AN INTEGRATED SYSTEM FOR ESTIMATING THE RISK PREMIUM OF INDIVIDUAL CAR MODELS IN MOTOR INSURANCE*

BY MALCOLM CAMPBELL

Skandia International, Stockholm†

ABSTRACT

The estimation of risk premium for individual car models is discussed. Cluster analysis is used to identify groups of car models with similar technical attributes. Credibility theory is used to combine estimates of risk premium from individual car model claim statistics, group claim statistics, and a technical assessment carried out by car experts. The procedure is applied to a small set of car models.

KEYWORDS

Credibility, motor insurance, cluster analysis, risk premium.

1. THE PROBLEM AND AN OUTLINE OF THE SOLUTION

In Sweden the premium charged for a private motor insurance policy depends, principally, on four rating factors: geographical area; mileage; no-claims bonus; and car model. In this paper the problem of using car model as a rating factor is discussed. Each car model is allocated to a car-model insurance class, and the premium charged depends on this classification. The allocation of car models to insurance classes presents few difficulties for car models that are common and have been in existence for a few years. However, for cars which are so new or are so uncommon that little (or no) claim statistics are available, the solution is less trivial.

The allocation of car models to insurance classes depends on the risk premium, i.e., expected claim amount in relation to the exposure to risk. The basic problem can thus be defined as estimating a risk premium.

In order to estimate the risk premium three sources of information can be defined.

1. Claim statistics for the car model itself.
2. Claim statistics for car models which are similar to the car model in question.
3. A technical assessment of the car from an insurance point of view.

Provided that the amount of claim statistics is sufficient, the first of these sources yields the best objective estimate of risk premium. However, in the event of a limited amount of claim statistics, for example if the car is new or uncommon, the other two sources must be brought into play.

* A previous version of this paper was presented to the Astin Colloquium at Biarritz, France.

† The work was carried out whilst the author was working at the Swedish Actuarial Research Board.

In the past these three sources of information have been combined in a fairly subjective manner. Although this has probably resulted in satisfactory results, an objective method for combining the sources of information is to be preferred. Figure 1 outlines the solution proposed.

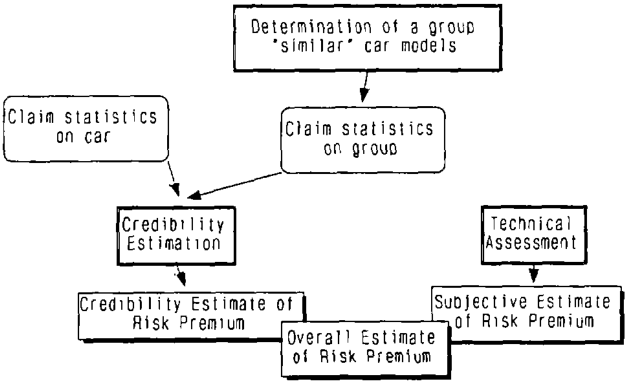
Determination of a group 'similar' car models
Claim statistics on car
Claim statistics on group
Credibility Estimation
Technical Assessment
Credibility Estimate of Risk Premium
Subjective Estimate of Risk Premium
Overall Estimate of Risk Premium

FIGURE 1. Schematic representation of the solution proposed

In this paper we first address the problem of finding "similar" car models, in particular cluster analysis methods are mentioned. This allows us to identify groups (or clusters) where car models within the group have similar attributes (such as engine power, weight etc.).

The intention is to use known technical attributes of car models to form groups where the risk premiums are relatively similar. A small sample of fairly common car models has been taken, and some cluster analyses applied (using attributes engine power, and weight of car). The resulting groups have been examined, with particular reference to the estimated risk premium for third party insurance. This showed that use of the grouping structure went a long way towards differentiating between risk premiums.

Having identified an estimate of risk premium from each of the three sources, the problem remains to combine these estimates. Linear combinations of risk premium have been considered, in other words a weighted combination such as:

$a \times$ estimated risk premium from claim statistics for the car model

$+ b \times$ estimated risk premium from claim statistics for a group of similar car models

$+ c \times$ estimated risk premium from a technical assessment.

Estimates for a, b and c need to be found. In fact for reason of presentation (of both methodology and of results) the process is split into two parts. Firstly a linear combination of the first two sources of information is considered. This problem can be solved by using credibility theory, and use is made of the Bühlmann-Straub model to attain a credibility estimate of risk premium (from now on called a pure credibility estimate in this paper).

This estimate can then be linearly combined with the estimate from the technical assessment. Appropriate weights for the pure credibility estimate and the technical estimate are found by means of a simple extension of the credibility model.

2. ESTIMATION OF RISK PREMIUM FROM A SINGLE SOURCE

2.1. *Motivation*

Before considering how risk premium estimates can be combined, a few words should first be said about how we expect to estimate risk premium from each of the "sources of information".

