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.

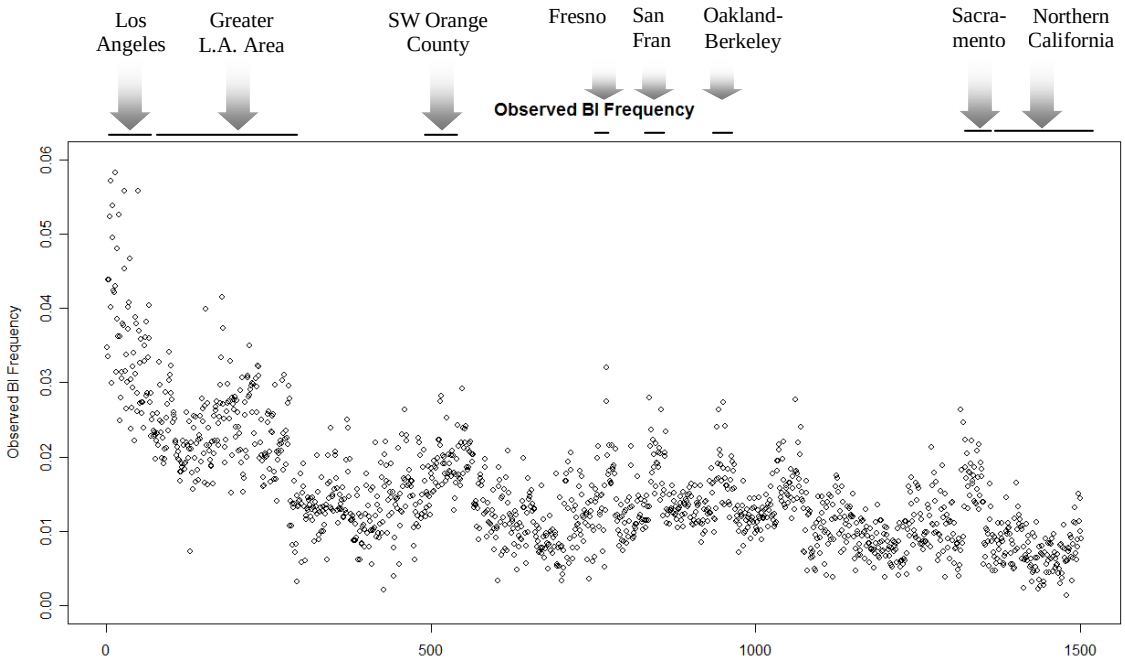
Acknowledgment

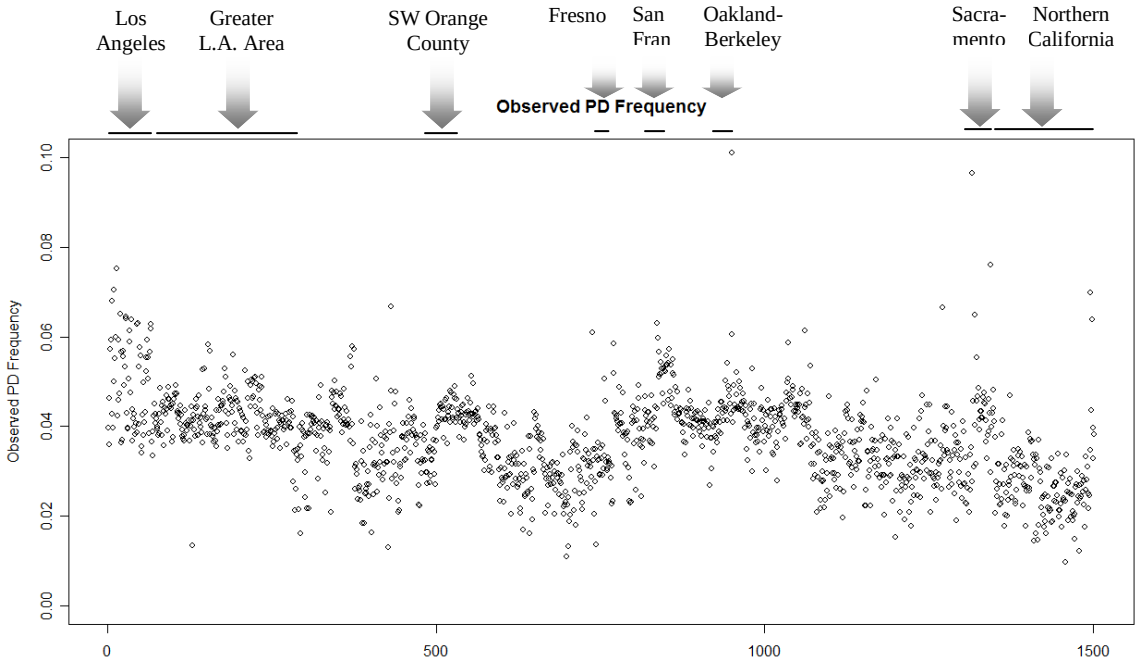
The authors thankfully acknowledge the useful comments provided by the reviewers. Any errors that may remain are their own.

