2008 Casualty Actuarial Society Discussion Paper Program

Applying Multivariate Statistical Models

Presented June 15-18, 2008
Fairmont Le Château Frontenac
Québec City, Québec, Canada

2008 CAS Discussion Paper Program
Applying Multivariate Statistical Models

These papers have been prepared in response to a call for papers by the Casualty Actuarial Society to provide discussion material for the 2008 CAS Spring Meeting, June 15-18 at Fairmont Le Château Frontenac in Québec City, Québec, Canada. The CAS Professional Education Policy Committee coordinated this Discussion Paper Program.

TABLE OF CONTENTS
2008 CASUALTY ACTUARIAL SOCIETY DISCUSSION PAPER PROGRAM
Applying Multivariate Statistical Models

The Impact of Rate Regulation on Claims Evidence from Massachusetts Automobile Insurance

Richard A. Derrig and Sharon Tennyson, Ph. D.

Abstract

Rate regulation has a long history in insurance markets. In many states an important goal of regulation is to reduce price variation across purchasers, and specifically to reduce price levels for high-risk purchasers. That feature of rate regulation leads to price cross-subsidies from low-risk purchasers to high-risk purchasers. Consumers who are charged higher prices in order to finance cross-subsidies to high-risks may be less likely to purchase insurance and to reduce participation in insured activities. These adverse selection effects will lead to a higher proportion of high-risk consumers and a higher proportion of insurance purchased by high-risks.[1] In addition, because cross-subsidies reduce the links between insurance risk and insurance prices, all consumers face reduced incentives for loss prevention and safety investments due to moral hazard.

The article tests the hypothesis that insurance price subsidies lead to higher insurance cost growth. To squarely focus on the impact of regulatory price subsidies rather than that of price regulation more generally, the paper makes use of data from the Massachusetts private passenger automobile insurance market. Cross-subsidies were explicitly built into the rate structure through rules that limit rate differentials and differences in rate increases across driver rating categories. Two approaches were taken to study the potential loss cost reaction to the Massachusetts cross-subsidies that began in systematic form in 1977 and continued through 2007. The first approach compared Massachusetts to all other states on demographic, regulatory and liability coverage levels. Loss cost levels that were 44 to 50 percent above the expected level were found for Massachusetts during the 1978-1995 periods when premiums charged were those fixed by the state and included explicit cross subsidies from low risk drivers to high risk drivers. A second approach considered changing cost levels across Massachusetts by studying loss cost changes by town and relating those changes to subsidy providers and subsidy receivers. Subsidy data for 1999-2007, with underlying accident year data for 1993-2004, showed a significant and positive (relative) growth in loss costs for towns that were subsidy receivers in line with the theory of underlying incentives for adverse selection and moral hazard.

Keywords: Auto Insurance, Subsidies, Adverse Selection, Moral Hazard

1. INTRODUCTION

Rate regulation has a long history in insurance markets. In the United States, regulation arose due to concerns about monopoly pricing if insurers were allowed to pool information for rating purposes, coupled with concerns about price instability and insolvencies if they were not. With the growth of large national insurers and advances in information technology, such concerns have eased and insurance markets are widely considered to be workably competitive. Large numbers of firms operate in most markets and rates of return are within competitive norms. As a consequence, price regulation has become confined to selected markets and state approaches vary significantly.

Rate regulation is most commonly employed in health insurance, automobile insurance, and workers compensation insurance, markets in which insurance is mandatory or universal insurance coverage is thought to be desirable. In many states an important goal of regulation is to reduce price variation across purchasers, and specifically to reduce price levels for high-risk purchasers. For example, some state rate regulations place limits on risk classes, restrict price differences across consumers, and restrict insurers' ability to deny coverage to high-risk purchasers to achieve these goals. These features of rate regulation lead to price cross-subsidies from low-risk purchasers to high-risk purchasers.

Although consumers who receive subsidies should be more likely to purchase insurance, thereby promoting one goal of rate regulation, price cross-subsidies may have many unintended consequences in insurance markets. For example, consumers who are charged higher prices in order to finance cross-subsidies to high-risks may be less likely to purchase insurance and to reduce participation in insured activities. These adverse selection effects will lead to a higher proportion of high-risk consumers and a higher proportion of insurance purchased by high-risks.[2] In addition, because cross-subsidies reduce the links between insurance risk and insurance prices, all consumers face reduced incentives for loss prevention and safety investments due to moral hazard (Shavell, 1982).

These combined adverse selection and moral hazard effects of insurance price cross-subsidies lead to efficiency losses in the insurance market, and will lead inevitably to higher insured losses and thus to higher insurance prices in the long run. This article explores the economic importance of this hidden cost of insurance rate-leveling.

The article tests the hypothesis that insurance price subsidies lead to higher insurance cost growth. To squarely focus on the impact of regulatory price subsidies rather than that of price regulation more generally, the paper makes use of data from the Massachusetts private passenger automobile insurance market, a state with unique regulatory characteristics. During the period of study, insurance prices in this market were determined annually through a state hearing process, and the state-determined rate grid formed the basis for pricing by all insurers in the state. Cross-subsidies were explicitly built into the rate structure through rules that limit rate differentials and differences in rate increases across driver rating categories. Market forces that might work to undermine intended

[2] Harrington and Doerpinghaus (1993) provide a useful exposition of these ideas.

cross-subsidies – including the definitions of driver rate classes and the underwriting criteria used by insurers – are also highly regulated by the Commonwealth. Price subsidies to high risk drivers are thus a hallmark of the Massachusetts regulatory system since 1978, providing an ideal venue for the study. Table 1 lists the timeline of major developments in the regulatory environment in the Massachusetts private passenger auto insurance market.[3]

Table 1
Major Regulatory Changes, 1970-2007
Massachusetts Private Passenger Automobile Insurance

Year	Regulation
1971	No-fault auto insurance effective
1975	State rate-setting extended to all auto coverages
1977	Competitive rate-setting allowed
1978	State rate-setting reinstituted
1989	Automobile Insurance Reform Law effective
1991	Insurance Fraud Bureau began operation
1996	Competitive Discounts and Deviations begin at -7.4%
2006	Competitive Discounts and Deviations stabilize at -1.7%
2007	Competitive rate-setting allowed 4/1/08

We conduct two distinct sets of analysis to present evidence on the relationship between price subsidies and insurance cost growth. In the first analysis, annual state-level data on loss costs per insured car for Massachusetts are compared to those in other states during the time period 1972-1998. This analysis uses variation in Massachusetts' regulations over time to demonstrate that cross-subsidies lead to unusually high loss costs in the state. The empirical results show that loss costs are significantly higher in Massachusetts after the cross-subsidy system is introduced (in 1978) but not before; and that the Massachusetts cost differential decreases during the 1990s after the state passed an insurance reform law (1989) and established an insurance fraud bureau (1991).[4] These reforms led to lower loss costs and induced insurers to offer some drivers premiums below those set by the state, reducing effective cross-subsidies. Table 2 reports the average annual percentage discounts offered by insurers, subject to prior approval, beginning in 1996.

[3] State set Massachusetts PPA insurance rates are expected to come to an end after 30 years as of April 1, 2008, to be replaced by a form of managed competition (Burnes, 2007).
[4] See Weisberg and Derrig (1992) for the effects of the tort reform, Derrig (1997) for the effects of the auto property reform, and Derrig, et al. (2006) for the more recent effects of the Insurance Fraud Bureau.

Table 2
Massachusetts Private Passenger Automobile
Historical Summary of Industry Discounts/Deviations

Year	Average Discount		Annual Change in Discount
1996	-7.4%		
1997	-9.2%		-1.8%
1998	-9.2%		+0.0%
1999	-6.5%		+2.7%
2000	-5.5%		+1.0%
2001	-3.0%		+2.5%
2002	-2.2%		+0.8%
2003	-1.9%		+0.3%
2004	-1.7%		+0.2%
2005	-1.8%		-0.1%
2006	-1.7%		+0.1%
2007	-1.7%	est'd	+0.0%

Source: Automobile Insurers Bureau of Massachusetts

In order to test the hypothesis that price subsidies contribute to cost growth, we undertake a second analysis using data by Massachusetts towns for the time period 1999-2007. This analysis makes use of the variation in subsidies across towns to identify the effects of subsidies on loss cost growth. The empirical results show that loss cost growth is significantly higher among towns in which the average driver receives a premium subsidy.

Section 2 describes the regulated Massachusetts auto insurance system in more detail and documents the extent of price subsidies. Section 3 develops the theoretical arguments regarding incentive effects of regulation and discusses prior research on the impact of insurance regulation and premium subsidies. Section 4 presents our analysis of state-level average annual loss costs for the time period 1972-1998. Section 5 introduces the Massachusetts town level data and presents results of analysis of those data. The final section of the paper provides conclusions and discusses the implications of our findings.

2. MASSACHUSETTS AUTOMOBILE INSURANCE REGULATION

In Massachusetts, regulated automobile insurance rates are determined annually by the state insurance commissioner as the outcome of a public hearing process. The rates determined through the hearing process must be charged by all firms writing in the state – irrespective of differences in operating costs or loss experience – unless an insurer obtains approval from the insurance commissioner to charge lower rates.[5] Massachusetts is the only state that used this form of rate regulation for automobile insurance, until quite recently.[6]

Massachusetts has regulated automobile insurance rates since 1927, but the regulatory features of primary interest to this paper took shape in the late 1970s, after a brief experiment with a more competitive system in 1977. [7] In that year legislation introduced file-and-use rate regulation, which allowed insurers to set their own rates subject to light regulatory review. The new system led to dramatic increases in premiums and reduced insurance availability for some drivers, producing a record number of consumer complaints to the Division of Insurance (Stone, 1977). In response, the state returned to the state-made rates and new legislation and regulatory decisions imposed even further state controls over pricing.[8] The legislature ordered rebates on 1977 premiums for many policyholders, and passed legislation that prohibited premium surcharges to policyholders insured through the residual market facility. In determining the 1978 rates, the insurance commissioner rejected the use of age, gender and marital status as risk classification variables, and required all insurers to utilize the same classification variables (Stone, 1978).

The resulting regulations mandate a common set of rating territories and driver rating classes for all insurers. Rating territories are determined by town and the assignment of towns to territories is determined through a periodic hearing process. Only nine driver rating classes are allowed, with

[5]Historically, such deviations were not common; however, most insurers sought significant rate deviations for selected groups of drivers in 1996-2004 as shown in Table 2.

[6] This regulatory system has recently been overturned (see Burnes, 2007) and the state will employ a file-and-use regulation called "managed competition," effective April 2008. The recently adopted change of December 24, 2006, to the California DOI regulation on rating class differentials (10 CCRs2632.8, Factor Weights) provides for indirect subsidies for high-risk towns that will differ by insurer through the suppression of the true cost of location by lowering the "importance" of territory in the final rate differentials.

[7] See Derrig (1993), Yelen (1993), Rottenberg (1989), and Tennyson, Weiss, and Regan (2002) for more discussion of the history and process of Massachusetts regulation. Details on the current regulations are available at the Web site of the state of Massachusetts http://www.mass.gov/.

[8] The state still operates under the law that allows file-and-use regulation. State-set rates were reintroduced due to a provision in the law that allows the insurance commissioner to hold an annual hearing to determine whether competition

drivers classified only by driving experience, drivers' training, and use of car. Experienced drivers are defined as those with more than six years of driving experience. There are an additional six classes consisting of four types of inexperienced drivers, senior citizens, and business-use drivers. Age, gender, and marital status are specifically prohibited from use as rating variables (Mass GL c. 175E, s4(b)). The restrictions on rating classes produce a far coarser rate matrix than used in other states' automobile insurance markets, leading to cross-subsidies in rates across drivers.[9]

Additional cross-subsidies are built into the rates through a systematic leveling process known as *tempering* and *capping*. Tempering restricts the differences in average rate levels across the class-territory rating cells. Capping restricts the average annual increase in rates for any rating cell to be no more than a pre-specified percentage above the average statewide rate increase. Capping thus restricts the change in average rate levels over time for each rating category, reinforcing the cross-subsidies generated by tempering.

A final set of inter-class rate constraints is applied within each territory to assure that the lower risk classes do not pay more than a given percentage of the rate paid by higher risk classes. For example, traditionally the experienced driver rate is set at no greater than 95 percent of the lowest inexperienced driver rate; the business use rate is at least as great as the experienced driver rate, and so forth.[10] The application of these constraints introduces additional, sometimes substantial, cross-subsidies across drivers.[11]

This complex set of restrictions produces significant variation in premium charges relative to those based on costs alone. Some drivers receive substantial premium subsidies, with the remaining drivers paying relatively smaller premium increases. Table 3 summarizes the direction and extent of premium subsidies received and premium subsidies paid, using data for 2004.[12] The table reports the average premium, average subsidy (or surcharge) amount and the percent of class-territory rating

---

is feasible, and to impose state-set rates if it is not. In every year since 1978 competition has been found not to be viable and state-set rates have been imposed. File and use is expected to return for rates effective April 1, 2008 (Burnes, 2007).

[9] Finger, (2006), notes that the standard industry classification plan contains 217 driver classes.

[10] See Docket R98-42, AIB Filing on 1999 rates, August 1998.

[11] For 2007 Boston Compulsory Rates, the inter-class constraints generated half of the average subsidy of 17.6% (AIB, Actuarial Notice 07-1, p.2).

[12]The premiums paid by each individual driver vary from the class-territory average due to experience rating adjustments based on accident experience, the type of car driven, and other factors. Experience rating adjustments are applied through the revenue neutral Safe Driver Insurance Plan (SDIP), which allows for discounts and surcharges to bodily injury liability (BIL), property damage liability (PDL), and collision rates, based upon one's recent driving record. The state also allows premium discounts for anti-theft devices, airbags, low mileage, multiple cars, or use of public transit. These adjustments are applied as percentage changes to the premium, but are relatively modest. The state determines

cells that are subsidized for compulsory insurance coverage. Boston drivers and inexperienced drivers are those most likely to receive a subsidy, but some proportion of drivers in every rating cell receive a subsidy.[13]

Table 3
Direction of Subsidies by Driver Class and Territory Compulsory Insurance Coverage
2004

		Experienced Classes	Inexperienced Classes	Business Classes
Non-Boston Territories	**Average Premium**	$527.15	$1,220.54	$500.67
	Average Subsidy	-$26.00	$138.29	-$46.43
	Cells Subsidized (%)	12.50%	42.71%	6.25%
Boston Territories	**Average Premium**	$813.33	$1,434.04	$751.98
	Average Subsidy	$253.77	$520.09	$32.30
	Cells Subsidized (%)	64.65%	72.73%	36.36%

Source: Authors' calculations using data from Actuarial Notice 04-1, Automobile Insurers Bureau of Massachusetts, 2004. Compulsory coverages are Bodily Injury Liability (20/40), Personal Injury Protection (8,000), Property Damage Liability (5,000) and Uninsured Motorist (20/40)

Within Boston, 72.73% of inexperienced driver rating cells, 64.65% of experienced driver rating cells, and 36.36% of business cells receive a subsidy. Outside of Boston, 42.71% of inexperienced driver rating cells receive a subsidy. In contrast, only 12.5% of experience driver cells and 6.25% of business driver cells outside of Boston receive a subsidy. Both experienced and inexperienced Boston drivers receive substantial premium subsidies (averaging $253.77 and $520.09 respectively), while Boston business classes receive only a modest subsidy ($32.30).

experience rating adjustments and other adjustments, which are subject to prior approval regulation. Premium surcharges to drivers insured through the state's residual market facility are prohibited.

[13] The number of drivers receiving subsidies or paying surcharges differs from the number of insured vehicles in the rating cells, due to cross-subsidies across drivers within class-territory rating cells.

It is natural to ask how the state can sustain an auto insurance market under the regulations described here. Several additional regulations promote the continued supply of insurance in the market. Access to insurance for high-risk drivers is protected by restrictions on insurers' ability to refuse insurance or cancel a policy. There are also strong restrictions placed on insurers' rights to exit the automobile insurance market. An insurer wishing to withdraw from the market must receive regulatory approval and must pay substantial penalties in order to exit the market (Yelen, 1993). Finally, insurance demand is bolstered by a strong compulsory insurance law.

Nonetheless, previous studies have documented a number of distortions to the Massachusetts' market that arise due to regulation (Rottenberg, 1989; Derrig, 1993; Tennyson, 1997; Tennyson, Weiss, and Regan, 2002). Most notably, these studies have found that the supply side characteristics of the Massachusetts automobile insurance market differ from those in comparable state markets, with far fewer insurance providers[14] and fewer national insurers in the state. Residual market size is also greater in Massachusetts, providing another indicator of lack of insurance availability. And, movement toward lower cost distribution systems has been much slower there than in other automobile insurance markets.

3. PREMIUM SUBSIDIES AND DRIVER INCENTIVES

3.1 Economic Theory

If insurance premiums reflect the expected marginal costs of coverage, consumers have appropriate information on which to base their decisions about insurance purchase and also their decisions to purchase or drive a car or both (Harrington and Doerpinghaus, 1993). However, consumers who receive premium subsidies face less than the true expected marginal cost of their decisions with respect to insuring and driving decisions and will be more likely to own a car, to drive and to purchase more insurance. Consumers who pay premium overcharges will have the opposite response, tending to be less likely to drive and to purchase less insurance.

Drivers also make choices about the amounts and types of driving and other actions that are correlated with expected loss costs.[15] These choices will also be distorted by cross-subsidies. Simply put, the Massachusetts rate structure and the cross-subsidies built into the rates reduce the penalties

[14] A total of 19 insurers had (non-specialty) Massachusetts private passenger automobile insurance written premium in 2006.

[15] See Brockett and Golden, (2007), for a discussion of risky behavior by drivers and its relation to credit scores as a proxy measurement of that risk.

for risk-taking. The rate categories are few and will, of necessity, permit cross-subsidies of identifiable subgroups.[16] Policies cannot be cancelled based on loss or accident experience. Higher premium charges due to higher accident costs of any one driver will be partly shared by the driver under a Safe Driver Insurance Plan,[17] partly across all members of the class-territory rating cell, and partly by all subsidy-paying drivers in the remainder of the Commonwealth. Those subsidies will dampen individual incentives to reduce costs. Risky choices may move a driver into the residual market but at no premium differential. These regulatory restrictions will increase the risky choices of all drivers. As a result, average loss costs are predicted to be higher under the Massachusetts regulatory system than otherwise.

Perhaps as important, there may be an additional upward shift in *insured* claims and loss costs due to the greater incentives for claims filing introduced by the rate structure. Consumers consider the marginal costs and the marginal benefits when deciding whether to file a claim. The restrictions on policy cancellation, the relatively small premium penalties for high losses, and the tempering of rate increases across time imply that the future adverse consequences of filing a claim or of filing many claims are lessened. This will increase the propensity of consumers to file claims. The same arguments apply not only to legitimate claims, but also to fraudulent or exaggerated claims.[18] Under Massachusetts law insurers may cancel a policy due to fraud, but the fraud must be proven which may be costly and difficult. These forces underlie the authors' prediction that there will be a greater incidence of fraudulent claiming in Massachusetts.

The prediction of fraudulent claim behavior was observed soon after the 1988 Reform Law provision that raised the monetary threshold to file a tort claim from $500 to $2,000 in claimed medical bills. Weisberg and Derrig (1992) document the increase in numbers and intensity of medical provider visits with the result being a much larger-than-anticipated 1989 proportion of auto injury claims with medical bills in excess of $2,000, the new tort threshold. More recently, Derrig, et

[16] The largest such classification is Experienced Driver, representing about 89% of the 2005 exposure in each rating territory and consisting of all non-business drivers with more than 6 years licensed. By law, the subgroup of drivers 65 and older pay 75% of the rate that the remaining risks pay who will range in age from about 22-64 years.

[17] The Massachusetts Safe Driver Insurance Plan sets forth differentials within each rate class based upon at-fault accidents and traffic violations.

[18] In addition, the no-fault compensation system increases the marginal benefits of building up bodily injury claims. First-party insurance for automobile-related injuries is mandatory in Massachusetts under PIP coverage, which pays a maximum of $8,000 in losses, which can be offset by up to $6,000 through private health insurance. However, injured parties may be eligible for compensation under bodily injury liability (BIL) in addition to PIP if their medical expenses exceed $2,000. BIL claimants may be compensated for medical and wage losses, plus "pain and suffering." This provides significant incentives for fraudulent BIL claiming, and the medical expense threshold for BIL claiming provides significant incentives for PIP claims build-up (Weisberg and Derrig, 1991, 1992; Cummins and Tennyson, 1992, 1996).

al. (2007) discuss auto injury claims with the appearance of fraud and/or buildup[19] countrywide through the Insurance Research Council (IRC) Study of 2002 Claims and in Massachusetts through the developments in the town of Lawrence. In Lawrence, Insurance Fraud Bureau (IFB) activities reduced injury claims per 100 accidents (PDL claims) from 141 in 2002 to 60 in 2004 and claim payments from $48.6 million in 2002 to $19.8 million in 2004. Granted, Lawrence was an exceptional case identified as far back as 1991 (Weisberg and Derrig, 1991), but reductions on a lesser scale in other towns have been realized by IFB efforts since 2004.

The combined effect of the incentive distortions from premium subsidies is to increase the average cost of insurance coverage relative to that under a cost-based system of rates. The extent to which this increases average costs will depend upon the sensitivity of driving, insuring and claiming decisions in relation to insurance prices, and the extent to which cross-subsidies change prices.

3.2 Empirical Evidence

The empirical literature on insurance price cross-subsidies has mainly examined their impact on insurance purchase decisions, with particular focus on adverse selection. Cross-state studies in health insurance find small or insignificant effects on insurance coverage due to state regulations that impose price cross-subsidies (Simon, 2005; Buchmueller and DiNardo, 2002; Monheit, Steinberg, and Schore, 2003). Case studies of New Jersey's community rated program for individual health insurance find more mixed evidence (Swartz and Garnick, 1999; Monheit, Cantor, Koller and Fox, 2004). Studies in automobile insurance find that insurance purchase decisions are sensitive to price cross-subsidies, with low risk drivers reducing insurance purchase (Dahlby, 1983, 1992).

There is a growing empirical literature relating loss cost growth to insurance rate regulation more generally, but rarely has the focus been on cross-subsidies. Using data on state-level average loss costs, studies in both workers compensation insurance and automobile insurance find evidence that rate regulation is associated with higher loss costs.[20] Using data by individual rating classes from eight states, Danzon and Harrington (2001) find that workers compensation loss cost growth is higher for classes with a larger proportion of risks insured in the residual market.

[19] "Build-up" is the term of art for excessive treatment for the injury (if any) sustained in an auto accident with treatment usually provided by a chiropractor or physical therapist.

[20] Barkume and Ruser, (2001), and Harrington and Danzon, (2000) analyze state-level data on workers compensation losses; Weiss, Tennyson, and Regan, (2007) present a similar analysis of automobile insurance losses.

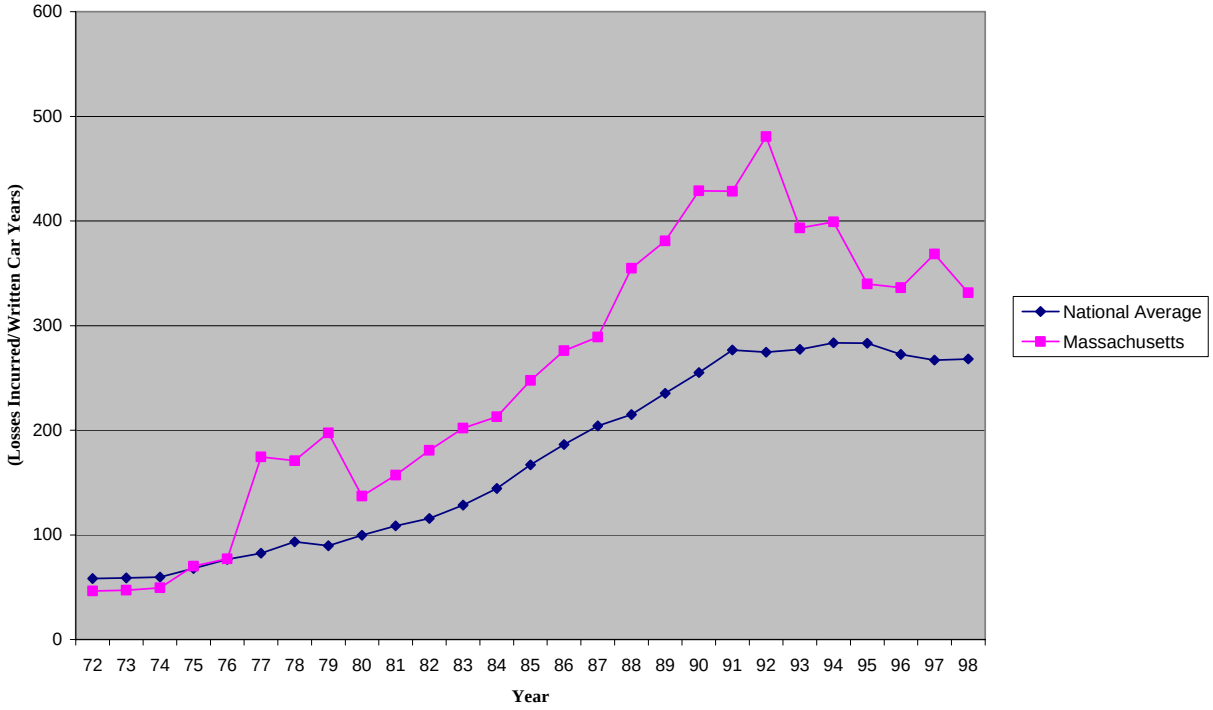
The current study uses data from Massachusetts automobile insurance to examine the relationship between regulatory price cross-subsidies and loss costs. The study adds to existing research by providing a more direct analysis of the relationship between regulatory premium subsidies and loss cost growth than has been possible previously. Taking advantage of the explicit and transparent system of premium cross-subsidies under Massachusetts regulations, we estimate the effects of the subsidies by first comparing loss cost levels in Massachusetts to those in other states before and after the subsidies are introduced; and by comparing loss costs and risk characteristics for Massachusetts towns that receive subsidies to those that do not.

4. REGULATION AND LOSS COST GROWTH IN MASSACHUSETTS

State-level data are used to compare Massachusetts loss costs to those that would be predicted in the absence of the regulations. The data encompass the time period before regulation, the period during the enforcement of the most stringent regulations, and a period during which regulatory stringency was relaxed.

Figure 1 plots Massachusetts annual loss costs over the time period 1972-1998, along with the average losses for all other states. The loss costs reported in the figure are statewide average liability losses per insured car, constructed as total liability losses incurred divided by number of written car years. The liability losses are those reported in the insurers' annual statements and include those paid under all automobile bodily injury insurance, including bodily injury liability, first party personal injury protection (PIP) in no-fault states, medical payments, and uninsured/underinsured motorist coverages; along with losses incurred under property damage liability coverage. Losses not included in the liability data series are first-party property damage losses paid under collision and comprehensive insurance coverages. Liability losses are used in this comparison because liability coverages are compulsory in most states. Thus there is anticipated to be less variation due to differences in insurance purchase in automobile liability insurance losses than in total automobile insurance losses.

Figure 1: Average Annual Liability Loss Costs per Insured Car
Massachusetts vs National Average



The figure shows a striking difference between Massachusetts loss costs and those of other states after 1977, and virtually no difference between Massachusetts and other states prior to 1977.[21] The loss cost gap appears just at the time the more stringent regulations are enacted in Massachusetts, and loss cost growth appears to be greater in Massachusetts throughout the 1980s. The distance between Massachusetts and other states is reduced in the 1990s, at a time when the state introduced antifraud measures and when regulatory stringency was relaxed sufficiently that insurers offered discounted rates to select low-risk drivers (Table 2).

Of course, other features of the state may account for the difference in insured loss costs. Massachusetts is a largely urban state with high traffic density, which may contribute to higher rates of accidents. The state population also has high per capita income and high cost of living, which may contribute to higher costs per claim. Moreover, the increase in relative loss costs in Massachusetts in the early 1990s coincides with an increase in the maximum first-party PIP compensation (from $2,000 to $8,000) and compulsory bodily injury liability limits (from

$10,000/$20,000 to $20,000/$40,000 per person and per accident, respectively). This pattern reinforces the idea that factors such as coverage amounts must be controlled for when comparing loss costs across states.[22]

4.1 Regression Analysis

To further explore whether Massachusetts loss costs are higher than expected we estimate a regression model using 1972-1998 annual loss cost data from 50 states, including control variables for time-varying state characteristics as well as state fixed effects. The methodology for modeling the effect of regulation is to examine whether the difference between Massachusetts' loss costs and those of other states is significantly different before and after the regulations that introduced substantial price cross subsidies in Massachusetts.

Because there are many other determinants of loss costs at the state aggregate level, the empirical model includes state characteristics as control variables. The empirical model takes the basic form:

$$L_{st} = \beta_0 + \beta_1 CSYears_t + \beta_2 MA_s CSYears_t + \beta_3 StateRegs_{st} + \beta_4 StateRegs_{st} CSYears_t + \delta' X_{st} + a_s + \varepsilon_{st} \quad (1.0)$$

where subscript s denotes state and t denotes year, and L_{st} is the logarithm of statewide average liability losses per car. The variable $CSYears_t$ is an indicator variable set equal to 1 in the years of cross-subsidy regulation in Massachusetts. The impact of subsidies in Massachusetts is identified by including the interaction of a Massachusetts dummy variable (MA_s) with the cross-subsidy years variable: $MA_s CSYears_t$. If premium cross-subsidies in Massachusetts are a significant determinant of higher loss costs in the state, then – after controlling for other determinants of losses – we expect to find a significant increase in loss cost differences in Massachusetts during the regulatory periods denoted by $CSYears_t$, and thus a significantly positive value for the coefficient β_2. Including the dummy variable $CSYears_t$ for the years in which Massachusetts cross-subsidy regulation occurs means that we are testing for differences in Massachusetts loss costs relative to any nationwide

[21] Loss costs for 1977 were incurred under a mixed regulatory system that began as open competition then collapsed under an average 14.5% increase and ended with a rate rollback for 1977 and return to fix and establish rates for 1978 (Burnes, 2007).

[22] Weisberg and Derrig (1992) argue that this change also increased incentives for claims exaggeration, due to the need to exceed the threshold in order to become eligible to file a bodily injury liability claim. Their analysis shows that PIP claim amounts cluster at the value of the tort threshold.

effects on auto insurance loss costs that occur in those years. We first estimate models in which the regulated period is defined to be all years 1978-1998; we also explore models in which the regulated period is defined as 1978-1995.

$StateRegs_{st}$ is a vector of variables reflecting each state's legal and regulatory environment for auto insurance in state s and year t; included are indicators of a state's use of rate regulation and no-fault auto insurance. Massachusetts has both rate regulation and no-fault throughout the sample period, and so we include these variables and interact them with $CSYears_t$. This specification means that the interaction term $MA_sCSYears_t$ distinguishes the effects of Massachusetts cross-subsidy regulations from these more general regulatory features.

The variable denoted X_{st} is a vector of other time-varying state characteristics. Other state characteristics included in the model are traffic density, defined as total vehicle miles driven divided by total miles of roadway in the state; per capita income in the state; and the statewide average cost per day of hospital stay. All three variables are entered in the model in log form. Higher traffic density should be associated with higher accident rates, and thus is expected to be positively related to loss costs. Both per capita income and hospital costs will affect the costs associated with accidents, holding other charactaristics that affect accident severity constant. Higher per capita income may also be positively associated with insurance purchase amounts, which will affect loss payments. Thus both are expected to be positively related to loss costs.

We also recognize that differences in loss costs across states and time will be affected by differences in insurance purchase amounts. For example, the average insurance coverage limits may vary across states and years, and states with no-fault insurance may provide different coverage limits for compulsory (or optional) first party PIP limits. To partially control for these differences, our models include the minimum required coverage limits (if any) for BIL and PDL coverages in each state and year, and the maximum PIP limits in no-fault (and add-on) states.

The term a_s represents a state-fixed effect, and the term ε_{st} is a stochastic error term. Including state-fixed effects implies that we are testing for differences in Massachusetts' loss costs under the cross-subsidy regulation relative to average loss costs in the state over the sample period. In estimating the model, standard errors are corrected to allow for arbitrary forms of heteroskedasticity and for clustering by state. Clustering takes into account the fact that the regression errors are likely to be correlated within each state across years (Bertrand, Duflo, and Mellainathan, 2004).

Table 4 reports summary statistics of the variables included in the state regression models. The table reports the mean and standard deviation of each variable for the full set of states and for Massachusetts alone. The data confirm the conclusion from Figure 1 that average loss costs in Massachusetts are higher than the national average. However, the data also reveal that traffic density, per capita income, and costs of medical care – factors that could contribute to automobile loss costs – are also higher in Massachusetts. As mentioned previously, Massachusetts is a no-fault state and has a compulsory insurance law; however, Massachusetts' mandatory coverage limits were lower than the national average.

Table 4
Summary Statistics
Annual State-Level Data, 50 States
1972-1998

	All Other States		Massachusetts		
Variable	**Mean**	**S.D.**	**Mean**	**S.D.**	
Losses per Insured Car	176.73	111.30	251.06	132.41	**
Traffic Density	0.53	0.40	1.21	0.21	**
Average Cost of Hospitalization	492.72	326.45	618.04	387.05	
Per Capita Income	13,686	6,815	16,728	8,866	
Rate Regulation Dummy	0.60	0.49	1.00	0.00	**
No-fault Dummy	0.27	0.44	1.00	0.00	**
Add-on Dummy	0.19	0.39	0.00	0.00	**
Person Minimum Limit (000)	18.43	6.80	12.59	4.25	**
Property Minimum Limit (000)	9.56	5.19	5.00	0.00	**
PIP Coverage Limit	13,078	45,667	4,000	2,882	**
Add-on Coverage Limit	802.96	2565.85	0.00	0.00	**

Source: Authors' calculations from state-level data. *** indicates Massachusetts mean is significantly different from other-states' average at the 5 percent confidence level.

Table 5
Regression Analysis of Statewide Annual Average Liability Losses per Car
1972-1998

Dependent Variable = Ln(Liability Losses/Written Car Years)

	All Years				1977-1979 Omitted			
Explanatory Variable	**CSYears 78-98**		**CSYears 78-95**		**CSYears 78-98**		**CSYears 78-95**	
MA x CS Years	0.3781	***	0.3663	***	0.4874	***	0.4051	***
	0.1543		0.1101		0.0579		0.0519	
Reg x CS Years	0.1097	***	0.0418		0.1311		0.0547	
	0.0338		0.0260		0.0820		0.0464	
No-fault x CS Years	0.0114		0.0046		0.0038		0.0029	
	0.0316		0.0245		0.0682		0.0431	
CS Years Dummy	-0.0807	***	0.0523	***	-0.0595		0.0657	***
	0.0291		0.0181		0.0501		0.0249	
Ln(Traffic Density)	0.2017	***	0.2663	***	0.1566		0.2170	*
	0.0616		0.0636		0.0099		0.1191	
Ln(Hospitalization Cost)	0.2927	***	0.2557	***	0.2588	***	0.2116	***
	0.0786		0.0647		0.0802		0.0681	
Ln(Per Capita Income)	0.4876	***	0.4809	***	0.5165	***	0.5416	***
	0.1057		0.0876		0.1149		0.1057	
Rate Regulation Dummy	-0.0889	***	-0.0206		-0.1115	*	-0.0315	
	0.0268		0.0234		0.0614		0.0401	
No-fault Dummy	-0.0179		-0.0421		-0.0190		-0.0501	
	0.0303		0.0270		0.0380		0.0396	
Add-on Dummy	0.0504		0.0319		0.0320		0.0122	
	0.0499		0.0498		0.0931		0.0966	
Person Minimum Limit	0.0082	***	0.0087	***	0.0082	**	0.0086	**
	0.0021		0.0020		0.0034		0.0034	
Property Minimum Limit	0.0089	***	0.0091	***	0.0115	*	0.0112	*
	0.0027		0.0027		0.0063		0.0065	
PIP Maximum	5.0E-07		7.1E-07	**	6.10E-07		8.1E-07	*
	3.3E-07		3.2E-07		4.70E-07		4.7E-07	
Add-on Maximum	2.4E-05	***	2.4E-05	***	2.50E-05		2.5E-05	
	6.7E-06		6.7E-06		1.90E-05		2.0E-05	
Intercept	-1.3720	**	-1.1359	**	-1.5147	**	-1.5065	*
	0.5946		0.5149		0.7480		0.8103	
State Fixed Effects	Yes		Yes		Yes		Yes	
R-squared	0.8742		0.8769		0.8898		0.8936	
N	1334		1334		1190		1190	

Standard errors appear below the coefficient estimates and are adjusted to allow for arbitrary forms of heteroskedasticity and arbitrary correlation across years within a state. *** indicates statistical significance at the 1% confidence level; ** at the 5% confidence level; and * at the 10% confidence level; all are two-sided tests.

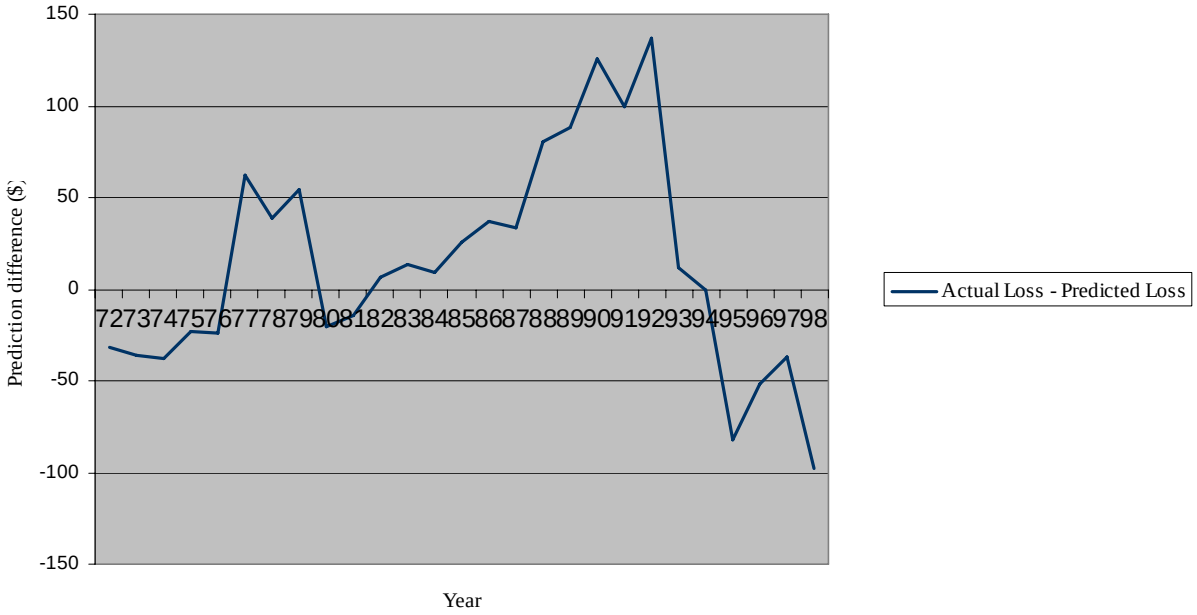
4.2 Estimation Results

The estimation results are shown in Table 5 The first two columns of the table report the estimates obtained using all years of data 1972-1998, and columns three and four report the estimates when the years 1977-1979 are omitted from the dataset (due to concerns that these years are outliers for Massachusetts). Columns 1 and 3 show the estimates when the cross-subsidy years are defined as the entire sample period after regulatory enactment (1978-1998), and columns 2 and 4 show the estimates when the cross-subsidy years are defined as 1978-1995 due to the insurer discounting below state-set rates that began in 1996.

The estimates demonstrate that Massachusetts' loss costs were significantly higher than expected during the period of cross-subsidy regulation, based on other characteristics of the state insurance market. All four specifications show $MA_sCSYears_\tau$ significant and positive at the 1% confidence level. The two specifications that define the cross-subsidy era as ending in 1995 show more consistent results than those that define the era through 1998. The model estimates using all years of data show about a 44 percent (exp (0.3663)) increase in liability loss costs over the levels experienced in the overall U.S. market with the same demographics and liability coverages. The model estimates about a 50 percent increase when 1977-1979 are omitted.

An alternative estimation approach is to exclude Massachusetts from the sample to obtain the beta coefficients of the model for the 49 other states, and use those beta coefficients combined with Massachusetts data to predict Massachusetts loss costs for each year. Comparing actual Massachusetts loss costs with predicted loss costs from this model then yields the estimated excess loss costs in Massachusetts for each year. Using this approach we find similar results, with loss costs in Massachusetts higher than predicted in the years of cross-subsidy regulation. Figure 2 shows the plot of actual minus predicted loss costs in Massachusetts for each year. The estimated average excess loss cost per year over the years 1978-1995 is 32.14 percent using this methodology.

Figure 2
Actual Losses Incurred vs Predicted Losses from Regression Model



5. RELATING PREMIUM SUBSIDIES TO LOSS COST GROWTH

The state-level data provide evidence that loss costs are higher in Massachusetts under cross-subsidy regulations. More detailed data on insurance costs and premiums by Massachusetts town are used to link these trends more specifically to the incentive effects that arise from premium subsidies. Economic theory predicts that premium subsidies distort incentives in a variety of ways that lead to greater interest in insurance among consumers in subsidized groups, and to more rapid growth in loss costs among subsidized groups. This prediction can be tested with the data available.

5.1 Data by Massachusetts Town

The data on insured loss costs and insured driver risk characteristics by town are obtained from documentation of the biennial Massachusetts regulatory hearings for territory determination. These data include actuarial estimates of the pure premiums and aggregate risk factors for each town. The available data used in our analysis span the time period 1999-2007 for territory determination filings, at two-year intervals. In each case, the filings include four prior accident years of loss data. For example, towns are grouped into rating territories for policies issued as of 4/1/2007 based upon

data reported for accident years 2001-2004, each as of 24 months' development. Overall, five datasets are available for this study, reflecting loss experience for the years 1993-2004.

The analysis of town assignments to rating territories has two major elements. First, the relative loss potential of each town is estimated. Second, towns having similar estimates of loss potential are grouped into territories (AIB, 2006). Specifically,

> [The] estimation of each town's loss potential begins with the actual insurance experience (vehicle exposures, claim counts, loss dollars) of each town. The towns' loss cost experiences are dissected into claim frequency and claim severity components and the two compontents are analyzed separately. This information alone is not sufficient, however, since less than complete credibility can be attributed to the actual experiences of the towns. The partially credible actual data must be supplemented by additional information or judgments or both.

In the analysis of town claim frequencies, a mathematical model of frequency potential by town is constructed using data related to the four-year insurance experiences of the towns, and is used to supplement less than fully credible actual town frequency data. The model parameters and quantification of town credibilities are based on an analysis of patterns in loss experience across towns and years (Conger, 1988).

In the analysis of town claim severities, partially credible actual town average cost data are supplemented by data from larger geographic regions (countywide and statewide data). As in the claim frequency analysis, the quantification of town credibilities is based on an analysis of patterns in loss experience across towns, counties, and years.

The result of the claim frequency and severity analysis for each town is used to calculate a loss cost (pure premium) index value for each town as a simple product. The index value expresses the town's loss potential relative to the statewide average. For example, a town index of 1.25 indicates that the per vehicle insurance loss costs for a typical driver in the town are expected to be 25 percent greater than the statewide average.

The town index procedure has been in place with few changes since the early 1980s. The initial credibility procedures were developed by William DuMouchel, a statistician at MIT (DuMouchel, 1984). As a result, the town index for each of the five coverage groups (BIL, PIP, PDL, collision, and comprehensive, all at compulsory or standard deductible levels) is a best Bayesian estimate of

the combined four accident year data for each town, normalized for other rating variables, underlying each new territory definition year. We use these town data indices in analyzing the effect of subsidies on the realized accident years 1993-2004 that underlie the territory definition years of 1999-2007.

Some of the town-to-town variation in pure premiums may be captured already by other rating variables. For example, a town with a heavy concentration of inexperienced operators will tend to exhibit a high pure premium, but should not necessarily be put in a high-rated territory, since the classification pricing scheme will already account for this high pure premium (because inexperienced operators are charged higher than average experienced operator rates in each territory). Therefore, the procedure removes from each town's pure premium index the effects of the mix of insured drivers by driver classification as measured by the average class rating factor (ACRF). The ACRF is a town exposure weighted average of relative pure premium by class compared to statewide. The resulting town net pure premium indices are rebalanced to unity within each coverage and provide for a normalized index that measures comparable loss costs by town.

Table 6 provides a comparison of towns that are indicated to receive subsidies and those that are indicated to pay subsidies in the territory definitions for 2005. The table shows that only 19 of the 360 towns in Massachusetts receive a subsidy. On average, these towns have higher pure premium index values (by design). The table also shows that these towns tend to be larger and more densely trafficked, and have a different mix of insured drivers than the subsidy paying towns.

Table 6
Massachusetts Town Data—Characteristics of Towns Receiving Subsidy in 2005

	Subsidy-Receiving (N=19)		Subsidy-Paying (N=341)		
	Mean	**S.D.**	**Mean**	**S.D.**	
Pure Premium Index	2.0262	0.5248	0.7267	0.1733	**
Insured Exposures (PDL)	30,346.00	20,889.00	9,969.00	9,933.00	**
Traffic Density	210.10	59.39	93.56	60.38	**
BIL ACRF	1.1150	0.1034	0.9804	0.0330	**
PDL ACRF	1.0538	0.0581	0.9936	0.0255	**

Source: Authors' calculations based on data from AIB. Traffic density = exposures per mile of road. ** indicates means for subsidized and unsubsidized towns are significantly different at the 5 percent confidence level.

5.2 Cost Growth by Town

To provide a preliminary look at the empirical relationships, Figures 3 and 4 compare trends in pure premium growth in the underlying accident year data for high-cost and low-cost towns over the territory definition years 1999-2007 for BIL and PDL insurance coverages, respectively. The data are grouped into high-cost and low-cost with classifications based on each town's overall pure premium index in 1999, using a value of 1.20 or greater to indicate high cost and a value of 1.00 or lower to indicate low cost. Only towns above (or below) the cut-off values in every year of the sample period are included in the figures. The averages for each group of towns are weighted by the number of insured exposures in each town. The data are normalized by setting the 1999 pure premium index for each coverage and each group of towns equal to 1.00, to facilitate comparisons in the average annual growth rates over time.

The figures show a clear upward trend in pure premiums among high-cost towns, and a mirroring decrease in low-cost towns as the overall average is about 1.00. The greater growth in the pure premium index among towns that were high-cost in 1999 is consistent with the hypothesis that subsidies positively affect loss cost growth.

Figure 3: BIL Pure Premium Index Growth

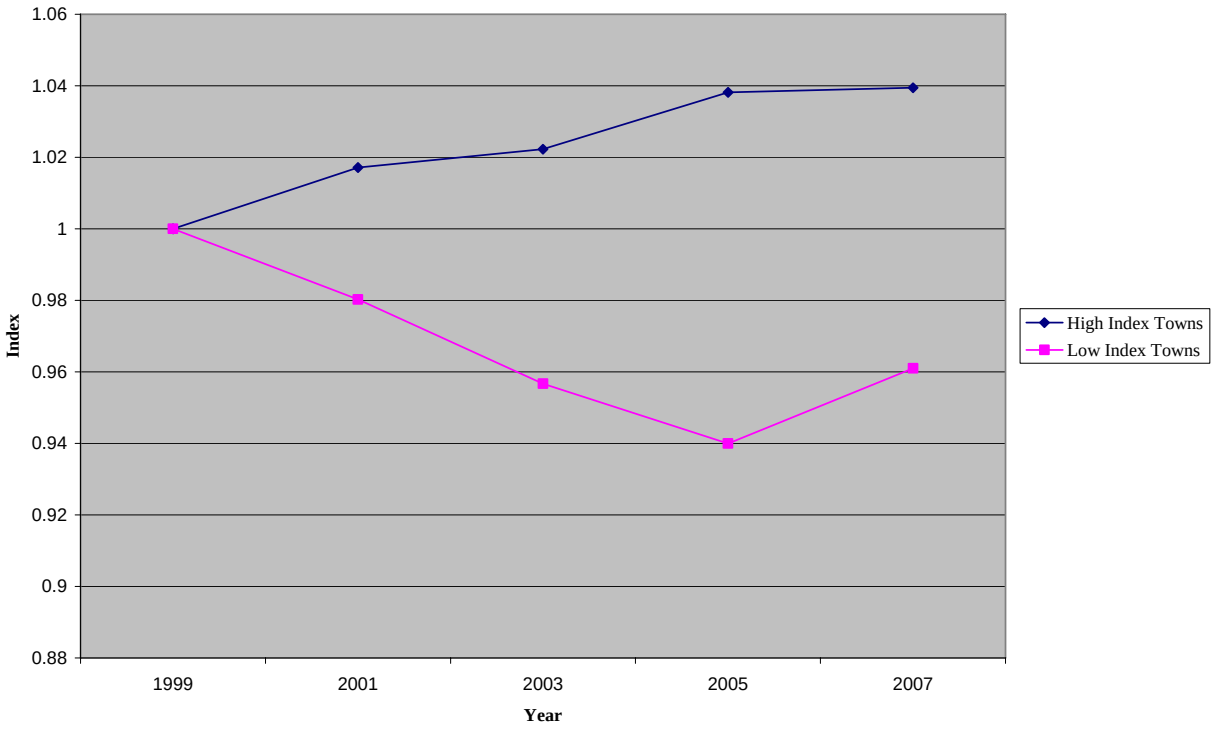
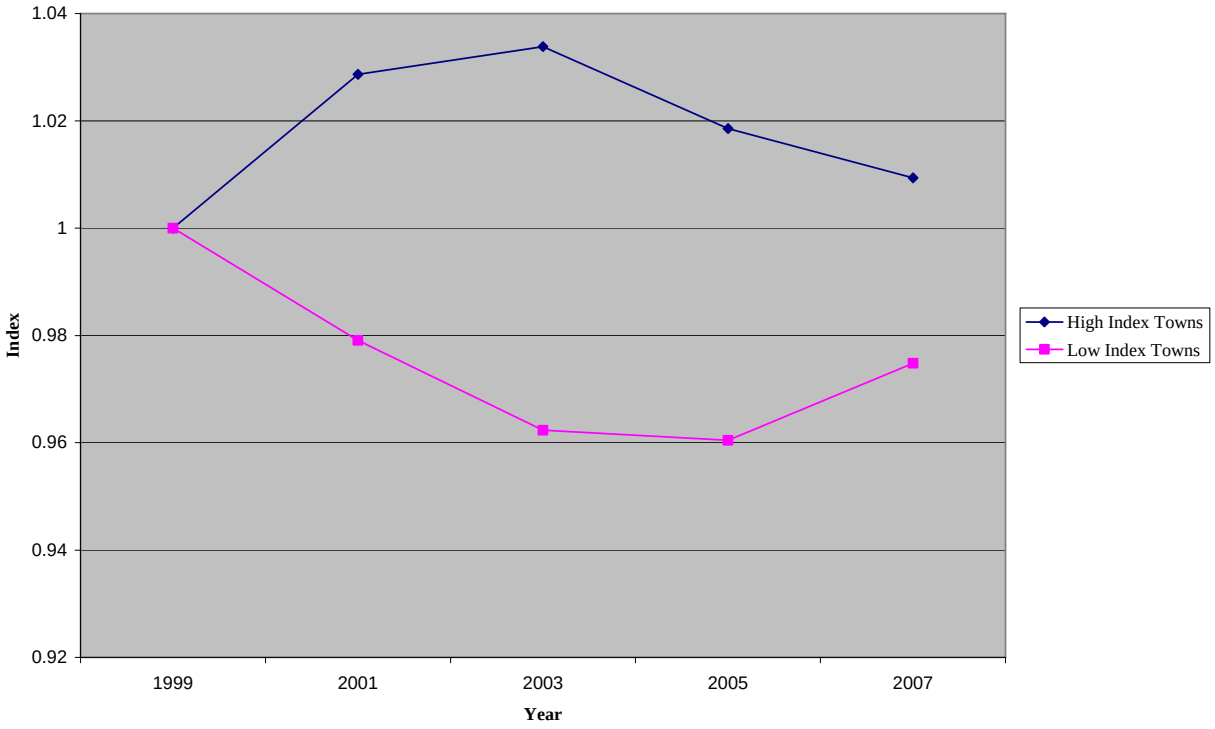


Figure 4: PDL Pure Premium Index Growth



5.3 Regression Analysis

To further explore the relationship between price subsidies and changes in loss costs, simple regression models of the period-to-period changes in pure premiums are estimated. The pure premium regression model takes the following form:

$$\text{Chg}(PP_{it}) = \beta_0 + \beta_1(Subsidy_{it-\tau}) + \beta_2(PP_{it-1}) + \beta_3(Boston_i) + \delta'(Chg\ X_{it}) + T_t + \varepsilon_{it} \qquad (2)$$

where Chg denotes percentage biennial change, subscript i denotes town and t denotes year, and PP_{it} is the pure premium index for town i in year t. The biennial change in the average pure premium index for each town is modeled as depending on the receipt of subsidy in the years during which the loss data were generated, allowing for unspecified differences between Boston-area towns and other towns, and after controlling for changes in other relevant characteristics of each town. The lagged value of the pure premium index is included in the model to allow for the tendency for loss costs to regress toward the mean, and year indicators T_t are included to allow for loss shocks that are common to all towns in a year. The term ε_{it} is a stochastic error term. In estimating the model standard errors are corrected to allow for arbitrary forms of heteroskedasticity and for clustering by town. Clustering takes into account the fact that the regression errors are likely to be correlated within each town across years.[23]

The hypothesis to be tested is that pure premium growth is positively related to the receipt of a subsidy in the years in which the loss data are generated. The period indicator for the subsidy variable is denoted by t-τ to make clear that the relevant period is not the previous territory definition year but rather the years in which the current pure premium data were generated. For example, the 2007 pure premium index is based on loss data from 2001-2004. Because the loss data years overlap for our territory definition years, we use the the subsidy values generated from the two earliest years (using either 2001 or 2002 or both year's subsidy status to analyze the impact on 2007, for example). We obtain the subsidy status and percentage subsidy received by town and year by matching each town to its assigned territory in each year that the underlying loss data were generated.

