UNIVERSITÄT HAMBURG

Nearest-Neighbour Methods for Reserving with respect to Individual Losses

J.M. Dittmer

Preprint–No. 2006-05 August 2006

FACHBEREICH MATHEMATIK

SCHWERPUNKT MATHEMATISCHE STATISTIK
UND STOCHASTISCHE PROZESSE

BUNDESSTR. 55, D – 20146 HAMBURG

Nearest-Neighbour Methods for Reserving with respect to Individual Losses

Jens M. Dittmer[1] (Hamburg)

1 Introduction

Non-life insurance contracts often have to manage the problem that losses cannot be settled immediately but rather cause indemnity payments throughout the following years. The insurer has to estimate these amounts in order to allocate provisions and to calculate the business profit as well as the premiums appropiately.

The commonly used technique for this purpose is the Chain-Ladder method. It allows to predict the required total reserves if the portfolio is sufficiently homogeneous. This article, however, focuses on the situation with large losses dominating the aggregate payments. Here, it appears reasonable to treat these major claims separately.

In a closely related context, Mack (2002), p. 312ff., has proposed to apply certain nearest-neighbour methods: The future payments are forecasted by comparing the claims experience of the particular loss event under consideration with the corresponding history of those claims observed the years before.

At first, the most similar past loss (with respect to the hitherto provided compensation) is identified. The multiplicative or additive increment of the accumulated indemnities during the particular development period is then used to extrapolate the concerned claim accordingly.

Mack (2002) also suggests to improve this approach using several (i.e. $k \in \mathbb{N}$) similar losses of recent years to predict the future late claims. In the following section, this k-nearest-neighbour estimator is interpreted in the framework of a nonparametric regression model. By means of its asymptotic normality, confidence intervals for the expected future payments can be constructed for every single loss. Deriving confidence intervals for the aggregate loss of the portfolio remains an open problem, though.

In Section 3, an empirical comparison with the Chain-Ladder method shows that (for a large range of values of k) relatively precise predictions are obtainable.

Afterwards, the practical application of the k-nearest-neighbour estimator is discussed in Section 4.

The paper closes with a brief conclusion.

[1] E-Mail: dittmer@math.uni-hamburg.de

2 Asymptotic properties of the k-nearest-neighbour estimator

2.1 The model

For analytical purposes, the prognosis of future indemnities based on the claims experience of $p \in \mathbb{N}$ *development years* is interpreted as a nonparametric regression problem.
Concretely, let $X : (\Omega, \mathcal{A}, P) \longrightarrow (\mathbb{R}^p, \mathbb{B}^p)$ describe the payments of compensation within the first p years. Moreover, $Y : (\Omega, \mathcal{A}, P) \longrightarrow (\mathbb{R}, \mathbb{B})$ denotes the amount of a certain later period.
Usually, one is interested in predicting the losses accumulating either in a particular year (e.g. the following) or within the next $q \in \mathbb{N}$ years.[2]
Below, a model of the type

$$Y = m(X) + \sigma(X) \cdot \varepsilon \tag{1}$$

is presumed, where m and σ denote real-valued functions and ε a random variable (independent of X) with $E(\varepsilon) = 0$ and $\mathrm{Var}(\varepsilon) = 1$. This ensures

$$m(x) = E(Y|X = x)$$

and

$$\sigma^2(x) = \mathrm{Var}(Y|X = x)$$

for P^X-almost every $x \in \mathbb{R}^p$.
For the sake of predicting the expected future payments due to a damage characterised by a (fixed) claims experience $x \in \mathbb{R}^p$, i.e. $m(x)$, the already observed losses are represented by $n \in \mathbb{N}$ independent random variables

$$(X_1, Y_1), \dots, (X_n, Y_n), \quad \text{with} \quad (X, Y), (X_i, Y_i) \text{ iid.}$$

Furthermore, let $||\cdot||$ be an arbitrary norm on $\mathbb{R}^p$. For a sequence $\{k_n\}_{n \in \mathbb{N}}$ of natural numbers (with $k_n \leq n$), the k_n-nearest-neighbour estimator (for $m(x)$) is then defined as

$$m_n(x) := \frac{\sum_{i=1}^n K(\frac{X_i - x}{R_n}) Y_i}{\sum_{i=1}^n K(\frac{X_i - x}{R_n})}. \tag{2}$$

Here, R_n denotes the distance between x and the k_n-th nearest neighbour among the X_i, $i \in \{1, \dots, n\}$,

$$R_n := \inf\{t \geq 0 \mid \sum_{j=1}^n 1_{\{||X_j - x|| \leq t\}} \geq k_n\}.$$

[2] In principle, the proposed estimator in (2) may analogously be defined for a multivariate Y (in order to forecast the amounts of the following years in a single step). Such an approach would go beyond the scope of this paper, though.

$K : \mathbb{R}^n \longrightarrow [0, \infty)$ is a kernel function with $K(u) > 0$ for $u \in B_1(0)$ and $K(u) = 0$ else.[3]

Assuming m as a smooth function, $m(x)$ is estimated as weighted average of the Y_i. The weights $K(\frac{X_i - x}{R_n})$ are positive for the $k_n - 1$ most similar losses (to x). Thus, the choice of k_n determines the number of loss events that contribute to the estimation. This way, k_n directly serves as a smoothing parameter.

The selection of $||\cdot||$ defines the measure of the similarity of the claims experiences X_i and x and therefore determines *which* past losses are taken into account.

Furthermore, the kernel K is often chosen such that those observations are the further downweighted the more X_i deviates from x. For example, K may be defined via

$$K(u) := \kappa\big((1 - ||u||^2) + \delta\big)1_{B_1(0)}(u), \quad u \in \mathbb{R}^p, \tag{3}$$

with $\delta \geq 0$ and $\kappa > 0$ fixed that way to ensure $\int K(u)du = 1$. To fulfill assumption (4) of Theorem 2.1, $\delta > 0$ is required. With $\delta = 0$, (3) would describe the well-known *Epanechnikov* kernel.

The estimator $m_n(x)$ is a weighted average of such values Y_i for which the distance $||X_i - x||$ falls below the bandwidth R_n. Estimating $m(x)$ assuming m to be locally constant leads to a systematic error since $m(X_i)$ actually deviates from $m(x)$. As R_n increases with k_n, this bias becomes more important when the number of included neighbours rises. Additionally, for a major part of the losses, $\frac{||X_i - x||}{R_n}$ tends to zero choosing k_n too large. This would undermine the idea of weighting the Y_i according to the similarity between X_i and x.