2.2. *Estimation of Risk Premium from Claim Statistics*

Risk premium is the ratio between the total claims cost and the exposure to risk. The total claims cost is well defined, but a definition of exposure to risk is not so clear cut. Some of the more common measures of exposure used in actuarial studies include: number of policies; number of policy-years; and total premium charged. Each measure has its own advantages and disadvantages, and care must be taken when making a choice. In Sweden we are fortunate to have available in our database a very good measure of exposure: that is a measure called "number of normalised insurance years". To obtain this, the amount of time that each policy is effective is multiplied by the appropriate rating factors (for bonus/ mileage/geographical area) to give a normalised exposure. This is then summed over the appropriate policies to give the total number of normalised insurance years.

The method of obtaining suitable rating factors is not discussed here. One method would be to use a classical factor analysis model (see for example VAN EEGHEN *et al.* 1983) on a sample of car models, using bonus, mileage, geographical area and car model as factors.

Risk premium can then be defined as

$$\text{Risk premium} = \frac{\text{Total amount of claims}}{\text{Number of normalised insurance years}}.$$

The chief advantage of this measure of exposure is that it allows for the fact that certain types of cars are driven by drivers in high (/low) risk groups.

2.3. *Estimation of Risk Premium from a Technical Assessment*

There are a number of studies which have tried to establish a link between the risk premium and a detailed technical description of a particular car model. There are, however, problems in applying such methods not least of which the problem of data availability. For example the costs of relevant spare parts are not necessarily available when a new car model is allocated to an initial insurance class. It is thus preferable to use the present practice where experts conduct a technical assessment of the car and use their judgement to estimate a risk premium. Note that the usual practice is to express this judgement in terms of an assignment to an insurance class. As discussed in the following section, the insurance class is just a one-to-one transformation of risk premium. The assigned insurance class can thus be used to construct an estimated risk premium.

2.4. *Relationship Between Risk Premium and Insurance Class*

For the purpose of calculating premium, each car model is assigned to a car model insurance class. This allotment is quite simply a transformation of the risk premium for the car model in question.

$$i = [f(r)]$$

where: i is the insurance class, r is the risk premium, $[x]$ denotes the nearest integer to x, $f(r)$ is a monotonically increasing function of r.

Note that $f(r)$ should include some method of indexing the risk premium.

One example of a suitable function for $f(r)$ is the linear function:

$$f(r) = k_1 + k_2 r$$

where k_2 is changed from year to year to reflect inflationary increases in claim costs, and both k_1 and k_2 are used to produce a reasonable spread of insurance classes.

The choice of transforming function is essentially a political decision. Its importance for this paper is that technical experts find it easier to express their judgement of risk by reference to insurance class instead of "raw" risk premium. Knowledge of $f(r)$ is thus essential if we are to convert a judgement of insurance class to an estimated risk premium.

3. CLUSTER ANALYSIS OF CAR MODELS

3.1. *Motivation*

One possible source of information to aid in the allocation of car models to insurance classes, is the claim statistics of similar car models. The first question to ask is: what do we mean by similar? It is, of course, possible for experts to look at and compare various car models and say how similar they are. This has several drawbacks:

(a) The measure of similarity is a qualitative instead of a quantitative judgement.
(b) The creation of a group of similar car models will be done on subjective and not objective grounds.
(c) The exercise is time consuming.

In order to achieve a useful quantitative judgement of similarity we need to seek suitable measurable technical data on car models. For example it has long been claimed that there is a relationship between risk premium and the car weight/power.

Having formed a quantitative measure of similarity, cluster analysis methods give a way of forming groups (or clusters) or car models where car models in the same group are in some sense similar.

A full discussion of cluster analysis is not possible here, and for further information the reader is referred to HARTIGAN (1975). The intention in this paper is to illustrate how cluster analysis can be used, rather than to give an exhaustive analysis of the data.

3.2. *A Pilot Study*

For the purposes of illustration, a sample of 50 car models has been taken. The car models chosen are all fairly common, and have a relatively large amount of claim statistics. For each car model we have extracted the following information:

(a) Manufacturer.
(b) Model name.
(c) Code (as defined by the central car classification authority—BKK).
(d) Weight in kg (averaged over cars given the same model code).
(e) Engine power in kW (averaged over cars given the same model code).
(f) Estimated Risk Premium (third party insurance).

Note that we have defined a "car model" as all cars with the same code number.

Figure 2 examines the effect of both weight and engine power on risk premium. Those car models in the top third have been given the symbol +, those in the middle third the symbol 0, and those in the bottom third the symbol −. These symbols have then been plotted on a graph with axes weight and engine power. The relationship between weight/engine power and risk premium is clearly seen.