[23] We do not have enough time series observations to estimate models specifically corrected for autocorrelation.

The model is estimated separately for each of the five auto insurance coverages. The control variables included in the models are the percentage change in the ACRF, the percentage change in the number of cars insured for compulsory coverages, and the change in traffic density in each town. These variables measured using data from the most recent loss data year, for example 2004 data are used for the 2007 pure premium index year.

A second set of models analyzes changes in the average class rating factors (ACRF). These models provide a test of the hypothesis that subsidies lead to greater entry of high-risk drivers into the insurance pool. The model estimated is

$$\text{Chg}(ACRF_{it}) = \beta_0 + \beta_1(Subsidy_{it-\tau}) + \beta_2(ACRF_{it-1}) + \beta_3(Boston_i) + \delta'(ChgX_{it}) + T_t + \varepsilon_{it}, \qquad (3)$$

using the same notation defined for the previous model. The ACRF models are estimated for each of the five automobile insurance coverages separately. Changes in the ACRF may reflect either changes in the demographics of a town or changes in the characteristics of insured drivers in the town (irrespective of changes in town demographics). To capture changes in demographic characteristics that may lead to changes in the propensity to purchase insurance, the control variables include the change in the average age of cars and the change in the percentage of luxury cars registered, as well as the change in traffic density, for each town. As in the pure premium models, we also allow for unspecified differences between Boston-area and other towns.

5.5 Estimation Results

Summary statistics for the variables included in the Massachusetts town regression models are reported in Table 7, and the estimation results are presented in Tables 8, 9, 10, and 11. Tables 8 and 9 report estimates for the pure premium models, and those for the ACRF models are reported in Tables 10 and 11. In each case the first table of estimates use the lagged subsidy status indicators for each town, while the second table reports estimates using the lagged percentage subsidy amount for each town. The subsidy is measured as a percentage of the true cost-based premium for the town, and the subsidy percent variable is set equal to zero in towns that do not receive a subsidy.

Table 7
Summary Statistics
Annual Town-Level Data, Massachusetts
2001-2007 Index Years

Variable	N	Mean	S.D.
Pct change BIL PP Index	1439	0.0006	0.1588
Pct change PIP PP Index	1439	-0.0191	0.1351
Pct change PDL PP Index	1440	-0.0048	0.0524
Pct change COLL PP Index	1439	-0.0056	0.0635
Pct change COMP PP Index	1440	0.0210	0.0963
Lag BIL PP Index	1440	0.7476	0.4503
Lag PIP PP Index	1440	0.7203	0.5688
Lag PDL PP Index	1440	0.8404	0.2344
Lag COLL PP Index	1440	0.8777	0.2644
Lag COMP PP Index	1440	0.8516	0.3896
Pct change BIL ACRF Index	1440	-0.0036	0.0175
Pct change PIP ACRF Index	1440	-0.0048	0.0196
Pct change PDL ACRF Index	1440	-0.0020	0.0130
Pct change COLL ACRF Index	1440	-0.0018	0.0177
Pct change COMP ACRF Index	1440	-0.0011	0.0153
Lag BIL ACRF Index	1440	0.9951	0.0386
Lag PIP ACRF Index	1440	0.9889	0.0463
Lag PDL ACRF Index	1440	1.0000	0.0266
Lag COLL ACRF Index	1440	0.9982	0.0336
Lag COMP ACRF Index	1440	0.9974	0.0452
Subsidy Indicator (lagged)	1440	0.0688	0.2531
Subsidy Percent if subsidized (lagged)	1440	0.0094	0.0453
Pct change Luxury Cars	1433	0.1350	0.2414
Pct change Mean Age of Car	1425	0.0391	0.1023
Pct change Traffic Density	1440	0.0310	0.0454
Boston Dummy	1440	0.2194	0.4140
Pct change Exposures	1440	0.0399	0.0348

Source: Data on PP Indices, ACRF Indices, subsidies, traffic density and insured exposures from AIB; data on luxury cars and mean car age from Massachusetts Division of Motor Vehicles.

The results of estimating changes in pure premium by town are reported in Table 8. The lagged subsidy indicator is positive and significant at the 5% level for PDL coverage and at the 1% level for the remaining four coverages. The estimates support the hypothesis that the growth in pure premiums is positively related to previous subsidies received. This suggests that cost increases are greater in subsidized towns than in unsubsidized towns – consistent with the existence of significant incentive effects on entry into driving and/or riskier driving behaviors created by premium subsidies.

The estimates using subsidy percents rather than subsidy indicators show a positive relationship between larger subsidy percents and pure premium growth for all five coverages. However, the effect of larger subsidies on pure premium growth is statistically insignificant in PD liability coverage, and there are only marginally significant effects (at the 10% confidence level) of larger subsidies on pure premium growth in collision coverage. The estimated effects of larger subsidies are significant at the 1% confidence level for BI liability, PIP, and comprehensive pure premium growth.

Table 8
Regression Analysis of Growth in Pure Premiums by Town
Index Years 2001-2007

Dependent variable = PP Index(t)/PP Index(t-1) - 1

Explanatory Variable	BIL PP		PIP PP		PDL PP		COLL PP		COMP PP	
Lag subsidy indicator	0.0717	***	0.0566	***	0.0112	**	0.0277	***	0.0206	***
	0.0220		0.0184		0.0061		0.0088		0.0134	
Lag pure premium	-0.0598	***	-0.0105		-0.0035		-0.0284	***	-0.0616	***
	0.0192		0.0076		0.0083		0.0101		0.0114	
Pct change exposures	-0.0467		0.0243		-0.0560		-0.1808		-0.0444	
	0.2328		0.1617		0.0797		0.1423		0.1509	
Pct change density	0.3893	*	0.2726	***	0.0686		0.0680		-0.0701	***
	0.1567		1.1E-01		0.0461		0.0456		0.0722	
Boston dummy	0.0010		1.8E-02	**	-0.0050		0.0041		0.0097	*
	0.0134		8.9E-03		0.0039		0.0057		0.0056	
Pct change BIL ACRF	0.3434									
	0.7024									
Pct change PIP ACRF			-0.2124							
			0.3699							
Pct change PDL ACRF					-0.1408					
					0.1835					
Pct change COLL ACRF							0.4936			
							0.3670			
Pct change COMP ACRF									0.2906	
									0.2223	
Intercept	0.0097		-0.0763	***	-0.0145		0.0146		0.0647	***
	0.0211		0.0105		0.0091		0.0148		0.0115	
R-squared	0.0396		0.0794		0.0625		0.0701		0.0841	
F-statistic	2.95	***	19.22	***	7.54	***	7.51	***	8.90	***
N	1439		1439		1440		1439		1439	

Standard errors appear below the coefficient estimates, and are adjusted to allow for arbitrary heteroskedasticity and for arbitrary correlation in errors across years within each town. *** indicates statistically significant at the 1% confidence level ** at the 5% confidence level, and * at the 10% level; all are two-sided tests.

Table 9
Regression Analysis of Growth in Pure Premiums by Town
Index Years 2001-2007

Dependent variable = PP Index(t)/PP Index(t-1) – 1

Explanatory Variable	BIL PP		PIP PP		PDL PP		COLL PP		COMP PP	
Lag subsidy percent	0.5832	***	0.3649	**	0.0328		0.1006	*	0.3504	***
	0.1767		0.1670		0.0355		0.0546		0.1110	
Lag pure premium	-0.0759	***	-0.0162		0.0007		-0.0223	**	-0.0850	***
	0.0213		0.0128		0.0081		0.0106		0.0120	
Pct change exposures	-0.0462		0.0214		-0.0554		-0.1810		-0.0277	
	0.2304		0.1605		0.0795		0.1422		0.1504	
Pct change density	0.3867	***	0.2798	***	0.0727		0.0718		-0.0753	***
	0.1562		0.1087		0.4622		0.0456		0.0712	
Boston dummy	-0.0048		0.0152	*	-0.0045		0.0042		0.0064	*
	0.0141		0.0089		0.0040		0.0059		0.0059	
Pct change BIL ACRF	0.2513									
	0.7074									
Pct change PIP ACRF			-0.2287							
			0.3768							
Pct change PDL ACRF					-0.1370					
					0.1853					
Pct change COLL ACRF							0.5038			
							0.3663			
Pct change COMP ACRF									0.2664	
									0.2221	
Intercept	0.0253		-0.0694	***	-0.0177	*	0.0106		0.0836	***
	0.0219		0.0125		0.0092		0.0152		0.0123	
R-squared	0.0424		0.0782		0.0615		0.0667		0.0913	
F-statistic	3.01	***	21.42	***	7.48	***	6.75	***	13.01	***
N	1439		1439		1440		1439		1440	

Standard errors appear below the coefficient estimates, and are adjusted to allow for arbitrary heteroskedasticity and for arbitrary correlation in errors across years within each town. *** indicates statistically significant at the 1% confidence level ** at the 5% confidence level, and * at the 10% level; all are two-sided tests.

The estimation results relating town subsidy status to changes in ACRF are reported in Table 10. The estimates support the hypothesis that previous subsidies received are significantly related to changes in a town's average class rating factor. The estimated effect is significant at the 1% confidence levels for all five coverages and is greatest in magnitude for BI liability and PIP, even after controlling for changes in other characteristics of towns' auto insurance environments. The estimates reported in Table 11, using percentage subsidies by town rather than subsidy indicators, yield similar conclusions. Consistent with the predictions of theory, these estimates suggest that insurance decisions are sensitive to the receipt of premium subsidies, with subsidies leading to a greater proportion of high-risk drivers in the insurance pool.

Table 10
Regression Analysis of Growth in ACRF by Town
Territory Assignment Years 2001-2007

Dependent variable = ACRF Index(t)/ACRF Index(t-1) - 1

Explanatory Variable	BIL ACRF		PIP ACRF		PDL ACRF		COLL ACRF		COMP ACRF	
Lag subsidy indicator	0.0258	***	0.0326	***	0.0156	***	0.0180	***	0.0142	***
	0.0033		0.0040		0.0022		0.0027		0.0020	
Lag ACRF	-0.0272		-0.0343		-0.0470	**	-0.1256	***	-0.1007	***
	0.0186		0.0222		0.0194		0.0374		0.0209	
Pct change density	7.3E-02	***	7.6E-02	***	4.7E-02	***	5.7E-02	***	7.5E-02	***
	1.9E-02		2.1E-02		1.6E-02		1.9E-02		1.8E-02	
Boston dummy	-2.7E-03	***	-2.1E-03	**	-2.3E-03	***	-2.0E-03	*	-1.4E-03	*
	1.1E-03		1.1E-03		9.0E-04		1.1E-03		8.0E-04	
Pct change luxury cars	3.0E-03		2.3E-03		2.5E-03		-5.9E-03		3.3E-03	
	2.2E-03		2.8E-03		1.7E-03		3.8E-03		3.0E-03	
Pct change mean car age	-3.1E-03		-1.6E-03		-3.7E-03		-8.5E-03	**	-2.4E-03	
	4.2E-03		4.5E-03		3.2E-03		3.5E-03		2.8E-03	
Intercept	0.0186		0.0232		0.0025	**	0.1220	***	0.0930	***
	0.0188		0.0223		0.0011		0.0372		0.0208	
Year fixed effects	Yes		Yes		Yes		Yes		Yes	
R-squared	0.1928		0.2268		0.1341		0.1320		0.1861	
F-statistic	13.14	***	20.21	***	7.81	***	11.32	***	28.88	***
N	1418		1418		1418		1418		1418	

Standard errors appear below the coefficient estimates, and are adjusted to allow for arbitrary heteroskedasticity and for correlation in errors across years within each town. *** indicates statistically significant at the 1% confidence level ** at the 5% confidence level, and * at the 10% level; all are two-sided tests.

Table 11
Regression Analysis of Growth in ACRF by Town
Territory Assignment Years 2001-2007

Dependent variable = ACRF Index(t)/ACRF Index(t-1) - 1

Explanatory Variable	BIL ACRF		PIP ACRF		PDL ACRF		COLL ACRF		COMP ACRF	
Lag subsidy percent	0.1845	***	0.2390	***	0.1061	***	0.1195	***	0.0870	***
	0.0251		0.0344		0.0142		0.0201		0.0136	
Lag ACRF	-0.0689	***	-0.0847	***	-0.0726	***	-0.1501	***	-0.1087	***
	0.0177		0.0232		0.0167		0.0385		0.0220	
Pct change density	7.0E-02	***	7.3E-02	***	4.6E-02	***	5.6E-02	***	7.5E-02	***
	1.9E-02		1.9E-02		1.7E-02		1.9E-02		1.8E-02	
Boston dummy	-4.4E-03	***	-4.3E-03	**	-3.2E-03	***	-2.9E-03	***	-2.0E-03	***
	1.1E-03		1.1E-03		9.0E-04		1.1E-03		8.0E-04	
Pct change luxury cars	3.3E-03		2.6E-03		2.6E-03		-5.7E-03		3.3E-03	
	2.3E-03		3.0E-03		1.8E-03		3.8E-03		3.0E-03	
Pct change mean car age	-3.5E-03		-2.0E-03		-4.0E-03		-8.8E-03	***	-2.8E-03	
	3.8E-03		4.1E-03		3.0E-03		3.4E-03		2.8E-03	
Intercept	0.0618		0.0752	***	0.0686	***	0.1475	***	0.1019	***
	0.0174		0.0227		0.0167		0.0382		0.0218	
Year fixed effects	Yes		Yes		Yes		Yes		Yes	
R-squared	0.2497		0.2931		0.1722		0.1509		0.1942	
F-statistic	14.34	***	21.18	***	8.44	***	11.08	***	28.25	***
N	1418		1418		1418		1418		1418	

Standard errors appear below the coefficient estimates, and are adjusted to allow for arbitrary heteroskedasticity and for correlation in errors across years within each town. *** indicates statistically significant at the 1% confidence level ** at the 5% confidence level, and * at the 10% level; all are two-sided tests.

6. CONCLUSION

Rate regulation in the United States had its origins in the twin concerns of excessive monopoly pricing on the one hand and potential insolvency from inadequate pricing and capital commitment on the other hand. Under the state-by-state regulatory scheme in the United States, rate regulation evolved to address local concerns such as price levels for high-risk insurance consumers, risk classifications (and price differentials) based upon socially unacceptable or controversial

characteristics of insurance consumers, and mandatory levels of coverage. That evolution has led to a patchwork of state-specific laws and regulations with varying levels of stringency and enforcement.

Nowhere is this variety more prominent than in private passenger automobile insurance, where rating classifications and regulatory restraints have promoted cross-subsidies among the insured populations. The most common of subsidy-inducing regulatory actions are (1) restriction of risk classification plans and (2) restrictions on pricing for allowed classification. The strict regulation of classification and pricing of Massachusetts private passenger automobile insurance during 1978-2007 serves here as a test of whether the reduction in efficiency from these cross-subsidy-providing restrictions result in excessive cost growth through overuse of the insurance system by high-risk drivers.

Two approaches were taken to study the potential loss cost reaction to the Massachusetts cross-subsidies that began in systematic form in 1977 and continue through 2007. The first approach compared Massachusetts to countrywide on demographic, regulatory, and liability coverage levels. Loss cost levels that were 44 percent above the expected level were found for Massachusetts during the 1978-1995 period, when premiums charged were those fixed by the state. A second approach considered changing cost levels across Massachusetts by studying loss cost changes by town and relating those changes to subsidy providers and subsidy receivers. Subsidy data for 1999-2007, with underlying accident year data for 1993-2004, showed a significant and positive (relative) growth in loss costs for towns that were subsidy receivers. These results are in line with the theory of underlying incentives for adverse selection and moral hazard created by premium cross-subsidies.

Acknowledgment

The authors are grateful for the historical data and assistance provided by the Automobile Insurers Bureau, especially Kim A. Scott, William Scully, and Eilish Browne and to anonymous referees.

7. REFERENCES

[1] Automobile Insurers Bureau of Massachusetts, "Actuarial Notice - 2: Subsidies in the Rates," Various Years, Boston, MA.

[2] Automobile Insurers Bureau of Massachusetts, AIB Recommendations for 2007 Private Passenger Automobile Insurance Territory Definitions, MA DOI Docket R2006-03, May 15, 2006.

[3] Barkume, A. and J. Ruser, "Deregulating Property-Casualty Insurance Pricing: The Case of Workers' Compensation," *Journal of Law and Economics,* 2001, Vol. 44, 37-64.

[4] Bartlett, D.K. III, R.W. Klein and D.T. Russell, "Attempts to Socialize Costs in Voluntary Insurance Markets: The Historical Record," *Journal of Insurance Regulation*, 1999, Vol. 17.

[5] Bertrand, M., E. Duflo and S. Mullainathan, "How Much Should We Trust Differences-in-Differences Estimates," *Quarterly Journal of Economics,* 2004, 249-275.

[6] Blackmon, B.G. Jr. and R. Zeckhauser, "Mispriced Equity: Regulated Rates for Auto Insurance in Massachusetts," *American Economic Review,* 1991, Vol. 81, 65-69.

[7] Brockett, P.L. and L.L. Golden, "Biological and Psychobehavioral Correlates of Credit Scores and Automobile Insurance Loss: Toward an Explication of Why Credit Scoring Works," *Journal of Risk and Insurance*, March 2007, Vol. 74, No. 1, 22-63.

[8] Buchmueller, T. and J. DiNardo, "Did Community Rating Introduce an Adverse Selection Death Spiral? Evidence from New York, Pennsylvania, and Connecticut," *American Economic Review*, 2002, Vol. 92, 280-294.

[9] Burnes, N.S., "Opinion, Findings, and Decision on the Operation of Competition in Private Passenger Motor Vehicle Insurance in 2008," Massachusetts Division of Insurance Docket No. R2007-03, July 16, 2007.

[10] Cohen, A. and R. Dehejia, "The Effect of Automobile Insurance Accident Liability Laws on Traffic Fatalities," *The Journal of Law and Economics*, 2004, Vol. 47, 357-393.

[11] Conger, R.F., "The Construction of Automobile Rating Territories in Massachusetts," *Proceedings of the Casualty Actuarial Society*, 1988, Vol. 71, Part 1, 1-74.

[12] Cummins, J.D., "Property-Liability Insurance: Price Deregulation: The Last Bastion," in *Deregulating Property-Liability Insurance,* J.D. Cummins, ed. (Washington, D.C.: Brookings Institution Press, 2002) 1-24.

[13] Cummins, J.D. and S. Tennyson, "Controlling Automobile Insurance Costs," *Journal of Economic Perspectives*, American Economic Association, 1992, Vol. 6, No. 2, 95-115.

[14] Cummins, J.D. and S. Tennyson, "Moral Hazard in Insurance Claiming: Evidence from Automobile Insurance," *Journal of Risk and Uncertainty*, 1996, Vol. 12, No. 1, 29-50.

[15] Dahlby, B., "Adverse Selection and Statistical Discrimination: An Analysis of Canadian Automobile Insurance," *Journal of Public Economics*, 1983, Vol. 20, 121-130.

[16] Dahlby, B., "Testing for Asymmetric Information in Canadian Automobile Insurance," in *Contributions to Insurance Economics,* Georges Dionne, ed. (Boston: Kluwer Academic Publishers, 1992).

[17] Danzon, P.M. and S.E. Harrington, "Workers' Compensation Rate Regulation: How Price Controls Increase Costs," *Journal of Law and Economics,* 2001, Vol. 44, 1-36.

[18] Derrig, R.A., D.J. Johnston, and E.A. Sprinkel, "Auto Insurance Fraud: Measurements and Efforts to Combat It," *Risk Management and Insurance Review*, 2006, Vol. 9, No. 2, 109-130, Fall.

[19] Derrig, R.A., "Auto Property Damage Cost Containment - A Billion Dollar Decade of Progress in Massachusetts," Automobile Insurers Bureau of Massachusetts Cost Containment and Fraudulent Claims Payment Filing, Docket R97-37, July 1997, Boston, Massachusetts Division of Insurance.

[20] Derrig, R.A., "Price Regulation in US Automobile Insurance: A Case Study of Massachusetts Private Passenger Automobile Insurance 1978-1990," *The Geneva Papers on Risk and Insurance,* 1993, Vol. 18, 158-173.

[21] DuMouchel, W.H., "The Massachusetts Automobile Insurance Classification Scheme," *The Statistician,* 1983, Vol. 32, 69-81.

[22] Finger, R.J., "Risk Classification," Chapter 6 in *Foundations of Casualty Actuarial Science* (Arlington, Va.: Casualty Actuarial Society, 2001) http://www.casact.org/admissions/syllabus/ch6.pdf.

[23] Greene, W.E., *Econometric Analysis*, 4th edition, (Upper Saddle River: Prentice-Hall, Inc., 2002).

[24] Grace M.F., R.W. Klein, and R.D. Phillips, "Auto Insurance Reform: Salvation in South Carolina," in *Deregulating Property-Liability Insurance,* J.D. Cummins, ed. (Washington, D.C.: Brookings Institution Press, 2002) 148-194.

[25] Harrington, S.E., "Automobile Insurance in Michigan: Regulation, No-fault and Affordability," *Journal of Insurance Regulation*, 1991, Vol. 10, 144-183.

[26] Harrington, S.E., "Effects Of Prior Approval Rate Regulation of Auto Insurance," in *Deregulating Property-Liability Insurance*, J.D. Cummins, ed. (Washington, D.C.: Brookings Institution Press, 2002) 285-314.

[27] Harrington, S.E., "The Relationship Between Voluntary and Involuntary Market Rates: Regulation in Automobile Insurance," *Journal of Risk and Insurance*, 1990, Vol. 57, 9-27.

[28] Harrington, S.E. and P.M. Danzon, "Rate Regulation, Safety Incentives, and Loss Growth in Workers' Compensation Insurance," *Journal of Business*, 2000, Vol. 73, 569-95.

[29] Harrington, S.E. and H.I. Doerpinghaus, "The Economics and Politics of Automobile Insurance Rate Classification," *Journal of Risk and Insurance*, 1993, Vol. 60, 59-84.

[30] Insurance Research Council, "Uninsured Motorists," 1999, Malvern, PA.

[31] Keeton, W.R. and E. Kwerel, "Externalities in Automobile Insurance and the Underinsured Driver Problem," *Journal of Law and Economics,* 1984, Vol. 27, 149-179.

[32] Jaffee, D.M. and T. Russell, "Regulation of Automobile Insurance in California," in *Deregulating Property-Liability Insurance: Restoring Competition and Increasing Market Efficiency,* (Washington, DC: AEI-Brookings Joint Center for Regulatory Studies, 2002).

[33] Jaffee, D.M. and T. Russell, "The Causes and Consequences of Rate Regulation in the Auto Insurance Industry", 2002, 81-112, in *The Economics of Property-Casualty Insurance,* D. F. Bradford, ed. (Chicago, IL: University of Chicago Press, 1998).

[34] Ma, Y.L. and J.T. Schmit, "Factors Affecting the Relative Incidence of Uninsured Motorists Claims," *Journal of Risk and Insurance*, 2000, Vol. 67, 281-294.
[35] Monheit, A.C., Steinberg Schone, B., "How Has Small Group Market Reform Affected Employee Health Insurance Coverage?" *Journal of Public Economics,* 2003, Vol. 88, 237-254.
[36] Monheit, A.C., et al, "Community Rating and Sustainable Individual Health Insurance Markets in New Jersey," *Health Affairs,* 2004, Vol. 23, 167-175.
[37] Puelz R., and W. Kemmsies, "Implications for Unisex Statutes and Risk Pooling: The Costs of Gender and Underwriting Attributes in the Automobile Insurance Market," *Journal of Regulatory Economics*, 1993, Vol. 5, 289-301.
[38] Riley, J.G., "Informational Equilibrium," *Econometrica,* 1979, Vol. 47, 331-359.
[39] Rothschild, M. and J.E. Stiglitz, "Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information," *Quarterly Journal of Economics,* 1976, Vol. 90, 629-649.
[40] Rottenberg, S., "The Cost of Regulated Pricing: a Critical Analysis of Auto Insurance Premium Rate-Setting in Massachusetts," 1989, Boston, Pioneer Institute for Public Policy Research.
[41] Shavell, S., "On Liability and Insurance," *Bell Journal of Economics*, 1982, 13-120.
[42] Simon, K. "Adverse Selection in Health Insurance Markets? Evidence from State Small-Group Health Insurance Reforms," *Journal of Public Economics*, 2005, Vol. 89, 1865-1877.
[43] Stone, J.M., "Opinion and Findings on the Operation of Competition Among Motor Vehicle Insurers," Rendered June 1977, Boston, MA., Massachusetts Division of Insurance.
[44] Stone, J.M., "Opinion, Findings and Decision on the Operation of Competition Among Motor Vehicle Insurers," Rendered June 1978, Boston, MA: Massachusetts Division of Insurance.
[45] Swartz, K. and D.W. Garnick, "Can Adverse Selection be Avoided in a Market for Individual Health Insurance?" *Medical Care Research and Review,* 1999, Vol. 56, 373-388.
[46] Tennyson, S., "The Impact of Rate Regulation on State Automobile Insurance Markets," *Journal of Insurance Regulation*, 1997, Vol. 15: 502-523.
[47] Tennyson, S., M. Weiss and L. Regan, "Automobile Insurance Regulation: The Massachusetts Experience," in *Deregulating Property-Liability Insurance: Restoring Competition and Increasing Market Efficiency,* J. David Cummins, ed. (Washington, D.C.: AEI-Brookings Joint Center for Regulatory Studies, 2002).
[48] Weisberg, H.I. and R.A. Derrig, "Fraud and Automobile Insurance: A Report on the Baseline Study of Bodily Injury Claims in Massachusetts," *Journal of Insurance Regulation*, 1991, Vol. 9, 497-541.
[49] Weisberg, H.I. and R.A. Derrig, "Massachusetts Automobile Bodily Injury Tort Reform," *Journal of Insurance Regulation*, 1992, Spring, Vol. 10, 384-440.
[50] Weiss, M., S. Tennyson and L. Regan, "Incentive Effects of Automobile Insurance Rate Regulation on Loss Costs and Accident Frequency," working paper, 2007.
[51] Worrall, J.D., "Private Passenger Auto Insurance in New Jersey: A Three-Decade Advertisement for Reform," in *Deregulating Property-Liability Insurance,* J.D. Cummins, ed. (Washington, D.C.: Brookings Institution Press, 2002) 81-134.
[52] Yelen, S., "Withdrawal Restrictions in the Automobile Insurance Market," *Yale Law Journal*, 1993, Vol. 102, 1431-1455.

Biographies of the Authors

Richard A. Derrig is president of Opal Consulting, LLC in Providence, Rhode Island. He can be reached at Richard@derrig.com.

Sharon Tennyson, Ph.D., is Associate Professor of Policy Analysis and Management at Cornell University. She can be reached at st96@cornell.edu.

Using Cluster Analysis to Define Geographical Rating Territories

Philip J. Jennings, FCAS, MAAA

Abstract. Geographic risk is a primary rating variable for personal lines insurance in the United States. Creating homogeneous groupings of geographic areas is the goal in defining rating territories. One methodology that can be used for creating these groupings with similar exposure to the risk of insurance losses is cluster analysis. This paper gives a description of an application to define rating territories using a *k*-means partition cluster analysis. Several of the key decisions made during the analysis are detailed including the following: the choice of building blocks, what variables to cluster on, choice of complement of credibility, and what clustering method is appropriate. In addition to the choice I made for each of these, I offer alternative choices that should be considered throughout the process. The method outlined here is based on Michael J Miller's presentation at the 2004 CAS Ratemaking Seminar titled "Determination of Geographical Territories." The measure of homogeneity used for this analysis is the within cluster variance as a percentage of the total variance. It will be shown that for the particular analysis that I describe in this paper, the within cluster variance as a percentage of the total variance was significantly reduced from 29.4% to 5.3%. This was also a more powerful result in comparison to the territory definitions of any of the major writers in this state.

Keywords. Rating territory definitions, cluster analysis, personal lines.

1. INTRODUCTION

Current actuarial ratemaking methodologies for the pricing of personal lines automobile and homeowners insurance in the United States include a geographical component. Almost all personal lines insurers incorporate geography by varying price by rating territory. These rating territories are typically defined by groupings of geographical regions. Most insurers use zip code boundaries to define the geographical areas. Zip codes are grouped together based on similar expected loss costs (expected losses for an individual policy for a policy term). In the past, rating territory definitions were based on subjective information such as agent feedback or loss ratios that may have lacked credibility. It is clear now that historical territory definitions used by some companies lacked statistical support and may have lost meaning over time.

The technique for defining rating territories described in this paper was inspired by, and is primarily derived from, a presentation by Michael J. Miller at the 2004 Casualty Actuarial Society Ratemaking Seminar titled "Determination of Geographical Territories." Miller [4] defines homogeneity in terms of risk classification stating, "A risk classification is homogeneous if all risks in the class have the same or similar degree of risk with respect to the specific risk factor being

measured." And as an example Miller states, "A territory is considered homogeneous if all risks in the territory represent the same, or approximately the same, geographical risk."

One methodology that lends itself quite well to performing this grouping of geographical areas is cluster analysis. Kaufman and Rousseeuw [3] define cluster analysis as "the art of finding groups in data." I include this definition because it conveys the idea that although the methodology is scientific and technical in nature, there is still an element of art involved in a cluster analysis application.

The statistical test of homogeneity presented by Miller and used in this analysis is the within cluster variance as a percentage of the total variance. The within cluster variance is based on the squared difference between each building block's pure premium in the cluster and the average pure premium for the specific cluster being tested. The building blocks for the analysis presented in this paper are zip codes. The between clusters variance is based on the squared difference between each cluster's pure premium and the statewide average pure premium. The total variance is equal to the sum of the within cluster variance and the between clusters variance. The goal is to achieve a low within cluster variance percentage and a high between clusters variance percentage to the total variance.

1.1 Research Context

This paper covers material that falls under CAS Research Taxonomy I.G.12.g Actuarial Applications and Methodologies/Ratemaking/Trend and Loss Development/Territory Analysis. Defining rating territories using cluster analysis was outlined by Miller in his presentation at the 2004 CAS Ratemaking Seminar. Other creative approaches to defining rating territories or addressing the geographic risk component in ratemaking are given by Christopherson and Werland [2] and Brubaker [1]. Werner [6] highlights the disadvantages and hazards of using a building block that can change over time for territory definitions.

1.2 Objective

This paper will provide a guideline to performing a cluster analysis in order to define rating territories. There are many decisions to consider during the process. The goal of this paper is not to give a rigid set of steps to follow but rather to present one application of this type of analysis and to offer various options at each step throughout the process.

1.3 Outline

The following section details several of the decisions that need to be made to perform this type of analysis. I will describe the choice that I made at each step during one application of this methodology along with some alternatives that could have been used and some issues that might be encountered at each step. Finally, I will discuss some of the implementation issues that may arise once the analysis is completed.

2. BACKGROUND AND METHODS

Each section below highlights one of the fundamental choices that need to be made during the course of the analysis. This includes the choice of building blocks to use in the analysis, what data to use, what variables to cluster on, what data to use for the complement of credibility, and issues to be aware of when choosing the clustering method.

2.1 Building Blocks

One of the first considerations is what to choose for the geographic building blocks for the proposed territories. This choice may be constrained by the company's technology resources available. Typically, companies use postal zip codes for defining rating territories. Zip codes are a convenient geographical area to use in this type of analysis since they are readily available and well known in the general population. However, zip codes in the United States were never designed to group homogeneous risks for exposure to insurance losses. In addition, zip codes are subject to change over time.

Other alternative territory building blocks include all of the census geographical boundary definitions such as minor civil divisions, census county divisions, census tracts, block groups, or even census blocks. These options have the advantage of being more stable than postal zip codes over time and, at the census tract, block group, and block level, contain relatively homogeneous units with respect to population characteristics and living conditions at the time they are established. In order to use any of these census geographies, a company would need to have a front-end system in place in order to assign the policy to the correct grouping based on the address location since these geographical boundaries are generally not known by the average consumer. With the growth in the availability and sophistication of geographic information systems (GIS), using these geographical

areas for territorial building blocks has become easier to implement.

Werner [6] describes the disadvantages of choosing as a building block a geographical unit whose boundaries can change over time. He also provides the following list of considerations when deciding which geographic risk unit to use:

- The building block must be small enough to be homogeneous with respect to geographic risk.
- The unit should be large enough to produce credible results.
- The collected company loss and premium data should be easily assigned to the chosen unit.
- All competitive and/or external data should be easily mapped to the geographical unit.
- It should be easy for the insured and company personnel to understand.
- The unit must be politically acceptable.
- The unit should be verifiable.
- The geographic unit should not change over time.

The focus of Werner's paper is on the last bullet point but he provides details of the other elements of this list in the appendix to his paper.

Brubaker describes a method for assessing geographic risk without defining territories or territory boundaries. His method assigns a geographic rate to a set of grid points. Then for a specific location the rate is interpolated from the nearest grid points. He suggests that it may be desirable to vary the spacing of these grid points having smaller grids where expected loss varies over relatively short distances and allow for greater spacing in rural areas where expected losses may not vary as much over short distances.

2.2 Data

For the example presented in this paper I used five years of private passenger automobile accident year data for State X, including premium, exposures, incurred losses, and incurred claim counts. The incurred losses were developed to ultimate and trended to the average settlement date. This was done by coverage using standard actuarial techniques. Liability losses were capped at a predefined amount to minimize the impact of large losses.

As stated above the data used should be easily assigned to the chosen building block. Zip codes were used for this analysis since the company's data was easily assigned to zip code. With data at the policy level and given clean addresses associated with each policy, a good GIS can geocode (assign

latitude and longitude) each policy record. The data can then be aggregated within the GIS tool to any geographical region used as a building block. External data that can be geocoded can also be aggregated to any geographical region.

In this step of the process it may become necessary for some level of manual cleansing of the data. This is particularly relevant if the building blocks are subject to change over time as is true with zip codes. Zip codes are added and deleted periodically by the U.S. Postal Service. The final proposed territories should be defined using the current active zip codes. Any zip codes in your experience period data that have been deleted need to be examined and the data for those zip codes reassigned to the current zip codes for that area. If your policy level data is geocoded and you have digitized zip code boundaries, the assignment of historical data to zip codes is a straightforward point in polygon assignment within a GIS. However, if you are lacking geocoded policy data and digitized zip code boundaries, this assignment of historical data to the current zip codes can become a difficult and labor intensive project. For example, if a zip code has been split into two new zip codes the optimal process would involve obtaining street maps and updated zip code maps to correctly assign each policy's data to the correct zip code based on the street address. This may not be a reasonable approach depending on the volume of data that needs to be investigated and any particular time constraints for your project. Alternatively, your historical data could be allocated based on a population density estimate or the size of the geographical area for the new zip codes.

Another data issue that may require manual intervention relates to zip codes that are in fact post office box (P.O. box) zip codes. In this case the location of the post office for that PO box zip code can be used to allocate the historical data to the correct currently active surrounding zip codes.

2.3 Variables to Cluster On

One significant benefit of clustering methods is that they allow for the inclusion of as many variables as desired. This means that clusters could be created separately based on similar claim frequencies and based on similar claim severities or variables can be included to create one set of clusters based on both components. In order to capture both a frequency and a severity component of geographic risk, this analysis used a credibility-weighted frequency and a credibility-weighted pure premium for each zip code. The derivation of these variables will be discussed below.

Traditionally, rating territory definitions are based on large contiguous geographical areas defined

by groups of zip codes. To increase the acceptability of the new territory definitions resulting from this analysis from both a regulatory and a sales agent perspective, I wanted to maintain the contiguous nature of territory definitions. One way to accomplish this is to include the zip code's centroid (geographic center) latitude and longitude as variables in the clustering routine. This step is not necessary if there are no constraints on the number of rating territories allowed. In fact, the measure of homogeneity this methodology is based on, the within variance as a percentage of the total variance, is minimized at zero if each building block becomes its own rating territory. However, as is shown in Exhibit 1, the within cluster variance percentage has a decreasing marginal rate of improvement as the number of territories increases beyond a certain point. So the optimal number of territories may be influenced by this decreasing marginal improvement as well as acceptability to sales agents and regulators.

Clustering methods create groups of building blocks based on a similarity (or dissimilarity) measure. The degree of influence a certain variable carries in the analysis is driven by the range of values for that variable. A variable with a wide range of values will have more influence in the resulting clusters than a variable with a narrow range of values. For this reason, if we want all the clustering variables to carry the same weight in the resulting clusters, it is important to standardize or transform each of the variables before performing the cluster analysis. Some software packages that perform cluster analysis automatically perform this standardization while others do not.

I chose to standardize all of the variables to the same mean and standard deviation. In addition, this step of transforming the variables can also allow the researcher the flexibility of ranking the influence of variables if desired. By transforming the variables to have differing variability you can control the influence a given variable will have on the resulting clusters. Those with wider variability will have a greater influence on the final clusters than those with a narrower swing. Caution should be exercised regarding standardization of variables because some of the similarity measures available for use require non-negative values for all variables.

2.4 Complement of Credibility

Data can be thin at the fundamental building block-level and the smaller the building block, the less credible it can become. To supplement my zip code-level data, I used a form of the principle of locality that can be stated as follows: the expected loss experience at a given location is similar to the loss experience nearest to that location.

The creation of a credibility-weighted pure premium for each zip code proceeded as follows. I started out with a pure premium for each zip code, the total losses divided by total bodily injury liability exposures for each zip code. Then for each zip code, I used the latitude and longitude of the centroid to determine the group of zip codes whose centroid is within a five-mile radius of this zip code. Next I computed a pure premium for this group. I used a Visual Basic script and macro to compare zip code centroids but most GIS software can create these groupings for you. The grouping of zip codes and calculation of pure premium were repeated using 10-, 15-, 20-, 25-, and 50- mile radius circles. The statewide average pure premium was also calculated. For each zip code, credibility was assigned to the zip code pure premium and the six groupings associated with that zip code. This credibility value was calculated using earned premium and the formula $z = P / (P+K)$ where z is the credibility assigned, P = Earned Premium, and K = a credibility constant of \$2,500,000. For the five-mile radius grouping pure premium, the credibility assigned to the zip code was subtracted out to get the credibility assigned to this grouping's pure premium. For the 10-mile radius grouping, the credibility previously assigned to the zip code and the five-mile radius grouping were subtracted out of the formula credibility for the 10-mile radius group to get a credibility value to assign to the 10-mile radius pure premium. This process continued through the 15-, 20-, 25-, and 50-mile radius groupings, each time subtracting out previously assigned credibility. If the sum of the assigned credibilities was not at 100%, then any remaining credibility was assigned to the statewide average pure premium. Now a credibility weighted average pure premium has been calculated for each zip code.

The process described in the preceding paragraph was repeated for the claim frequency of each zip code. The only difference in methodology was that claim counts were used for credibility to assign to the frequencies using the formula z=minimum (1, $\sqrt{(n/k)}$) where n = the number of incurred claims and k= 1,082. At this point we now have a credibility weighted pure premium and frequency for each zip code.

Miller, in his analysis, uses a normalized zip code pure premium to cluster on. His measure is defined as:

$$\frac{\text{State Average Premium}}{\text{State Average Base}} \div \frac{\text{Zip Code Average Premium}}{\text{Zip Code Base}}$$

For a credibility constant Miller suggests the use of 3,000 claims.

The credibility formulas used in my analysis are widely accepted methods for assigning partial credibility and are well documented in CAS literature. There are many choices for credibility and the complement of credibility. Miller lists several choices for the complement of credibility including data grouped based on population density groups, vehicle density, accidents per vehicle, injuries per accident, or thefts per vehicle for whatever building block you may be using. The method of assigning credibility described above was designed to pick up the information from the surrounding geographical areas of a zip code. For most zip codes in this study, almost all credibility was assigned within a 10-mile radius. However, there are some drawbacks or potential dangers to using this method. You may be calculating the credibility-weighted pure premium for a rural zip code with a low volume of experience in your data. If most of the credibility gets assigned to a 50-mile radius grouping, you could pick up experience from very different areas that are in fact not homogeneous to the conditions of the zip code you are evaluating. An inverse distance weighting approach may be more appropriate.

Christopherson and Werland [2] incorporate a form of inverse distance weighting by using a linear weighting function to weight data from zip codes within a 35-km radius of a given zip code's centroid with less weight given as the zip code's centroid gets farther away. They offer the following function that is simple but effectively gives greater weight to nearer data.

Distance	Weight
0<= *d* <= 5 km	1
5 km < *d* < 35 km	(35-*x*)/30
35 km <= *d*	0

They weight the exposures in the nearby zip codes and combine these with the given zip code's exposures to assign a credibility value to the zip code. To arrive at an adjusted pure premium for a local zip code center they do a three-way credibility weighting using the zip code, the metropolitan statistical area grouping (rural vs. non-rural), and the statewide pure premium.

Miller also includes as one choice for a complement of credibility the use of a distance based criteria. He presents a sigmoid curve of the form:

$$Y = 1 / (1 + \exp(-a(b-x-c)) \quad (2.1)$$

This curve will provide decreasing weights as the distance, x, increases. It also provides flexibility in its shape through the choices for the *a*, *b*, and *c* parameters.

Another consideration regarding the approach of using concentric rings of zip code groupings becomes apparent when considering zip codes that fall along a state's border or coastline. In this particular application of this methodology I made no adjustment for this issue. One adjustment that could be made for non-coastline state border zip codes is to incorporate historical data from neighboring states. Caution should be exercised here and adjustments may need to be made if there are significant differences in the regulatory and legislative environments between the state being analyzed and the neighboring state. For example, differences in tort law or minimum liability financial responsibility limits may have an influence on your claims data. An adjustment that could be made for coastal zip codes is to use similar coastal zip code data for credibility complements rather than concentric circles. In effect oval bands along the coast could be created rather than using circles.

My analysis was performed on an all coverages combined basis. Given adequate time to complete a thorough analysis one would probably want to perform the analysis by coverage. It is reasonable to assume that the resulting territory boundaries would vary by coverage. If system resources could

support this level of detail a company could have territory definitions by coverage. Or the intersections of the by coverage territory definitions could be used to define an overall set of territory definitions. It also seems reasonable to expect that the chosen credibility complement could, and probably should, vary by coverage. For example, relating to the use of external data, a medical cost index might be used for bodily injury liability while a theft rate might be used for comprehensive coverage.

2.5 Clustering Method

It has been said that there are as many cluster analysis methods as there are people performing cluster analysis. This is a gross understatement! There exist infinitely more ways to perform a cluster analysis than people who perform them. StataCorp [5].

Several general types of cluster analysis methods exist. For each of these general types there are a number of specific methods and most of these cluster analysis methods can use a wide array of similarity or dissimilarity measures. The statistical analysis software tool I used for this clustering analysis is Stata [5]. Two of the general types of clustering methods are available in Stata: hierarchical and partition. Hierarchical clustering methods create, by combining or dividing, hierarchically related sets of clusters. Partition clustering methods separate the observations into mutually exclusive groups. Of the many different partition methods, Stata has two of them available, *k*-means and *k*-medians. The partition cluster method I used for this analysis is *k*-means. The number of clusters to create (*k*) is specified by the user. These *k* clusters are formed iteratively. Starting with *k* means, or centers, each observation is assigned to the group whose mean is closest to that observation's mean. New group means are then calculated. This continues until no observations change groups. With this method of cluster analysis, for the similarity measure I used the Euclidean distance metric (also known as the Minkowski distance metric with argument 2).

$$\left\{\sum_{m=1}^{p}\left(X_{mi}-X_{mj}\right)^2\right\}^{1/2} \tag{2.2}$$

where X_{mi} = value of observation (zip code) i and variable m. A general form for the distance metric between observation i and centroid j using p variables is given by

$$\left\{\sum_{m=1}^{p}\left|X_{mi}-X_{mj}\right|^{N}\right\}^{1/N} \tag{2.3}$$

This is called the L_N norm or the Minkowski distance metric with argument N. When $N = 1$ this is known as the absolute, cityblock, or Manhattan distance. There are also several variations on this formula along with other distance metrics that are available. Note that in these formulas the summation is over the p variables—in this case latitude, longitude, pure premium, and frequency. Latitude and longitude were included to make the territories as contiguous as possible.

Using Stata, groupings were generated for $k = 1$ to 100, where k is now the number of proposed territories. For each k the cluster variance as a percentage of the total variance was calculated. Exhibit 1 shows a graph of the within cluster variance percentage for each value of k, the number of proposed territories. This graph shows that the within cluster variance percentage drops off quickly as the number of territories increases and then levels off considerably indicating a decreasing marginal improvement in this measure of homogeneity. We also took our current territory definitions as well as the territory definitions from several major competitors and calculated the within cluster variance percentage for those groupings of zip codes. These values are also plotted on Exhibit 1 for reference.

When using k-means clustering the starting values, or initial centers, are an important consideration and can affect the resulting clusters. Stata has several built-in options for the starting values. These options include choosing k unique observations at random with an optional seed; using the first k or last k observations; randomly forming k partitions and using the means of these k groups; using group centers formed from assigning observations 1, $1+k$, $1+2k$,…to the first group; assigning observations 2, $2+k$, $2+2k$,… to the second group and so on to form the k groups; also one can group on a variable in your dataset to form k groups and use the mean of these for starting values. Another option available is to create k nearly equal partitions by taking the first N/k observations for the first group, the second N/k observations for the second group, and so on and using the means for these groups as the starting values. This is the option I used after sorting the data by the credibility-weighted pure premium.

Although the within cluster variance as a percentage of the total variance results for each k were similar for different starting values as shown in Exhibit 2, the groupings of zip codes did display some differences. Exhibit 3 shows a histogram of the differences in pure premiums using two different sets of clusters created using different methods to obtain starting values. The first set of clusters was created by setting k=90, sorting by pure premium then using starting values with the

mean of k nearly equal partitions taking the first N/k observations for the first group, the second N/k observations for the second group, and so on. These results are compared with a second set of clusters obtained by setting k=90 and using k random initial group centers chosen from a uniform distribution over the range of the data. This comparison shows that 87% of the zip codes end up in clusters that have resulting pure premiums from both methods within +/- 5%. However, there are some zip codes, 3%, that fall outside of a +/- 10% difference. These results show that consideration should be given to the choice of starting values and the results of several choices should be evaluated.

2.6 Implementation Issues

For my first implementation of this methodology I had the luxury of being able to define rating territories for a new company. This new company existed by license but had no current business written in it. Therefore, there was no need to be concerned with rate disruption to an existing book of business. Subsequent applications of this methodology did not come with this luxury. A great deal of effort may be needed to analyze the full extent of rate disruption and make the appropriate adjustments to the resulting clusters to bring the impacts into an acceptable range. State restrictions on overall rate increases or differences within prior territories or counties within a state may require additional adjustments.

The disruption resulting from creating new territory definitions not only affects customers and potential customers but also may have an impact on sales management and the sales agents. Even a simple re-numbering of territory codes may cause great consternation with your sales force. What seemed reasonable to the researchers at the time, to re-number the territories in order of their within cluster variance percentage indicating the analysts confidence level with the results, may invoke many questions and concerns why the historical territories 1 through 4 are now territories 5 , 9, 26, and 38.

Another set of implementation issues deal with the choice of building blocks to define rating territories. The optimal building block may be grids defined by latitude and longitude boundaries as used by Brubaker. This would require each address to be geocoded corresponding to the address of the location of garaging for an automobile policy or the exact location of the insured dwelling for a homeowners policy. Or the building blocks may be census blocks, block groups, or tracts that, again, would require the assignment of the correct census geography. In today's environment of GIS

capabilities these are not unreasonable expectations but may require significant yet worthy company investments to integrate into production environments.

3. RESULTS AND DISCUSSION

The results of this particular cluster analysis are shown in Exhibit 1. The graph shows the results for 4 competitors along with the results for the writing company we were using in this state prior to this analysis (labeled as current on Exhibit 1). The results show that we were able to reduce the within cluster variance percentage from 29.4% to 5.3%. After the final cluster analysis was run, there were still some manual adjustments done to get to the final proposed territory definitions. This involved considerations of contiguity of territories, competitive concerns, and sales presence. This is why the final proposed point on the graph with 90 territories lies slightly above the within cluster variance percentage curve.

The competitors shown on Exhibit 1 ranged from a low of 28.2% up to 31.6%. A fifth competitor we measured is not shown on this exhibit but even with 140 territories had a within cluster variance of 24.6%. This example demonstrates the significant improvement in this measure of homogeneity that can be achieved.

From Exhibit 1 it is graphically evident that there is a decreasing marginal improvement in the within cluster (territory) variance percentage and that we could have obtained similar results, based on this measure, by choosing fewer than 90 territories. However, by creating a greater number of territories as a result of this analysis the company was now positioned to grow the book of business and allow each territory's rates to move in the appropriate direction in the future, based on its own emerging loss experience, without having to repeat a territorial re-alignment analysis as often as might be necessary otherwise. As in any rating or pricing analysis, business judgment plays a key role in interpreting and implementing the final statistical results of the analysis. Statistical results should be used in conjunction with the company's growth and profitability objectives to implement the optimum pricing program within each state.

4. CONCLUSIONS

This paper has presented an application of one technique to define geographic rating territories.

Cluster analysis can be a valuable tool to use towards the goal of determining homogeneous groups of geographic areas. It has many options associated with the choices of clustering methods, similarity measures, and starting values. The art of applying this technique lies in the investigation of the impact that each step of the analysis has on the resulting clusters. The power of this technique is revealed by the dramatic increase in the homogeneity of the building blocks inside each of the resulting territories compared to the current definitions as measured by the within cluster variance as a percentage of the total variance.

