Bernhard Christian Kübler

Risk Classification by Means of Clustering

Schriften zum Controlling, Finanz- und Risikomanagement

Herausgegeben von Andreas Brieden, Thomas Hartung,
Bernhard Hirsch und Andreas Schüler

PETER LANG
Internationaler Verlag der Wissenschaften

Determining risk-adequate insurance premiums is a core issue in actuarial mathematics. This study is specifically concerned with identifying convenient partitions of (general) insurance collectives such that the resulting tariff classes are homogeneous to a maximum extent and – on the other hand – yet large enough to allow for the occurrence of the group balance concept and to end up with reliable estimates of the moments of the claim size distributions. Therefore, the author develops an innovative classification algorithm utilizing a multidimensional cluster approach combined with credibility-theoretical implications. Its construction stems from involving the entire claim information of risks simultaneously and in a suitable manner, and particulary from obtaining optimality regarding the cluster criterions. Under certain conditions, commonly used cross classification schemes are shown to be a particular case of the new approach. Besides desirable theoretical benefits like its generalizing established cross classification systems, an empirical investigation also suggests the practical superiority of the new algorithm.

Bernhard Christian Kübler, born in 1978 in Backnang; 1998–2004 University of Bonn, University of Cologne: Diploma in Economics; 2004–2005 University of Hull: Master of Science in Mathematical Finance; 2005–2009 Free University Hagen: Intermediate Diploma in Mathematics; 2005–2009 Institute of Statistics and Risk Management, Universität der Bundeswehr München: Scientific Assistent.

www.peterlang.de

Risk Classification by Means of Clustering

Schriften zum Controlling, Finanz- und Risikomanagement

Herausgegeben von Andreas Brieden, Thomas Hartung,
Bernhard Hirsch und Andreas Schüler

Band 4

PETER LANG
Frankfurt am Main · Berlin · Bern · Bruxelles · New York · Oxford · Wien

Bernhard Christian Kübler

Risk Classification by Means of Clustering

PETER LANG
Internationaler Verlag der Wissenschaften

Bibliographic Information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data is available in the internet at http://dnb.d-nb.de.

Zugl.: München, Univ. der Bundeswehr., Diss., 2009

D 706
ISSN 1867-027X
ISBN 978-3-653-00279-9

© Peter Lang GmbH
Internationaler Verlag der Wissenschaften
Frankfurt am Main 2010
All rights reserved.

www.peterlang.de

Vorwort

Die vorliegende Dissertation entstand während meiner Tätigkeit als wissenschaftlicher Mitarbeiter an der Professur für Statistik, insbesondere Risikomanagement an der Universität der Bundeswehr München.

Mein Dank gilt meinem Doktorvater Herrn Professor Dr. Andreas Brieden, der durch seine erstklassige Betreuung entscheidend zum Gelingen der Arbeit beigetragen hat. Herrn Professor Dr. Thomas Hartung danke ich für die Übernahme des Zweitgutachtens.

Besonderer Dank gilt meinen Eltern, denen diese Arbeit auch gewidmet ist. Sie haben durch ihre großzügige Unterstützung die Anfertigung dieser Arbeit ermöglicht.

Contents

List of Figures

List of Abbreviations

a.s.	almost surely
CC	Cross classification
cf.	compare
e.g.	for example
fig.	figure
i.e.	that is
i.i.d.	independent and identically distributed
MAE	Mean absolute error
MC	Multidimensional credibility-based algorithm
MSE	Mean squared error
TWSS	Total within group sum of squares
vs.	versus
w.l.o.g.	without loss of generality

Chapter 1

Introduction

1.1 Exploratory Focus

The crucial aspect of pricing insurance premiums is – compared to other products such as cars or computers – that the cost of the good protection/insurance is not known beforehand; the cost (namely the claim sizes) for the insurance company will not be known until some date in the future. Thus, insurers have to develop alternative approaches to determine adequate prices for their products. Accurate pricing is a critical issue for at least two reasons. First, the *overall* level of premiums has to guarantee safety and profit objectives: Too low premiums do not ensure the insurer's liability to cover the claims and thus guarantee its solvency or are not able to ensure the compliance of certain profit margins, whereas too high premiums banish customers. A liquid insurer is not only important to the insurer itself but also to the policyholders since they are given the guarantee to receive payments in case of a claim event. Second, it is particularly important to get the *relative* premium structure right. An important aspect within this context is **adverse selection**: If an insurer charges too little for high risk groups and too much for low risk groups, it will eventually lose low risk customers and gain high risk customers (such customer fluctuations are very likely since insurance markets have become highly competitive in the past decades, see [Völ08]). This obviously has a substantial influence on portfolio performance as the high risks cost too much and yield too low accruals of funds and the low risk groups are likely to terminate their policies. Increasing the overall level of premiums to raise profitability again even worsens the situation as it compounds the drawback of too low and too high premiums. This leads to a spiral of losing low risk customers and attracting high risk customers. Thus the relative levels of premiums have to be fixed correctly in order to be competitive, and this is the main aspect of statistical approaches to premium pricing.

All ideas being developed in the course of this investigation are valid for all branches of non-life insurance. Typically, we shall illustrate our arguments by examples from motor third party insurance since tariffs within this branch partition their portfolios to a – compared to other branches of insurance – greater extent and since we shall carry out an empirical analysis related to car insurance. So our first observation regarding adverse selection relates to (German) motor insurance as well, cf. [Sie71]: As is known, the Nazi regime intended to increase the level of motorization of the German population. Against this background, it enforced a uniform tariff[1] ("flat rate") for motor insurance.[2] As the civil road traffic almost vanished due to the outbreak of war in 1939, one cannot judge whether or not this uniform tariff would have coped with the situation before 1939. After the war, the existing uniform tariff had been adapted by insurance firms initially. However, it turned out quite early that one had to charge increased contributions. The problem of adverse selection described above was compounded by such an increasement. The differentation of rates in motor insurance has become inevitable.