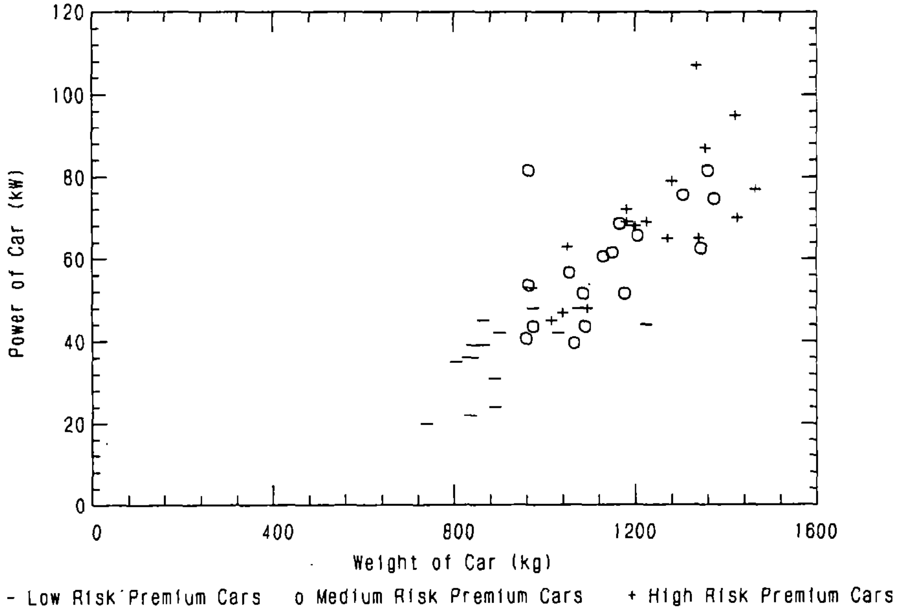

FIGURE 2. The effect of engine power/weight on risk premium

3.3. *Cluster Analysis of Car Models*

As indicated before, it is not possible to fully discuss here the application of all types of cluster analysis. Instead we content ourselves with one particular method, which illustrates how cluster analysis can be used.

The first step in a cluster analysis is to define similarity (or dissimilarity) between objects. In this case we use sum of squares:

Let x_{ij} = measure of attribute j on object i; $i = 1, n$; $j = 1, m$. Define *dis*similarity between objects k and l:

$$d_{kl} = \sum_{j=l}^{m} w_j (x_{kj} - x_{lj})^2.$$

For weights w_j we have used the inverse of attribute variance, i.e.,

$$w_j = \frac{1}{\dfrac{1}{(n-1)} \sum_{i=1}^{n} (x_{ij} - \bar{x}_{.j})^2}; \qquad \bar{x}_{.j} = \frac{1}{n} \sum_{i=1}^{n} x_{ij}.$$

Having formed a matrix of dissimilarities between objects the next stage is to define a criterion for forming groups.

In this case we use Wards method, i.e., we find a partition into G groups such that

$$\sum_{g=1}^{G} \frac{1}{n_g} \sum_{k,l \in S_g; k<l} d_{kl} \quad \text{is minimized}$$

(where n_g = number of objects in group g and S_g = set of indices of objects allocated to group g).

We can prove that this is equivalent to minimizing

$$\sum_{g=1}^{G} \sum_{k \in S_g} \sum_{j=1}^{m} w_j (x_{kj} - \bar{x}_{.j}^{(g)})^2$$

where

$$\bar{x}_{.j}^{(g)} = \frac{1}{n_g} \sum_{i=1}^{n_g} x_{ij}.$$

In other words we are looking to minimize the within group weighted sum of squares.

The method according to Ward is carried out stepwise. We start with n separate groups, each group containing a single object. At each stage we amalgamate that pair of groups which leads to the minimum increase in sum of squares.

The results of such a method can be illustrated by means of a dendogram, or tree diagram. Figure 3 gives the dendogram for the analysis of the sample set of car models (using the attributes engine power and car weight). Note that reading from left to right we can see at what level groups are amalgamated.

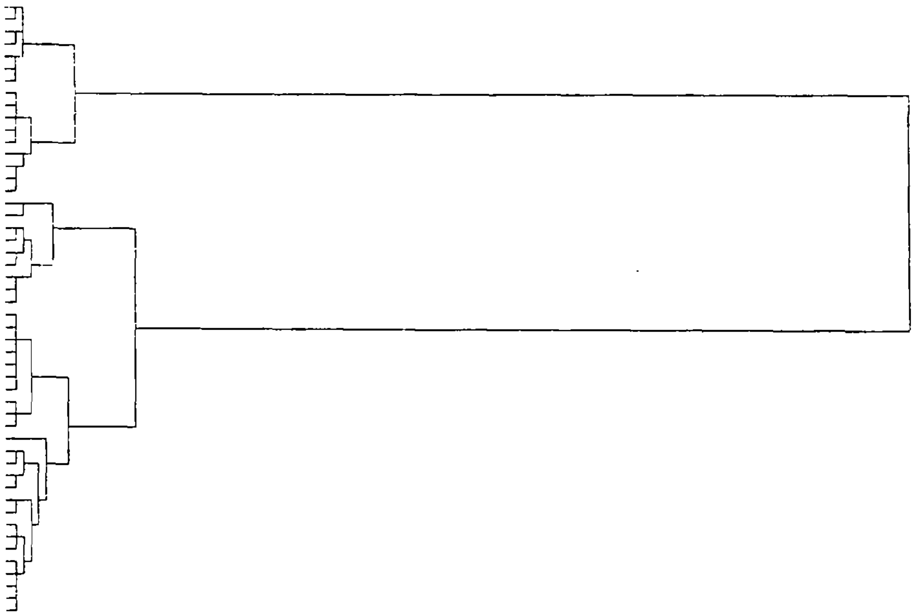

FIGURE 3. Dendogram of cluster analysis

From the dendogram we can see how the sample set of car types can be partitioned into five groups. These five groups are used in the subsequent analysis. The choice of how many groups to take is fairly arbitrary, but in the practical implementation the emphasis would be on producing groups with sufficient numbers of car types to give a good group risk premium estimate.