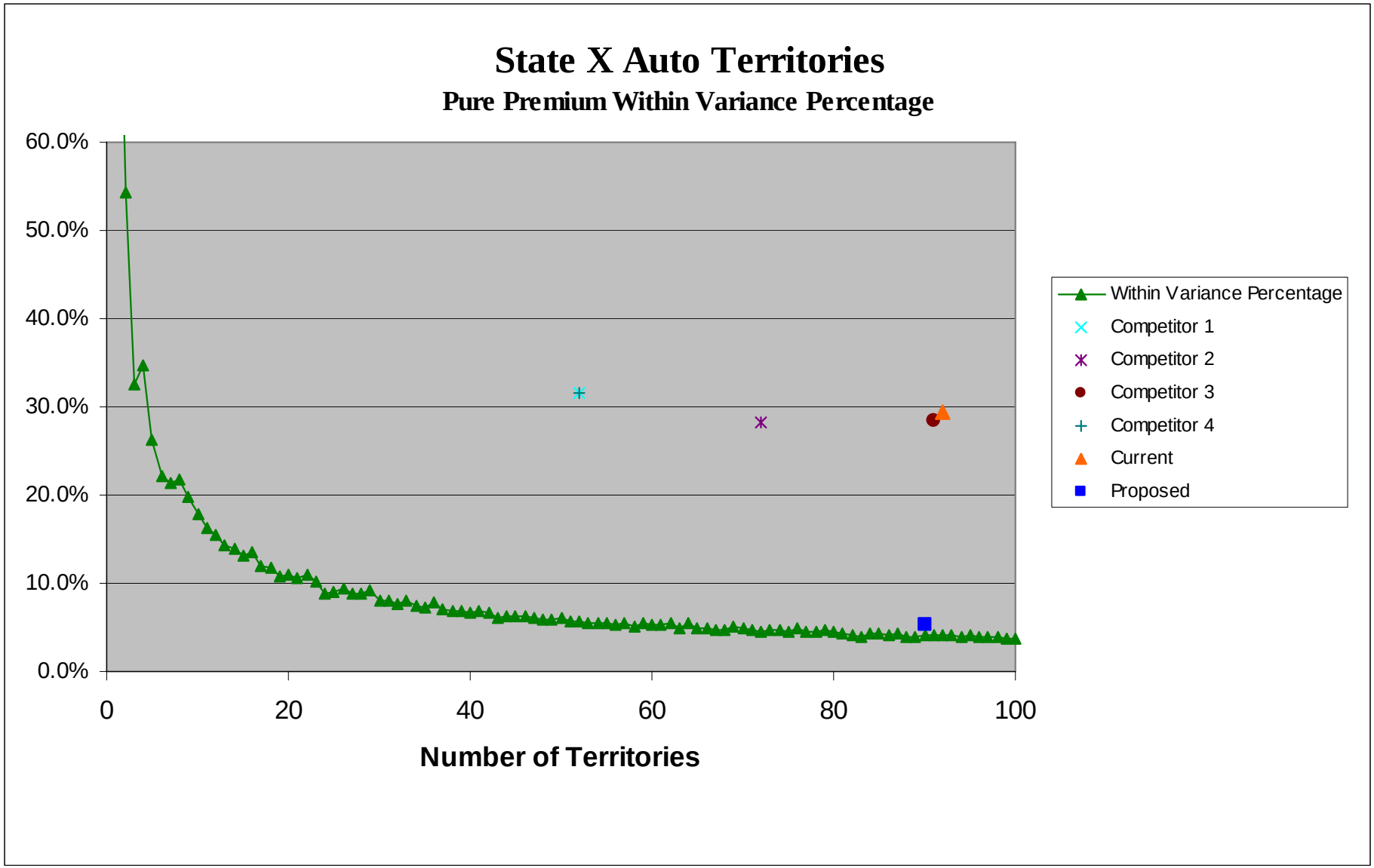
5. REFERENCES

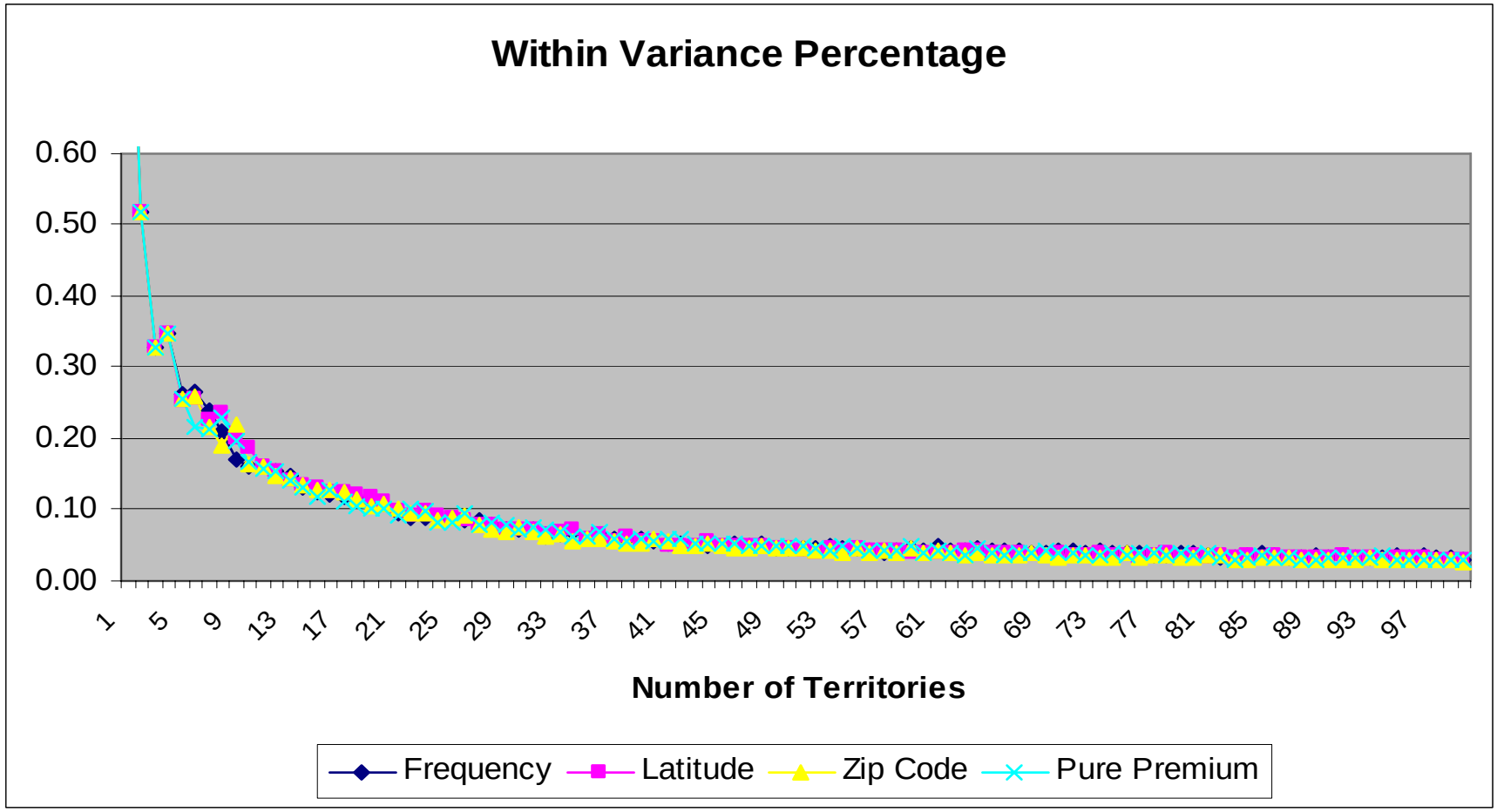
[1] Brubaker, Randall E., "Geographic Rating of Individual Risk Transfer Costs Without Territorial Boundaries," *Casualty Actuarial Society Forum*, 1996, Winter, 97-127.
[2] Christopherson, Steven, and Debra L. Werland, "Using a Geographic Information System to Identify Territory Boundaries," *Casualty Actuarial Society Forum*, 1996, Winter, 191-211.
[3] Kaufman, L. and P.J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis* (Hoboken, New Jersey: John Wiley & Sons, 1990).
[4] Miller, Michael J., "Determination of Geographical Territories," Presented at the 2004 CAS Ratemaking Seminar.
[5] *Stata 8 Cluster Analysis Reference Manual*, (College Station, TX: StataCorp, 2003), 5. (Parts reprinted by permission of the publisher.)
[6] Werner, Geoffrey, "The United States Postal Service's New Role: Territorial Ratemaking," *Casualty Actuarial Society Forum*, 1999, Winter, 287-308.

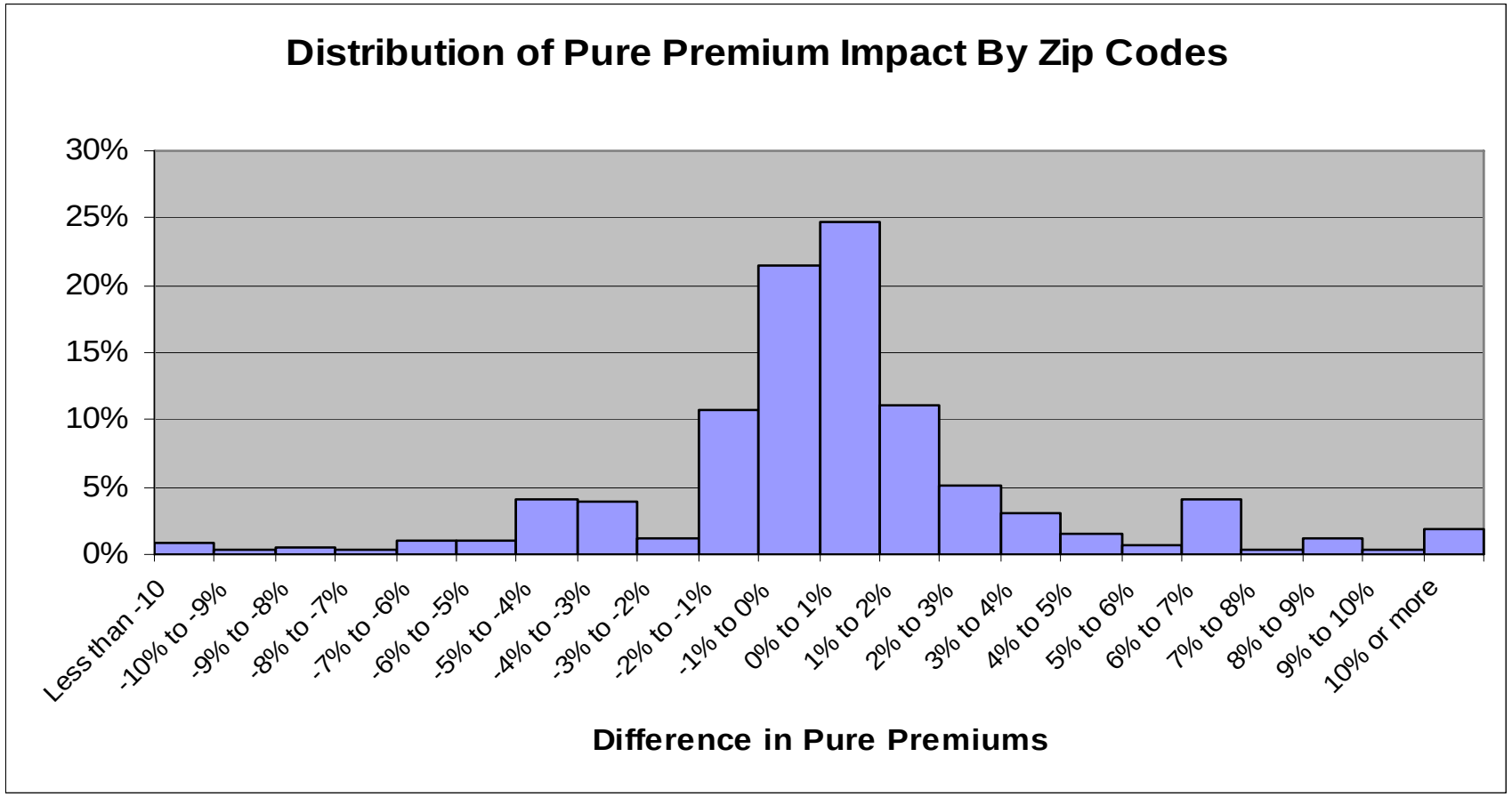
Abbreviations and notations
GIS, Geographic Information System

Biography of the Author

Phil Jennings is a director on the Quantitative Research and Modeling Team of the Actuarial Department at MetLife Auto & Home Insurance Company in Warwick R.I. He has a bachelors degree in mathematics from Regis University in Denver, Colorado and a masters degree in mathematics from the University of Arkansas. He is a Fellow of the CAS and a Member of the American Academy of Actuaries. He also participates on the CAS Examination Committee.


State X Auto Territories
Pure Premium Within Variance Percentage
60.0%
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
0
20
40
60
80
100
Number of Territories
Within Variance Percentage
Competitor 1
Competitor 2
Competitor 3
Competitor 4
Current
Proposed


Within Variance Percentage
0.60
0.50
0.40
0.30
0.20
0.10
0.00
1
5
9
13
17
21
25
29
33
37
41
45
49
53
57
61
65
69
73
77
81
85
89
93
97
Number of Territories
Frequency
Latitude
Zip Code
Pure Premium



This histogram shows an example of the impact on the resulting cluster pure premiums for zip codes assigned to different groups depending on the starting values used. This comparison shows that 87% of the zip codes end up in clusters that have resulting pure premiums from both methods within +/- 5%. However, there are some zip codes, 3%, that fall outside of a +/- 10% difference. These results show that consideration should be given to the choice of starting values and the results of several choices should be evaluated.

A Multivariate Model for Predicting the Efficiency of Financial Performance for Property and Liability Egyptian Insurance Companies

Osama Hanafy Mahmoud

Abstract: This paper uses the financial data of some property-liability insurance companies in Egypt to develop a multivariate model that reflects the efficiency of financial performance. Data will be classified statistically among three categories of financial performance based on the results of fuzzy clustering. The predictive variables in the multivariate model are represented as 25 financial ratios, which are more commonly used in describing the financial performance of the insurance company. Factor analysis has been utilized as a data reduction technique. The paper also examines the effect of the insurance company's ownership type (public vs. private). A multivariate model should be better able to identify the efficiency of financial performance by comparing the results of discriminant analysis and logistic regression.
Key Words: Financial performance, factor analysis, fuzzy clustering, discriminant analysis, logistic regression, property and liability insurance, Egyptian insurance companies

1. INTRODUCTION

Insurance companies sell protection to policyholders against many types of risks: property damage or loss, health and casualty, financial losses, etc. In return for this risk protection, insurance companies receive a premium from the policyholder that is used to cover expenses and the expected risk. For longer-term risk protections, part of the premiums is invested to get higher yields. Although the protection buyer mitigates the individual risk to the large and better-diversified portfolio of the insurer, the risk is not completely reduced because the insurer may default his obligations. Insurers need to have sufficient equity or buffer capital to meet their obligations in adverse conditions when their losses on the diversified portfolio exceed the expected losses. Ratings provide an assessment of the ability of the insurer to meet its obligations to policyholders and debt holders.

This study presents a model for identifying the financial performance of insurance companies. In this paper, financial ratios are used to describe and predict the financial performance of insurers. Beaver (1966) uses the financial ratios as predictors of failure and states that the usefulness of ratios can only be tested with regard to some particular purpose. Ratios are currently in widespread use as predictors of failure. While this is not the only possible use of ratios, it is a starting point from which to build an empirical verification of ratio analysis. Van Gestel et al. (2007) analyze the relationship between financial ratios and the rating for different types of insurance companies by using advanced statistical techniques that are able to detect non-linear relationship.

Various statistical models for the classification and prediction of financial performance have been presented in prior studies. Harrington and Nelson (1986) used regression analysis to estimate the

relationship between premium-to-surplus ratios and insurer characteristics, including asset and product mix variables. Analysis of the regression residuals then can be used to identify insurers with ratios that are substantially higher than those for insurers with similar characteristics. The method is illustrated using data for solvent and insolvent insurers. The method's ability to identify insurers that later became insolvent is compared to that of the National Association of Insurance Commissioners (NAIC) Insurance Regulatory Information System. BarNiv and Hershbarger (1990) presented models that incorporate variables designed to identify the financial solvency of life insurers. Three multivariate analyses (multidiscriminant, nonparametric, and logit) have been used to examine the applicability and efficiency of alternative multivariate models for solvency surveillance of life insurers. Thus, decision tools are developed based on real data, and systematic statistical frameworks are applied for evaluating the financial viability of life insurers. A comprehensive review of insurer insolvency literature is given in BarNiv and McDonald (1992).

Ambrose and Carroll (1994) examined the efficiency of Best's recommendations, Insurance Regulatory Information System (IRIS) ratios, and other financial measures for their statistical ability to classify solvent and insolvent life insurers. Ambrose and Carroll estimated classification models for a sample of insurers for 1969 through 1986 and applied the models to a holdout sample for 1987 through 1991. The financial variables and IRIS ratios outperformed Best's recommendations in distinguishing between the two groups in a logit model. Lee and Urrutia (1996) compared the performance of the logit and hazard models in predicting insolvency and detecting variables that have a statistically significant impact on the solvency of property-liability insurers. The empirical results indicated that the hazard model identifies more significant variables than the logit model and that both models have comparable forecasting accuracy.

Using estimations across 18 lines of insurance for the years 1984 through 1993, Chidambaran et al. (1997) presented an empirical analysis of the economic performance of the U.S. property-liability insurance industry. They adopted an industrial organization approach, focusing on the economic loss ratio as a measure of pricing performance. The concentration ratio for the line and the share of direct writers in the line were both found to be significant determinants of performance. The results were consistent with shortcomings in competition in some insurance lines. On a sample of forty-eight insolvent life insurers over the period 1990 to 1992, Pottier (1998) compared the predictive abilities (1) ratings, rating changes and total assets; (2) financial ratios; and (3) financial ratios combined with ratings and rating changes. Based on the expected cost of misclassification, the predictive abilities of ratings, rating changes, and total assets are comparable to financial ratios combined with ratings and rating changes. For most cost ratios, combining ratings and rating changes with financial ratios improved predictive ability compared to using financial ratios alone. Another interesting finding is that adverse rating changes are important predictors of insolvency.

BarNiv et al. (1999) illustrated a method that constructs confidence intervals for insolvency probabilities and examined various measures of the confidence intervals, such as their minimum lengths and minimum upper bounds. Lai and Limpaphayom (2003) examined the impact of organizational structure on firm performance, incentive problems, and financial decisions in the Japanese nonlife (property-casualty) insurance industry.

The remainder of this paper is organized as follows:

An introduction for clustering is presented in section 2. The methodology for clustering and fuzzy classification is presented in section 3. The proposed multivariate models for financial performance (discriminant analysis model and logistic regression model) are presented in sections 4 and 5, respectively. Section 6 provides the application that develops a multivariate model based on financial data of Egyptian property-liability insurance companies. This application reflects the efficiency of financial performance.

2. INTRODUCTION To CLUSTERING

Cormack (1971) presented a review of classification by proposing the definitions of similarity and of cluster. The principles, but no details of implementation, of the many empirical classification techniques currently in use are discussed. Limitations and short comings in their development and practice are also pointed out.

Clustering is the classification of objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters) so that the data in each subset (ideally) share some common trait, often proximity according to some defined distance measure. Data clustering is a common technique for statistical data analysis that is used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics [51]. Therefore, the cluster is a group of cases representing an access intensity within the population at risk that is unlikely to be due to chance (Alexander (1999)).

Cluster analysis is an important self-study tool for solving the problem of finding groups in data without the help of a response variable (Tibshirani et al. (2001)).

The goal of clustering or classification is to decide which of two or more populations a particular observation or set of observations belongs. In this type of problem, an error is made if an observation is assigned to any population other than the one to which it belongs. Thus, the relative value of a classification or assignment procedure can be measured in two ways: (1) assume that the underlying distributions are known and compute the probability of assigning an observation to the wrong population, or (2) use the procedure on a set of observations for which the correct assignments are known and calculate the percentage of observations that are assigned to a wrong

population (Mayer (1971)).

2.1 Common Distance Functions

An important step in any clustering is to select a distance measure that will determine how the similarity of two elements is calculated. This will influence the shape of the clusters, as some elements may be close to one another according to one distance and further away according to another.

2.1.1 Euclidean distance

Euclidean distance is the straight line distance between two points. In a plane with p_1 at (x_1, y_1) and p_2 at (x_2, y_2), it is

$$\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}.$$

The Euclidean distance function measures the distance between a point $\boldsymbol{X}$ ($\boldsymbol{X_1}$, $\boldsymbol{X_2}$, etc.) and a point $\boldsymbol{Y}$ ($\boldsymbol{Y_1}$, $\boldsymbol{Y_2}$, etc.) as

$$d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2},$$

where n is the number of variables, and $\boldsymbol{X_i}$ and $\boldsymbol{Y_i}$ are the values of the ith variable, at points $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively.

2.1.2 Manhattan distance

Manhattan distance is the distance between two points measured along axes at right angles. In a plane with p_1 at (x_1, y_1) and p_2 at (x_2, y_2), it is $|x_1 - x_2| + |y_1 - y_2|$.

The Manhattan distance function computes the distance that would be traveled to get from one data point to the other if a grid-like path is followed. The Manhattan distance between two items is the sum of the differences of their corresponding components.

The formula for this distance between a point $\boldsymbol{X}$=($\boldsymbol{X_1}$, $\boldsymbol{X_2}$, etc.) and a point $\boldsymbol{Y}$=($\boldsymbol{Y_1}$, $\boldsymbol{Y_2}$, etc.) is:

$$d = \sqrt{\sum_{i=1}^{n} |x_i - y_i|}$$

2.1.3 Mahalanobis distance

Mahalanobis distance is most commonly used as a multivariate outlier statistic. This measure is recommended for examining data profiles such as learning curves, serial position effects, and group

profiles.

The Mahalanobis distance has the advantage of utilizing group means and variances for each variable, and the correlations and covariance between measures. The Mahalanobis distance matrix algebra equation is written as follows:

$$\left(X_i - Y_i\right)' S_i^{-1}\left(X_i - Y_i\right)$$

where S^{-1} is the inverse covariance matrix.

Estimating the number of clusters represent the main problem when using the cluster analysis. Authors interested by this problem include Tibshirani et al. (2001) and Peck et al. (1989). Tibshirani et al. proposed a method (the "gap statistic") for estimating the number of clusters (groups) in a set of data. Peck et al. developed a bootstrap-based procedure for obtaining approximate confidence bounds on the number of clusters in the "best" clustering. The effectiveness of this procedure is evaluated in a simulation study. An application is presented.

2.2 Clustering Algorithms Used for Classification

2.2.1 *K*-Means Clustering

By assuming *k* clusters and defining *k* centroids, one for each cluster, this algorithm aims to minimize the objective function, which represents the squared error function in the following form:

$$S = \sum_{j=1}^{k}\sum_{i=1}^{n}\left\|x_i^{(j)} - C_j\right\|^2.$$

Where $\left\|x_i^{(j)} - C_j\right\|$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster center C_j.

2.2.2 Hierarchical Clustering

Hierarchical algorithms find successive clusters using previously established clusters, whereas partitional algorithms determine all clusters at once. Hierarchical algorithms can be agglomerative ("bottom-up") or divisive ("top-down"). Agglomerative algorithms begin with each element as a separate cluster and merge them into successively larger clusters. Divisive algorithms begin with the whole set and proceed to divide it into successively smaller clusters.

2.2.3 Fuzzy Clustering

Clustering involves the task of dividing data points into homogeneous classes or clusters so that items in the same class are as similar as possible and items in different classes are as dissimilar as possible. In non-fuzzy or hard clustering, data is divided into crisp clusters, where each data point

belongs to exactly one cluster. In fuzzy clustering, the data points can belong to more than one cluster. Associated with each of the points are membership grades that indicate the degree to which the data points belong to the different clusters.

Many authors in insurance use clustering. Some examples include Jensen (1971), who studied the financial performance of selected business firms using cluster analysis and Samson (1986), who depended on the classification system for designing an automobile insurance. Fiegenbaum and Thomas (1990) studied the performance of the U.S. insurance industry by using clustering to make the strategic groups. Yeo et al. (2001) used clustering technique for classifying risks and predicting claim costs in the automobile insurance industry. Wagstaff and Lindelow (2008) studied the effect of insurance on increased financial risk for health insurance in China.

3. FUZZY CLUSTERING

There are many text books that present the concept of fuzzy clustering and its algorithms and applications, such as Oliveira and Pedrycz (2007), Lazzerini et al. (2000), Pedrycz (2005), Sato et al. (1997), and the *NCSS Manual* (2007).

Fuzzy clustering generalizes partition clustering methods by allowing an individual to be partially classified into more than one cluster. In regular clustering, each individual is a member of only one cluster. Suppose we have K clusters and we define a set of variables, $m_{i1}, m_{i2}, \ldots, m_{ik}$, that represent the probability that object i is classified into cluster k. In partition clustering algorithms, one of these values will be one and the rest will be zero. This represents the fact that these algorithms classify an individual into one and only one cluster (Kaufman and Rousseeuw (1990)).

In fuzzy clustering the membership is spread among all clusters. The m_{ik} can now be between zero and one, with the stipulation that the sum of their values is one. We call this a fuzzification of the cluster configuration. It has the advantage that it does not force every object into a specific cluster.

The fuzzy algorithm seeks to minimize the following objective function, C, made up of cluster memberships and distances.

$$C = \sum_{L=1}^{k} \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} m_{iL}^2 m_{jL}^2 d_{ij}}{2\sum_{j=1}^{N} m_{jL}^2}$$

where m_{iL} represents the unknown membership of the object i in cluster L and d_{ij} is the dissimilarity between objects i and j.

The memberships are subject to constraints that they all must be non-negative and that the memberships for a single individual must sum to one. That is, the memberships have the same constraints that they would if they were the probabilities that an individual belongs to each group.

To test goodness of fit for fuzzy clustering, Kaufman and Rousseeuw (1990) proposed the silhouette statistic for assessing clusters and estimating the optimal number. For observation *i*, let *a*(*i*) be the average distance to other points in its cluster, and b(i) the average distance to points in the nearest cluster besides its own nearest is defined by the cluster minimizing this average distance. Then the silhouette statistic is defined by

$$S(i) = \frac{b(i) - a(i)}{Max \{a(i), b(i)\}}$$

A point is well clustered if *s*(*i*) is large. Kaufman and Rousseeuw (1990) proposed to choose the optimal number of clusters $\hat{k}$ as the value maximizing the average *s*(*i*) over the data set.

Various studies used the fuzzy clustering in insurance such as Lemaire (1990), who indicated the role of the fuzzy theory in decision making for calculating net premiums and reinsurance polices. Ebanks et al. (1992) presented how to use the measures of fuzziness to risk classification for life insurance in two phases—the first determinate is the degree of risk and the second determinate is the membership for each risk. Cummins and Derrig (1993) used the fuzzy trends in estimating the property-liability insurance claim costs. Carreno and Jani (1993) developed expert system based on fuzzy approach to insurance risk assessment.

Young (1996 and 1997) introduced the fuzzy approach for adjusting and changing the insurance rates. Verrall and Yakoubov (1998) used fuzzy clustering to grouping by policyholder age in general insurance. The fuzzy approach is used for estimating the reserves by Sanchez and Gomez (2003). Shapiro (2004) presented an overview of insurance studies that use fuzzy logic in numerous of applications for classification, underwriting, projected liabilities, fuzzy future and present values, pricing, asset allocations and cash flows, and investments. Sánchez (2006) and (2007) used the fuzzy regression to calculate insurance claim reserves.

4. DISCRIMINANT ANALYSIS

The statistical pattern recognition for discriminant analysis and how to apply it proposed by various studies as the *NCSS Manual* (2007), Huberty and Olejnik (2006), McLachlan (2004) and Huberty (1994).

Discriminant analysis finds a set of prediction equations based on independent variables used to classify individuals into groups. There are two possible objectives in a discriminant analysis: finding a predictive equation for classifying new individuals or interpreting the predictive equation to better

understand the relationships that may exist among the variables.

In many ways, discriminant analysis parallels multiple regression analysis. The main difference between these two techniques is that regression analysis deals with a continuous dependent variable, while discriminant analysis must have a discrete dependent variable. The methodology used to complete a discriminant analysis is similar to regression analysis. You plot each independent variable versus the group variable. You often go through a variable selection phase to determine which independent variables are beneficial. You conduct a residual analysis to determine the accuracy of the discriminant equations.

Discriminant analysis assumes linear relations among the independent variables. Suppose you have data for K groups, with N_k observations per group. Let N represent the total number of observations. Each observation consists of the measurements of p variables. The i^{th} observation is represented by X_{ki}. Let M represent the vector of means of these variables across all groups and M_k the vector of means of observations in the k^{th} group.

Define three sums of squares and cross products matrices, S_T, S_W, and S_A, as follows:

$$S_T = \sum_{L=1}^{K} \sum_{i=1}^{N_L} (X_{Li} - M)(X_{Li} - M)^{\backslash}$$

$$S_W = \sum_{L=1}^{K} \sum_{i=1}^{N_L} (X_{Li} - M_L)(X_L i - M_L)^{\backslash}$$

$$S_A = S_T - S_W \ .$$

A discriminant function is a weighted average of the values of the independent variables. The weights are selected so that the resulting weighted average separates the observations into the groups. High values of the average come from one group; low values of the average come from another group. The problem reduces to one of finding the weights that, when applied to the data, best discriminate among groups according to some criterion. The solution reduces to finding the eigenvectors, V, of $S_W^{-1} S_A$. The canonical coefficients are the elements of these eigenvectors.

A goodness-of-fit parameter, Wilks' Lambda, is defined as follows:

$$\Lambda = \frac{|S_W|}{|S_T|} = \prod_{j=1}^{m} \frac{1}{1+\lambda_j}$$

where is the j^{th} eigenvalue corresponding to the eigenvector described above and m is the minimum of K-1 and p.

The canonical correlation between the j^{th} discriminant function and the independent variables is related to these eigenvalues as follows:

$$r_{ci} = \sqrt{\frac{\lambda_i}{1+\lambda_i}}$$

The linear discriminant functions are defined as:

$$LDF_k = W^{-1} M_k$$

Where

$$W = \left(\frac{1}{N-K}\right) S_W$$

Many studies in insurance have used discriminant analysis. For example, Trieschmann and Pinches (1973) used multiple discriminant analysis to classify firms into two groups (solvent or distress). The model developed was able to classify correctly forty-nine out of fifty-two firms included in the study. One solvent firm was classified as being in distress while two of the distressed firms were classified as belonging to the solvent group. The six variables used to classify firms were (1) agents balances/total asset ratio, (2) stock cost (preferred and common)/stock-market ratio (preferred and common), (3) bond cost/bonds-market ratio. (4) loss adjustment expenses paid -*k* underwriting expenses paid/net premiums written ratio, (5) combined ratio, and (6) premiums written direct/surplus ratio. Trieschmann and Pinches (1974) later examined the efficiency of alternative models for solvency surveillance of property-liability insurance firms employing financial ratios. The two models they investigated were (1) financial ratios individually or in groups on a univariate basis, and (2) a set of financial ratios in a multivariate context based on a multiple discriminant model. Through the use of statistical tests, it is shown that the multiple discriminant model does a better job of identifying firms with a high probability of distress than the univariate models. Ambrose and Seward (1988) incorporated Best's ratings into the discriminant analysis through a system of dummy variates. Best's ratings are then compared to the results obtained by the use of financial variables. Finally, a two-stage discriminant technique is introduced and its results are

shown to be better for predicting insolvency for property-liability firms.

5. LOGISTIC REGRESSION

The application of logistic regression in the social sciences and its properties are proposed by many authors such as Kleinbaum et al. (2005), Menard (2001), Jaccard (2001), and Hosmer and Lemeshow (2000), as well as by the *NCSS Manual* (2007).

Logistic regression analysis studies the association between a categorical dependent variable and a set of independent (explanatory) variables. The name logistic regression is often used when the dependent variable has only two values. The name multiple-group logistic regression (MGLR) is usually reserved for the case when the dependent variable has three or more unique values.

Logistic regression competes with discriminant analysis as a method for analyzing discrete response variables. In fact, the current feeling among many statisticians is that logistic regression is more versatile and better suited for most situations than is discriminant analysis because it does not assume that the independent variables are normally distributed, as discriminant analysis does.

In multiple-group logistic regression, a discrete dependent variable Y having G unique values ($G \geq 2$) is regressed on a set of p independent variables, $X_1, X_2, \ldots, X_p$. Y represents a way of partitioning the population of interest. For example, Y may be the condition of the financial performance for the insurance company.

The logistic regression model is given by the G equations as follows:

$$\log it\,(y = g \mid X) = X\beta$$

$$X = (X_1, X_2, \ldots, X_p)$$

where

$$\beta_g = \begin{bmatrix} \beta_{g1} \\ \vdots \\ \beta_{gp} \end{bmatrix}$$

and p_g is the probability that an individual with values $X_1, X_2, \ldots, X_p$ is in group g.

$$p_g = \frac{\exp(X\beta)}{1 + \exp(X\beta)}$$

We can use the maximum likelihood method to estimate the parameters for the logistic regression (β) as follows:

Let

$$\pi_{gj} = \text{prob}(Y = g \mid X_j) = \frac{\exp(X_j \beta_g)}{\sum_{s=1}^{G} \exp(X_j \beta_g)}.$$

The likelihood function is:

$$L = \prod_{j=1}^{N} \prod_{g=1}^{G} \pi_{gj}^{\,y_{gj}}$$

Where y_{gj} is one if the j^{th} observation is in group g and zero otherwise.

The log likelihood function is given by:

$$\ln L = \sum_{j=1}^{N} \sum_{g=1}^{G} y_{gj} Ln(\pi_{gj})$$

$$= \sum_{j=1}^{N} \left[\sum_{g=1}^{G} y_{gj} X_j \beta_g - \ln\left(\sum_{g=1}^{G} \exp(X_j \beta_g) \right) \right].$$

Maximum likelihood estimates of the βs are found by finding those values that maximize this log likelihood equation. This is accomplished by calculating the partial derivatives as:

$$\frac{\partial \ln L}{\partial \beta_{ik}} = \sum_{j=1}^{N} X_{kj} (y_{ig} - \pi_{ig})$$

$$g = 2,3,\ldots G \quad \text{and} \quad k = 1,2,\ldots, p$$

Setting these equations to zero solves it. Because of the nonlinear nature of the parameters, there is no closed-form solution to these equations and they must be solved iteratively by using Newton-Raphson method.

Several studies have used the logistic regression in the area of insurance. Beirlant et al. (1992) applied logistic regression to determine different subportfolios and adjusted insurance premiums for

contracts belonging to a more or less heterogeneous portfolio, based on a representative sample from Belgian car insurance data from 1989. Devaney (1994) illustrated the usefulness of financial ratios as predictors of household insolvency by applying the logistic regression for forecasting. Ambrose and Carroll (1994) applied logistic regression analysis to matched-pair samples of life insurers and found that financial variables combined with NAIC Insurance Regulatory Information System ratios outperformed A.M. Best's recommendations in distinguishing between solvent and insolvent insurers. Combining all three types of predictors into one model provided the most accurate classification.

Steven and David (2000) analyzed the ability to forecast for NAIC and Best's company solvency measures by using logistic regression. Esteban and Jose (2001) used logistic regression for insurance risk classification. Mosley (2005) concentrated on the use of predictive modeling as logistic regression in the insurance industry. Cooper and Zheng (2007) and Zheng at al. (2007) used the logistic regression for estimating the contributions of speeding and impaired driving to insurance claim cost.

6. ApplicationS for Property-Liability Insurance Companies In Egypt

The initial data set consists of the financial ratios of property and liability insurance firms that represent the predictive variables for six insurance companies, three of which are public sector (Misr Insurance Company, Al-Chark Insurance Company and National Insurance Company). The others represent private sector companies (Suez Canal Insurance Company, El-Mohandes Insurance Company, and Delta Insurance Company). Data are derived from the annual statements of the period from 1992/1993 to 2005/2006.

6.1 Predictive Variables

Financial ratios provide a quick and relatively simple means of examining the financial condition of a business. A ratio simply expresses the relation of one figure appearing in the financial statements to some other figure appearing there or perhaps to some resource of the business. Ratios can be very helpful when comparing the financial health of different businesses. By calculating a relatively small number of ratios, it's often possible to build up a reasonably good picture of the position and performance of a business. There is no generally accepted list of ratios that can be applied to financial statements, nor is there a standard method of calculating many of them. In this paper, the predictive variables are the financial ratios for property and liability insurance firms, which are defined in the following table.

Variable	Definition
x_1	Net premiums written/surplus Measures the underwriting risk where a high ratio indicates greater underwriting exposure and more risk.
x_2	Adjusted surplus/surplus Measures the effects in the reinsurance policies and its reflection on insurer's retention rates.
x_3	Surplus/investment income Measures the ratio between the income from insurance operations and income from investments.
x_4	Net premiums written/premiums written Measures the retention policy of insurer.
x_5	Return on investments Measures the efficiency of insurer's investment portfolio.
x_6	Net profit/total assets Expresses the relationship between the net profit the business generates and total assets in the balance sheet.
x_7	Net profit/surplus Reflects the returns of an insurance company and the effect of its investments and underwriting policies.
x_8	Total liabilities/liquid assets Measures the ability of an insurer to have sufficient liquid recourses available to meet maturing obligations.
x_9	Debtors and sundry debtors/surplus Measures an insurer's ability to pay its policyholders' claims and to meet its other financial obligations.
x_{10}	Debtors and sundry debtors/total assets Describes the proportion of other people's money to the total claims against the assets of the business. The higher the ratio, the greater the likely risk for lender.
x_{11}	Reserves/surplus Measures the ability of an insurer to estimate reserves.
x_{12}	(Reserves + Surplus)/Net premiums written Measures the insurer's ability to cover the exposed money to risk and it's the underwriting policy.
x_{13}	Reserves/liquid assets Measures the ability of insurers to own assets that are easily converted to cash and to pay its liabilities on time.
x_{14}	Total liabilities/total assets Measures the portion of assets that are financed by others; its complement is the equity ratio.

Variable	Definition
x_{15}	Underwriting expenses paid/premiums written Measures an insurer's fund flow from insurance operations.
x_{16}	Commissions paid/premiums written Measures an insurer's fund flow from production operations.
x_{17}	Underwriting expenses paid total/premiums written Measures an insurer's fund flow from underwriting and production operations.
x_{18}	Loss ratio Measures the ratio between the premiums paid to an insurance company and the claims settled by the company.
x_{19}	Reinsurance ceded premiums/direct premiums written Measures an insurer's reinsurance policy and its acceptable appetite to cover different risks.
x_{20}	Cash/adjusted surplus This variable examines the ratio of adjusted surplus to convert to cash which help in planning for liquidity insurer's policy.
x_{21}	Total liabilities/adjusted surplus Measures the contribution of the insurer's adjusted surplus used to cover its total liabilities.
x_{22}	Net premiums written/total assets Examines the effectiveness of the assets a business employs in retaining premiums.
x_{23}	Direct premiums written/surplus Measures the underwriting risk for direct insurance operations.
x_{24}	Claims paid/liquid assets Compares liquid assets of a business with the current liabilities.
x_{25}	Claims paid/reserves Measures adequate the estimated reserves the meet claims paid.

Where

liquid assets = cash + deposits + reserves

Adjusted surplus = surplus – (Ceded reinsurance unearned premium × (Reinsurance ceded/ premiums written ceded)).

To test of the effect of an insurance company's ownership type (public vs. private) on the efficiency of financial performance, we need to know if the public and private sectors have the same strategies for developing their financial performance efficiencies or if their strategies are different. Thus, we will test the following hypotheses:

H_0: There is no difference between the mean of efficiency of financial performance ratios in the

public sector and private sector.

H_1: There is a difference.

Table 1--The *t*-test for financial performance efficiency ratios in Egyptian insurance sectors

Ratios	*t*	Sig.
x_1	-4.360	0.000
x_2	6.245	0.000
x_3	-2.516	0.014
x_4	4.366	0.000
x_5	**1.356**	**0.179**
x_6	**-0.995**	**0.322**
x_7	**0.732**	**0.466**
x_8	-3.149	0.002
x_9	-3.650	0.001
x_{10}	-6.572	0.000
x_{11}	5.465	0.000
x_{12}	12.043	0.000
x_{13}	5.331	0.000
x_{14}	**-1.052**	**0.296**
x_{15}	**-1.963**	**0.053**
x_{16}	-10.089	0.000
x_{17}	-6.923	0.000
x_{18}	2.036	0.045
x_{19}	-2.635	0.010
x_{20}	-3.893	0.000
x_{21}	-2.490	0.015
x_{22}	-6.319	0.000
x_{23}	-5.291	0.000
x_{24}	-3.259	0.002
x_{25}	-2.820	0.006

From Table 1, we can conclude that there is not a significance difference between the mean for efficiency of financial performance ratios in the public sector and the private sector for the following ratios:

Return on investments

Net profit/total assets

Net profit/surplus

Total liabilities/Total assets

Underwriting expenses paid/premiums written

6.2 Factor Analysis

The main applications of factor analytic techniques are to reduce the number of variables and detect structure in the relationships between variables, that is, to classify variables. Therefore, factor analysis is applied as a data reduction or structure detection method.

We used the statistical package SPSS to perform a factor analysis on efficiency and financial performance data, consisting of 25 ratios. Our results concluded that the data can be reduced to six factors when applying extraction method principle component analysis and Varimax with Kaiser Normalization rotation method, the most popular orthogonal technique.

Table 2 shows the rotated component and component score coefficient for each variable with related factor and its percent of variance.

Table 2—Rotated Component and Component Score Coefficient for Each Variable with Related Factors

Factor	% of Variance	Variable	Rotated Component	Component Score Coefficient
F1	19.430	x_{19}	0.934	0.293
		x_4	-0.857	-0.269
		x_2	-0.777	-0.164
		x_{21}	0.707	0.134
		x_{23}	0.662	0.099
		x_{20}	0.626	0.166
F2	18.568	x_{22}	0.931	0.286
		x_{24}	0.889	0.303
		x_{12}	-0.708	-0.136
		x_1	0.630	0.082
F3	18.335	x_{11}	0.899	0.269
		x_{13}	0.820	0.225
		x_3	-0.775	-0.199
		x_{14}	0.628	0.127
		x_8	0.572	0.123
		x_5	0.551	0.076
F4	14.268	x_{17}	0.891	0.301
		x_{16}	0.844	0.267
		x_{15}	0.661	0.248
		x_{10}	0.625	0.194
		x_9	0.566	0.244
F5	8.458	x_6	0.934	0.459
		x_7	0.882	0.427
		x_{18}	-0.562	-0.261
F6	4.741	x_{25}	-0.698	-0.583

We can estimate each factor's predictive value using regression analysis through the score coefficient of each variable belonging to its factor, which is a method we will use in the following section to cluster our cases according to their financial performance efficiencies.

6.3 Fuzzy Clustering

When distance is measured in Euclidean distance, we used NCSS (Number Cruncher Statistical Software) to determine our data's optimal number for fuzzy clustering. Table 3 shows our results.

Table 3—The Numbers of Fuzzy Clustering and Its Silhouette Statistic

Number of clusters	s(i)
2	0.161224
3	0.186804
4	0.049518
5	-1.00000

Table 3 shows that the value maximizing the average s(i) over the data set is 0.186804 at 3, which is the optimal number of clusters.

Therefore, we can organize our data into three clusters that can be classified according to its medoids cluster into three levels of efficiency for financial performance (low, moderate, and high), as shown in the Table 4.

Table 4—The Number and Percent of Cases According To Fuzzy Classification for Efficiency of Financial Performance

Efficiency of financial performance	No.	%
Low	29	34.5
Moderate	23	27.4
High	32	38.1
Total	**84**	**100**

Table 5 represents the cross-tabulation between the ownership type of the insurance company (public or private) and the fuzzy classification for efficiency of financial performance.

Table 5—Number and Percentage of Cases Based on Fuzzy Classification for Efficiency of Financial Performance and the Insurance Company's Ownership Type

Sector	Efficiency of Financial Performance						Total
	Low		Moderate		High		
	No.	%	No.	%	No.	%	
Public	28	66.7	2	4.8	12	28.6	42
Private	1	2.4	21	50	20	47.6	42
Total	**29**	**34.5**	**23**	**27.4**	**32**	**38.1**	**84**

Table 5 shows that 66.7% of public sector cases lie in the low-efficiency cluster of financial performance, 4.8% lie in the moderate-efficiency cluster, and 28.6% are in the high-efficiency cluster. Private sector cases are composed of 50% moderate-efficiency clusters and 47.6% high-efficiency clusters of financial performance.

To test of the impact of the insurance company's ownership type (public or private) on the fuzzy classification for financial performance efficiency, we applied the Chi-square test with the following hypotheses:

H_0: There is no relationship between the fuzzy classification for financial performance efficiency and the insurance company's ownership type.

H_1: There is a relationship.

	Value	df	Sig.
Person Chi-Square	42.834	2	0.000

The Sig. of the test is 0.000, which is less than 0.05; therefore, there is a relationship between the fuzzy classification for financial performance efficiency and the insurance company's ownership type.

Table 6 cross-tabulates the insurance companies with the fuzzy classification for efficiency of financial performance.

Table 6—Frequency and Percentage of Cases Based on Fuzzy Classification For Efficiency of Financial Performance and the Insurance Companies

Insurance Company	Efficiency of Financial Performance						Total
	Low		Moderate		High		
	No.	%	No.	%	No.	%	
Misr	13	92.9	1	7.1			14
Al-Chark	6	42.9	1	7.1	7	50	14
National	9	64.3			5	35.7	14
Suez Canal			14	100			14
El-Mohandes			4	28.6	10	71.4	14
Delta	1	7.1	3	21.4	10	71.4	14
Total	**29**	**34.5**	**23**	**27.4**	**32**	**38.1**	**84**

Table 6 shows that the best companies with the highest financial performance efficiencies are El-Mohandes and Delta. Both companies have 71.4% of its cases in high-efficiency clusters of financial performance. Misr has the worst showing with 92.9% of its cases in the low-efficiency cluster.

To test insurance companies and the fuzzy classification for the efficiency of financial performance, we applied the Chi-square test with the following hypotheses:

H_0: There is no relationship between the fuzzy classification for the efficiency of financial performance and insurance companies.

H_1: There is a relationship.

	Value	df	Sig.
Person Chi-Square	84.928	10	0.000

The Sig. of Chi-square test is 0.000, which is less than 0.05; therefore, there is a relationship between the fuzzy classification for efficiency of financial performance and insurance companies.

6.4 Multiple Discriminant Analysis

When the dependent variable has more than two values, there will be more than one regression equation. In fact, the number of regression equations is equal to one less than the number of values.

Our dependent variable is categorical and we have three fuzzy clusters; therefore, there are two dependent variables LDF_1 and LDF_2.

The independent variables represent the six factors (f_1, f_2, f_3, f_4, f_5, f_6) resulting from factor analysis for the efficiency of financial performance ratios.

Therefore, if we have three fuzzy clusters for efficiency of financial performance of the Egyptian insurance companies, then the two discriminant functions are:

$$LDF_1 = 0.580\ f_1 + 1.007\ f_2 + 0.946\ f_3 - 0.460\ f_4 + 0.195\ f_5 - 0.466\ f_6$$

$$LDF_2 = 0.675\ f_1 + 0.545\ f_2 - 0.549\ f_3 + 1.017\ f_4 + 0.412\ f_5 - 0.195\ f_6$$

6.4.1 A Goodness of Fit for Discriminant Analysis Model

There are many measures to determining the efficiency of the discriminant analysis model.

6.4.1.1 Wilks' Lambda

This measure represents the percentage of the sum of squares within the group and the total sum of squares, which are valued between zero and one. If the percentage is close to one, there is no difference between groups; alternatively, if the percentage is close to zero, there is a difference between the groups.

The value of Wilks' Lambda statistic for the two discriminant functions are 0.067 and 0.278 respectively, which indicates that there are differences among the groups of the fuzzy clusters for efficiency of financial performance.

6.4.1.2 Chi-Square Test

This measure is used to test the significance of the discriminant analysis model as follows:

Discriminant Analysis Model	Chi-Square Value	df	Sig.
LDF_1	212.554	12	0.000
LDF_2	100.380	5	0.000

The Sig. of the Chi-square test is 0.000, which is less than 0.05, thus indicating the significance of the discriminant analysis model and its dependability in predicting the classifications of financial performance efficiencies for the Egyptian insurance companies.

6.4.1.3 Canonical Correlation

This measure refers to the correlation between the value of the discriminant function and the

independent variables in the function. For the first discriminant function, LDF_1, the canonical correlation coefficient value is 0.872; the value for the second discriminant function, LDF_2, is 0.849. Thus, there is a strong relationship between the discriminant functions and the independent variables.

6.4.1.4 Percentage of Variance

Using this measure, the first discriminant function yields 55% of variance and the second discriminant function yields 45%. Thus, the independent variables in the two discriminant functions can explain the variability in discriminant scores by 100%.

6.4.1.5 Percentage of Correct Classification

This measure applies the proposed discriminant model on our data and classifies the cases into three categories of low-, moderate-, or high-efficiency of financial performance. Table 7 yields the following results:

Table 7—Classification Results of Discriminant Model for Efficiency of Financial Performance

Original Group Membership	Predicted Group Membership						Total	
	Low		*Moderate*		*High*			
	No.	**%**	**No.**	**%**	**No.**	**%**	**No.**	**%**
Low	28	96.6	0	0	1	3.4	29	100
Moderate	1	4.3	22	95.7	0	0	23	100
High	0	0	1	3.1	31	96.9	32	100

From Table 7, we can conclude that the discriminant model for predicting the efficiency of financial performance for Egyptian insurance companies is correctly classified by 96.4%.

6.5 The Logistic Regression Model

Using the logistic regression model as a tool for predicting the financial performance efficiency of property and liability Egyptian insurance companies, yields the following results:

The estimated logistic regression models are:

$$\log it(y_2) = 0.3043 + 2.9716\ f_1 + 2.198\ f_2 - 0.9773\ f_3 + 4.2337\ f_4 + 1.323\ f_5 - 1.476\ f_6$$

$$\log it(y_3) = 0.2124 + 1.508\ f_1 - 1.5203\ f_2 - 6.311\ f_3 + 6.798\ f_4 + 0.851\ f_5 - 0.894\ f_6$$

The two equations above can be used to predict the probability that an individual company belongs to each group of financial performance efficiency (low, moderate, high). Whereas the probability that an insurance company belongs to moderate-efficiency cluster of financial performance can be written as

$$p(y = 2 \mid X) = \frac{1}{(1 + \exp(\log\ it(y_2)))}$$

and the probability of an insurance company belongs to high-efficiency cluster of financial performance can be written as

$$p(y = 3 \mid X) = \frac{1}{(1 + \exp(\log\ it(y_3)))}.$$

If the case does not belong to the moderate- or high-efficiency clusters of financial performance, then it belongs to the low-efficiency cluster.

6.5.1 A Goodness of Fit for Logistic Regression Model

6.5.1.1 Likelihood Ratio Test

This is the test of choice in logistic regression. The likelihood ratio test statistic is -2 times the difference between the log likelihoods of two models. The -2 log likelihood ratio test is approximately the chi- square distribution. Therefore, we have the following:

	Value	**Sig.**
Person Chi-Square	15.180	0.000

The Sig. of the Chi-square test is 0.000, which less than 0.05, which in turn indicates the significance of logistic regression as predictive model for determining the financial performance efficiency of the Egyptian insurance companies.

6.5.1.2 *R*-Square

The *R*-square of logistic regression model is 0.9567, which means that the independent variables $(f_1, f_2, f_3, f_4, f_5, f_6)$ in the logistic regression model can interpret the changes in the efficiency of financial performance of the Egyptian insurance companies by 95.67%.

6.5.1.3 Percent of Correctly Classification

This measure applies the proposed logistic regression model to our data and classifies the cases into three categories (low, moderate, high) of financial performance efficiencies. Table 8 shows our results.

Table 8—Classification Results of Logistic Regression Model for Efficiency of Financial Performance

Original Group Membership	Predicted Group Membership						Total	
	Low		*Moderate*		*High*			
	No.	%	No.	%	No.	%	No.	%
Low	29	100	0	0	0	0	29	100
Moderate	0	0	23	100	0	0	23	100
High	0	0	0	0	32	100	32	100

From Table 8, we can conclude that the logistic regression model predicts the of financial performance efficiencies for Egyptian insurance companies correctly 100% of the time.

Conclusions

Comparing the public and private sectors, the mean of efficiency of financial performance ratios does not vary significantly for the following ratios: return on investments, net profit to total assets, net profit to surplus, total liabilities to total assets, and underwriting expenses paid to premiums written.

Predictive variables, which represent 25 ratios measuring efficiency and financial performance data, can be reduced into six factors by using factor analysis.

Fuzzy cluster procedure indicates that the data involving the efficiency of financial performance for the Egyptian insurance companies can be classified into three groups: low, moderate, and high.

The number of financial performance cases exhibiting high efficiency is 32, or 38.1% of the total cases; the number of moderate-efficiency cases is 23, or 27.4 of total cases; and the number of low-efficiency cases represent 34.5% of total cases.

Public sector cases represent 66.7% of the low-efficiency clusters of financial performance, while private sector cases comprise 47.6% of high-efficiency clusters for financial performance. Thus, there is a relationship between the fuzzy classification of the insurance company's financial performance efficiency and its ownership type.

The best companies in efficiency of financial performance are El-Mohandes and Delta, while the worst company is Misr. There is relationship between the fuzzy classification for efficiency of financial performance and insurance companies.

Wilks' Lambda and Chi-square test indicate that discriminant analysis model is significant in determining the efficiency of financial performance of the Egyptian insurance companies.

The canonical correlation coefficients for the two discriminant functions are 0.872 and 0.849, which reflect the strong relationship between the independent variables in the model and the

efficiency of financial performance.

The discriminant analysis model can predict a correct classification for the efficiency of financial performance of Egyptian insurance companies by 96.4%.

Likelihood ratio test indicates that the logistic regression model is significant for determining the efficiency of financial performance of Egyptian insurance companies.

The independent variables in the logistic regression model can interpret the efficiency of financial performance of Egyptian insurance companies by 95.67%.

The logistic regression model can predict a correct classification for the efficiency of financial performance of Egyptian insurance companies by 100%.

The results indicate that the logistic regression model can more accurately forecast the financial performance efficiencies of Egyptian insurance companies than the discriminant analysis model.

Acknowledgment

The author would like to express his appreciation to Casualty Actuarial Society and the reviewers of his submission.

References

[1] Alexander, F. E. 1999. "Cluster and clustering of childhood cancer: A review." *European Journal of Epidemiology* 15:847-852.

[2] Ambrose, Jan M. and Anne M. Carroll. 1994. "Using Best's ratings in life insurer insolvency prediction." *Journal of Risk and Insurance*. 61, no. 2:317-327.

[3] Ambrose, Jan M. and Seward J. Allen. 1988. "Best's ratings, financial ratios and prior probabilities in insolvency prediction." *Journal of Risk and Insurance* 55, no. 2:229-244.

[4] BarNiv, Ran and Hershbarger Robert A. 1990. "Classifying financial distress in the life insurance industry." *Journal of Risk and Insurance* 57, no. 1:110-136.

[5] BarNiv, Ran and McDonald James B. 1992. "Identifying financial distress in the insurance industry: A synthesis of methodological and empirical issues." *Journal of Risk and Insurance* 59, no. 4:543-573.

[6] BarNiv, Ran, John Hathorn, Abraham Mehrez, and Douglas Kline. 1999. "Confidence intervals for the probability of insolvency in the insurance industry." *Journal of Risk and Insurance* 66, no. 1:125-137.

[7] Beaver, William H. 1966. "Financial ratios as predictors of failure." *Journal of Accounting Research* 4:71-111.

[8] Beirlant, J., V. Derveaux, M. De Meyer, J. Goovaerts, E. Labie and B. Maenhoudt 1992, "Statistical risk evaluation applied to (Belgian) car insurance." *Insurance: Mathematics and Economics* 10, no. 4:289-302.

[9] Carreno, L. and Y. Jani. 1993. "A fuzzy expert system approach to insurance risk assessment using Fuzzy Clips." In *WESCON/1993 Conference Record:* 536-541

[10] Chidambaran, N. K., Thomas A. Pugel, and Anthony Saunders. 1997. "An investigation of the performance of the U.S. property-liability insurance industry." *Journal of Risk and Insurance* 64, no. 2:371-382.

[11] Cooper, Peter J. and Yvonne Y. Zheng. 2007. "Estimating the contributions of speeding and impaired driving to insurance claim cost." *Journal of Safety Research* 38, no. 1:17-23.

[12] Cormack, R. M. 1971. "A review of classification." *Journal of the Royal Statistical Society*. Series A (General) 134, no. 3:321-367.

[13] Cummins, J. David and Richard A. Derrig. 1993. "Fuzzy trends in property-liability insurance claim costs." *Journal of Risk and Insurance* 60, no. 3:429-465.