Having noticed the necessity to install a risk-differentiating

[1] In fact, there was a very course differentation according to *type of vehicle*, *power* and *insured sum*

[2] Legal foundation: *Pflichtversicherungsgesetz*

tariff, we pose the question how this can be achieved. The key is to make use of the (strong) law of large numbers. In our context, we think of random variables $X_i \quad (i \in \mathbb{N})$ as **claim sizes**. Before we formulate the laws of large numbers, let us review two important concepts of convergence coming up in this context: *P-almost sure convergence* of the sequence $(X_n)_{n\in\mathbb{N}}$ of real random variables on a probability space $(\Omega, \mathcal{A}, P)$ towards a real random variable X on $(\Omega, \mathcal{A}, P)$ means

$$P\{\lim_{n\to\infty} X_n = X\} = 1$$

and *P-stochastic convergence* means

$$\lim_{n\to\infty} P\{|X_n - X| \geq \varepsilon\} = 0 \quad (\varepsilon > 0).$$

Definition 1.1.1 *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of integrable real random variables on a probability space $(\Omega, \mathcal{A}, P)$.*

(i) *$(X_n)_{n\in\mathbb{N}}$ is said to satisfy the* **weak law of large numbers** *if*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (X_i - E(X_i)) = 0$$

holds in the sense of P-stochastic convergence.

(ii) *$(X_n)_{n\in\mathbb{N}}$ is said to satisfy the* **strong law of large numbers** *if*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (X_i - E(X_i)) = 0 \qquad P - a.s.$$

Of course, the implication (ii) $\Longrightarrow$ (i) holds. There are various conditions which can be shown to be sufficient for the occurrence of a law of large numbers, cf. [Bau02]. The most prominent results are the theorems of Kolmogorov, the theorem of Khinchin and the theorem of Etemadi,[3] according to which the random variables $X_i \quad (i \in \mathbb{N})$ particularly are assumed to be **identically distributed**.

[3] We mention another sufficient condition in the so called *production law of the insurance technology* in Chapter 3.

Proposition 1.1.2 (Khinchin) *If the sequence $(X_n)_{n\in\mathbb{N}}$ of integrable and pairwise uncorrelated random variables with variances $V(X_n)$ obeys*

$$\lim_{n\to\infty} \frac{1}{n^2} \sum_{i=1}^{n} V(X_i) = 0,$$

it satisfies the weak law of large numbers.

Proof. See [Bau02]. □

We now state a very general condition.

Proposition 1.1.3 (Etemadi) *Each sequence $(X_n)_{n\in\mathbb{N}}$ of real, integrable, identically distributed and pairwise independent random variables obeys the strong law of large numbers.*

Proof. See [Ete81]. □

Actually, there are two pertinent theorems of Kolmogorov. We cite that one requiring identically distributed random variables and thus providing a link to cluster analysis; it follows immediately from the theorem of Etemadi.

Corollary 1.1.4 (Kolmogorov) *Each independent sequence $(X_n)_{n\in\mathbb{N}}$ of real i.i.d. random variables satisfies the strong law of large numbers.*

Aiming to make use of these statements, one has to lay the foundations for applying them. For instance, one of the (sufficient) conditions which is common to all stated theorems is to consider (a) infinitely many (b) identically distributed random variables (the law of large numbers is a statement on convergence, i.e. the behaviour of a finite number of risks is not decisive, instead, a sequence of infinitely many risks is required). **Cluster analysis** is an adequate aid to generate reasonably *large* families of risks which we assume to have the *same distribution.* By considering such – in some degree – large families ("clusters") of identically distributed risks we expect the strong law of large numbers to work. If we would like to make a statement on the expected claim size of a risk (this quantity serves as fundament of many premium calculation principles), the laws of large numbers justify to look at a (sufficiently large) group of risks having the same distribution. We hope to obtain such a group by a suitable clustering. So in other words, our considerations mean to raise the statistical basis for estimation purposes. We will continue discussing these ideas in conjunction with determining suitable collective sizes later on.

So far, we have been concerned with the *expected value* of the claims size distribution. We will see, however, that some premium principles such as the variance principle require to know *higher moments* of the claim size distribution as well. The so-called quantile principle even requires the actuary to know the *entire* claim size distribution. To deal with these issues, the **Glivenko-Cantelli theorem** (see [GS77]), an application of the strong law of large numbers, produces relief. According to the Glivenko-Cantelli theorem, the distribution function F (which is independent of n) of X_n can be determined approximately in the sense of P-a.s. uniform convergence by means of samples, i.e. by realizations of the sequence $(X_n)_{n\in\mathbb{N}}$ and the corresponding empirical distributions. Observe that the latter statement again requires a sequence $(X_n)_{n\in\mathbb{N}}$ of i.i.d. real random variables, so performing a cluster analysis is appropriate here, too.

Having recognized that the way to generate large homogeneous risk groups leads over the law of large numbers and hence via cluster analysis, we pose the question in which manner one ought to cluster in order to obtain preferable (or in some sense "optimal") classification results. This question constitutes the subject matter of our analysis. There are indefinitely many methods to design a classification scheme assigning homogeneous risks the same tariff class. We put forward arguments in favour of a particular approach to be developed in the course of this research. Our new proposal will be referred to as MC (***m**ultidimensional **c**redibility-based classifica-*