The important question to now ask is: has this partition helped differentiate risk premiums?

3.4. *Has the Cluster Analysis Helped?*

Our aim in conducting the cluster analysis was to form groups of car models with similar technical characteristics, which hopefully have similar risk premiums. Having conducted a cluster analysis of a sample set of car models, let us now look at the risk premiums in those groups. Note that we have chosen car models with a relatively high amount of claim statistics, thus yielding risk premium estimates which can be regarded as fairly accurate.

To investigate the within group risk premiums, Box Plots have been drawn (see Figure 4). For those not familiar with Box Plots, a quick explanation might be useful.

A Box Plot is a diagrammatic representation of the location and spread of a set of data. The data is first ordered, and the following key values registered.

(a) Max value.

(b) Upper quartile (i.e., $\frac{1}{4}$ way down the list).

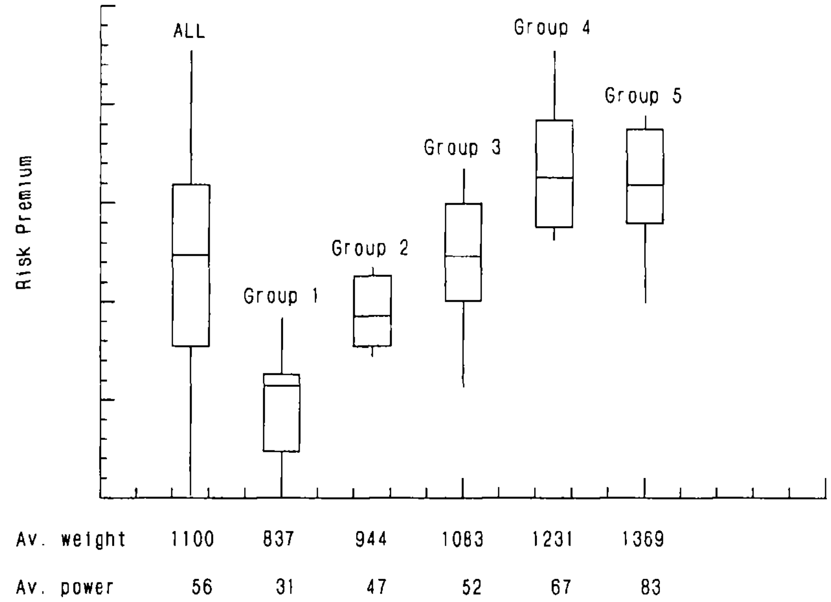

FIGURE 4. Box Plots of estimated risk premium for all cars and individual groups

(c) Median (i.e., $\frac{1}{2}$ way down the list).
(d) Lower quartile (i.e., $\frac{3}{4}$ way down the list).
(e) Min value.

On a vertical scale the spread between quartiles is illustrated by a box and the median is plotted inside this by a straight line. The spread to max/min values is then plotted by simple straight lines (or "whiskers"). More information can be found in VELLEMAN and HOAGLIN (1981).

A Box Plot of all estimated risk premiums, together with Box Plots of estimated risk premiums for individual groups, are given in Figure 4. It is clear that, in general, the groups differentiate quite well between risk premiums, although the difference between groups 4 and 5 is debatable. The average within group weight and engine power are also shown, and we can see that groups 4 and 5 represent the heavier more powerful cars. It could well be that the relationship between risk premium and weight/engine power tails off for higher values.

The fact that two groups have similar risk premiums does not adversely effect the ensuing analysis. The cluster analysis will be a positive aid provided that there are some groups with differing risk premiums.

Further to the analysis with Box Plots we can also conduct an analysis of variance (of risk premiums). Clearly there is a significant difference between the groups in terms of risk premium.

Source	Sum Squares	Degrees Freedom	Mean Square Error	F-Ratio
Between Group	11 030.25	4	2757.56	18.63
Within Group	6 660.77	45	148.02	
Total	17 691.02	49		

3.5. *Further Remarks*

So far, only the two attributes weight and engine power have been considered. Other attributes may also be of help, for example:

(a) average age of car model,
(b) manufacturer,
(c) breadth/length etc.,
(d) form of car (estate/saloon/sports/...),
(e) engine type (diesel/petrol).

The value of weights w_j in the calculation of similarity could also be investigated further to see if better results an be obtained. Indeed other measures of similarity and clustering criterior could be considered.

4. APPLICATION OF THE BÜHLMANN-STRAUB MODEL

4.1. *Motivation*

The relationship between individual risk and collective risk is a key problem for the insurance mathematician. If we knew enough about an individual risk in terms of expected future claims, there would be little problem in calculating an insurance premium. This is rarely the case, and the insurance mathematician needs to consider a collection of risks when estimating the insurance premium for an individual risk.

Credibility theory gives us a way of combining the little information that we *do* have on an individual risk with the information on the collective.

In this case we want to consider how to combine claim statistics for an individual car model with claim statistics for the relevant group of car models.