[14] Devaney, Sharon A. 1994. "The usefulness of financial ratios as predictors of household insolvency: Two perspectives." *Financial Counseling and Planning* 5:5-24.

[15] Ebanks, B., W. Kanvowski, and K. Ostaszewski. 1992. "Application of measures of fuzziness to risk classification in insurance." In Proceedings of the Fourth International Conference on Computing and Information, 290-291. Los Alamitos, California: IEEE Computer Society Press.

[16] Esteban, F. and G. Jose. 2001. "Robust logistic regression for insurance risk classification." Business

Economics Working Papers wb016413, Universidad Carlos III, Departamento de Economía de la Empresa.
[17] Fiegenbaum, A. and H. Thomas. 1990. "Strategic groups and performance: The U.S. insurance industry, 1970-84." Strategic Management Journal 11, no. 3:197-215.
[18] Harrington, Scott E. and Jack M. Nelson. 1986. "Regression-based methodology for solvency surveillance in the property-liability insurance industry." *Journal of Risk and Insurance* 53, no. 4:583-605.
[19] Hosmer, D. W. and S. Lemeshow. *Applied Logistic Regression.* 2000. 2nd ed. New York: Wiley-Interscience/Wiley Series in Probability and Statistics.
[20] Huberty, C. J. 1994. *Applied Discriminant Analysis.* New York: Wiley-Interscience/Wiley Series in Probability and Statistics.
[21] Huberty, C. J. and S. Olejnik. 2006. *Applied MANOVA and Discriminant Analysis.* New York: Wiley-Interscience/Wiley Series in Probability and Statistics.
[22] Jaccard, J. J. 2001. *Interaction Effects in Logistic Regression.* Series: Quantitative Applications in the Social Sciences. Thousand Oaks, California: Sage Publications, Inc.
[23] Jensen, R. 1971. "A cluster analysis study of financial performance of selected business firms." *The Accounting Review* 46, no. 1:36-56.
[24] Kaufman, L. and P. Rousseeuw. 1990. "Finding groups in data: an introduction to cluster analysis." New York: Wiley-Interscience/Wiley Series in Probability and Statistics.
[25] Kleinbaum, D. G., M. Klein, and E. Rihl Pryor. 2005. *Logistic Regression.* 2nd ed. New York: Springer.
[26] Lai, Gene C. and Piman Limpaphayom. 2003. "Organizational Structure and Performance: Evidence from the Nonlife Insurance Industry in Japan." *Journal of Risk and Insurance* 70, no. 4:735-757.
[27] Lazzerini, B., D. Dumitrescu, and L. C. Jain. 2000. *Fuzzy Sets & their Application to Clustering & Training.* Boca Raton, Florida: CRC Press International Series on Computational Intelligence.
[28] Lee, Suk Hun and Jorge L. Urrutia. 1996. "Analysis and prediction of insolvency in the property-liability insurance industry: A comparison of logit and hazard models." *Journal of Risk and Insurance* 63, no. 1:121-130.
[29] Lemaire, Jean. 1990. "Fuzzy insurance." *ASTIN Bulletin* 20, no.1:33-56.
[30] McLachlan, G. J. 2004. *Discriminant Analysis and Statistical Pattern Recognition.* New York: Wiley-Interscience/Wiley Series in Probability and Statistics.
[31] Menard, S. W. 2001. "Applied Logistic Regression Analysis (Quantitative Applications in the Social Sciences). 2nd ed. Thousand Oaks, California: Sage Publications, Inc.
[32] Mosley, Roosevelt. 2005. *The Use of Predictive Modeling in the Insurance Industry.* Bloomington, Indiana: Pinnacle Actuarial Resources, Inc.
[33] NCSS Statistical & Power Analysis Software. *NCSS Manual.* 2007. http://www.ncss.com.
[34] Oliveira, J. V. and W. Pedrycz, eds. 2007. *Advances in Fuzzy Clustering and its Applications.* New York: John Wiley and Sons.
[35] Peck, R., Lloyd Fisher, and John Van Ness. 1989. "Approximate confidence intervals for the number of clusters." *Journal of the American Statistical Association* 84, no. 405:184-191.
[36] Pedrycz, W. 2005. *Knowledge-Based Clustering: From Data to Information Granules.* New York: John Wiley and Sons/Wiley-Interscience.
[37] Pottier, Steven W. 1998. "Life insurer financial distress, Best's ratings and financial ratios." *Journal of Risk and Insurance* 65, no. 2:275-288.
[38] Samson, Danny. 1986. "Designing an automobile insurance classification system." *European Journal of Operational Research* 27, no. 2:235-241.
[39] Sánchez, J. and A. Gomez. 2003. "Applications of fuzzy regression in actuarial analysis." *Journal of Risk and Insurance* 70, no. 4:665-699.
[40] Sánchez, Jorge de Andrés. 2006. "Calculating insurance claim reserves with fuzzy regression." *Fuzzy Sets and Systems* 157, no. 23:3091-3108.
[41] Sánchez, Jorge de Andrés. 2007. "Claim reserving with fuzzy regression and Taylor's geometric separation method." *Insurance: Mathematics and Economics* 40, no. 1:145-163.
[42] Sato, M., Y. Sato, and L. C. Jain. 1997. *Fuzzy Clustering Models and Applications.* Heidelberg: Physica-Verlag.
[43] Shapiro, Arnold F. 2004. "Fuzzy logic in insurance." *Insurance: Mathematics and Economics* 35, no. 2:399-424.
[44] Steven, W. and W. David. 2000. "Capital ratios and property insurer insolvencies." *Journal of Risk and Insurance* 70.
[45] Tibshirani, R., Walther G. Guenther and T. Hastie. 2001. "Estimating the number of clusters in a data set via the gap statistic." *Journal of the Royal Statistical Society,* Series B (Statistical Methodology) 63, no. 2:411-423.
[46] Trieschmann, James S. and George E. Pinches. 1973. "A multivariate model for predicting financially distressed P-L insurers." *Journal of Risk and Insurance* 40, no. 3:327-338.
[47] Trieschmann, James S. and George E. Pinches. 1974. "The efficiency of alternative models for solvency

surveillance in the insurance industry." *Journal of Risk and Insurance* 41, no. 4:563-577.

[48] Van Gestel, Tony, Martens David, Baesens Bart, Feremans Daniel, Huysmans Johan and Vanthienen Jan. 2007. "Forecasting and analyzing insurance companies' ratings." *International Journal of Forecasting* 23, no. 3:513-529.

[49] Verrall, R. and Y.H. Yakoubov. 1998. "A fuzzy approach to grouping by policyholder age in general insurance." Actuarial Research Paper no. 104, Department of Actuarial Science and Statistics, City University, London.

[50] Wagstaff, A. and M. Lindelow. 2008. "Can insurance increase financial risk? The curious case of health insurance in China." *Journal of Health Economics* (forthcoming) (accepted manuscript, available online 9 February 2008).

[51] Wikipedia, the Free Encyclopedia. "Cluster Analysis." http://en.wikipedia.org/wiki/Data_clustering (accessed 2007-2008).

[52] Yeo, A., K. Smith, J. Robert, R. Willis and M. Brooks. 2001. "Clustering technique for risk classification and prediction of claim costs in the automobile insurance industry." *International Journal of Intelligent Systems in Accounting, Finance* & *Management* 10, no. 1:39-50.

[53] Young, Virginia R. 1996. "Insurance rate changing: A fuzzy logic approach." *Journal of Risk and Insurance* 63, no. 3:461-484.

[54] Young, Virginia R. 1997. "Adjusting indicated insurance rates: fuzzy rules that consider both experience and auxiliary data." *Proceedings of the Casualty Actuarial Society* 84:734-765.

[55] Zheng, Yvonne Y., Peter J. Cooper, and C.B. Dean. 2007. "Modeling the contribution of speeding and impaired driving to insurance claim counts and costs when contributing factors are unknown." *Journal of Safety Research* 38, no. 1,:25-33.

Biography of the Author

Osama Hanafy Mahmoud is an assistant professor in the department of mathematics, statistics, and insurance at the Sadat Academy for Management Sciences in Egypt. He is also an assistant professor in the department of statistics and quantitative methods in the College of Management Sciences & Planning at King Faisal University in Saudi Arabia. Professor Mahmoud can be reached at K.F.U, PO Box 1760, Al Hofuf, Al HASSA 31982, Saudi Arabia; (E- Mail) Oshanafy@hotmail.com.

Principle Component Analysis and Partial Least Squares: Two Dimension Reduction Techniques for Regression

Saikat Maitra and Jun Yan

Abstract: Dimension reduction is one of the major tasks for multivariate analysis, it is especially critical for multivariate regressions in many P&C insurance-related applications. In this paper, we'll present two methodologies, principle component analysis (PCA) and partial least squares (PLC), for dimension reduction in a case that the independent variables used in a regression are highly correlated. PCA, as a dimension reduction methodology, is applied without the consideration of the correlation between the dependent variable and the independent variables, while PLS is applied based on the correlation. Therefore, we call PCA as an unsupervised dimension reduction methodology, and call PLS as a supervised dimension reduction methodology. We'll describe the algorithms of PCA and PLS, and compare their performances in multivariate regressions using simulated data.

Key Words: PCA, PLS, SAS, GLM, Regression, Variance-Covariance Matrix, Jordan Decomposition, Eigen Value, Eigen Factors.

Introduction

In large-scale data mining and predictive modeling, especially for multivariate regression exercises, we often start with a large number of possible explanatory/predictive variables. Therefore, variable selection and dimension reduction is a major task for multivariate statistical analysis, especially for multivariate regressions. A well-known method in regression analysis for dimension reduction is called stepwise regression algorithm, which is covered by many statistical softwares such as SAS and SPSS. One of the major limitations of the algorithm is that when several of the predictive variables are highly correlated, the tests of statistical significance that the stepwise method is based on are not sound, as independence is one of the primary assumptions of these tests.

Often, many variables used as independent variables in a regression display a high degree of correlation, because those variables might be measuring the same characteristics. For example, demographic variables measuring population density characteristics or weather characteristics are often highly correlated.

A high degree of correlation among the predictive variables increases the variance in estimates of the regression parameters. This problem is known as multi-colinearity in regression literature (Kleinbaum et al. [4]). The parameter estimates in a regression equation may change with a slight change in data and hence are not stable for predicting the future.

In this paper, we will describe two methodologies, principle component analysis (PCA) and

partial least square (PLS), for dimension reduction in regression analysis when some of the independent variables are correlated. We'll describe what algorithm is used in each methodology and what the major differences are between the two methodologies.

Principal Component Analysis

PCA is a traditional multivariate statistical method commonly used to reduce the number of predictive variables and solve the multi-colinearity problem (Bair et al. [3]). Principal component analysis looks for a *few* linear combinations of the variables that can be used to summarize the data without losing too much information in the process. This method of dimension reduction is also known as "parsimonious summarization" (Rosipal and Krämer [6]) of the data. We will now formally define and describe principal components and how they can be derived. In the process we introduce a few terms for the sake of completeness.

Data Matrix

Let $Xn\times p$ denote the matrix of predictive variables (henceforth referred to as *data-matrix*), where each row denotes an observation on p different predictive variables, $X1, X2, \ldots, Xp$. We will denote a random observation from this matrix by $x1\times p$. The problem at hand is to select a subset of the above columns that holds most of the information.

Variance-Covariance Matrix

Let σ_{ij} denote the co-variance between X_i and X_j in the above data-matrix. We will denote the matrix of $((\sigma_{ij}))$ by $\sum$. Note that the diagonal elements of $\sum$ are the variances of X_i. In actual calculations σ_{ij}s may be estimated by their sample counterpart's s_{ij} or sample covariance calculated from the data. The matrix of standard deviations $((s_{ij}))$ will be denoted by S. Note both $\sum$ and S are $p\times p$ square and symmetric matrices.

Linear Combination

A linear combination of a set of vectors $(X1, X2, \ldots, Xp)$ is an expression of the type $\sum \alpha iXi$ (i=1 to p) and αis are scalars. A linear combination is said to be normalized or standardized if $\sum|\alpha i|=1$ (sum of absolute values). In the rest of the article, we will refer to the standardized linear combination as SLC.

Linear Independence

A set of vectors are said to be linearly independent if none of them can be written as a linear

combination of any other vectors in the set. In other words, a set of vectors $(X_1, X_2, \ldots, X_p)$ is linearly independent if the expression $\sum \alpha_i X_i = 0 \rightarrow \alpha_i = 0$ for all values of i. A set of vectors not linearly independent is said to be linearly dependent.

Statistically, correlation is a measure of linear dependence among variables and presence of highly correlated variables indicate a linear dependence among the variables.

Rank of a Matrix

Rank of a matrix denotes the maximum number of linearly independent rows or columns of a matrix. As our data-matrix will contain many correlated variables that we seek to reduce, rank of the data-matrix, $X_{n\times p}$, is less than or equal to p.

Jordan Decomposition of a Matrix

Jordan decomposition or spectral decomposition of a symmetric matrix is formally defined as follows.

Any symmetric matrix $A_{p\times p}$ can be written as $A=\Gamma\Lambda\Gamma^T=\sum\lambda_i\gamma'_{(i)}\gamma_{(i)}$ where $\Lambda_{p\times p}$ is a diagonal matrix with all elements 0 except the diagonal elements and $\Gamma_{p\times p}$ is an orthonormal matrix, i.e., $\Gamma\ \Gamma'=I$ (identity matrix).

The diagonal elements of Λ are denoted by λ_i (i=1 to p) and the columns of Γ are denoted by $\gamma_{(i)}$ (i=1 to p). In matrix algebra, λ_is are called eigen values of A and $\gamma_{(i)}$s are the corresponding eigen vectors.

If A is not a full rank matrix, i.e., rank(A) = $r < p$, then there are only r non-zero eigen values in the above Jordan decomposition, with the rest of the eigen values being equal to 0.

Principal Components

In principal component analysis, we try to arrive at a suitable SLC of the data-matrix X based on the Jordan decomposition of the variance-covariance matrix $\sum$ of X or equivalently based on the correlation matrix Φ of X. We denote the mean of the observations as $\mu_{1\times p}$.

Let $x_{1\times p}=(x_1,x_2,\ldots,x_p)$ denote a random vector observation in the data-matrix (i.e., transpose of any row of the $n\times p$ data matrix), with mean $\mu_{1\times p}$ and covariance matrix $\sum$. A principal component is a transformation of the form $x_{1\times p} \rightarrow y_{1\times p} = (x\text{-}\mu)_{1\times p}\ \Gamma_{p\times p}$, where Γ is obtained from the Jordan decomposition of $\sum$, i.e., $\Gamma^T\sum\ \Gamma = \Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)$ with λ_is being the eigen values of the decomposition. Each element of $y_{1\times p}$ is a linear combination of the elements of $x_{1\times p}$. Also each

element of y is independent of the other.

Thus we obtain p independent principal components corresponding to the p eigen values of the Jordan decomposition of $\sum$. Generally, we will only use the first few of these principal components for a regression. In the next section, we will list the major properties of the principal components as obtained above. This will help us to understand why the first few of the principal components may hold the majority of the information and thus help us reduce the dimension in a regression without losing too much information.

Properties of Principal Components (Anderson [2])

The following result justifies the use of PCA as a valid variable reduction technique in regression problems, where a first few of the principal components are used as predictive variables.

Let x be a random p dimensional vector with mean μ and covariance matrix $\sum$. Let y be the vector of principal components as defined above. Then the following holds true.

(i) $E(y_i) = 0$

(ii) $\text{Var}(y_i) = \lambda_i$

(iii) $\text{Cov}(y_i, y_j) = 0$

(iv) $\text{Var}(y_1) \geq \text{Var}(y_2) \geq \ldots \geq \text{Var}(y_p)$

(v) No SLC of x has variance larger than λ_1, the variance of the first principal component.

(vi) If $z = \sum \alpha_i x_i$ be a SLC of x, which is uncorrelated with first k principal components, then variance of z is maximized if z equals the $(k+1)^{th}$ principal component.

Item (iii) above justifies why using principal components instead of the raw predictive variables will remove the problem of multi-colinearity.

Items (iv), (v), and (vi) indicate that principal components successively capture the maximum of the variance of x and that there is no SLC that can capture maximum variance without being one of the principal components. When there is high degree of correlation among the original predictive variables, only the first few of the principal components are likely to capture majority of the variance of the original predictive variables. The magnitude of λ_is provides the measure of variance captured by the principal components and should be used to select the first few components for a regression.

A Numerical Example of PCA

In this section we describe the process of building principle components in a multivariate regression set up using a simulated data for line of business of business owners policies (BOP). The simulated data has been used for the 2006 and 2007 CAS Limited Attendance Predictive Modeling Seminars.

Description of the Data

The simulated data is in a policy-year level. That means each data record contains information of a twelve-month BOP policy. In the data, we simulated claim frequency, claim count over per $000 premium, and six correlated policy variables.

The policy variables in this example are:

fireProt – Fire Protection Class

numBldg – Number of Building in Policy

numLoc – Number of Locations in Policy

bldgAge – Maximum Building Age

bldgContents – Building Coverage Indicator

polage – Policy Age

All the predictive variables are treated as continuous variables including the bldgContents variable. Both the multivariate techniques described in this paper works only with continuous and ordinal variables. Categorical variables cannot be directly analyzed by these methods for variable reduction.

The correlation matrix of the above predictive variables is:

	fireProt	**numBldg**	**numLoc**	**bldgAge**	**bldgContents**	**polAge**
fireProt	1.0000	-0.3466	0.0020	0.2921	-0.0945	-0.0328
numBldg	-0.3466	1.0000	0.8012	-0.2575	0.1216	0.0494
numLoc	0.0020	0.8012	1.0000	-0.0650	0.0619	0.0417
bldgAge	0.2921	-0.2575	-0.0650	1.0000	-0.0694	0.0287
bldgContents	-0.0945	0.1216	0.0619	-0.0694	1.0000	0.0068
polAge	-0.0328	0.0494	0.0417	0.0287	0.0068	1.0000

As can be seen, numBldg and numLoc are highly correlated and the variable fireProt has significant correlation with two other variables.

Principle Components Regression

As described earlier, the principle components are obtained by eigen-value decomposition of the covariance or correlation matrix of the predictive variables under consideration. Generally, most statistical software can compute the principle components once we specify the data set and the variables with which we want to construct principle components. SAS, for example, provides outputs of the linear coefficients (eigen vectors) along with mean and standard deviations of each predictive variables. These can be used to compute the principle components on the data set for regression.

Step 1: Compute the Jordan decomposition of the correlation matrix and obtain the eigen vector ($\Gamma i = \{\gamma 1i, \gamma 2i, \ldots, \gamma pi\}$) corresponding to each eigen value (λi).

The six eigen values of the eigen value decomposition of the above correlation matrix are as follows:

Eigen Values	Proportion of Total	Cumulative Proportion of Total
2.00767648	0.294943617	0.294943617
1.9965489	0.293308887	0.588252504
1.00066164	0.147005141	0.735257644
0.96103098	0.141183082	0.876440726
0.71945588	0.105693782	0.982134508
0.12161012	0.017865492	1

As we can see the first four eigen values capture about 90% of the information in the correlation matrix.

The eigen vectors (columns of matrix Γ in the Jordan decomposition) corresponding to each of the eigen values above are:

Eigen Vector 1	Eigen Vector 2	Eigen Vector 3	Eigen Vector 4	Eigen Vector 5	Eigen Vector 6
(0.336140)	0.589132	(0.135842)	0.167035	0.654102	0.256380
0.664985	0.178115	(0.053062)	(0.050656)	(0.097037)	0.715033
0.561060	0.501913	(0.109841)	0.005781	0.065075	(0.645726)
(0.313430)	0.558248	0.087962	0.212191	(0.729197)	0.075033
0.168213	(0.204757)	0.127973	0.953512	0.061786	0.020003
0.059014	0.125363	0.970851	(0.123655)	0.151504	0.002265

Step 2: Construct the principle components corresponding to each eigen value by linearly combining the standardized predictive variables using the corresponding eigen vector.

Hence the first principle component can be computed as:

PrinComp1 =

-0.336139581 * (fireProt - 4.55789) / 2.4533790858

+ 0.6649848702 * (numBldg - 1.10179) / 0.6234843087

+ 0.5610599572 * (numLoc - 1.16947) / 0.4635645241

+ -0.313430401 * (bldgAge - 48.5329) / 17.719473959

+ 0.1682134808 * (bldgContents - 2.36607) / 0.8750945166

+ 0.0590138772 * (polage - 4.81878) / 3.1602055599

Note that each variable is standardized while computing the principal components.

Now, we'll use the principle components we constructed above in a generalized linear model (GLM) type of regression. There are lot of papers and presentations on GLM ([1], [5]), and we will not spend effort here to describe the related concepts and details. The only two characteristics of GLM that we like to mention are error distribution and link function. Unlike the traditional ordinary regressions, a GLM can select any distribution within the exponential family as the model for the distribution of the target variable. GLM also allows us to use a non-linear link function that permits us to incorporate a non-linear relationship between the target variables and the predictive variables. For example, while fitting a severity curve often the LOG of the loss value can be modeled more easily than the actual loss value in a linear model. GLM allows us to accomplish this by specifying a LOG as the link function. However, it is to be noted that GLM is still linear in terms of the regression parameters.

In this numerical example for PCA, we choose Poisson distribution for regression error and choose IDENTITY as a link function. We used the claim frequency, claim count over $000 premium, as the dependent variable and used the principle components constructed above as independent variables. The summary of the regression is displayed below:

Obs	Source	DF	ChiSq	Prob ChiSq
1	Prin1	1	435.73	<.0001
2	Prin2	1	543.36	<.0001
3	Prin3	1	135.78	<.0001
4	Prin4	1	120.90	<.0001
5	Prin5	1	0.32	0.5737
6	Prin6	1	60.67	<.0001

The *P*-values and chi-square-statistics demonstrate that the first three principle components explained about 75% the predictive power of the original six policy variables. But, we also noticed the rank of the predictive power didn't line up with the order of the principle components. For example, the first principle component is less explanatory for the target than the second, even though the first principle component contains more information on the six original policy variables. In the next section, we'll describe another dimension reduction technique, partial least squares (PLS), which can be used to solve the problem.

PARTIAL LEAST SQUARES

In the last section we discussed applying PCA in regression as a dimension reduction technique as well as using it to deal with multi-colinearity problems. One drawback of PCA technique in its original form is that it arrives at SLCs that capture only the characteristics of the *X*-vector or predictive variables. No importance is given to how each predictive variable may be related to the dependent or the target variable. In a way it is an unsupervised dimension reduction technique. When our key area of application is multivariate regression, there may be considerable improvement if we build SLCs of predictive variables to capture as much information in the raw predictive variables as well as in the relation between the predictive and target variables. Partial least square (PLS) allows us to achieve this balance and provide an alternate approach to PCA technique. Partial least squares have been very popular in areas like chemical engineering, where predictive variables often consist of many different measurements in an experiment and the relationships between these variables are ill-understood (Kleinbaum et al. [4]). These measurements often are related to a few underlying latent factors that remain unobserved. In this section, we will describe PLS technique and discuss how it can be applied in regression problems by demonstrating it on our sample data.

Description of the Technique

Assume X is a $n \times p$ matrix and Y is a $n \times q$ matrix. The PLS technique works by successively extracting factors from both X and Y such that covariance between the extracted factors is maximized. PLS method can work with multivariate response variables (i.e., when Y is an $n \times q$ vector with $q>1$). However, for our purpose we will assume thatwe have a single response (target) variable i.e., Y is $n \times 1$ and X is $n \times p$, as before.

PLS technique tries to find a linear decomposition of X and Y such that $X = TP^T + E$ and $Y = UQ^T + F$, where

T $n \times r$ = X-scores U $n \times r$ = Y-scores

P $p\times r$= X-loadings Q $1\times r$ = Y-loadings

E $n\times p$ = X-residuals F $n\times 1$ = Y-residuals (1)

Decomposition is finalized so as to maximize covariance between T and U. There are multiple algorithms available to solve the PLS problem. However, all algorithms follow an iterative process to extract the X-scores and Y-scores.

The factors or scores for X and Y are extracted successively and the number of factors extracted (r) depends on the rank of X and Y. In our case, Y is a vector and all possible X factors will be extracted.

Eigen Value Decomposition Algorithm

Each extracted x-score are linear combinations of X. For example, the first extracted x-score t of X is of the form $t=Xw$, where w is the eigen vector corresponding to the first eigen value of X^TYY^TX. Similarly the first y-score is $u=Yc$, where c is the eigen vector corresponding to the first eigen value of Y^TXX^TY. Note that X^TY denotes the covariance of X and Y.

Once the first factors have been extracted we deflate the original values of X and Y as,

$X_1=X-tt^TX$ and $Y_1=Y-tt^TY$. (2)

The above process is now repeated to extract the second PLS factors.

The process continues until we have extracted all possible latent factors t and u, i.e., when X is reduced to a null matrix. The number of latent factors extracted depends on the rank of X.

A NUMERICAL EXAMPLE FOR PLS

In this section we will illustrate how to use the PLS technique to obtain X-scores that will then be used in regression. The data we used for this numerical example is the same as we used for the last numerical example of PCA. The target variable and all the predictive variables used in the last numerical example will be also used in this numerical example.

Partial Least Squares

As we described in the last section, PLS tries to find a linear decomposition of X and Y such that $X=TP^T + E$ and $Y=UQ^T + F$, where

T = X-scores U = Y-scores

P = X-loadings Q = Y-loadings

E = X-residuals F = Y-residuals

Decomposition is finalized so as to maximize covariance between T and U. The PLS algorithm works in the same fashion whether Y is single response or multi-response.

Note that the PLS algorithm automatically predicts Y using the extracted Y-scores (U). However, our aim here is just to obtain the X-scores (T) from the PLS decomposition and use them separately for a regression to predict Y. This provides us the flexibility to use PLS to extract orthogonal factors from X while not restricting ourselves to the original model of PLS.

Unlike PCA factors, PLS factors have multiple algorithms available to extract them. These algorithms are all based on iterative calculations. If we use the eigen value decomposition algorithm discussed earlier, the first step is to compute the covariance X^TY. The covariance between the six predictive variables and the target variable are:

2,208.72
9,039.18
9,497.47
2,078.92
2,858.97
(2,001.69)

As noted, the first PLS factor can be computed from the eigen value decomposition of the matrix X^TYY^TX. The X^TYY^TX matrix is:

4,878,441	19,965,005	20,977,251	4,591,748	6,314,657	(4,421,174)
19,965,005	817,067,728	85,849,344	18,791,718	25,842,715	(18,093,644)
20,977,251	85,849,344	90,201,995	19,744,478	27,152,967	(19,011,011)
4,591,748	18,791,718	19,744,478	4,321,904	5,943,562	(4,161,355)
6,314,657	25,842,715	27,152,967	5,943,562	8,173,695	(5,722,771)
(4,421,174)	(18,093,644)	(19,011,011)	(4,161,355)	(5,722,771)	4,006,769

The first eigen vector of the eigen value decomposition of the above matrix is:

{ -0.1588680, -0.6501667, -0.6831309, -0.1495317, -0.2056388, 0.1439770}.

The first PLS *X*-scrore is determined by linearly combining the predictive variables using the above values.

$$
\begin{aligned}
Xsr1 = \; & -0.1588680 * (\text{fireProt} - 4.55789) / 2.4533790858 \\
& - 0.6501667 * (\text{numBldg} - 1.10179) / 0.6234843087 \\
& - 0.6831309 * (\text{numLoc} - 1.16947) / 0.4635645241 \\
& - 0.14953171 * (\text{bldgAge} - 48.5329) / 17.719473959 \\
& - 0.2056388 * (\text{bldgContents} - 2.36607) / 0.8750945166 \\
& + 0.1439770 * (\text{polage} - 4.81878) / 3.1602055599
\end{aligned}
$$

Once the first factor has been extracted, the original *X* and *Y* is deflated by an amount ($Xscr1 * Xscr^T$) times the original *X* and *Y* values. The eigen value decomposition is then performed on the deflated values, until all factors have been extracted (refer to formula 2).

Obs	Source	DF	ChiSq	Prob ChiSq
1	xscr1	1	1131.04	<.0001
2	xscr2	1	141.42	<.0001
3	xscr3	1	20.96	<.0001
4	xscr4	1	24.23	<.0001
5	xscr5	1	4.06	0.0439
6	xscr6	1	0.11	0.7379

We next perform a GLM using the same claim frequency as the dependent variable and the six PLS components, *xscr*1 – *xscr*6, as independent variables. Same as we did in the numerical example for PCA, we still choose Poisson distribution for error the term and an IDENTITY link function. The regression statistics are displayed below.

Comparing to the ChiSq statistics derived from the GLM using PCA, we can see how each PLS factors are extracted in order of significance and predictive power.

Further Comparison of PCA and PLS

In this section, we have done a simulation study to compare principal components method against the partial least squares methods as a variable reduction technique in regression. A number

of simulated datasets were created by re-sampling from original data. PCA and PLS analysis were performed on these data samples and ChiSq statistics of the extracted PCA factors and PLS factors were compared. The exhibit below shows the results on three such samples.

	Simulated Sample 1		Simulated Sample 2		Simulated Sample 3	
Extracted Factor #	**ChiSq Statistics for PCA Factors**	**ChiSq Statistics for PLS Factors**	**ChiSq Statistics for PCA Factors**	**ChiSq Statistics for PLS Factors**	**ChiSq Statistics for PCA Factors**	**ChiSq Statistics for PLS Factors**
1	79.79	190.73	71.62	160.35	51.44	144.03
2	101.65	24.55	65.18	25.61	43.28	19.21
3	4.78	9.06	34.73	7.72	35.99	0.53
4	17.19	3.58	4.61	5.13	22.65	1.86
5	0.75	0.44	0.21	0.24	2.11	1.16
6	17.91	0.3	20.29	0.14	4.66	0.15

We can see from the above table that the chi-squared statistics of the first two PLS factors are always more than the corresponding two PCA factors in capturing more information.

Summary

PCA and PLS serve two purposes in regression analysis. First, both techniques are used to convert a set of highly correlated variables to a set of independent variables by using linear transformations. Second, both of the techniques are used for variable reductions. When a dependent variable for a regression is specified, the PLS technique is more efficient than the PCA technique for dimension reduction due to the supervised nature of its algorithm.

References

[1] Anderson, D., et al., "A Practitioner's Guide to Generalized Linear Models," CAS Discussion Paper Program (Arlington, Va.: Casualty Actuarial Society, 2004).

[2] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, 2nd Edition (New York: John Wiley & Sons, 1984).

[3] Bair, Eric, Trevor Hastie, Paul Debashis, and Robert Tibshirani, "Prediction by Supervised Principal Components," *Journal of the American Statistical Association* 101, no. 473, 2006, pp. 119-137(19).

[4] Kleinbaum, David G., et al., *Applied Regression Analysis and Multivariable Methods,* 3rd Edition (Pacific Grove, Ca.: Brooks/Cole Publishing Company, 1998).

[5] McCullagh, P. and J.A Nelder, *Generalized Linear Models,* 2nd Edition (London: Chapman and Hall, 1989).

[6] Rosipal, Roman, and Nicole Krämer, "Overview and Recent Advances in Partial Least Squares," in *Subspace, Latent Structure and Feature Selection*, Saunders, C., et al. (eds.) (Heidelberg: Springer-Verlag, 2006) vol. 3940, pp. 34-51.

Territory Analysis with Mixed Models and Clustering

Eric J. Weibel and J. Paul Walsh

Abstract

Motivation. Territory as it is currently implemented is not a causal rating variable. The actual causal forces that drive the geographical loss generating process (LGP) do so in a complicated manner. Both the loss cost gradient (LCG) and information density (largely driven by the geographical density of exposures and by loss frequency) can change rapidly, and at different rates and in different directions. This makes the creation of credible homogenous territories difficult. Auxiliary information that reflects the causal forces at work on the geographical LGP can provide useful information to the practitioner. Furthermore, since the conditions that drive the geographical LGP tend to be similar in proximity, the use of information from proximate geographical units can be helpful. However, to date procedures for incorporating auxiliary information involve the subjective consideration of conditions. And the use of proximate experience as a complement is complicated by complex patterns taken on by the LCG in relation to information density. Spline and graduation methods implicitly incorporate this information, but they tend to be applied ad-hoc to different regions. Incorporating a complement of credibility via proximate geographical units is only discussed formally in two papers, and is fairly undeveloped as a method. Another problem involves determining the relative value of information obtained via proximity versus the information provided by auxiliary variables. Separately, the implementation of territory as a categorical variable has prevented the integration of Territory Analysis with the parameterization of the remainder of the classification plan. In addition to these actuarial problems, territory's lack of causality creates acceptability problems. Lack of causality and increasingly complex territorial definitions have also reduced jurisdictional loss control incentives. The newly promulgated Proposition 103 regulations in California provide a useful venue for investigating solutions to these problems.

Method. Using the same data that was employed to create the California Private Passenger Automobile Frequency and Severity Bands Manual under Proposition 103, we employ a Mixed Model approach that combines the local zip code indication, an arithmetic model of causal geographical variables, and a proximity complement to determine the ultimate frequency and severity indication for each zip code. We then use constrained cluster analysis to assign these atomic geographical units into objectively determined and optimally configured frequency and severity bands. The constrained cluster analysis involves formulating the problem in terms of Nonlinear Programming.

Results. In three out of four cases, our approach, which is a rudimentary implementation of the mixed models with clustering concept that we introduce here, outperforms the existing Proposition 103 Frequency and Severity Bands Manual in terms of mean absolute deviation.

Conclusions. A mixed model approach is objective and efficient, and can substantially improve accuracy. The use of constrained cluster analysis on the result further achieves these ends. Furthermore, the development and analysis of the mixed model, particularly the arithmetic model of causal geographical variables, can be used to lay the groundwork for the introduction of causal geographical rating variables. These variables, such as traffic density, could eliminate complaints about the lack of causality. Furthermore, since these variables are typically continuous, they could be incorporated directly into the parameterization of the remaining classification plan. In California, such variables could be introduced to progressively supplant relative frequency and severity, improving accuracy and furthering the goals of Proposition 103.

Availability. The R programming language was used in preparing the data and mixed model. R is available free of charge at www.r-project.org. The constrained cluster analysis, employed the Premium Solver™ and KNITRO™ Solver Engine. This software is distributed by Frontline Systems, Inc. Order information, including free 15-day trials, are available at www.solver.com.

Keywords. Territory Analysis; Rate Regulation; Predictive Modeling; Credibility; Personal Automobile; Classification Plans.

1. INTRODUCTION

This paper introduces an objective two staged approach to Territory Analysis. In the first stage, a mixed model is applied to determine the expected loss frequency or severity for each zip code. The second stage applies cluster analysis to the results to arrive at objectively determined territorial groupings. We also introduce the use of constraints in the cluster analysis to reflect non-actuarial risk classification criterion.

1.1 Research Context

Territory Analysis has been subject to numerous risk classification challenges. Actuarial challenges include a particularly thorny opposition between homogeneity and credibility, and integration with the parameterization of the rest of the class plan. Non-actuarial risk classification challenges include the difficulty in creating objective methods, a perceived lack of causality and controllability, and affordability issues.

1.1.1 Homogeneity versus Credibility

In Territory Analysis, the classical tension between homogeneity and credibility expresses itself in the choice of an atomic geographical unit, and in the subsequent application of complimentary data if that atomic unit is not fully credible.

Selection of Atomic Geographical Unit

Over time, atomic geographical units have gone from those that correspond to jurisdictions to individual zip codes. Most recently, following the proliferation of GPS technology in the 1990s, there has been research into the treatment of territory as a continuum.[1] But, that research can be seen largely as paving the way for the future as opposed to seeing widespread implementation today. Much of the emphasis so far has been on methods that make use of indications and proximity only, with no consideration of auxiliary information.

Determining the Credibility Complement

When partially credible atomic units are elected in territory analysis, the problem then becomes how to group them to create credible homogeneous groupings. Kirkpatrick (1921) [5] first noted the problem, stating that while significant differences in loss costs between nearby cities may exist, those individual nearby cities typically would lack the credibility necessary to be properly recognized.

[1] Boskov (1994) [10]. Brubaker (1996) [11]. CAS (1997) [12]. Christopherson *et al.* (1996) [14]. Guven (2004) [17]. Taylor (2001) [22]. Taylor (1994) [23]. Wang and Zhang (2003) [24].

Contrary to Barber (1929) [1], Kirkpatrick argues that the solution is to group cities with similar conditions.

A fairly substantial divide has continued down throughout the years between fairly subjective systems that give consideration to auxiliary information and objective systems that do not. In the middle of the century the subjective approach might be typified by Stern (1956) [9], while the objective approach employed in Massachusetts is typified by McDonald (1955) [6].

Another approach is to objectively select and use complimentary data from proximate geographical units. California has led the way with this approach, publishing the only two papers that formally treat the subject[2].

1.1.2 Objectivity

The use of auxiliary information to help configure territories typically involves subjective judgment. This opens up the risk classification process to criticism. As early as Barber (1929) [1], it was argued that subjective approaches would not be accepted by the public. Shayer (1978) [34] claimed that the "Massachusetts" approach likely led to more accurate territorial groupings because it only gave consideration to the pure indication, as opposed to other information about the territory, which she termed "geographical considerations." Casey, Pezier, and Spetzler (1976) [26] note that subjective procedures, including the use of judgment in drawing territorial boundaries, is undesirable, and could be unfairly discriminatory. Phase I (1978) [19] seconded this concern.

Riegel (1920) [8] actually did propose a fairly workable objective system of incorporating exposure to traffic density into territory analysis. This was accomplished by using concentric circles around large city centers. Relative experience for concentric circles drawn around such cities could then be aggregated and used to guide in the selection of rate differentials around each individual city within the same size category.

1.1.3 Causality

Territory as a rating variable is often criticized on the basis of causality. Shayer, for instance, claims that territory is not causal but a mere proxy, and that this decreases its desirability as a rating variable. The Phase I authors criticized the industry's inability to explain *why* territory was an important rating variable.

A long list of factors has been posited to influence the geographical LGP. Traffic density

[2] Hunstad (April, 1996) [18] and Tang (2005) [21].

probably has the longest and most distinguished pedigree. Others include the configuration and maintenance of roads and highways, laws and regulations, attitudes of the public and court toward claims, population of drivers (distinct from population of potential claimants), enforcement of traffic regulations, population density, climate, driver education, and topography. More recently, variation in bodily injury liability loss costs has been attributed to triangles of attorneys, medical providers, and claimants, primarily in urban areas.[3]

1.1.4 Controllability

Territory has been criticized as not being reasonably under the control of the insured (Shayer). In addition to some conceptions of fairness, controllability is desirable because a self-elected reduction in exposure can reduce losses (Finger (2001) [30]). Shayer calls this the variable's *incentive value*.

1.1.5 Mobility and Automobile Territory Analysis

As early as Riegel (1920) [8], attention was given to the fact that vehicles may not be driven in only one territory. McDonald (1955) [6] noted that interests in one district argued that vehicles garaged in another district were responsible for accidents in their own district. Zoffer (1959) [10] notes a commercial automobile system that computed a weighted average rate based upon the proportion of time a vehicle was driven in each territory. In Stone (1978) [35], the commissioner acted on concern that suburban commuters contributed to congestion in the urban center. The Phase II (1979) [20] authors recommended that the occurrence zip code be coded by the DMV in order to study the problem.

1.1.6 Integration with the Remainder of the Class Plan

The determination of territorial boundaries has not been integrated with the application of modeling techniques such as generalized linear models. Furthermore, even after boundaries have been selected, the sheer number of territories often exceeds the number that can be supported by the data with the modeling process. When the disjoint process is used, authors have suggested fitting the model to the other classification factors, perhaps with a crude territorial component included, and then normalizing the territorial indications using those classification factors and the distribution of classifications in each territory.[4]

---

[3] Conners and Feldblum (1997) [15], Feldblum (1993) [16].

[4] Another problem which is largely outside of the scope of our paper is the fact that classification relativities may vary by territory. Early on, classification experience was often tabulated by general territory type. Spellwagen (1925) claimed that hazard by class did not vary much between territories. And Barber (1929) [1] argued against grouping "similar" territories together to arrive at separate classification factors. Alternatively, Stern (1956) [9] presents data showing significant

1.1.7 Affordability

Casey et al. note that affordability became a concern with territorial rates after the demographic shift of middle income persons out of many inner city neighborhoods. This left relatively low income persons to pay the high premiums that exist there. Chang and Fairley (1978) [27] showed that high-rated classes in high-rated territories may be charged too much when a purely multiplicative algorithm is used. This drew attention to the affordability issue because young urban drivers also tend to have the lowest incomes.

1.1.8 California Personal Automobile Insurance and Proposition 103

Proposition 103 was enacted by California voters in a 1988 referendum. The Proposition, and the sequence of regulations (and related court challenges) promulgated to implement it have profoundly impacted personal automobile ratemaking in California.

The Proposition allows the establishment of a relative frequency classification dimension and a relative severity classification dimension for each coverage part. No other geographical rating variables are currently approved for use. Originally, up to ten levels were allowed in each such classification dimension. Each zip code or other geographical unit must be assigned to one of the bands.

Because the credibility of an individual insurer's experience in a particular zip code is limited, the California Department of Insurance (CDI) created the *California Private Passenger Auto Frequency and Severity Bands Manual*, along with the data used to produce it. Carriers are allowed to make use of the CDI band assignments or the raw data if they need a complement of credibility. Hunstad (April, 1996) [18] presents the raw data, the methodology, and the final band assignments and factors.

Most recently, former Commissioner John Garamendi promulgated new regulations that may decrease the scope that territory can play in the overall rating plan. These regulations are being phased in, with full implementation to occur shortly. Consequently, personal automobile classification plan ratemaking in general, and personal automobile territory analysis in particular is the subject of intense focus in California currently.

1.1.9 Cluster Analysis

Cluster analysis has only entered the literature twice. Recently, Sanche and Lonergan (2006) [50]

---

differences in classification relativities by territory. Even larger differences were found in Phase II (1979) [20]. And Chang and Fairley (1978) [27] noted the inaccuracies introduced by purely multiplicative rating algorithms, when only one set of classification relativities are employed. Phase II also found that the impact of age and gender on geographical loss costs to be fairly negligible, but argued that an off-balance for the classification distribution should be applied.

introduced the actuarial use of cluster analysis. However, their treatment involved consideration of cluster analysis as a data reduction or mining technique.

Our focus is on the use of cluster analysis to group *objects,* not variables. This use in territory analysis has been brought up once before, in Phase I (1978) [19]. Although the authors cited works on cluster analysis and proposed its use, in the end, they grouped zip codes into contiguous territories by manually considering credibility-weighted indications.

1.2 Objective

Our objective is to strengthen the position of territory analysis as an accepted and accurate means of developing rating variables by confronting the primary risk classification challenges it is subject to.

As it stands now, territory is criticized as not being a causal variable. At the same time, the treatment of territory as a purely dichotomous categorical variable largely precludes the use of an integrated approach in the parameterization of the remaining class plan.

The mixed model approach that we propose includes the development of an arithmetic model of causal geographical variables. This can be considered a first step toward actually implementing new geographically based rating variables, such as *traffic density*, *legal environment*, and *traffic enforcement*. To a large extent, these variables can be expressed quantitatively. Thus, in addition to addressing concerns about causality, their ultimate introduction as rating variables can facilitate the integration of territory analysis with the parameterization of the remainder of the classification plan.

More centrally, our mixed model approach confronts the primary actuarial risk classification challenge, which involves the opposition of credibility and homogeneity. At the same time, the objectivity of the approach addresses concerns that territory analysis incorporates too much subjective judgment in configuring territorial definitions. The subsequent use of cluster analysis to group zip codes into territories adds further objectivity to the process, and should more completely inoculate territory analysis from such claims. Furthermore, we show how non-actuarial risk classification criterion can be incorporated objectively into the cluster analysis process itself.

1.3 Outline

The remainder of the paper proceeds as follows. In Section (2.1), we discuss our source of experience data from the California Department of Insurance, and introduce the context in which the data was produced. In Section (2.2), we discuss the primary actuarial risk classification challenge

in territory analysis, which is particularly thorny opposition between credibility and homogeneity. In that Section, we more precisely define the problem, and we discuss possible means of resolution, including the one we are proposing in this paper.

In Section (2.3), we introduce the mixed model. In Section (2.4) we conduct a search for causal variables related to geography. In Section (2.5), we discuss cluster analysis of the mixed model results.

We discuss our development of the regression model, and present the final model form in Section (3.1). The final model parameters and statistics are presented in Appendix B. We discuss the proximity complement in Section (3.2). An analysis of the regression model and proximity complement by region occurs in Section (3.3), including our monitoring of the credibility weighting of the three mixed model components. Plots of the mixed model components are presented in Appendix A. A comparison of our proximity complement to the existing proximity complement of Hunstad (April, 1996) [18] is given in Appendix C, giving mean absolute deviation by California Automobile Assigned Risk Plan (CAARP) territory. We discuss our constrained cluster analysis in Section (3.4). A comparison of our final result with Hunstad's final result occurs in Section (3.5) using mean absolute deviation as a metric. We also introduce the associated factor weights and their method of computation. The results are discussed in Section (3.6). In (3.7) we summarize potential avenues of future research. In Section (3.8) we discuss potential refinements of the mixed model. In Section (3.9) we discuss a possible alternative to the mathematical method we used in our cluster analysis, and we also discuss the possibility of automating our sequential cluster analysis procedure. We discuss the potential for introducing new causal geographical rating variables in Section (3.10) and potential enhancements to California personal automobile ratemaking in Section (3.11). Conclusions are presented in Section 4.

2. BACKGROUND AND METHODS

2.1 *The 1996 California Frequency and Severity Bands Manual*

In 1996, after much debate and legal fighting, an approach to territorial rating was arrived at, at least for the time, as we outlined in (1.1.8). The method of creating "bands" appears to have drawn on Phase I (1978) [19]. Members of each band were not required to be contiguous, but did need to exceed twenty square miles in aggregate.

Because the credibility of an individual insurer's experience in a particular zip code is limited, the Department of Insurance (CDI) created the *Private Passenger Auto Frequency and Severity Bands Manual.*

Carriers are allowed to make use of the CDI band assignments or the supporting raw data if they need a complement of credibility.

We conduct our analysis on this raw data. Hunstad (April, 1996) [18] presents this raw data, the methodology, and the final band assignments and factors. The data consists of exposures, claim counts, and capped losses for each zip code and for each coverage part. The data was aggregated between 1988 and 1993. Results were adjusted for relative amounts of coverage purchased, using auxiliary data taken from a subsequent data call of the major carriers in 1994.

2.2 Homogeneity versus Credibility in Territory Analysis

This is a particularly thorny problem in territory analysis. Urban loss costs can change dramatically over relatively short spans. When this occurs, it can be quite difficult to arrive at territorial definitions that are large enough to be credible, but yield a homogenous grouping. This phenomenon can occur in more subtle and insidious forms. Consider a series of small towns each separated by large sparsely populated expanses of land. What if costs did vary between these areas, albeit more modestly? While the gradient in costs might be flatter, the geographical density of information might be reduced even further. The gradual erosion of the existence and size of "remainder of state" territories provides some evidence that this situation has existed.

This phenomenon occurs when the loss cost gradient (LCG) overwhelms the density of information. This problem can certainly occur in other rating variables as well. Consider the 19-year-old driver.

It is quite likely that there will not be enough data to support the indication for 19-year-old drivers on their own. On the other hand, the LCG is so steep that if we widen the class we will introduce a substantial degree of heterogeneity. The common sense technical solution is to create a class for 19-year-olds, and then bracket the indication with the indications from 18- and 20-year-olds, either manually through the "avoidance of reversals," or more formally by fitting a line or curve through the indications, or some similar approach.[5]

[5] There is a non-technical difficulty with this approach. Regulators, consumerists, and the public will typically perceive the cells of the classification plan as completely dichotomous when in fact they rarely are. While this difficulty can be overcome, it can take the expenditure of some effort. When combined with the lack of a causal relationship and the additional dimension in territory, it can become quite an impediment. A similar etiology may lie at the root of allegations against sophisticated classification plans; specifically that they cannot generate credible indications and are thus necessarily undesirable. Implicit in this allegation is the conception that each cell is completely dichotomous; data from cells that even common sense would tell us are similar but not identical are ascribed no predictive value for the original cell.

The solution in territory analysis is not as easy because, despite statements that age is not a causal variable, age has a much more direct causal relation to loss propensity. Spatial loss gradients do not run one way or the other. If we had a patch of land analogous to our 19-year-old drivers, and we wanted to find the equivalent of the bracketing 18- and 20-year-old drivers, which patch of land would be equivalent to the 18-year-old and which patch of land would be equivalent to the 20-year-old? Our immediate response to that question might be to ask which way the center of the city is, and to assign the equivalent of the 18-year-old driver to the patch of land in that direction, and the equivalent of the 20-year-old driver to the patch of land in the opposite direction. This clearly demonstrates the lack of a direct relationship between geographical coordinates and loss costs. Other measures that are embedded geographically, such as traffic density and legal environment are the operative factors.

2.2.1 Resolution without Auxiliary Data

Without reference to auxiliary data, all that we have is proximity and the indication unadjusted for credibility.

McDonald Approach

This approach, which we referenced in Section (1.1.1), is one means that does not use auxiliary data and is purportedly objective. When the task at hand is only to revise a reasonably well functioning set of territorial boundaries and associated relativities, the approach is reasonable, although perhaps not optimal. Considerable information is thrown out when auxiliary information and the information from similar and adjacent geographical units is simply ignored. This may result in less accurate rates. In a competitive environment, the firm that used such techniques would be subject to adverse selection by carriers that employed techniques that used all the information at their disposal to arrive at more accurate territorial rates. Using the technique in conjunction with a reorganization caused by the imposition of new regulatory constraints, such as is occurring now in California, would be dangerous, as would the use of the analog to this technique when initially forming territories as opposed to revising.

Proximity Complement Approach

Another approach is to ignore auxiliary information, but to make use of additional data through the use of a proximity complement, as we discussed in (1.1.1). This is the approach that was employed in developing the California Personal Automobile Frequency and Severity Bands Manual (Hunstad, April 1996 [18]). It was also employed in the analysis by Tang (2005) [21].

Outside of those two papers, the literature is fairly silent about the construction of such complements.

The specific implementation of a proximity complement by Hunstad (April, 1996) [18] is subject to bias when the zip code to be complemented falls on the outside edge of the CAARP territorial boundary. Complements that are not uniquely constructed for each zip code are subject to this problem.

Ideally, proximity complements would be dynamically determined for each atomic geographical unit. Both the selection and the weighting of complementary units would be determined based upon numerous pieces of information. The amount of information present in the unit being complemented, along with the distance, land area, density of experience information, and the dispersion or spatial pattern in loss costs might all contribute.

The implementation by Tang is somewhat dynamic in this respect. The first complement is determined using the weighted average of all contiguous zip codes (the atomic units). In the event that the data is still not credible at this point, the indication for the CAARP territory is used as a second complement.

Hunstad (April, 1996) suggests that the indications for nearby zip codes could be weighted by their distance from the zip code being complemented. It is also suggested there that zip codes could be added to the complement one by one until full credibility is achieved.

Spline and Graduation Approaches

Another alternative, which is ostensibly objective and does not make use of auxiliary information, would be to use the spline and graduation techniques introduced by the authors we referenced in Section (1.1.1). While such approaches are continuous in nature, they can be converted for use with zip codes or other such discrete geographical units. These approaches can be somewhat ad hoc in that different analyses might be selected for different areas. As a result they might be difficult to justify to regulators and the public. They may be quite useful as analytic tools, however.

2.2.2 Subjective Resolution with Auxiliary Information

While this may well be the predominant approach, many aspects may not be frequently made explicit. For example, a carrier might present its groupings after the fact. With the subjective approach, the causal factors we identified in Section (1.1.3) may be incorporated in the process using professional judgment. Ad hoc consideration may also be given to proximate complementary data.

2.2.3 Objective Resolution with Auxiliary Information

Riegel's Approach

Objective resolution of the problem with auxiliary information has been largely nonexistent. In automobile insurance, a noteworthy objective procedure was the one proposed by Riegel (1920) [8], where concentric territories were established by radial distance from the city center, and the radial pattern of loss cost decay fit to similarly sized cities to develop uniform differentials from similar city centers. The auxiliary information is distance from city center, which is correlated with exposure to traffic density.

2.3 A Mixed Model Approach to Territory Analysis

We propose using an analog of the mixed model approach. Mixed models were first introduced by Bishop, Fienberg, and Holland (1975) [25], and were later discussed in general terms in Chang and Fairley (1978) [27], Venter (1990) [36], and Mildenhall (1999) [33].

The mixed model will consist of three components. The indication for the zip code is the first component. The arithmetic model predicted value for the zip code is the second element, and a proximity complement is the third element. We will examine the three resulting components, and arrive at a means of credibility weighting the three elements to arrive at a predicted value for each zip code.

Conceptually, we think this approach has tremendous promise to increase the accuracy of territorial rates. The specific implementation is preliminary, and we would expect improved means of implementing the general concept to be developed.

The specification of the arithmetic model and the identification of auxiliary variables will yield substantial benefits in addition to accuracy. By modeling continuous causal variables, we may put territory analysis on a firmer footing in terms of acceptability, and promote the integration of territory analysis with the parameterization of the remainder of the classification plan.

2.3.1 Selecting an Arithmetic Model

Because the purpose of this paper is to introduce the mixed model approach to territory analysis as a concept, and then introduce the use of cluster analysis in handling the result, we did not devote an inordinate amount of attention on the specific arithmetic model applied to the problem.

For simplicity, we elected to use a simple multiple regression model of auxiliary variables. To be sure, there are more appropriate models. We leave the search for the most appropriate arithmetic

models to future work that more specifically focuses on that element.