In contrast, for very small values of k_n, the variance of the estimation error turns out to be very large. Thus, we have to deal with a typical bias-variance tradeoff concerning the choice of k_n.

In the context of nonparametric statistics, the discussed kind of k_n-nearest-neighbour estimator has been introduced by Collomb (1979) and has been analysed by Mack (1981) and Liero (1987), among others. The theory is partially based on results in the related field of kernel density estimation, see Loftsgaarden/Quesenberry (1965), Mack/Rosenblatt (1979) and Mack (1980).

[3] In this paper,

$$B_r(w) := \{u \in \mathbb{R}^p \mid ||u - w|| < r\}$$

denotes the open ball in $\mathbb{R}^p$ with radius r centered at w.

2.2 Asymptotic normality of the k_n-nearest-neighbour estimator

In the following, an expression for the approximative normal distribution of $m_n(x)$ is stated - allowing to construct confidence intervals for $m(x)$. As before, $x \in \mathbb{R}^p$ is fixed.

2.1 Theorem. *Let X be absolutely continuous with Lebesgue density f and $f(x) > 0$. The functions f and m are assumed bounded and (in a neighbourhood of x) twice continuously differentiable. Let*

$$l(u) := m^2(u) + \sigma^2(u) = E(Y^2|X = u), \quad u \in \mathbb{R}^p,$$

be continuous at x and bounded. Assume that

$$E(|Y|^3|X = u) \leq M$$

for some $M < \infty$ and P^X-almost every $u \in \mathbb{R}^p$. Let $\{k_n\}_{n\in\mathbb{N}}$ satisfy

$$k_n \leq n \quad \forall\, n \in \mathbb{N}, \quad \lim_{n\to\infty} k_n = +\infty, \quad \lim_{n\to\infty} \frac{k_n}{n} = 0 \quad \textit{and}$$

$$\eta := \lim_{n\to\infty} k_n \cdot n^{-4/(4+p)} < \infty.$$

Finally, assume

$$\exists\, N_1,\, N_2 > 0 : \quad N_1 \leq K(u) \leq N_2 \quad \forall\, u \in B_1(0), \tag{4}$$

$$K(u) = 0 \quad \forall\, u \notin B_1(0), \tag{5}$$

$$\int K(u)du = 1 \qquad \textit{and} \tag{6}$$

$$\int K(u)u_\alpha du = 0 \quad \forall\, \alpha \in \{1, \ldots, p\}. \tag{7}$$

Then,

$$\sqrt{k_n}\big(m_n(x) - m(x)\big) \longrightarrow \mathcal{N}\Big(B,\, c \cdot \sigma^2(x)\int K^2(v)dv\Big), \tag{8}$$

in distribution as $n \to \infty$, with

$$c := \lambda_p\big(B_1(0)\big)$$

(Lebesgue-measure of the unit ball in $\mathbb{R}^p$ with regard to $||\cdot||$) and

$$B := \eta^{1/2+2/p}\tilde{B} := \eta^{1/2+2/p}\frac{Q(mf)(x) - m(x)Q(f)(x)}{2f(x)(cf(x))^{2/p}}.$$

Here, for a function $\psi : \mathbb{R}^p \longrightarrow \mathbb{R}$, twice continuously differentiable at x, $Q(\psi)(x)$ is defined as

$$Q(\psi)(x) := \sum_{\alpha=1}^{p}\sum_{\beta=1}^{p}\int v_\alpha v_\beta \frac{\partial^2\psi(x)}{\partial\alpha\partial\beta}K(v)dv.$$

The proof is subdivided into two parts. At first, $m_n(x)$ and the underlying random variables are analysed conditioned under R_n. The convergence

$$\sqrt{k_n}\Big(m_n(x) - E\big(m_n(x)|R_n\big)\Big) \overset{n\to\infty}{\longrightarrow} \mathcal{N}\Big(0,\ c\cdot\sigma^2(x)\int K^2(v)dv\Big) \tag{9}$$

is established using the central limit theorem of Berry-Esséen. The basic procedure corresponds to the methods applied in Mack (1981). Examining the referenced proof more precisely, however, one finds that certain additional assumptions[4] are required in order to derive the desired results.
Afterwards, it suffices to show

$$\sqrt{k_n}\Big(E\big(m_n(x)|R_n\big) - m(x)\Big) \overset{n\to\infty}{\longrightarrow} B \quad \text{in probability.} \tag{10}$$

For further details see Dittmer (2005), p. 33ff.

Approximative confidence intervals can hence be derived in the usual way. For any $\alpha \in (0,1)$, Theorem 2.1 yields

$$P\Big\{m(x) \in \Big[m_n(x) - \frac{B}{\sqrt{k_n}} - \tau_n(x),\ m_n(x) - \frac{B}{\sqrt{k_n}} + \tau_n(x)\Big]\Big\} \overset{n\to\infty}{\longrightarrow} 1-\alpha, \tag{11}$$

with

$$\tau_n(x) := z_{1-\alpha/2}\sqrt{\frac{c}{k_n}\sigma^2(x)\int K^2(v)dv} \tag{12}$$

and $z_{1-\alpha/2}$ denoting the $(1-\alpha/2)$-quantile of the standard normal distribution. Since B and $\sigma^2(x)$ are usually unknown, they have to be estimated from the data. Replacing both of them (in (11) and (12)) by consistent estimators $\hat{B}$ and $\hat{\sigma}^2(x)$, asymptotic confidence intervals with level $(1-\alpha)$ are obtained again.

In general, the performance of the normal approximation is better for more symmetric distributions of ε featuring less heavy tails.

Under weak additional assumptions, $n^{4/(4+p)}E\big(m_n(x) - m(x)\big)^2$ converges in $(0,\infty)$ and is asymptotically minimized if we choose

$$k_n = \lfloor \eta^* \cdot n^{4/(4+p)} \rfloor \tag{13}$$

for

$$\eta^* := \Big(\frac{p}{4}\cdot\frac{c\sigma^2(x)\int K^2(v)dv}{\tilde{B}^2}\Big)^{p/(p+4)}.$$

[4] In particular, Mack (1981) aims at proving (9) without presuming $N_1 > 0$ in (4). However, his proof is flawed and it has turned out that it cannot be corrected without assuming that K is bounded away from zero on $B_1(0)$.

2.3 Prediction accuracy of the k_n-nearest-neighbour estimator

Aiming at assessing the forecast of the future values of Y, the influence of ε (in (1)) has to be kept in mind. For this purpose, we also consider *prediction intervals.* They cover a proportion of *future observations* of at least $1 - \alpha$.

