Vehicle Postcode Zoning in Personal Lines Rating

Vehicle Postcode Zoning in Personal Lines Rating

Working Party

VEHICLE POSTCODE ZONING IN PERSONAL LINES RATING WORKING PARTY

Duncan Anderson	Roger Massey
Saideh Charlton	Chris Spiller
Dave Coughlan (Chair)	Kim Pape
Mark Harrison	James Tanser
Stephen Jones	Richard Verrall

Summary

This paper examines the methods and information that can be used to group individual postcodes in to rating areas for pricing.

We have tried to highlight where companies can obtain additional useful information to help them in this exercise and have set out a description of some methods that can be used to complete a zoning exercise.

We have also examined how companies currently rate postcodes. This has highlighted some inconsistencies in our approach as a market to individual postcode rating. This may suggest this is an area which insurers could usefully review.

CONTENTS

1. Introduction

The working party set out to examine how postcodes were allocated to rating districts. We felt that this was an area often overlooked within the statistical rating of a portfolio. Existing structures are often historic, based on previously unknown or undocumented assumptions. Structures tend to be reviewed regularly with small adjustments to the existing basis being made.

Our aim in this paper is

- to outline areas where additional information can be obtained to help further refine the risk assessment for postcodes
- to discuss some practical methods that can be used to zone postcodes
- to investigate how market rates vary by postcode

Owing to the large scale of the exercise, a limited scope had to be defined and we decided to concentrate on motor theft frequency. We realised that theft represents a decreasing percentage of motor claims in the current environment but we felt that the methods and discussions that apply for theft risk can also apply to a number of other areas. We hope they will give a starting point for actuaries and statisticians to improve methods currently in use. The opinions expressed in this paper are those of the working party members and are not necessarily the views of the companies for which they work or of the Institute of Actuaries.

2. Sources of information

We aimed to outline sources from which additional information could be obtained. We set out below at a high level the information that is available and where appropriate a fuller description and contacts are available in Appendix 1.

2.1 Internal data

For a company with a large amount of claims and exposure experience, using internal data to group postcodes into homogeneous groupings has several advantages over using data from other sources. These include:

- there is no purchase cost
- the data may be directly relevant to the job in hand
- the data will probably be at the required level of detail and accuracy

The fact that the data may be directly relevant to the job in hand is an important point. For example, when deriving postcode groupings for a motor rating structure the company's own motor claims experience may be used. That is not to say that other internal data is of no value. If it was

shown that for the same postcode there was a correlation between the risk of theft on motor policies and the risk of theft on household contents policies, then the claims experience from the household business could be used to support grouping postcodes for motor. This may be of particular value to a company with a large household book of business and a small motor book.

2.1.1 The main drawback of using internal data is the company may have no data or the data is sparse.

2.1.2 Data is available on how other companies group postcodes from several sources. One such source is other companies rating/underwriting guides which may be issued to sales staff or brokers. These may be freely exchanged between companies as part of an 'exchange of market information agreement'. An alternative source is from a competitive position quotation package such as 'What-if?' or 'Prémium'. For a given risk profile this gives premium quotations from several companies. It is possible from such systems to derive how they group postcodes.

2.2 External provider information

External data providers may be able to provide information that can be used to assist in grouping postcodes. This information has to be purchased and therefore has a cost that needs to be outweighed by the additional benefit to the company. ISL has kindly supplied information as one method of groupings postcodes in to rating areas.

2.2.1 The main advantages in using this type of information essentially complement the disadvantages of using your own data. Where data is sparse this information may add extra credibility to the results obtained or provide information over and above that held in existing internal data.

2.2.2 Drawbacks can include the data not being an ideal fit to the purpose and the fact that it is often in the format of a risk score, this score being a relative measure of how much risk there is within a particular postcode. It is often not clear how such scores have been constructed. Details of current information available and providers are shown in Appendix 1.

2.3 The Census

A census has been carried out at least every 10 years since 1801 with the exception of 1941. The amount of data collected has increased over the years. This makes it a potentially excellent source of external data to bring into the rating process.

2.3.1 Data collected in the 1991 Census is available at postcode sector level. This data is in the form of tables known as the Small Area Statistics. The Office for National Statistics (ONS) produces Small Area Statistics and there are 95 tables available. The ONS uses the postcode sectors that were valid at the time the census was taken. There were 7,722 sectors in 1991. Considerable adjustments are needed to produce anything fully relevant to today's postcode geography. There are 1.6 million postcodes in the UK and every year around 10% of them change. *[source: postcode update number 28]*

2.3.2 Apart from changes in postcode geography the main problem with the Small Area Statistics is that the data is very old (the 1991 Census was completed on 22 April 1991). Also some of the questions were only asked to 10% of the population and responses were imputed to the whole population. The next Census will take place on 29 April 2001 and the outputs from it will be available in the financial year 2002/03. All the questions will be asked to the whole population so the problem of 10% data will be eliminated. Improvements to the 2001 census are listed in Appendix 1

2.4 The ABI Motor and Household Risk Statistics Schemes (MRSS and HRSS)
An alternative to using your own company's data as the basis for a postcode zoning system is to participate in an industry data-sharing arrangement, which may typically provide access to significantly greater volumes of data, and to a preliminary analysis of that data based on appropriate actuarial techniques. Two industry data-pooling arrangements, one (the MRSS) relating to private motor business and the other (the HRSS) to household business, are operated by the Association of British Insurers for the benefit of participating members. Each scheme currently provides a detailed analysis of standardised claims experience, including theft claims frequencies and severities, at the postcode sector level.