2.3.2 Selecting Causal Geographical Variables as Independent Variables

In addition to model form, the specific auxiliary variables to be included in the analysis must be identified.

2.3.3 Proximity Complement

Our proximity complement consists of all zip codes whose population weighted latitude and longitude falls within ten miles of the same measure for the zip code being complemented. The experience for all such zip codes in relation to each zip code being complemented was aggregated and a proximity complementary indication generated, along with the number of claims and exposures from which credibility figures could be derived.

We gave consideration to the use of contiguous zip codes, but deemed the effort to be too great, given that the zip code definitions we would be using would be somewhat dated, and thus of limited use on an ongoing basis.

2.3.4 Assigning Credibility Weight to Each Mixed Model Component

We start with a relatively simplistic credibility weighting procedure as our base, and then modify it when the data clearly show that one of the components is performing inadequately in a particular region.

We use the simple 1,082 claim rule to assign credibility z to the experience of the zip code in question. The proximity complement is assigned credibility via the following formula:

$$z_p = \frac{\left(\sqrt{\frac{c}{1082}}\right)(1-z)}{\left(\sqrt{\frac{c}{1082}} + R^2\right)} \tag{2.1}$$

where c is the number of claims in complement, and R^2 is the corresponding statistic for the arithmetic model fit to the frequency or severity of that coverage. The arithmetic model receives the remaining credibility, or

$$z_m = \frac{(R^2)(1-z)}{\left(\sqrt{\frac{c}{1082}} + R^2\right)} \tag{2.2}$$

Once again, our purpose was to introduce the use of mixed models in territory analysis as a concept, and so we did not devote attention to arriving at an optimal means of assigning credibility to each component. We leave this fine tuning to future researchers.

2.4 Causal Variables in the Geographical LGP

In this section we mention all of the variables that have been posited as being causal in the geographical LGP. We discuss the most immediately promising variables and sources of data.

Before proceeding with our variable search, we discuss the problem of *spatial interaction,* which is fairly unique to automobile territory analysis.

2.4.1 The Problem of *Spatial Interaction* in Automobile Insurance

In Section (1.1.5), we discussed the problem of mobility in automobile insurance. In geography, "the movement of people, materials, capital and information between geographic locations" is referred to as *spatial interaction*, Miller and Han (2001) [48]. Due to spatial interaction, the conditions that hold in a particular geographical unit such as a zip code do not fully describe the conditions to which vehicles garaged in that zip code will be exposed.

2.4.2 Causal Geographical Variables

Our literature review covered all of the geographical factors that have been thought to influence the geographical LGP. We summarize them below. As we will discuss in the succeeding sections, we have elected to include three of these variables in our arithmetic model.

We will Model	We will Discuss	Others
Traffic Density	Traffic Density	Medical Costs
Legal Climate	Legal Climate	Topography
Population Density	Population Density	Roads
	Nature of Population	Regulation
	Enforcement	Education
	Weather	Repair Costs

2.4.3 Traffic Density

- Population
- Number of Vehicles Used in Commuting

- Number of Vehicles
- Time Spent on the Road to Work
- Time Leaving to go to Work Each Day
- Total Road Surface Area
- Total Land Area
- Populated Land Area

With the exception of populated land area and total road surface area, all of these measures are available at the zip code level from the decennial census. And populated land area is a figure we derived by only including the land area of census blocks that were populated. Industrial, agricultural, and wilderness areas without population were thus not included in this measure.

Traffic density has been studied by the California Department of Motor Vehicles in models that include the driving record variables for the individual. Such a study was included in Phase II (1979) [20]. Traffic density had modest predictive value for individuals. Unfortunately, because miles of road lane were not available below the county level, the measure for the county had to be used. This simplification most certainly reduced the predictive power of traffic density.

Population density can be used as a proxy for traffic density. However, for our purposes we wanted to segregate the two elements. We discuss population density in (2.3.6).

We elected to focus on the commuter measures rather than the vehicle measures. In particular, the number of minutes spent commuting one-way by each commuter.

As we stated earlier, road surface area only appeared to be available at the county level in 1990, so we will not consider it as a candidate for a spatial denominator in our density measure. Rather, we give consideration to land area and populated land area in that role.

2.4.5 Legal Climate

- History and Current Philosophy of Local Court Jurisdiction
- Friendliness of Potential Juror Pool to Claimants
- Nature and Level of Activity of Local Bar
- Existence of Networks of Physicians and Lawyers who Cooperate
- Lawyer Density

The first four measures are not easily quantifiable. Lawyer density can be computed using the number of employees employed in legal offices, which is reported in the 2005 Survey of Economic Conditions from the Census Bureau. The denominator in the measure can be square miles of land, or population count. In addition to quantitative measures, there is the possibility of measuring the impact of the legal climate by examining experience within each superior court district as a binary variable.

Legal climate has a pedigree as long as traffic density. It has been difficult to measure, however. Most recently, in Conners and Feldblum (1997) [15] and Feldblum (1993) [16], it has been suggested that the density of lawyers contributes to liability loss costs, and that the impact of the legal environment can be measured by taking the ratio of bodily injury liability claims to property damage liability claims. The idea behind this is that this represents the percentage of property damage claims that were converted to bodily injury claims. Since the severity of accidents actually increases in rural areas due to higher speeds, the observed increase in the ratio in urban areas is posited to reflect an adverse claims environment, with an increased prevalence of soft-tissue injury claims.

We examine the number of legal employees, divided by land area, populated land area, and general population in our arithmetic model.

It is possible that numbers of actual lawyers could be obtained from bar associations, but for our purposes we felt that the number of employees is a sufficient proxy. It might also be useful to identify the number of personal injury attorneys or the number of medical specialists like chiropractors. Also, jurisdictions might be graded by experts in terms of the claims environment, and such measures might be tested in a similar model.

2.4.6 Population Density

Population density will be included in our model as well, and will be evaluated by the same measures with one exception: we will include a block weighted measure of average density. If density in the very immediate proximity of one's residence is more relevant, then this measure might be able to reflect that.

2.4.7 Population Characteristics

Variables of Interest

- Class Plan Off-Balance Effects
- Externality Effects from Variables Reflected in Class Plan

- Externality Effects from Variables not Reflected in Class Plan

It is important to remember that these are three different effects.

The first effect involves removing the influence of the other rating variables from the drivers in a given geographical unit.

Next are externality-like effects, which refer to synergistic or dampening effects that might be caused by the distribution of drivers. For instance, it is possible that if there are a lot of young inexperienced drivers in a particular area, loss costs in that area might increase *more* than the classification factor effects indicate. It is not difficult to imagine that for each at-fault accident that a bad driver is involved in, there might be one or more accidents that were at least partially caused by the driver's actions, even though the driver may not be recorded as the at-fault driver or even have been physically involved in the accident itself. On the other hand, the opposite might be true. In any case, there is the possibility the class plan off-balance would not fully reflect the impact of driver distribution on geographical loss costs.

Finally, there may well be population-related factors not even measured in the classification plan that influence geographical loss costs.

Risk Classification Issues

We can think of no objection to removing classification plan off-balances from territorial indications. In fact in California the sequential analysis procedure mandated under Proposition 103 regulations essentially require it.

The reflection of synergistic or externality effects has not been discussed much, so expectations with regard to potential acceptability are unclear.

With respect to variables not reflected in the classification plan itself, there is nothing that says per se that such items could not be modeled, either in the mixed model context or as an entirely separate geographical rating variable. An interesting question would be whether some variables that might not be acceptable for use on a personal basis would be deemed acceptable on a geographical basis, for example, average income.

Existing Data

Unfortunately, classification distributions are typically not provided in publicly available loss cost data at the zip code level. The data supporting the *California Frequency and Severity Bands Manual* is no exception. And, despite the fact that the sequential analysis procedure requires the removal of the

influence of all other rating factors from territorial loss cost indications, this influence is not removed from the Hunstad (April 1996) data.

In Phase II, the CDI found that externality or synergy effects were negligible. The authors did argue that class plan off-balances should be removed from zip code indications, however. Phase II employed DMV data that included some driver classification information.

Usage in our Study

We could have attempted to remove the effect of classification factors from the raw indications provided in Hunstad (April, 1996) [18], imputing the classification distribution from decennial census bureau data. We were reluctant to do so because of likely variations between the insured distribution and the population. The proportion of the population that is uninsured increases for younger drivers, due to their lower average incomes and higher average premiums. Furthermore, even if the overall proportion of uninsured motorists of various ages were provided, there are still probably unequal geographical variations in the rate of uninsured motorists by age. For instance, although the overall proportion of uninsured motorists might increase for a low income area with high premiums, the increase might be greater for younger drivers than more experienced drivers.

Since the adjustment could potentially introduce more error than it would eliminate, we elected not to adjust using decennial census bureau data.

To some degree, the fact that age, experience, gender, and marital status are not provided is mitigated by the fact that these variables tend to be fairly evenly distributed. However, this is obviously not the case for a variable such as driving record. Drivers in high frequency zip codes are going to have accident records that are worse than average, and vice versa.

We should note that we do include a temporal measure of commute distance in our models as a standalone variable and as a contributor to our measure of traffic density. Our intention was to account for spatial interaction to the extent possible, not to remove the effect of the mileage rating variable from the indications. However, our approach does have the impact of, to some degree, adjusting for average mileage driven.

Suggestions for Future Research

While we will not venture to tackle the problems enumerated here, future research should attempt to resolve them. And wherever possible, including the *California Frequency and Severity Bands Manual* case, the classification distribution should be provided at the zip code level when such data is

published for ratemaking purposes. Failing that, the impact of classification off-balance effects should be removed from the indications, using some agreed upon classification factors. It might also be useful to study how accurately census bureau data could be used to correct for class distribution for an insurance dataset where the actual insured distribution is known.

2.4.8 Implementation and Enforcement

- Traffic Enforcement

We found no data sources sufficient for use in our study. However, the measure known as the *enforcement ratio*, which was employed in Phase II (1979) [20], is a good first attempt at measuring how different levels of enforcement might affect loss rates. The Phase II enforcement ratio related the total number of all accidents and violations to the number of injury accidents in a zip code. The authors noted that the results might have been confounded by claims-consciousness. We would concur. Given that it is thought that bodily injury liability claim frequencies vary considerably based not upon accident conditions but on the legal environment, the use of injury accidents in the denominator appears problematic.

2.4.9 Weather

Weather data are certainly available in quite granular form. Although it is beyond the scope of the present study, the impact that weather and climate has on accident statistics may be worthy of further study. Given the tension between credibility and homogeneity that exists in territory analysis, smoothing of this significant source of variation could actually improve our ability to further improve the specificity our study of geographical loss costs. If an accurate weather model could be constructed, and the time and impact of that weather on losses could be derived, then the random noise created by annual fluctuations in the weather could be removed and replaced with a continuous cost variable similar to what is produced in geographical catastrophe models.

2.5 Grouping Mixed Model Results with Cluster Analysis

2.5.1 The Primary Objective

Our goal is to objectively group zip codes into bands that accurately reflect their expected relative frequency and severity rates. Additionally, we wish to be able to impose various social and regulatory acceptability constraints on the grouping process. One of the reasons for grouping in the first place, a complement of credibility, is less of a concern for us because we have already incorporated complimentary information from the arithmetic model and from the surrounding zip codes.

As we have stated earlier, the use of professional judgment in assigning zip codes to territories is a frequent source of criticism.[6]

In basic terms, we would like to specify our problem as follows:

Let x_{ij} be our decision variables, where the first dimension represents the zip code. The second dimension represents the frequency or severity band. So, under the 1996 regulations and our data, i can range from 1 to 1,502, while j can range from 1 to 10.

A particular piece of land can only be assigned once. So, it would seem that we should define x as a binary variable.

$$x_{ij} \in [0,1] \in \mathbb{N} \tag{2.3}$$

$$\sum_i x_{ij} = 1 \tag{2.4}$$

A desirable objective function for frequency might be of the form:

$$\min \sum_i \sum_j \left[\left(R_i - \frac{\sum_b x_{bj} R_b E_b}{\sum_c E_c x_{cj}} \right) x_{ij} E_i \right]^2 \tag{2.5}$$

Where R_i is the computed *mixed model* relativity for the zip code (as opposed to the *raw* computed relativity for the zip code). E_i is defined as the number of exposures in the zip code.

Or, alternatively,

$$\min \sum_i \sum_j \left[abs \left(R_i - \frac{\sum_b x_{bj} R_b E_b}{\sum_c E_c x_{cj}} \right) x_{ij} E_i \right] \tag{2.6}$$

2.5.2 Constraints

In addition to the number of bands, an initial constraint we would be interested in is the requirement that each band consist of at least 20 square miles. To incorporate such a constraint, we

[6] Barber (1929) [1], Casey et al. (1976) [26], Phase I (1978) [19], Shayer (1978) [34].

would define L_j as the number of square miles of land area contained in the zip code and impose the following:

$$\sum_i L_i \, x_{ij} \geq 20 \tag{2.7}$$

We would also be interested in developing constraints upon the size of the factor weight, as computed via the proxy[7] method. Affordability constraints are also of interest.

2.5.3 Basic Cluster Analysis

Cluster analysis comes immediately to mind as an appropriate means of accomplishing the task at hand.

The literature on cluster analysis is vast and diverse because for some time it developed somewhat independently under the auspices of different academic disciplines. The two standard textbooks on the subject are Kaufman and Rousseeuw (1990) [46] and Everitt, Landau, and Leese (2001) [43]. Han, Kamber, and Tung (2001) [45] also provide a remarkably brief introduction. The use of cluster analysis for our task was mentioned once in the actuarial literature (Phase I). However, it was ultimately not employed.

Accuracy for Selected Number of Clusters

Partitioning (Kaufmann and Rousseeuw) techniques, otherwise known as optimization methods (Everitt et al.), tend to create more accurate partitions for a given number of clusters according to Kaufmann and Rousseeuw. Sanche and Longergan focused immediately on hierarchical methods, which are more suited to the task they were concerned with. We are predisposed toward choosing the more accurate, computationally demanding methods.

Robustness

In selecting a methodology and algorithm, we could elect an L_2 objective function (2.4) that more severely penalizes misclassification but is less robust. Or we could apply an L_1 objective function like (2.5), which is robust. Kaufmann and Rousseeuw strongly advocated robust methods of clustering.

[7] See Title 10, California Code of Regulations, Section 2632.8(c) , which was filed on 11/1/2002

2.5.4 Constrained Clustering

The imposition of constraints is a very new topic in cluster analysis. Kaufmann and Rousseeuw do not even mention it. The more recent Everitt et al. discuss constrained cluster analysis. However it quickly becomes apparent that the types of constraints we are interested in are not covered. Everitt et al. devote their discussion to spatial constraints, such as proximity and contiguity[8], and certain constraints related to hierarchy.

Han, Kamber, and Tung's (2001) [45] excellent and concise survey discusses constraint-based cluster analysis, and pioneering work being done. Of interest to us is Tung et al. (2001) [52]. Those authors discuss constrained cluster analysis generally, and introduce a solution for one particular form of constraint.

Tung et al. (2001) [52]

The authors introduce the following classes of constraints: 1) *Existential*, 2) *Universal*, 3) *Existential-Like*, 4) *Parameter*, 5) *Summation*, and 6) *Averaging*.

Existential Constraints

Existential constraints focus on the particular qualities of the individual atomic geographical units being grouped. In terms of our problem, an example of existential constraint would be a requirement that each cluster contain at least two zip codes that *each* have a land area of at least two square miles.

Unfortunately for us, this is the only type of constraint for which the authors construct a specific solution. This particular form of constraint is not of immediate concern to us, although it is possible it could be a concern in some type of territorial assignment problems.

Universal Constraints

Universal constraints require each member of a *particular* cluster to meet a particular condition. In our example, this might be the requirement that our highest rated cluster only contain zip codes with per capita income levels in excess of a certain measure. This constraint is simply solved by running separate cluster analyses. Although not of immediate concern to us, this could be of use in formulating a cluster analysis that incorporates affordability constraints.

---

[8] The recentness of the literature cited may provide part of the reason cluster analysis was not employed in Phase I. One of the constraints imposed on CAARP territories, in addition to the minimum twenty square miles rule, is the requirement that each territory be contiguous. Everitt *et al.* mentions the following research in regards to contiguity: Maravalle *et al.* (1997) [47], Ferligoj and Batagelj (1982) [44], Murtagh (1995) [49], and Wojdyla *et al.* (1996) [53]

Existential-Like Constraints

This type of constraint focuses on the *number* of objects contained in each cluster. In our case, such a requirement might be that each band contains at least three zip codes. These constraints are similar to existential constraints, and can be handled by fairly simple modifications to algorithms. Unfortunately, these constraints are not of particular interest to us either.

Parameter Constraint

This is a constraint on the number of clusters.

Summation Constraint

This is the particular form of constraint we are interested in. It is concerned with the sum of a quantity of the members of each cluster. In our example, the minimum land area of twenty square miles is a summation constraint. Again, unfortunately the authors do not provide a method for solving the problem.

Averaging Constraint

Averaging constraints are similar summation constraints.

Berkhin (2006) [38]

This author provides a survey of very recent advances in cluster analysis. Included is a discussion of recent advances in constraint-based cluster analysis.

Unfortunately, with respect to the constraints we are interested in, the author refers to sources we have already covered, in particular Han et al. and Tung et al.

Since this is a very recent survey, and since Han et al. and Tung et al. note the difficulty in solving the summation constraint problem, this leaves us in a bit of a pinch with respect to the cluster analysis literature.

Teboulle et al. (2006) [51]

Teboulle et al. indicates that most optimization problems in cluster analysis involve non-convex objective functions. The author claims that the *k-means* method of cluster analysis can sometimes be configured as a nonlinear programming gradient-type method.

2.5.5 Constrained Cluster Analysis Using Nonlinear Programming

A review of our objective function and the initial constraints indicates it can be considered a

nonlinear programming problem from operations research. (See Hillier and Lieberman (1995) [60]).

Since we are predisposed toward an optimization method as opposed to a hierarchical method, and since optimization cluster analysis is related to nonlinear programming, we elected to look here for a solution to our problem, which includes the imposition of summation constraints that are not currently well-handled in traditional cluster analysis.

3. RESULTS AND DISCUSSION

3.1 Regression Models

3.1.1 Modeling Objectives

Our primary objective is prediction; we want to create a model that will provide the best credibility complement. A secondary objective is to provide groundwork for further research into the introduction of causal geographical variables. Given our primary objective, we built more complex models than we might have if our primary concern was to establish the use of causal geographical variables. Any project to directly introduce causal geographical variables for the first time might need to use relatively simple models whose coefficients are easy to explain.

3.1.2 Spatial Interaction

We previously mentioned the problem of *spatial interaction* in automobile insurance. When geographical variables are introduced in automobile insurance, careful consideration must be given to how proximate geographical units will interact.

10, 20, and 50 Mile Radii

Our general approach was to compute values for our variables within the zip code itself, and for zip codes within three mutually exclusive radii of 10, 25 and 50 miles. Distances were computed using the Haversine formula. Zip code latitudes and longitudes were computed by population weighting census blocks (without using the Haversine formula).

Jaggedness

One problem with this general approach is the jaggedness of the zip code rings created by the procedure. With more time and computing power, the information fed into the model might be taken at the decennial census block level rather than the zip code level. This would prevent the jaggedness that occurs when zip codes of different sizes are included. California contains well over

300,000 census blocks, so this would be very computationally intense. If this level of granularity were to be used, experiments could be run on the appropriate number and length of radii. A less computationally intensive approach could employ census tracts rather than blocks.

Variable Exposure Density in Presence of Gradient

Although it was beyond the scope of our study, future researchers may wish to consider mitigating variation in exposure density via an arithmetic average model variables or some other weighting scheme.

Commute Times

We gave careful attention to commute length when considering how to structure the models with respect to spatial interaction.

3.1.3 Final Variables

Commute Distance of Drivers in the Zip Code

Our focus on commute distance is motivated not by a desire to incorporate mileage into the model per se; rather, it is to accurately reflect spatial interaction within the framework of our radial defined variables. Commute distance is a key contributor to our traffic density measure, and it can be expected to interact with geographical conditions within its range.

- CT_i:= We estimated average time spent commuting to work, one-way, for commuters in the zip code being modeled, using the decennial census variable that presents the temporal commute distance distribution.

Traffic Density

We carefully considered how to reflect spatial interaction in this variable. For the numerator, we elected to use the total number of minutes one way to work, in aggregate for all commuters. The density of this combination was computed by dividing by the involved land area. Thus the measure is called commute-time-space-density. We computed the three radial versions of this variable at 10, 25, and 50 miles.

- $TD10_i$:= Commute length (in minutes) for commuters in zip codes within 10 miles/land area for zip codes within 10 miles. Unlike our standard procedure in computing radial measures, we *did* include commuters and land area contained in the zip code being modeled, within the 10-mile variable.

- $TD25_i$:= Same measure. Computed for zip codes between 10 and 25 miles from the zip code being modeled.
- $TD50_i$:= Same measure. Computed for zip codes between 25 and 50 miles from the zip code being modeled.

Legal Environment

As we have stated, quantitative variables should be exhausted before binary geographical variables are employed. Proceeding along the lines suggested in Conners and Feldblum (1997) [15], and Feldblum (1993) [16], we attempted to incorporate lawyer density where it made sense. A priori, we suspected it to be most important to bodily injury (BI) liability frequency, followed by bodily injury liability severity and *perhaps* property damage (PD) liability severity respectively. We did not anticipate it to be a causal variable in property damage liability frequency.

Superior court jurisdiction could be a fairly substantial causal binary geographical variable, although this conflicts with our desire to minimize the use of categorical variables. Additionally, at the time the data was generated, the tort liability system operated under a different jurisdictional scheme. Since then jurisdiction has been reorganized.

As we discuss in the section on geographical binary variables, we do allow, as a last resort, the introduction of major metropolitan binary variables, which could to some extent be thought to correspond to general legal environment. We discuss this further there.

In computing our most favored measure of legal environment, lawyer density, we have elected to use population as the denominator rather than land area. Either land or population are plausible denominators, but given that so many of our other measures include land area as a denominator in a density measure, we gave a priori preference to population. This might reduce multicolinearity somewhat.

- $LD25_i$:= Lawyer Density 25 miles: Number of persons employed in legal offices in zip codes within 25 miles/total population in zip codes within 25 miles. *Includes* the zip code being modeled.
- $LD50_i$:= Lawyer Density 50 miles: Same but includes zip codes greater than 25 miles but less than 50miles radius.

Population Density

Population and traffic density overlap. Because it seems more plausible that traffic density is a

directly causal variable, and because it would likely be seen as a somewhat more acceptable measure than population density, we gave it preference, and took care not to use population density when traffic density would suffice.

As it turned out, population density was a fairly important factor, particularly for property damage liability severity.

In attempting to model population density as distinct from traffic density, we hypothesized that very local density conditions (in the precise neighborhood where the vehicle was garaged), might influence claimant behavior. In this regard, we did introduce a *block weighted* measure of population density, which measured average density at the census block level. So, a zip code that is highly dense on one side, and very sparse on the other would have a very high block weighted measure of population density, while the measures using simple land area as a denominator, would have an intermediate value. Falling between these two measures we created a measure that included only land area from census blocks that had a population of at lease one. In testing these variables, we were surprised to find that the block weighted measure did not perform well at all. The measure using only populated census blocks performed about as well as the normal measure of population density. Given the rough equivalence of the two, we have elected to employ the standard measure of population density in our final model.

- PD_i := Population density within modeled zip code. Population divided by total land area for the zip code being modeled.
- $PD10_i$:= Population density 10 miles: Same measure but for all zip codes (except the zip code being modeled) that are less than or equal to 10 miles radius from the zip code being modeled.
- $PD25_i$:= Population density 25 miles: Same measure but for zip codes between 10 and 25 miles radius from modeled zip code.
- $PD50_i$:= Population density 50 miles: Same measure but for zip codes between 25 and 50 miles radius from modeled zip code.

Geographical Binary Variables

Geographical binary variables can be criticized with respect to causality. When considered alone, these variables reflect current territory analysis practice. To the extent that the boundaries of the region correspond to factors thought to be causal, such as jurisdictional boundaries for the superior

court or local governments, they could to some extent be identified with those factors.

Our primary interest with respect to these variables is legal jurisdiction. But other unexplained differences may be reflected as well, particularly for PD liability frequency. As we stated above, there has been a significant judicial reorganization since the time our experience data was generated. Additionally, to the extent possible we wish to measure causal forces in terms of quantitative variables, as opposed to categorical ones.

For this reason, we only introduced the major metropolitan areas as binary variables, to account for the most major regional differences we would anticipate a priori. A priori, we anticipate Los Angeles, San Francisco, and the remainder of state to have different environments.

We only introduced these two metropolitan areas as a last resort, when combinations of variables could not produce nearly as good a fitting model. During the course of the model-fitting exercise, we found that the city of Los Angeles and the remainder of Los Angeles county behaved somewhat differently, and hence we introduced two binary variables for Los Angeles, one for the central city and one for the remainder of county.

- LA_i := Los Angeles: A binary variable that is coded "1" for all zip codes in central Los Angeles, which is defined as zip codes from 90001 to 90077.
- LAC_i := Los Angeles area: A binary variable that is coded "1" for all zip codes in Los Angeles County with the exception of central Los Angeles, which consists of zip codes from 90001 to 90077.
- SF_i := San Francisco: A binary variable that is coded "1" for all zip codes in the city of San Francisco.

Results

Appendix B contains the model parameters and statistics. Appendix A contains plots of observed frequency/severity, model predicted frequency/severity, and model residuals. The *x*-axis is arrayed by observation, rather than listing individual zip codes, which number 1,502 in our overall data set, and usually a few less in each individual instance due to missing independent variable values that prevented us from computing a model estimate. To help orient the reader, ranges associated with particular cities, counties or regions are denoted with arrows at the top of each plot.

3.1.4 Bodily Injury Liability Frequency

Final Model

$$BIFQ_i = \hat{\alpha} + \hat{\beta}(CT_i) + \hat{\gamma}(TD10_i) + \hat{\delta}\,(TD25_i) + \hat{\varepsilon}(TD50_i) + \hat{\epsilon}(LD25_i) + \hat{\theta}(LD50_i) + \hat{\vartheta}(LA_i) + \hat{\pi}(LAC_i) + \hat{\rho}(SF_i) + \hat{\tau}(CT_i TD25_i) + \hat{\varphi}(CT_i LA_i) + \hat{\omega}(LD25_i LAC_i) + \hat{\varsigma}(LD50_i LAC_i) + \hat{\xi}(CT_i LD25_i) \quad (3.1)$$

3.1.5 Property Damage Liability Frequency

Final Model

$$PDFQ_i = \hat{\alpha} + \hat{\beta}(CT_i) + \hat{\gamma}(TD10_i)^{0.5} + \hat{\delta}\,(TD25_i)^{0.5} + \hat{\vartheta}(LA_i) + \hat{\pi}(PD_i) + \hat{\rho}(PD10) + \hat{\theta}(PD25_i)^{0.5} + \hat{\epsilon}(CT_i TD10_i) + \hat{\tau}(CT_i TD25_i) + \hat{\varsigma}(CT_i PD10_i) + \hat{\varphi}(CT_i PD25_i) + \hat{\omega}(CT_i LA_i) \quad (3.2)$$

3.1.6 Bodily Injury Liability Severity

Final Model

$$BISV_i = \hat{\alpha} + \hat{\beta}(CT_i) + \hat{\epsilon}\,(LD25_i) + \hat{\theta}(LD50_i) + \hat{\gamma}(TD10_i) + \hat{\varepsilon}\,(TD50_i) + \hat{\vartheta}(LA_i) + \hat{\pi}(LAC_i) + \hat{\tau}(CT_i LD25_i) + \hat{\varphi}(CT_i LD50_i) + \hat{\omega}(LD50_i LA_i) \quad (3.3)$$

3.1.7 Property Damage Liability Severity

Final Model

$$PDSV_i = \hat{\alpha} + \hat{\beta}(LD25_i)^{0.5} + \hat{\pi}(PD_i)^{0.5} + \hat{\varsigma}(PD10_i)^{0.5} + \hat{\theta}(PD25_i)^{0.5} + \hat{\gamma}(PD50_i)^{0.5} + \hat{\vartheta}LA_i + \hat{\pi}LAC_i + \hat{\rho}SF_i + \hat{\epsilon}(CT_i * LD25_i)^{0.5} \quad (3.4)$$

3.2 The Proximity Complement

Our goal once again was to introduce the concept of a mixed model, using model and proximity defined complements to the zip code indication. As a result we introduced a relatively simple proximity complement. We discuss potential avenues of future research later in Section 3.

The proximity complement we elected can be considered dynamic in that a separate measure is

computed for each zip code being complemented. This is as compared to the Hunstad (April, 1996) [18] CAARP complements, which were pre-defined and static.

The proximity complement employed in Tang (2005) [21], however, can be considered even more dynamic. Immediately contiguous zip codes are used as a first complement for each zip code, which is similar to our ten-mile radius measure. Tang's complement is also dynamic in that it responds to the amount of information contained in the zip code being complemented, and the contiguity complement, and then determines whether the CAARP complement is necessary for any unfulfilled credibility.

We considered use of a contiguous proximity complement. But as we stated earlier, it was deemed to be too laborious given that the zip code definitions are from 1990, and so creating or procuring the contiguity definitions would serve no useful future purpose.

Our proximity complement appears to fare best in less densely populated areas and areas where the LCG does not appear to be particularly steep. This is as we would have expected. Our complement fared poorly in the most densely populated urban areas. Particularly in central Los Angeles, where many of the zip codes are not completely credible, this is a serious problem.

In Appendix C, we present a table comparing our proximity complement to Hunstad's CAARP complement, using mean absolute deviation within each CAARP territory as a statistic. We also included the number of zip codes in each CAARP territory that required a complement, since the primary concern should be with areas where complementary information is needed. We analyze the regionally specific performance of our complement against the other two elements of the mixed models in the following section.

3.3 Analysis and Credibility Weighting of Mixed Model Components

In this section we evaluate the relative regional performance of the proximity and model complements for each coverage part. Ideally, the relative credibility for each mixed model component would be determined by its relative local performance. Our purpose here is to introduce the concept, not necessarily to arrive at the best possible implementation. For this reason, we did not devote significant attention to the determination of the credibility weighting formula. Because both the individual mixed model components and the credibility weighting formulas are preliminary in nature, we did intervene in the credibility weighting process (from our formulas (2.1) and (2.2)) when there were particularly serious problems with the local fit of a measure. We discuss each such instance as it occurs below.

Once again we leave the determination of optimal credibility weighting schemes to future researchers.

Appendix A contains plots of each mixed model component and the regression model residuals. The attached plots include arrows that denote geographical regions of interest. The *x*-axis is simply the zip code, so too much meaning should not be ascribed to changed patterns in ranges outside of the arrows without further investigation.

3.3.1 Bodily Injury Liability Frequency

Bodily injury liability frequency is certainly the most interesting of the four analyses. As evidenced by the plots of observed values, the range is much wider. The local legal and claimant environment is thought to significantly influence geographical variation in BI frequency. In central Los Angeles, BI frequency is almost equal to PD frequency, while in rural areas BI liability frequencies are much lower than PD liability frequencies. Since rural accidents tend to be more serious in nature, this would seem to point to substantial differences in claiming behavior.

We expected and found legal variables (lawyer density and the geographical binary variables) to significantly influence frequency.

Urban Metropolitan Areas

These include Los Angeles, San Francisco, and Oakland/Berkeley.

Los Angeles

Central Los Angeles exhibited the highest frequencies. The zip codes here tend to be smaller and densely packed. In this sort of an environment, we would expect a lack of performance from our proximity complement. The radius of ten miles used in our proximity complement is static. It is not responsive to local heterogeneity or exposure density. In central Los Angeles, ten miles is probably too much, since geographical information density is extremely high. Adequate quantities of information can be obtained in smaller radii. And, given the steep LCG, using a wider than necessary radius introduces heterogeneity. This can be observed by comparing the plot of observed frequency with the proximity complement plots. The proximity complements are densely packed at about 0.03. Each proximity complement contains a massive amount of data, and each complement contains mostly the same zip codes, as the size of each zip code probably dramatically increases as one leaves the center city.

CAARP territory 39 roughly corresponds to the most central part of Los Angeles. In Appendix

C, we can see that the Hunstad complement fares much better in terms of mean absolute deviation than our ten-mile complement in this territory.

Our arithmetic model includes a binary variable for central Los Angeles and for its remainder so there is little regional bias in the residuals. The higher observed heterogeneity in central Los Angeles is probably due both to actual heterogeneity in expected frequencies, and also to the fact that many of the zip codes in central Los Angeles are not fully credible, because many drivers are uninsured due to affordability.

Because of the extreme lack of fit for the proximity complement here, we have elected to intervene in the credibility weighting process. No credibility is assigned to the proximity complement in and around central Los Angeles (zip codes 90001 to 91108). All of the credibility that would have been assigned to the proximity complement was instead assigned to the model complement.

San Francisco

San Francisco is subject to much lower BI frequency than would be expected given its density. The BI/PD ratio is relatively low for an urban area. This is likely due to the legal environment. Slow average speeds associated with density could have contributed, but this could be counterbalanced by more collisions with pedestrians.

The residuals for the model complement indicate good performance for San Francisco, while the proximity complement is tightly bunched, although not particularly biased. The adjacent bay and ocean may contribute to this bunching.

Future researchers might wish to include an investigation into the impact that the bay and ocean have on the performance of mixed model components.

Oakland/Berkeley

Next rightmost is Oakland/Berkeley, which exhibits a modest positive residual bias. Such a bias is not discernable in PD frequency residuals.

Suburban Areas

These consist of southwest Orange County, Fresno, and Sacramento, as well as a modest proportion of the remainder of the plot.

Although several residual spikes are clearly noticeable, and indicate places where a geographical

binary variable would significantly improve fit, no interventions were made in these areas, so the credibility formulas (2.1) and (2.2) were left to operate freely.

Fresno

Moving from left to right, the first such spike is for Fresno. Clearly the model is underestimating frequency here. This bias also exists for property damage liability, but to a much less significant degree, so it would appear that legal environment might be to blame, as opposed to some unexplainable increase in the overall level of accident frequency. An investigation into the claims environment would be of interest. And a binary geographical variable would clearly improve fit here.

San Jose

A fairly surprising residual spike occurs for San Jose, which is not denoted on the graph but can be quickly identified between 1000 and 1100 on the x-axis. There is little corroborating evidence in the property damage frequency plot to indicate a general unexpected spike in the overall accident rate. There appears to be a somewhat stronger uptick in raw bodily injury frequency observations for San Jose. And there would not seem to be any obvious reason why accidents in San Jose would be relatively more likely to result in real injury. So there is some basis for an investigation of differences in claimants and the courts. A binary geographical variable would clearly improve fit here.

Sacramento

The most striking residual spike occurs for the city of Sacramento, which sits to the far right of the plot. Such a spike only occurs in muted form in the property damage liability residual. The spike is clearly visible in the raw frequency plot also. An analysis of the legal environment here would clearly be in order. And, clearly a binary geographical variable would dramatically improve fit.

Rural Areas

This includes extreme northern California, which falls to the immediate left and right of Sacramento on the plot. And, the majority of the remaining unlabeled plot consists of rural zip codes, many of them in central California and the southern inland empire area.

Northern California

This label actually refers to extreme Northern California away from the coast, while the area immediately to the left of Sacramento occurs in extreme northern California along the coastline.

Zip codes in this area are relatively sparsely populated. Hence the plots in this area contain more

dispersion, and it takes a few more seconds to see the bias in residuals. Comparing the residuals to the 0.0 line on the *y*-axis it becomes clear that the inland extreme northern California is significantly overestimated by the arithmetic model. It is possible that this is a significantly less litigious environment. A similar, but less extreme situation can be observed in Coastal northern California, which falls to the immediate left of Sacramento on the plot. The same pattern exists for property damage liability, but to a significantly reduced degree. Clearly geographical binary variables would improve fit here.

The proximity complement performs well here with respect to bias. This is to be expected given the lack of geographical information density and the shallow LCGs likely to be present here. But the precision of estimates could probably be improved by increasing the geographical scope of the proximity complement. So in this area we observe the opposite situation from central Los Angeles. Clearly a more dynamic complement would improve things.

Upon inspection, it would appear that the performance of the mixed model could be improved here if a higher relative credibility weight were awarded to the proximity complement. And perhaps the model complement could be assigned zero credibility here. A better arithmetic model, combined with a dynamic complement is probably the best solution. Ultimately we elected not to intervene in the credibility weighting procedure here.

Remainder of State

Rural areas in southern and central California did not appear to be subject to the same degree of model bias. These areas are probably less sparsely populated than in extreme northern California. So while larger proximity complements might be in order, the need is not as pronounced as in the extreme north.

Conclusions

To conclude, we only intervened in the limited instances we discussed above. However, this was partly due to the nature of this paper, which is to introduce the concept in simple form, allowing later researchers to more finely tune each element of the mixed model and cluster analysis. It would appear that major increases in fit could be gained by dividing the state into a few additional regions and assigning binary random variables. A suggestion would be binary variables for Fresno, Sacramento, San Jose, the remainder of state north of the bay area, and the remainder of state falling south and east of there, perhaps including significant suburban and urban (Oakland/Berkeley) populations in the east Bay Area. Another alternative, which we will discuss later, is to use a spatial

autocorrelation model.

3.3.2 Property Damage Liability Frequency

The proximity complement performs similarly for property damage liability frequency. But the problems in the central city areas are not pronounced.

From the perspective of regional bias, the model complement performs much better. The model similarly over-predicts for inland and coastal extreme northern California, but error is smaller. The model tends to modestly over-predict for rural areas. There appears to be modest over-prediction for San Jose.

No interventions in the credibility weighting process were urgently necessary.

3.3.3 Bodily Injury Liability Severity

The model modestly under-predicts for central Orange County. A moderate over-prediction occurs for the Oakland/Berkeley area. Part of Marin County is underestimated immediately below 1000. The Santa Rosa area at about 1150 is underestimated. There is an overestimate in the area around 1200. There is a modest underestimate for Sacramento. The extreme Northern California coastal area is underestimated. The desert area immediately before 500 is underestimated. Santa Barbara, which occurs in the 590s is underestimated.

The proximity complement performs similarly to the property damage liability frequency case. No credibility weighting interventions were urgently necessary.

3.3.4 Property Damage Liability Severity

The most striking bias occurs for southwest Orange County. A less severe spike occurs for Sacramento. There is a slight overestimate for part of San Diego County, which is plotted to the immediate right of the greater Los Angeles area. And Oakland/Berkeley is modestly underestimated. Inland extreme northern California appears to be modestly over-predicted.

The proximity again performs similarly. No credibility weighting interventions were urgently necessary.

3.3.5 Regression Model Conclusions

It would appear that, even with this very simplistic multiple regression approach, three of the four loss quantities were well handled with relatively few binary geographical variables. And, even for the somewhat more complicated BI frequency case, the model would do an adequate job with

the addition of a few more binary geographical variables and perhaps some reorganization of the model. It is quite likely that much of the regional bias in geographical BI frequency is due to unmeasured differences in the legal environment.

Obviously, a spatially autoregressive approach has the potential to improve the results, which we again leave to future researchers.

3.3.6 Proximity Complement Conclusions

Clearly the quality of the complement would be improved through a dynamically determined radius and weighting procedure. Larger radii appear to be in order for rural areas and smaller ones appear to be in order for urban areas.

3.4 A Nonlinear Programming Approach to Constrained Clustering

3.4.1 Introduction

As we stated in Section (2.2.4), Teboulle et al. noted the similarity between optimization cluster analysis and nonlinear programming. Given the lack of solutions available in the cluster analysis literature for *summation* and *averaging* constraints, we looked to nonlinear programming as a means of formulating constrained cluster analysis problems because operations research, of which nonlinear programming is a part, has constrained optimization as one of its central objects of analysis.

A simple description of the difference in the types of nonlinear mathematical programming programs can be found in the appropriate chapter of Hillier and Lieberman (1995) [60]. Both of our proposed objective functions, (2.5) and (2.6), are nonlinear and non-convex. Additionally, (2.6) is non-smooth in a small finite number of places corresponding to the breaking point for the absolute value function. These factors generally make the problem difficult to solve and guarantees of a globally optimal solution hard to come by.

Additionally, our decision variables are defined as *binary*. So what we have is a constrained non-convex pure integer programming problem.

Computationally intense approaches are required to ensure good solutions for this class of problems. It is the modeler's task to creatively specify the model in a manner that makes maximum usage of the structure present, increasing chances of success and decreasing computational demands.

3.4.2 Large Non-Convex Integer Programming Problems

As originally configured in (2.3), (2.4) or (2.5), (2.6) and (2.7), our problem is generally too large

to be solved in a reasonable amount of time.

The size of the problem can be significantly reduced and its structure made clearer with a few additional steps. First, the zip codes should be sorted by the mixed model indication, from smallest to largest. In that configuration, our decision variable x_{ij} runs from i=1 being the zip code with the smallest mixed model indication, to i=1,502 being the zip code with the largest indication. The fact that we are not giving consideration to the relative credibility of our mixed model indications is significant here. Credibility considerations would make the problem difficult to solve, although it might improve the end result.

We already have constraint (2.4), which ensures that only one decision variable in a row (for a zip code) can take on a "1" value, and all the remaining decision variables have to take on a "0". This means that the zip code can only be assigned to one band.

Combining this fact with the new sorted nature of the matrix, it also becomes clear that the column of "1"s in a good solution should generally march in discrete columns from left to right. Except in very limited instances, there should be no reason for the column of "1"s to move backward to the left.

After making this realization, we see that certain portions of the matrix are irrelevant. For instance, for low i values, the right hand part of the matrix is irrelevant, since in a good solution those values will always be "0".

It would seem that the size and complexity of the problem could already be reduced considerably given these considerations.

3.4.3 The Frontline Premium Solver[tm]

The R language we have been using up until this time does not currently have ready-made packages for dealing with non-convex optimization problems. And, the size of our problem exceeds the number of variables allowable in the Microsoft Excel Solver.

But as it turns out, the maker of Microsoft's Excel Solver has also made a commercial package available that handles larger problems. We elected to employ Frontline's KNITRO[tm] Solver using the Frontline Premium Solver Platform[tm].

This Solver employs one of three methods each time it conducts a minimization step. The first two are interior point algorithms, which are also known as barrier methods. The third method is known as an active-set method.

The conjugate gradient iteration interior point method employs a step to improve feasibility and a tangential step to improve optimality using a projected conjugate gradient iteration. The direct interior point method solves the primal-dual KKT system using direct linear algebra. The interior-point methods employed are described in Byrd, Gilbert, and Nocedal (2000) [56] and Byrd, Nocedal, and Waltz (2003) [58].

The active set method is a sequential linear quadratic programming technique. The first stage identifies those constraints that are "active" for a first solution of the problem. This solution involves solving a linear approximation within a trust region. The second stage involves an equality constrained quadratic approximation that incorporates only those constraints that were identified as active in the first stage. A projected conjugate gradient method is employed in the second stage. The active set methodology employed is outlined in Byrd, Gould, Nocedal, and Waltz (2004) [57].

Integer and binary problems also involve the use of the branch and bound method. As we shall discuss, the constraints imposed with the interior point methods sometimes lead to an overly-restrictive feasibility region when used in conjunction with the branch and bound method, and as a result the active-set method might need to be employed.[9]

3.4.4 Experimentation with Model Formulations

Reducing the Size of the Decision Variable Matrix

Starting with BI frequency, we began by dividing the matrix of decision variables into roughly equal length sections in terms of the number of zip codes. Then we pre-assigned the decision variables "0" or "1" values in discrete columns. The first set of zip codes, numbered i=1 to 148, were assigned to frequency band "1", which means that the first of the ten columns (j=1) were assigned the value "1" while the remaining columns (j=2 to 10) were assigned "0" values. For i=149 to 296, the column j=2 was assigned values of "1" while the columns corresponding to j=1 and j=3 to 10 were assigned values of "0". And so forth.

We found the problem was far too large to be solved so we began to pair down the number of variables by inspection eliminating those variables that would never be "1" in an optimal solution. This involved removing variables more than a certain distance from the "1" in its row. So, for instance, the cell at (1,10) was among the first removed, since certainly the zip code with the lowest mixed model indication was not going to be assigned to the highest frequency band. We removed

[9] See Frontline Systems, Inc., [59].

close to half of the variables using this approach and attempted to solve the problem, but it was still far too large.

Non-Decreasing Band Assignment Constraint

Through successive experimentation we found that the problem had to be restricted both in terms of width around the "trial solution" represented by our columns of "1" values, and in terms of the number of zip codes considered at one time (we could not consider all 1,502 zip codes at one time). We finally arrived at a system that yielded solutions in a reasonable amount of time, and which were relatively certain not to be significantly affected by the restrictions in the size of the individual problems solved.

In the process of successive experimentation, we also found that it was useful to require that the frequency band assignments march forward in the column-like fashion we expected. Imposing this constraint takes into account our knowledge of what an optimal solution has to look like, and saves computational time, since the algorithm will not have to investigate solutions that clearly are out of the range of an optimal solution.

We prevent the band assignments from moving "backwards" through the following system of constraints. Mathematically, we represent these constraints as

$$0 \leq \sum_{j=1}^{10} j\left[x_{(i+1),j} - x_{i,j}\right] \leq 1 \; for \; i \; from \; 1 \; to \; 1{,}501 \tag{3.5}$$

This corresponds to our entire original range of decision variables. When we reduce the size of the problem as we just outlined, we only need to consider constraint (3.5) in terms of this reduced range of possible *i,j* values.

3.4.5 The Final Model Formulation

Our final method of solution is a sequential one. We present our first model formulation next and then show the logic behind the sequential progression.

Initial Problem Formulation

We began by only considering the decision variable in the following limited range:

$$x_{ij} \; for \; i \leq 148, j \leq 2 \; and \; for \; 149 \leq i \leq 296 \; j \leq 3, and \; 297 \leq i \leq 444, 2 \leq j \leq 4 \tag{3.6}$$

The actual values in the initial solution we provide remain unchanged, that is in the first of the three ranges enumerated above, "1" values are assigned to the decision variable when $j = 1$, and "0" values are assigned to all the other decision variables in the range. In the second range, the decision

variables are assigned "1" values when $j = 2$, while the other decision variables in the range are assigned "0" values. And for the third range of variables, "1" values were assigned when $j = 3$, with "0" values assigned to all the remaining variables in the range.

Throughout the process, we elected to use the L_1 objective function (2.6), which converted to the range specified in (3.6) is

$$\min\left[\begin{array}{l}\sum_{i=1}^{148}\sum_{j=1}^{2}\left[abs\left(R_i-\frac{\sum_b x_{bj}R_bE_b}{\sum_c E_c x_{cj}}\right)x_{ij}E_i\right] \\ +\sum_{i=149}^{296}\sum_{j=1}^{3}\left[abs\left(R_i-\frac{\sum_b x_{bj}R_bE_b}{\sum_c E_c x_{cj}}\right)x_{ij}E_i\right] \\ +\sum_{i=297}^{444}\sum_{j=2}^{4}\left[abs\left(R_i-\frac{\sum_b x_{bj}R_bE_b}{\sum_c E_c x_{cj}}\right)x_{ij}E_i\right]\end{array}\right] \tag{3.7}$$

In our initial attempts, we thought we would wait before incorporating the minimum land area constraint (2.7). Should a solution ever be arrived at that violated or threatened that constraint, we could always move back a step and add it.

Sequential Procedure

The sequential procedure essentially involves moving downward and to the right through our original range of decision variables.

Initial Solution Stage

The first stage involves running the problem as formulated immediately above, using the Frontline KNITRO™ Solver on the Microsoft Excel™ implementation. We will discuss the parameters selected for the KNITRO™ Solver a little later.

As an example, our first solution of the problem as formulated immediately above assigned BI frequency band 1 to zip codes corresponding to *i* values of 1 to 116. BI frequency band 2 was assigned to zip codes corresponding to *i* values of 117 to 275. Band 3 was assigned to *i* values of 276-444.

Solution Check Stage

After the previous run of the KNITRO™ Solver, we check the stability of the solution under a different set of constraints. We keep the same band assignments ("1" values), but we modify the

range of decision variables somewhat.

First, we ensure that the leftmost (the lowest band under consideration) column of "1" values has no decision variables defined to its left.

For the next column of "1" values, or the next assigned band, we ensure that there is only one decision variable defined immediately to its left, and one to its immediate right.

We do the same for the proceeding columns of "1" values. So, the leftmost column under consideration cannot in the future move backward, while it can move forward one band. The remaining band assignments from the previous solution can move forward or backward a maximum of one band assignment.

With the same range of *i*-values under consideration, and a somewhat reconfigured set of *j*-values, we rerun the problem.

If we get the same result, then we move on to the next step in the "sequence". As we will explain further, moving forward in the sequence involves "dropping" the leftmost band from consideration, and adding a new segment of *i,j* values for consideration, corresponding to a downward and possible rightward movement on the right-hand side.

As it turns out, the solution check stage was unnecessary. The solution to the problem under new constraints always was the same as the previous solution. We conducted the solution check stage through the entire process for BI frequency, but abandoned it for the remaining frequency and severity analyses.

As an example, our solution check of our first initial solution was formulated as follows: For *i* from 1 to 116, and for *j* from 1 to 2, the decision variables were defined, with "1" values assigned when *j*=1 and "0" values assigned when *j*=2. For i from 117 to 275, decision variables were defined for *j* from 1 to 3. "1" values were assigned when *j*=2, and "0" values were assigned when *j*=1 or *j*=3. For *i* from 276 to 444, decision variables were defined from *j*=2 to *j*=4. "1" values were assigned when *j*=3, and "0" values were assigned when *j*=2 or *j*=4.

When we reran the problem, the same solution was generated; the first BI frequency band was assigned to *i* from 1 to 116, the second to *i* from 117 to 275, and the third from 276 to 444.

Sequential Advancement Stage

When the solution check yielded the same solution (which it always did), we essentially moved downward and to the right, dropping the lowest band (furthest to the left) from consideration and

adding a new range of decision variables to consider downward and to the right.

For the returning "bands", the band assignments from the previous solution remain unchanged. New zip codes, which have already been assigned values in our original trial solution are then added.

Defined decision variables follow the same general pattern, with the leftmost column of "1" values not having any decision variables defined to their left, thus restricting consideration to solutions that either maintain the band assignment, or increase it by one (moving one column over to the right). The remaining band assignments are allowed one decision variable to the right and left, so they are free to move forward or backward a band from their existing position.

After the solution is run for this problem, we move to the solution check stage and test this new result. If the result is the same we move forward again, dropping the lowest band and picking up one new one.

As an example, our first sequential advancement from the previous solution was as follows: for i=117 to 275, the decision variable was defined for j=2 and j=3 with "1" values assigned when j=2 and "0" values being assigned when j=3. For i=276 to 444, the decision variable was defined for j= 2 to j=4; when j=3 the decision variable was assigned a value of "1", and when j=2 or j=4 a value of "0" was assigned to the decision variable. For i=445 to 593, the decision variable was defined for j= 3 to 5, with "0" values being assigned when j=3 and j=5, and "1" values being assigned when j=4.

Reaching the Final Band

When sequentially advancing to the stage where final decision variables (when the rightmost and lowest cell under consideration is (max(i), max(j)) then a slight modification of the problem setup is in order. The treatment of all the ranges is the same except for the last one. Those cells that were previously assigned "1" values in the rightmost column are only allowed to have one decision variable defined to their left. And, by definition there are no decision variables defined to their right.

The result is checked once and that gives the final result.

3.4.6 Elected KNITRO™ Solver Parameters

Solution Method

As we discussed earlier, there are three solution methods available: the Direct Interior Point Method, the Conjugate Gradient (CG) Interior Point Method, and the Active Set Method. Interior Point Methods are also known as "barrier" methods.

The default setting is to allow the Solver itself to choose the best method as it proceeds during

the iterative solution process. We elected to keep this setting. As we will discuss later, there were two occasions where we had to modify our reliance on the default and make use of a particular solution method.

Global Optimization of Non-Convex Problems

When the problem is non-convex, as ours is, a truly optimal solution can often not be guaranteed, or can often not be guaranteed in a reasonable period of computation time (for integer programming problems). Integer programming problems can sometimes be solved with guarantees of global optimality, but often the amount of computing time necessary would be too high.

We will discuss integer programming in a moment. With respect to the non-convex aspect of our problem, a kind of brute-force method can be used to help increase the likelihood that the solution obtained is optimal or near-optimal. In KNITRO[tm] these parameters are known as Multi-Start Search and Topographic Search. The Multi-Start Search involves trying different randomly selected points from which to attempt solution of the problem. The Topographic Search option is essentially an add-on to the Multi-Start Search. From the point generated by the Multi-Start Search, the Topographic Search attempts to map the local terrain to determine the best starting point.

We elected both the Multi-Start Search and Topographic Search for the solution of our problems.

Automatic Scaling

Poor scaling in the problem formulation can reduce the precision with which the Solver can operate. Automatic scaling helps to handle some scaling problems, but is not a guarantee. We used the automatic scaling option when solving our problems.

Derivatives

The interior point methods work best when they can use analytic second derivatives. Analytic second derivatives could not be found for our problem, probably because of the absolute value used in the objective function. We did test our L_2 objective function early in the process and KNITRO[tm] was not able to find the analytic second derivatives to that problem either.

When analytic second derivatives cannot be found, KNITRO[tm] offers the option of using analytic first derivatives or finite differences. We elected analytic first derivatives.

The user is also given the option of using the default selection of forward derivatives or selecting central derivatives. We used the default.