Under the assumptions stated above, the prediction error

$$m_n(x) - Y = \big(m_n(x) - m(x)\big) - \sigma(x) \cdot \varepsilon$$

is the sum of the independent random variables $m_n(x) - m(x)$ and $-\sigma(x) \cdot \varepsilon$. Hence, prediction intervals can be determined by calculating the corresponding convolution. They are always wider than confidence intervals. Moreover, we obtain

$$\operatorname{Var}\big(m_n(x) - Y\big) = \operatorname{Var}\big(m_n(x) - m(x)\big) + \sigma^2(x). \tag{14}$$

The asymptotic limit distribution of $m_n(x) - m(x)$ is given by Theorem 2.1, while the behaviour of ε has to be examined analysing the standardised residuals

$$\frac{Y - m_n(x)}{\sqrt{\hat{\sigma}^2(x)}}.$$

For a non-normally distributed ε, the distribution of $m_n(x) - Y$ cannot be calculated explicitly which complicates the computation of the prediction intervals.

Under weak additional assumptions,

$$\frac{1}{k_n} c\sigma^2(x) \int K^2(v)dv$$

can be considered as an approximation to $\operatorname{Var}\big(m_n(x) - m(x)\big)$. Then, by (14), we have

$$\sqrt{\operatorname{Var}\big(m_n(x) - Y\big)} \approx \sigma(x)\sqrt{1 + \frac{c}{k_n}\int K^2(v)dv}.$$

With K from (3) and the Euclidean norm $||\cdot||$ we have $c\int K^2(v)dv \leq 2$ for any $p \in \mathbb{N}$ and $\delta > 0$. Thus, for larger numbers of neighbours the influence of $m_n(x) - m(x)$ and k_n on the prediction error decreases rapidly. For very small values of k_n (as in the extreme example $k_n = 1$) the estimation error becomes important, though.

3 Accuracy of methods for reserving for individual losses

Considering third party liability insurance data as an example, the performances of the aforementioned methods are compared empirically. We consider claims experiences of an established german insurance company. The data is provided by courtesy of *AON Rück* (in a perturbed form).

Concretely, 1346 losses above a certain threshold are available that incurred between 1973 and 2004. The number of loss events tends to increase over the years - from 10 (before 1979) up to 96 (in 1996). Typically, damages result in annual indemnities with the largest amounts during the first three development years.[5] For few losses, the compensation does not commence in the year of reporting the damage. Then we define the period of the initial payment as accident year and thus as first development year.

Since information about loss increments of former years is used to predict future late claims, trend effects caused by inflation would falsify the results. Therefore, the payments are converted into values of 2004.[6] The effect that some amounts of the earlier years may have been fallen below the threshold mentioned above, partly explains the increasing number of loss events taken into account.

Figure 1 displays the structure of the average annual loss increments for the particular development years. In addition to mean and standard deviation, the estimated linear time trend is illustrated. The latter is the slope coefficient of a linear least squares regression of payments (within the fixed development year) versus time. It is negative for most of the early development years (2 to 10) and behaves contrarily afterwards. Apparently, structural changes took place. The settlement of losses seems to shift to the later development years. As these effects are neglected below, it does not surprise that all of the methods under consideration slightly overestimate the indemnities of the first development years.

This section primarily deals with the comparison of the discussed k-nearest-neighbour estimator with several of the other techniques proposed by Mack (2002) and a version of the Chain-Ladder method (applied to individual losses). For this purpose, an adequate backtesting procedure is used.

Concretely, we consider the prediction of compensation payments of the following year on the basis of the claims development hitherto observed. For the sake of simplicity, the involved calendar years are renamed as $1, \ldots, I$ (i.e. $I = 32$). For every year $p \in \{2, \ldots, I-1\}$ and for every individual loss that has not been closed by then, the future indemnities are predicted using the information so far available. Afterwards, the

[5] In case of refunding excessive compensation, the indemnification in later development years can be negative on occasion.

[6] For this purpose, the consumer price indices for Germany, published by the *Federal Statistical Office (www.destatis.de)*, have been applied.