2.5 Reinsurers' Data
Reinsurers collect data from the companies they reinsure. Depending on the covers bought this data will be at varying levels of detail. Often they aggregate this data, model it and then sell it to primary insurers. As this is an amalgamation of several companies data it is less relevant than an insurer's own data but may be more relevant than other sources of data.

2.5.1 Reinsurers' data is probably more useful in Household insurance where some perils (eg subsidence and flood) are such that any insurer is unlikely to have enough claims and exposure data to form an adequate risk assessment

at postcode level. For each peril, models are available which give a postcode to district allocation together with a risk score.

2.6 Information not available at postcode level

We also investigated a number of other sources that described theft risk. Information was not available at individual postcode level. These sources did contain a wealth of information on risk and could be used as a high level reasonability check on answers derived from any zoning analysis. The main items examined were the British Crime Survey, police records and information from the Claims Underwriting Exchange.

3. Postcode grouping methods

We attempted to offer a number of methods of combining postcodes into rating areas and each of these is discussed in the following section. We have tried to give an overall discussion of methods and practical hints for completing and debugging each. Some sample code and formulae for each method have been included in the Appendices and only a high level description has been outlined here. In total 5 of the 6 methods below were run on actual data. These methods were as follows:

- Weighted distance smoothing method
- External provider information method
- Credibility based method
- Spatial model method
- Market average method
- Modern Heuristic method

Data files were prepared by a number of companies on the working party and each company applied a selection of the above methods to this data. The results of each method were compared to assess the goodness of fit of each of the methods in order to assess the best method to use for zoning of postcodes. Details of the data preparation are shown in Appendix 3

3.1 Weighted distance smoothing method

This method attempts to produce a smoothed theft risk based on a weighted average of the individual postcode district and all other postcodes based on their proximity.

This method is defined as:

$$r_i^* = Ar_i + (1 - A)\frac{\sum_{j \neq i} e_j r_j / (d_{ij})^n}{\sum_{j \neq i} e_j / (d_{ij})^n},$$

where r_i^* is the adjusted risk for postcode district i, r_i is the unadjusted residual risk for postcode district i, e_i is the exposure in vehicle years for postcode district i, and d_{ij} is some measure of the distance between postcode district i postcode district j.

In our analysis we defined d_{ij} as the Euclidean distance between the centroids of the postcodes. This method could, however, be refined by including other factors in to this measure of distance. For instance, the difference in the urban densities of the two postcodes could be incorporated in to the measure. This would, for example, result in the adjusted risk in a rural area being more affected by the experience in nearby rural areas than by the experience in nearby urban areas, reflecting the hypothesis that the nearby rural areas were more likely to be similar in nature.

A and n are parameters which need to be determined by investigating the predictive properties of the method on a sample of the experience.

In our investigation, we treated A and n as being fixed parameters. The method could, however, be significantly improved if A varied by postcode, and was defined as being a function of the exposure of the postcode in question.

3.2 External provider information

ISL, an external provider, supplied a theft score for districts in Northern England for assessment. The scoring depends on a number of factors and includes information from police forces and insurers in the UK. This information is then modified by a range of 255 demographic, socio-economic, behavioural and built environment descriptors for each unit postcode including:

- the level of unemployment compared to the UK average
- the proportion of households with bad debts compared to the UK average
- the density of housing and proximity to town centres and pedestrian thoroughfares
- access to road and rail network
- a neighbourhood risk index

3.3 Credibility Method

This method calculates the risk associated with the postcode based on a mixture of the postcodes own experience, the experience of the sector that the postcode is in and the experience of its postcode area. For the data exercises we used the two levels of postcode district and postcode area. This method can be defined as,

$$\tilde{f}_{X12} = c_{X12} f_{X12} + (1 - c_{X12}) f_X$$

where c_{X12} is the postcode district credibility,

f_X is the postcode area standardised theft claim frequency,

$\tilde{f}_{X12}$ is the postcode district credible theft claim frequency,

f_{X12} is the postcode district standardised theft claim frequency.

Derivation of the postcode district credibility factors are given in Appendix 2.

3.4 Average Market Method

This method was based on output from a standard broker quotation system. Postcodes were categorised based on the average premium quoted for a single example risk by a number of leading insurance companies. This method gives a view of how the market on average rates a postcode. It could be modified to include only a selection of insurance companies. It may be useful in determining how much you have to change your rates as a company in order to make your product competitive for certain postcodes. Details of the method used in this exercise is given in Appendix 2.

3.5 Spatial model method

Spatial models rely on the assumption that points that are close together are more similar than those that are far apart. This method therefore assumes that the risk of a postcode is associated with that of its neighbours. The models that we examined in the working party were based on the Bayesian approach to statistics. A full description of approach and method is in [Boskov M., Verrall R.J., (1994) *Premium Rating by Geographic Area using Spatial Models* Astin Bulletin Volume 24 No. 1]

3.6 Modern Heuristic Technique

Due to the growing complexity and the increasing size of combinatorial optimisation problems, researchers have moved to using and developing heuristic search techniques to achieve acceptable results. A heuristic is a technique that seeks good (i.e. near optimal) solutions at a reasonable computational cost without being able to guarantee either feasibility or optimality or even in some cases, how close to optimality a particular solution is. Heuristics techniques, such as genetic algorithms, neural networks, simulated annealing and tabu search can be used to solve problems involving the categorisation of postcodes to districts. Heuristic

methods are based on expansion of local search methods. The basic problem that heuristic techniques attempt to solve can be set out as follows:

Let
Q be the set of all objects to be clustered,
$n = |Q|$ be the number of objects in Q,
$k <= n$ be the maximum number of clusters,
$P = \{\mathbf{p}: \forall\ i \in \{1,\ldots,n\}, p_i \in \{1,\ldots,k\}\}$ be the set of all partitionships,
$J: P \rightarrow \Re$ be the internal clustering condition;
Then
Minimize J(**p**)
Subject to
$\mathbf{p} \in P$.