Sparse Optimization

Our problem is quite large. For large sparse problems, the KNITRO[tm] solver "sparse" option can improve performance considerably. The Solver indicated our problems were all sparse, with a sparsity measure always well under 1%, so we always elected the sparse option.

Integer Tolerance

When solving integer programming problems, the branch & bound method can solve to a predetermined level of tolerance from true integer values, when testing for optimality. The default setting is 0.05, which we did not change. If one were to select "0", it is possible that the Solver could arrive at a guaranteed globally optimal solution, although it might take quite a while.

Remaining Parameters

We employed all the remaining default parameters. The most significant of these involve tolerance levels.

3.4.7 An Example of the Process

To illustrate the solution process, we present the complete sequence of problem setups and solutions below in a simplified tabular form for BI frequency:

	i range	FB1	FB2	FB3	FB4	FB5	FB6	FB7	FB8	FB9	FB10
Setup1	1 to 148	1	0								
	149 to 296	0	1	0							
	297 to 444		0	1	0						
Solution1	1 to 116	1									
	117 to 275		1								
	276 to 444			1							
Setup2	117 to 275		1	0							
	276 to 444		0	1	0						
	445 to 592			0	1	0					
Solution2	117 to 276		1								
	277 to 453			1							
	454 to 592				1						
Setup3	277 to 453			1	0						
	454 to 592			0	1	0					
	593 to 740				0	1	0				
Solution3	277 to 474			1							
	475 to 628				1						
	629 to 740					1					

	i range	FB1	FB2	FB3	FB4	FB5	FB6	FB7	FB8	FB9	FB10
Setup4	475 to 628				1	0					
	629 to 740				0	1	0				
	741 to 888					0	1	0			
Solution4	475 to 637				1						
	638 to 766					1					
	767 to 888						1				
Setup5	638 to 766					1	0				
	767 to 888					0	1	0			
	889 to 1036						0	1	0		
Solution5	638 to 794					1					
	795 to 927						1				
	928 to 1036							1			
Setup6	795 to 927						1	0			
	928 to 1036						0	1	0		
	1037 to 1184							0	1	0	
Solution6	795 to 928						1				
	929 to 1067							1			
	1068 to 1184								1		
Setup7	929 to 1067							1	0		
	1068 to 1184							0	1	0	
	1185 to 1332								0	1	0
Solution7	929 to 1084							1			
	1085 to 1220								1		
	1221 to 1332									1	
Setup8	1085 to 1220								1	0	
	1221 to 1332								0	1	0
	1333 to 1485									0	1
Solution8	1085 to 1223								1		
	1224 to 1339									1	
	1340 to 1485										1

The solutions contain only the values of those decision variables that were assigned to a particular band (only those with "1" values). We did not include setups and solutions for the "solution check" stage since in each case, the solution found did not change from the previous solution.

The setups contain all of the defined decision variables and their pre-assigned values (before the solver is applied). The ten columns, corresponding to the ten bands, also correspond to *j* values from 1 to 10.

3.4.8 Clustering the Remaining Frequency and Severity Bands

For the most part, we were able to use the same KNITRO[tm] parameters, and the same general process in solving the other three problems. There were a few problem areas, some of which actually serve to highlight the relative strengths of the interior point methods versus the active set method.

The PD frequency clustering was as uneventful as the BI frequency clustering.

Severity Band Clustering

The severity clustering processes developed two complications not encountered with frequency: four band solutions and the inability to find feasible solutions.

Four Band Solutions

First, solutions moved to take up four bands in many of the solution steps. For example, the first solution for BI severity was as follows: *i* from 1 to 76 was assigned to Severity Band 1, *i* from 77 to 212 was assigned to Severity Band 2, *i* from 213 to 352 was assigned to Severity Band 3, and *i* from 353 to 450 was assigned to Severity Band 4.

This did not really present a challenge. We formulated the next setup in the same way, with Severity Band 1 dropped, and both the previous solution values for Severity Band 4 and the new segment, which was assigned to four, being introduced into the setup. The system of surrounding all but the leftmost column of "1" values with "0" values corresponding to defined decision variables, and the leftmost column of "1" values having only a single column of decision variables, coded to "0" immediately to its right.

Inability to Find Feasible Solutions

One of the drawbacks of using the two interior point methods in combination with integer programming problems is that the feasible region drawn by the algorithm may be too restrictive for the branch & bound method to operate properly. While we elected the default value for the solution method, which allows the Solver to choose the best of the three methods, there were two instances where we did have to intervene.

For PD severity, on setup3, repeated attempts yielded the result that a feasible solution could not be found. This must have been an instance where one of the two interior point methods was being used but was drawing too tight a boundary for the branch & bound algorithm to operate in.

In response, we manually selected the active set methodology. Under that method, the algorithm

ran for a much longer period than we had seen before for our reduced-size problems. We could see that, at each iteration, the solver was making very slow progress, but measurable progress nonetheless. At that point, we elected to stop the algorithm, maintaining the intermediate solution that it had come to at that point. We then ran the algorithm with the default solution parameter set that allows the Solver to choose the appropriate method. That approach yielded a solution in a reasonable amount of time.

The problem repeated itself on the eighth and final setup, and we used the same procedure, but only allowing the active set method to run for a shorter period of time to an interim solution.

3.5 Final Results

Detailed information for BI frequency, PD frequency, BI severity, and PD severity has all been placed on the CAS Web Site. This detailed information includes each of the mixed model components, the credibility assigned to each component, the mixed model estimate, and a comparison of the new band assignment with the Hunstad (April, 1996) [18] band assignment. In the present section, we present summary statistics to evaluate the performance of our approach. In the following tables, we present a comparison of the mixed model average indication for each band with the actual indication, to indicate bias that exists at respective hazard levels. Furthermore, we present the indicated relativity for corresponding Hunstad bands. Below this, we compare the mean absolute deviation for the new bands against the same statistic for the Hunstad bands. A discussion follows.

3.5.1 Statistics for Final Band Assignments

BI Frequency

	FB1	FB2	FB3	FB4	FB5	FB6	FB7	FB8	FB9	FB10
					Relativities					
Mixed Model	0.5438	0.6180	0.6730	0.7253	0.7866	0.8602	0.9870	1.1386	1.3374	1.7544
Actual	0.4895	0.5775	0.6589	0.7232	0.7882	0.8619	0.9940	1.1488	1.3472	1.7708
Hunstad	0.5334	0.6715	0.7456	0.8037	0.8767	0.9795	1.0752	1.1856	1.3425	1.7393
					MAD					
New Cell	0.00105	0.00092	0.00047	0.00039	0.00037	0.00045	0.00058	0.00071	0.00109	0.00315
Hunstad Cell	0.00121	0.00041	0.00034	0.00029	0.00048	0.00035	0.00052	0.00052	0.00086	0.00319
New Total					0.00087					
Hunstad Total					0.00083					

PD Frequency

	FB1	FB2	FB3	FB4	FB5	FB6	FB7	FB8	FB9	FB10
					Relativities					
Mixed Model	0.6548	0.7265	0.7853	0.8423	0.9171	0.9663	1.0127	1.0598	1.1247	1.3036
Actual	0.6132	0.7137	0.7827	0.8423	0.9173	0.9671	1.0140	1.0613	1.1271	1.3102
Hunstad	0.7301	0.8634	0.9297	0.9642	0.9965	1.0219	1.0492	1.0740	1.1117	1.2430
					MAD					
New Cell	0.00223	0.00094	0.00081	0.00074	0.00067	0.00049	0.00047	0.00044	0.00114	0.00299
Hunstad Cell	0.00261	0.00129	0.00048	0.00042	0.00030	0.00027	0.00027	0.00029	0.00060	0.00318
New Total					0.00082					
Hunstad Total					0.00097					

BI Severity

	SB1	SB2	SB3	SB4	SB5	SB6	SB7	SB8	SB9	SB10
					Relativities					
Mixed Model	0.8297	0.8777	0.9026	0.9267	0.9499	0.9805	1.0136	1.0422	1.0761	1.1268
Actual	0.8224	0.8728	0.8985	0.9253	0.9508	0.9833	1.0154	1.0427	1.0765	1.1293
Hunstad	0.8380	0.8902	0.9202	0.9525	0.9792	1.0049	1.0232	1.0445	1.0675	1.1156
					MAD					
New Cell	207.61	129.62	91.92	87.93	87.16	124.18	90.86	92.81	100.82	206.48
Hunstad Cell	229.64	100.22	113.12	158.01	210.82	171.97	139.16	144.30	145.46	243.90
New Total					117.85					
Hunstad Total					168.71					

PD Severity

	SB1	SB2	SB3	SB4	SB5	SB6	SB7	SB8	SB9	SB10
					Relativities					
Mixed Model	0.8387	0.8770	0.9078	0.9346	0.9615	0.9905	1.0181	1.0423	1.0803	1.1487
Actual	0.8355	0.8755	0.9076	0.9349	0.9625	0.9909	1.0181	1.0421	1.0807	1.1503
Hunstad	0.8505	0.8989	0.9406	0.9771	0.9983	1.0155	1.0283	1.0449	1.0700	1.1303
					MAD					
New Cell	28.94	11.79	11.40	12.11	13.73	12.68	8.33	9.46	19.58	35.53
Hunstad Cell	29.83	18.28	20.34	12.95	10.06	5.18	7.54	8.00	14.25	42.84
New Total					14.67					
Hunstad Total					17.01					

3.5.2 Other Quantities

Basic Constraints

The minimum 20-square mile requirement for bands came nowhere near being reached. We also checked to ensure that each band contains a credible amount of experience, and again nothing even close to a problem emerged.

Factor Weights

Under the proxy weighting methodology promulgated by the CDI, a "relative" factor weight can be computed for our results and related to relative factor weights on the marketplace and also that might be projected to be necessary under the regulations that are soon to take full effect.

"Relative" factor weights can be computed in terms of our formulation as follows:

$$\frac{\sum_i \sum_j \left[abs\left(\frac{\sum_b x_{bj} R_b E_b}{\sum_c E_c x_{cj}} - 1 \right) x_{ij} E_i \right]}{\sum_d E_d} \tag{3.8}$$

Using this formula our relative factor weights for each of the four bands are as follows:

- BI frequency: 0.2701
- PD frequency: 0.1014
- BI Severity: 0.0705
- PD Severity: 0.0629

An individual company's factor weights can be converted to relative factor weights by dividing out the base rate and the total number of exposures. Relative factor weights can then be compared on an apples-to-apples basis with our relative factor weights.

3.6 Analysis of Final Results

3.6.1 Mean Absolute Deviation Comparison

It would appear that the mixed model with clustering approach outperformed the Hunstad (April, 1996) approach for PD frequency, and both measures of severity, using mean absolute deviation as the basis of comparison.

For BI frequency the Hunstad assignments modestly outperform mixed models with clustering.

The mixed model outperforms the Hunstad result for bands 1 and 10, with results for the first band significantly better.

For PD frequency, mixed models with clustering moderately outperformed the Hunstad result. Our approach again outperformed for bands 1 and 10.

For BI severity, our approach significantly outperformed the Hunstad result, and again outperformed in bands 1 and 10.

For PD severity, our approach moderately outperformed the Hunstad result, and again outperformed for bands 1 and 10.

3.6.2 Mixed Model Bias by Hazard Level

The mixed model shows bias in the first band for BI frequency, and to a lesser extent, for PD frequency. However, for BI frequency, the net effect of the mixed model and clustering appears however to be greater separation and greater accuracy for the lowest hazard band. Much of this band comes from extreme northern California, where we observed significant bias in the regression model, and where the proximity complements might be made larger.

The bias for the first band in PD frequency was moderate. There was little regional bias to speak of for the remaining frequency and severity bands.

3.6.3 Constraints

Each band vastly exceeds the minimum required land area of 20-square miles. Each band also contained extremely credible quantities of data.

So it would appear that the introduction of constraints made our search for a solution easier rather than more difficult (in particular (3.5)).

3.7 Directions for Future Research

We can see several separate prongs of research emanating from this paper.

First, within the scope of the existing framework of territory analysis, the implementation of the concept we have introduced could certainly be improved. This would require the devotion of individual attention to the arithmetic model, the proximity complement, credibility weighting of the three mixed model components, and the automation and possible methodological improvement of the cluster analysis technique.

Next, it would seem to make sense to begin to move territory analysis forward with the

introduction of new causal geographical rating variables. As the arithmetic model and proximity complement are improved within the existing framework of territory analysis, it seems to us that the groundwork could be laid for the introduction of new rating variables that would address several issues in territory analysis. The introduction of rating variables such as *traffic density*, *claims environment*, and *traffic enforcement* could strengthen geographical rating from claims that is not a causal rating variable. Furthermore, by introducing these variables as *continuous* measurements, say for each zip code, they could be properly integrated into the parameterization of the remaining parts of the classification plan, helping to alleviate the current disjointed relationship between the two.

The use of constrained cluster analysis as a potential alternative to pumping and tempering to achieve factor weight compliance could also be investigated in California. It is possible that a procedure could be arrived at that would not be viewed as arbitrary by the courts.

Also in California in particular, it would seem to make great sense to introduce new geographical rating variables under the new Proposition 103 regulations soon taking full effect.

3.8 Discussion of Mixed Models in Territory Analysis

With even this crude implementation of our concept, we see that for three out of four territory bands, our method outperformed the method initially used to form the California Personal Automobile Frequency and Severity Bands Manual under Proposition 103. Furthermore, the implementation was completely objective.

Individual attention to three elements of the mixed model could substantially improve the result. We would suggest the following separate lines of research.

3.8.1 Refinement of the Arithmetic Model

As we discussed earlier, even within the framework of the simplistic and somewhat inappropriate multiple regression model form, substantial improvements for BI frequency could probably be obtained by identifying better variables related to legal environment. And even failing that, the simple introduction of a handful of binary geographical variables could substantially improve the result. In particular we feel that there is the potential to dramatically improve the territory analysis of the lowest frequency, sparsely populated areas of northern California.

Perhaps an area of even greater promise is the spatially autoregressive model. This type of model is used in geography, and is almost certainly more appropriate than the multiple linear regression we employed. See Bailey (1995) [37] for a fairly gentle introduction to spatial statistics with software and

data. Another such introduction, in the R language, is provided in Crawley (2007) [40]. For a very theoretical treatment, see Cressie (1993) [41].

Given further refinements in the variables employed and the form of the model, we are quite confident that the results we reported here can be substantially improved upon.

In addition to improving the mixed model result, improved models of causal geographical variables could hasten the introduction of new, causal and continuous rating variables that will be both more acceptable to regulators and the public and easily integrated into the parameterization of the remaining elements of the classification plan.

3.8.2 Refinement of the Proximity Complement

The proximity complement should become a formal area of study under territory analysis. Up until now Hunstad (April, 1996) [18] and Tang (2005) [21] provided the only two papers dealing with the topic substantially. In our study, we found that in sparsely populated areas optimal complements should have more than a ten-mile radius, while in the center of the city a shorter radius is in order. This really is a matter of common sense when looking at these extremes. However, deriving methods that will dynamically generate *optimal* proximity complements for *each* atomic geographical unit based on all of the relevant local information would seem to be a nontrivial task deserving of some future research.

Another possible approach would be to employ a form of cluster analysis that allows for overlapping clusters (where an individual zip code may be included in more than one proximity complement).

Another approach, would be to use the results of a spatially autoregressive model to arrive at indications for the zip codes in the proximity complement. The model would not need to incorporate auxiliary variables. The model could incorporate gradients, which could be a great advantage.

Another possibility, although perhaps it would be too unwieldy, would be to incorporate spline or graduation information into a proximity complement. One problem to solve would be how to treat the data from the geographical unit being complemented, since without adjustment it would be counted twice.

In addition to the actual selection of units to include in the complement and their valuation, attention should be given to the means by which the resulting complementary indication is computed, as it might make sense to weight the results on some non-obvious basis or perhaps use

an arithmetic average instead of a weighted average. Perhaps a population weighted latitude and longitude (at the block level) should be computed for the proximity complement, and used to adjust the complement, either in terms of the geographical units included or the method in which they are weighted. It would seem that ideally such a measure should fall at or near a similarly computed center for the geographical unit being complemented.

3.8.3 Refinement of the Credibility Weighting Scheme

A formal credibility analysis should be conducted to arrive at better methods of credibility weighting the results. Ideally the local geographical fit should influence the weight for both the arithmetic model and the proximity complement. Additionally perhaps the local fit in terms of the auxiliary variables in the arithmetic model should also influence the result.

3.9 Refinement & Automation of Constrained Cluster Analysis

3.9.1 Refinement

One competing method of nonlinear programming should be investigated. Although we did not have much luck in our initial experiments, the Large-Scale SQP[tm] solver engine from Frontline Systems, Inc. has a particular feature of interest for problems with our structure.

The structure in question involves binary decision variables constrained in the manner of (2.4). This type of constraint is known as a special ordered set (SOS). Williams (1999) [62] indicates that this method was introduced in Beale and Tomlin (1969) [54].

This methodology is incorporated in the Large-Scale SQP[tm] (Sequential Quadratic Programming) Solver, along with several other methods associated with integer and binary programming. Particularly if one wants to attempt to increase the size of the problems we analyzed sequentially (particularly widthwise), or even solve the whole thing at one time, such methods should be investigated to see how they perform against the methodology employed in the KNITRO[tm] Solver. Generally speaking, outside of the SOS method we mention, the Large-Scale SQP[tm] Solver operates on principles similar to those employed under the active-set methodology under KNITRO.[tm]

3.9.2 Automation

We employed the Frontline Systems, Inc., implementation that plugs in directly into the Microsoft Excel[tm] Solver. This made it easy for us to learn and experiment. The sequential procedure involving the solution and setup of sub-problems is cumbersome when performed manually. Frontline offers implementations where we are certain the procedure we employed, or any variant,

could be fully automated relatively easily. When so automated, the method would be incredibly efficient, dramatically improving the productivity of those involved in large-scale territorial revisions for many states.

3.10 Introduction of New Geographical Rating Variables

3.10.1 Traffic Density

Traffic density is probably the leading candidate for introduction as a causal geographical variable. It has consistently been considered a causal factor in accidents for at least ninety years.[10] This includes instances where territory has been criticized as a simple proxy for truly causal factors like traffic density.

Arriving at the best measure of density is the main challenge in implementing this as a rating variable.

In the past, measures of the denominator typically employed quantities of road lane. However, these figures were only tabulated at the county level, which introduces a significant degree of inaccuracy. At this point, we do not see a better alternative to using simple land area. In the course of our study we investigated the use of populated land area (defined as the land area for all census blocks that contain at least one inhabitant), on the suspicion that a simple land based measure would not reflect land on which essentially no commuting takes place. However, we found the standard measure and the new measure to perform at about the same levels.

The best current source of a numerator of for any density measure is probably the census bureau. However, new sources of information could soon become available.

Census Bureau Data

Our measure of the numerator focused on the total commute minutes one-way, partly because it was easier to measure and partly because it is most relevant. All of this information came from the 1990 decennial census.

Significant traffic congestion generally only occurs during the common commute hours. Furthermore, the relative traffic density between geographical units during the commute hours is probably maintained to some degree by density at other times of the day.

[10] Michelbacher (1918) [7], Dorweiler (1930) [4], Whitney (1941), Phase I (1978) [19], Phase II (1979) [20], Stone (1978) [35], Shayer (1978) [34].

If commute traffic density is to be introduced as a unitary measure, as opposed to the three radial measures we employed, then the appropriate radius will need to be selected. Determining the average commute density for a geographical unit like a zip code by using only data from the commuters within that zip code would likely be inferior to a broader measure that incorporates the true spatial interaction that exists.

Several additional decennial census measures should be considered to potentially improve the numerator. The census contains information on the number of vehicles used in commuting, including those used in car-pools, and also the number of persons taking mass transit. A measure that is responsive to such variations would be ideal. However, the vehicle measures are not cross-tabulated against the temporal measures of time spent commuting one-way. We feel time spent commuting is the most important measure, since geographical variation in temporal length of commute is probably much greater than variation in the use of public transport and car-pools. To combine the measures, one would either have to assume they are independently distributed, or one would have to find some means of imputing the cross-tabulated distributions.

In addition to the vehicle-related measures, the decennial census contains an additional commuter variable of interest; the hour in which each commuter leaves for work. Variations in this measure could also influence density to some degree. However, once again it is not cross-tabulated with the other measures. Furthermore, incorporation would be exceedingly complex. The same comments apply with regard to independence or imputation.

Another interesting measure from the census bureau tabulates the number of workers in various industries at the location where they work as opposed to where they reside. We obtained this data from the 2005 survey of economic conditions, and used it to derive the numerator of our lawyer density measure. This survey could also be used to essentially compute the "demand" for workers. This could be laid in some relation to the "supply" of commuters taken from the decennial census to arrive at an improved estimate of average density. Aside from the mismatch between the 2005 date of the survey and our data (which we deemed to be tolerable in our measure of lawyer density), the level of complexity of such an analysis exceeds the scope of this paper. But it may well be worth investigating whether the "directional" nature of the information that could be gleaned from such a study could be used to improve measures of density.

Finally, our density measure involved "rings" around each zip code being modeled, and considers only commuters who reside in those rings when computing the quantities. It is possible that considerably more complex models could be used to compute traffic density. It is important to

remember that the quantity of interest is the traffic density to which vehicles inside the zip code being modeled will be exposed to, not necessarily the traffic conditions that exist in their own zip code. Coming up with a way to more accurately model the flow of vehicles might involve the use of spatial statistics.

GIS Data

Although it may not quite be ready yet, it is likely that accurate traffic density measures will soon be computable from the vast and growing information storehouse being created by position-aware devices in cell phones, vehicles, and the like.

Remote Sensing

Remote sensing data that physically measures density at various sites may also soon become more widely available.

Ensuring Acceptable Measurement

In a competitive marketplace, there will be the obvious incentives to determine the most effective measurement of traffic density. In heavily regulated markets that may restrict the use of territorial rating variables, there may well be a need for regulators to either determine standard measures of density for each geographical unit, or the appropriate standards by which such measures can be created.

3.10.2 Traffic Enforcement

It is commonly accepted that increased enforcement reduces accidents. Phase II (1979) [20] attempted to measure the impact that enforcement has on accident rates through a measure called the *enforcement ratio*. The enforcement ratio, as computed in Phase II, involved measuring the relative frequency of bodily injury accidents to all violations and accidents.

Since that time, authors such as Feldblum (1993) [16] and Conners and Feldblum (1997) [15] have pointed to data that show that many bodily injury liability claims appear to be elective soft-tissue injury claims, and that the propensity to make such claims successfully varies significantly by area.

We think it would be extremely worthwhile to re-investigate an enforcement using property damage liability accidents in lieu of bodily injury liability accidents. And perhaps other measures of enforcement could be derived.

A relative index of traffic enforcement might well be considered a causal variable and be deemed

controllable through the local government. Additionally, it would provide economic incentives for actions that reduce the number of accidents.

The use of a measure like the enforcement ratio has advantages over other measures such as local citations issued or enforcement expenditures. The enforcement ratio already implicitly reflects spatial interaction, so no adjustments in that regard would be necessary from an actuarial perspective.

Although such an undertaking would be laborious if done manually, all of the data necessary to conduct such a study using property damage liability accidents is available in the appendices of the Phase II study. Obtaining a fresh data set from the DMV would be an even better alternative.

Were a good measure of enforcement be shown to have a significant relationship to loss we think it would be an excellent candidate for early introduction as a geographical rating variable.

3.10.3 Legal Environment

We were the least successful with our approach in dealing with BI frequency. And this is the problem most affected by the legal environment.

Although there are remaining difficulties with the introduction of legal or claims environment as a causal geographical rating variable, we mention it because it likely has such a great impact on bodily injury liability loss costs.

Legal or claims environment might only be a good candidate for introduction in heavily regulated jurisdictions after several other causal geographical variables have successfully been introduced. In the meantime, improved measures of lawyer density, perhaps using the actual number of personal injury attorneys or perhaps using certain forms of medical specialists, should be researched. Additionally, analysis of differences by court jurisdiction might be useful, although the use of binary geographical variables corresponding to legal jurisdictions would not promote integration of territory analysis with the parameterization of the remainder of the classification plan.

3.10.4 Medical and Repair Cost Indices

These factors probably influence losses less, but may be easier to implement quickly in heavily regulated jurisdictions. If a relationship can be established to an accurate index, we think it would be relatively difficult to argue against their causality. A search for granular indices of these costs would be of interest in developing these causal geographical variables for BI and PD severity (in addition to severity for other coverage parts not addressed in the present study).

3.11 Refinements to California Personal Automobile Ratemaking

3.11.1 A New Frequency and Severity Bands Manual For California

In California, it would seem that the *Private Passenger Automobile Frequency and Severity Bands Manual* could be updated with the release of more recent data from the same source, such as was used in Tang (2005) [21]. In addition to the use of new data, the use of a mixed model technique, or Tang's new proximity complement might be in order. To promote stability and give carriers time to adjust, carriers could be given a choice of using either the new *Manual* or the old *Manual* as a credibility complement for a short period of time.

3.11.2 An Alternative to Pumping and Tempering in California

When the new Proposition 103 regulations take full effect in the near future, the factor weights for frequency and severity bands will have to fall below the factor weight for years of driving experience. This may force some insurers to reduce the scope of influence of relative frequency or severity in their rating plan.

Currently, a procedure exists called pumping and tempering, which provides a means by which the years of driving experience (or any other mandatory factor with a weight that is "too small" under the regulations) factor can be increased (pumped) in its scope, and/or relative frequency or severity (or any other factor with a weight that is "too large" under the regulations) can be decreased (tempered) in scope. The courts have criticized this procedure as arbitrary.

Introducing factor weight as a constraint in the cluster analysis procedure is an alternative. In this case we would set an upper bound on the relative frequency or severity factor weight equal to the factor weight for years of driving experience.

A factor weight constraint in our formulation would simply involve constraining (3.8) as follows:

$$\frac{\sum_i \sum_j \left[abs\left(\frac{\sum_b x_{bj} R_b E_b}{\sum_c E_c x_{cj}} - 1 \right) x_{ij} E_i \right]}{\sum_d E_d} \leq M \tag{3.9}$$

where M is the constant. A difficulty would be involved in that the constraint should be incorporated over the entire range of the problem. For computational reasons we solved the original problem in a series of steps that breaks the problem into pieces. This sort of approach might not work to arrive at an optimal constrained solution since the constraint should operate on the whole range of zip codes at the same time. An investigational attempt to implement this form of constraint would be of interest.

3.11.3 The Introduction of New Causal Geographical Rating Variables in California

The Underpinnings of Proposition 103

Proposition 103, which passed as a referendum in 1988, can be thought of as the California culmination of events that began with the publication of Casey et al. (1976) [26]. The Proposition's intellectual underpinnings seem to be traceable to Casey et al., the subsequent events and publications[11] associated with the revisions to the Massachusetts ratemaking procedures in 1978, and the publication of Phase I (1978) [19] and Phase II (1979) [20] by the CDI.

Proposition 103 requires that driving record be made the most important rating variable, and suggests that territory should be made much less important. This clearly mirrors the proposal in Ferreira (1978a) [28] and the subsequent Massachusetts experience. Proposition 103's use of years of driving experience as a rating variable, and prohibition of the use of age clearly mirrors the proposal in Shayer (1978) [34] and subsequent adoption in Massachusetts. The system of factor weights, which sometimes requires that rates be tempered, bares resemblance to the asymmetrical pricing introduced in Ferreira (1978b). And clearly these Massachusetts papers and developments drew heavily on the SRI Report of Casey et al. So the link seems pretty clear. Phase I can clearly be seen as a precursor in that it developed a "band" system of territorial rating that was emulated in the regulations used to implement the Proposition. And portions of Phase II were clearly in direct response to Casey et al.

Objections to Territory Immediately Preceding Proposition 103

The central argument against territorial rating by Proposition 103's precursors was its lack of causality and perceived arbitrariness.

Casey et al. argued, among other things, that territorial ratemaking was easy to criticize because of the subjective procedures used in grouping together geographical units into territories. It was argued that this arbitrariness could result in unfairly discriminatory rates, which did not reflect actual loss propensity. Phase I (1978) [19] seconded the concern about the arbitrariness with which territorial definitions were drawn up.

Shayer (1978) [34] criticized territory for not being a causal rating variable, stating that it was a surrogate for truly causal forces such as traffic density and road quality. We can see that this criticism of a lack of causality is crucial by examining the paper's discussion of other rating variables. For

[11] For instance, Shayer (1978) [34], Ferreira (1978a) [28], and Ferreira (1978b) [29], Change and Fairley (1978) [27] and Stone (1978) [35].

instance, despite the fact that years of driving experience is largely beyond the control of the insured just as age is, Shayer advocates its use in lieu of age, arguing the plausibility of the causal relationship between experience and loss propensity. Also, years of driving experience was deemed acceptable despite the fact that, by the line of reasoning she used in relation to age, years of driving experience would have no incentive effect. So it seems clear that causality was a determinative factor.

In California itself, Phase I (1978) [19] largely took the industry to task because it had failed to explain *why* geography had a significant impact on loss costs, and essentially argued that the industry had brought the then present state of affairs upon itself by not responding to the public's growing demand to know why they were being charged particular premiums. The authors of Phase I even unsuccessfully tried to relate geographical loss costs to causal geographical variables such as population density, topography, road quality, and weather. Although not explicitly arguing for the introduction of causal geographical rating variables, Phase I was arguing that if some form of analysis showing the true causal geographical forces at work on territorial loss costs were not forthcoming, the existence of territorial rating might be imperiled.

Also in California, in Phase II (1979) [20], the authors attempted to draw a link between a causal geographical variable—traffic enforcement—and geographical loss costs. They also attempted to analyze the impact of differences in the classification distribution on geographical loss costs.

Basis for Introducing Causal Geographical Rating Variables

By eliminating the determinative objection, which involved a lack of causality, on the basis of Proposition 103 and its associated regulations themselves and on the basis of factors we have pointed out earlier in our study, it seems that it would be worthwhile to investigate the introduction of new causal geographical variables into the personal automobile classification plan.

Under the Proposition, the California Insurance Commissioner has the power to introduce new rating variables that have been demonstrated to have a "substantial relationship to the risk of loss." Currently, two such geographical rating variables exist – relative claims frequency and relative claims severity.

Since in Shayer, and virtually everywhere else, it is explicitly recognized that traffic density is a causal geographical rating variable, and since lack of causality seems to have been such an important concern in the prelude to Proposition 103, if a suitable method of measuring traffic density at the zip code level could be agreed on, it could be introduced as a rating variable by the commissioner. Zip codes with similar traffic densities could be grouped via an objective means like cluster analysis,

or the existing manual methods of grouping frequency and severity bands could continue to be employed. Sequential analysis of the resulting bands would seem to be an easy enough process.

Since the CDI itself commissioned the earlier study of the enforcement ratio in Phase I, an investigation and enforcement ratio based upon property damage liability claims would seem to be in order. The non-actuarial rationale for the introduction of such a causal geographical rating variable is overwhelming because of the potential for loss prevention incentives.

The introduction of medical and repair cost indices at the zip code level, if they could be related to loss severity, would also seem to be uncontroversial candidates for introduction as causal geographical variables for the appropriate coverage parts.

As causal geographical variables are introduced, the more "undesirable" geographical variation in frequency and severity, with no known cause, would be captured in the relative frequency and severity bands. Perhaps in tandem with or shortly following the introduction of causal geographical rating variables, the scope of relative frequency and relative severity, which would become nothing short of unexplainable geographical variation in loss costs, could be reduced even further than it is now, in effect even further achieving the objective of the Proposition in the first place. For instance, the sum of the factor weights for relative frequency and relative severity could be required not to exceed the factor weight for years of driving experience. Or, perhaps the relative frequency and severity factor weights could be restricted in relation to the size of the smallest causal geographical rating variable.

What seems clear is that the introduction of causal geographical rating variables, combined with reductions in the scope of relative frequency and severity, would improve accuracy and further achieve the objectives of Proposition 103.

4. CONCLUSIONS

Our mixed model with clustering approach to territory analysis, which is entirely objective, generally outperformed the existing Proposition 103 California Frequency and Severity Band Manual in terms of mean absolute deviation. This is impressive because the implementation of the new concept was rudimentary.

Significant further work can be done on improving each of the elements of the mixed model, which would substantially improve the accuracy of the result. Modest improvements in the constrained cluster analysis may also yield additional marginal improvements in accuracy.

And after the method is fine-tuned and has matured, it would be a relatively easy matter to automate the sequential piecewise procedure employed in the cluster analysis. In that format, the approach could become extremely efficient, relative to the manual procedures currently involved when extensive territorial refinements are conducted.

The causal analysis of geographical variation in loss costs, which could ensue from our approach, could pave the way for the introduction of new causal geographical rating variables. In addition to eliminating criticisms regarding causality and potentially invigorating local loss prevention initiatives, this group of largely continuous variables could be integrated with the parameterization of the remaining classification plan via the extensive array of predictive modeling procedures that are being employed for that purpose.

Moving forward to a more causally based method of territory analysis will in turn better prepare us for the revolutionary ratemaking changes in automobile insurance that are sure to come as the means for incorporating data from mobile position-aware devices come into being.

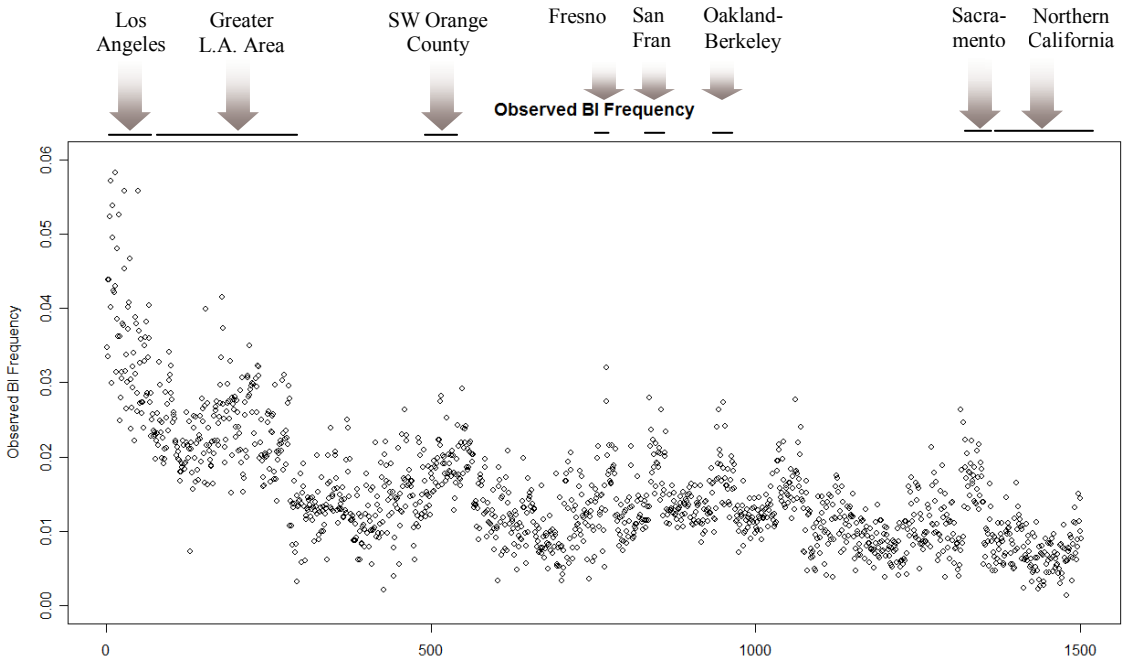
Acknowledgment

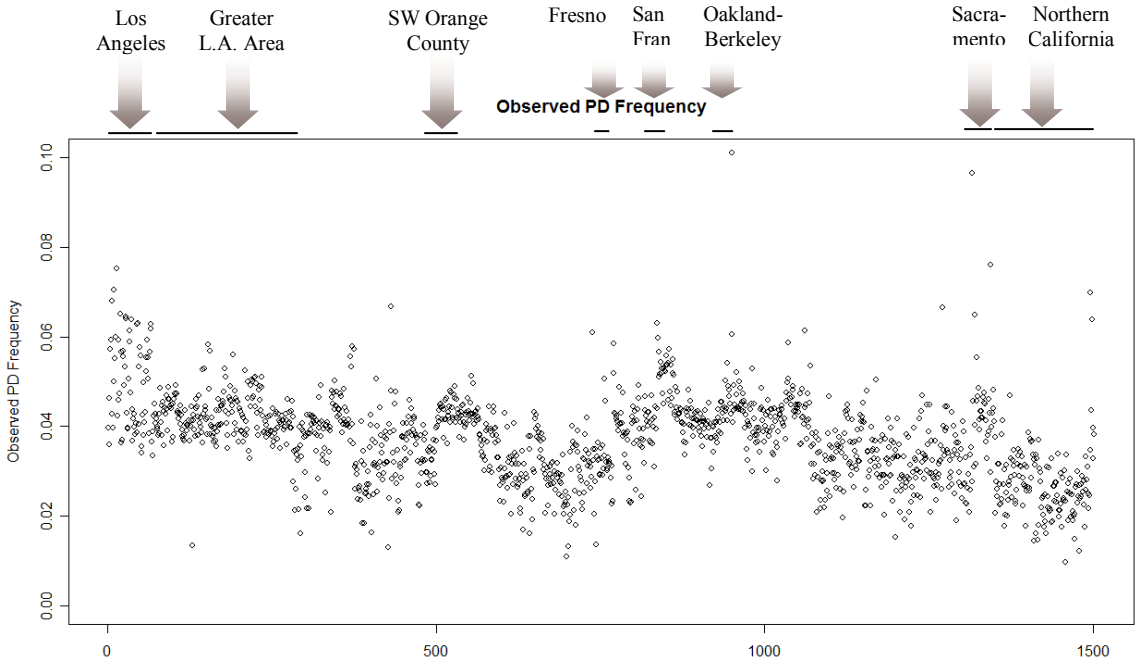
The authors thankfully acknowledge the useful comments provided by the reviewers. Any errors that may remain are their own.

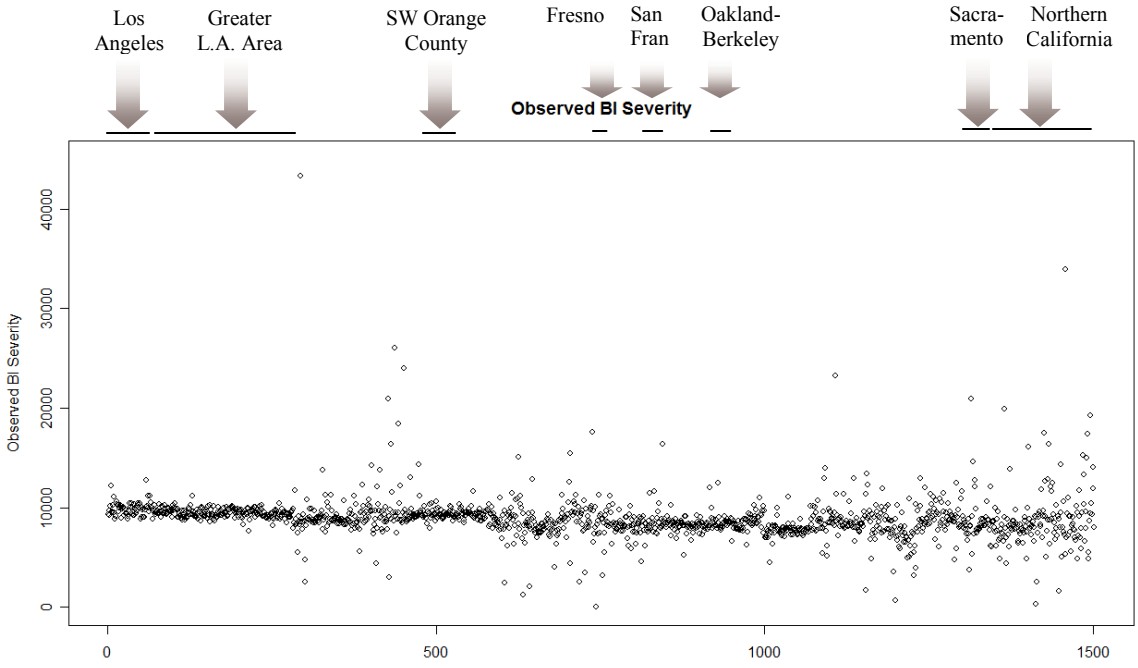
Supplementary Material

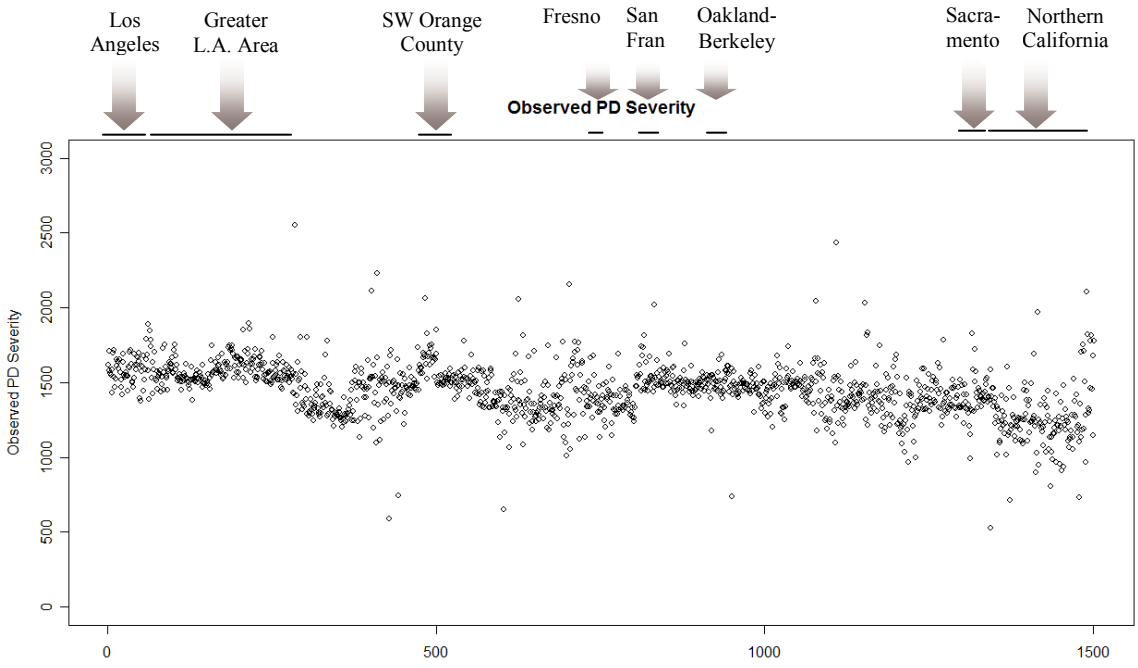
Tables containing a side by side comparison of the frequency and severity bands assigned in the present study and in Hunstad (April, 1996) [18] are available electronically on the CAS Web Site.

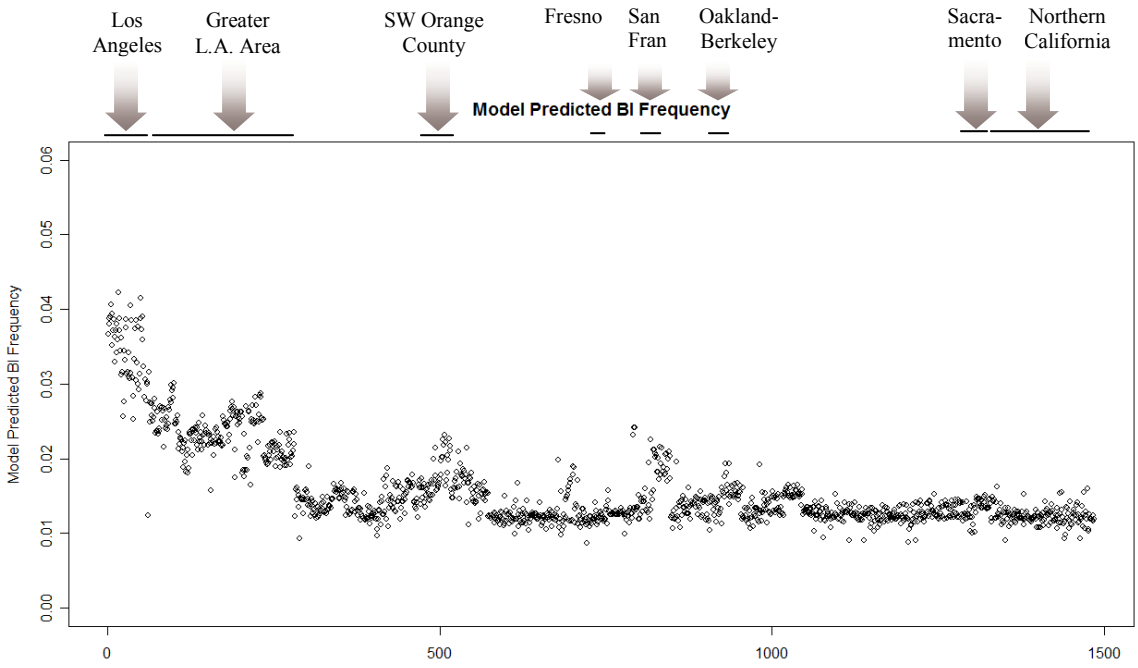
Appendix A

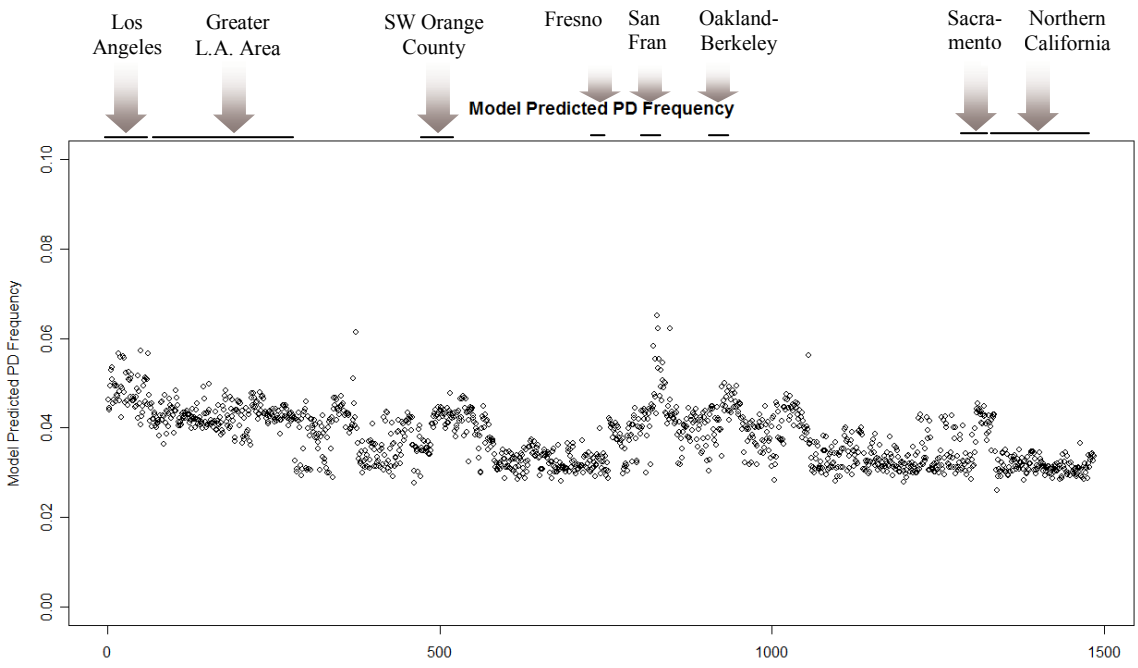

Los Angeles
Greater L.A. Area
SW Orange County
Fresno
San Fran
Oakland-Berkeley
Sacra-mento
Northern California
Observed BI Frequency
Observed BI Frequency
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0
500
1000
1500

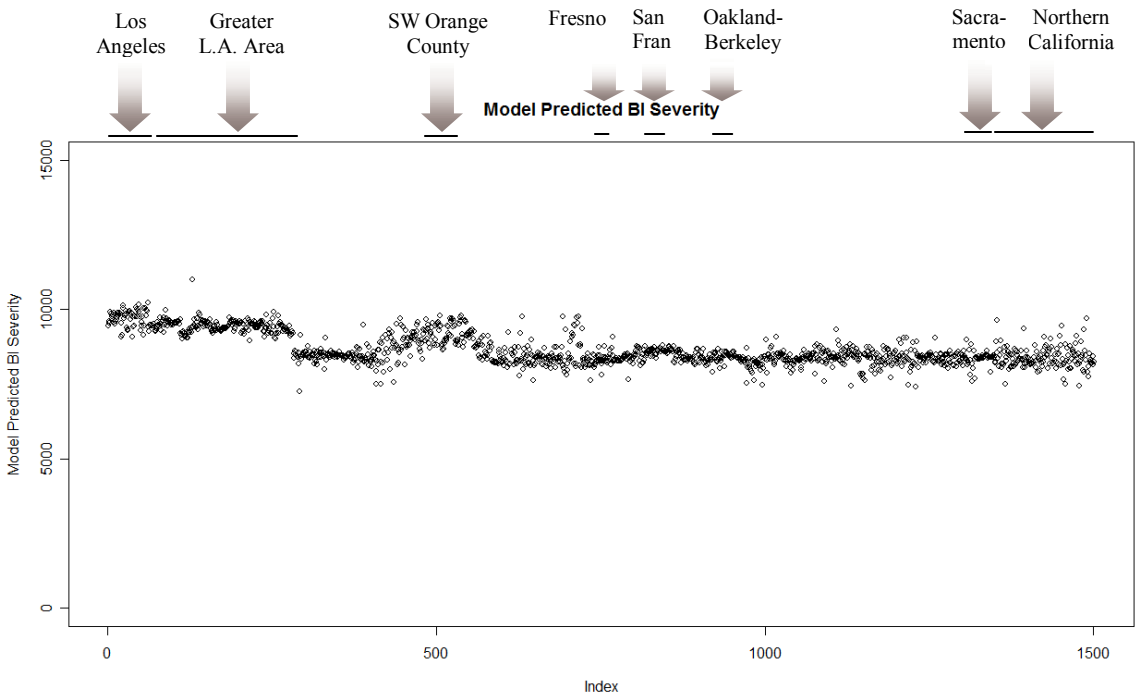

Los Angeles
Greater L.A. Area
SW Orange County
Fresno
San Fran
Oakland-Berkeley
Sacra-mento
Northern California
Observed PD Frequency
Observed PD Frequency
0.00
0.02
0.04
0.06
0.08
0.10
0
500
1000
1500

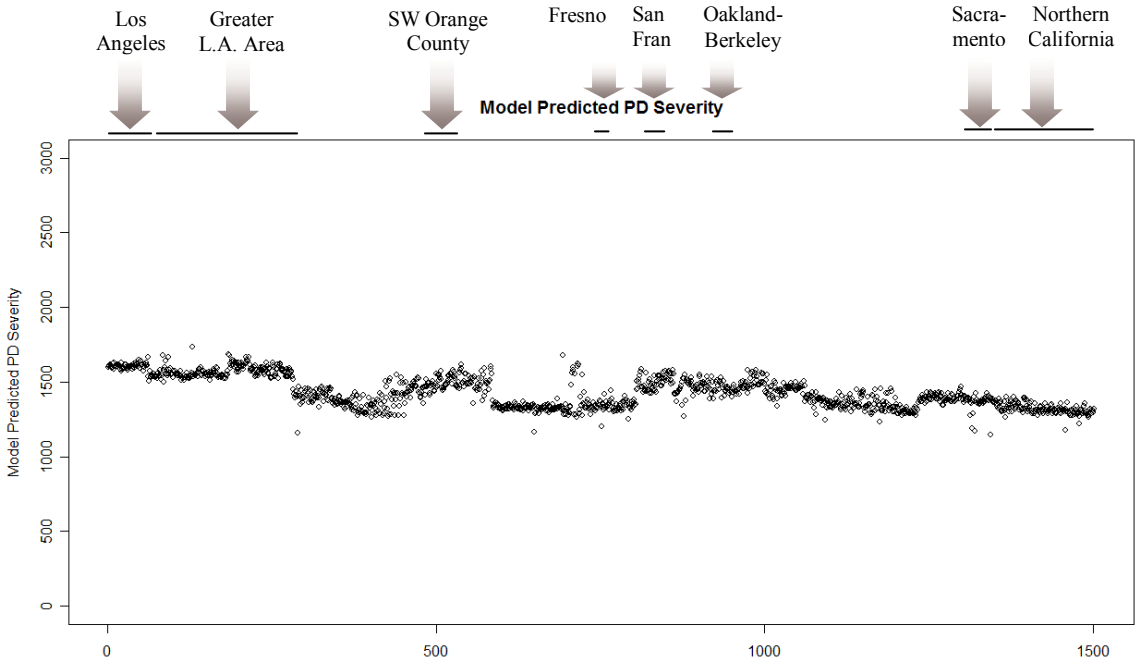

Los Angeles
Greater L.A. Area
SW Orange County
Fresno
San Fran
Oakland-Berkeley
Sacra-mento
Northern California
Observed BI Severity
Observed BI Severity
0
10000
20000
30000
40000
0
500
1000
1500