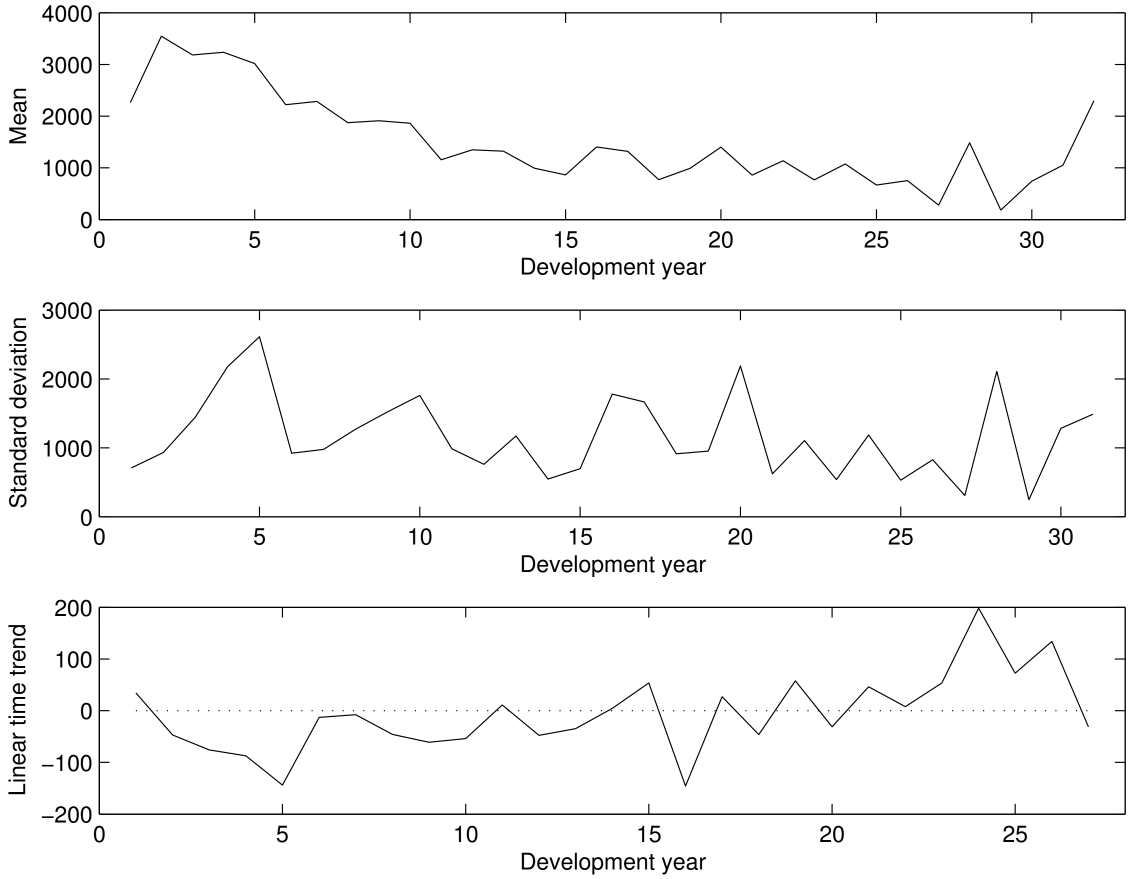

Figure 1: Averaged annual loss increments of the distinct development years: Progression of mean, standard deviation and estimated linear time trend

prognosis is compared with the actually payed amount.

In practice, predicting the accumulated payments from the presence up to a specified final year L is similarly relevant. The corresponding estimators can obviously be defined analogously to the case treated in this section. However, the application of a backtesting procedure is more problematic. On the one hand, claims experiences can only be used as historical data if the L-th development year has already been observed. On the other hand, only such predictions can be included for which the actual payments are available.

Therefore, only those losses can be used for assessment that incurred in the calendar years L to $I-L+1$. Hence, for large values of L, the number of involved data decreases rapidly. An examination for $L = 5$ and $L = 10$ (in Dittmer (2005), p. 25ff.) yields qualitatively similar results as in Figure 2, though.

3.1 Specification of the regarded reserving methods

In the following, a brief survey of the reserving techniques discussed in this paper is presented.

For the purpose of predicting future late claims with respect to individual losses, a natural modification of the Chain-Ladder method provides a possible solution:
Let S_{ik} denote the aggregated indemnities of the accident year i and development year k (i.e. $S_{ik} := \sum_{m=1}^{n_{ik}} X_{ikm}$) and C_{ik} the correspondent accumulated claims

$$C_{ik} := \sum_{j=1}^{k} S_{ij}.$$

Carrying out the original Chain-Ladder estimation, for every development year an incremental factor is determined as

$$\hat{f}_k := \frac{\sum_{i=1}^{I-k} C_{i,k+1}}{\sum_{i=1}^{I-k} C_{ik}}, \quad k \in \{1, \ldots, I-1\}. \tag{15}$$

Future cumulated claims C_{ik} (with $k \in \{1, \ldots, I\}$ and $i + k > I + 1$) are then forecasted by means of

$$\hat{C}_{ik} := C_{i,I+1-i} \cdot \hat{f}_{I+1-i} \cdot \ldots \cdot \hat{f}_{k-1}.$$

Considering individual losses, it suggests itself to calculate the factors $\hat{f}_k$ just as in (15) but to apply these to the separate data.

Following a proposal of Mack (2002), p. 312ff., one may predict late claims belonging to a certain loss by determining the most similar claims experience among the historical data. The observed (additive or multiplicative) increment of the indemnities during the particular development year is used to carry forward the considered loss. In this regard and concerning the distance between multivariate losses (stating the degree of similarity) there exist certain choices.
As in Section 2, let $x \in \mathbb{R}^p$ (fixed) and X_i, $i \in \{1, \ldots, n\}$, describe the claims experiences of the considered and the various historical damages, respectively. More precisely, denote the accordant components as $x^{(k)}$ and $X_i^{(k)}$, $k \in \{1, \ldots, p\}$, representing the *accumulated* claims up to development year k. The associated indemnifications of the $(p+1)$-th period are described by Y (to be estimated) and Y_i, $i \in \{1, \ldots, n\}$ (observed).
As an alternative, $x^{(k)}$ and $X_i^{(k)}$ could characterise the actual (non-accumulated) indemnity payments within the particular year. Such a proceeding, though, leads to less accurate forecasts and is not taken into consideration any further.

In order to measure the similarity between different losses (considering the distance $||X_i - x||$), an adequate norm $||\cdot||$ on $\mathbb{R}^p$ has to be specified. Two natural approaches are used by either defining $||\cdot||$ as the Euclidean norm,

$$||u|| := \sqrt{\sum_{i=1}^{p} u_i^2}, \quad u \in \mathbb{R}^p, \tag{16}$$

or as the absolute value of the p-th component,

$$||u|| := |u_p|, \quad u \in \mathbb{R}^p. \tag{17}$$

Utilising the Euclidean distance, the accumulated payments of the entire claims experience are compared after every development year. In the one-dimensional case (17), however, only the single difference of the indemnities accumulated up to the present year is considered. That claims experience (among the X_i, $i \in \{1, \dots, n\}$) with minimal distance to x is denoted as $\tilde{X}$ whereas $\tilde{Y}$ characterises the correspondent amount of the $(p+1)$-th period. The value of $\tilde{Y}$ is used to extrapolate x by means of an *additive* or *multiplicative* continuation in order to forecast the future indemnification Y,

$$\hat{Y}_{(\text{add})} := \tilde{Y} \tag{18}$$

and

$$\hat{Y}_{(\text{mult})} := x^{(p)} \cdot \frac{\tilde{Y}}{\tilde{X}^{(p)}}, \tag{19}$$

respectively. The terms *additive* and *multiplicative* are used due to the relationships (with $\tilde{X}^{(p+1)} := \tilde{X}^{(p)} + \tilde{Y}$ denoting the accumulated loss after $p+1$ periods)

$$x^{(p)} + \hat{Y}_{(\text{add})} = x^{(p)} + (\tilde{X}^{(p+1)} - \tilde{X}^{(p)})$$

and

$$x^{(p)} + \hat{Y}_{(\text{mult})} = x^{(p)} \cdot \frac{\tilde{X}^{(p+1)}}{\tilde{X}^{(p)}}.$$

As a different approach, the k-nearest-neighbour estimator presented in (2) is interpreted as a forecast of the future indemnities,

$$\hat{Y}_{\text{knn}} := m_n(x).^{7} \tag{20}$$

For this purpose, the kernel function K and the norm $||\cdot||$ have to be specified. In the following, let K be defined by (3), with $\delta = 0.05$. Furthermore, we define $||\cdot||$ as the Euclidean norm.

