model confidence sets; and assessed by way of sensitivity analyses, Monte Carlo and other comparisons)

6.3 Conclusions

Conclusions, some of which are data or model dependent (i.e. not necessarily applicable in every situation) include:

- Severity distributions, based on data breach costs, were heavy tailed in the main, although *D*, representing business interruption, often affiliated with issues such as interdependence in the realm of insurance, was found to be light tailed
- Correlation between *D* and other classes (i.e. *A*–*C*) was found to have the greatest impact on the *ALD* in its tail (in the case where the aggregate loss model was used, the peak of the second mode of a bimodal distribution was intensified). The *Value at Risk*, however, was less affected by this compared to other risk measures (e.g. standard deviation)
- Empirical evidence suggests insurers are indeed avoiding volatile severity risk associated with increased cover limits, not only through low upper limits, but through increasing implied risk margins. Reducing *Riebesell* parameters support this view; in some (isolated) cases, this led to *ILF consistency* not being observed

Enriched empirical data, as a basis for actuarial experience rating, may represent a source of value, despite the notion that it 'quickly goes stale' due to the dynamic nature of the technological environment. This is demonstrated by reconciling modelled (i.e.

'experience-based') and insurer (exposure-based) *ILF*s, and introduces the following recommendations.

6.4 Recommendations

As mentioned, onus should be placed upon all stakeholders concerned to establish a unified approach to deal with common cyber-risk management issues – whilst industry groups and international initiatives are reportedly underway; actions to 'better' address basic data issues are still highly anticipated.

Developing an anonymised 'community-wide' data base (with key elements for quantifying *cyber-risk*) may be fraught with wider issues concerning cooperation, funding, administration, and governance. However, there would appear to be some incentive to collaborate more effectively, given the $600bn (and growing) cyber-cost estimate previously mentioned (§1.1).

This would align with academic interests in support of such an initiative – although a unified approach may also be required here – possibly through a multidisciplinary academic interest group. Such cross-pollination would accelerate the development of *cyber-risk* and associated pricing models.

There were only two 'actuarial' contributions (according to title) that featured in the *model review* (§2.2), neither of which appeared to have emerged from that domain. Given this, it is worth emphasising that further actuarial contribution to this specialised field of academia is essential. Specific recommendations, in this regard, are provided in the following section.

6.5 Future directions

There are several areas that require attention and much greater input – those specific to the present research are followed by comprehensive recommendations (summarised by possible approach).

Specific areas

- *Correlation* and *interdependence*: risks within a class were assumed to be independent – simulation (e.g. common shock model) would be useful for understanding interdependence with respect to business interruption
- *Information asymmetry*: anti-selection (e.g. different limits attracting different types, levels of risk) could be explored using *ILF*s by turnover band (as in Table 5.10, Hanover (2015), possibly based on *D* divided by customer churn); empirical insight into the notion of secondary loss (Bandyopadhyay, Mookerjee & Rao, 2010) and associated asymmetries (e.g. insureds' claiming strategy, §2.2) could be investigated in terms of 'retention factors' (for pricing different deductibles) possibly in combination with *ILF*s (4.28)

Regarding cyber-loss distributions, parameter uncertainty was not considered in terms of intraclass severity *cdf*s (only for loss count, through a *Negative Binomial*). For this, *gamma-mixed Exponential* (Reshetar, 2008) or other relevant mixture could be considered. There was also the case of thresholds for *composite severity distributions*, which were 'fine-tuned' to the degree possible under an approach such as *Maximum Likelihood*, ML, (Ralucavernic, 2009; Scollnik & Sun, 2012).

Additional insight into *Value at Risk, VaR*, can also be gained. Chapter 5 demonstrated *VaR* appeared to be most resilient to the bimodal distortion accompanying correlated loss (Figure 5.6: 1–2, Scenarios 1–3). Further, several data sources remain '*untapped*' (§2.4.1) – further research in regard to available data sources would be an invaluable catalyst for subsequent research and *cyber-risk* model development.

Approach 1: building upon extant (cyber) models

In terms of the *model review* (§2.2), areas, by type of model, include the following:

- *Economic*: mainly considered, thus far, from the perspective of an organisation; further contribution from an insurer's perspective is needed
- *Correlation based* (e.g. Bayesian networks; latent factor, beta-binomial models): models reviewed require reformulation to incorporate realistic severity distributions (i.e. non-uniform); this represents an opportunity to leverage 'embedded value' within existing *cyber-risk* models

- *Operational Risk* (e.g. *GPD*, *EVT*): greater insight into aggregate loss models is still required, only a few methods for dealing with correlation in this regard were considered in the present research; recursive methods (Panjer, 1981; Panjer & Willmot, 1992), for instance, are still to be explored

In terms of *epidemiological* (e.g. *SIR*) models, few contributions have been made in cyber-specific academia, although developments can be found elsewhere (Feng & Garrido, 2011); a point at issue which is discussed in the following (and final) set of recommendations appertaining specifically to the Actuarial community.

Approach 2: reframing existing (non-cyber) methods

In the spirit of contributing towards a genuine multidisciplinary approach, the following may motivate applications that are relevant to various Actuarial disciplines:

- *Reserving*: stochastic reserving techniques could be deployed to study the effect of correlation relating to *cyber-risk*; development in the number of records breached (Chapter 3) could be based on ITRC (2018); LMA (2008) is a potential empirical source for consideration (although access permission would be required)
- *Capital Modelling* the highly topical systemic *cyber-risk* could be modelled in terms of aggregate risk, cyber-disaster (UK Government and Industry, 2015) and accumulation scenarios (Risk Management Solutions, 2016)
- *Pricing*: Generalised Linear Models – the *model review* (§2.2) identified two related contributions (Liu, Tanaka & Matsuura, 2007; Wang & Kim, 2009), both of which were based on empirical findings, none of which represented Actuarial contributions. Mapping relativities from sources such as Verizon (2019), referred to as VER (2019) in §2.4, and various online 'risk assessment' tools may be of value; machine-learning techniques could also be of use

~ The End ~

REFERENCES

Aggarwal, A., Beck, M.B., Cann, M., Ford, T., Georgescu, D., Morjaria, N., Smith, A., Taylor, Y., et al. 2015. Model risk – daring to open up the black box. *British Actuarial Journal*. 21(2):229–296. DOI: 10.1017/S1357321715000276.

Akaike, H. 1998. Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*. E. Parzen, K. Tanabe, & G. Kitagawa, Eds. New York, NY: Springer. 199–213. DOI: 10.1007/978-1-4612-1694-0_15.

Akerlof, G.A. 1970. The market for "lemons": quality uncertainty and the market mechanism. *The Quarterly Journal of Economics*. 84(3):488–500. DOI: 10.2307/1879431.

Allaben, M., Diamantoukos, C., Dicke, A., Gutterman, S., Klugman, S.A., Lord, R., Luckner, W., Miccolis, R., et al. 2008. Principles underlying actuarial science. In *Actuarial Practice Forum (July 2008, working paper)*. Available: https://www.casact.org/research/wp/.

American Association for the Advancement of Science. 2019. *AAAS home*. Available: https://www.aaas.org [2017, August 29].

Anderson, R.D. 2013. Insurance coverage for cyber attacks. *Insurance Coverage Law Bulletin (Part 1 - May, Part 2 - June)*. 12(4). Available: https://www.jdsupra.com/legalnews/insurance-coverage-for-cyber-attacks-26290/.

ArborNetworks. 2019. *DDoS & network visibility solutions*. Available: www.arbornetworks.com [2017, September 21].

Arthur J. Gallagher. 2017. *Cyber insurance summary*. Available: https://www.ajginternational.com/media/97747/mrp-cyber-insurance-summary.pdf [2017, October 22].

Association for Computing Machinery. 2019. *ACM digital library*. Available: https://dl.acm.org/.

Baer, W.S. & Parkinson, A. 2007. Cyberinsurance in IT security management. *IEEE Security & Privacy Magazine*. 5(3):50–56. DOI: 10.1109/MSP.2007.57.

Bahnemann, D. 2015. *Distributions for actuaries*. (CAS Monograph Series - 2). Available: https://www.casact.org/pubs/index.cfm?fa=monographs [2017, May 12].

Baldwin, A., Gheyas, I., Ioannidis, C., Pym, D. & Williams, J. 2012. *Contagion in cyber security attacks*. Berlin, DE. Available: https://web.archive.org/web/20150813192331/http://infosecon.net/workshop/bibliography.php [internet archive]; [2016, July 23].

Bandyopadhyay, T., Mookerjee, V. & Rao, R. 2010. A model to analyze the unfulfilled promise of cyber insurance: the impact of secondary loss. University of Texas at Dallas (working paper). Available: https://web.archive.org/web/20190331231242/http://www.utdallas.edu/~rrao/CyberBMR[1].pdf [internet archive]; [2015, July 24].

Bank for International Settlements. 2003. *The 2002 loss data collection exercise for operational risk: summary of the data collected*. Available: https://www.bis.org/bcbs/qis/ldce2002.htm [2015, June 25].

Barracchini, C. & Addessi, M.E. 2014. Cyber risk and insurance coverage: an actuarial multistate approach. *Review of Economics & Finance*. 4(1):57–69. Available: https://econpapers.repec.org/article/bapjournl/140105.htm [2015, March 22].

Betterley Risk Consultants. 2017. Betterley Report: cyber / privacy insurance market survey 2017. Available: http://betterley.com/ordering.php [2017, October 22].

Biener, C., Eling, M. & Wirfs, J.H. 2015. Insurability of cyber risk: an empirical analysis. *Geneva Papers on Risk and Insurance - Issues and Practice*. 40(1):131–158. DOI: 10.1057/gpp.2014.19.

BNY Mellon. 2016. Available: http://www.actuarialpost.co.uk/downloads/cat_1/BNY - insurance-linked-securities-cyber-risk,-insurers-and-the-capital-markets.pdf [2018, July 20].

Böhme, R. 2005. Cyber-insurance revisited. In *4th Workshop on the Economics of Information Security (WEIS 2005)*. Cambridge, MA. Available: https://web.archive.org/web/

20150813192331/http://infosecon.net/workshop/bibliography.php [internet archive]; [2015, August 05].
Böhme, R. & Kataria, G. 2006. Models and measures for correlation in cyber-insurance. In *5th Workshop on the Economics of Information Security (WEIS 2007 working paper)*. Cambridge, MA. Available: https://web.archive.org/web/20150813192331/http://infosecon.net/workshop/bibliography.php [internet archive]; [2015, July 21].
Böhme, R. & Schwartz, G. 2010. Modeling cyber-insurance: towards a unifying framework. In *9th Workshop on the Economics of Information Security (WEIS 2010)*. Harvard University, USA. Available: https://web.archive.org/web/20150324183146/http://infosecon.net/workshop/bibliography.php [internet archive]; [2015, March 22].
Boor, J.A. 1997. The complement of credibility. In *Casualty Actuarial Society (158 - May 1996, congress catalog - HG9956.C3)*. V. LXXXIII. 1–40. Available: https://www.casact.org/pubs/proceed/proceed96/ [2017, January 02].
Boutin-Dufresne, F. 2003. Between the individual and collective models, revisited. *Actuarial Research Clearing House (ARCH)*. Available: https://www.soa.org/search/publication-browse/ [2016, November 25].
Box, G.E.P. 1979. Robustness in the strategy of scientific model building. In *Robustness in Statistics*. R.L. Launer & G.N. Wilkinson, Eds. USA: Academic Press. 201–236. DOI: https://doi.org/10.1016/B978-0-12-438150-6.50018-2.
Boyd, T.A. & Sloan, A.P. 2002. *Charles F. Kettering: a biography*. Washington, DC: BeardBooks.
Bridwell, L. 2004. *ICSA labs 9th annual computer virus prevalence survey*. Available: https://www.icsalabs.com/ [2015, March 22].
Bryant, M. 2011. *20 years ago today, the World Wide Web was born*. Available: https://thenextweb.com/insider/2011/08/06/20-years-ago-today-the-world-wide-web-opened-to-the-public/ [2017, October 31].
Buch-Kromann, T. 2009. Large loss models for general insurance. University of Copenhagen. Available: https://www.math.ku.dk/english/outreach/phd-theses/.
Bühlmann, H. 1985. Premium calculation from top down. *ASTIN Bulletin (Journal of the IAA)*. 15(2):89–101. DOI: 10.2143/AST.15.2.2015021.
Burnecki, K., Janczura, J. & Weron, R. 2011. Building loss models. In *Statistical Tools for Finance and Insurance*. Berlin, Heidelberg: Springer Berlin Heidelberg. 293–328. Available: https://doi.org/10.1007/978-3-642-18062-0_9.
Burnham, K.P. & Anderson, D.R. 2002. *Model selection and multimodel inference: a practical information-theoretic approach*. 2nd ed. Springer.
California legislative information. 2016. *Law section*. United States of America, California. Available: https://leginfo.legislature.ca.gov [2016, March 25].
California Office of Privacy Protection. 2012. Recommended Practices on Notice of Security Breach Involving Personal Information. *California Office of Privacy Protection*. (January):1–32. Available: www.privacy.ca.gov.
Campbell, R., Francis, L., Prevosto, V.R., Rothwell, M. & Sheaf, S. 2006. *Report of data quality working party*. Available: actuaries.org.uk [2016, January 14].
CAS Data Management Educational Materials Working Party. 2008. *Actuarial I.Q. (information quality)*. Available: https://web.archive.org/web/20160328015149/https://www.casact.org/pubs/forum/08wforum/ [internet archive]; [2016, March 28].
Cashell, B., Jackson, W.D., Jickling, M. & Webel, B. 2004. *Economic impact of cyber-attacks*. Washington DC. Available: https://fas.org/sgp/crs/misc/RL32331.pdf [2018, January 30].
Cebula, J.J. & Young, L.R. 2010. *A taxonomy of operational cyber security risks*. Available: https://apps.dtic.mil/dtic/tr/fulltext/u2/a537111.pdf [2015, September 29].
Cerchiara, R.R. & Acri, F. 2016. Aggregate loss distribution and dependence: composite models, copula functions and fast Fourier transform for the Danish fire insurance data

(working paper). Available: https://web.archive.org/web/20190331232859/https://www.ivass.it/pubblicazioni-e-statistiche/pubblicazioni/att-sem-conv/2017/conf-131407/CerchiaraAcri_Paper.pdf [internet archive]; [2018, February 16].

Chon, G. 2015. *Cyber attack risk requires $1bn of insurance cover, companies warned*. Available: https://www.ft.com/content/61880f7a-b3a7-11e4-a6c1-00144feab7de [2017, September 12].

Chubb. 2017. *Quarter century club*. Available: https://www.chubb.com/us-en/about-chubb/quarter-century-club.aspx [2017, September 16].

Cirillo, P. 2013. Are your data really Pareto distributed? (June, 1). DOI: 10.1016/j.physa.2013.07.061.

Columbia university. 2015. *Fair use - copyright advisory services*. Available: https://copyright.columbia.edu/basics/fair-use.html [2019, March 02].

Cope, E. & Antonini, G. 2008. Observed correlations and dependencies among operational losses in the ORX consortium database. *Journal of Operational Risk*. 3(4):47–74. DOI: 10.21314/JOP.2008.052.

Cresenzi, C. & Alibrio, C. 2016. *ACE filling - Cyber and Digitech (SERFF tracking: ACEH-130778328)*. Washington State. Available: https://www.insurance.wa.gov/insurers-regulated-entities [2017, September 16].

CRO Forum. 2014. Available: https://www.thecroforum.org/wp-content/uploads/2015/01/Cyber-Risk-Paper-version-24-1.pdf [2018, July 20].

Cullina, M. 2017. *How actuaries can vet out the real risk from cyber threat*. Available: https://web.archive.org/web/20170528034421/http://www.actuarialpost.co.uk/article/how-actuaries-can-vet-out-the-real-risk-from-cyber-threat-10667.htm [internet archive]; [2017, May 28].

Currency.me.uk. 2008. *Currency*. Available: http://www.currency.me.uk [2016, March 30].

Daigle, D. & Cresenzi, C. 2018. *Federal filing - Forefront Portfolio 3.0 (SERFF tracking: ACEH-131628857)*. District of Columbia. Available: http://serff.disb.dc.gov/.

Dasu, T. & Johnson, T. 2003. *Exploratory data mining and data cleaning*. (Wiley Series in Probability and Statistics). Hoboken, NJ, USA: John Wiley & Sons, Inc. DOI: 10.1002/0471448354.

Defense Communications Agency. 1985. *Arpanet information brochure*. Available: http://www.dtic.mil/dtic/tr/fulltext/u2/a164353.pdf [2016, March 19].

Devroye, L. 1986. *Non-uniform random variate generation*. New York, NY: Springer New York. DOI: 10.1007/978-1-4613-8643-8.

Digital Attack Map. 2013. *Top daily DDOS attacks worldwide*. Available: http://www.digitalattackmap.com [2016, March 16].