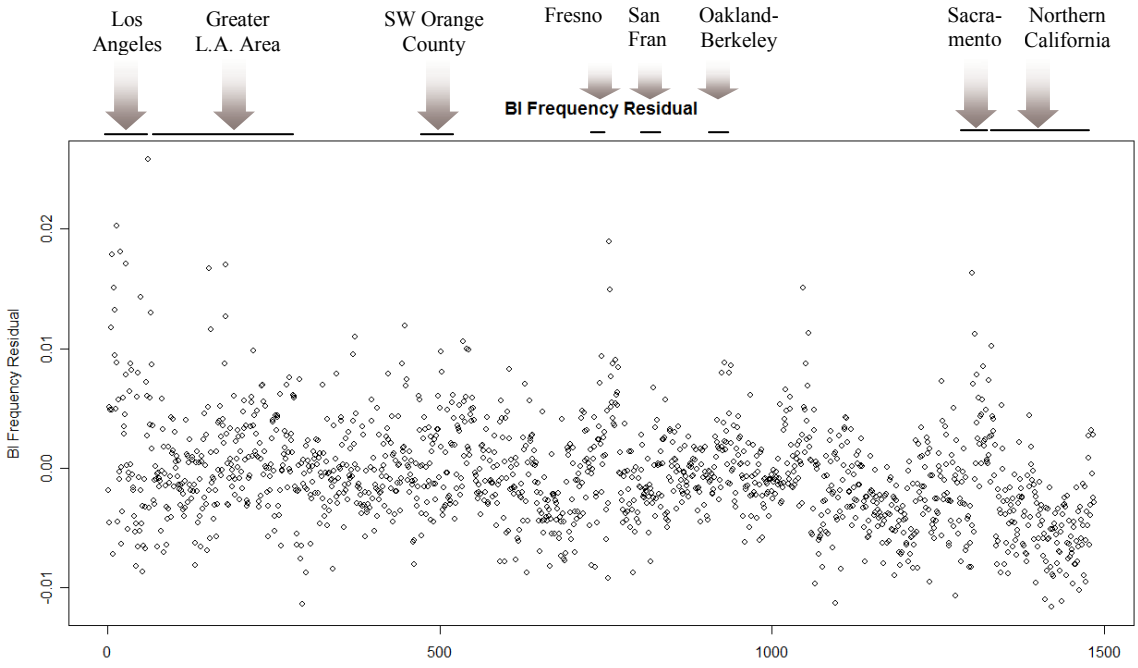

Los Angeles
Greater L.A. Area
SW Orange County
Fresno
San Fran
Oakland-Berkeley
Sacra-mento
Northern California
Observed PD Severity
Observed PD Severity
0
500
1000
1500
2000
2500
3000
0
500
1000
1500

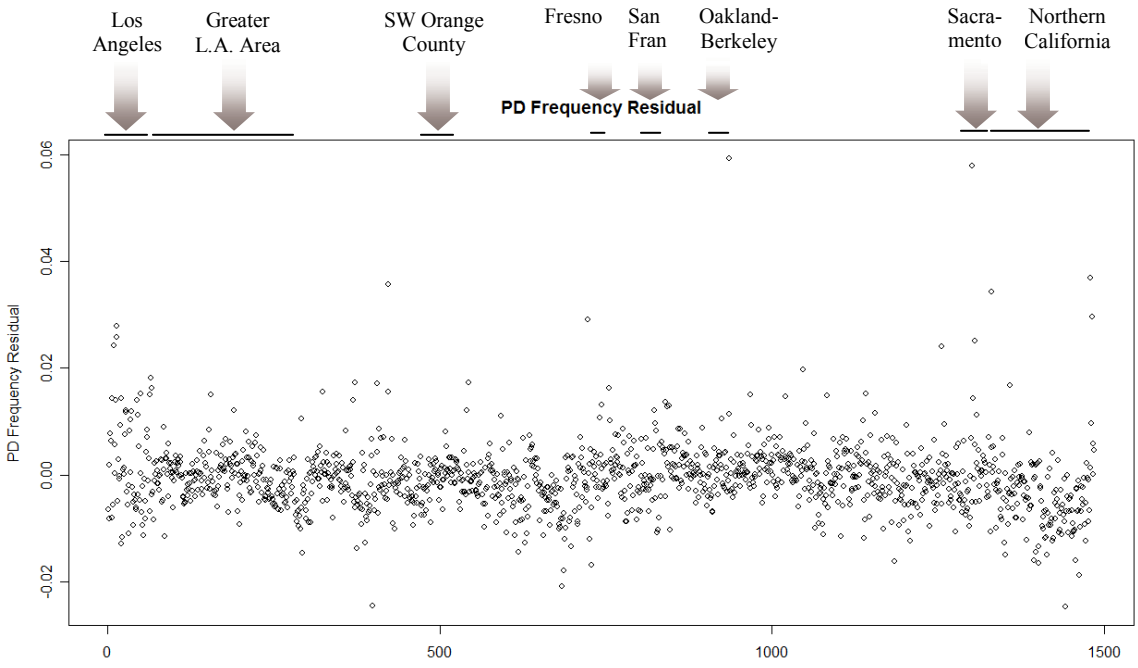

Los Angeles
Greater L.A. Area
SW Orange County
Fresno
San Fran
Oakland-Berkeley
Sacra-mento
Northern California
Model Predicted BI Frequency
Model Predicted BI Frequency
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0
500
1000
1500

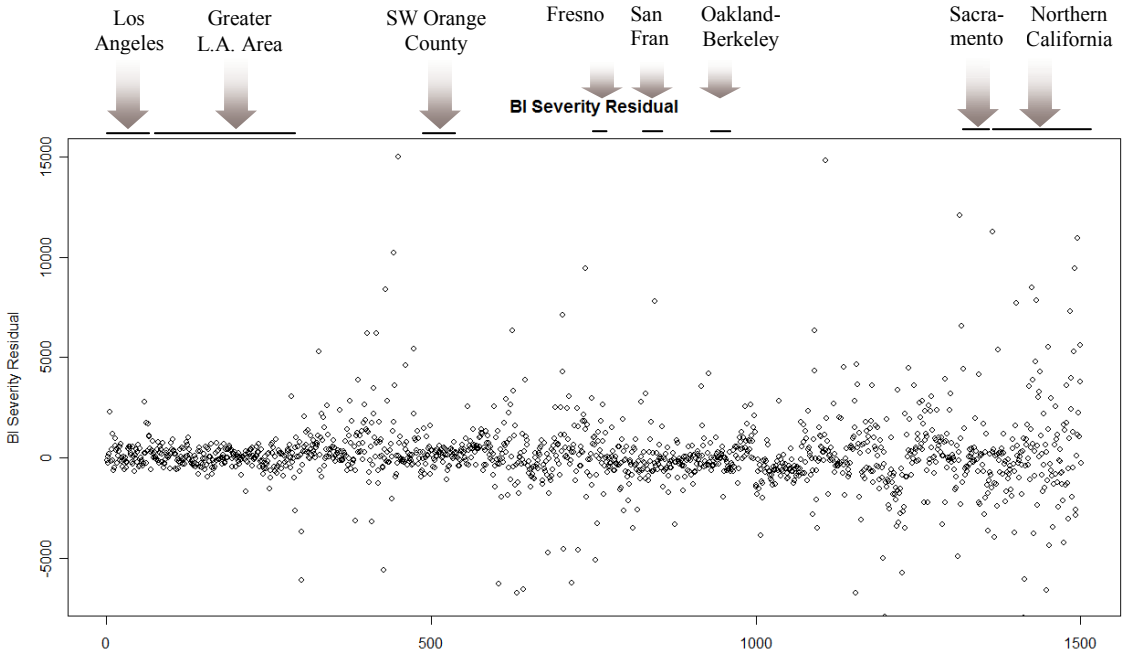

Los Angeles
Greater L.A. Area
SW Orange County
Fresno
San Fran
Oakland-Berkeley
Sacra-mento
Northern California
Model Predicted PD Frequency
Model Predicted PD Frequency
0.00
0.02
0.04
0.06
0.08
0.10
0
500
1000
1500

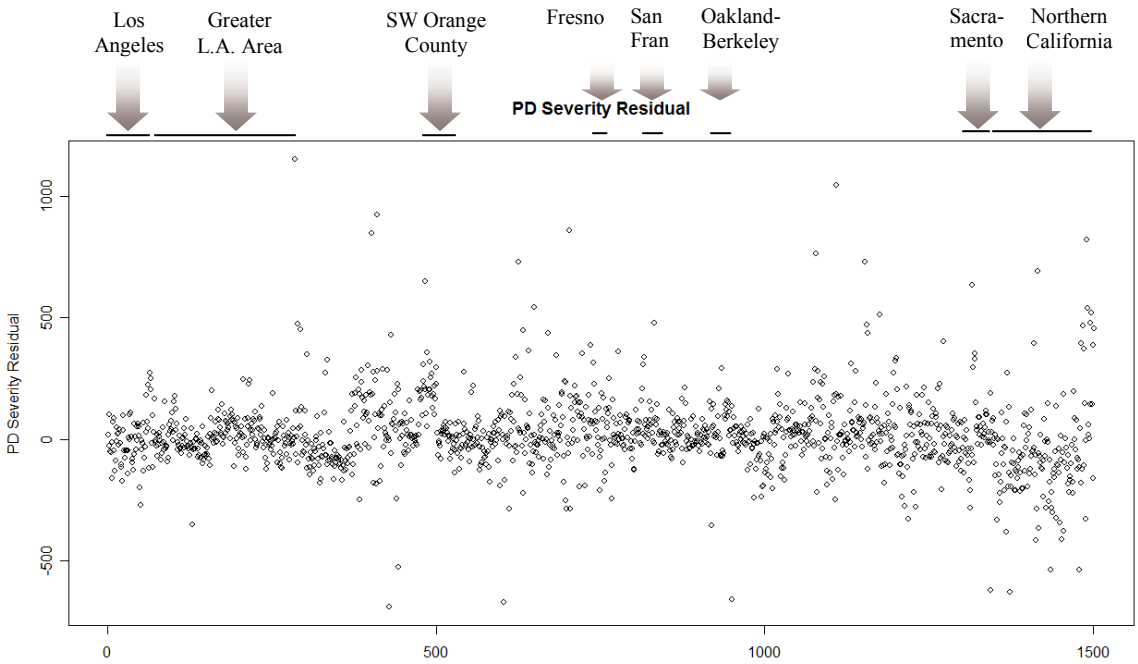

Los Angeles
Greater L.A. Area
SW Orange County
Fresno
San Fran
Oakland-Berkeley
Sacra-mento
Northern California
Model Predicted BI Severity
Model Predicted BI Severity
0
5000
10000
15000
0
500
1000
1500
Index

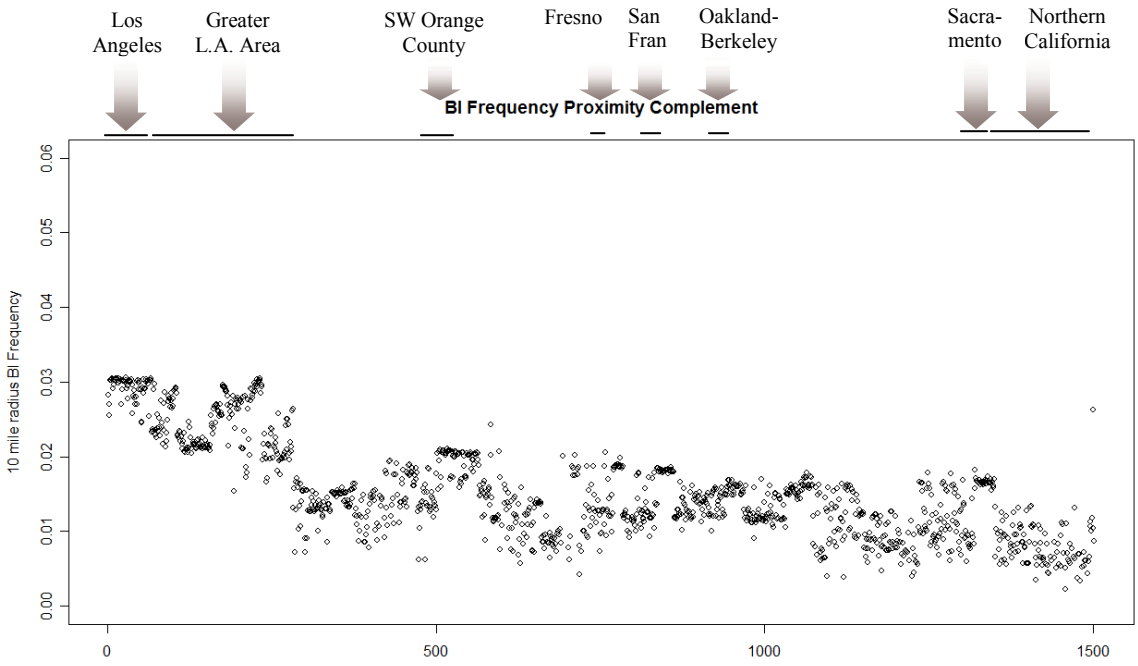

Los Angeles
Greater L.A. Area
SW Orange County
Fresno
San Fran
Oakland-Berkeley
Sacra-mento
Northern California
Model Predicted PD Severity
Model Predicted PD Severity
0
500
1000
1500
2000
2500
3000
0
500
1000
1500

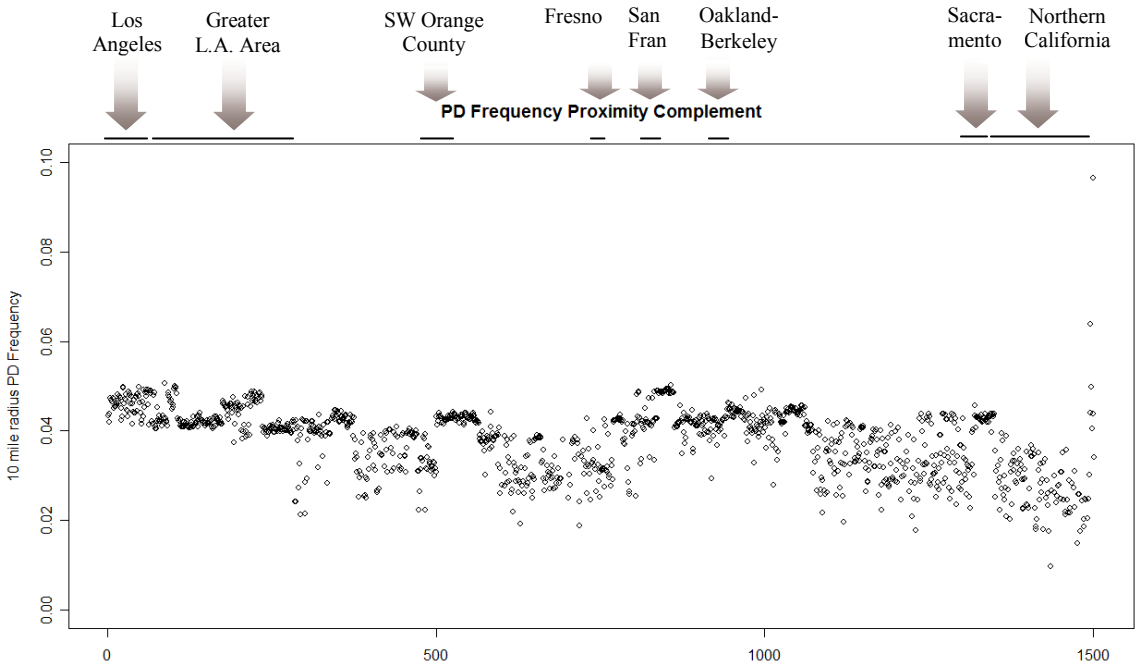

Los Angeles
Greater L.A. Area
SW Orange County
Fresno
San Fran
Oakland-Berkeley
Sacra-mento
Northern California
BI Frequency Residual
BI Frequency Residual
0.02
0.01
0.00
-0.01
0
500
1000
1500

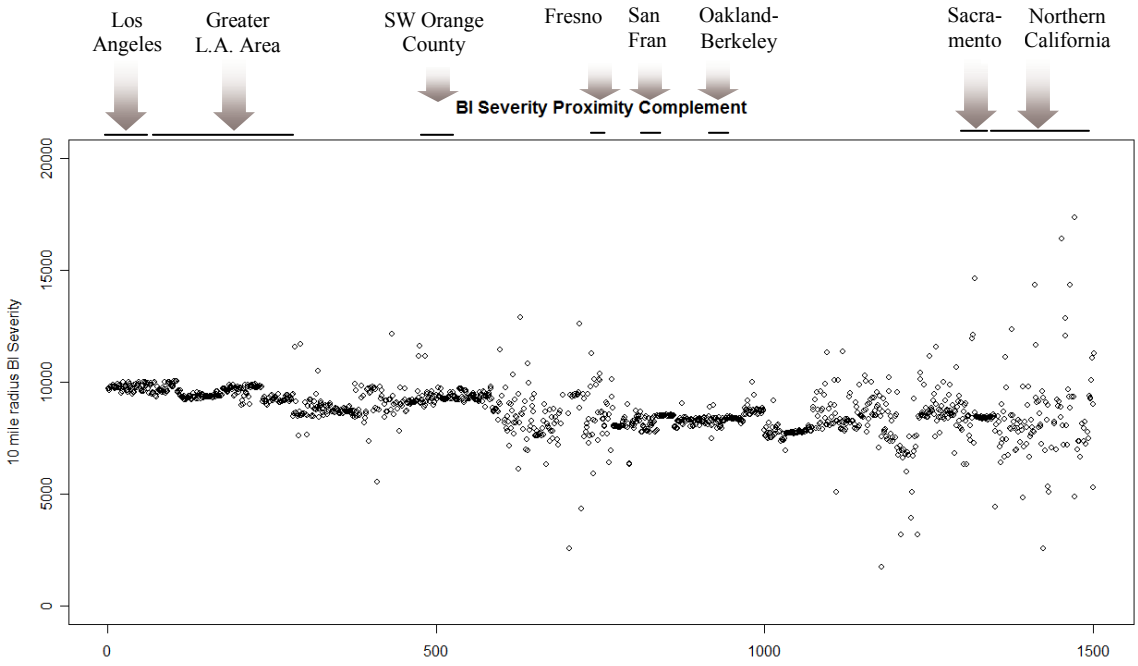

Los Angeles
Greater L.A. Area
SW Orange County
Fresno
San Fran
Oakland-Berkeley
Sacra-mento
Northern California
PD Frequency Residual
PD Frequency Residual
0.06
0.04
0.02
0.00
-0.02
0
500
1000
1500

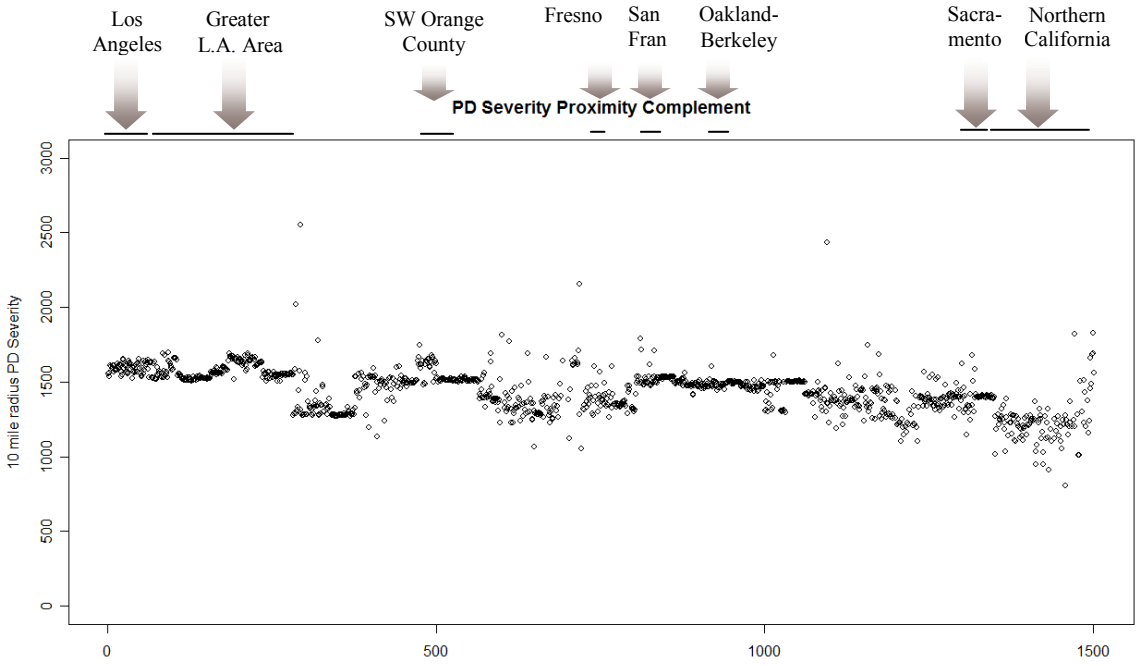

Los Angeles
Greater L.A. Area
SW Orange County
Fresno
San Fran
Oakland-Berkeley
Sacra-mento
Northern California
BI Severity Residual
BI Severity Residual
15000
10000
5000
0
-5000
0
500
1000
1500


Los Angeles
Greater L.A. Area
SW Orange County
Fresno
San Fran
Oakland-Berkeley
Sacra-mento
Northern California
PD Severity Residual
PD Severity Residual
1000
500
0
-500
0
500
1000
1500


Los Angeles
Greater L.A. Area
SW Orange County
Fresno
San Fran
Oakland-Berkeley
Sacra-mento
Northern California
BI Frequency Proximity Complement
10 mile radius BI Frequency
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0
500
1000
1500


Los Angeles
Greater L.A. Area
SW Orange County
Fresno
San Fran
Oakland-Berkeley
Sacra-mento
Northern California
PD Frequency Proximity Complement
10 mile radius PD Frequency
0.00
0.02
0.04
0.06
0.08
0.10
0
500
1000
1500


Los Angeles
Greater L.A. Area
SW Orange County
Fresno
San Fran
Oakland-Berkeley
Sacra-mento
Northern California
BI Severity Proximity Complement
10 mile radius BI Severity
0
5000
10000
15000
20000
0
500
1000
1500


Los Angeles
Greater L.A. Area
SW Orange County
Fresno
San Fran
Oakland-Berkeley
Sacra-mento
Northern California
PD Severity Proximity Complement
10 mile radius PD Severity
0
500
1000
1500
2000
2500
3000
0
500
1000
1500

Appendix B

Property Damage Liability Severity

```
Call:
lm(formula = PDSV ~ sqrt(LawDensePopInc25) + sqrt(CommuteMinutes *
    LawDensePopInc25) + sqrt(POPDENSE) + sqrt(PopDense10) +
    sqrt(PopDense25) + sqrt(PopDense50) + LosAngeles +
    LosAngelesArea + SanFrancisco, data = Data11,
    weights = PDExposure)

Residuals:
   Min     1Q Median     3Q    Max
-71323  -7633   1124  11427  67306

Coefficients:
                                Estimate   Std. Error  t value  Pr(>|t|)
(Intercept)                     1272.9902     8.2572   154.168  < 2e-16 ***
sqrt(LawDensePopInc25)         -2467.5218   455.3205    -5.419  6.97e-08 ***
sqrt(CommuteMinutes
* LawDensePopInc25)             640.4759    79.3663     8.070  1.43e-15 ***
sqrt(POPDENSE)                   -0.6594     0.1529    -4.313  1.71e-05 ***
sqrt(PopDense10)                  1.0725     0.2975     3.605  0.000322 ***
sqrt(PopDense25)                 -1.7157     0.3584    -4.786  1.87e-06 ***
sqrt(PopDense50)                  8.7453     0.4097    21.346  < 2e-16 ***
LosAngeles                      144.6067    13.1650    10.984  < 2e-16 ***
LosAngelesArea                   84.9154     6.0392    14.061  < 2e-16 ***
SanFrancisco                     61.4017    16.9720     3.618  0.000307 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 16860 on 1492 degrees of freedom
Multiple R-Squared: 0.6164,     Adjusted R-squared: 0.6141
F-statistic: 266.4 on 9 and 1492 DF,  p-value: < 2.2e-16
```

Bodily Injury Liability Frequency

```
Call:
lm(formula = BIFQ ~ CommuteMinutes + CommTimeSpaceDensity10 +
   CommTimeSpaceDensity25 + CommTimeSpaceDensity50 +
   CommuteMinutes * CommTimeSpaceDensity25 + +LawDensePopInc25 +
   LawDensePop50 + LosAngelesArea + LosAngeles +
   SanFrancisco + CommuteMinutes * LosAngeles +
   LosAngelesArea * LawDensePopInc25 + LosAngelesArea *
   LawDensePop50 + CommuteMinutes * LawDensePopInc25,
   data = Data11, weights = BIExposure)
Residuals:
   Min     1Q  Median     3Q    Max
-3.3214 -0.4119 -0.1400  0.2401  3.6836
Coefficients:
                                Estimate        Std. Error      t value  Pr(>|t|)
(Intercept)                      9.803e-03      7.985e-04       12.277   < 2e-16 ***
CommuteMinutes                   1.178e-04      3.047e-05       3.867    0.000115 ***
CommTimeSpaceDensity10           2.557e-07      1.402e-08       18.236   < 2e-16 ***
CommTimeSpaceDensity25           -4.986e-07     8.352e-08       -5.970   2.97e-09 ***
CommTimeSpaceDensity50           4.256e-07      4.235e-08       10.049   < 2e-16 ***
LawDensePopInc25                 3.717e-01      2.131e-01       1.745    0.081265 .
LawDensePop50                    -1.977e-01     4.176e-02       -4.734   2.41e-06 ***
LosAngelesArea                   5.829e-03      1.620e-03       3.597    0.000333 ***
LosAngeles                       -1.188e-02     4.391e-03       -2.705   0.006908 **
SanFrancisco                     -3.772e-03     8.008e-04       -4.711   2.70e-06 ***
CommuteMinutes:
CommTimeSpaceDensity25           1.072e-08      2.875e-09       3.728    0.000200 ***
CommuteMinutes:
LosAngeles                       8.224e-04      1.570e-04       5.240    1.84e-07 ***
LawDensePopInc25:
LosAngelesArea                   5.635e-01      1.646e-01       3.424    0.000634 ***
LawDensePop50:
LosAngelesArea                   -9.075e-01     2.159e-01       -4.203   2.80e-05 ***
CommuteMinutes:
LawDensePopInc25                 -1.772e-02     7.293e-03       -2.430   0.015204 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7196 on 1470 degrees of freedom
  (17 observations deleted due to missingness)
Multiple R-Squared: 0.719,      Adjusted R-squared: 0.7163
F-statistic: 268.6 on 14 and 1470 DF,  p-value: < 2.2e-16
```

Property Damage Liability Frequency

```
Call:
lm(formula = PDFQ ~ sqrt(CommuteMinutes) + sqrt(CommTimeSpaceDensity10) +
    sqrt(CommTimeSpaceDensity25) + LosAngeles +
    sqrt(CommuteMinutes * CommTimeSpaceDensity10) +
    sqrt(CommuteMinutes * CommTimeSpaceDensity25) +
    POPDENSEPOP + PopDensePop10 + sqrt(PopDensePop25) +
    sqrt(CommuteMinutes * PopDensePop10) + sqrt(CommuteMinutes *
    PopDensePop25) + sqrt(CommuteMinutes * LosAngeles),
    data = Data11, weights = PDExposure)

Residuals:
    Min      1Q  Median      3Q     Max
-2.9499 -0.4954 -0.1054  0.3493  3.6748

Coefficients:
                                  Estimate   Std. Error  t value  Pr(>|t|)
(Intercept)                     3.345e-02  1.869e-03   17.896  < 2e-16 ***
sqrt(CommuteMinutes)           -6.481e-04  3.662e-04   -1.769  0.077018 .
sqrt(CommTimeSpaceDensity10)    2.235e-04  3.764e-05    5.938  3.59e-09 ***
sqrt(CommTimeSpaceDensity25)   -6.401e-04  2.505e-04   -2.555  0.010731 *
LosAngeles                      4.305e-02  9.978e-03    4.314  1.71e-05 ***
sqrt(CommuteMinutes *
CommTimeSpaceDensity10)        -2.238e-05  7.159e-06   -3.126  0.001807 **
sqrt(CommuteMinutes *
CommTimeSpaceDensity25)         1.417e-04  4.752e-05    2.981  0.002918 **
POPDENSEPOP                     1.048e-06  7.161e-08   14.632  < 2e-16 ***
PopDensePop10                  -6.204e-06  3.021e-07  -20.539  < 2e-16 ***
sqrt(PopDensePop25)             1.659e-03  8.264e-04    2.007  0.044943 *
sqrt(CommuteMinutes *
PopDensePop10)                  4.890e-05  7.432e-06    6.580  6.53e-11 ***
sqrt(CommuteMinutes *
PopDensePop25)                 -3.966e-04  1.567e-04   -2.531  0.011466 *
sqrt(CommuteMinutes *
LosAngeles)                    -7.302e-03  1.892e-03   -3.859  0.000119 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8041 on 1472 degrees of freedom
  (17 observations deleted due to missingness)
Multiple R-Squared: 0.6166,     Adjusted R-squared: 0.6134
F-statistic: 197.2 on 12 and 1472 DF,  p-value: < 2.2e-16
```

Bodily Injury Liability Severity

```
Call:
lm(formula = BISV ~ LawDensePopInc25 + LawDensePop50 +
    LawDensePop50 * LosAngeles + CommuteMinutes +
    CommuteMinutes * LawDensePopInc25 + CommuteMinutes *
    LawDensePop50 + CommTimeSpaceDensity10 + CommTimeSpaceDensity50 +
    LosAngelesArea, data = Data11, weights = BIExposure)

Residuals:
   Min     1Q  Median     3Q    Max
-577104  -76520    7645   91232  852072

Coefficients:
                                 Estimate   Std. Error   t value  Pr(>|t|)
(Intercept)                    7.333e+03  1.898e+02    38.639  < 2e-16 ***
LawDensePopInc25               6.512e+04  3.685e+04    1.767    0.07736 .
LawDensePop50                  7.103e+04  3.622e+04    1.961    0.05009 .
LosAngeles                     2.560e+03  9.074e+02    2.822    0.00484 **
CommuteMinutes                 5.159e+01  7.561e+00    6.824    1.28e-11 ***
CommTimeSpaceDensity10         4.546e-03  1.691e-03    2.689    0.00724 **
CommTimeSpaceDensity50         1.039e-01  7.567e-03    13.737   < 2e-16 ***
LosAngelesArea                 3.892e+02  4.994e+01    7.794    1.21e-14 ***
LawDensePop50:LosAngeles       -6.066e+05  3.051e+05   -1.988   0.04700 *
LawDensePopInc25:CommuteMinutes -3.607e+03  1.236e+03   -2.918   0.00358 **
LawDensePop50:CommuteMinutes   -6.545e+03  1.365e+03   -4.795   1.79e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 141000 on 1490 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-Squared: 0.4232,     Adjusted R-squared: 0.4194
F-statistic: 109.3 on 10 and 1490 DF,  p-value: < 2.2e-16
```

Appendix C

Bodily Injury Liability Frequency MAD Comparison by CAARP Territory

CAARP	Zip Codes not fully credible	Zip Codes	CAARP MAD	10Mile MAD	Frequency
1	4	5	0.0004	0.0006	0.0109
2	84	84	0.0016	0.0009	0.0072
3	7	9	0.0008	0.0006	0.0131
4	46	47	0.0015	0.0010	0.0096
5	21	31	0.0021	0.0020	0.0169
6	31	32	0.0012	0.0017	0.0086
7	21	39	0.0015	0.0012	0.0136
8	17	19	0.0008	0.0019	0.0121
9	7	11	0.0006	0.0013	0.0114
10	7	13	0.0022	0.0032	0.0206
11	8	16	0.0020	0.0021	0.0188
12	0	3	0.0003	0.0031	0.0144
13	14	29	0.0011	0.0012	0.0119
14	1	3	0.0004	0.0036	0.0187
15	10	12	0.0030	0.0043	0.0200
16	11	29	0.0010	0.0009	0.0144
17	5	16	0.0027	0.0024	0.0170
18	56	64	0.0065	0.0030	0.0123
19	5	6	0.0021	0.0025	0.0189
20	7	11	0.0016	0.0011	0.0149
21	2	3	0.0011	0.0013	0.0166
22	9	11	0.0048	0.0034	0.0192
23	9	10	0.0091	0.0013	0.0100
24	10	11	0.0009	0.0012	0.0124
25	28	31	0.0008	0.0011	0.0090
26	21	28	0.0011	0.0014	0.0131
27	5	7	0.0007	0.0017	0.0154
28	5	18	0.0053	0.0016	0.0234
29	19	31	0.0051	0.0020	0.0173
30	5	23	0.0039	0.0028	0.0282
31	1	12	0.0040	0.0031	0.0277
32	8	35	0.0036	0.0033	0.0259
33	3	8	0.0027	0.0028	0.0237
34	3	22	0.0021	0.0020	0.0206
35	5	12	0.0023	0.0027	0.0266

CAARP	Zip Codes not fully credible	Zip Codes	CAARP MAD	10Mile MAD	Frequency
36	6	10	0.0058	0.0108	0.0409
37	7	10	0.0055	0.0046	0.0323
38	6	11	0.0025	0.0040	0.0325
39	21	21	0.0053	0.0072	0.0335
40	6	9	0.0029	0.0034	0.0278
41	1	6	0.0031	0.0025	0.0236
42	1	5	0.0023	0.0037	0.0224
43	1	6	0.0008	0.0024	0.0195
44	2	11	0.0011	0.0015	0.0216
45	1	11	0.0024	0.0032	0.0249
46	7	37	0.0022	0.0024	0.0199
47	4	19	0.0015	0.0019	0.0207
48	5	16	0.0010	0.0022	0.0184
49	0	5	0.0021	0.0046	0.0250
52	8	18	0.0014	0.0011	0.0171
54	56	65	0.0019	0.0018	0.0143
57	5	8	0.0012	0.0015	0.0130
59	9	17	0.0015	0.0019	0.0203
64	75	92	0.0026	0.0015	0.0130
65	10	13	0.0013	0.0015	0.0131
66	35	44	0.0025	0.0013	0.0159
67	36	37	0.0020	0.0007	0.0117
68	28	32	0.0048	0.0015	0.0143
71	8	10	0.0032	0.0055	0.0192
74	20	33	0.0014	0.0012	0.0132
75	19	30	0.0018	0.0013	0.0135
76	10	15	0.0011	0.0019	0.0118
77	43	45	0.0015	0.0010	0.0097
80	35	37	0.0020	0.0016	0.0122
89	3	8	0.0011	0.0014	0.0152
93	6	12	0.0006	0.0008	0.0137
94	3	14	0.0012	0.0010	0.0167
95	2	6	0.0027	0.0060	0.0215
96	6	16	0.0010	0.0013	0.0134
97	1	6	0.0007	0.0007	0.0127
98	7	15	0.0014	0.0015	0.0145
99	12	14	0.0016	0.0026	0.0150

5. REFERENCES

5.1 Historical

[1] Barber, Harmon T., "A Suggested Method for Developing Automobile Rates," *PCAS*, 1929, Vol. XV, No. 32, 191-222.

[2] Constable, William J., "Compulsory Automobile Insurance," *PCAS*, 1927, Vol. XIII, No. 28, 188-216.

[3] Constable, William J., "Massachusetts Compulsory Automobile Liability Insurance," *PCAS*, 1929, Vol. XV, No. 32, 171-190.

[4] Dorweiler, Paul, "Notes on Exposure and Premiums Bases," *PCAS*, 1930, Vol. XVI, No. 34, 319-343.

[5] Kirkpatrick, A. L., "The Development of Public Liability Insurance Rates For Automobiles," *PCAS*, 1921, Vol. VIII, No. 17, 35-53.

[6] McDonald, M. G., "Compulsory Automobile Insurance Rate Making in Massachusetts," *PCAS*, 1955, Vol. XLII, No. 77, 19-69.

[7] Michelbacher, G. F, "Casualty Insurance for Automobile Owners," *PCAS*, 1918, Vol. V, No.12, 213-242.

[8] Riegel, Robert, "Automobile Insurance Rates," *Journal of Political Economy*, February 17, 1920, 561-579.

[9] Stern, Phillip K., "Current Rate Making Procedures for Automobile Liability Insurance," *PCAS*, 1956, Vol. XLIII, No. 80, 112-165.

[10] Zoffer, H. Jerome, *The History of Automobile Liability Insurance Rating*, 1959.

5.2 Territory Analysis

[11] Boskov, M, R. J. Verrall, "Premium Rating by Geographic Area Using Spatial Models," *ASTIN* Bulletin, 1994, Vol. 24, No. 1, 131-143.

[12] Brubaker, Randall E, "Geographic Rating of Individual Risk Transfer Costs Without Territorial Boundaries," CAS Winter *Forum*, 1996, 97-127.

[13] CAS Committee on Management Data and Information, "1996 Geo-Coding Survey," CAS Winter *Forum*, 1997, 169-186.

[14] Christopherson, Steven, Debra L. Werland, "Using a Geographic Information System to Identify Territory Boundaries," CAS *Forum*, Winter, 1996, 191-211.

[15] Conners, John B, Sholom Feldblum, "Personal Automobile: Cost Drivers, Pricing, and Public Policy," CAS Winter *Forum*, 1997, 317-341.

[16] Feldblum, Sholom, "Workers' Compensation Ratemaking," *CAS Exam Study Note*, 1993.

[17] Guven, Serhat, "Multivariate Spatial Analysis of the Territory Rating Variable," CAS *Discussion Paper Program*, 2004, 245-260.

[18] Hunstad, Lyn, "Methodology and Data Used to Develop the California Private Passenger Auto Frequency and Severity Bands Manual," California Department of Insurance, April 1996.

[19] Rate Regulation Division, California Department of Insurance, "Study of California Driving Performance by Zip Code (Phase I)," November 1978.

[20] Rate Regulation Division, California Department of Insurance, "Study of California Driving Performance (Phase II)," November 1979.

[21] Tang, Max, C., "Auto Insurance in California: Differentials in Industrywide Pure Premiums and Company Territory Relativities between Adjacent Zip Codes," *Policy Research Division, California Department of Insurance*, 2005.

[22] Taylor, Greg C, "Geographic Premium Rating by Whittaker Spatial Smoothing," *ASTIN* Bulletin, 2001, Vol. 31, No. 1, 147-160.

[23] Taylor, Greg C, "Use of Spline Functions for Premium Rating by Geographic Area," *ASTIN* Bulletin, 1994, Vol. 19, No. 1, 91-122.

[24] Wang, H. H, Hao Zhang, "On the Possibility of a Private Crop Insurance Market: A Spatial Statistics Approach," *The Journal of Risk and Insurance*, 2003, Vol. 70, No. 1, 111-124.

5.3 Risk Classification

[25] Bishop, Yvonne M, Stephen E. Fienberg, and Paul W. Holland, *Discrete Multivariate Analysis: Theory and Practice*,

(Boston: The MIT Press, 1975).
[26] Casey, Barbara, Jacques Pezier and Carl Spetzler, "The Role of Risk Classifications in Property and Casualty Insurance: A Study of the Risk Assessment Process," *Stanford Research Institute*, May 1976, SRI Project 4253-4.
[27] Chang, Lena, William B. Fairley, "An Estimation Model for Multivariate Insurance Rate Classification," *Automobile Insurance Classification: Equity & Accuracy*, Massachusetts Division of Insurance, 1978, 25-55.
[28] Ferreira, Joseph Jr., "Merit Rating and Automobile Insurance," *Automobile Insurance Risk Classification: Equity & Accuracy*, Massachusetts Division of Insurance, 1978a, 56-73.
[29] Ferreira, Joseph Jr., "Identifying Equitable Insurance Premiums for Risk Classes: An Alternative to the Classical Approach," *Automobile Insurance Risk Classification: Equity & Accuracy*, Massachusetts Division of Insurance, 1978b, 74-120.
[30] Finger, Robert J., "Risk Classification," Chapter 6 in *Foundations of Casualty Actuarial Sciences*, 4th edition, (Arlington, Va.: Casualty Actuarial Society, 2001) 287-342.
[31] Hunstad, Lyn, "Sequential Analysis Guidelines," *California Department of Insurance*, September 1996.
[32] Hunstad, Lyn, Robert Bernstein, and Jerry Turem, "Impact Analysis of Weighting Auto Rating Factors to Comply with Proposition 103," *Office of Policy Research, California Department of Insurance*, December 1994.
[33] Mildenhall, Stephen J., "A Systematic Relationship Between Minimum Bias and Generalized Linear Models," *PCAS*, 1999, Vol. LXXXVI, 393-487.
[34] Shayer, Natalie, "Driver Classification in Automobile Insurance," *Automobile Insurance Risk Classification: Equity & Accuracy*, Massachusetts Division of Insurance, 1978, 1-24.
[35] Stone, James M., "Excerpt from the Opinion, Findings and Decision on 1978 Automobile Insurance Rates," *Automobile Insurance Risk Classification: Equity & Accuracy*, Massachusetts Division of Insurance, 1978, 144-205.
[36] Venter, Gary G., "Discussion: Minimum Bias with Generalized Linear Models," *PCAS*, 1990, Vol. LXXVII, 337-349.

5.4 Clustering, Classification, Spatial Statistics, and Geography

[37] Bailey, Trevor C, Anthony C. Gatrell, *Interactive Spatial Data Analysis*, Longman Scientific & Technical, 1995.
[38] Berkhin, P., "A Survey of Clustering Data Mining Techniques," *Grouping Multidimensional Data: Recent Advances in Clustering*, Springer, Edited by Kogan, Jacob, Charles Nicholas and Marc Teboulle, 2006, 26-71.
[39] Chawla, Sanjay, Shashi Shekhar, Weili Wu, and Uygar Ozesmi, "Modeling spatial dependencies for mining geospatial data: an introduction," *Geographic Data Mining and Knowledge Discovery*, 2001, 131-159.
[40] Crawley, Michael J., *The R Book*, (New York: Wiley, 2007).
[41] Cressie, Noel A. C., *Statistics for Spatial Data*, Wiley, 1993.
[42] Ester, Martin, Hans-Peter Kriegel, and Jorg Sander, "Algorithms and applications for spatial data mining," *Geographic Data Mining and Knowledge Discovery*, 2001, 160-187.
[43] Everitt, Brian, Sabine Landau, and Morven Leese, *Cluster Analysis*, Oxford University Press, 2001, 4th edition.
[44] Ferligoj, A. and V. Batagelj, "Some types of clustering with relational constraints," *Psychometrika*, 1982, Vol. 47, 541-552.
[45] Han, Jiawei, Micheline Kamber and Anthony K. H. Tung, "Spatial Clustering methods in data mining: A survey," *Geographic Data Mining and Knowledge Discovery*, 2001, 188-217.
[46] Kaufman, Leonard, Peter J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*, Wiley, 1990.
[47] Maravalle, M., B. Simeone, and R. Naldini, "Clustering on trees," *Computational Statistics and Data Analysis*, 1997, Vol. 24, 217-234.
[48] Miller, Harvey J., Jiawei Han, "Geographical data mining and knowledge discovery: an overview," *Geographic Data Mining and Knowledge Discovery*, 2001, 3-32.
[49] Murtagh, F.D., "Contiguity-constrained hierarchical clustering, *Partitioning Data Sets*, Edited by Cox, I., P. Hansen and B. Julesz, *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, American Mathematical Society, 1995, Vol. 19, 143-152.
[50] Sanche, Robert, Kevin Lonergan, "Variable Reduction for Predictive Modeling with Clustering," CAS *Forum*, Winter, 2006, 89-100.
[51] Teboulle, M., P. Berkhin, I. Dhillon, Y. Guan, and J. Kogan, "Clustering with Entropy-Like k-Means Algorithms," *Grouping Multidimensional Data: Recent Advances in Clustering*, Edited by Kogan, Jacob, Charles Nicholas and Marc Teboulle, Springer, 2006, 127-160.
[52] Tung, A.K.H., J. Han., R. Nu and L. Lankershmanan, "Constrainted clustering on large database," *Proceedings of*

the 2001 International Conference on Database Theory (ICDT '01), January 2001.

[53] Wojdyla, D., L. Poletto, C. Cuesta, C. Badler and M.E. Passamonti, "Cluster analysis with constraints: Its use with breast cancer mortality rates in Argentina," *Statistics in Medicine*, 1996, Vol. 15, 741-746.

5.5 Mathematical Programming

[54] Beale, E. M. L., J. A. Tomlin, "Special Facilities in general mathematical programming system for non-convex problems using ordered sets of variables," in J. Lawrence (Ed.), *Proceedings of the 5th International Conference on Operations Research*, Tavestock. London, 1969.

[55] Bertsekas, D.P., *Nonlinear Programming*, Athena Scientific, 1999.

[56] Byrd, Richard H., Jean Charles Gilbert, and Jorge Nocedal, "A Trust Region method based on interior point techniques for nonlinear programming," *Mathematical Programming*, Vol. 89, No. 1, 2000, 149-185.

[57] Byrd, Richard H., Nicholas I.M. Gould, Jorge Nocedal, and Richard A. Waltz, "An algorithm for nonlinear optimization using linear programming and equality constrained subproblems," *Mathematical Programming, Series B*, Vol. 100, No. 1, 2004, 27-48.

[58] Byrd, Richard H., Jorge Nocedal, and Richard A. Waltz, "Feasible interior methods using slacks for nonlinear optimization," *Computational Optimization and Applications*, Vol. 26, No. 1, 2003, 35-61.

[59] Frontline Systems, Inc., *Premium Solver Platform, Solver Platform SDK, Field-Installable Solver Engines User Guide*.

[60] Hillier, Frederick S., Gerald J. Lieberman, *Introduction to Operations Research*, McGraw-Hill, 1995, 6th edition.

[61] Li, Duan L., Xiaoling Sun, *Nonlinear Integer Programming*, Springer, 2006.

[62] Williams, Paul W., *Model Building in Mathematical Programming*, John Wiley and Sons, 1999, 4th edition.

Abbreviations and notations

BI, bodily injury
PD, property damage
LCG, Loss Cost Gradient
LGP, Loss Generating Process
FB, Frequency Band
SB, Severity Band

Biography(ies) of the Author(s)

Mr. Weibel is the President of Alta Financial & Insurance Services, LLC, and Alta Program Management. He is in the process of starting wholesaling and general insurance agency operations. Immediately prior, he was a founding member and Vice President of Cabrillo General Insurance Agency. While there he developed and managed several innovative property and automobile insurance products, which generated substantial profits for the insurance carriers. Prior to this he served in actuarial positions at Tower Hill Insurance Group, Arrowhead General Insurance Agency, and the ICW Group. As a college student he did statistical work for Cadence Design Systems, Inc. He has a degree in Statistical Sciences from the University of California at Santa Barbara. He is licensed (0D14281) to transact Fire and Casualty, Life, Accident and Health, and Surplus Lines in the state of California. He heads Adult Stem Cell Therapies and Research, which disseminates information and conducts advocacy on behalf of Adult Stem Cell research, and has over one thousand members. Eric is currently conducting several research projects involving personal automobile and professional liability insurance.

Eric can be reached at EJWeibel@msn.com

Mr. Walsh is an Actuarial Analyst for the Enterprise Risk Management department of the ICW Group in San Diego, California. He holds a bachelor's degree in Mathematics/Economics from the University of California at Santa Barbara and studied econometric modeling at the London School of Economics. Paul is involved in catastrophe modeling, ceded and assumed reinsurance, enterprise risk management and commercial property underwriting at the ICW Group. Prior to the ICW Group, Paul worked as an Actuarial Analyst for Cabrillo General Insurance Agency.

Paul can be reached at jpaulwalsh@yahoo.com

The authors' opinions expressed herein do not necessarily reflect the views of the authors' employers or clients.

Clustering in Ratemaking: Applications in Territories Clustering

Ji Yao, Ph.D.

Abstract: Clustering methods are briefly reviewed and their applications in insurance ratemaking are discussed in this paper. First, the reason for clustering and the consideration in choosing clustering methods in insurance ratemaking are discussed. Then clustering methods are reviewed and particularly the problem of applying these methods directly in insurance ratemaking is discussed. An exposure-adjusted hybrid (EAH) clustering method is proposed, which may alleviate some of these problems. Results from EAH approach are presented step by step using the U.K. motor data. The limitations and other considerations of clustering are followed in the end.

Keywords: Clustering, ratemaking, generalized linear modeling, territory analysis, data mining.

1. INTRODUCTION

Clustering is the unsupervised classification of patterns into groups [1]. It is widely studied and applied in many area including computer science, biology, social science, and statistics. A significant number of clustering methods were proposed in literature [1]-[6]. In the context of actuarial study, [7]-[9] studied possible application of clustering in insurance. As to the territory ratemaking, [10] considered the use of geographical information system. However, a thorough analysis of clustering in insurance ratemaking is not known to this author.

The purpose of this paper is two-fold. The first part of the paper introduces the basic idea of clustering and state-of-the-art clustering methods. Due to the large amount of methods, however, it is not intended to give a detailed review of every clustering method in the literature. Rather, the focus is on the key idea of each method and, more importantly, their advantages and disadvantages when applied in insurance ratemaking.

In the second part, a clustering method called exposure-adjusted hybrid (EAH) clustering is proposed. The purpose of this section is not to advocate one certain clustering method, but to illustrate the general approach that could be taken in territory clustering. Because clustering is subjective, it is well recognized that most details should be modified to accommodate the feature of data-set and the purpose the clustering.

The remainder of this paper proceeds as follows: Section 2 introduces clustering and its application in insurance ratemaking; section 3 reviews clustering methods and their applicability in insurance ratemaking; section 4 proposes the EAH clustering method and illustrates this method step-by-step

using U.K. motor data; section 5 discusses some other considerations; and section 6 draws conclusions. Some useful references are listed in Section 7.

2. OVERVIEW OF CLUSTERING

2.1 Introduction to Clustering

The definition of clustering is not unique. Generally, *clustering* is the process of grouping a set of data objects into a cluster or clusters so that the data objects within the cluster are very similar to one another, but are dissimilar to objects in other clusters [3]. Usually a *similarity measure* is defined and the clustering procedure is to optimize this measure locally or globally.

It is important to understand the difference between clustering and discriminant analysis. In discriminant analysis, we have a set of pre-classified samples, which could be used to train the algorithm to learn the description of each class. For example, we have a set of claims, some of which are fraud claims. These fraud cases are used to train the algorithm to find a rule that predicts the probability of fraud claims in future cases. However, in the case of clustering, these pre-classified samples are not available. So all these rules have to be derived solely from data, indicating that clustering is subjective in nature.

With so many clustering methods available in literature [1]-[6], it is a very difficult task to choose the appropriate method. Two considerations are the purpose of clustering and the feature of dataset.

2.2 Purpose of Clustering in Insurance

There are many reasons to use clustering in insurance ratemaking. The first reason is to better understand the data. After grouping data object into clusters, the feature of each cluster is clearer and more meaningful. For example, it is useful to cluster similar occupations and analyze their claim experience together.

The second reason is to reduce the volatility of data and to make the rates stable over time. Because the amounts of data are usually limited over a certain period, historical data in each segment may show high volatility. In ratemaking, if analysis is only based on experience of each single segment, the resulting rates will be volatile as well. Appropriate clustering alleviates this problem.

The third reason is to reduce the number of levels in rating factors. For example, in ratemaking for vehicles, it is possible to have rates for each individual vehicle type, probably because enough historical data have been collected over a long period. However, this may be difficult to implement and usually similar vehicles will be clustered together.

And lastly, the fourth reason is to make the rate are reasonable and smooth the rates. For example, in territory ratemaking there may be marketing, regulatory, or statute limitations requiring that adjacent territories have similar rates. Some clustering methods may reduce the probability that the rate of one location is substantially different than a neighboring area.

2.3 Nature of Insurance Dataset

The nature of data is also critical in choosing clustering method. The type of insurance data is usually numerical, such as claim frequency and severity. Some information that is originally expressed in non-numerical format, such as postcode, can be translated into a numerical format. However, in some cases such translation may be not possible; one example is occupation. The focus of this paper is on numerical data or data that could be translated numerical format.

Insurance data usually is multi-dimensional; some are related to risk characteristics and some are related to the rating factors that need to be clustered. For example, in territory clustering there may be one dimension of claim frequency and two dimensions of longitude and latitude. However, in most cases the dimension would not be too high. It is well understood that high-dimension clustering is very different to low-dimension [1]-[6]. The focus of this work is on low-dimension.

The data usually have a large noise because of the uncertainty of insurance results. Ideally, we should use expected claim frequency or expected loss in clustering. However, only the observed claim experience is available for analysis. This uncertainty should be considered in designing the similarity measure.

The insurance data also may not be well-separated and the change between clusters could be gradual. For example, in territory clustering, the difference in risk characteristics between two adjacent areas usually is quite small. So the task is to find the boundaries where the difference is relatively large. This indicates that some methods that require data well-separated may be not suitable.

3. CLUSTERING METHODS

There are many clustering methods in literature and detailed reviews are in references [1]-[6]. In this section, each method is briefly introduced, focusing on its applicability in insurance ratemaking.

3.1 Partitioning Methods

Broadly speaking, partitioning method organizes the data objects into a required number of clusters that optimizes certain similarity measure. However, it is usually narrowly defined as a method that is implemented by an iterative algorithm where the similarity measure is based on the distance between

data objects. In the context of insurance ratemaking, the distance could be the difference in claim frequency/severity or the difference in the numerical rating factors or a combination of these two.

Generally, the algorithm of partitioning methods is as follows:

(i) choose initial data objects randomly as a center or a representation of clusters;

(ii) calculate the membership of each data object according to the present center or a representation of clusters;

(iii) update the center or representation of clusters that optimizes the total similarity measure;

(iv) repeat step (ii) if there is a change in the center or representation of clusters; otherwise stop.

There are different methods to be used depending on how the similarity measure is chosen and how the center or the representation of clusters is defined.