[7] Note that $m_n(x) = \hat{Y}_{(\text{add})}$ for $k_n = 2$ (only one included neighbour).

More generally, we could determine the underlying norm as $||u|| := \sqrt{\sum_{i=1}^{p} a_i u_i^2}$, $u \in \mathbb{R}^p$, for a fixed $a \in [0,\infty)^p$. Thus, the influence of the different development years can be controlled by weighting the correspondent components of the claims experience. Since the consideration is based on accumulated payments, the amounts of later development years also include information about preceding periods and may reasonably be incorporated to a higher degree. In Dittmer (2005), p. 25ff., the special case $a_i := 2^{i-1}$, $i \in \{1,\ldots,p\}$, is treated as well. The performance of this method, however, is only slightly better than in the Euclidean case ($a = (1,\ldots,1)^T$). Hence, the choice of $||\cdot||$ supposably does not affect the outcome crucially.

The considered techniques (including the Chain-Ladder method) can be improved by only including claims experiences of the latest ten years - partially eliminating the trend effects stated above. Compared to each other, the performances of the methods remain qualitatively unchanged.

3.2 Assessment criteria for reserving methods

Let X_{ikm} denote the m-th among the n_{ik} losses belonging to the i-th accident year and to the k-th development year. The correspondent backtesting estimators (assessing these payments on the basis of the hitherto observed data) are described by $\hat{X}_{ikm}$.
Depending on the prevailing objective, different performance measures are useful to compare the methods presented in 3.1.

The sum of squared residuals over the separate indemnities of every accident and development year in the backtesting period,

$$SSR_{\text{ind}} := \sum_{i=2}^{I-1} \sum_{k=2}^{I-i+1} \sum_{m=1}^{n_{ik}} (\hat{X}_{ikm} - X_{ikm})^2, \tag{21}$$

indicates the precision of the prognosis with respect to individual losses.

One's main interest may lie, however, in a satisfactory prediction of the total losses accumulated over the payments of fixed accident and development years (allowing effects of balancing out between the n_{ik} payments). In this case,

$$SSR_{\text{ann}} := \sum_{i=2}^{I-1} \sum_{k=2}^{I-i+1} \Big(\sum_{m=1}^{n_{ik}} \hat{X}_{ikm} - \sum_{m=1}^{n_{ik}} X_{ikm}\Big)^2 \tag{22}$$

should preferably be taken into account.

In addition, calculating certain quantiles of the empirical distribution of the absolute residuals $|\hat{X}_{ikm} - X_{ikm}|$, we receive supplementary information about the adequacy of the estimation.

3.3 Comparison of alternative reserving methods

Table 1 summarises the comparison of the estimators $\hat{Y}_{(\text{add})}$ and $\hat{Y}_{(\text{mult})}$ with the Chain-Ladder method. Apparently, the latter is superior to all of the considered alternatives with respect to the precision of forecast - regarding both assessment criteria.

Method	Distance	Continuation	SSR_{ind}	SSR_{ann}
Nearest-neighbour	(16)	Additive	$4.849{\cdot}10^{11}$	$5.476{\cdot}10^{11}$
Nearest-neighbour	(16)	Multiplicative	$5.569{\cdot}10^{11}$	$6.443{\cdot}10^{11}$
Nearest-neighbour	(17)	Additive	$5.088{\cdot}10^{11}$	$5.407{\cdot}10^{11}$
Nearest-neighbour	(17)	Multiplicative	$5.187{\cdot}10^{11}$	$5.606{\cdot}10^{11}$
Chain-Ladder			$3.090{\cdot}10^{11}$	$4.063{\cdot}10^{11}$

Table 1: Prediction accuracy of the Nearest-neighbour estimators $\hat{Y}_{(\text{add})}$ and $\hat{Y}_{(\text{mult})}$ and the Chain-Ladder method

Multiplicative variants appear to deliver inadequate predictions in particular. This is attributed to the fact that some losses feature only minor compensation during the first development years but substantial amounts later on. In this situation, multiplicative increments of the accumulated claims can turn out to be very large. Additive methods, in contrast, are less affected by this problem.

For the same reason, another attempt fails that intends to standardise the claims experiences x and X_i by their respective means (in case of $p \geq 2$ observed development years) in order to eliminate multiplicative differences in level,

$$\frac{x}{\overline{x}} \quad \text{and} \quad \frac{X_i}{\overline{X_i}}, \quad \text{with} \quad \overline{x} := \frac{1}{p}\sum_{k=1}^{p} x^{(k)} \quad \text{and} \quad \overline{X_i} := \frac{1}{p}\sum_{k=1}^{p} X_i^{(k)}.$$

Evidently, additive approaches are unemployable in this context.

The k-nearest-neighbour estimators perform substantially better in this situation. As it can be seen from Figure 2, for a wide range of values of k, Y_{knn} yields more precise predictions than the Chain-Ladder method (with respect to SSR_{ind}). Even the prognosis of the *aggregate* loss within fixed accident and development years (the original intention of the Chain-Ladder method - assessed by SSR_{ann}) is more accurate applying the k-nearest-neighbour technique. While SSR_{ind} and SSR_{ann} decrease rapidly with k_n for small numbers of neighbours (e.g. $SSR_{\text{ind}}(k_n = 5) \approx 0.67 \cdot SSR_{\text{ind}}(k_n = 1)$), both criteria are slowly falling until approximately $k_n = 75$ and slightly increasing afterwards.[8]

In particular, the amounts of the first ten development years are predicted comparatively well. The Chain-Ladder method, though, performs marginally better in forecasting the payments of the following years where the number of underlying data decreases.[9]