4.2. *The Model*

The model suggested by BÜHLMANN and STRAUB (1970) can be defined as follows: Let P_{ij} = a fixed (known) measure of the volume of data or risk exposure of car model j in year i; x_{ij} = the loss ratio (i.e., claims paid divided by P_{ij}) of car model j in year i; n = the number of observation years, N = the number of different car models.

We expect car model j to differ in some respect from other car models, and assume that this difference can be characterised by a risk parameter which differs from car model to car model.

$$\theta_j = \text{the risk parameter for car model } j.$$

In the Bühlmann-Straub model the following assumptions are made:

BS1. Conditionally, for fixed θ_j, the random variables $X_{1j}, X_{2j}, \ldots, X_{nj}$ are independent. There exist functions μ and σ^2 of θ, and known positive constants P_{ij} such that

$$E[X_{ij} \mid \theta_j] = \mu(\theta_j)$$

and

$$\operatorname{var}[X_{ij} \mid \theta_j] = \sigma^2(\theta_j)/P_{ij}.$$

BS2. The vectors $(\theta_j, X_{1j}, X_{2j}, \ldots, X_{nj})$, $j = 1, \ldots, N$ are independent and the random variables $\theta_j, j = 1, 2, \ldots, N$ are independent and identically distributed. Define

$$P_j = \sum_{i=1}^{n} P_{ij}$$

$$P = \sum_{j=1}^{N} P_j$$

$$X_j = \sum_{i=1}^{n} P_{ij} X_{ij}/P_j$$

$$X = \sum_{j=1}^{N} P_j X_j / P$$

$$\mu = E[\mu(\theta_j)]$$

$$v = E[\sigma^2(\theta_j)]$$

$$w = \operatorname{var}[\mu(\theta_j)].$$

Bühlmann and Straub show that the greatest accuracy linear estimator of $\mu(\theta_j)$, i.e., the estimator that minimises

$$E[\{E[X_{(n+1)j} \mid \theta_j] - g_0 - g_1 X_{1j} - \cdots - g_n X_{nj}\}^2]$$

is

$$\hat{\mu}(\theta_j) = \alpha_j X_j + (1 - \alpha_j)\mu$$

where

$$\alpha_j = P_j w/(P_j w + v).$$

For the purposes of this paper this estimate is called the pure credibility estimate.

Assuming we have data for X_{ij} and P_{ij}, it just remains to find estimators for the parameters μ, v and w.

Note that μ represents the overall mean value, v represents the within car model variance and w represents the between car model variance.

4.3. *Parameter Estimators*

There is much discussion in the literature on parameter estimation in the Bühlmann-Straub model (see for example Dubey and Gisler 1981). I suggest the following:

$$\hat{\mu} = \frac{1}{\hat{\alpha}} \sum_{j=1}^{N} \hat{\alpha}_j X_j$$

where

$$\hat{\alpha} = \sum_{j=1}^{N} \hat{\alpha}_j$$

$$\hat{\alpha}_j = P_j \hat{w} / (P_j \hat{w} + \hat{v})$$

$$\hat{v} = \frac{1}{N} \sum_{j=1}^{N} \frac{1}{n-1} \sum_{i=1}^{n} P_{ij} (X_{ij} - X_j)^2$$

$$\hat{w} = \frac{1}{\sum_{j=1}^{N} \frac{P_j}{P}\left(1 - \frac{P_j}{P}\right)} \sum_{j=1}^{N} \frac{P_j}{P} (X_j - X)^2 - (N-1) \frac{\hat{v}}{P}.$$

Note that $\hat{w}$ can be less than 0 in which case the estimator of $\mu(\theta_j)$ is taken as:

$$\hat{\mu}(\theta_j) = \sum_{j=1}^{N} \frac{P_j}{P} X_j$$

i.e., we assume that there is no heterogeneity in the portfolio.

4.4. *Additional Comments to the Model*

4.4.1. *Exposure*

The model requires a measure of exposure for each car model. The measure used in this study was "number of normalised insurance years". Further details can be found in Section 2.2.

4.4.2. *An Inflation Free Measure of Loss*

Implicit in the model assumptions is that the claim process is stationary and free from the effects of inflation. This is obviously not true if raw claim costs are used, and a method of indexing is required. One suggestion for getting around the problem is to divide all estimates of risk premium by the risk premium for the collective of all cars.

5. INCLUSION OF TECHNICAL EXPERTS' JUDGEMENT

5.1. *Motivation*

A regular problem in the allocation of car models to insurance classes is that a new car model must be assigned to a class before any such cars are sold, let alone before any claim statistics have been gathered. At present the problem is solved by experts conducting a technical assessment of the vehicle. The result of the assessment is allocation to an insurance class, which is equivalent to an (approximate) estimation of risk premium. This technical assessment can also be of use in the case where the amount of claim statistics is small.

Although the assessment of the experts is very much a subjective judgement, we can still study the past performance of such experts and arrive at an estimate of the accuracy of their assessment. The statistical estimate of risk premium (following the method described in Sections 3 and 4) also has an associated accuracy, and by comparing the accuracy of these two estimates an optimal combination of the two estimates of risk premium can be found.