3.1.1 *K*-Means Method

The center of the cluster m_i, is defined as the mean of each cluster C_i, that is,

$$m_i = \frac{1}{n_i} \sum_{\mathbf{x} \in C_i} \mathbf{x}$$

where $\mathbf{x}$ is the data object and is n_i the number of data objects within the cluster C_i. The total similarity measure is the squared-error function around the center of each cluster, i.e.

$$f = \sum_{i=1}^{k} \sum_{\mathbf{x} \in C_i} \left| \mathbf{x} - m_i \right|^2 ,$$

where k is the number of clusters.

This method is easy to understand and apply. The time complexity of this method is also lower than *k*-Medoids method. Generally it is one of the most popular clustering methods. However, it is very sensitive to noise and outliers because the mean of data objects in each cluster is used to represent each cluster. This is a big problem for insurance data as outliers are always expected. It is also difficult to choose the appropriate number of clusters. This may not be critical, however, because in insurance ratemaking the number of cluster may be determined by other factors such as IT limitations. The silhouette coefficient [1]-[4] is also introduced to solve this problem. The results of *K*-Means method tend to be sphere-shaped because the squared-error function is used as similarity measure. This applies to most methods that use distance as a similarity measure. This drawback is quite critical in territory clustering, as the nature cluster is not necessarily sphere-shaped. This method does not work very well when the density of data changes. Finally, the efficiency of this method is greatly affected by the initial setting and sometimes it may only converge to a local optimal. In practice, this may be solved by running the program several times with different initial

settings.

3.1.2 *K*-Medoids Method

K-Medoids method is similar to *K*-Means method but it defines the most centrally located data object of cluster C_i as the cluster center to calculate the squared-error function. Because of this, this method is less sensitive to noise and outliers. However, the procedure to find the most centrally located object requires a much higher run time than *K*-Means method [1]-[4]. This basic method is named partition around medoids (PAM) method. Clustering large application (CLARA) and clustering large applications based upon RANdomized search (CLARANS) were later proposed to reduce the time complexity [1]-[4]. However, these methods are still subject to other problems as *K*-Means method.

3.1.3 Expectation Maximization (EM)

Rather than representing each cluster by a point, this method represents each cluster by a probability distribution. In step (i), each cluster is represented by a default probability distribution. In step (ii), the probabilities of each data object belonging to every cluster C_i are calculated by the probability distribution representing cluster C_i. Then every data object is assigned to the cluster that gives the highest probability. In step (iii), the probability distribution is then re-calculated for each cluster based on new members of each cluster. If there is change in probability distribution that represents each cluster, then go to step (ii); otherwise, stop the iteration.

The time complexity of EM is lower than *K*-Medoids, but it has most of the problem *K*-Means suffers. What's more, the choice of probability distribution gives rise to more complexity.

3.2 Hierarchical Methods

Hierarchical method creates a hierarchical decomposition of the given set of data objects forming a dendrogram, a tree that splits the dataset recursively into smaller subsets. So the number of clusters is not chosen at the early stage of analysis in this method.

3.2.1 AGglomerative NESting (AGNES) and DIvisia ANAlysis (DIANA)

Both of these methods are earlier hierarchical clustering methods, where AGNES is bottom-up method and DIANA is top-down method. In AGNES, clustering starts from sub-clusters that each includes only one data object. The distances between any two sub-clusters are then calculated and the two nearest sub-clusters are combined. This is done recursively until all sub-clusters are merged into one cluster that includes all data objects. In DIANA, clustering starts from one cluster that includes all data objects. Then it iteratively chooses the appropriate border to split one cluster into

two smaller sub-clusters that are least similar.

Slightly different from the object-to-object definition of similarity measure in partitioning methods, the similarity measure in hierarchical method should be cluster-to-cluster. Different similarity measures of two clusters can be defined and common ones are

1. Min distance: $d_{\min}(C_i, C_j) = \min_{p_i \in C_i, p_j \in C_j} d(p_i, p_j)$, where $d(\cdot,\cdot)$ is a similarity measure of two data objects p_i, p_j and C_i, C_j are two clusters;
2. Max distance: $d_{\max}(C_i, C_j) = \max_{p_i \in C_i, p_j \in C_j} d(p_i, p_j)$;
3. Average distance: $d_{\text{avg}}(C_i, C_j) = \frac{1}{n_i n_i} \sum_{p_i \in C_i, p_j \in C_i} d(p_i, p_j)$,

where n_i and n_j are the size of clusters C_i and C_j, respectively.

The concept is easy to understand and apply. The resulting clusters are less sphere-shaped than partitioning methods, but still have that tendency because distance is used as similarity measure. The number of clusters is also chosen at a later stage, which is better than partitioning methods. The performance regarding noise and outlier depends on the similarity measure chosen. However, the most critical problem of the AGNES and DIANA methods is that the over-simplified similarity measure often gives erroneous clustering results, partly because the hierarchical method is irreversible. Another problem is that the complexity of time, which depends on the number of data objects, is much higher than that of *K*-Means method.

3.2.2 Balanced Iterative Reducing and Clustering using Hierarchies (BIRTH)

The method's key idea is to compress the data objects into small sub-clusters in first stage and then perform clustering with these sub-clusters in the second stage. In the second stage, the AGNES or DIANA methods could be used, while in actuarial literature BIRTH is specifically named after the method that uses a tool called clustering feature (CF) tree [3].

One advantage is that it greatly reduces the effective number of data objects that need to cluster and reduces the time complexity.

However, it still tends to have spherical shape clustering because similarity measure has the same definition as AGNES or DIANA.

3.2.3 Clustering Using REpresentatives (CURE) and CHAMELEON

The key idea of CURE is to use a fixed number of well-scattered data objects to represent each cluster and shrink these selected data objects towards their cluster centers at a specified rate. Then the two clusters with the closest "distance" will be merged. The "distance" can be defined in any way as in Section 3.2.1.

Compared with AGNES and DIANA, CURE is more robust to outliers and has a better performance when clusters have non-spherical shape. However, all parameters, such as number of representative data points of a cluster and shrinking speed, have a significant impact on the results, which makes this method difficult to understand and apply.

In CHAMELEON method, instead of distance, more sophisticated measures of similarity such as *inter-connectivity* and *closeness* are used. CHAMELEON also uses a special graph partitioning algorithm to recursively partition the whole data objects into many small unconnected sub-clusters.

CHAMELEON is more efficient than CURE in discovering arbitrarily shaped clusters of varying density. However, the time complexity, which is on order of square of number of data objects, is quite high [1]-[4].

The ability to create arbitrarily shaped clusters makes these two methods quite attractive in territory clustering. However both methods are too complicated to apply and therefore not further developed in this paper.

3.3 Density-Based Methods

Most partitioning and hierarchical methods use the similarity measure based on distance. However, density could be used as similarity measure. For example, an intuitive understanding of this method is that satellite towns around a big city can often be clustered with the big city while rural areas are not clustered.

3.3.1 Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

This method defines the density of a data object as the number of data objects within a certain distance of the data object. If the data object's density is high, the data object is very similar to its neighbors and should be clustered with those neighboring data objects. This is exactly the basic idea in DBSCAN method. After calculating the density of every data object, clusters are generated by several rules—the basic idea is to expand every cluster as long as the density of the neighboring data object is higher than the threshold. Outliers are discarded and not grouped to any clusters.

The advantage of this method is that it could find arbitrary shape of clusters. However, the efficiency of this method largely depends on parameters chosen by the user, so it requires a high level of expertise to apply this method successfully. Also it does not work very well for a large or high-dimensional dataset, because the time complexity is very high in finding all those neighboring data objects and any intermediate results are not an approximation to final results.

3.3.2 Ordering Points To Identify the Clustering Structure (OPTICS)

Rather than producing a clustering of data objects for certain chosen parameters as in DBSCAN, this method produces a cluster ordering for a wide range of parameter settings. The user then can do clustering interactively by using the cluster ordering results.

Other than finding arbitrary shape clusters, this method solves the problem of dependency on parameters as in DBSCAN. However, it still has other problems as DBSCAN does.

3.3.3 DENsity-based CLUstEring (DENCLUE)

This method is efficient for large datasets and high-dimensional noisy datasets. It can also find arbitrarily shaped clusters, which makes it suitable for insurance ratemaking. However, there are many parameters to set and it may be difficult for the non-expert to apply. Details of this algorithm are discussed in [1]-[4].

3.4 Grid-Based Methods

These methods quantize the space into a finite number of cells that form a grid structure on which all of the clustering operations are performed. The basic grid-based algorithm defines a set of grid-cells, assigns data objects to the appropriate grid cell, and computes the density of each cell. After cells with densities below a certain threshold are eliminated, clusters are generated by combining adjacent groups of cells with similar densities or minimizing a given objective function.

The advantage of these methods is fast processing time, which is typically independent of the number of data objects and only dependent on the number of cells in each dimension of the quantized space. However, its disadvantage is that the shape of the cluster is limited by the shape of grid. But this problem can be reasonably overcome by smaller grid, which has become feasible because of the rapid development of computer. So this is a promising clustering method for insurance ratemaking.

STING, WaveCluster, and CLIQUE are three advanced grid-based methods, that differ in how information about data objects is stored in each grid or what cluster principle is used. STING explores statistical information, WaveCluster uses wavelet transform to store the information, and CLIQUE discovers sub-clusters using the a priori principle [1]-[4].

3.5 Kernel and Spectral Methods

Both of these are relatively new methods. Although they originated from different backgrounds, recent studies indicate that there is a possible connection between these two methods [5], [6].

The key idea of kernel method is to map the data into high-dimensional space called *feature space*, so

that non-linear feature in the low-dimensional space becomes linear in the feature space. The conventional clustering methods introduced in previous sections are then applied in the feature space.

The main tools for spectral clustering methods are graph Laplacian matrices [6] and associated eigenvectors, which are widely studied in spectral graph theory. The original data is first transformed into the *similarity matrix*, which is defined as the matrix of similarity measure and its eigenvectors. Then the conventional clustering methods, such as *K*-Means method, are applied on the similarity matrix or eigenvectors.

Although these methods appear to be easy to implement [6], they actually are not that easy for the non-expert to use. What's more, they seems to give no more advantages than other methods in the context of insurance ratemaking. Thus, these two methods will not be further explored.

4. EXPOSURE-ADJUSTED HYBRID (EAH) CLUSTERING METHOD

The choice of clustering method depends on the feature of the data and the purpose of clustering. Most of the methods introduced in Section 3 could be used in an appropriate situation. However, in insurance ratemaking, another consideration is that the method be easy to understand and use. Based on this philosophy, this paper is focused on how to modify the partitioning and hierarchical methods to accommodate the needs of insurance ratemaking.

The proposed exposure-adjusted hybrid (EAH) method is a combination of the partitioning and hierarchical methods. This method also adjusts the similarity measure by exposure to take account of the volatility of insurance data. The whole procedure, which has been customized to territory clustering, is as follows:

1. Use the generalized linear model (GLM) technique to model the claim experience;
2. Calculate the residual of the GLM results as the pure effect of territory;
3. Use the partitioning method to generate small sub-clusters that contain highly similar data points;
4. Use the hierarchical method to derive the dendrogram clustering tree;
5. Choose an appropriate number of clusters and get corresponding clusters;
6. Repeat steps 3-5 with different initial setting to find a relatively consistent pattern in clusters;
7. Use the territory clustering results to re-run GLM and compare the results with that of Step 1. If there is large difference in the resulting relativities from GLM, then start again from Step 1; otherwise stop.

4.1 Comments on the EAH Method

The purpose of steps 1 and 2 is to calculate the "pure" effect of territory. Because of the correlation between rating factors, the effect of territory cannot be calculated by simple one-way analysis. A common approach is to use the generalized linear model (GLM) [11]. However, because the output from territory clustering usually will be fed into GLM again to calculate the final relativities, there are two possible approaches:

Approach One: Include all rating factors other than territory in the first GLM and consider the residual as the pure effect of territory. Perform the clustering analysis and have the resulting clusters fed into second GLM, which includes all rating factors.

Approach Two: Include all rating factors, including a high-level group of territories in the first GLM and consider the residual as the pure effect of territory. Do the clustering analysis and have the resulting cluster fed into second GLM.

The problem with Approach One is that the relativities of other rating factors will change between the first and the second GLM because the second GLM includes a new rating factor. Approach Two has the same problem although to a less extent. In both case, it is necessary to compare the relativities of all the other rating factors between the two GLM. If it changes significantly the whole procedure should be repeated from Step 1, and this iteration stops when the relativities don't change much between the first and the second GLM. Because in most cases Approach Two has a lesser number of iterations, it is better to take this approach if possible.

Steps 3 to 6 are the clustering procedures. They could be replaced by any other methods discussed in Section 3. However, whichever methods are used, the definition of measure similarity must be modified to accommodate the feature of insurance data. Usually, each data point has at least two types of data: one is the measure of geographical information and the other is the measure of claim potential/risk characteristics. The most common measure for geographical information is Euclidean distance:

$$g\left(x_i, y_i, x_j, y_j\right) = \left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2,$$

Where (x_i, y_i) are longitude and latitude of data object *i* and (x_j, y_j) are those of the data object *j*. However, because of the curvature of the Earth's surface, some other definitions could be used. One such formula is the Haversine formula, which gives the shortest distance over the Earth's surface between two locations.

As for the claim potential/risk characteristics, one can use claim frequency, severity or burning cost. However, since the claim severity can be quite volatile in most cases, claim frequency is most commonly used. The similarity measure can be defined as Euclidean distance $(\mu_1 - \mu_2)^2$ where μ_1 and

μ_2 are claim frequencies. However, while Euclidean distance should be calculated by the expected claim frequency, only actual claim frequency is available for analysis. The uncertainty of the data must be considered in the definition of the similarity measure. One solution is that, if it is assumed that every risk in both territories have same variance σ^2, then the observed actual claim frequency μ_1 and μ_2 are approximately normal distributed with variance σ^2/E_1 and σ^2/E_2, where E_1 and E_2 are exposures in each territory, respectively. This assumption could be justified by the Central Limit Theorem. So the variance of μ_1- μ_2 is $\sigma^2/E_1 + \sigma^2/E_2$, which is used to adjust the Euclidean distance

$$f\left(\mu_1, E_1, \mu_2, E_2\right) = -\frac{\left(\mu_1 - \mu_2\right)^2}{\left(1/E_1 + 1/E_2\right)},$$

where σ^2 is dropped as it will be merged into the weight parameter introduced next.

Another question is how to combine the two measures. The solution proposed in this paper is to use the weighed sum of two similarity measures:

$$g\left(\cdot\right) + w \cdot f\left(\cdot\right) \quad .$$

This weight *w* has to be chosen tentatively and subjectively.

In step 3, the user chooses the number of small sub-clusters. Because of the high time complexity of hierarchical clustering method in step 4, this number cannot be too high. On the other hand, if the number of small sub-clusters is too low, the performance of the EAH method will deteriorate to a partitioning method. Numbers around one hundred could be used but it also depends on the purpose of clustering and the dataset feature.

The choice of the number of clusters in step 5 is also largely subjective and usually affected by other considerations, such as IT limitations or market practice. However, the general rule is not to put the threshold at the place where there is only a small change in the similarity measure. This will be illustrated later in a case study.

4.2 Case Study

In this case study we consider the territory clustering for ratemaking in motor insurance. The data we have are the geographical information in the form of postcode, other rating factors, exposures, and actual claim numbers.

In the U.K., it is a normal practice to use the postcode as a rating factor. The postcode is in a hierarchical structure and there are about 2 million postcodes. This amount is too large to analyze, so the data is first aggregated at the postcode district level, which has about 2,900 districts. All these postcode districts are then translated into longitude and latitude.

The difference between GLM-predicted claim frequency and the actual claim frequency is the

residual that will be clustered. This is plotted in Fig. 1. The color of black and blue shows the area where actual claim frequency is worse than predicted and the color of green and yellow shows where actual claim frequency is better. Other colors mean that actual frequency is similar to GLM prediction. Although overlapping, it is quite clear that there are clusters: the London and Midlands areas are the worst risks while the North and Wales are much better.

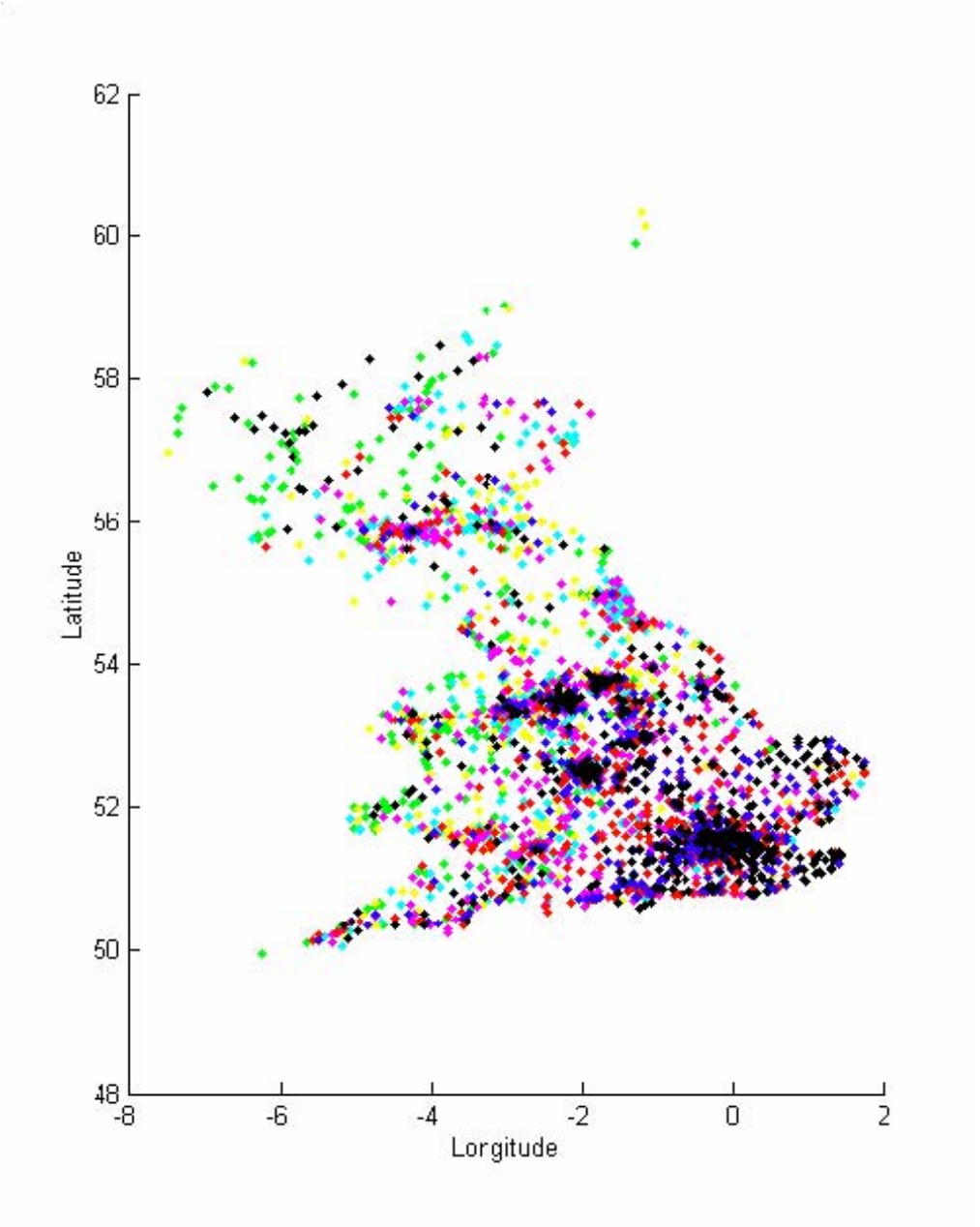


Fig. 1 Residual of GLM results that will be clustered

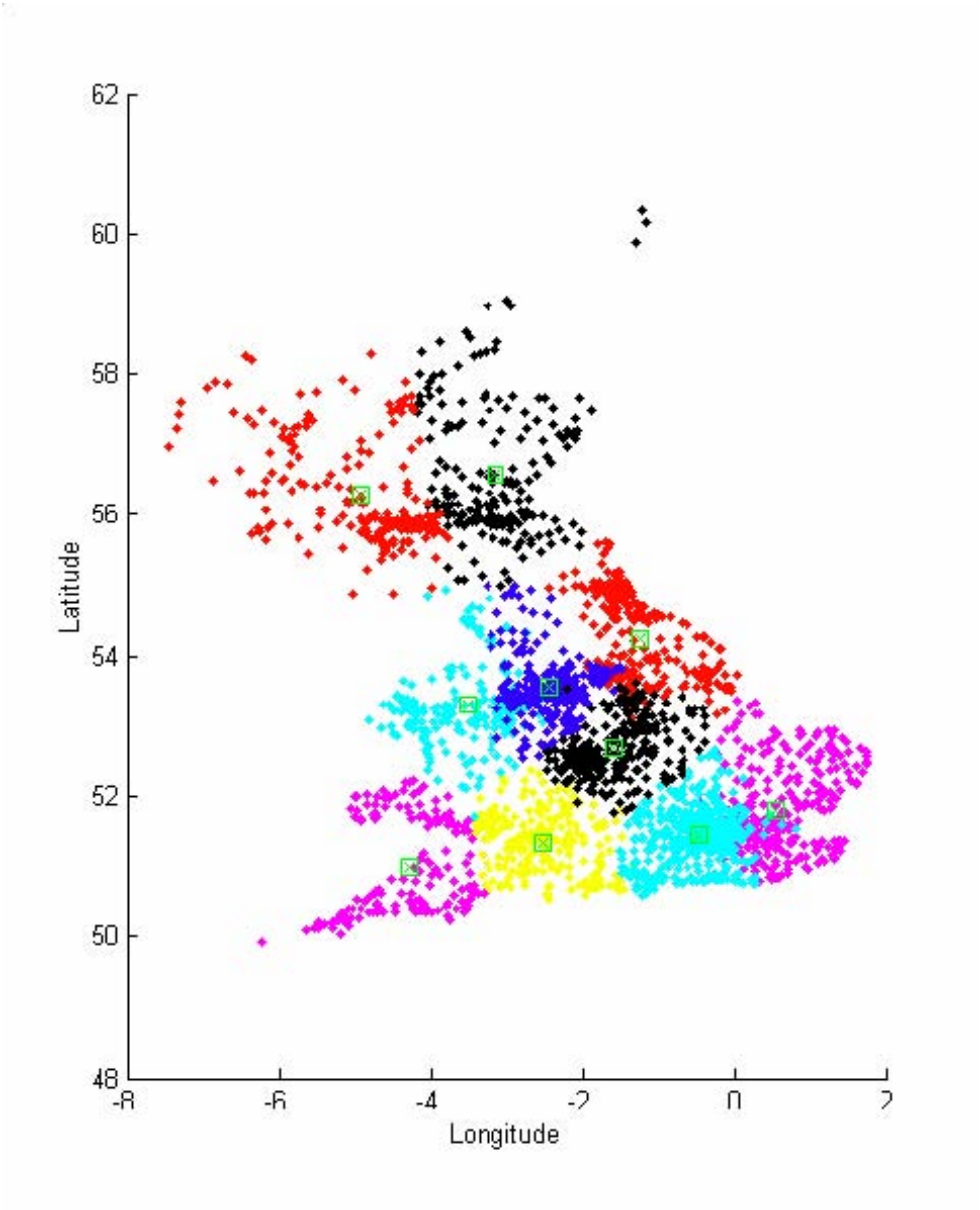


Fig. 2. Results of *K*-Means clustering method with $w = 1$.

4.2.1 Results of *K*-Means Method

The *K*-Means method is first applied and the results are plotted in Fig. 2-5 for different settings. In all cases, 10 clusters are generated and Fig. 2 plots the clustering results for the weighting parameter $w = 1$. Fig. 3 gives the result for the same weighting parameter $w = 1$ but with different initial settings. Although similar in the South, the results are significantly different in the North. It is very difficult to determine which one is better by looking at the initial dataset in Fig. 1.

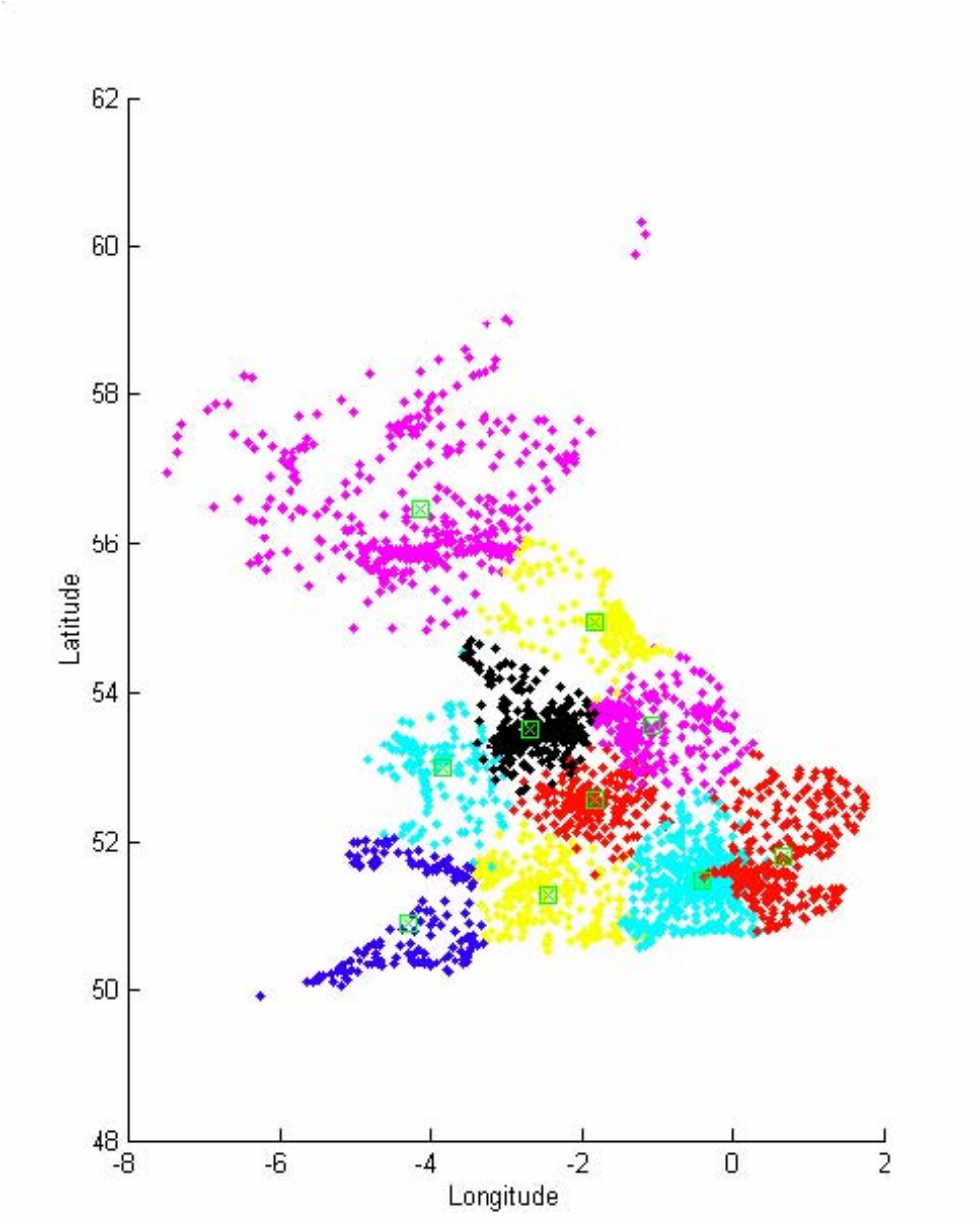


Fig. 3 Results of *K*-Means clustering method with $w = 1$ and different initial setting from Fig. 2.

The result of $w = 0.1$ is plotted in Fig. 4 and $w = 10$ is in Fig. 5. They each have the same initial settings as Fig. 2. The larger the parameter w is, the more weight that is put on the similarity measure in the claim experience and the less weight is put on the geographical closeness. Fig. 4 shows a much clearer border than Fig. 2, while Fig. 5 shows more overlapping. Probably, the result in Fig. 5 is not acceptable, but the choice between Fig. 2 and Fig. 4 is quite subjective.

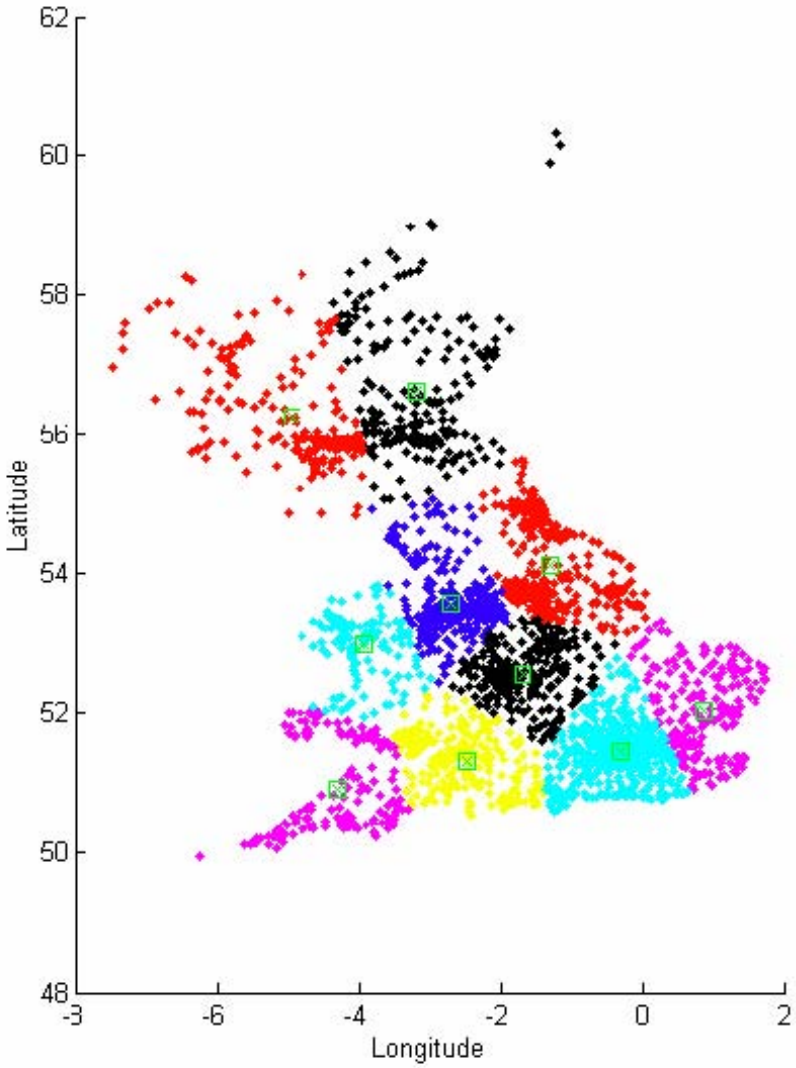


Fig. 4. Results of *K*-Means clustering method with $w = 0.1$.

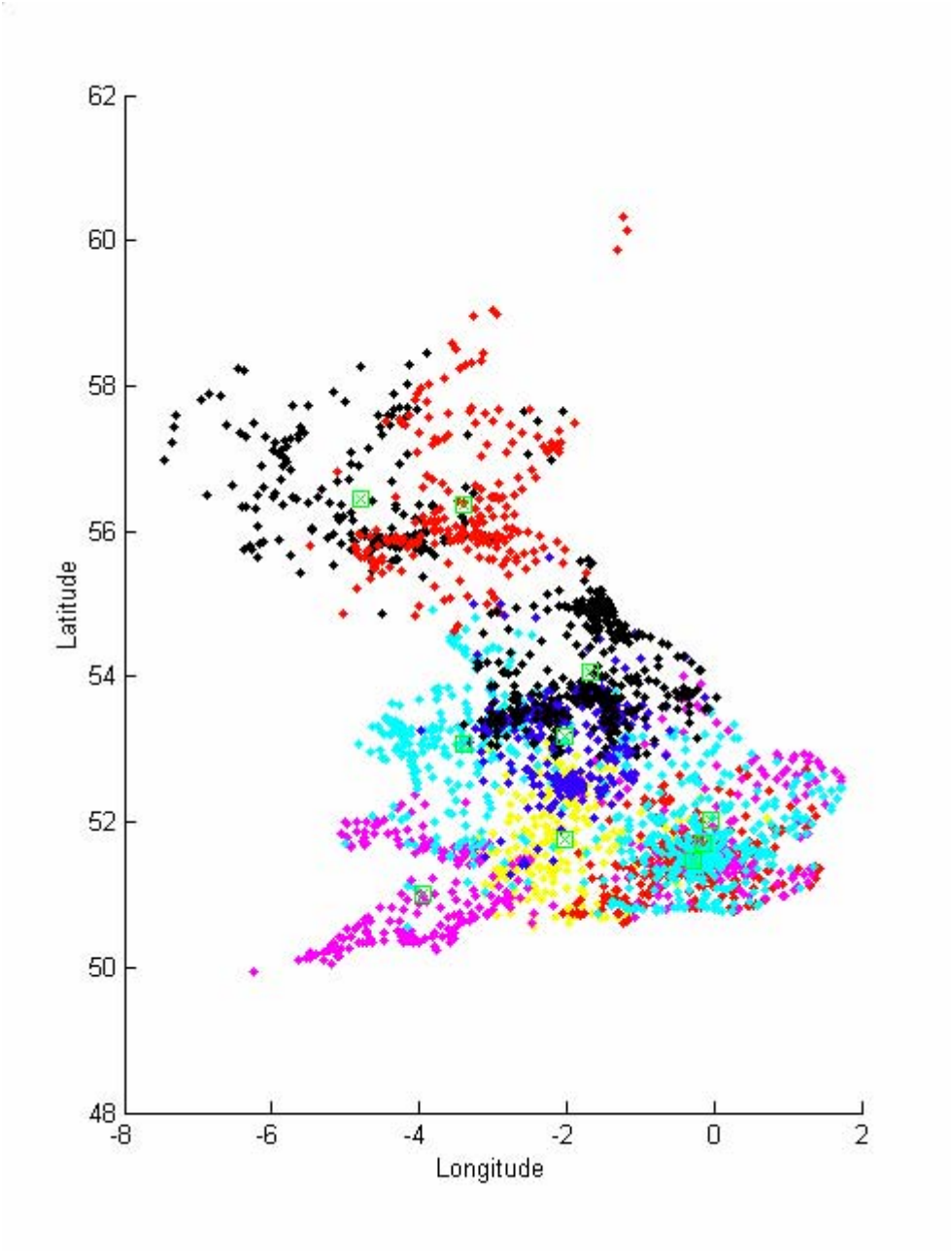


Fig. 5. Results of *K*-Means clustering method with $w = 10$.

These results highlight some features of *K*-Means methods. The sensitivity to initial settings is a big problem and the choice of parameters is also difficult. However, the dependence on parameters may be not a big problem as Fig. 2 looks very similar to Fig. 4.

4.2.2 Results of EAH Method

Then the results from each steps of the EAH method are presented. In step 3, 200 small sub-

clusters are generated. The output of the center of each cluster is shown in Fig. 6.

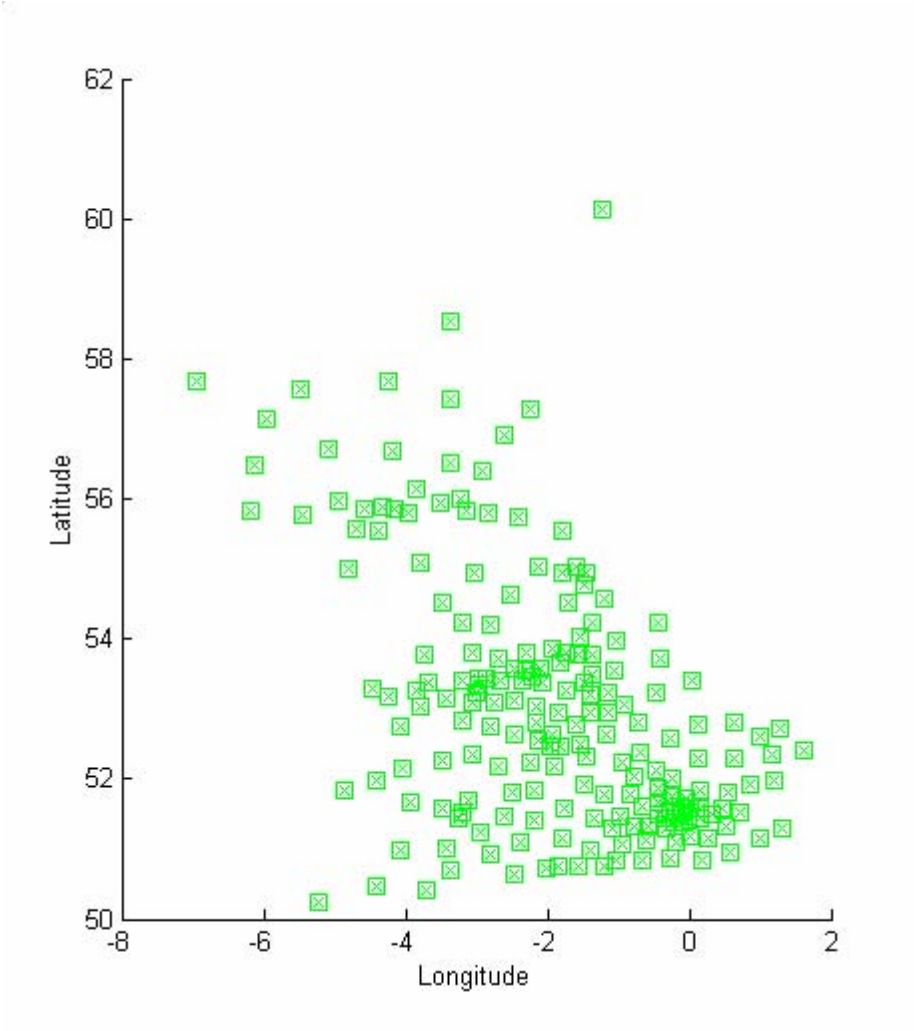


Fig. 6. Output from step 3 of 200 small clusters.

In step 4 of the hierarchical clustering method AGNES, the average distance defined in Section 3.2 is used in the weighted similarity measure proposed in Section 4.1. The resulting dendrogram is shown in Fig. 7, where the numbers of sub-clusters are not shown on the *x*-axis because there are too many to be shown in a readable format. The *y*-axis is the value of similarity measure that two sub-clusters are merged, which is termed *merging point*. The first 20 merging points are listed in Table 1. The first value is 29.0071, which means that if any two clusters with a similarity measure less than 29.0071 can be merged there will be only one cluster. Similarly, if any two clusters with similarity measures less than 4 can be merged, then there will be 8 clusters (because 4 is between 3.8875 and 4.7887).

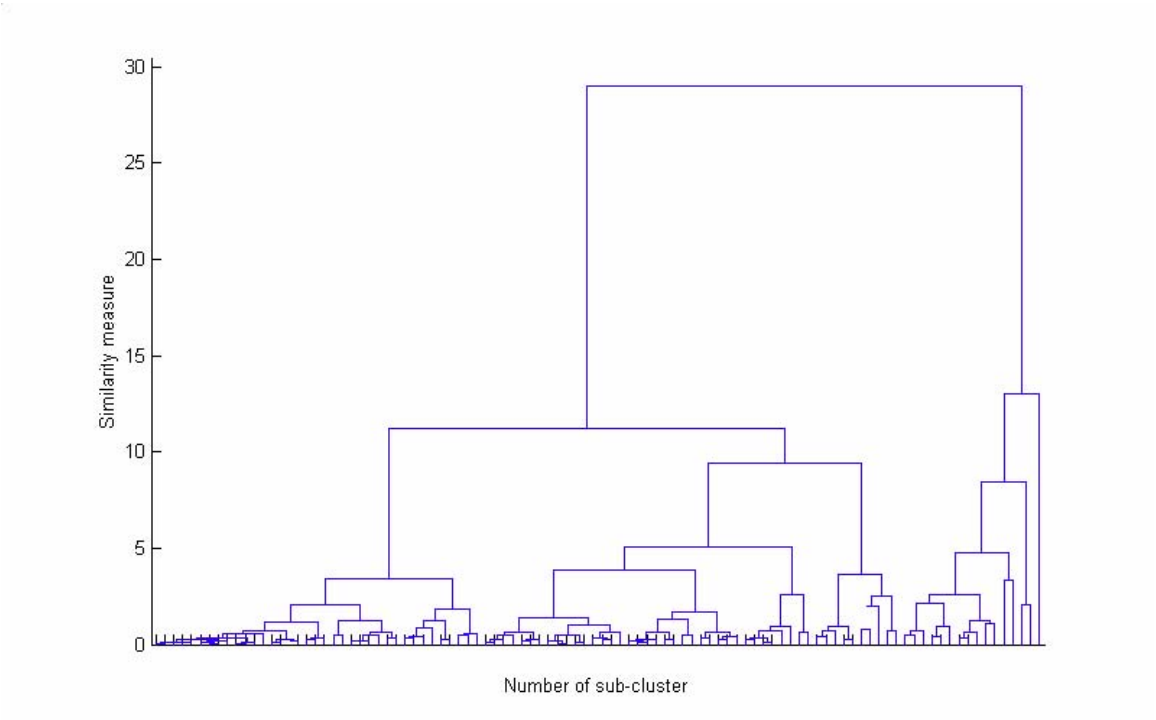


Fig.7 Dendrogram of clustering

In step 5, the number of clusters has to be chosen based on the dendrogram output. The general idea is not to put the threshold at the place where the change in similarity measure is small. If the change in similarity measures is quite small (for example, such as the gaps between numbers 10 and 11, 12 and 13, or 13 and 14 in Table 1) it is not very clear which two sub-clusters should be merged, making the results less reliable. Based on this rule, it is better to have 8 or 12 clusters in this case. The result of 12 clusters is plotted in Fig. 8.

Table 1. First 20 merging points.

Number	Similarity Measure	Change in Similarity Measure
1	29.0071	16.0212
2	12.9859	1.805
3	11.1809	1.7825
4	9.3984	0.9691
5	8.4293	3.3747
6	5.0546	0.2659
7	4.7887	0.9012
8	3.8875	0.2687
9	3.6188	0.2026
10	3.4162	0.0561
11	3.3601	0.7895
12	2.5706	0.0205
13	2.5501	0.0402
14	2.5099	0.3665
15	2.1434	0.1108
16	2.0326	0.0013
17	2.0313	0.034
18	1.9973	0.177
19	1.8203	0.1282
20	1.6921	0.2814

The whole procedure could be re-run from step 3 with different initial settings in the *K*-Means method. Another possible result is plotted in Fig. 9. There is still an apparent difference between Fig. 8 and Fig. 9, which means that this method still converges to local optimal. However, the difference is much smaller than that between Fig. 2 and Fig. 3, in this case.

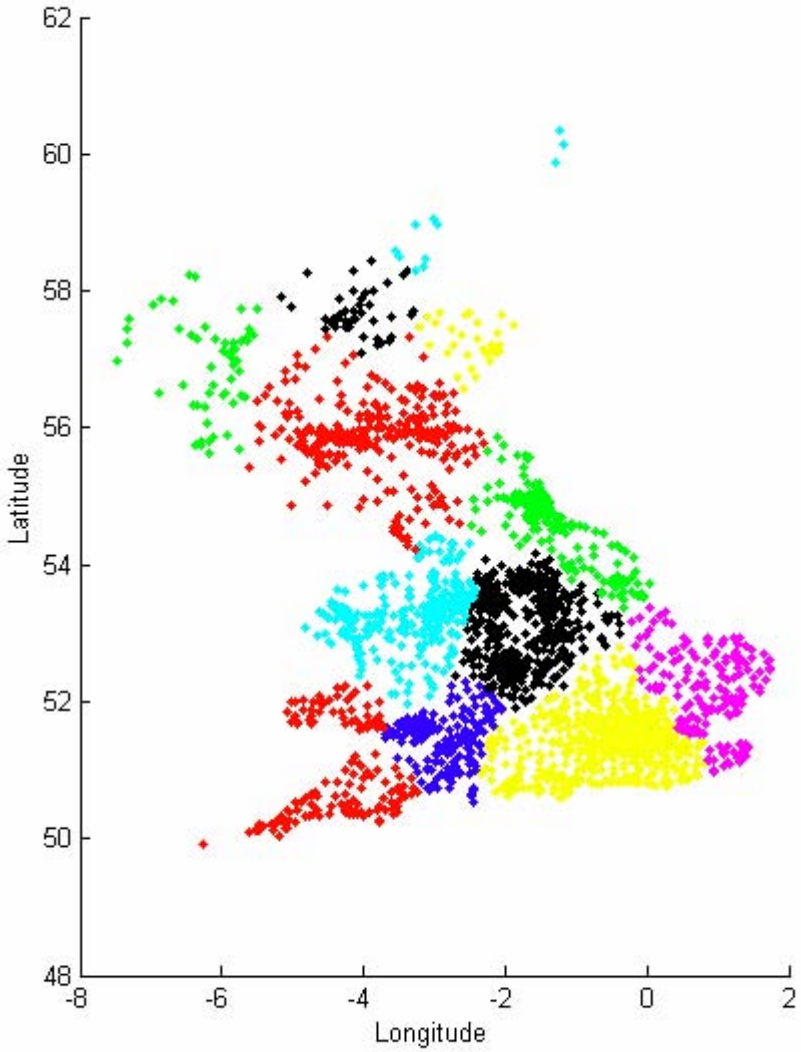


Fig.8 Clustering result by EAH method with 12 clusters

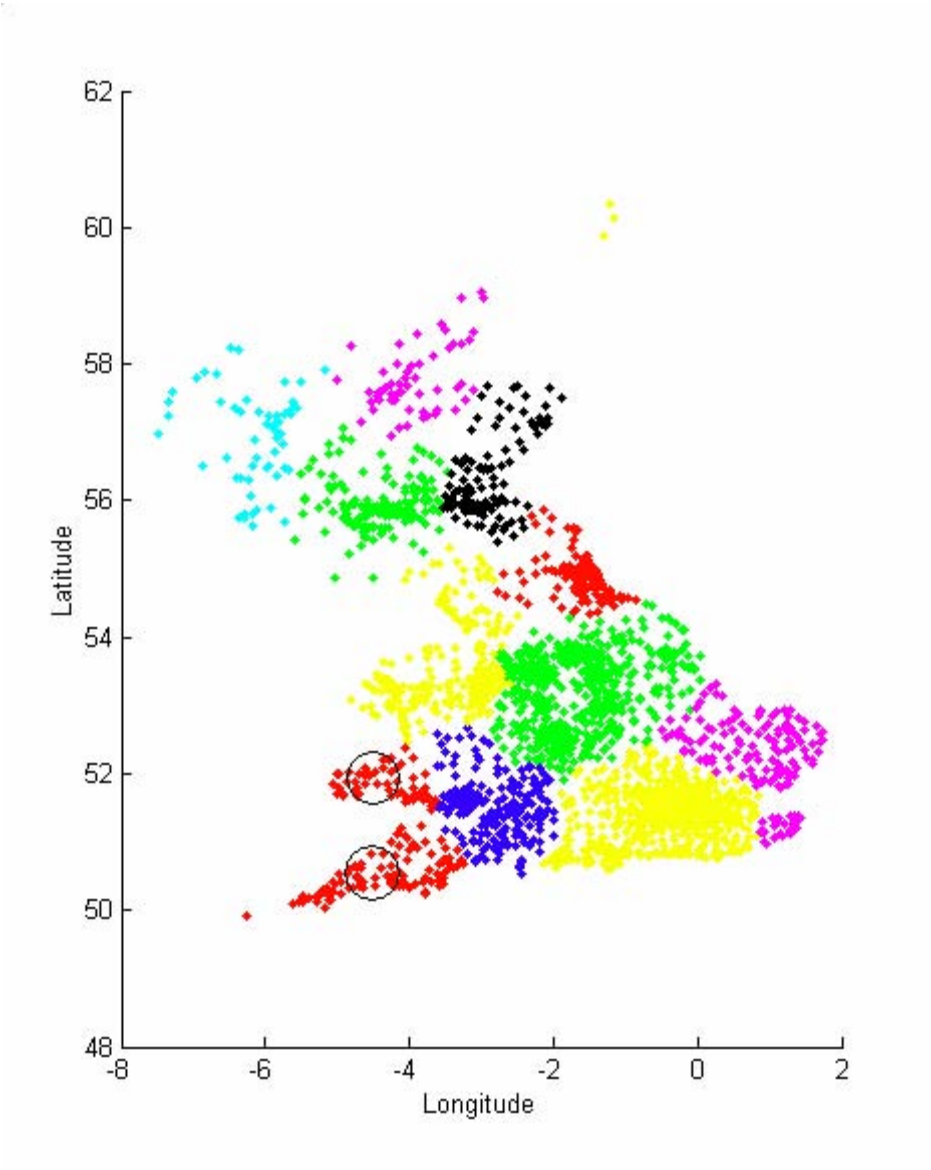


Fig. 9 Clustering result by EAH method with 12 clusters using different initial setting to Fig. 8

5. OTHER CONSIDERATIONS

In this section some other considerations in clustering are introduced and briefly explained. One common problem is the existence of obstacles and constraints in clustering. For example, in Fig. 9 the two circled areas are not adjacent because of the gulf. However, when distance is used to define the similarity measure, they are very close. One solution is to introduce a distance function that includes this information: for example, the distance between any two points across the gulf doubles the normal Euclidean distance. There are other methods in [1]-[4].

Whether to use claim frequency, severity or burning cost is also interesting. As explained in Section

4, there is strong argument for using claim frequency. However, in the case where claim severity is different between territories, it may be more reasonable to use claim severity or burning cost. In such a situation, the variance adjustment to the Euclidean distance could be different and it could also affect the magnitude of weighting parameter w.

Finally, checking the effectiveness of clustering is also difficult. As illustrated in the case study, it is very difficult to compare the results from different clustering methods or the same method with different initial settings. One solution is to repeat the clustering procedure by a large number of times and find the consistent pattern. Another is to check with external information, such as industry benchmarks. The third option is to split the data into half, using half of the data to do clustering analysis and the other half to test whether the same pattern appears.

6. CONCLUSIONS

Clustering is an important tool in data mining for insurance ratemaking. However, choosing clustering methods is a difficult task because there are a large number of clustering methods in literature and there is no conclusion as to which method is always best. The philosophy suggested in this paper is to use the simplest method possible, as long as there is no critical drawback.

Broad review of clustering methods shows that using partitioning methods without proper modifications are not suitable for insurance ratemaking. Hierarchical methods have much better performance but they are limited by highly complex calculations, so they struggle with a large dataset. Advanced methods could improve the efficiency of clustering but may be difficult to understand and apply. So from a practical point of view, this paper emphasizes modifying the partitioning and hierarchical methods to accommodate the needs of insurance ratemaking.

In the proposed exposure-adjusted hybrid (EAH) clustering method, the exposure-adjusted similarity measure is used to take account of the uncertainty of insurance data and the *K*-Means method is applied first to generate sub-clusters to reduce the time used in the hierarchical methods. Case study results show that this method could alleviate some problems of basic partitioning and hierarchical methods.

By its unsupervised nature of clustering, there is no definite choice for best clustering method; other methods introduced in this paper could give reasonable solutions in appropriate situations. However, it is hoped that various considerations mentioned in this paper could provide some practical help to users of clustering in insurance ratemaking.

Acknowledgment

The author would like to thank the reviewers for comments that greatly improved the paper, in technical aspects as well as in spelling and grammar.

7. REFERENCES

[1] Jain, A.K., M.N. Murty, and P.J. Flynn. "Data clustering: A review." *ACM Computing Surveys* 31, no. 3:264-323.
[2] Xu, R. and D. Wunsch. "Survey of clustering algorithms." *IEEE Transactions on Neural Networks* 16, no. 3:645-678.
[3] Han, J. M. Kamber, and A.K.H. Tung. "Spatial clustering methods in data mining: A survey." In: Miller, H., Han, J. (Eds.), *Geographic Data Mining and Knowledge Discovery* (London: Taylor and Francis, 2001).
[4] P. Berkhin. "Survey of clustering data mining techniques," Technical Report, Accrue Software, 2002.
[5] Filippone, M., F. Camastra , F. Masulli, S. Rovetta. 2008. "A survey of kernel and spectral methods for clustering," *Pattern Recognition* 41, no. 1:176-190.
[6] Von Luxburg, U. 2007. "A Tutorial on Spectral Clustering." *Statistics and Computing* 17, no. 4:395–416.
[7] Pelessoni, R. and L. Picec. 1998. "Some applications of unsupervised neural networks in rate making procedure." Presented at the 1998 General Insurance Convention & ASTIN Colloquium, 549-567.
[8] Sanche, R. and K. Lonergan. "Variable reduction for predictive modeling with clustering." *Casualty Actuarial Society Forum*, Winter 2006, 89-100.
[9] Guo, L. "Appling data mining techniques in property/casualty insurance." *Casualty Actuarial Society Forum*, Winter 2003, 1-25.
[10] Christopherson, S. and D. L. Werland. "Using a geographic information system to identify territory boundaries." *Casualty Actuarial Society Forum*, Winter 1996, 191-212.
[11] Anderson, D., S. Feldblum, C. Modlin, D. Schirmacher, E. Schirmacher, and N. Thandi. "A practitioner's guide to generalized linear models." *Casualty Actuarial Society Discussion Paper Program*, 2004, 1-116

Biography of the Author

Ji Yao is an actuarial analyst for Zurich Financial Services in the United Kingdom. He works in the commercial lines pricing department. Before his position at Zurich, he had two and a half years of personal lines experience. He has a BEng degree in electronic information from Shanghai's Jiao Tong University in the People's Republic of China and a Ph.D. in mathematics and statistics from the University of Birmingham, U.K. He is a Fellow of the Royal Statistical Society.