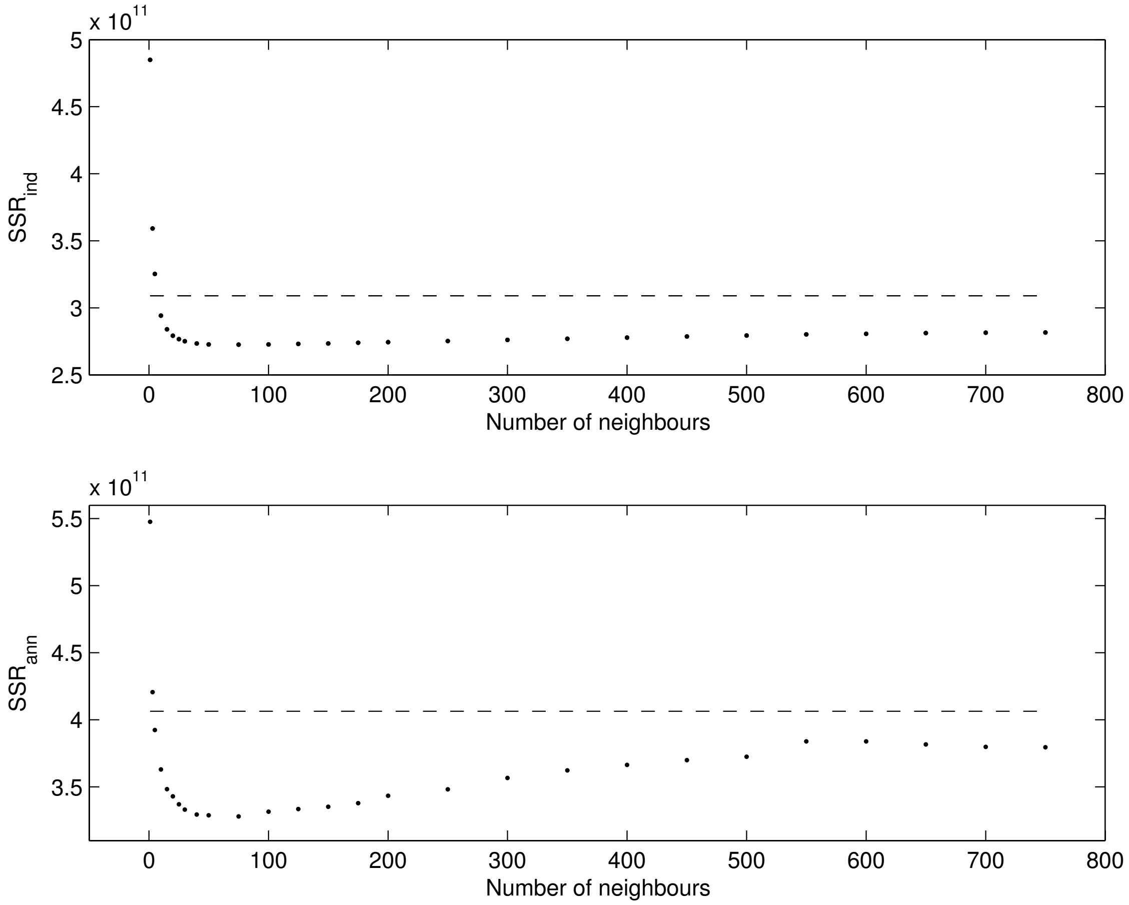

Figure 2: Performance measures for the k-nearest-neighbour estimator $\hat{Y}_{\text{knn}}$ in comparison with the Chain-Ladder method (dashed)

[8] If k_n exceeds n, we set $k_n = n$. Therefore, the illustrated sums of squared residuals in Figure 2 may undervalue the effect of an increasing number of neighbours. A further backtesting analysis only including those forecasts where at least 200 preceding observations have been available does not show any major qualitative changes, though.

[9] Thus, a correspondent combination of these procedures can be reasonable in practice.

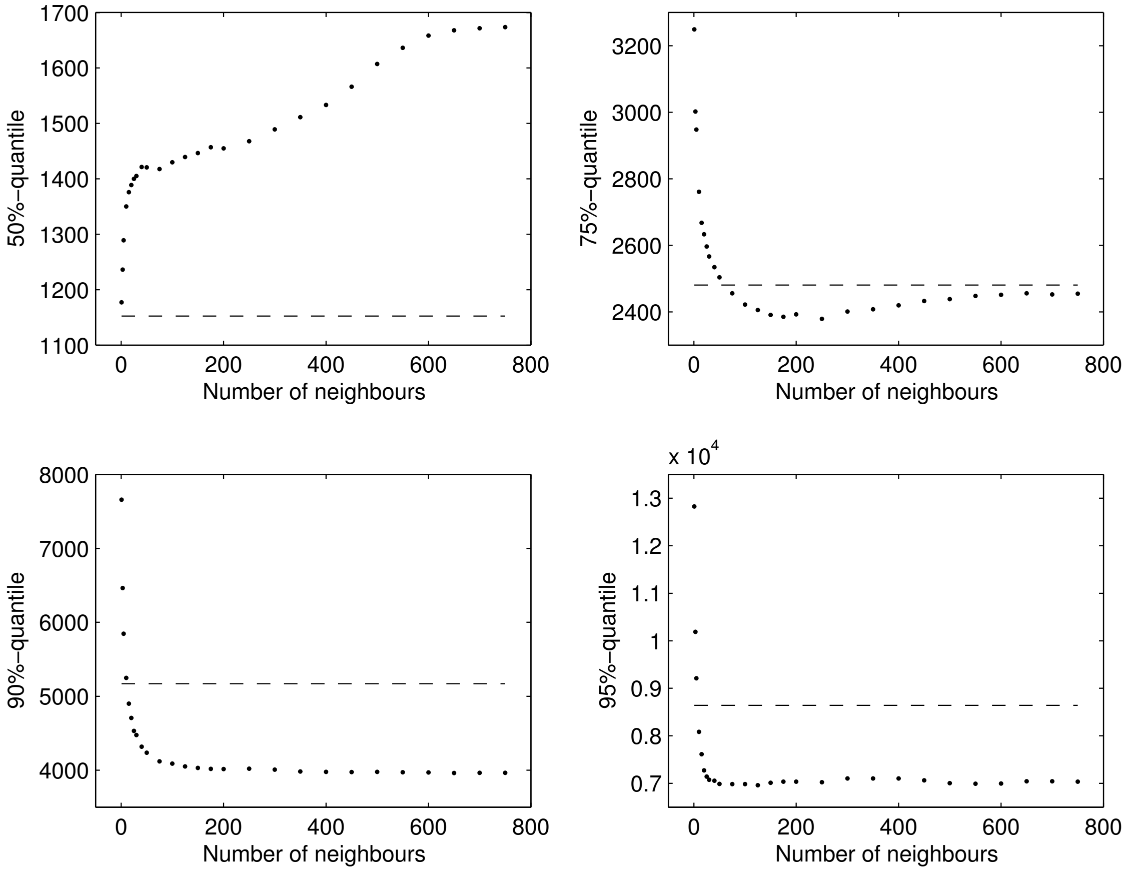

Figure 3: Quantiles of absolute residuals of the k-nearest-neighbour estimator $\hat{Y}_{\mathsf{knn}}$ in comparison with the Chain-Ladder method (dashed)

Figure 3 displays the quantiles of the absolute residuals $|\hat{X}_{ikm} - X_{ikm}|$. A $(1-\alpha)$-quantile is the smallest value that is exceeded by less than $\alpha \cdot 100\%$ of the data. The lower 50%-quantile shows that the Chain-Ladder method forecasts a large number of future indemnities slightly more precisely than the k-nearest-neighbour estimator. However, the comparison of the 75%-, 90%- and 95%-quantiles shows that it generates substantial discrepancies between predicted and observed payments more often, in return. A further inspection reveals that larger absolute residuals tend to emerge together with major losses.