Digital Guardian. 2018. *Definitive guide to U.S. state data breach laws*. Available: https://web.archive.org/web/20180905170224/https://info.digitalguardian.com/rs/768-OQW-145/images/the-definitive-guide-to-us-state-data-breach-laws.pdf [internet archive]; [2019, March 02].

Doyle, A.C. 1901. *A study in scarlet*. London: Ward, Lock & Co., Ltd. Available: https://hdl.handle.net/2027/mdp.39015083421969 [permanent link]; [2019, August 19].

Edwards, B., Hofmeyr, S. & Forrest, S. 2016. Hype and heavy tails: a closer look at data breaches. *Journal of Cybersecurity*. 2(1):3–14. DOI: 10.1093/cybsec/tyw003.

Eling, M. & Wirfs, J.H. 2015. Modelling and management of cyber risk. *Colloquium of the International Actuarial Association (IAA)*. (June). Available: https://web.archive.org/web/20160402165510/http://actuaries.org/oslo2015/scientificprogram.cfm [internet archive]; [2016, June 20].

Eling, M. & Wirfs, J.H. 2019. What are the actual costs of cyber risk events? *European Journal of Operational Research*. 272(3):1109–1119. DOI: 10.1016/j.ejor.2018.07.021.

European Commission. 2017. *Banking and finance*. Available: https://ec.europa.eu/ [2016, February 08].

European Commission. 2018. *EU member states notifications to the European Commission under the GDPR*. Available: https://ec.europa.eu/info/law/law-topic/data-protection/data-protection-eu/eu-countries-gdpr-specific-notifications_en [2019, March 02].
Federal Bureau of Investigation. 2006. *Internet crime complaint center (IC3)*. Available: http://www.ic3.gov/media/annualreports.aspx [2016, March 24].
Feenberg, A. & Friesen, N. 2012. *(Re)inventing the internet: critical case studies*. Sense. Available: https://www.sensepublishers.com/media/835-reinventing-the-internet.pdf [2017, August 29].
Feldblum, S. 1993. Risk loads for insurers. *Insurance: Mathematics and Economics*. 12(1):77. DOI: 10.1016/0167-6687(93)91044-U.
Feng, R. & Garrido, J. 2011. Actuarial applications of epidemiological models. *North American Actuarial Journal*. 15. DOI: 10.1080/10920277.2011.10597612.
Fitch Ratings. 2016. *Fitch: U.S. cyber insurance premiums total $1B per new supplemental filing*. Available: https://www.fitchratings.com/site/pr/1010744 [2019, December 29].
Floresca, L. 2014. *Cyber insurance 101: the basics of cyber coverage*. Available: https://wsandco.com/cyber-liability/cyber-basics/ [2017, December 10].
Forfar, D. & Raymont, D. 2002. Glossary of terms - general insurance.
Freifelder, L.R. 1979. Exponential utility theory ratemaking: an alternative ratemaking approach. *The Journal of Risk and Insurance*. 46(3):515–530. DOI: 10.2307/252462.
Gara, T. & Warzel, C. 2014. *A look through the Sony pictures data hack: this is as bad as it gets*. Available: https://www.buzzfeednews.com/article/tomgara/sony-hack [2017, September 23].
Gharib, A., Davies, E., Goss, G. & Faramarzi, M. 2017. Assessment of the combined effects of threshold selection and parameter estimation of generalized Pareto distribution with applications to flood frequency analysis. *Water*. 9(9):1–17. DOI: 10.3390/w9090692.
Ghosh, S. & Resnick, S. 2010. A discussion on mean excess plots. *Stochastic Processes and their Applications*. 120(8):1492–1517. DOI: 10.1016/j.spa.2010.04.002.
Google. 2016. *Jigsaw*. Available: https://jigsaw.google.com/ [2017, September 21].
Gordon, L.A. & Loeb, M.P. 2002. The economics of information security investment. *ACM Transactions on Information and System Security*. 5(4):438–457. DOI: 10.1145/581271.581274.
Gordon, L.A., Loeb, M.P. & Sohail, T. 2003. A framework for using insurance for cyber-risk management. *Communications of the ACM*. 46(3):81–85. DOI: 10.1145/636772.636774.
Great American Insurance. 2011. *Great American filing - plan limit options, commercial cyber (SERFF tracking: GACX-G127206148)*. Wisconsin. Available: https://filingaccess.serff.com/sfa/search/filingSearchResults.xhtml.
Greenberg, P. 2012. *Security breach legislation*. Available: http://www.ncsl.org/research/telecommunications-and-information-technology/security-breach-legislation-2012.aspx [2019, March 02].
Greenberg, P. 2014. *Security breach legislation*. Available: http://www.ncsl.org/research/telecommunications-and-information-technology/2014-security-breach-legislation.aspx [2019, March 02].
Greenberg, P. 2015. *Security breach legislation*. Available: http://www.ncsl.org/research/telecommunications-and-information-technology/2015-security-breach-legislation.aspx [2019, March 02].
Grübel, R. & Hermesmeier, R. 1999. Computation of compound distributions I: aliasing errors and exponential tilting. *ASTIN Bulletin (Journal of the IAA)*. 29(02):197–214. DOI: 10.2143/AST.29.2.504611.
Halliwell, L.J. 2009. Mixing collective risk models. *Casualty Actuarial Society e-forum (Fall)*. 1–12. Available: https://www.casact.org/pubs/forum/09fforum/1.pdf [2018, May 09].
Halliwell, L.J. 2013. Classifying the tails of loss distributions. In *Casualty Actuarial Society E-*

Forum (Spring). V. 2. E.A. Smith, Ed. 1–27. Available: https://www.casact.org/pubs/forum/13spforumv2/ [2017, April 24].

Haney, W. 1972. Available: http://legaciesofwar.org/files/The_Pentagon_Papers_and_the_United_States_Involvement_in_Laos.pdf.

Hanover Insurance. 2015. *Hanover religious institutions filling (SERFF tracking: HNVX-G129911675)*. District of Columbia. Available: http://serff.disb.dc.gov/.

Hawkes, A.G. 1971. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*. 58(1):83–90. DOI: 10.1093/biomet/58.1.83.

Heal, G. & Kunreuther, H. 2004. *Interdependent security : a general model*. Cambridge, MA. DOI: 10.3386/w10706.

Herath, H. & Herath, T. 2011. Copula-based actuarial model for pricing cyber-insurance policies. *Insurance Markets and Companies: Analyses and Actuarial Computations*. 2(1):7–20. Available: https://businessperspectives.org/author/hemantha-s-b-herath [2015, March 21].

Hess, C. 2011. The impact of the financial crisis on operational risk in the financial services industry: empirical evidence. *Journal of Operational Risk*. 6(1):23–35. DOI: 10.21314/JOP.2011.087.

Hilbert, M. & López, P. 2011. World's technological capacity to store, communicate, and compute information. *Science*. 332(6025). Available: http://science.sciencemag.org/content/332/6025/60?variant=full-text&sso=1&sso_redirect_count=1&oauth-code=35c031a5-9a62-46b4-aebe-5f075125d0f0 [2017, August 29].

Hisakado, M., Kitsukawa, K. & Mori, S. 2006. Correlated binomial models and correlation structures. *Journal of Physics A: Mathematical and General*. 39(50):15365–15378. DOI: 10.1088/0305-4470/39/50/005.

Hiscox. 2017. *Cyber and data - policy wording*. Available: https://www.hiscox.co.uk/sites/uk/files/documents/2017-04/13388-cyber-and-data-uk-wording.pdf [2017, October 29].

Homer, D.L. & Rosengarten, R.A. 2011. Method for efficient simulation of the collective risk model. In *Casualty Actuarial Society E-Forum (Spring)*. 1–41. Available: https://www.casact.org/pubs/forum/11spforum/ [2016, November 23].

Identity Theft Resource Center. 2018. *Data breaches*. Available: https://www.idtheftcenter.org/data-breaches/.

Institute and Faculty of Actuaries. 2019. *Heritage online*. Available: https://actuaries.cirqahosting.com [2019, February 25].

Inter-university consortium for political and social research. 2012. *National crime victimization survey*. Available: http://www.icpsr.umich.edu/icpsrweb/landing.jsp.

Jacobs, J. 2014. *Analyzing ponemon cost of data breach*. Available: https://datadrivensecurity.info/blog/posts/2014/Dec/ponemon/ [2016, December 11].

Jensen, F. & Rosenthal, S. 2015. Cyber insurance from a quantum perspective. In *Property Insurance Claims Group (PICG) conference (14 May)*. London, UK. Available: https://web.archive.org/web/20170519105929/http://www.picgconference.com/presentations-2/ [internet archive]; [2017, May 06].

Jobs, S. 2010. Available: https://news.stanford.edu/news/2005/june15/jobs-061505.html [2019, June 23].

Kaas, R., Goovaerts, M., Dhaene, J. & Denuit, M. 2008. *Modern actuarial risk theory*. Berlin, Heidelberg: Springer Berlin Heidelberg. DOI: 10.1007/978-3-540-70998-5.

Kampstra, P. 2008. Beanplot: a boxplot alternative for visual comparison of distributions. *Statistical Software*. 28(Code Snippet 1):1–9. DOI: 10.18637/jss.v028.c01.

Kardoulaki, E. 2018. *Cyber insurance pricing: quantifying the unknown in a multibillion dollar market*. Available: https://www.finextra.com/blogposting/15278/cyber-insurance-pricing-quantifying-the-unknown-in-a-multibillion-dollar-market [2018, May 28].

Kesan, J., Majuca, R. & Yurcik, W. 2005. Cyberinsurance as a market-based solution to the problem of cybersecurity: a case study. In *4th Workshop on Economics of Information*

Security. Cambridge, MA. 1–46. Available: https://web.archive.org/web/20180429142742/http://infosecon.net/workshop/bibliography.php [internet archive]; [2016, July 30].

Kesan, J., Majuca, R. & Yurcik, W. 2008. Three economic arguments for cyberinsurance. In *Chapter 16, Securing Privacy in the Internet Age, A. Chander, L. Gelman, & M. J. Radin*. (U Illinois Law & Economics Research Paper). Stanford University Press. 345–366. Available: https://ssrn.com/abstract=577862.

Kirsch, C. & Greenberg, P. 2013. *Security breach legislation*. Available: http://www.ncsl.org/research/telecommunications-and-information-technology/2013-security-breach-legislation635200257.aspx [2019, March 02].

Klugman, S.A., Panjer, H.H. & Willmot, G.E. 2004. *Loss models: from data to decisions*. 2nd ed. (Wiley Series in Probability and Statistics). Hoboken, NJ: Wiley Interscience.

Kullback, S. & Leibler, R.A. 1951. On information and sufficiency. *Annals of Mathematical Statistics*. 22(1):79–86. DOI: 10.1214/aoms/1177729694.

Kumar, C. 2017. *9 interesting ways to watch cyberattack in real-time worldwide*. Available: https://geekflare.com/real-time-cyber-attacks/ [2017, September 20].

Kunreuther, H. & Heal, G. 2002. *Interdependent security: the case of identical agents*. Cambridge, MA: National Bureau of Economic Research. DOI: 10.3386/w8871.

LaCroix, K. 2016. *Guest post: cyber-liability insurance and the retroactive date exclusion*. Available: https://web.archive.org/web/20171029042516/http://www.dandodiary.com/2016/05/articles/cyber-liability/guest-post-cyber-liability-insurance-and-the-potentially-severe-limitations-of-the-retroactive-datepolicy-inception-date-exclusions/ [internet archive] [2017, October 29].

Laffont, J.-J. & Martimort, D. 2009. *The theory of incentives*. Princeton University Press. DOI: 10.2307/j.ctv7h0rwr.

Laszka, A., Felegyhazi, M. & Buttyan, L. 2014. A survey of interdependent information security games. *ACM Computing Surveys*. 47(2):1–38. DOI: 10.1145/2635673.

Laube, S. & Böhme, R. 2016. The economics of mandatory security breach reporting to authorities. *Journal of Cybersecurity*. 2(1):29–41. DOI: 10.1093/cybsec/tyw002.

Lawrence, I. & Lin, K. 1989. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*. 255–268.

Lawson, D.M. 2004. *Posterity: letters of great Americans to their children*. New York: Doubleday.

Lee, Y. 1988. The mathematics of excess of loss coverages and retrospective rating - a graphical approach. *PCAS LXXV*. 49–77. Available: http://casact.net/pubs/proceed/proceed88/88049.pdf.

Levene, H. 1960. Contributions to probability and statistics. In *Essays in honor of Harold Hotelling*. I. Olkin, S. Ghurye, W. Hoeffding, W. Madow, & H. Mann, Eds. Stanford, California: Stanford University Press California. 278–292.

List, H.-F. & Lohner, N. 1998. Extreme value techniques part 3: pricing high-excess property and casualty layers. *General Insurance Convention & ASTIN Colloquium (7-10 Oct)*. 301–339. Available: https://www.actuaries.org.uk/.

Liu, H. & Wang, R. 2017. Collective risk models with dependence uncertainty. *ASTIN Bulletin*. 47(2):361–389. DOI: 10.1017/asb.2017.4.

Liu, W., Tanaka, H. & Matsuura, K. 2007. Empirical-analysis methodology for information-security investment and its application to reliable survey of Japanese firms. *IPSJ Digital Courier*. 3:585–599. DOI: 10.2197/ipsjdc.3.585.

Lloyd's. 2015. *A quick guide to cyber risk*. Available: https://www.lloyds.com/news-and-insight/news-and-features/emerging-risk/emerging-risk-2015/a-quick-guide-to-cyber-risk [2017, December 12].

Lloyd's. 2019. *Glossary and acronyms*. Available: https://www.lloyds.com/common/help/glossary [2017, October 12].

Lloyd's Market Association. 2008. *Risk code triangulation reports*. Available: www.lmalloyds.com/ [restricted access].

Lyman, P., Varian, H.R., Charles, P., Good, N., Jordan, L.L. & Pal, J. 2003. *How much information?* Available: http://groups.ischool.berkeley.edu/archive/how-much-info-2003/ [2006, February 25].

Mack, T. & Fackler, M. 2003. Exposure-rating in liability reinsurance. *Blätter der DGVFM*. 26(2):229–238. DOI: 10.1007/BF02808374.

Marker, J.O. & Mohl, F.J. 1980. Rating Claims-Made Insurance Policies. *Pricing Property and Casualty Insurance Products*. 265–304.

Marks, P. 2011. *Dot-dash-diss: The gentleman hacker's 1903 lulz | New Scientist*. Available: https://www.newscientist.com/article/mg21228440-700-dot-dash-diss-the-gentleman-hackers-1903-lulz/ [2016, March 16].

Marsh. 2014. Cyber gap insurance. *Global Energy Practice*.

Marsh. 2015. *United States insurance market report*. Available: http://www.oliverwyman.com/content/dam/marsh/Documents/PDF/US-en/United States Insurance Market Report 2015-02-2015.pdf [2018, July 21].

McAfee & Center for Strategic and International Studies. 2018. *Economic impact of cybercrime — no slowing down*. Available: https://www.mcafee.com/enterprise/en-us/assets/reports/restricted/economic-impact-cybercrime.pdf [2018, July 16].

Meyers, G. & Heckman, P. 1984. The calculation of aggregate loss distributions from claim severity and claim count distributions. In *Casualty Actuarial Society (163 - May 1993)*. V. LXX. 22–61. Available: https://www.casact.org/pubs/proceed/proceed83/ [2016, June 11].

Meyers, C.A., Powers, S.S. & Faissol, D.M. 2009. *Taxonomies of cyber adversaries and attacks: a survey of incidents and approaches*. Available: https://e-reports-ext.llnl.gov/pdf/379498.pdf.

Miccolis, R.S. 1978. On the theory of increased limits excess of loss pricing. In *Casualty Actuarial Society (121 - May 1977)*. V. LXIV. 27–59. Available: https://www.casact.org/pubs/proceed/proceed77/.

Michaelides, N., Brown, P., Chacko, F., Graham, M., Haynes, J., Hindley, D., Howard, S., Johnson, H., et al. 1997. The premium rating of commercial risks. In *General Insurance Convention*. 397–491. Available: https://www.actuaries.org.uk/learn-and-develop/conference-paper-archive/1997.

Microstrategy. 2016. Available: www.microstrategy.com.

Mildenhall, S.J. 2005. *Correlation and aggregate loss distributions with an emphasis on the Iman-Conover method*. Available: https://www.casact.org/pubs/forum/06wforum/06w107.pdf.

Milne-Thomson, L.M. 2000. *The calculus of finite differences*. American Mathematical Soc.

Ministry of Economy Trade Industry. 2004. *Information processing survey report*. Available: https://web.archive.org/web/20071130225516/http://www.meti.go.jp/statistics/zyo/zyouhou/result-2/h15jyojitsu.html [internet archive]; [2016, February 28].

Moher, D., Liberati, A., Tetzlaff, J. & Altman, D.G. 2009. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLOS Medicine*. 6(7):1–6. DOI: 10.1371/journal.pmed.1000097.

Moore, T. 2012. Theory of and research on cyber insurance. In *Presented at the Department of Homeland Security Workshop: Defining Challenges to Today's Cybersecurity Insurance Market*.