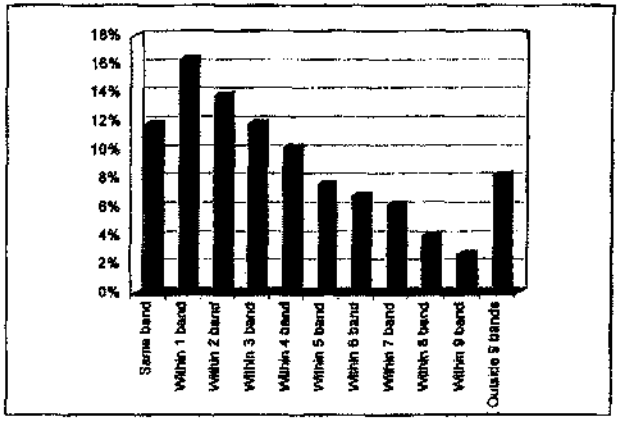
A variety of algorithms can then be used to solve this problem. The algorithm chosen determines the criteria for clustering for example, minimising the total squared distance of the objects to their associated cluster means.

Sample code for the simulated annealing technique is shown in the Appendix and a full description of this approach is given in [Brown D.E., Huntley C.L. (1991), *A Practical Application of Simulated Annealing to Clustering*]

3.7 Evaluating results

The results of the group's data analysis showed that there was little to choose between the credibility and weighted distance methods tested, with the goodness of fit measures similar. All methods produced, as expected, a better goodness of fit than the base model. The base model did not include a postcode rating variable.

3.7.1 The market method used appeared to place postcodes in significantly different zones than a number of the other methods.



From the above graph these two methods only placed 41% of postcodes within 2 postcodes bands of each other, with 8% of cases placed in a band that was 9 bands higher or lower depending on the method chosen. The bandings were also highly dependent on the parameters chosen for the model or the clustering criteria.

3.7.2 The group's conclusion from our initial analysis suggests that a mixture of the methods described above would produce the best results. A considerable amount of time and effort is required to decide on the most appropriate method based on a given set of data

3.8 Practical issues

There are a number of practical issues that need to be resolved when completing a zoning exercise and we have outlined just two of these in what follows. We have not given a full description of how these issues can be overcome but have tried to address them at a high level.

3.8.1 Allocation of postcodes where the insurer has no exposure

There will be some valid postcodes where an insurer will not have any historic exposure. These postcodes need to be allocated to a zone in a sensible and pragmatic manner. A number of approaches could be considered to solve this problem as follows:

- Allocate based on the category of the postcode's neighbours
- Allocate based on other characteristics of the postcode such as house type, density, car ownership etc
- Allocate based on the market's perception of the postcode's risk

None of the above solutions are ideal as they introduce an element of subjectivity in to the categorisation process, however the nil exposure postcodes must be allocated to a rating area.

3.8.2 Implementing the findings

The implementation of a new postcode zoning method could lead to significant changes in premium for some customers and the full extent of the new pricing structure should be assessed before implementing. Key areas that need to be investigated include:

- How will the company's competitive position change?
- How will the changes affect the company's existing customers?
- Can pricing systems incorporate the new rating structure?

Each of the above questions is complex and we have not discussed them further in the paper. They need to be quantified before any premium changes are passed on to customers based on a re-zoning exercise. However, the additional knowledge gained on the true risk of a postcode following a zoning exercise enables direction to be set to manage a portfolio of business towards more profitable areas.

4 UK market overview

In addition to considering statistical methods which can be used in postcode rating analyses, we also carried out a limited investigation of how UK motor insurers rate by postcode in practice.

4.1 We considered an example risk (details given in appendix 2) and with the co-operation of EL Systems Ltd, to whom the group expresses its thanks, used the broker quotation analysis tool *Prémium* to obtain comprehensive motor quotations from 18 large UK motor insurers for that risk in every postcode district in the UK. The resulting quotations were analysed with the following conclusions.

4.1.1 Most of the 18 insurers appeared to categorise postcodes into 20 or so groups, and then to apply multiplicative premium adjustments for each category, with the highest rated category generally being charged premiums of the order of twice that of the lowest rated category (all other factors being equal).

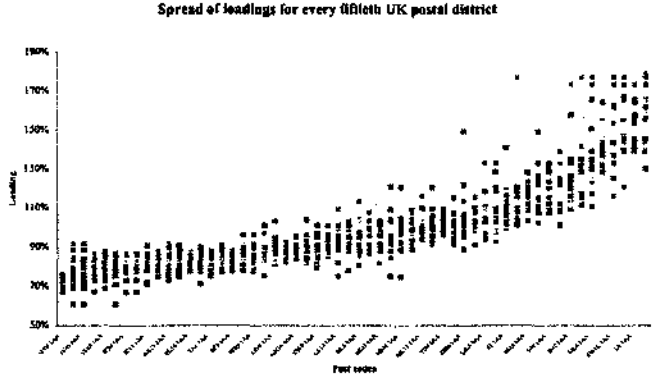
Details of the exact number of categories used by each insurer, together with the lowest and highest rated category multipliers, are set out in the below table.