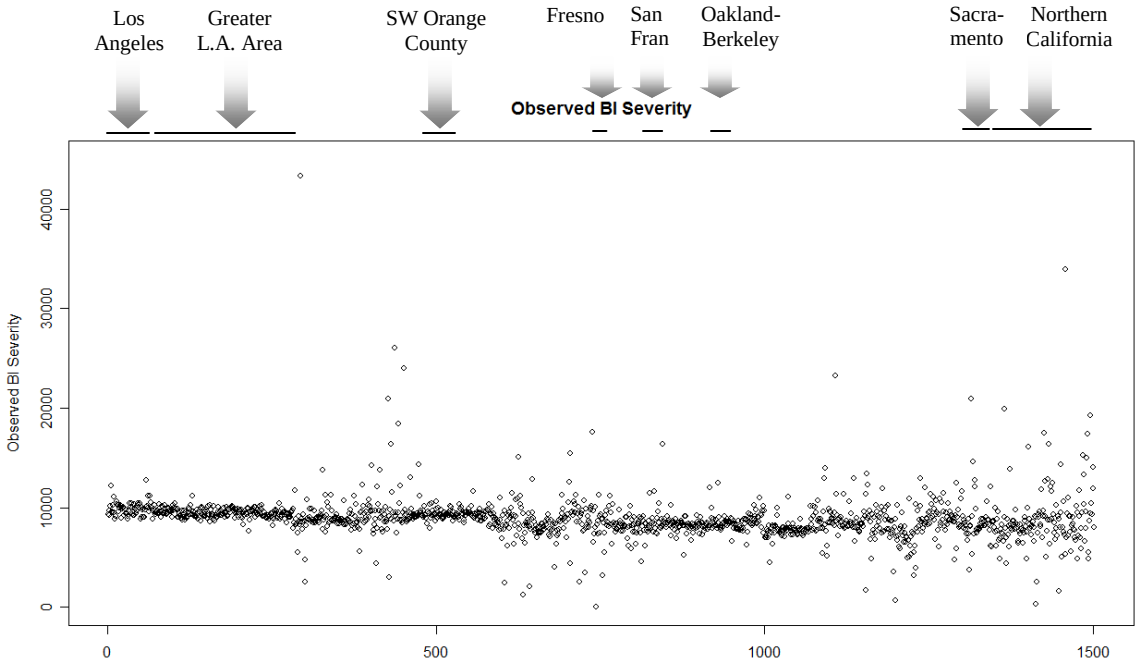
Supplementary Material

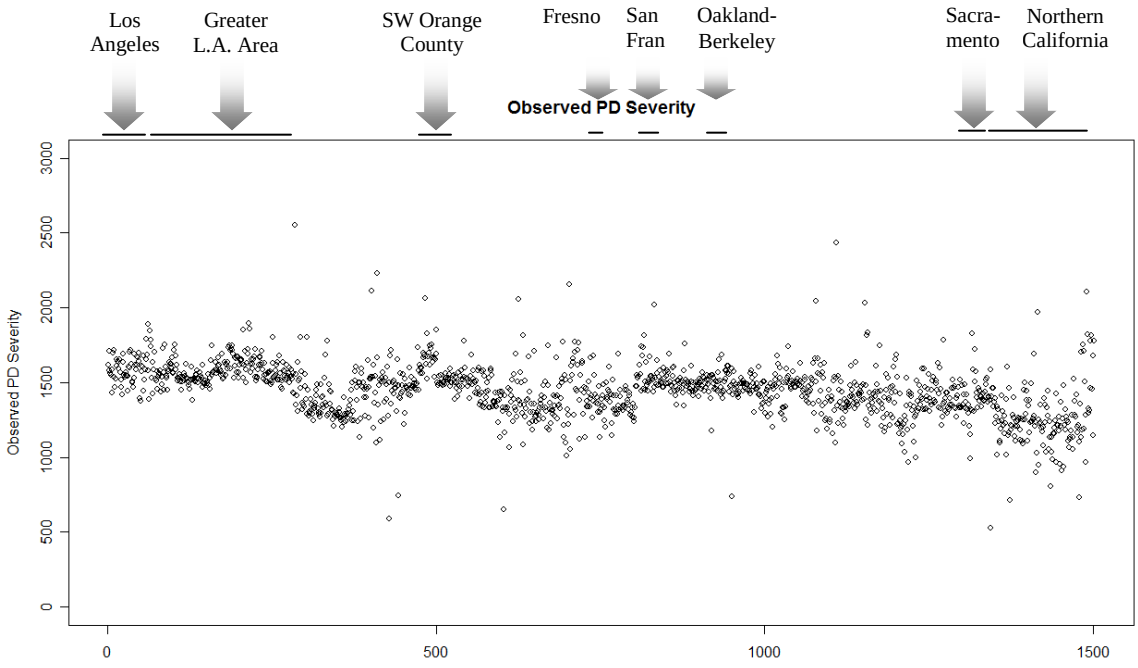
Tables containing a side by side comparison of the frequency and severity bands assigned in the present study and in Hunstad (April, 1996) [18] are available electronically on the CAS Web Site.