Motulsky, H. & Christopoulos, A. 2004. *Fitting models to biological data using linear and nonlinear regression: a practical guide to curve fitting*. Oxford University Press. Available: https://global.oup.com/academic/product/fitting-models-to-biological-data-using-linear-and-nonlinear-regression-9780195171792?cc=us&lang=en& [2019, June 03].

Mukhopadhyay, A., Saha, D., Chakrabarti, B.B., Mahanti, A. & Podder, A. 2005. Insurance for

cyber-risk: a utility model. *Decision*. 32(1):153–169. Available: https://www.academia.edu/19443334/Insurance_for_cyber-risk_A_Utility_Model.

Mukhopadhyay, A., Chatterjee, S., Saha, D., Mahanti, A. & Sadhukhan, S.K. 2013. Cyber-risk decision models: To insure IT or not? *Decision Support Systems*. 56(1):11–26. DOI: 10.1016/j.dss.2013.04.004.

Müller, J., Polansky, D., Novak, P., Polivaev, D., Foltin, C. & Lavarde, E. 2004. Available: http://freemind.sourceforge.net/wiki/index.php/Main_Page [2019, February 10].

Munich Re. 2015. *Cyber risk underwriting guidelines for Nationwide Mutual Insurance company*. Available: www.hsb.com/HSBExt/nationwide_data-security_Cyber-Risk-Underwriting-Guidelines/.

Murphy, W. 2013. *Illinois National filing - PortfolioSelect (SERFF tracking: AGNY-129243435)*. South Dakota. Available: https://apps.sd.gov/cc57serffportal/ [2016, September 04].

National Association of Insurance Commissioners (NAIC). 2019. *System for electronic rate and form filing*. Available: https://www.serff.com/ [2015, November 21].

NetDiligence. 2016. *Cyber risk management services*. Available: https://netdiligence.com/portfolio/cyber-claims-study/ [2016, March 26].

Obama, B. 2009. Available: https://web.archive.org/web/20170126105102/https://obamawhitehouse.archives.gov/the-press-office/remarks-president-securing-our-nations-cyber-infrastructure [internet archive]; [2017, February 01].

Ogut, H., Raghunathan, S. & Menon, N. 2005a. Cyber insurance and IT security investment: impact of interdependent risk. In *4th Workshop on the Economics of Information Security (WEIS 2005)*. Cambridge, MA. Available: https://web.archive.org/web/20180429142742/http://infosecon.net/workshop/bibliography.php [internet archive]; [2015, March 02].

Ogut, H., Raghunathan, S. & Menon, N. 2005b. Available: https://www.researchgate.net/publication/229014862_Information_security_risk_management_through_self-protection_and_insurance [2018, June 28].

Operational Riskdata eXchange Association. 2017. *ORX loss data*. Available: https://www.orx.org.

Organisation for Economic Co-operation and Development. 2018. *OECD data*. Available: https://data.oecd.org/ [2016, March 24].

Osborne, C. 2015. *Anthem data breach cost likely to smash 100 million barrier*. Available: https://web.archive.org/web/20160111234028/http://www.zdnet.com/article/anthem-data-breach-cost-likely-to-smash-100-million-barrier/ [internet archive]; [2015, December 29].

Panjer, H.H. 1981. Recursive evaluation of a family of compound distributions. *Astin Bulletin (Journal of the IAA)*. 12(01):22–26.

Panjer, H.H. & Willmot, G.E. 1992. *Insurance risk models*. JSTOR.

Parker, D. & Farkas, C. 2011. Modeling estimated risk for cyber attacks: merging public health and cyber security. *Journal of Information Assurance and Security (JIAS)*. 2:32–36. Available: http://www.cse.sc.edu/~farkas/publications/j18.pdf [2015, March 22].

Parodi, P. 2014. *Pricing in general insurance*. New York: Chapman and Hall/CRC. DOI: 10.1201/b17525.

Ponemon Institute. 2012a. *2011 cost of data breach study: Australia (sponsored by Symantec)*.

Ponemon Institute. 2012b. *2011 cost of data breach study: France (sponsored by Symantec)*.

Ponemon Institute. 2012c. *2011 cost of data breach study: Germany (sponsored by Symantec)*.

Ponemon Institute. 2012d. *2011 cost of data breach study: global analysis (sponsored by Symantec)*. Available: http://www.ponemon.org/library/.

Ponemon Institute. 2012e. *2011 cost of data breach study: India (sponsored by Symantec)*.

Ponemon Institute. 2012f. *2011 cost of data breach study: Italy (sponsored by Symantec)*.

Ponemon Institute. 2012g. *2011 cost of data breach study: Japan (sponsored by Symantec)*.

Ponemon Institute. 2012h. *2011 cost of data breach study: United Kingdom (sponsored by*

Symantec).
Ponemon Institute. 2012i. *2011 cost of data breach study: United States (sponsored by Symantec).*
Ponemon Institute. 2013a. *2013 cost of data breach study: Australia (sponsored by Symantec).*
Ponemon Institute. 2013b. *2013 cost of data breach study: Brazil (sponsored by Symantec).*
Ponemon Institute. 2013c. *2013 cost of data breach study: France (sponsored by Symantec).*
Ponemon Institute. 2013d. *2013 cost of data breach study: Germany (sponsored by Symantec).*
Ponemon Institute. 2013e. *2013 cost of data breach study: global analysis (sponsored by Symantec).* Available: http://www.ponemon.org/library/.
Ponemon Institute. 2013f. *2013 cost of data breach study: India (sponsored by Symantec).*
Ponemon Institute. 2013g. *2013 cost of data breach study: Italy (sponsored by Symantec).*
Ponemon Institute. 2013h. *2013 cost of data breach study: Japan (sponsored by Symantec).*
Ponemon Institute. 2013i. *2013 cost of data breach study: United Kingdom (sponsored by Symantec).*
Ponemon Institute. 2013j. *2013 cost of data breach study: United States (sponsored by Symantec).*
Ponemon Institute. 2014a. *2014 cost of data breach study: Arabian region (sponsored by IBM).*
Ponemon Institute. 2014b. *2014 cost of data breach study: Australia (sponsored by IBM).*
Ponemon Institute. 2014c. *2014 cost of data breach study: Brazil (sponsored by IBM).*
Ponemon Institute. 2014d. *2014 cost of data breach study: France (sponsored by IBM).*
Ponemon Institute. 2014e. *2014 cost of data breach study: Germany (sponsored by IBM).*
Ponemon Institute. 2014f. *2014 cost of data breach study: global analysis (sponsored by IBM).* Available: https://www.ponemon.org/library.
Ponemon Institute. 2014g. *2014 cost of data breach study: India (sponsored by IBM).*
Ponemon Institute. 2014h. *2014 cost of data breach study: Italy (sponsored by IBM).*
Ponemon Institute. 2014i. *2014 cost of data breach study: Japan (sponsored by IBM).*
Ponemon Institute. 2014j. *2014 cost of data breach study: United Kingdom (sponsored by IBM).*
Ponemon Institute. 2014k. *2014 cost of data breach study: United States (sponsored by IBM).*
Ponemon Institute. 2015a. *2015 cost of data breach study: Arabian region (sponsored by IBM).*
Ponemon Institute. 2015b. *2015 cost of data breach study: Australia (sponsored by IBM).*
Ponemon Institute. 2015c. *2015 cost of data breach study: Brazil (sponsored by IBM).*
Ponemon Institute. 2015d. *2015 cost of data breach study: Canada (sponsored by IBM).*
Ponemon Institute. 2015e. *2015 cost of data breach study: France (sponsored by IBM).*
Ponemon Institute. 2015f. *2015 cost of data breach study: Germany (sponsored by IBM).*
Ponemon Institute. 2015g. *2015 cost of data breach study: global analysis (sponsored by IBM).* Available: https://www.ponemon.org/library.
Ponemon Institute. 2015h. *2015 cost of data breach study: India (sponsored by IBM).*
Ponemon Institute. 2015i. *2015 cost of data breach study: Italy (sponsored by IBM).*
Ponemon Institute. 2015j. *2015 cost of data breach study: Japan (sponsored by IBM).*
Ponemon Institute. 2015k. *2015 cost of data breach study: United Kingdom (sponsored by IBM).*
Ponemon Institute. 2015l. *2015 cost of data breach study: United States (sponsored by IBM).*
Ponemon Institute. 2019. *Measuring trust in privacy and security*. Available: https://www.ponemon.org/ [2019, February 24].
Pouget, F., Dacier, M. & Pham, V.H. 2005. Leurre.com: on the advantages of deploying a large scale distributed honeypot platform. In *E-Crime and Computer Conference (ECCE).* Monaco. Available: http://www.eurecom.fr/publication/1558 [2016, February 28].
Princeton university. 2009. *WordNet search (3.1).* Available: http://wordnetweb.princeton.edu/perl/webwn [2016, March 29].
Privacy Rights Clearinghouse. 2016. *Data breaches*. Available: https://web.archive.org/web/

20161029235437/https://www.privacyrights.org/data-breaches [internet archive]; [2015, November 21].
Rachev, S.S.T., Chernobai, A. & Menn, C. 2006. Empirical examination of operational loss distributions. In *Perspectives on Operations Research*. M. Morlock, C. Schwindt, N. Trautmann, & J. Zimmermann, Eds. Wiesbaden: DUV. 379–401. DOI: 10.1007/978-3-8350-9064-4_21.
Radcliff, D. 2001. *Calculating e-risk*. Available: https://www.computerworld.com/article/2591459/data-privacy/calculating-e-risk.html [2018, July 19].
Ralucavernic, S.T. 2009. Some composite exponential-Pareto models for Actuarial prediction. *Romanian Journal of Economic Forecasting –*. 4(12):82–100. Available: http://www.ipe.ro/rjef/rjef4_09/rjef4_09_5.pdf [2018, February 16].
Reshetar, G. 2008. Dependence of operational losses and the capital at risk. *SSRN Electronic Journal*. (January, 7). DOI: 10.2139/ssrn.1081256.
Reynkens, T., Verbelen, R., Beirlant, J. & Antonio, K. 2016. Modeling censored losses using splicing: a global fit strategy with mixed Erlang and extreme value distributions. 1–45. Available: http://arxiv.org/abs/1608.01566.
Risk Management Solutions. 2016. Cyber insurance Exposure Data Schema v1.0. *Cambridge Centre for Risk Studies*. Available: http://static.rms.com/email/documents/managing-cyber-insurance-accumulation-risk-rms-crs-jan2016.pdf.
Rohatgi, A. 2013. Available: https://automeris.io/WebPlotDigitizer.
Romanosky, S., Ablon, L., Kuehn, A. & Jones, T. 2017. Content analysis of cyber insurance policies: how do carriers write policies and price cyber risk? (draft paper). *SSRN Electronic Journal*. (March, 7). DOI: 10.2139/ssrn.2929137.
Sagan, C. 1983. *Cosmos*. 1st pbk. e ed. A. Freedgood, Ed. New York & Canada: Random House. Available: http://www.worldcat.org/oclc/12736033 [permalink]; [2019, August 18].
SAS. 2015. Available: https://web.archive.org/web/20150807100828/https://www.sas.com/resources/product-brief/sas-oprisk-globaldata-brief.pdf [internet archive]; [2015, July 11].
Scollnik, D.P.M. & Sun, C. 2012. Modeling with Weibull-Pareto models. *North American Actuarial Journal*. 16(2):260–272. DOI: 10.1080/10920277.2012.10590640.
Secretariat of the Security and Defence Committee Eteläinen. 2013. Finland's cyber security strategy: government resolution. *Finland´s Cyber security Strategy*. Available: https://www.defmin.fi.
Sedano, S. & Rodriguez, E. 2015. *Samsung fire and marine filing - package endorsement (SERFF: PERR-129994727)*. District of Columbia. Available: http://serff.disb.dc.gov/ [2015, June 08].
Selleck, C. 2015. *National Liability & Fire (NLF) filing - CyberSecurity (SERFF tracking: PRFL-130307253)*. District of Columbia. Available: http://serff.disb.dc.gov/ [2017, September 12].
Sharp, M. 2016. *First party vs. third party liability insurance coverage*. Available: https://www.myinsurancequestion.com/first-party-vs-third-party-liability/ [2017, December 10].
Shevchenko, P. V. 2010. Calculation of aggregate loss distributions. *Journal of Operational Risk*. 5(2):3–40. DOI: 10.21314/JOP.2010.077.
Solomon, M. 2017. Cyber risk is opportunity. In *Cybersecurity: impact on insurance business and operations*. (Joint Ruisk Management Section Essays on Cybersecurity). 4–8.
Soo Hoo, K.J.K.J. 2000. *How much is enough? A risk management approach to computer security (working paper)*. California, Stanford University: Citeseer. Available: https://doc.uments.com/h-how-much-is-enough-a-risk-management-approach-to-computer-security.pdf [2015, September 05].
Statsoft. 2016. Available: https://www.tibco.com.
Stieltjes, T.J. 1995. Recherches sur les fractions continues. *Annales de la faculté des sciences*

de Toulouse Mathématiques. 4(1):1–35. DOI: 10.5802/afst.789.
Sundt, B. 1999. On multivariate Panjer recursions. *ASTIN Bulletin (Journal of the IAA)*. 29(1):29–45. DOI: 10.2143/AST.29.1.504605.
SysAdmin Audit Admin and Security Technology Institute. 2019. *Internet Storm Center - DShield*. Available: https://web.archive.org/web/20151112153715/https://isc.sans.edu/ [internet archive]; [2015, November 12].
Treasury, H. 2011. Internal Audit Records Management.
Tse, Y.-K. 2009. *Nonlife actuarial models*. Cambridge University Press. DOI: 10.1017/cbo9780511812156.
UK Government and Industry. 2015. Cyber Risk Report. (March).
United States Department of Justice. 2001. *Department of justice*. Available: https://www.justice.gov/ [2016, February 27].
University of Cape Town. 2019. *UCT home*. Available: https://www.uct.ac.za/ [2016, March 06].
US Department of Commerce. 2019. *US bureau of economic analysis*. Available: https://www.bea.gov/ [2017, September 22].
US Department of Homeland Security. 2012. *Cybersecurity insurance*. Arlington, Virginia. Available: https://www.dhs.gov/cisa/cybersecurity-insurance.
van der Vaart, A.W. 1998. *Asymptotic statistics*. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511802256.
van Mieghem, P., Omic, J. & Kooij, R. 2009. Virus spread in networks. *IEEE/ACM Transactions on Networking*. 17(1):1–14. DOI: 10.1109/TNET.2008.925623.
Verisk analytics. 2017. *ISO*. Available: https://www.verisk.com/insurance/brands/iso/ [2017, April 08].
Verizon. 2019. *Data breach investigation report*. Available: https://enterprise.verizon.com/resources/reports/dbir/ [2016, March 25].
Vernic, R. & Sundt, B. 2009. *Recursions for convolutions and compound distributions with insurance applications*. (EAA Lecture Notes). Berlin, Heidelberg: Springer. DOI: 10.1007/978-3-540-92900-0.
Villaseñor-Alva, J.A. & González-Estrada, E. 2009. A bootstrap goodness of fit test for the generalized Pareto distribution. *Computational Statistics and Data Analysis*. 53(11):3835–3841. DOI: 10.1016/J.CSDA.2009.04.001.
Vose. 2019. *Risk analysis software for excel*. Available: https://www.vosesoftware.com/products/modelrisk/.
Wang, S. 1995. Insurance pricing and increased limits ratemaking by proportional hazards transforms. *Insurance: Mathematics and Economics*. 17(1):43–54. DOI: 10.1016/0167-6687(95)00010-P.
Wang, S. 1996. Premium calculation by transforming the layer premium density. *ASTIN Bulletin (Journal of the IAA)*. 26(1):71–92. DOI: 10.2143/AST.26.1.563234.
Wang, S.S. 1998. *Aggregation of correlated risk portfolios: models & algorithms (research contract with COTOR)*. Available: https://www.casact.org/research/cotor/ [2015, May 16].
Wang, S.S. 1999a. Aggregation of correlated risk portfolios: models and algorithms. In *Casualty Actuarial Society (163 - Nov 1998, congress catalog - HG9956.C3)*. V. LXXXV. E. C. Connell, D.A. Crifo, W.F. Dove, D.R. Edlefson, E.M. Gardiner, J.F. Golz, G.R. Josephson, K.E. Kufera, M. Lewis, R.A. Moody, D. Schwab, & T.A. Turnacioglu, Eds. Toronto. 848–939. Available: https://www.casact.org/pubs/proceed/proceed98/ [2015, October 13].
Wang, S.S. 1999b. Implementation of proportional hazard transforms in ratemaking. In *Casualty Actuarial Society (163 - Nov 1998, congress catalog - HG9956.C3)*. V. LXXXV. E. C. Connell, D.A. Crifo, W.F. Dove, D.R. Edlefson, E.M. Gardiner, J.F. Golz, G.R. Josephson, K.E. Kufera, M. Lewis, R.A. Moody, D. Schwab, & T.A. Turnacioglu, Eds.