Insurer	Number of post code categories	Multipliers from	To	Maximum price difference
1	10	0.76	1.61	212%
2	11	0.68	1.34	197%
3	12	0.75	1.72	229%
4	14	0.43	1.42	330%
5	15	0.74	1.63	220%
6	16	0.68	1.48	218%
7	16	0.68	1.49	219%
8	17	0.58	1.55	267%
9	17	0.68	1.44	212%
10	18	0.89	1.92	216%
11	19	0.60	1.62	270%
12	19	0.74	1.43	193%
13	19	0.77	1.78	231%
14	20	0.58	1.84	317%
15	20	0.81	1.76	217%
16	21	0.64	2.00	313%

17	22	0.63	1.74	276%
18	125	0.74	1.67	226%

4.1.2 Insurer number 18 appears to use 125 postcode categories. This could result from using different categorisations for different claims elements – perhaps, for example, a 5 level categorisation for theft together with a 25 level categorisation for other claim types.

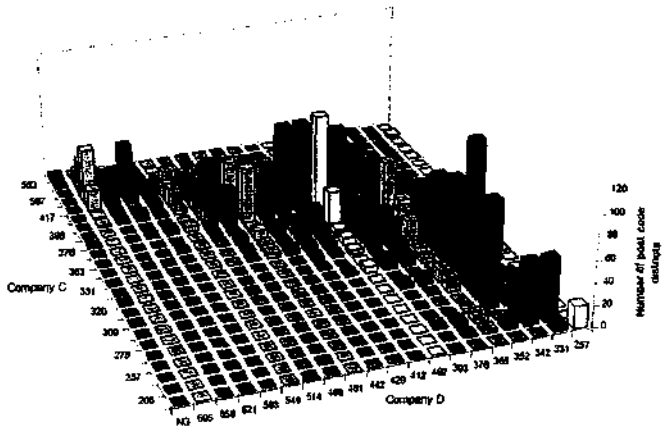
4.1.3 There seems to be some difference in the way in which any given postcode district is treated by the market. For example, we considered the multiplicative loading which each insurer allocated to a postcode district (relative to each insurer's "average" premium for the risk in question which was estimated from an unweighted average of the quotations given across all UK postcode districts). For each district the average loading (across companies) was calculated and districts were ranked by this. Every fiftieth postcode was considered, and the loadings for these postcodes used by each of the 18 insurers were plotted on a graph.



It can be seen that although there is a certain degree of consistency in the way in which postcodes are rated, for any given district there can be quite a significant difference between the minimum and maximum loadings applied in the market.

4.1.4 To investigate this further we went on to consider how consistently different pairs of insurers treated each postcode. For each of the 153 possible pairs of the 18 insurers a graph was produced showing the number of postcode

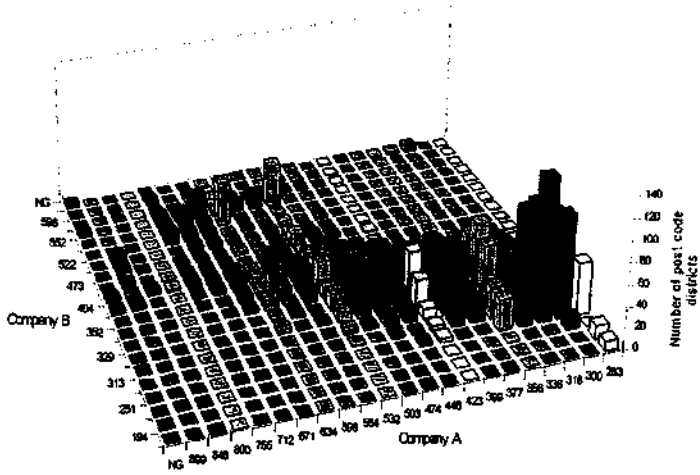
districts which, for the standard risk in question, had different premiums quoted by each insurer. One such graph is shown below.

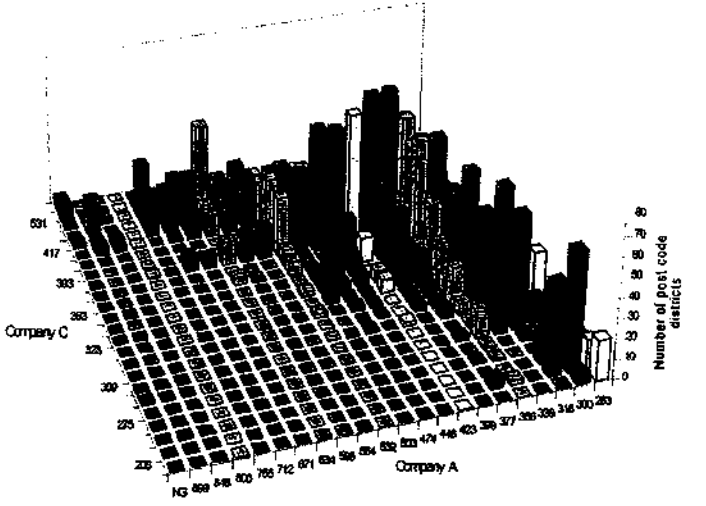


It can be seen from the previous graph, for example, that there are around 20 districts which, for the risk in question, are charged both £206 by Company C and £257 by Company D.