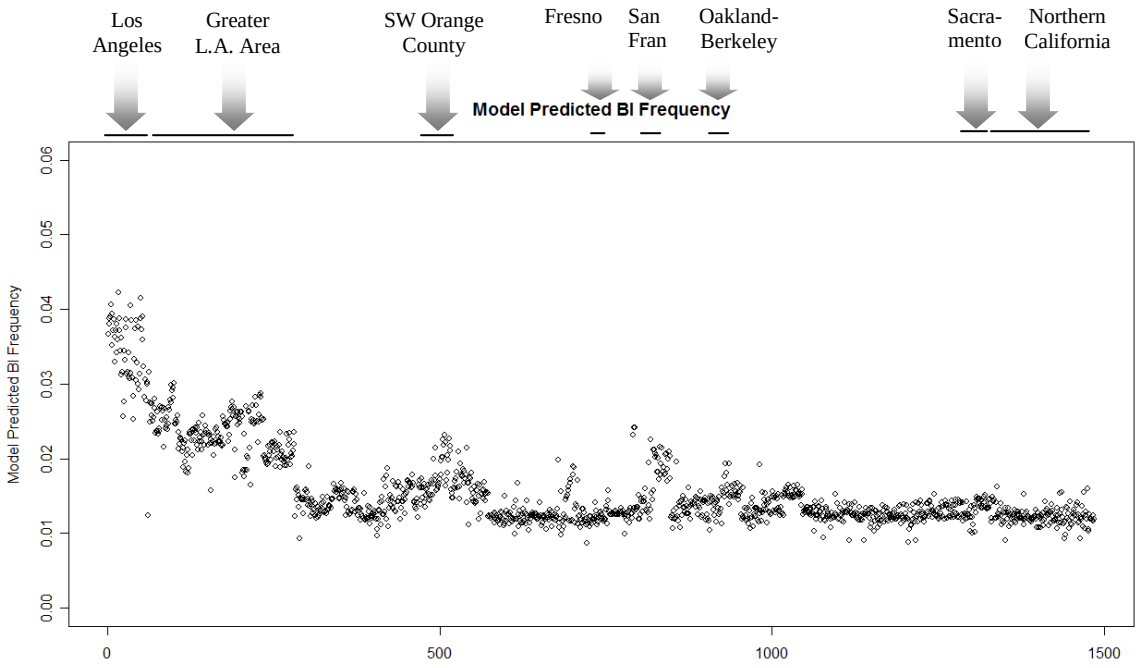
Appendix A

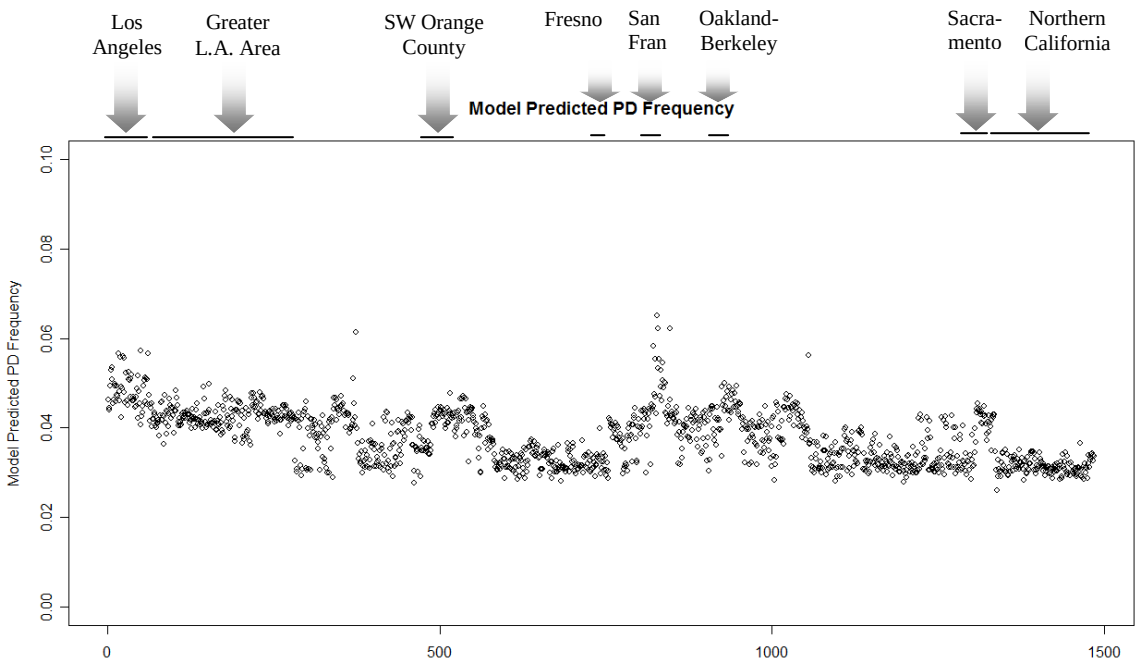

Los Angeles
Greater L.A. Area
SW Orange County
Fresno
San Fran
Oakland-Berkeley
Sacra-mento
Northern California
Observed BI Frequency
Observed BI Frequency
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0
500
1000
1500