United Book Press. 940–979. Available: https://www.casact.org/pubs/proceed/proceed98/ [2017, December 28].

Wang, Q. & Kim, S. 2009. Cyber attacks: cross-country interdependence and enforcement. In *8th Workshop on the Economics of Information Security (WEIS) 2009 working paper, University College London, UK 24-25 June 2009*. (Research Collection School Of Information Systems). 1–16. Available: https://ink.library.smu.edu.sg/sis_research/3301; [WEIS internet archive: https://web.archive.org/web/20180429142742/http://infosecon.net/workshop/bibliography.php] [2015, November 15].

Werner, G. & Modlin, C. 2010. *Basic ratemaking*. 4th ed. Available: https://web.archive.org/web/20150616074102/https://www.casact.org/library/studynotes/Werner_Modlin_Ratemaking.pdf [internet archive]; [2015, July 05].

Whatsapp. 2019. . Available: https://www.whatsapp.com.

Wolfrom, B., Little, J. & Rielley, J. 2015. Filing Fees. 2012377.

Workshop on the Economics of Information Security. 2019. *WEIS*.

World Bank. 2019. *Open data*. Available: https://web.archive.org/web/20151231235345/https://data.worldbank.org/ [internet archive]; [2015, December 27].

World Economic Forum. 2015. *Global risks report - 10th edition*.

World Economic Forum. 2016. *Global risks report - 11th edition*.

World Economic Forum. 2018. *Global risks report - 13th edition*. Available: https://www.weforum.org/reports/the-global-risks-report-2018 [2018, July 16].

WorldCat. 2019. *World's largest library catalog*. Available: http://www.worldcat.org/.

Yannacopoulos, A.N., Lambrinoudakis, C., Gritzalis, S., Xanthopoulos, S.Z. & Katsikas, S.N. 2008. Modeling privacy insurance contracts and their utilization in risk management for ICT firms. In *Computer Security - ESORICS 2008, 13th European Symposium on Research in Computer Security*. S. Jajodia & J. López, Eds. (Lecture Notes in Computer Science, vol 5283). Málaga, Spain: Springer Berlin Heidelberg. 207–222. DOI: 10.1007/978-3-540-88313-5_14.

Zweifel, P. & Eisen, R. 2012. *Insurance economics*. (Springer Texts in Business and Economics). Berlin, Heidelberg: Springer Berlin Heidelberg. DOI: 10.1007/978-3-642-20548-4.

APPENDICES

Appendix A Literature and data sources

This appendix is relevant for Chapter 2 – in particular, A.1 describes the search strategy for identifying literary sources considered in the *model review* (§2.2); A.2 summarises underlying components used to calculate *PSS*s for ranking sources in §2.4.2; and A.3 provides a reference guide pertaining to various data sources considered in Chapter 3.

A.1 Literary search strategy

The search strategy used to identify studies in the *model review* (Figure 2.1) is illustrated in Figure A.1. This incorporates various filters (e.g. language, content, etc.) and utilises the University of Cape Town [UCT] (2019) online search engine.

Titles and keywords are searched using strings that are made up of one word from each of the following groups:

- Group 1: 'cyber', 'information', and 'interdependent'
- Group 2: '*risk management*', '*insurance*' (and derivatives, such as insurability), and '*security*'

The UCT (2019) online search, used to generate these results, accesses databases such as WorldCat (2019), which is self-proclaimed as 'world's largest network of libraries'.

Incorporated in Figure A.1 are supplementary sources to compliment this search, such as Workshop on the Economics of Information Security [WEIS] (2019) – (archives of papers on information security and privacy), and Association for Computing Machinery [ACM] (2019) – (an international society for learned computing). The library catalogue of Institute and Faculty of Actuaries (2019) was also considered.

The 22 studies that are identified in this figure constitute studies in the *model review* (Figure 2.1) – this excludes the study Edwards, Hofmeyr & Forrest (2016), which fell outside the *review period* (2000–mid-2006).

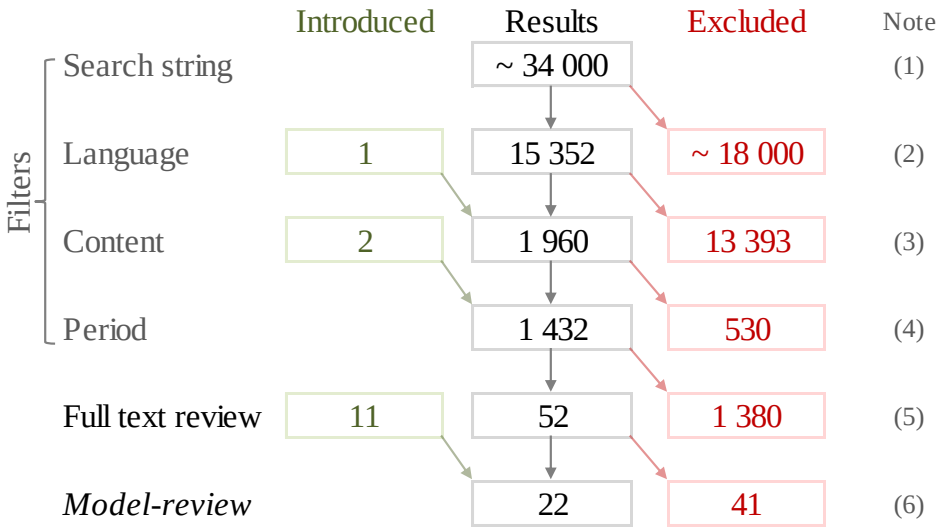


Figure A.1 Identification of studies Notes: (1) Search string: "*ti:((cyber | information | interdependent) + (risk management | insur* | security)) kw: (model | empirical)*" - which applies to titles (i.e. '*ti*') and keywords (i.e. '*kw*'), through the UCT (2019) search engine. (2) English-only; identified Barracchini & Addessi (2014) from a similar (but excluded) Italian manuscript. (3) Full-text, peer-reviewed (re-included Soo Hoo (2000), Liu, Tanaka & Matsuura (2007) - not peer-reviewed). (4) Period: 2000–mid 2016. (5) 52 studies identified for full-text review by scanning titles, then abstracts, and introduced 11 new studies from on-line searches, references, and archived libraries such as WEIS (2019). (6) Eliminated 41 studies based on full-text review, leaving 22 for the *model review*. Motivated by *Preferred Reporting Items for Systematic Reviews and Meta-Analyses*, PRISMA, (Moher et al., 2009) and Biener, Eling & Wirfs (2015) search strategy for cyber-related losses.

A.2 Data-source ranking

A summary of backing calculations for *PSS*s in Figure 2.5 are summarised in Table A.1 – the top portion of this table represents 'previously modelled' data sources; and the bottom portion includes '*untapped*' sources that did not feature in the *model review* §2.2 (in particular, Figure 2.1). Green (highest), yellow, and grey (lowest) colour coding represents scores associated with factors 1–3 in Figure 2.4 (a red cross indicates where minimum requirements are not met in this regard).

Comparable sources in Figure 2.5 (i.e. with common icon and text colour) are indicated by *A*–*C* in the data source column of this table. Sources that fail prespecified minimum criteria are indicated with a red cross mark.

		1) **Content & level of detail**			2) **Credibility**	3) **Relevance**	**Potential Suitability**
Modelled (author, year)	**Data source**	Count (N)	Severity (X_1,..., X_N)	Exposure	Years spanned	Age (most recent yr)	Score = 1) + 2) + 3)
Bohme (2006)	C Honeypot [1]	Individual	None	None	× [0,3)	× (2,∞)	1.0
Rachev, Chernobai & Menn (2006)	B BIS (2003)	Aggregate	Aggregate	Aggregate	× [0,3)	× (2,∞)	1.5
Liu, Tanaka & Matsuura (2007)	METI (2004)	Aggregate	None	None	× [0,3)	× (2,∞)	0.5
Cope & Antonini (2008)	B ORX (2017)	None	Individual	Aggregate	[5,∞)	× (2,∞)	2.5
Wang & Kim (2009)	C SANS (2019)	Individual	None	None	[5,∞)	(0,1]	3.0
	D WDID [2]	None	None	Aggregate	[5,∞)	(0,1]	2.5
H. Herath & T. Herath (2011)	ICSA [3]	None	Aggregate	Aggregate	× [0,3)	× (2,∞)	1.0
Biener, Eling & Wirfs (2015)	B SAS (2015)	Individual	Individual	Individual	[5,∞)	× (2,∞)	4.0
Edwards, Hofmeyr & Forrest (2016)	A PVC (2016)	Individual	None	None	[5,∞)	(0,1]	3.0
'Untapped' (not featured in *model review*)	A PON (2019)	Individual	Individual	Aggregate	[5,∞)	(0,1]	4.5
	A ITRC (2018)	Individual	None	None	[5,∞)	(0,1]	3.0
	SERFF [4]	Aggregate	Aggregate	Aggregate	[5,∞)	(0,1]	3.5
	E NetD (2016)	Aggregate	Aggregate	Aggregate	[5,∞)	(0,1]	3.5
	IC3 [5]	Aggregate	Aggregate	Aggregate	[5,∞)	(0,1]	3.5
	C DIG (2013)	Individual	None	None	× [0,3)	(0,1]	2.0
	D BEA (2019)	None	None	Aggregate	[5,∞)	(0,1]	2.5
	VER (2019)	Aggregate	None	None	[5,∞)	(0,1]	2.5
	E LMA (2008)	None	Aggregate	Aggregate	[3,5)	(0,1]	2.5
	D OECD (2018)	None	None	Aggregate	[5,∞)	(0,1]	2.5
	ICPSR (2012)	Aggregate	Aggregate	Aggregate	[3,5)	× (2,∞)	2.0

Table A.1 Potential Suitability Score calculations Sources with comparable attributes or elements: *A – records breached*. *B – OR* and related loss data. *C –* online security attacks (e.g. DDoS). *D* – economic sources with exposure information (e.g. *GDPR*, *ICT* sector). Badges: green – individual level of detail; orange – aggregate; grey – no such data. Source fails minimum criteria (i.e. ×) if credibility less than 3 years; over 2 years out of date, or both. *PSS* range: 1–5. Notes: (1) (Pouget, Dacier & Pham, 2005). (2) WDID of the World Bank (2019). (3) ICSA reported by Bridwell (2004). (4) (NAIC, 2019). (5) (FBI, 2006).

A.3 Data sources

Table A.2 provides a summary of the abbreviated data sources considered in Chapter 3 and corresponding references (R.1).

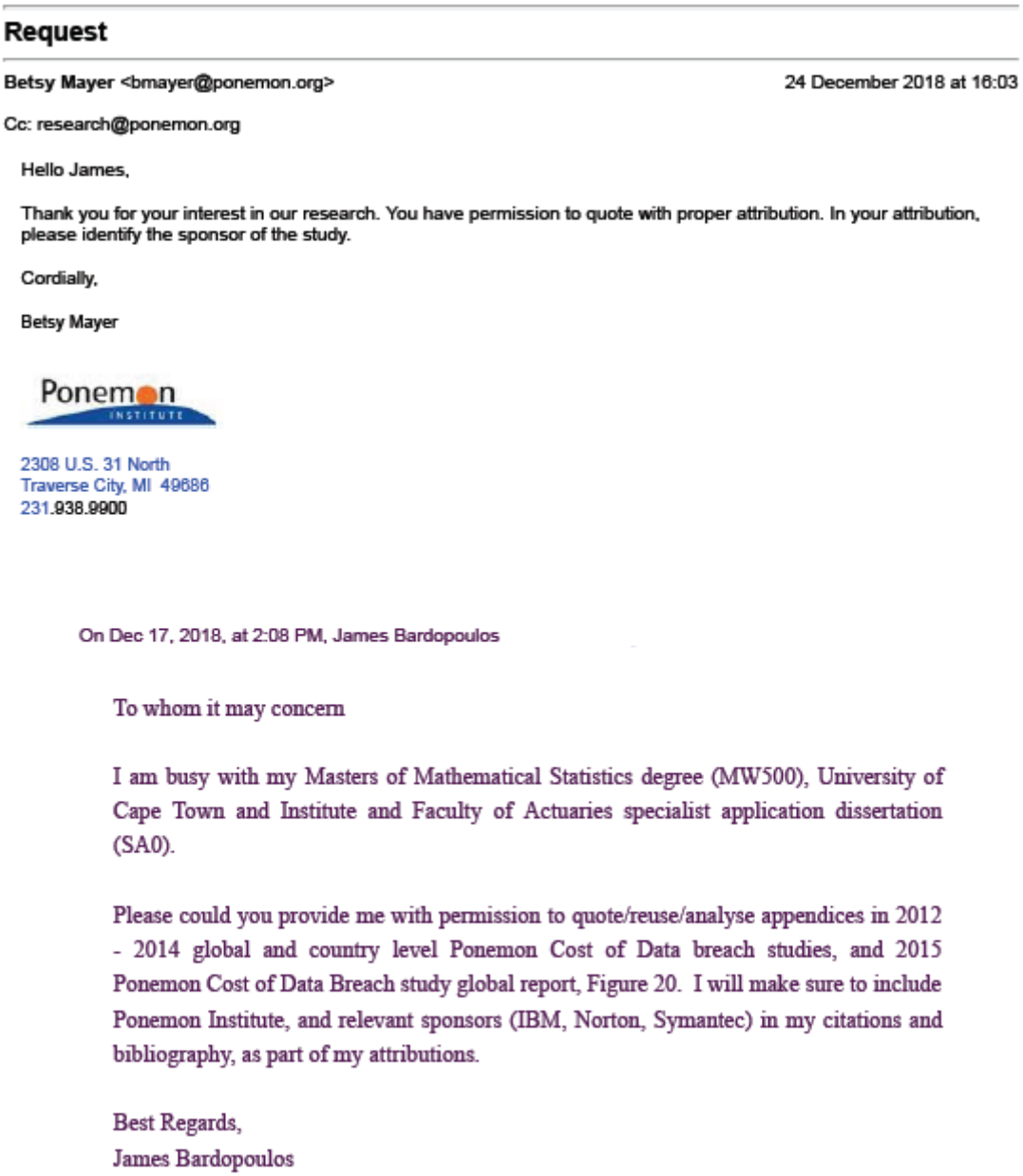
	Source [1]	Reference [2]
1	BEA	US Department of Commerce (2019)
2	BIS	Bank for International Settlements (2003)
3	DIG	Digital Attack Map (2013)
4	Honeypot	Pouget, Dacier & Pham (2005) [3]
5	IC3	Federal Bureau of Investigation (2006)
6	ICPSR	Inter-university Consortium for Political and Social Research (2012)
7	ICSA	Bridwell (2004) [4]
8	ITRC	Identity Theft Resource Center (2018)
9	LMA	Lloyd's Market Association (2008)
10	METI	Ministry of Economy Trade Industry (2004)
11	NetD	NetDiligence (2016)
12	OECD	Organisation for Economic Co-operation and Development (2018)
13	ORX	Operational Riskdata eXchange Association (2017)
14	PON	Ponemon Institute (2019)
15	PVC	Privacy Rights Clearinghouse (2016)
16	SANS	SysAdmin, Audit, Admin and Security (2019)
17	SAS	SAS (2015)
18	SERFF	National Association of Insurance Commissioners (2019)
19	VER	Verizon Data Breach Incident Response (2019)
20	WDID	World Bank (2019)

Table A.2 Reference guide for data sources Ordering: alphabetical (according to source, 1st column). Notes: (1) In-line, figure, and table citations or references; any of these that did not form part of the initial citation can be found in Abbreviations (p. x). (2) Reference used in R.1 list. (3) Courtesy Leurre.com, Eurecom, cited by Böhme & Kataria (2006). (4) Author of survey report.

Appendix B Selected data

Permission, from the copy-right holder, to quote Ponemon Institute (2012a–i, 2013a–j, 2014a–k, 2015g), is evidenced in Figure B.1.

B.1 Permission and fair usage



Request

Betsy Mayer <bmayer@ponemon.org> 24 December 2018 at 16:03

Cc: research@ponemon.org

Hello James,

Thank you for your interest in our research. You have permission to quote with proper attribution. In your attribution, please identify the sponsor of the study.

Cordially,

Betsy Mayer

Ponemon INSTITUTE

2308 U.S. 31 North
Traverse City, MI 49686
231.938.9900

On Dec 17, 2018, at 2:08 PM, James Bardopoulos

To whom it may concern

I am busy with my Masters of Mathematical Statistics degree (MW500), University of Cape Town and Institute and Faculty of Actuaries specialist application dissertation (SA0).

Please could you provide me with permission to quote/reuse/analyse appendices in 2012 - 2014 global and country level Ponemon Cost of Data breach studies, and 2015 Ponemon Cost of Data Breach study global report, Figure 20. I will make sure to include Ponemon Institute, and relevant sponsors (IBM, Norton, Symantec) in my citations and bibliography, as part of my attributions.

Best Regards,
James Bardopoulos

Figure B.1 Rights to quote Ponemon Institute (2012a–i, 2013a–j, 2014a–k, 2015g).