If the two companies being considered categorised all postcode districts in an identical way, these graphs would show a very high correlation, with all postcode districts falling along a thin (though not necessarily straight) line from the top left corner of the graph to the bottom right. In practice, it can be seen that whilst there is some correlation, many postcodes are treated in quite different ways. For example, (for the risk in question) districts for which Company C charges the same premium of £387 are charged anything from £360 to £500 by Company D.

This example considered only one pair of insurers, but very similar results can be seen for most pairs of insurers. Further examples are set out below.


Company B
Company A
Number of post code districts
NG
596
552
522
473
434
392
339
313
251
194
859 848 800 755 712 671 634 596 554 532 503 474 446 423 399 377 356 338 318 300 283
140
120
100
80
60
40
20
0


Company C
Company A
Number of post code districts
531
417
303
263
225
859 848 800 755 712 671 634 596 554 532 503 474 446 423 399 377 356 338 318 300 283
80
70
60
50
40
30
20
10
0

4.1.5 The analysis above considered one example risk only. In practice the loadings made for geographical area can vary depending upon the value of other rating factors. That is to say that rather than rating factors having an independent effect upon the premium, many insurers consider the combined effect or "interaction" of two or more rating factors in determining the premium.

In the case of geographical area, we noted that most insurers included an interaction between area and vehicle group in their rating structures, with high category vehicles in high risk areas generally being loaded more. Of the 18 insurers considered, 12 seemed to use such an interaction, although generally the size of the interaction effect over and above the simple area and vehicle group multipliers was relatively small (generally the additional effect corresponding to multipliers of between 0.95 and 1.05).

In addition, 7 of the 18 insurers also included an interaction term between area and the location in which the vehicle is left overnight. (The hypothesis being that garaged reduces car theft in urban and high risk areas, but has much less effect in rural and low risk areas). Further rating factor interactions were not investigated in this exercise.

It should be noted that this market analysis is subject to a number of limitations. For example, only postcode districts were considered. Many of the insurers considered may rate at sector level, which could distort some of the above analyses. Furthermore, the 18 insurance schemes considered were all schemes available to brokers – no direct writers were analysed.

Appendix 1

- The Census – detailed information

  Potentially useful data items available at sector level include:

  - number of adults in household, implying population density
  - economic activity (unemployed, full time employee etc.) which together with a number of the other tables could be used to create a social deprivation index
  - age
  - car availability by method of travelling to work
  - car availability by socio-economic group
  - number of student households that have cars
  - number of cars to which each household has access by number of people in the household
  - number of cars each household has access to by age of Household Reference Person (formerly called the head of household) and age of youngest child
  - information on the households which don't have a car (eg household composition, ethnic group)
  - information on means of transport to work (eg socio-economic group, number of cars in household, gender, age)
  - distance travelled to work by means of transport to work and gender
  - distance travelled to work by age and gender.

- Improvements to 2001 census

  Consultations about the outputs that will be created from the 2001 Census are already underway and packs detailing the proposals are in circulation. An interesting development is that the ONS is in consultation with geographical information systems providers to plan the enumeration areas and the areas for which output is available. The current proposal is that Output Areas would contain on average, about 100 to 125 households. An advantage of this could be that lower level information would be available. However paragraph 120 of the White Paper pertaining to the 2001 Census states: "Special precautions may apply particularly to statistical output for small areas. Measures to ensure disclosure control will include...

  - randomly modifying some data before the statistics are released."

It is proposed to make a small number of amendments to the database designed to prevent the recipient of any product making any deductions about a household or individual record. For example any cross-tabulation of one variable against another can disclose information if it contains a row filled with zeroes except for one cell. If the user knows a particular individual is in this particular row then they can deduce which column the individual is in. This would result in the disclosure of previously undisclosed information from the Census. The amendments may take the form of pairing records from a sample and swapping them or blanking a sample of items in some records and imputing values back in.

In the 2001 Census each person will be asked to provide information on their current gross income including earnings, pensions, benefits, interest from savings or investments, rent from property, maintenance payments and any grants received. The inclusion of this question is still under discussion but clearly the results would be a useful addition to the current range of data available from the Census.

- Census contacts

England & Wales	Scotland	2001 Census
Sue Bates	Census Customer Services	Neil Lander Brinkley
Census Marketing	GRO (Scotland)	2001 Census Programme
OPCS	Ladywell Road	Output Production Project
Segensworth Road	Corstorphine	Segensworth Road
Titchfield	Edinburgh	Titchfield
Fareham	EH12 7TF	Fareham
Hampshire		Hampshire
PO15 5RR		PO15 5RR
Tel 01329 813800	Tel 0131 314 4254	Tel 01329 813522

- HRSS and MRSS – detailed information

The Motor Risk Statistics Scheme (MRSS)

The MRSS is a motor insurance data-pooling arrangement that provides its participants with detailed statistical analyses of their own private car policy and claims data, and also that of the aggregate membership. The MRSS is open to insurance companies writing private car business in the United

Kingdom, and also to Lloyd's syndicates. Members benefit from a comparison of their own experience with market norms, and from access to analyses of claims experience based on large data volumes, which can be of considerable assistance in product pricing.