Thus, confirming the initial conjecture, the Chain-Ladder method should preferably be applied to homogeneous portfolios while the k-nearest-neighbour technique is advantageous in case of heterogeneity.

In order to gain a deeper insight into the k-nearest-neighbour procedure, the individual predictions are now analysed in more detail.

For example, we consider an insurance company that is interested in predicting the indemnification payments of the fifth development year, based on the observed losses of the four preceding periods. To ensure that the forecasts can be compared with the actually payed amounts and to include a satisfactory number of data, we consider the losses that incurred in 1998. Thus, the estimation is assumed to be carried out at the end of 2001 (with the corresponding level of information).

With respect to an application of the results derived in 2.2 (for X possessing a Lebesgue-density), only those losses are taken into account which feature loss increments distinct from zero in every year of the *past* claims experience.

Using the normal approximation (8) to the estimation error

$$m_n(x) - m(x) = \hat{Y}_{\text{knn}} - m(x),$$

confidence intervals for $m(x)$ can be specified according to (11).

Applying another backtesting procedure on the considered data and comparing the sums of squared prediction errors (analogously to (21) for fixed i and k), we find that $k_n = 130$ (for here $n = 536$) would be the optimal choice.[10] However, including less neighbours, the prediction accuracy is only scarcely worse. Choosing k_n clearly lower than according to (13), η (and therefore B) becomes zero or at least negligibly small. Thus, we presume that B is neglectable for $k_n = 30$.

Now, an appropriate consistent estimator of $\sigma^2(x)$ has to be selected. Maintaining the k-nearest-neighbour approach, it suggests itself to define

$$\sigma_n^2(x) := |l_n(x) - m_n^2(x)|$$

with

$$l_n(x) := \frac{\sum_{i=1}^n K\left(\frac{X_i - x}{R_n}\right) Y_i^2}{\sum_{i=1}^n K\left(\frac{X_i - x}{R_n}\right)}.$$

Let $u \mapsto E(Y^4|X = u)$ be continuous at x and bounded. Then, under the assumptions of Theorem 2.1, $\sigma_n^2(x)$ is consistent for $\sigma^2(x)$.

A comparison with the prediction error (Figure 4) reveals that only less than half of the (ex post) observed payments in fact belong to the respective 0.95-confidence intervals (displayed by the dashed line). This is not surprising, bearing in mind the argumentation in Subsection 2.3.

In the present situation prediction intervals cannot be derived analytically since ε is not normally distributed. In order to gain a rough impression of their widths, prediction intervals (represented by the solid line in Figure 4) are calculated *as if* ε was Gaussian, nevertheless.

A possible explanation for the apparent asymmetry of the prediction error is a skewness (to the right) of ε.

[10] Note that such a global selection of k_n is made by purpose of simplicity. In fact, the optimal k_n is given by (13) and depends on $x \in \mathbb{R}^p$.

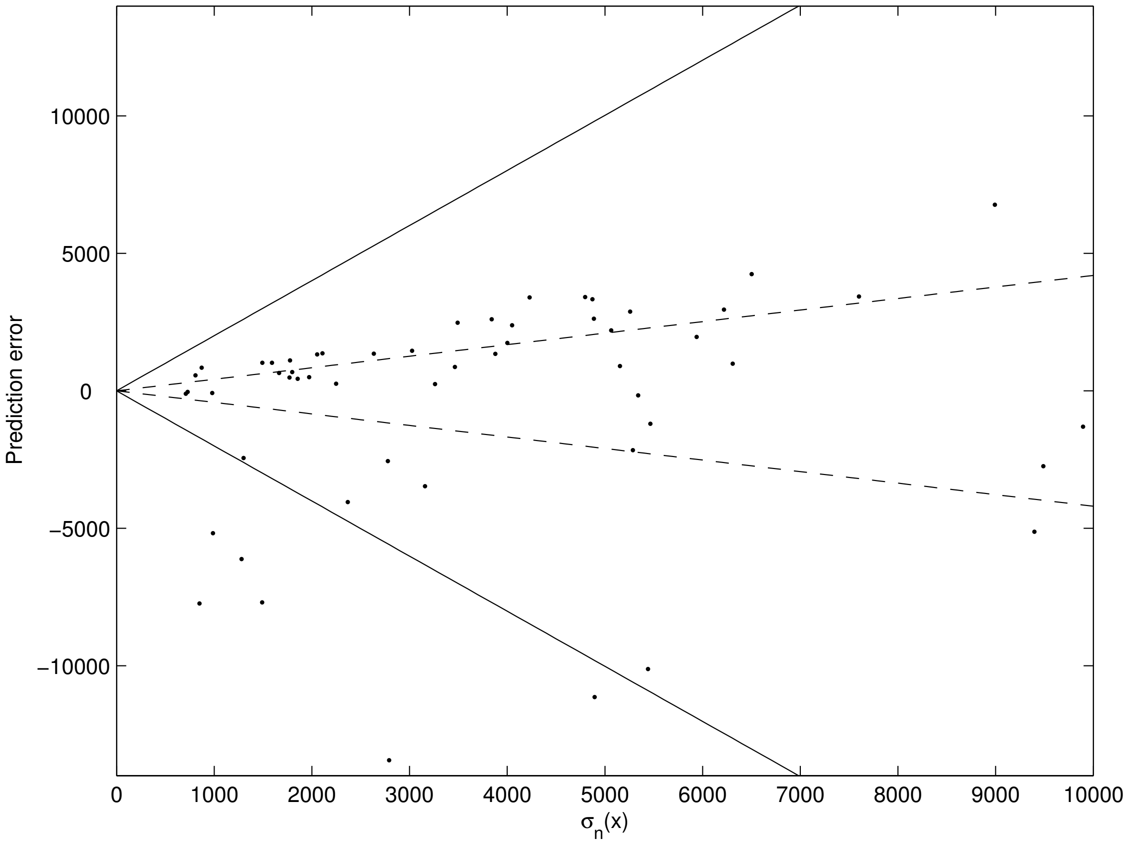

Figure 4: Prediction error in comparison to the estimated value of $\sigma(x)$; dashed: confidence interval, solid: prediction interval.