It can be noted that the accompanying request to "*reuse/analyse*" this data source was neither declined nor approved (Figure B.1) – as such, a four-factor checklist is considered to assess the extent of fair usage (Columbia university, 2015):

- *Purpose*: non-commercial research; largely 'transformative' – through a variety of statistical models and algorithms, new material has been produced and insight gained

(e.g. risk-adjusted *ILF*s; bimodal distortions due to correlated class effects); with the exception of the first 10 rows (which were rounded to the nearest thousand) – data from this source has not been reproduced in its original form

- *Nature*: factual, published, and legally accessible information
- *Amount*: extent of data utilised is appropriate for the intended educational purpose set out in §1.2 and does not represent a significant portion of the overall data and information embodied within this source
- *Effect*: considered to be low (or negligible) in terms of the copyright market, given the general limitations associated with the availability of (regarding restricted historical time periods over which such detail is available), and ability to assimilate, this data

B.2 Currency adjustments

Table B.1 compares $US (independent, online) rates online applied to Ponemon Institute (2012a–i, 2013a–j, 2014a–k, 2015g) with rates implied by country-level and global reports (which report the same values but in different currencies).

Currency	Implied rate	Online rate
Australian Dollar	0.847	0.798
Brazilian Real	0.446	0.328
Canadian dollar	0.828	0.767
Euro	1.388	1.137
Indian Rupee	0.016	0.016
Japanese Yen	0.010	0.008
Saudi Riyal	0.356	0.267
British Pound	1.404	1.526

Table B.1 US exchange rates Implied - according to Ponemon Institute (2015a–l), average cost per record data; online - as at 6 May 2015 (Currency.me.uk, 2008).

B.3 Key data fields

Table B.2 summarises relevant fields from the primary source (Chapter 3) in terms of their type (i.e. measure, factor), level of detail (aggregate, individual), and proposed definitions (interpreted from the descriptions and examples provided in the 2012–2015 reports).

	Field	Definition	Level of detail
	Company identifier	Surrogate key created and ascribed to each survey participant	Individual
Measures	Costs: *classes A–D* [1]	• Estimated financial loss in respect of (publicly disclosed) data breaches that occurred up to 12 months prior to surveys • Split by *classes A–D* (stages of a "data breach process") • *A–C* comprise elements of direct (related to defined activities) and indirect costs (*C* also includes costs associated with non-compliance) • *D* represents lost business costs: reduced sales due to diminished customer base due to reputational damage (over average customer lifetime) and business disruption (e.g. due to system outage)	Individual
Measures	Records	Number of lost or stolen items of personally identifiable information (typically ranges 1k–100k per organisation-year)	Individual
Measures	Churn	Percentage of customers that terminate their relationship due to breaches over the prior 12 months	Aggregate
Measures	Breach probability	Two-year probability of an organisation suffering at least one breach	Aggregate
Factors	Country	Country of establishment (up to 11 levels): Australia, Brazil, Canada, France, Germany, India, Italy, Japan, Middle East (Saudi Arabia and United Arab Emirates), UK, and US	Aggregate
Factors	Sector	Industry of operation (up to 16 levels, varies by country and year): communications, consumer, education, energy, financial, healthcare, hospitality, industrial, media, pharmaceutical, public, research, retail, services, technology, and transportation	Aggregate

Table B.2 Definitions for data fields (primary source) Individual detail (organisation level) tabulated for *A–D* (by year, 2012–2014) in Ponemon Institute (2012a–i, 2013a–j, 2014a–k) and graphed for *E* (note 1: total cost) in Ponemon Institute (2015g, fig. 20).

B.4 Inflation

Inflation periods (Table B.3) and methodology pertaining to the inflation rates by which costs are adjusted (Chapter 3) are included in this appendix.

Inflation Periods

As mentioned previously, inflation periods are determined by the average ‘interview’ date (by year, Table B.3) to 31st December 16.

Survey-year	Interview date	Inflation period
2012	30-Jul-11	5.4
2013	16-Jul-12	4.5
2014	14-Oct-13	3.2
2015	14-Oct-14	2.2

Table B.3 Inflation periods Years between average interview date and 31st December 2016, over which costs Ponemon Institute (2012a–i, 2013a–j, 2014a–k, 2015g) are inflated.

Average interview dates (Table B.3) are determined as the midpoint of respective 'interview periods' (given by survey-year; typically 10 months), based on the assumption that interviews are conducted uniformly.

Methodology

There are too few data points (i.e. years) to model inflation rates using conventional regression techniques (e.g. exponential, linear). Alternative methods such as constructing indices that reflect underlying cost drivers (e.g. class *C* – product discounts: CPI; credit monitoring and other fees; regulatory penalty adjustments; etc.) lie beyond the scope of the present research. Instead, annual inflation is determined using a simple and practical approach, based on the movement in the mean cost, by class, between mean interview dates for the 2012 and 2014 survey-years (i.e. for which *A*–*D* costs are available). In other words, compound inflation rates are assumed.

This implies an inflation rate for class *E* (i.e. average inflation weighted by uninflated *A*–*D* costs) which is applied to class *E* costs associated with the 2015 survey-year (i.e. to inflate from 14-Oct-16 to 31-Dec-16). More formerly, inflation, r_t, over $t \geq 0$ time units is derived as $r_t = (\frac{X_t}{X_0})^{\frac{1}{t}} - 1$, where X_0 and X_t are mean costs at time 0 and t respectively.

As mentioned, t, in this case, is taken as the number of years between average breach (or equivalently, 'interview') dates for survey years 2012 and 2014 (~ 2 years), with respective mean costs (for a given class) X_0 and X_t. Inflation rates derived this way were summarised, by class, in Table 3.3.

B.5 Homogeneity

Figure B.2 supports the notion that survey years are homoscedastic with respect to variance (noting the restricted range of the *x* and *y* axes), whilst Table B.4 compares the mix of countries (in terms of count, total cost) for 2012–2014 survey-years.

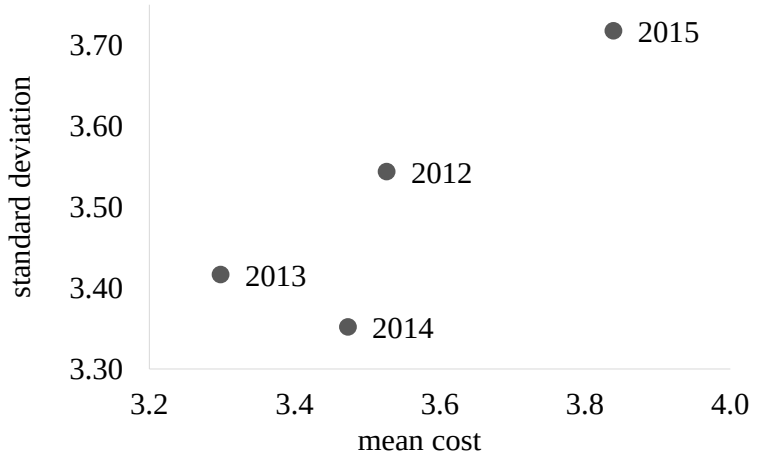


Figure B.2 Mean cost by standard deviation Based on *class E* costs from Ponemon Institute (2012a–i, 2013a–j, 2014a–k, 2015g), inflated to year 2016.

		Year				Year		
Country		2011	2013	2014		2011	2013	2014
US	Count	42%	36%	35%	Total cost	23%	19%	19%
UK		15%	15%	13%		17%	14%	13%
Germany		16%	16%	12%		12%	11%	10%
France		11%	10%	9%		11%	9%	9%
Australia		6%	6%	5%		11%	8%	7%
Italy		4%	5%	5%		9%	8%	7%
Japan		4%	6%	5%		7%	9%	8%
India		3%	3%	4%		10%	10%	9%
ME		0%	0%	7%		0%	0%	8%
Brazil		0%	3%	4%		0%	11%	10%
Total		100%	100%	100%		100%	100%	100%

Table B.4 Country mix by year Count % (number of participating organisations). Total cost % (based on *class E*, Ponemon Institute (2012a–i, 2013a–j, 2014a–k), inflated to year 2016).

Table B.4 shows broad similarity in the mix of countries by year (in support of §3.3), noting, however, the decline in (e.g. USA and UK) due to Brazil and *ME* joining.

Appendix C Supplementary theory

C.1 Rounding method (mass-dispersal)

The following method of *piecewise-linear* discretisation (for *FFT*, Chapter 5) is adapted from (Klugman, Panjer & Willmot, 2004: 182–183) and Wang (1998: 47) to allow for without limits and limited severities:

- Select $2^r, r > 0$ the number of points for *FFT* computation, and a suitable (constant) *span*, $h > 0$ ($h2^r$ should cover the maximum likely aggregate loss)
- Discretise severity X with *pdf*, f, and *cdf*, F, to calculate a vector of probabilities with 2^r elements $f(hk) = \Pr(X = hk), k = 0,1,...,2^r - 1$ as follows:

$$f(hk) = \begin{cases} F(\frac{h}{2}) & k = 0 \\ F(\frac{2k+1}{2}h) - F(\frac{2k-1}{2}h) & k = 1,2,...,2^r - 2 \\ 1 - \sum_{i=1}^{2^r-2} f(hi) & k = 2^r - 1 \end{cases} \quad \text{C.1}$$

Refer to Klugman, Panjer & Willmot (2004, sec. 6.6.5) for a description of an alternative method that preserves the mean of a continuous severity *cdf*.

Appendix D Results

D.1 Mean excess plots (supporting figures)

The *ME* plot for *class E* (Figure D.1) follow on from §5.2.1. As can be seen, there is an alternating positive and negative gradient, up to a threshold of ~93%.

This supports the light- and heavy- tailed Weibull *cdf*s in Table 5.1.

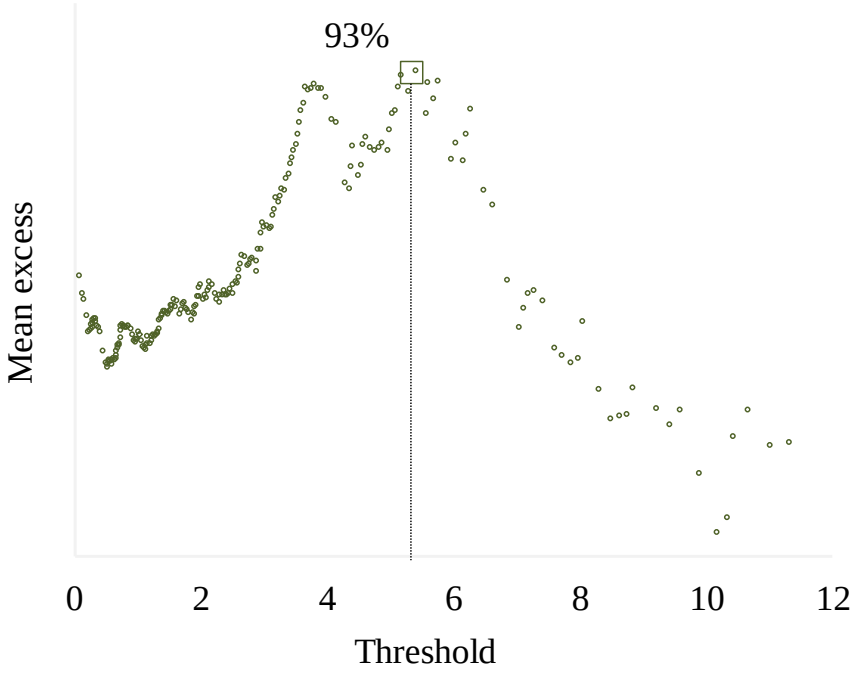


Figure D.1 Mean excess plot (*class E*) Separate investigation in support of a light-tailed *cdf* for large losses (for *spliced cdf* in respect of *class E*, §5.2.2). Costs based on Ponemon Institute (2012a–i, 2013a–j, 2014a–k), inflated to 2016.

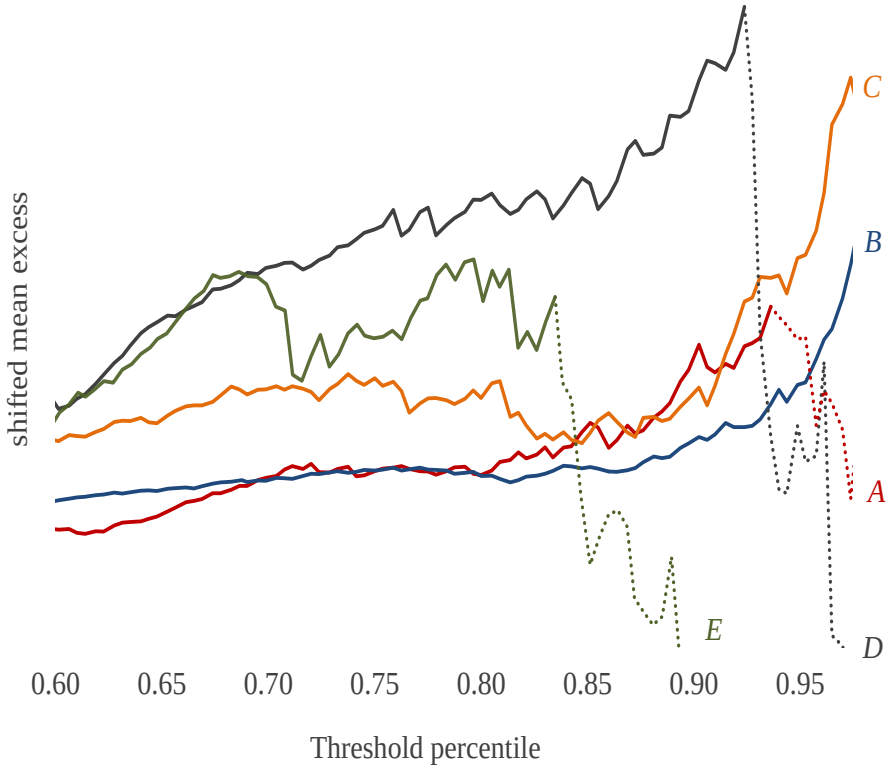


Figure D.2 Shifted mean excess (*classes A–E*) Follows considerations in §5.2.2; *ME*s are shifted to align with *A* at the 40% threshold. Costs based on Ponemon Institute (2012a–i, 2013a–j, 2014a–k), inflated to 2016. Dotted lines: above maximum percentiles identified (visually, using *ME*s) for large-loss *cdf*s.

Figure D.2 compares *ME*s in respect of all classes (*A*–*E*). For graphical convenience, these are vertically aligned by adding (or subtracting) a constant (across all percentiles for a given class) s.t. they intersect with *class A*'s *ME* at 40% threshold.

D.2 Candidate large-loss distributions

Distributions available in Vose (2019) software are used for running Algorithm 4.3 – these *cdf*s are fitted to large-loss severities for making selections for *classes A*–*E*.

Bradford	GEV	Log-Uniform
Burr	Inverse Gaussian	Maxwell
Chi	Johnson	Pareto
Chi-Squared	Kumaraswamy4	Pareto2
Dagum	Levy	Pearson5
Erlang	Lifetime2	Pearson6
Exponential	Lifetime3	Rayleigh
F distribution	Log-Laplace	Reciprocal
Fatigue	Lognormal	Weibull
Gamma	Log-Triangle	Weibull3

Table D.1 Candidate large-loss *cdf*s Number of parameters specified after *cdf* title (e.g. Weibull3 is the three parameter form of this model); number of parameters; based on available distributions in Vose (2019).

D.3 Large-loss model selection

Table D.2 follows from §5.2.2 and summarises the results of both runs of Algorithm 4.3 for the highest four combined scores. As can be seen, percentiles (i.e. splicing points) correspond to within 1%, and *cdf*s (for the top three) belong to the same family (e.g. Weibull light- or heavy- tailed, Burr).

	Rank	*A*	*B*	*C*	*D*	*E*
Percenitile	1	87.2%	77.5%	81.0%	92.1%	83.9%
	2	87.5%	78.4%	80.7%	92.0%	83.8%
	3	87.4%	78.0%	80.9%	91.9%	83.7%
	4	87.6%	78.1%	80.6%	92.4%	83.5%
Distribution	1	Weibull	Burr	Burr	Weibull*	Weibull*
	2	Weibull	Burr	Burr	Weibull*	Weibull*
	3	Weibull	Burr	Burr	Weibull*	Weibull*
	4	Weibull	Burr	Burr	Pearson	Weibull*

Table D.2 Top ranking percentiles and *cdf*s Rank refers to overall score (based on KS-ratio). Black font: both 1st and 2nd run (Algorithm 4.3); Red: only 2nd run. Asterisked are light-tailed Weibull *cdf*s. Underlying data: Ponemon Institute (2012a–i, 2013a–j, 2014a–k, 2015g), inflated to 2016.