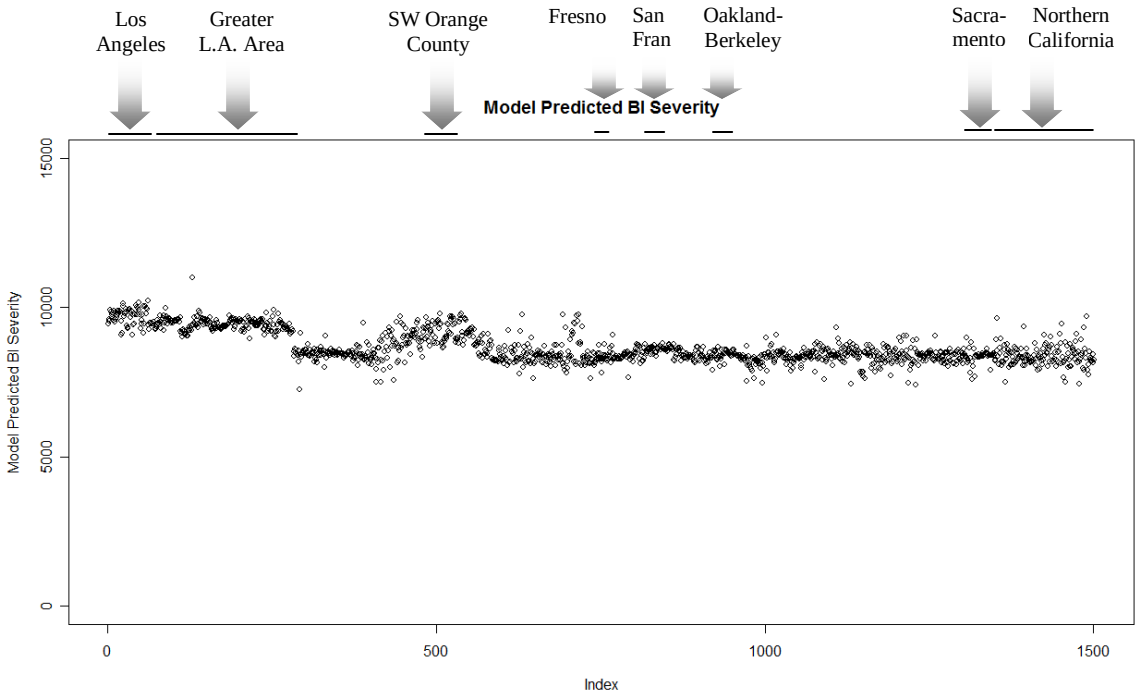

Los Angeles
Greater L.A. Area
SW Orange County
Fresno
San Fran
Oakland-Berkeley
Sacra-mento
Northern California
Observed PD Frequency
Observed PD Frequency
0.00
0.02
0.04
0.06
0.08
0.10
0
500
1000