The Scheme was established in 1968 by a group of the largest UK motor insurers, and has produced results continuously since that time. As part of a comprehensive review, completed in August 1998, the systems, methodologies and outputs of the Scheme were amended so as to reflect modern data collection and analysis techniques, and to enable results to be produced at a lower level of detail and with the flexibility now required by members. In particular, the Scheme now makes extensive use of SAS statistical analysis software, within a PC-environment, and results are increasingly produced in both hard-copy and electronic media. The Scheme management includes a technical unit composed of motor underwriters and actuaries, having responsibility for pricing motor business, which seeks to ensure that the analysis and presentation techniques adopted by the Scheme keep up with market best-practice.

The Scheme currently has 12 members, accounting for between four and five million vehicles insured, or around 25% of the UK private car market. It is able therefore, to base its statistical analyses on data volumes more than twice as great as those of the largest individual insurers. Such size is particularly valuable when examining claims experience differentials by factors having many different levels, e.g. postcode or vehicle model code. Data is collected quarterly, at the individual policy and claim level, and in accordance with data standards agreed by members. This approach allows for flexibility in the analyses undertaken, and the production of up-to-date results, whilst maintaining standards of accuracy and consistency.

Strict confidentiality of data and results is a key characteristic of Scheme operation. Although all member companies gain access to pooled data analyses, no member company is permitted access to the data or results of any other member. Members may, however, with the agreement of others, initiate new analyses designed to investigate aspects of the aggregate claims experience not previously investigated.

The Scheme is financed by members by means of an annual fee set at a level designed to cover the costs of operation only. Each member pays the same fee which, for 1999, is less than £9,000.

The Household Risk Statistics Scheme (HRSS)

The HRSS is an analogous data-pooling arrangement relating to domestic property business. Established in 1975, using the MRSS as its model, the Scheme has also recently been subject to a fundamental review resulting in the modernisation of the techniques adopted for data submission, manipulation and analysis.

Separate results are produced by the scheme for buildings covers and contents covers, with contents business being subdivided according to indemnity or replacement basis, and sum-insured or bedroom rating. The scheme currently has 17 members, contributing data relating to around 3 million covers per year.

- External data – detailed information

  Experian theft model

  The external data provider, Experian, has developed a Motor Theft model. The output of this database is a measure of propensity for vehicles to be stolen and a measure of propensity for theft from a vehicle. The crime data models are designed to predict the expected annual loss rates for each household within a postcode unit from motor vehicle theft. There are four core data sources

  - Experian has commissioned 35,000 interviews, which have been carried out by NOP and MORI. The questions asked consider the frequency and severity of actual theft/vandalism and attempted theft. Other areas covered include the location of the vehicle at the time that the car crime was experienced and whether anyone in the near neighbourhood has suffered burglary/attempted burglary in the past 2 years. Data is also being collected on home ownership and the number of vehicles per household.
  - Neighbourhood level (e.g. postcode unit, enumeration district) demographic and socio-economic data from Experian (e.g. lifestyle surveys, Electoral roll) and official sources (e.g. Census)
  - Detailed motor vehicle statistics from the Driver and Vehicle Licensing Agency (DVLA). This includes the complete make, model and year of registration for all motor vehicles in Great Britain.
  - "Accessibility" measures – household density, road network density and driver distance to major roads and motorways

  The characteristics of the interviewees were then extrapolated by multivariate analysis techniques to postcode unit demographic, socio-economic and accessibility data. Each postcode unit's vulnerability to

vehicle crime is then determined. Ultimately for each full postcode a score is available that will give an estimate of the frequency of motor theft. The propensity for repeat victimisation and vandalism at full postcode level is also be available.

Equcat household theft model

This risk database includes data on domestic burglary from police forces and insurers in the UK. A range of demographic, socio-economic and behavioural factors available at unit postcode level are used to adjust the data. Example factors include:

- unemployment
- number of households with bad debts compared to the UK average
- housing density
- proximity to town centres
- access to road and rail network

- External provider contacts

ISL contact	Experian contact
Adrian Lord	Philip Highland
Intermediary Systems Ltd	Experian
18 Mansell Street	39 Houndsditch
London	London
E1 8AA	EC3A 7DB
Tel 0171 3572312	Tel 0171 623 5551
Fax 0171 3571460	Fax 0171 397 6630

- Information not available at postcode level – detailed information

CUE score

A CUE household score has been developed. This score can be used to predict the claim performance of a customer. The CUE score will indicate the probability of a customer having a claim in the next 12 months and the type of claim including theft risk. The score has been derived from the CUE database of household claims and uses other lifestyle factors in addition to this data The score has been developed at a customer level so does not directly map to postcode level, therefore it is not of additional benefit in assessing postcode risk.

British Crime Survey

The British Crime survey offers a wealth of information on trends in claim levels. Vehicle property damage is included as a specific item in the analysis. The latest survey that was issued was the 1998 British Crime Survey and this covered crimes against people in private households during 1997. It is based on a nationally representative sample of 14,947 people aged 16 and above. Face to face interviews are carried out to assess the level of crime. Of the 16.5 million crimes against adults that the BCS estimates nearly 3.5 million (21%) of these related to vehicle related thefts. Most of these thefts (62%) involved theft from a vehicle. The survey also shows thefts have reduced by 25% over the last 2 years following a stable period, and an increasing theft rate in the 1980's and early 1990's. The data can be analysed by a number of factors such as age of head of household, physical disorder in area and region. The data is not available at individual postcode level so can not be included in the zoning method. The survey can be used as an overall review of zoning results to ensure consistency with an alternative source.