D.4 Densities, limited moments

For beta and gamma families (Table D.3) gamma, Γ, and beta, B functions, and respective lower incomplete variations are defined as follows:

$$\Gamma(a)=\int_0^{\infty}u^{a-1}\exp(-u)du,\qquad \Gamma(a;b)=\int_0^{b}u^{a-1}\exp(-u)du\,f$$
$$\mathrm{B}(a,b)=\int_0^{1}u^{a-1}(1-u)^{b-1}du=\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)},\quad \mathrm{B}(a,b;c)=\mathrm{B}(a,b)\int_0^{c}u^{a-1}(1-u)^{b-1}du \qquad \text{D.1}$$

where $a,b,c>0$; $c<1$ (Klugman, Panjer & Willmot, 2004: 102, 627–629), noting that in this case, the incomplete gamma, $\Gamma(a,b)$, is not 'standardised' with divisor $\Gamma(a)$. In this table, limited moments for continuous distributions do not incorporate a shift (i.e. location parameter). For this, an adjustment can be applied as described in the following. Suppose random variable $Y=X+\phi$ has a shifted *cdf*, based on (non-negative) random variable X with location (i.e. 'shift') parameter $\phi>0$ (i.e. $Y\geq\phi$). Then limited moments for Y, when limit $l>\phi$ applies, can be determined analytically using $\mathrm{E}Y^{(l)k}=\mathrm{E}(X^{(l-\phi)}+\phi)^k$, assuming respective limited moments for X exist. This follows from the fact that $\min(X+\phi,l)=\min(X,l-\phi)+\phi$. For $\phi>l\geq 0$, $\mathrm{E}Y^{(l)k}=l^k$ by definition.

	Model or family	Notation, parameters	Density, distribution, support	Discrete: *pgf*, $P[t]$; mean, μ; variance, σ^2 Continuous: limited moments ($EX^{(l)k}$; $l > 0,\ k \in \mathbb{Z}^+$)	
Discrete	Binomial	$Bin(n, p)$ $n \in \mathbb{Z}^+,\ p \in (0,1)$	$f(x) = C_{(n,x)} p^x (1-p)^{n-x},$ $x = 0,1,...,n$	$P[t] = (1 - p + pt)^n$ $\mu = np;\ \sigma^2 = np(1-p)$	D.2
	Poisson	$Pois(\lambda)$ $\lambda > 0$	$f(x) = \lambda^x \exp(-\lambda)(x!)^{-1},$ $x = 0,1,...,n$	$P[t] = \exp(\lambda(t-1))$; $\mu = \sigma^2 = \lambda$	D.3
	Negative binomial	$NB(a,b)$ $a, b > 0$	$f(x) = \frac{\Gamma(a+x)b^x}{\Gamma(a)x!(1+b)^{a+x}},\ x = 0,1,...$	$P[t] = (1-b(t-1))^{-a}$ $\mu = ab;\ \sigma^2 = ab(1+b)$	D.4
Continuous	Lognormal	$LN(\mu,\sigma),$ $\mu \in \mathbb{R}, \sigma > 0$	$f(x) = \frac{\exp(-\frac{1}{2}s^2)}{x\sigma(2\pi)^{\frac{1}{2}}};\ s = \sigma^{-1}(\ln(x) - \mu)$ $F(x) = 1 - S(x) = \Phi(s),\ x > 0$	$EX^{(l)k} = \exp(k\mu + \frac{1}{2}k^2\sigma^2)\Phi(s - k\sigma) + l^k S(l)$	D.5
	Transformed beta (four parameter excluding shift)	$a, b, c, d > 0$ • Dagum: a = 1, Burr(b,c,d)* • GPD (a,b,d): c = 1 • Pareto (a,b): $c = d$ = 1 • Log-logistic (b,c): $a = d$ = 1	$f(x) = \frac{\Gamma(a+d)cx^{cd-1}b^{-cd}}{\Gamma(a)\Gamma(d)(1+(xb^{-1})^c)^{a+d}}$ $F(x) = 1 - S(x) = \mathrm{B}(d, a; p(x)),$ $p(x) = (1+(x^{-1}b)^c)^{-1};\ x > 0$	$EX^{(l)k} = \frac{b^k\Gamma(m)\Gamma(q)\mathrm{B}(m,q;p(l))}{\Gamma(d)} + l^k S(l)$, $k > -cd; m = d + kc^{-1}; q = a - kc^{-1}$	D.6
	Transformed gamma (three parameters, excluding shift)	$a, b, c > 0$ • Gamma: c = 1, G(a,b) • Weibull: a = 1, Weib(b,c) • Exponential: $a = c$ = 1, Exp(b)	$f(x) = \frac{cx^{ac-1}}{b^{ac}\Gamma(a)}\exp(-x^c b^{-c})$ $F(x) = 1 - S(x) = \Gamma(a, x^c b^{-c});\ x > 0$	$EX^{(l)k} = \frac{b^k\Gamma(a+kc^{-1}; x^c b^{-c})}{\Gamma(a)} + l^k S(l)$, $k > -ac$	D.7

Table D.3 Discrete and continuous distributions Limit l > 0 applies to random variable *X* for limited moments D.5–D.7 (Klugman, Panjer & Willmot, 2004, sec. A.2.1.1, A3.1.1) . *Dagum represented as Burr(*b*,*c*,*d*) – (i.e. *a* = 1) throughout present research to align with Vose (2019) parameterisation of Burr (ordinarily *d* = 1 for Burr). Location parameter, for a shifted *cdf*, is included after other applicable parameters *a*-*d* (limited moments, D.5–D.7, need to be adjusted accordingly).

D.5 LAS means and standard deviations

The caption for Table D.4 provides relevant detail for this appendix.

	***LAS* means (unadjusted)**		***PH* (5% margin at $10m)**		***LAS* std dev, covariance (for *variance principle*)**				
Limit	4.3: *IR*,*CR*	4.5–4.6	*Poiss-Weib*	*Poiss- LN*	4.3: *IR*	4.3: *CR*	4.5: scen 1	$C_V(\$10^9)$	4.6: ω=9%
10k	100 000	391 096	101 353	101 301	0	31 623	62 428	115	134 815
20k	200 000	767 166	202 705	202 603	0	63 246	123 309	4 411	264 815
25k	250 000	950 712	253 382	253 253	0	79 057	153 274	6 772	328 373
50k	500 000	1 828 808	506 764	506 497	0	158 114	298 874	25 011	633 433
100k	999 849	3 455 199	1 013 205	1 012 759	1 624	316 184	576 617	88 934	1 202 169
150k	1 497 524	4 943 660	1 515 254	1 518 084	10 186	473 668	839 109	181 387	1 726 626
200k	1 991 735	6 319 490	2 011 484	2 021 526	24 836	630 331	1 088 711	295 388	2 214 825
250k	2 483 846	7 594 237	2 501 808	2 522 120	42 323	786 600	1 326 032	425 292	2 670 205
300k	2 969 411	8 778 382	2 986 200	3 018 988	63 126	941 130	1 552 086	566 760	3 096 124
350k	3 445 954	9 883 448	3 464 662	3 511 371	89 220	1 093 353	1 768 258	716 805	3 496 471
400k	3 916 577	10 917 867	3 937 212	3 998 638	119 951	1 244 325	1 975 338	872 945	3 873 692
450k	4 382 500	11 880 318	4 403 878	4 480 272	154 035	1 394 402	2 172 439	1 031 729	4 227 003
500k	4 845 960	12 777 132	4 864 695	4 955 858	190 499	1 544 222	2 360 276	1 191 277	4 558 446
550k	5 305 883	13 607 428	5 319 702	5 425 074	228 587	1 693 367	2 538 140	1 348 884	4 867 437
600k	5 758 374	14 387 882	5 768 944	5 887 670	268 165	1 840 598	2 709 120	1 505 579	5 159 941
650k	6 203 687	15 126 000	6 212 467	6 343 463	309 516	1 986 045	2 874 295	1 661 329	5 438 488
750k	7 069 185	16 468 129	7 082 558	7 234 163	397 613	2 270 558	3 183 983	1 963 085	5 701 637
1.00m	9 106 431	19 267 936	9 161 348	9 337 440	649 988	2 952 151	3 874 314	2 668 129	5 950 197
1.25m	11 010 615	21 487 091	11 107 928	11 268 239	933 036	3 604 708	4 470 077	3 297 982	7 042 990
1.50m	12 762 291	23 256 801	12 929 380	13 037 680	1 231 614	4 219 535	4 983 342	3 843 812	7 937 816
1.75m	14 377 486	24 710 450	14 632 744	14 659 607	1 541 969	4 800 925	5 435 711	4 319 415	8 674 609
2.00m	15 880 430	25 940 542	16 224 885	16 148 106	1 860 371	5 355 351	5 843 966	4 740 677	9 299 010
2.25m	17 265 710	26 992 096	17 712 428	17 516 472	2 180 008	5 879 023	6 214 045	5 113 983	9 843 583
2.50m	18 558 647	27 882 298	19 101 718	18 776 829	2 501 595	6 379 680	6 544 992	5 439 162	10 322 796
2.75m	19 748 586	28 643 135	20 398 806	19 940 052	2 818 224	6 851 500	6 842 939	5 723 232	10 740 102
3.00m	20 847 092	29 295 547	21 609 439	21 015 811	3 129 766	7 297 641	7 111 430	5 970 998	11 106 826
3.25m	21 852 656	29 856 910	22 739 064	22 012 675	3 432 619	7 716 005	7 353 760	6 187 037	11 430 041
3.50m	22 772 730	30 331 965	23 792 829	22 938 223	3 726 128	8 108 252	7 568 466	6 371 963	11 715 819
3.75m	23 606 518	30 744 337	24 775 593	23 799 165	4 007 312	8 472 622	7 763 281	6 533 794	11 964 225
4.00m	24 378 561	31 097 080	25 691 935	24 601 447	4 281 987	8 818 551	7 937 228	6 673 260	12 185 639
4.25m	25 115 791	31 406 205	26 546 167	25 350 346	4 557 117	9 156 834	8 096 207	6 796 104	12 380 057
4.50m	25 819 744	31 678 392	27 342 342	26 050 558	4 831 050	9 487 095	8 241 903	6 904 712	12 554 949
4.75m	26 472 873	31 930 136	28 084 267	26 706 268	5 095 380	9 800 214	8 381 937	7 005 268	12 712 906
5m	27 081 894	32 164 446	28 775 518	27 321 216	5 351 405	10 098 536	8 517 081	7 098 914	12 862 691
6m	29 105 552	32 936 856	31 097 772	29 435 522	6 280 459	11 142 598	8 999 704	7 408 120	13 005 491
7m	30 719 249	33 475 468	32 842 468	31 112 939	7 121 211	12 044 869	9 378 779	7 624 503	13 502 695
8m	32 003 872	33 829 258	34 150 576	32 466 577	7 865 742	12 817 748	9 656 186	7 767 400	13 880 024

	LAS means (unadjusted)		*PH* (5% margin at $10m)		*LAS* std dev, covariance (for *variance principle*)				
Limit	4.3: *IR*,*CR*	4.5–4.6	*Poiss-Weib*	*Poiss- LN*	4.3: *IR*	4.3: *CR*	4.5: scen 1	C_V (10^9)	4.6: ω=9%
9m	32 993 248	34 049 705	35 129 611	33 574 783	8 497 884	13 456 206	9 847 403	7 857 166	14 148 395
10m	33 743 132	34 181 568	35 861 206	34 493 339	9 023 415	13 974 331	9 973 221	7 911 516	14 328 830
11m	34 305 432	34 258 603	36 407 129	35 262 913	9 453 919	14 389 678	10 053 653	7 943 833	14 444 933
12m	34 723 629	34 303 628	36 813 983	35 913 761	9 802 422	14 719 392	10 104 830	7 963 185	14 517 675
13m	35 032 581	34 330 770	37 116 838	36 468 801	10 081 687	14 978 938	10 138 243	7 975 206	14 563 128
14m	35 259 537	34 348 172	37 342 037	36 945 664	10 303 453	15 181 721	10 161 331	7 983 164	14 592 326
15m	35 425 439	34 360 268	37 509 322	37 358 102	10 478 116	15 339 071	10 178 545	7 988 854	14 612 212
16m	35 546 182	34 369 377	37 633 472	37 716 976	10 614 641	15 460 392	10 192 389	7 993 227	14 626 850
17m	35 633 715	34 376 688	37 725 528	38 030 959	10 720 607	15 553 378	10 204 206	7 996 779	14 638 489
18m	35 696 944	34 382 818	37 793 730	38 307 043	10 802 314	15 624 249	10 214 705	7 999 777	14 648 326
19m	35 742 469	34 388 100	37 844 221	38 550 922	10 864 929	15 677 981	10 224 261	8 002 367	14 656 989
20m	35 775 147	34 392 729	37 881 571	38 767 266	10 912 635	15 718 515	10 233 082	8 004 640	14 664 810
21m	35 798 537	34 396 833	37 909 182	38 959 935	10 948 784	15 748 949	10 241 296	8 006 656	14 671 978
22m	35 815 235	34 400 500	37 929 579	39 132 145	10 976 035	15 771 698	10 248 989	8 008 458	14 678 610
23m	35 827 125	34 403 799	37 944 637	39 286 590	10 996 480	15 788 631	10 256 226	8 010 080	14 684 785
24m	35 835 573	34 406 782	37 955 747	39 425 540	11 011 747	15 801 184	10 263 057	8 011 546	14 690 562
25m	35 841 561	34 409 493	37 963 940	39 550 920	11 023 100	15 810 455	10 269 523	8 012 878	14 695 987
26m	35 845 797	34 411 966	37 969 977	39 664 368	11 031 507	15 817 277	10 275 661	8 014 094	14 701 100
27m	35 848 787	34 414 232	37 974 424	39 767 288	11 037 709	15 822 281	10 281 500	8 015 208	14 705 932
28m	35 850 894	34 416 314	37 977 698	39 860 886	11 042 267	15 825 938	10 287 067	8 016 232	14 710 511
29m	35 852 375	34 418 234	37 980 107	39 946 205	11 045 606	15 828 604	10 292 385	8 017 176	14 714 860
30m	35 853 416	34 420 011	37 981 879	40 024 148	11 048 043	15 830 540	10 297 473	8 018 050	14 719 001
31m	35 854 145	34 421 658	37 983 181	40 095 499	11 049 817	15 831 943	10 302 350	8 018 860	14 722 951
32m	35 854 655	34 423 190	37 984 138	40 160 947	11 051 104	15 832 957	10 307 032	8 019 613	14 726 725
33m	35 855 011	34 424 618	37 984 841	40 221 091	11 052 036	15 833 688	10 311 533	8 020 316	14 730 338
34m	35 855 260	34 425 952	37 985 357	40 276 461	11 052 708	15 834 213	10 315 866	8 020 972	14 733 802
35m	35 855 433	34 427 201	37 985 735	40 327 522	11 053 192	15 834 590	10 320 041	8 021 586	14 737 128
36m	35 855 553	34 428 372	37 986 013	40 374 685	11 053 539	15 834 860	10 324 069	8 022 163	14 740 327
37m	35 855 637	34 429 473	37 986 216	40 418 317	11 053 788	15 835 053	10 327 960	8 022 704	14 743 405
38m	35 855 695	34 430 510	37 986 366	40 458 742	11 053 966	15 835 190	10 331 722	8 023 214	14 746 373
39m	35 855 735	34 431 487	37 986 475	40 496 248	11 054 092	15 835 288	10 335 363	8 023 695	14 749 237
40m	35 855 763	34 432 410	37 986 555	40 531 094	11 054 183	15 835 357	10 338 890	8 024 150	14 752 003
41m	35 855 782	34 433 283	37 986 614	40 563 511	11 054 247	15 835 406	10 342 310	8 024 579	14 754 678
42m	35 855 795	34 434 111	37 986 657	40 593 706	11 054 292	15 835 441	10 345 628	8 024 986	14 757 267
43m	35 855 804	34 434 895	37 986 688	40 621 865	11 054 324	15 835 465	10 348 850	8 025 373	14 759 775
44m	35 855 811	34 435 640	37 986 711	40 648 157	11 054 347	15 835 482	10 351 980	8 025 739	14 762 207
45m	35 855 815	34 436 348	37 986 728	40 672 732	11 054 363	15 835 495	10 355 025	8 026 088	14 764 566
46m	35 855 818	34 437 023	37 986 740	40 695 727	11 054 374	15 835 503	10 357 987	8 026 420	14 766 858
47m	35 855 820	34 437 665	37 986 749	40 717 267	11 054 382	15 835 509	10 360 872	8 026 736	14 769 084
48m	35 855 821	34 438 279	37 986 756	40 737 464	11 054 388	15 835 513	10 363 682	8 027 038	14 771 250
49m	35 855 822	34 438 864	37 986 760	40 756 420	11 054 392	15 835 516	10 366 422	8 027 326	14 773 357
50m	35 855 823	34 439 424	37 986 764	40 774 228	11 054 394	15 835 518	10 369 094	8 027 602	14 775 408
51m	35 855 823	34 439 960	37 986 766	40 790 973	11 054 396	15 835 520	10 371 702	8 027 866	14 777 407