5.2. *Accuracy of the Experts' Assessment*

To date no comprehensive study of the accuracy of the experts' assessment has been carried out. However, a small study of 72 car models that were initially assessed in the mid 1970's has been carried out. A comparison of initial assessment with the insurance class in 1984 shows the following (for third party insurance class)

Number of car models in study:	72
Number with a higher insurance class 1984:	19
Number with a lower insurance class 1984:	9
Number unchanged:	44

All of the movements (up or down) were no more than one insurance class.

If we assume that the 1984 insurance classes are correct, and use insurance class as a measure of risk premium, we can estimate the bias an accuracy of the experts' assessment as follows:

$$\text{Average error} = \frac{19-9}{72} = 0.13$$

$$\text{Variance of error} = (19+9-10^2/72)/71 = 0.37$$

$$\text{Standard deviation} = 0.61.$$

These estimates cannot be treated as anything more than a coarse estimate of the accuracy of the experts' assessment, but they do give us a starting point.

One of the drawbacks with using this estimate of risk premium is that the experts' assessment is expressed as a whole number. There is no reason why the expert should not be encourage to give a decimal number, for example he might give a car model an insurance rating of 5.4 indicating that, although he would place the car model in insurance class 5, he considers the car model to be one of the more risky in that insurance class.

Ideally we would also like the expert to give an estimate of the accuracy of his own assessment. This is not such a practical idea, but we could ask an expert to classify his assessment as:

(1) Certain or
(2) Fairly certain or
(3) Not certain.

A follow up study would show how accurate experts are and give estimates for variance. As a start we propose:

Classification Estimate	Bias	Standard Deviation of Error
(1) Certain	0	0.5
(2) Fairly certain	0	1.0
(3) Not certain	0	2.0

5.3. *Combination of Technical Assessment and the Pure Credibility Estimate*

In order to combine the estimate of risk premium obtained from the technical assessment with the pure credibility estimate already formed (see Section 4.3), the credibility model must be extended.

Let $\tilde{\mu}(\theta_j)$ = the estimate or risk premium as obtained from the technical assessment. Assume that $E[\tilde{\mu}(\theta_j)|\theta_j] = \mu(\theta_j)$ and let $q = E\{\text{Var}[\tilde{\mu}(\theta_j)|\theta_j]\}$. Conditionally, given θ_j, the random variables $\tilde{\mu}(\theta_j), X_{1j}, \ldots, X_{nj}$ are independent.

It then follows (see appendix for proof) that the estimator that minimises

$$E[\{E[X_{(n+1)j}|\theta_j] - g_0 - g_1X_{1j} - \cdots - g_nX_{nj} - h\tilde{\mu}(\theta_j)\}^2]$$

is

$$\hat{\hat{\mu}}(\theta_j) = \alpha_jX_j + \beta_j\tilde{\mu}(\theta_j) + (1 - \alpha_j - \beta_j)\mu$$

where

$$\alpha_j = P_jwq/(p_jwq + vq + vw)$$

$$\beta_j = vw/(P_jwq + vq + vw)$$

and x_j and $\tilde{\mu}(\theta_j)$ are conditionally uncorrelated, given θ_j. This can be reexpressed as

$$\hat{\hat{\mu}}(\theta_j) = a\hat{\mu}(\theta_j) + (1 - a)\tilde{\mu}(\theta_j)$$

where $a = q(P_jw + v)/(P_jwq + vq + vw)$ and $\hat{\mu}(\theta_j)$ is the pure credibility estimate of Section 4.3.

5.4. *Estimation of the Parameter q*

There are a variety of possible estimators for the parameter q. For example by noting that

$$E\{\mathrm{Var}\,[\tilde{\mu}(\theta_j)\,|\,\theta_j]\} = \mathrm{Var}\,[\tilde{\mu}(\theta_j)] - \mathrm{Var}\,[\mu(\theta_j)]$$

we see that the following is a natural unbiased estimator:

$$\hat{q} = \frac{1}{(N-1)} \sum_{j=1}^{N} (\tilde{\mu}(\theta_j) - \bar{\tilde{\mu}})^2 - \hat{w}$$

where

$$\bar{\tilde{\mu}} = \frac{1}{N} \sum_{j=1}^{N} \tilde{\mu}(\theta_j).$$

I would however recommend that an estimate of q is obtained by means of some form of follow up study where the initial assessment made is compared with the risk premium obtained after several years claim experience. Such a study is important in particular to analyse any bias in the assessment made, since the credibility analysis is built on the assumption that

$$E[\tilde{\mu}(\theta_j)\,|\,\theta_j] = \mu(\theta_j).$$

In such a follow up study, one possible estimator is

$$\hat{q} = \frac{1}{N} \sum (\tilde{\mu}(\theta_j) - X_j)^2 - \frac{\hat{v}}{N} \sum \frac{1}{P_j}.$$

This can be justified by noting that

$$E\{\mathrm{Var}\,[\tilde{\mu}(\theta_j)\,|\,\theta_j]\} = E[\{\tilde{\mu}(\theta_j) - X_j\}^2] - \mathrm{E}\{\mathrm{Var}\,(X_j\,|\,\theta_j)\}$$

In the case where P_j is large, i.e., when X_j is an accurate estimate of risk premium this estimator of q approximates to the "variance of error" found in Section 5.2.