4 Application of the k-nearest-neighbour estimator

Usually, insurance companies are interested in estimating the expected future payments within every single development year. For this purpose, $\hat{Y}_{\text{knn}}$ (defined in (20)) can be used to predict these amounts in repeated (i.e. $L \in \mathbb{N}$) steps. Confidence intervals for these annual loss increments can be constructed according to (11).

Additionally, we can forecast the sum of the indemnities during these L years in one single step. This also allows us to derive confidence intervals for the expected payments accumulating in this period.

The empirical analysis in Section 3 has shown that the aggregate loss of the portfolio is predicted comparatively well. However, the distribution of the corresponding prediction error cannot be specified easily.

In any case the number of neighbours and the kernel function have to be determined.
The k_n-nearest-neighbour procedure seems to deliver relatively precise predictions for a large range of values of k_n. In the aforementioned example the best results have been found choosing k_n lower than 10-20 percent of the number of data but not smaller than ten neighbours. Such rules of thumb may have to be revised after applying the method to different portfolios, however.
It is more reliable to analyse the data carrying out a backtesting procedure (as described in Section 3) for varying numbers of neighbours. If comparatively few loss experiences are available, a cross-validation approach typically used for kernel estimation with fixed bandwidth can be suitable. Here, one loss at a time is left out of the portfolio and then estimated on the basis of the remaining observations. A comparison of the accumulated squared residuals for different values of k_n provides an indication about a reasonable choice of the number of included neighbours.

The kernel function K should be designed such that

$$K(u) > K(v) > 0 \quad \text{for} \quad ||u|| < ||v|| \leq 1.$$

This way, observations contribute the more to the forecast the more similar they have been to a considered loss experience (up to the present). A popular choice is the (modified) Epanechnikov kernel defined in (3) with $||\cdot||$ as the Euclidean norm.

5 Conclusion

In this article, the k-nearest-neighbour procedure is discussed as a method to predict future late claims in heterogeneous portfolios. A normal approximation for the estimator of the expected payments is provided and an empirical advantage over the Chain-Ladder method has been observed. In connection with this paper several interesting aspects arise.

The proof of the asymptotic normality of $m_n(x)$ is based on the condition that X possesses a Lebesgue density. Claims experiences, however, may comprise periods without incurred payments (although the loss has not yet been concluded). Therefore, it can be desirable to generalise Theorem 2.1 allowing X to have probability mass in zero for the particular development years.
Since the asymptotic limit distribution of the estimation error $m_n(x) - m(x)$ does not depend on f, a corresponding statement should be valid under adequate additional assumptions.

In Section 3 it has been shown that the k-nearest-neighbour method provides precise predictions for indemnities of the following period as well as for accumulated payments

up to a fixed development year. Carrying out the prognosis for every development year separately, one receives an estimation of the entire future loss process.
For this purpose one could alternatively consider an $\mathbb{R}^{I-p}$-valued random variable describing the particular amounts of the remaining periods $p+1$ to I. Completely analogously to the present case it can be forecasted by means of k similar claims experiences. The analysis of the prediction accuracy cannot be transferred directly, though.

Acknowledgements

The author is grateful to Professor Dr. Holger Drees for supervising the diploma thesis on which this paper is based. He also thanks AON Rück (particularly Irina Kaiser) for providing the data.

References

[1] Collomb, G. (1979) Estimation de la régression par la méthode des k points les plus proches avec noyau: Quelques propiétés de convergence ponctuelle. *Lecture Notes in Mathematics* 821, 159-175.

[2] Dittmer, J. M. (2005) *Nächste-Nachbarn-Verfahren zur Reservierung für Einzelschäden.* Diploma thesis, Universität Hamburg.

[3] Liero, H. (1987) On the asymptotic behaviour of a k_n-nearest neighbour estimate of the regression function. *Seminarbericht, Sektion Mathematik der Humboldt-Universität Berlin* 89, 184-196.

[4] Loftsgaarden, D. O. and Quesenberry, C. P. (1965) A nonparametric estimate of a multivariate density function. *The Annals of Mathematical Statistics* 36, 1049-1051.

[5] Mack, T. (2002) *Schadenversicherungsmathematik.* 2. Auflage; Verlag Versicherungswirtschaft, Karlsruhe.

[6] Mack, Y.-P. (1978) *k-nearest-neighbor estimation.* Dissertation, University of California, San Diego.

[7] Mack, Y.-P. (1980) Asymptotic normality of multivariate k-nn density estimates. *Sankhyā* 42, Series A, 53-63.

[8] Mack, Y.-P. (1981) Local properties of k-nn regression estimates. *SIAM Journal on Algebraic and Discrete Methods* 2, 311-323.

[9] Mack, Y.P. and Rosenblatt, M. (1979) Multivariate k-nearest neighbor density estimates. *Journal of Multivariate Analysis* 9, 1-15.

Summary

Nearest-Neighbour Methods for Reserving with respect to Individual Losses

This paper focuses on the problem of predicting individual future late claims within a heterogeneous portfolio. For this purpose, the k-nearest-neighbour method is analysed, which intends to carry forward a considered loss according to k appropriately weighted similar observed claims experiences. Using a nonparametric regression approach, a normal approximation and asymptotic confidence intervals for the expected future indemnities are derived. The application of a backtesting procedure to third party liability insurance data reveals that the k-nearest-neighbour estimator performs better than the Chain-Ladder method.

Zusammenfassung

Nächste-Nachbarn-Verfahren zur einzelschadenbezogenen Reservierung

Diese Arbeit behandelt die Problematik der einzelschadenbezogenen Vorhersage zukünftiger Spätschäden innerhalb eines heterogenen Risikokollektivs. Zu diesem Zweck wird das k-nächste-Nachbarn-Verfahren analysiert, bei dem ein betrachteter Schaden entsprechend k geeignet gewichteter ähnlicher beobachteter Schadenverläufe fortgeschrieben wird. Mithilfe eines nichtparametrischen Regressionsansatzes werden eine Normalapproximation und asymptotische Konfidenzintervalle für die erwarteten zukünftigen Entschädigungen hergeleitet. Die Anwendung einer Backtesting-Prozedur auf Daten aus der Haftpflichtversicherung zeigt, dass der k-nächste-Nachbarn-Schätzer genauer prognostiziert als das Chain-Ladder-Verfahren.