	LAS means (unadjusted)		*PH* (5% margin at $10m)		*LAS* std dev, covariance (for *variance principle*)				
Limit	4.3: *IR*,*CR*	4.5–4.6	*Poiss-Weib*	*Poiss-LN*	4.3: *IR*	4.3: *CR*	4.5: scen 1	C_V (10^9)	4.6: ω=9%
52m	35 855 824	34 440 472	37 986 768	40 806 731	11 054 398	15 835 521	10 374 248	8 028 118	14 779 356
53m	35 855 824	34 440 964	37 986 770	40 821 574	11 054 399	15 835 521	10 376 736	8 028 360	14 781 256
54m	35 855 824	34 441 436	37 986 771	40 835 567	11 054 399	15 835 522	10 379 167	8 028 592	14 783 111
55m	35 855 824	34 441 888	37 986 771	40 848 768	11 054 400	15 835 522	10 381 544	8 028 815	14 784 922
56m	35 855 824	34 442 323	37 986 772	40 861 232	11 054 400	15 835 522	10 383 869	8 029 029	14 786 691
57m	35 855 824	34 442 742	37 986 772	40 873 010	11 054 400	15 835 523	10 386 144	8 029 235	14 788 420
58m	35 855 824	34 443 144	37 986 773	40 884 146	11 054 400	15 835 523	10 388 372	8 029 434	14 790 111
59m	35 855 824	34 443 531	37 986 773	40 894 684	11 054 400	15 835 523	10 390 553	8 029 624	14 791 764
60m	35 855 824	34 443 905	37 986 773	40 904 663	11 054 400	15 835 523	10 392 690	8 029 808	14 793 382
61m	35 855 824	34 444 265	37 986 773	40 914 118	11 054 401	15 835 523	10 394 785	8 029 985	14 794 966
62m	35 855 824	34 444 612	37 986 773	40 923 083	11 054 401	15 835 523	10 396 839	8 030 156	14 796 518
63m	35 855 824	34 444 947	37 986 773	40 931 588	11 054 401	15 835 523	10 398 853	8 030 321	14 798 038
64m	35 855 824	34 445 270	37 986 773	40 939 663	11 054 401	15 835 523	10 400 829	8 030 481	14 799 527
65m	35 855 824	34 445 583	37 986 773	40 947 334	11 054 401	15 835 523	10 402 768	8 030 635	14 800 988
66m	35 855 824	34 445 886	37 986 773	40 954 624	11 054 401	15 835 523	10 404 671	8 030 784	14 802 420
67m	35 855 824	34 446 178	37 986 773	40 961 558	11 054 401	15 835 523	10 406 541	8 030 928	14 803 825
68m	35 855 824	34 446 461	37 986 773	40 968 155	11 054 401	15 835 523	10 408 377	8 031 067	14 805 204
69m	35 855 824	34 446 735	37 986 773	40 974 437	11 054 401	15 835 523	10 410 180	8 031 202	14 806 558
70m	35 855 824	34 447 001	37 986 773	40 980 420	11 054 401	15 835 523	10 411 953	8 031 333	14 807 887
71m	35 855 824	34 447 258	37 986 773	40 986 123	11 054 401	15 835 523	10 413 696	8 031 460	14 809 192
72m	35 855 824	34 447 508	37 986 773	40 991 561	11 054 401	15 835 523	10 415 409	8 031 583	14 810 475
73m	35 855 824	34 447 750	37 986 773	40 996 748	11 054 401	15 835 523	10 417 095	8 031 702	14 811 735
74m	35 855 824	34 447 985	37 986 773	41 001 700	11 054 401	15 835 523	10 418 753	8 031 818	14 812 975
75m	35 855 824	34 448 213	37 986 773	41 006 429	11 054 401	15 835 523	10 420 384	8 031 930	14 814 193
76m	35 855 824	34 448 435	37 986 773	41 010 946	11 054 401	15 835 523	10 421 989	8 032 039	14 815 391
77m	35 855 824	34 448 650	37 986 773	41 015 264	11 054 401	15 835 523	10 423 570	8 032 145	14 816 570
78m	35 855 824	34 448 859	37 986 773	41 019 393	11 054 401	15 835 523	10 425 126	8 032 249	14 817 730
79m	35 855 824	34 449 063	37 986 773	41 023 343	11 054 401	15 835 523	10 426 658	8 032 349	14 818 871
80m	35 855 824	34 449 261	37 986 773	41 027 124	11 054 401	15 835 523	10 428 168	8 032 446	14 819 995
81m	35 855 824	34 449 454	37 986 773	41 030 743	11 054 401	15 835 523	10 429 655	8 032 541	14 821 101
82m	35 855 824	34 449 641	37 986 773	41 034 210	11 054 401	15 835 523	10 431 121	8 032 634	14 822 191
83m	35 855 824	34 449 824	37 986 773	41 037 532	11 054 401	15 835 523	10 432 565	8 032 724	14 823 264
84m	35 855 824	34 450 002	37 986 773	41 040 716	11 054 401	15 835 523	10 433 989	8 032 811	14 824 321
85m	35 855 824	34 450 175	37 986 773	41 043 770	11 054 401	15 835 523	10 435 393	8 032 897	14 825 363
86m	35 855 824	34 450 344	37 986 773	41 046 700	11 054 401	15 835 523	10 436 777	8 032 980	14 826 390
87m	35 855 824	34 450 509	37 986 773	41 049 511	11 054 401	15 835 523	10 438 142	8 033 061	14 827 402
88m	35 855 824	34 450 669	37 986 773	41 052 210	11 054 401	15 835 523	10 439 489	8 033 140	14 828 400
89m	35 855 824	34 450 826	37 986 773	41 054 802	11 054 401	15 835 523	10 440 818	8 033 218	14 829 384
90m	35 855 824	34 450 979	37 986 773	41 057 292	11 054 401	15 835 523	10 442 129	8 033 293	14 830 355
91m	35 855 824	34 451 128	37 986 773	41 059 684	11 054 401	15 835 523	10 443 422	8 033 366	14 831 312
92m	35 855 824	34 451 274	37 986 773	41 061 985	11 054 401	15 835 523	10 444 700	8 033 438	14 832 256
93m	35 855 824	34 451 417	37 986 773	41 064 197	11 054 401	15 835 523	10 445 960	8 033 508	14 833 188
94m	35 855 824	34 451 556	37 986 773	41 066 325	11 054 401	15 835 523	10 447 205	8 033 577	14 834 108

	LAS means (unadjusted)		*PH* (5% margin at $10m)		*LAS* std dev, covariance (for *variance principle*)				
Limit	4.3: *IR*,*CR*	4.5–4.6	*Poiss-Weib*	*Poiss- LN*	4.3: *IR*	4.3: *CR*	4.5: scen 1	C_V (10^9)	4.6: ω=9%
95m	35 855 824	34 451 692	37 986 773	41 068 372	11 054 401	15 835 523	10 448 434	8 033 644	14 835 016
96m	35 855 824	34 451 824	37 986 773	41 070 343	11 054 401	15 835 523	10 449 648	8 033 709	14 835 912
97m	35 855 824	34 451 954	37 986 773	41 072 241	11 054 401	15 835 523	10 450 847	8 033 773	14 836 797
98m	35 855 824	34 452 081	37 986 773	41 074 068	11 054 401	15 835 523	10 452 032	8 033 836	14 837 671
99m	35 855 824	34 452 205	37 986 773	41 075 829	11 054 401	15 835 523	10 453 202	8 033 897	14 838 534
100m	35 855 824	34 452 327	37 986 773	41 077 526	11 054 401	15 835 523	10 454 358	8 033 957	14 839 386

Table D.4 Mean and standard deviation of *LAS*s Supporting means, standard deviations for Models 4.3 - 4.6 (variance principle) and compound-*Poisson PH-transform LAS*s (§5.3.3: Table 5.6, Figure 5.4; §5.3.5: Table 5.10, Table 5.11).

D.6 Risk-adjustment parameters

Table D.5 provides risk-adjustment parameters for *variance principle* and *PH-transform* methods, by class and risk environment, underlying risk-adjusted *LAS*s (§5.3.3).

		Risk environment		
	Model	*Low*	*Medium*	*High*
Variance principle	4.3 (IR)	2.1E-08	1.0E-07	2.1E-07
	4.3 (CR)	8.6E-09	4.3E-08	8.6E-08
	4.5 (1)	3.3E-08	1.6E-07	3.3E-07
	4.5 (2)	2.0E-08	1.0E-07	2.0E-07
	4.5 (3)	1.4E-08	7.2E-08	1.4E-07
	4.6	1.2E-08	6.1E-08	1.2E-07
PH transform	*Weibull*	1.05	1.24	1.49
	Lognormal	1.04	1.23	1.47
	Risk margin	5%	25%	50%

Table D.5 Risk-adjustment parameters *low–high* risk margins (i.e. risk-adjusted *LAS*, relative to mean) at limits: $2.5m (Models 4.5–4.6) and $10m (Model 4.3 and *PH transforms*); *PH parameters* applied to (Weibull, lognormal) severity cdfs (based on *MLE* fit to class *E*) and loss count *cdf* (Poisson mean 10, for all models and methods).

D.7 Lognormal vs. spliced-severity *cdf*s (*class E*)

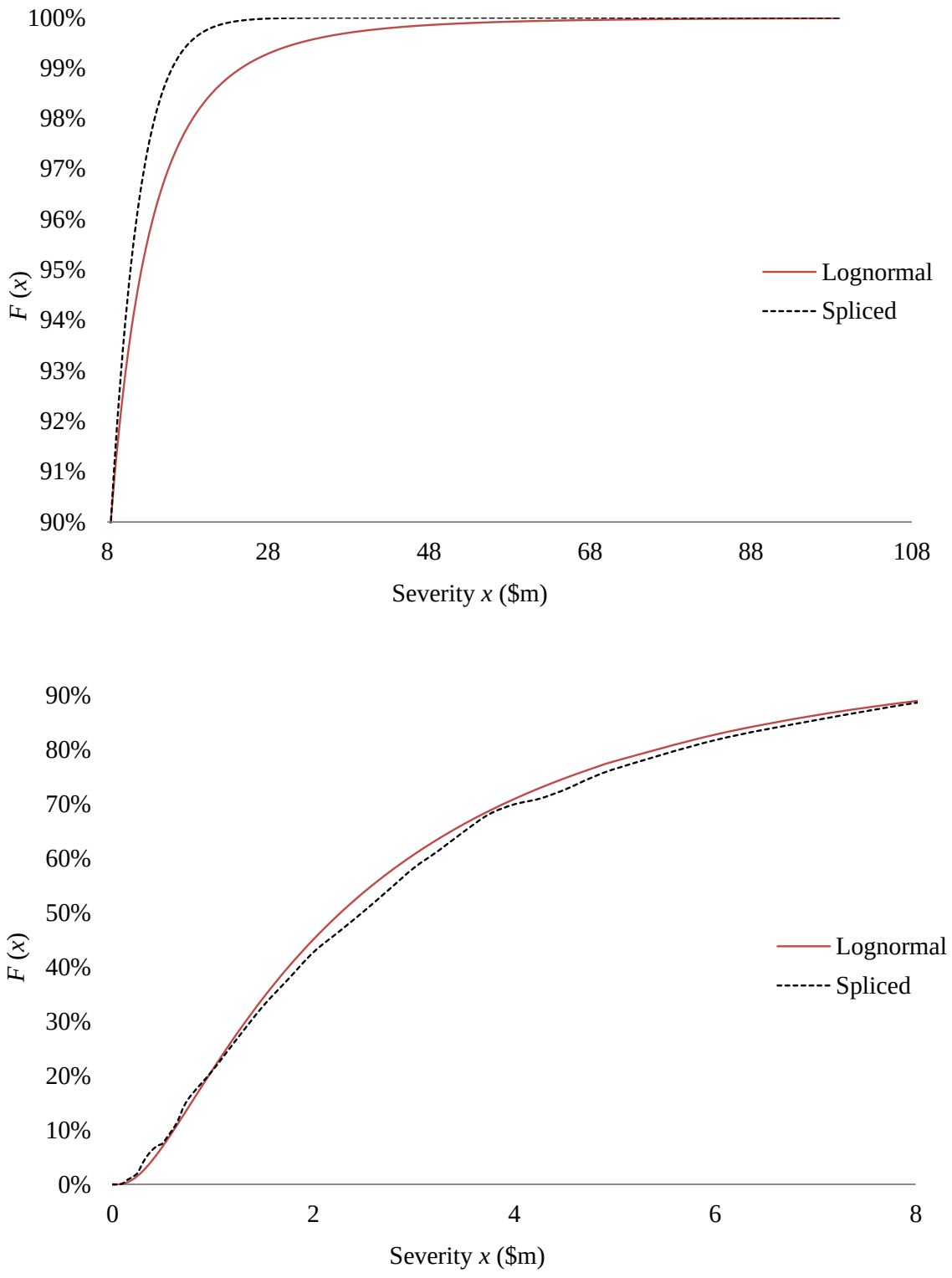


Figure D.3 Lognormal and spliced cdfs Based on *class E*; underlying discussions pertaining to observation 1 (§5.3.3, p. 5.22): top: above 90% (*spliced*: *light-tail* Weibull); bottom: below 90%.

D.8 Covariance and standard deviation (Model 4.5)

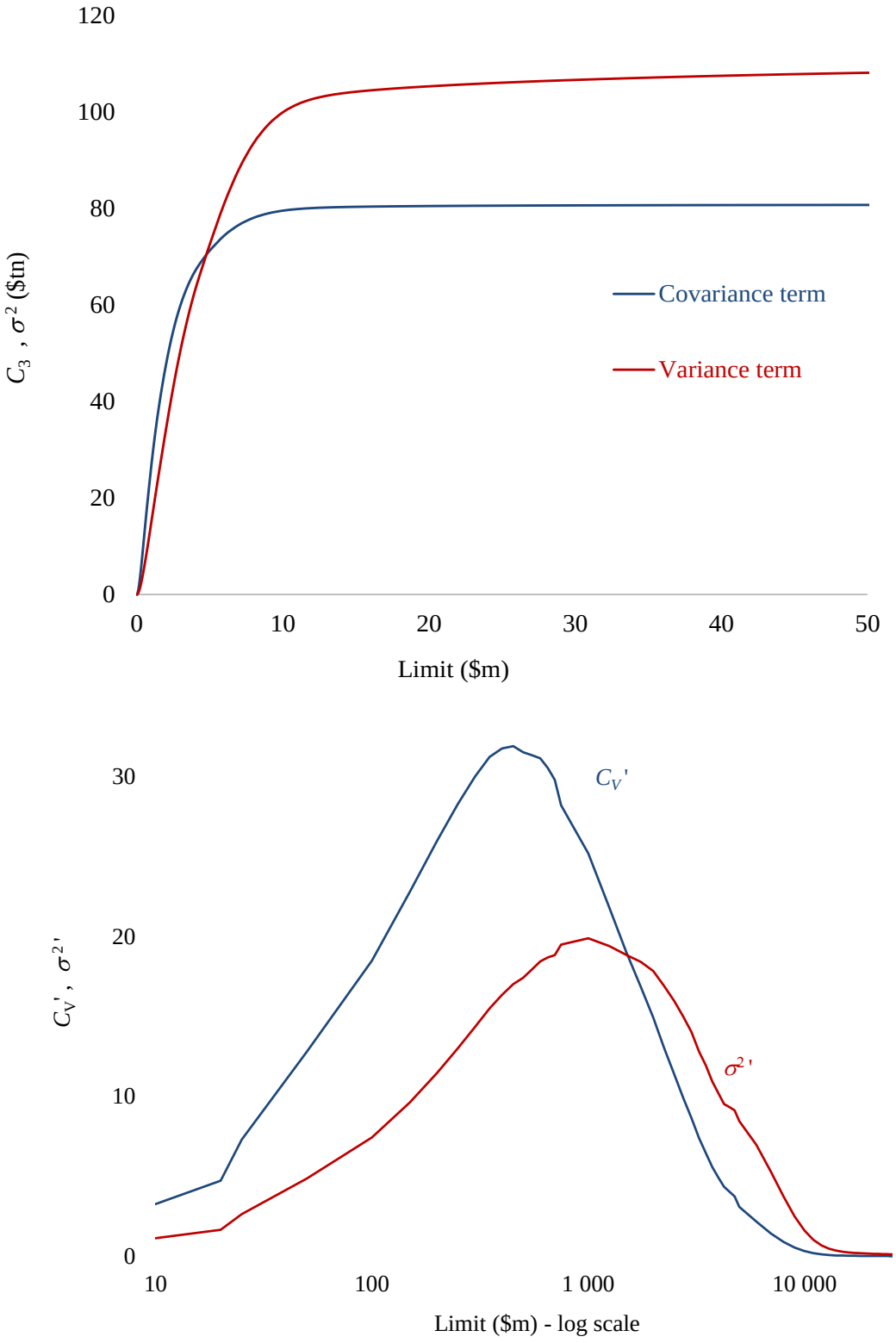


Figure D.4 Model 4.5 Covariance, variance, and gradients Top: *covariance* and *variance* associated with Model 4.5 Scenario 3; bottom: associated gradients for these terms. Related discussion: §5.3.3. Costs based on Ponemon Institute (2012a–i, 2013a–j, 2014a–k), inflated to end of 2016 year.