Police Records

Police records are available at individual police force area. There are 43 of these areas however unfortunately these "beat" areas do not match to postcodes and are therefore difficult to incorporate. Police records provide a good measure of well-reported crimes and are an important indicator of police workload although only crimes that are reported are included in police figures. The police provide monthly crime returns and figures are published every six months.

Appendix 2

- **POSTCODE DISTRICT CREDIBILITY**

  Credibility Criterion.

  The postcode district credibility describes the "believability" of the standardised claims experience in a postcode district. The credibility criterion provides a limit, above which the claims experience is deemed fully credible. Postcode districts whose experience lies below this limit are partially credible and a combination of the postcode district results and the postcode area results are used.

  We define:

  f_X - the postcode area standardised theft claim frequency,

  f_{X12} - the postcode district standardised theft claim frequency,

  $\tilde{f}_{X12}$ - the postcode district credible theft claim frequency,

  e_{X12} - the postcode district exposure,

  c_{X12} - the postcode district credibility.

  The credibility assigned to a postcode district depends on the postcode district exposure and the credibility criterion. The credibility criterion is allowed to vary by postcode area so that different criteria apply in low theft claim frequency areas and high theft claim frequency areas. The credibility criterion is defined through the exposure e_X^{40} where $e_X^{40} \cdot f_X = 40$. The derivation of the credibility criterion is given in Section C.

  Postcode District Credible Theft Claim Frequency $\tilde{f}_{X12}$.

  The postcode district credibility is calculated by comparing the postcode district exposure with the credibility criterion:

  - Full credibility, i.e. $c_{X12} = 1$, is assigned to f_{X12} in postcode districts where the exposure e_{X12} exceeds e_X^{40}.
  - For postcode districts such that f_{X12} is not fully credible, the postcode district credibility is defined as

  $$c_{X12} = 1 - \left(1 - \frac{e_{X12}}{e_X^{40}}\right)^2$$

  The postcode district credible frequency is calculated from the postcode area standardised frequency, the postcode district standardised frequency and the postcode district credibility:

$$\tilde{f}_{X12} = c_{X12} f_{X12} + (1 - c_{X12}) f_X$$

Derivation of the Credibility Criterion e_X^{40}.

If a postcode district has claims experience characteristic of the postcode area, the expected number of claims n has a Poisson distribution:

$$P(n) = \frac{n^{\lambda} \exp(-\lambda)}{n!}$$

where the mean number of claims is $\mathrm{E}(n) = \lambda = f_X e_{X12}$ and the variance of the distribution is $\sigma^2 = \mathrm{Var}(n) = \lambda$.

For a probability distribution, it is possible to calculate an exact two-sided $(1-\alpha)$ 100% confidence interval. If L and U are the lower and upper confidence limits respectively, then

$$P(n < L) = \alpha/2 \qquad P(n > U) = 1 - \alpha/2$$

As the district exposure increases ($\lambda > 30$), the Poisson distribution tends to a normal distribution. The lower and upper confidence limits associated with a normal distribution are related simply to the mean ($z_{\alpha/2}$ is a function of α):

$$L = \lambda - z_{\alpha/2}\sigma \qquad U = \lambda + z_{\alpha/2}\sigma$$

The postcode district zoning exercise attempts to group together postcode districts with similar standardised frequencies. This implies that there is an implicit allowed uncertainty in the standardised district theft frequency. This allowed error is defined to be a percentage of the mean i.e. error $\approx \delta\lambda/100$.

A postcode district is deemed to be fully credible when the distributional uncertainty is less than the allowed error i.e.

$$z_{\alpha/2}\sigma < \delta\lambda/100 \qquad \Rightarrow \qquad \lambda = e_{X12} f_X > \left(\frac{100 z_{\alpha/2}}{\delta}\right)^2 = C$$

The credibility limit C is a function of the level of confidence chosen and the allowed error. The exact credibility limit for different levels of confidence and allowed errors are given below.

Level of Confidence (%)	% error 10	20	30
95	384	96	43
90	271	68	30
80	164	41	18
70	107	27	12
60	71	18	8

This table is used to derive the credibility criterion:

- The credible postcode district frequency is to be categorised into broad bands. This implies the error associated with the district frequency can be 20% i.e. $\delta = 20$.
- The two-sided 80% confidence limits are chosen i.e. $z_{\alpha/2} = 1.28$.
- The credibility limit is therefore approximately 40.

The district credibility is based on the **area** frequency. This prevents high frequency postcode districts within a postcode area being more credible than low frequency districts with the same exposure.

- **Simulated Annealing sample code**

ProcedureSA($\delta, MaxIt, T_0, \alpha, T_f$)

Let C be the set of all feasible clusterings,

C,c′ ∈ C be the current and perterbed clusterings, respectively.

$\delta : C \to C$ be a randomized perturbation operator,

$J : C \to \Re^+$ be the internal clustering criterion

T ∈ $\Re^+$ be a "temperature" parameter that controls the "greediness",

U : $\Im^2 \to [0,1]$ be a function that returns a random number between 0 and 1,

MaxIt ∈ $\Im^+$ be the number of iterations of the Metroplis algorithm,

$\alpha \in \Re^+, \alpha < 1$ be an "attenuation" constant for reducing the temperature,

T_0 and T_f be the initial and final temperatures.