6. APPLICATION OF THE PROCEDURE

To illustrate the procedure outlined in this paper, seven uncommon car models have been taken and an attempt has been made to estimate their risk premiums (using the sample of "common" car models analysed in Section 3).

The group structure as presented in Section 3 has been preserved. The weight an engine power of each new car model was then used to allot the car model to the "nearest" group. The risk premium calculated from individual car model statistics was then combined with the group risk premium (following the method outlined in Section 5) to produce a pure credibility risk premium.

TABLE 1

APPLICATION OF THE PROCEDURE TO SOME UNCOMMON CAR TYPES

	Car Weight (kg)	Engine Power (kW)	Assigned to Group	Number of Claims	Risk Premiums (Car Statistics)	Risk Premiums (Group Statistics)	Risk Premium (Technical Expert)	Pure Credibility Estimate of Risk prem.	Overall Estimate Risk Prem.
1	950	35	3	8	6.62 (±11.31)	9.02 (±4.74)	8.00 (±3.92)	8.48 (±4.37)	8.15 (±2.92)
2	1330	53	4	73	17.03 (±6.62)	11.06 (±5.25)	10.00 (±3.92)	13.36 (±4.12)	11.14 (±2.84)
3	1200	74	4	29	23.14 (±11.68)	11.06 (±5.25)	10.00 (±3.02)	12.92 (±4.78)	10.80 (±3.03)
4	1470	104	5	25	14.89 (±11.17)	11.28 (±6.49)	12.00 (±3.92)	12.19 (±5.61)	12.04 (±3.21)
5	1660	140	5	53	20.79 (±9.82)	11.28 (±6.49)	11.00 (±3.92)	14.17 (±5.41)	11.71 (±3.17)
6	1550	101	5	73	15.94 (±7.66)	11.28 (±6.49)	14.00 (±3.92)	13.23 (±4.96)	13.80 (±3.07)
7	1430	70	5	115	11.28 (±5.45)	11.28 (±6.49)	10.00 (±3.92)	11.28 (±4.17)	10.42 (±2.86)

Notes: Claim statistics over a five year period were taken. For reasons of confidentiality Risk Premium is expressed in terms of artificial units. All estimates of Risk Premium are given with an estimated 95% confidence interval.

The initial risk premium as estimated by technical experts was then taken and combined with the credibility risk premium. Note that the value used for expected variance of a technical experts estimate of risk premium (i.e., $\hat{q}=4$) is not based on any factual study, but is given for illustrative purposes only.

The results of this exercise are given in Table 1. The table gives details of each car's weight and engine power together with the assigned group. In the fourth column the number of claims for the car model in question is given to give an idea of how few claim statistics are available on the individual car models. The remaining columns give the various estimates of risk premium, firstly from the various "sources of information" an then the combinations of those estimates. Each estimate of risk premium is given with an approximate 95% confidence interval (i.e., 1.96 × standard deviation).

The estimates used for confidence intervals of the risk premium estimates from individual "sources of information" were calculated using the formula:

(i) $1.96\sqrt{v/P_j}$ for risk premium calculated from car model claim statistics.
(ii) $1.96\sqrt{w}$ for risk premium calculated from the group claim statistics.
(iii) $1.96\sqrt{q}$ for risk premium estimates by technical experts.

In addition the following estimates were used for confidence intervals of the risk premium estimates using credibility and using the overall procedure:

(i) $1.96\sqrt{(vw/(P_jw+v))}$ for the pure credibility estimate
(ii) $1.96\sqrt{(vwq/(P_jwq+vq+vw))}$ for the overall estimate.

In addition Table 2 gives details of the effective weight given to each "source of information" in calculating the overall estimation of risk premium.

TABLE 2
WEIGHTS GIVEN TO THE VARIOUS SOURCES OF INFORMATION

	Individual Claim Statistics	Group Claim Statistics	Technical Expert
1	0.07	0.38	0.55
2	0.18	0.29	0.53
3	0.07	0.33	0.60
4	0.08	0.25	0.67
5	0.10	0.24	0.66
6	0.16	0.22	0.62
7	0.27	0.19	0.54

It can be seen that the estimates of risk premiums calculated from claim statistics for individual car models are far from accurate. This is, of course, not surprising given the few claims experienced during the five year period. A much improved accuracy in risk premium estimate can be found by combining with the risk premium as calculated from claim statistics for the whole group (see column 8). Further improvement can also be obtained by combining with the risk premium as estimated by technical experts.

The calculations carried out here are based on a small set of car models (just over 50). In Sweden there are more than 1500 car models on the road, and if we were to use all 1500 then we would expect to achieve the following benefits:

(1) A larger number of groups, where within group variance of risk premium was smaller (i.e., w is smaller).

(2) A larger number of car models in each group, leading to more accurate estimates of the credibility parameters (i.e., v, w, q and μ).

The assumed accuracy of technical experts judgement is also very conservative, although one should bear in mind that they are likely to have a lot more problems assessing an uncommon car model than a common car model.