Appendix E Cyber-risk and Insurance

E.1 Cyber-risk evolution

This section provides additional information pertaining to the historic evolution of *cyber-crime* (Figure E.1) in parallel with some of the following:

- The role played by cyber-criminals and the media with regard to data breaches and consequent legal actions
- Developments in insurance policy-wordings brought about by court appeals regarding legal liability as a result of '*hidden cyber-exposures*' (Appendix E.3)
- The effect of the internet on *cyber-risk*, and data available for modelling it statistically

Refer to (Meyers, Powers & Faissol, 2009) for a more detailed historic background of the development of *cyber-risk* and the internet.

Digital Age

The Whatsapp (2019) precursor of the 19th-20th centuries was Morse code – known as the first digital code – which could relay digital data using a discrete representation of information. Since then, IT has undergone a *Digital Revolution*, known as *The Third Industrial Revolution*, which follows the *Second Industrial Revolution* that brought with it petroleum, automobiles, airplanes, steel and electricity. A timeline of the *Digital Age*, before the internet, is depicted in Figure E.1.

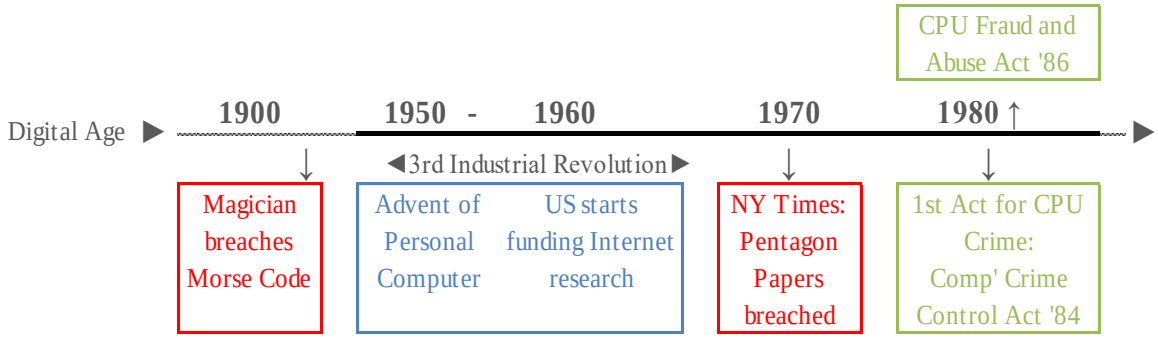


Figure E.1 Cyber-evolution during digital age Colour font indicates *cyber-crime* (Haney, 1972; Marks, 2011), technological developments (Defense Communications Agency, 1985), and regulation (Meyers, Powers & Faissol, 2009).

Figure E.1 highlights the following historical events in the evolution of *cyber-crime*:

- 1900s: The first report of a data breach was in 1903, when magician and scientist Maskelyne sent insulting Morse code out to an auditorium, in an attempt to disprove the concept of private and secure communication (Marks, 2011)
- 1960s: The US Department of Defense starts funding research into a technological precursor to the internet, known as the *ARPANET* (Defense Communications Agency, 1985)
- 1970s: the Pentagon breach is leaked to New York Times and widely distributed (Haney, 1972)
- 1980s: Meyers, Powers & Faissol (2009) regards the first legal attempt to address *cyber-crime* as the Comprehensive Crime Control Act (1984), followed by the Fraud and Abuse Act (1986) which formerly classifies breaking into a computer system as a crime in the USA

Information Age

It was only at the "beginning of last quarter of the 20th century" (Princeton university, 2009) that the Information Age erupted, with the emergence of computer telecommunication networks that allow the exchange of data between computers which are not directly connected, on account of the internet. Figure E.2 shows a timeline of events after the internet entered the public domain.

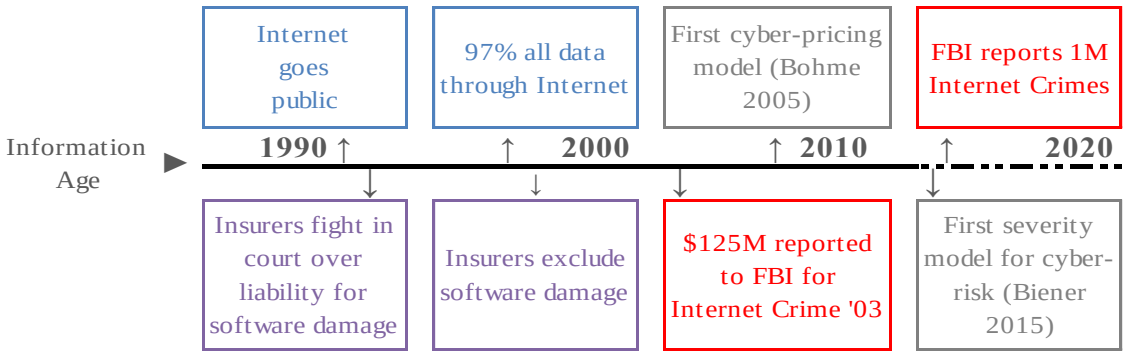


Figure E.2 Cyber-evolution during information age Colour font indicates *cyber-crime* (Federal Bureau of Investigation, 2006), technological developments (Hilbert & López, 2011; Feenberg & Friesen, 2012); insurance and legal implications (Baer & Parkinson, 2007; Anderson, 2013); cyber-risk model development (Böhme, 2005; Biener, Eling & Wirfs, 2015).

The following are key developments shown in Figure E.2:

- 1990: *ARPANET* – originally conceived as an indestructible global computer network that could not be destroyed by any single point of attack – is decommissioned (Feenberg & Friesen, 2012)
- 1991: The *World Wide Web* (WWW) goes public (Bryant, 2011), and policy wording exclusions for non-physical software damages start emerging to address *hidden* (i.e. unanticipated) *cyber-exposures* (e.g. due to gaps in specific cyber-related exclusions; implied coverage in 'all-risks' insurance policy) following legal disputes such as the court of Appeals of Minnesota in Retail Systems, Inc. v. CNA Insurance Co. (Anderson, 2013)
- 1990–2000: Specialised (standalone) *cyber-insurance* policies (evolving from Professional Liability covers) are developed with coverage against losses caused by computer viruses or other malicious code, destruction or theft of data, denial of service attacks, business interruption, and liability from e-commerce or other forms of network IT failure (Baer & Parkinson, 2007), although Moore (2012) describes cyber-security insurance as being commercially available from as early as the late 1970s
- 2000: According to Hilbert & López (2011) of the American Association for the Advancement of Science (2019), self-proclaimed as the "*world's largest general Scientific society*", research by Lyman et al. (2005) suggests that 97% of all digital data is communicated through the internet (i.e. within 10 years of it having gone public); also, cyber-media liability policies develop to cover perils such as viruses, network security failure, and unauthorised access
- 2003: Internet related crime costs an estimated $125m (not inflation adjusted) in the USA (Federal Bureau of Investigation, 2006) and the same year mandatory disclosure requirements for data breaches are introduced by California legislative information (2016)
- 2005: One million internet crimes in the US are reported to IC3 (FBI, 2006), around the same time some of the first frequency models (based on empirical *cyber-risk* data) can be found (Böhme, 2005), followed by the examples of empirical severity models (Biener, Eling & Wirfs, 2015)

Half a century in the making, from the antics of a magician to the epidemic global fear that a make-believe Y2K bug could mean the end of technological time; *cyber-risk* has become

an actuality. Society has encountered the classic catch-22 paradigm – one where the very defence mechanisms designed to protect against *cyber-risk* can also be a source of *cyber-risk* (§2.2- *correlated failures*). The world is utterly reliant upon what is believed to be the successor of an indestructible globally connected computer network: one that governs almost all telecommunication; one that breeds the indefatigable *cyber-risk*; one with *interdependent* security decisions (§2.2); one through which privacy of identity and economic livelihood are compromised on an everyday basis. This thing is called the *internet*.

E.2 Product features and coverage

As one might expect of a diverse range of product offerings, there are several types of coverage available in cyber-insurance. For instance, cyber-insurance policies often provide cover against *first-party* losses on *losses-occurring* or *losses-discovered* bases, and *third-party coverage* on a *claims-made* basis, as the following defines:

- *Losses-occurring* policies meet claims in respect of losses that occur during the policy period – this basis of cover is often used for *cyber-extortion* and *network interruption* insurance
- *Losses-discovered* policies meet claims in respect of losses that are first discovered during the policy period – coverage for loss of assets (e.g. due to computer fraud) and remediation costs (e.g. due to data breaches) can often be found on this basis
- *Claims-made* policies meet claims that are first made and reported during the policy period, irrespective of when the underlying incident occurred (Marker & Mohl, 1980), subject to other conditions such as the *discovery period* (specified within the *sunset clause*) within which the insured must notify the insurer of a claim, *retroactive date* (before which time, claim incidents are excluded) and *ERP* to cover claims reported after (in respect of incidents that occurred during) the original policy term

For specific considerations pertaining to cyber-liability (claims-made) policies, refer to US Department of Homeland Security (2012) on the topic of *retroactive dates*, and LaCroix (2016), who describes a possible coverage gap that can arise when this does not predate the point of 'failure to maintain IT security' (e.g. last computer configuration; software updates; etc.). For *ERP* (particularly relevant for claims-made cyber-liability), Betterley

Risk Consultants (2017) recommend, for insured parties, 'bilateral' provisions which are more flexible than 'one-way' provisions as insured can exercise the option to purchase (e.g. by cancelling the policy).

Product variations (Appendix E.3); perils (Appendix E.4); risk and rating factors (Appendix E.5); exclusions (Appendix E.6); and exposure measures (Appendix E.7).

E.3 Product variations

As mentioned there are numerous forms and types of products (§3.1) – according to Baer & Parkinson (2007), however, businesses generally purchase standalone coverage. Notwithstanding, a review by Risk Management Solutions [RMS] (2016), of 26 products, found virtually no commonality in terms of coverage (i.e. number, types) – indeed, 19 distinct ('primary') categories of coverage were identified in respect of these.

Some insurers provide *first-party coverage* to customers of the insured, whilst others offer 'services only'. Further, some products (e.g. 'cyber-security', 'privacy notification', or 'crisis management expense') only cover first-party losses; whilst others (e.g. technology *Errors and Omissions*, E&O) protect against third-party liability (e.g. clients' negligence claims; civil damages); others still (e.g. 'network security'; 'privacy liability') cover elements of both of these (Floresca, 2014; Sharp, 2016). For firms, the suitability of such products depends on numerous factors, key examples of which pertain to data (e.g. sensitivity, storage); IT infrastructure; nature of business, and regulatory environment. Indeed, increased uptake has been noted for businesses that hold confidential data; rely heavily on IT (e.g. systems, website) to transact; and that deal with electronic payments (Lloyd's, 2015). These and other factors, which have a bearing on the level of risk and thus insurance premium, are considered further in Appendix E.5.

Cyber-exposures

As for risk and coverage, (cyber-) exposure (i.e. exposure to *cyber-risk*) can be classified as *first-party* or *third-party*, both of which may represent what is referred to as '*hidden*' or '*silent*' *exposure*. To begin with, the concept of *first-party* and *third-party exposure* is described in the context of an example that considers various parties associated with a *data breach*. The following example introduced several terms (*records*, *data subject*, *data*

owner, *data custodian*) – precise definitions can be found in data breach legislation (California Office of Privacy Protection, 2012):

- A bank customer (*data subject*) entrusts its residential address (record) to a financial bank (*data owner*, in this case, also the insured)
- The bank stores this information on an information system comprising IT assets owned by the bank, but maintained by an outsourced third-party IT provider (*data custodian*)

In this case, the potential for the IT provider to suffer a cyber-attack can represent a *third-party exposure* for the bank, which can result in both *first-party losses* (due to damaged bank IT assets), as well as *third-party liability* (due to customer information being breached). Similarly, the potential for a breach of IT security within the data owner (representing a *first-party exposure*) can lead to both *first-party losses* and *third-party liability* insurance claims (under a cyber-insurance policy).

An alternative definition for *cyber-exposures*, provided by RMS (2016), are policies that could potentially trigger claims in the event of a cyber-incident. RMS (2016) then goes on to classify *cyber-exposures* in the market under headings such as standalone cyber-covers, endorsements (i.e. coverage extensions to traditional insurance products), and *silent* (or *hidden*) *cyber-exposures* (also known as *silent cyber*).

Silent cyber refers to potential cyber-related losses from policies not specifically designed to cover such losses, and can arise from gaps in specific cyber-exclusions and policies without cyber-exclusions (e.g. an *all-risks* insurance policy may not exclude specific *cyber-perils*). One might regard this as a type of *latent claim* exposure in that it can give rise to claims that "[result] *from perils or causes of which the insurer is unaware of at the time of writing a policy*" (Michaelides et al., 1997). *Latent claims,* however, are typically associated with much longer reporting and settlement delays (Forfar & Raymont, 2002).

E.4 Cyber-perils

This section describes *cyber-perils* (i.e. probable causes of *cyber-loss*) in the context of a taxonomy of operational cyber-security risks proposed by (Cebula & Young, 2010) and contemplated by Biener, Eling & Wirfs (2015) in related material. For this purpose, *cyber-*

perils are classified according to the human (malicious or otherwise) and technological interventions (e.g. system failure), processes and exogenous events.

Alternatively, *cyber-perils* could be classified according to *first-party* and *third-party risks*, for instance, *cyber-perils* that give rise to *first-party* claims may include: malicious or accidental destruction of data, denial of service attacks, cyber-extortion threats, and system failures; *third-party* claims may be caused by privacy or security breaches, misuse of personal data, defamation or slander, and transmission of malicious content. Other coverage triggers (i.e. *cyber-perils*) in respect of *data privacy* insurance (refer to *party coverage* examples, Appendix E.2), according to (Betterley Risk Consultants, 2017), include failure to secure data, loss attributable to an employee, and third-party acts.

Ponemon Institute (2012d: 6) reports, for many countries, *malicious attacks* and *negligent employees* as being the main causes of data breaches, and Ponemon Institute (2015g: 10) finds *malicious attacks* as being the most common (i.e. frequent) and costly. However, *negligent employees* only represent a subset of *inadvertent events*, which, according to the UK Government and Industry (2015), are more frequent but less severe (in terms of their impact on businesses) than *malicious* events. In terms of *malicious* events, *cyber-attacks* reportedly have a similar likelihood but a higher severity compared to *identity-theft* and *cyber-fraud* (World Economic Forum, 2015).

E.5 Risk and rating factors

Risk factors (factors that influence the level of risk) and rating factors (risk factors or proxies for risk factors that are practical, objective, measurable, and acceptable for market use) are used for reducing heterogeneity, and allow differential rates that are commensurate with the level of risk to be charged. Ideally, such factors should not be correlated with one another. An example of a risk factor could be 'network security breach risk'. However, as this would be difficult to objectively measure and verify proxies may be used instead, for instance:

- Type of firewall protection and other measurable factors relating to the level of IT security such as existence of IT security certificate and issuing authority
- Prior network security breaches

A number of rating factors are used in practice, some of which include security policy, *third-party exposures* (e.g. IT service provider, backup and archiving services), business continuity and incident response plans, intrusion testing, and the level of cover or optional coverages (e.g. defence costs, reward expenses, regulatory fines and penalties, etc.), (Selleck, 2015). Refer to *SERFF* (NAIC, 2019), where (publicly accessible) insurer rate plans can be found, and Romanosky et al. (2017) who provide information pertaining to the rating of cyber-insurance policies including other good examples of rating factors used in practice.

E.6 Common exclusions

Typically, *third-party liability* for death or bodily injury (cyber-related or otherwise) can be covered under a relevant traditional commercial general liability policy (as opposed to a cyber-insurance policy). A key misconception that may exist among many firms is that their traditional insurance arrangements provide suitable protection against cyber-losses; however, this is not necessarily the case due to common exclusions such as:

- *Electronic Data Exclusion* (*NMA2914*) excludes non-physical damage (Marsh, 2014)
- Network downtime caused by *cyber-crime* (i.e. resulting in lost business costs) is typically excluded in the business interruption *coverage section* of a traditional commercial fire policy (Anderson, 2013; UK Government and Industry, 2015)

E.7 Exposure measures

An exposure measure is a quantity that represents the basic unit of risk underlying an insurance premium (which, oftentimes, is expressed as a 'premium rate', per unit of exposure, per term of policy). Examples for cyber-insurance (which can also be used as rating factors, Appendix E.5) include:

- Internet revenue: as this relates to the potential for lost business costs and reputational damage (Great American Insurance, 2011; Wolfrom, Little & Rielley, 2015)
- Size of the workforce: this relates to handling customer complaints and notifications; however, the level of risk can vary for firms with the same workforce size, and so additional rating factors would be needed to explain this risk, for instance, manual

workers would not typically be involved in handling customer notifications (Philadelphia Indemnity, 2015)

- The number of records of confidential customer information: (Sedano & Rodriguez, 2015); the number of records should be verifiable, especially if stipulated in specific audit requirements (Treasury, 2011)