T ← T_0

REPEAT

FOR $i \leftarrow 1$ TO MaxIt DO
$c' \leftarrow \delta(c)$
$\Delta \leftarrow J(c') - J(c)$
IF $\Delta < 0$ OR $(e^{-\Delta/T} \geq$ U[0.1]) THEN
    c ← c'
ENDFOR
T ← αT
UNTIL T ≤ T_f

FUNCTION δ (p)

Let n =|Q| be the number of objects to be clustered
L= $\{i \in \{1,...,k\} : \exists\ m\ \{1,...,n\} \ni p_m = i\}$ be the set cluster labels in **p**,
$L^c = \{i \in \{1,...,k\} : i \notin L\ \}$ be the set of cluster labels unused in **p**,
SELECT (range) be a function that returns a random element from the set range,
p,p' ∈ *P* be the original and perturbed partitionings, respectively.

p'←p
$i \leftarrow SELECT(1,...,n)$
REPEAT

$M \leftarrow$ SELECT(0,...|L|)

IF|L| = k OR $m > 0$ THEN
    $p_i \leftarrow$ SELECT(L)
ELSE
$p_i \leftarrow$ SELECT(L^c)
ENDELSE
UNTIL $p_i \neq p_i$
RETURN p'

- **Weighted distance method sample code**

```
/*****************************************************************
*/
/* WPPCeg.SAS - Example of how to do smoothing method I          */
/*****************************************************************
*******/
libname zone 'c:\zone\';
/* Summary of method */
/* In order to make the program run as quickly as possible it is'*/
/* designed as follows:  */
/* The data is read into an array */
/* All the processing is performed in memory with no writing to disk */
/* The answer is written to disk */
/* Please note that it can take some time for the arrays to be created */
/* Program follows */
data zone.smooth;
/* Set the values of the parameters */
APARM = 0.1;
NPARM = 3;
/* Set up the arrays*/
/* The arrays should be the same size as the number of postcode districts */
array x {2761};
array y {2761};
array p {2761} $ 10;
array e {2761};
array r {2761};
array s {2761};
/* Read in the data */
/* Please note that SAS compression must be turned off for this to work */
do i = 1 to 2761;
/* This reads in the ith record from the file */
 set zone.indata point=i;
 x(i) = XCOORD;
 y(i) = YCOORD;
 p(i) = POSTCODE;
 e(i) = EXPY;
 r(i) = RESRISK;
end; /* Data input loop */
/* We now have all the information we need in memory */
```

```
/* The calculations follow */
/* i is the postcode we are smoothing */
do i = 1 to 2761;
/* Reset the running totals to zero */
 tot1=0;
 tot2=0;
/* Now scan through the other postcodes */
 do j = 1 to 2761;
 /* skip the current postcode */
  if j ne i then do;
   /* Calculate the distance */
   d=sqrt((x(i)-x(j))**2 + (y(i)-y(j))**2);
   /* Calculate the things we are interested in */
   tot1 = tot1 + e(j)*r(j)/(d**NPARM);
   tot2 = tot2 + e(j)/(d**NPARM);
  end; /* if j ne i */
end;  /* do j loop */
/* Now calculate the smoothed risk */
s(i) = APARM*r(i) + (1-APARM)*tot1/tot2;
end; /* do i loop */
/* Now output only those fields in which we are interested to a file */
do i = 1 to 2761;
 POSTCODE = p(i);
 SMTHRISK = s(i);
 keep POSTCODE SMTHRISK;
 output;
end; /* output loop */
/* Now stop the processing before it tries to do everything again for the next
record */
stop;
run;
/* End of program */
```

- **Market method**

  One sample quote was chosen based on an average person. This quote was then calculated for all insurers for every postcode district. An average loading was then derived based on the overall average premium for that company. This is based on an un-weighted average of premiums across all UK postcodes. Postcode are then categorised in to groups based on the average loading across all company's for that district.

- An average person was set to a 35 year old male driving a 1994 group 10 car worth £6000. The cover was fully comprehensive with 5 years protected No Claims Discount. The policy had an excess of £100 and the car was insured for the driver only and for social, domestic and pleasure

Appendix 3

- Data preparation method
  The data files were analysed at postcode district level. This enabled us to ensure we could get a statistically significant fit that was not possible at individual postcode level. The files were randomly split into a 70% training file and a 30% validation file. The approach to fitting the models was quite time consuming and involved the calculation of a standardised residual theft risk for each postcode used in the company's sample. The approach to standardisation of the theft scores is outlined below. The zoning methods were then applied to these standardised residual theft frequencies for each postcode. Postcodes were then banded in 20 groups and the model was refitted allowing for the effect of this 20 level geographical category.

- Goodness of fit statistics
  The goodness of fit of the methods was assessed by comparing actual theft frequencies in the 30% test file with expected frequencies calculated from the statistical fit. The test statistic used to assess the best model was the sum of differences squared.

- Standardisation method
  A model for claim frequency was produced using all the factors typically used by the company, but excluding area as a rating factor using standard generalised linear model approach

  The expected number of claims from the above model for each policy is then calculated.

  The residual risk for a postcode district is the actual number of claims divided by the expected number of claims for that postcode district.

  It is these residual risks that are smoothed or clustered in the postcode analysis.