Overall we would expect the improvements in risk premium estimate illustrated in Table 1 to be surpassed when the method is applied to all car models.

7. CONCLUSIONS

This paper sets out the methodology for combining three estimates of risk premium for a particular car model from three sources of information: claim statistics on the car model in question; claim statistics on similar car models; and a technical assessment by experts. Moreover methods for identifying "similar car models" are discussed.

Application of the methods to a limited set of car models give very encouraging results. Technical attributes such as car weight and engine power appear to be helpful in identifying groups of car models which have similar risk premiums. It is also apparent that risk premium estimates for car models with few (or no) claim statistics can be dramatically improved.

Although the results are encouraging, further work is needed before the procedures are used in practice. Analysis using a larger sample of car models is necessary, and particular attention should be made to the following points.

(1) How well the assumptions behind the credibility model are fulfilled.

(2) Whether improved credibility parameter estimates (i.e., of v, w, q and μ) can be found.

(3) Whether other technical attributes and/or other clustering technics are useful.

(4) How many years of claim statistics should be used, and whether other forms of credibility estimation (such as the evolutionary models suggested by SUNDT (1983)) give better results.

(5) How well technical experts can assess the risk premium for car models.

(6) How the method can be applied to the different elements of insurance (fire, theft, third party etc.).

The results of using the procedures described here should also be compared with present practice, where estimates of risk premium from the three sources are, in effect, *subjectively* combined. It is not expected that a dramatic improvement will be discovered, but the point is that the procedures described here enable the process of car model insurance classification to be put on a more *objective* footing.

8. ACKNOWLEDGEMENTS

I am extremely thankful for the discussions I have had with my colleagues regarding this paper. In particular Jan Jung for his continuous encouragement and Bjørn Sundt for his comments on Sections 5.3 and 5.4.

APPENDIX

Proof of the result stated in Section 5.3.

We wish to minimise

$$E[\{E[X_{(n+1)}|\theta_j]-g_0-g_1X_{1j}-\cdots-g_nX_{nj}-h\tilde{\mu}(\theta_j)\}^2]$$

note first that

$$E[(Z-k)^2]=\mathrm{Var}\,(Z)+[E(Z)-k]^2 \tag{1}$$

where k is a constant and Z is a random variable.

The second term is minimised (at zero) when $E(Z)=k$. Applying this we obtain

$$g_0=\left(1-\sum_{i=1}^{n} g_i-h\right)\mu.$$

Secondly note that:

$$\mathrm{Var}\,(Z)=\mathrm{Var}\,\{E(Z|\,Y)\}+E\{\mathrm{Var}\,(Z\,|\,Y)\}.$$

Applying this to the first term in (1), we are required to minimise

$$\mathrm{Var}\left[E\left\{\mu(\theta_j)-\sum_{i=1}^{n} g_iX_{ij}-h\tilde{\mu}(\theta_j)\,\middle|\,\theta_j\right\}\right]+E\left[\mathrm{Var}\left\{\mu(\theta_j)-\sum_{i=1}^{n} g_iX_{ij}-h\tilde{\mu}(\theta_j)\,\middle|\,\theta_j\right\}\right]$$

$$=\mathrm{Var}\left[\left(1-\sum_{i=1}^{n} g_i-h\right)\mu(\theta_j)\right]+E\left[\sum_{i=1}^{n}\mathrm{Var}\,\{g_iX_{ij}\,|\,\theta_j\}+\mathrm{Var}\,\{h\tilde{\mu}(\theta_j)\,|\,\theta_j\}\right]$$

$$=\left(1-\sum_{i=1}^{n} g_i-h\right)^2 w+\sum_{i=1}^{n} g_i^2v/P_{ij}+h^2q$$

by differentiating with respect to h and g_i we find the minimum when

$$h=vw/(P_jwq+vq+vw)$$

$$g_i=P_{ij}wq/(P_jwq+vq+vw)$$

thus

$$\sum_{i=1}^{n} g_iX_{ij}=P_jwqX_j/(P_jwq+vq+vw)$$

hence the result.

REFERENCES

BÜHLMANN, H. and STRAUB, E. (1970) Glaubwürdigkeit für Schadensätze. *Mitteilungen der vereinigung Schweizerischer Versicherungsmatematiker* **70** (1), 111-133.

DUBEY, A. and GISLER, A. (1981) On parameter estimators in credibility. *Mitteilungen der Vereinigung Schweizerischer Versicherungsmatematiker* **81** (1), 187-212.

VAN EEGHEN, J., GREUP, E. K. and NIJSSEN, J. A. (1983) *Surveys of Actuarial Studies, No 2, Rate Making.* Nationale-Nederlanden N.V., Rotterdam.

HARTIGAN, J. A. (1975) *Clustering Algorithms.* John Wiley & Sons, New York.

SUNDT, B. (1983) Finite Credibility Formulae in Evolutionary Models. *Scandinavian Actuarial Journal* 106-116.

VELLEMAN, P. F. and HOAGLIN, D. C. (1981) *Applications, Basics and Computing of Exploratory Data Analysis.* Duxbury Press, Boston, Massachusetts.

MALCOLM CAMPBELL
Skandia International, Box 7693, S-103 95 Stockholm, Sweden.