1500

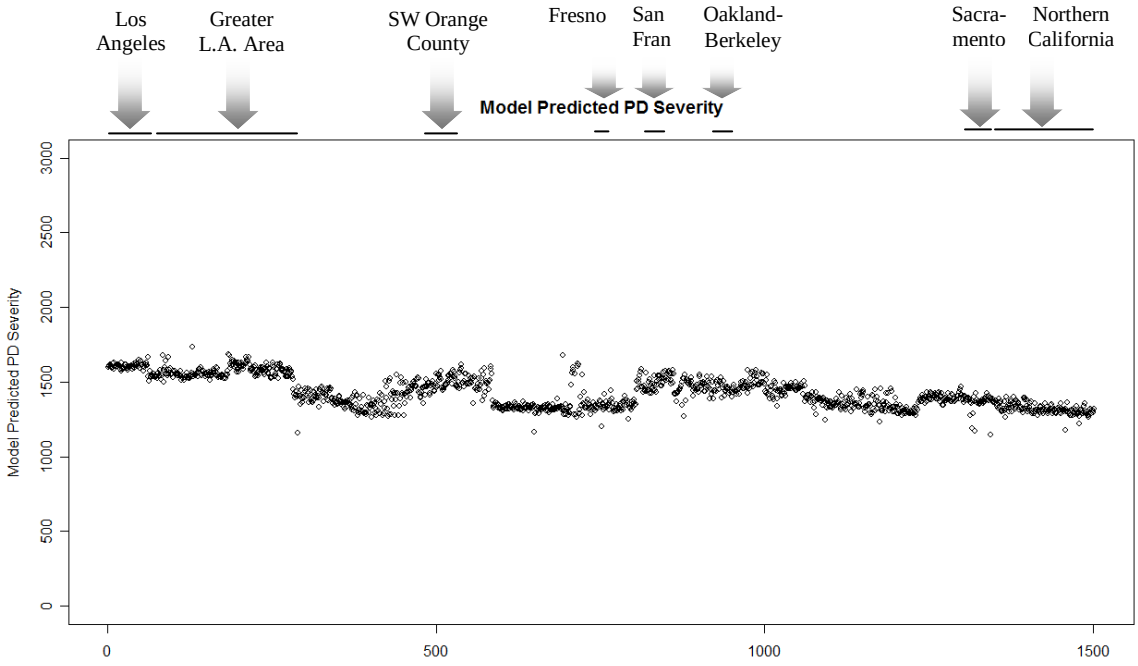

Los Angeles
Greater L.A. Area
SW Orange County
Fresno
San Fran
Oakland-Berkeley
Sacra-mento
Northern California
Observed BI Severity
Observed BI Severity
0
10000
20000
30000
40000
0
500
1000
1500

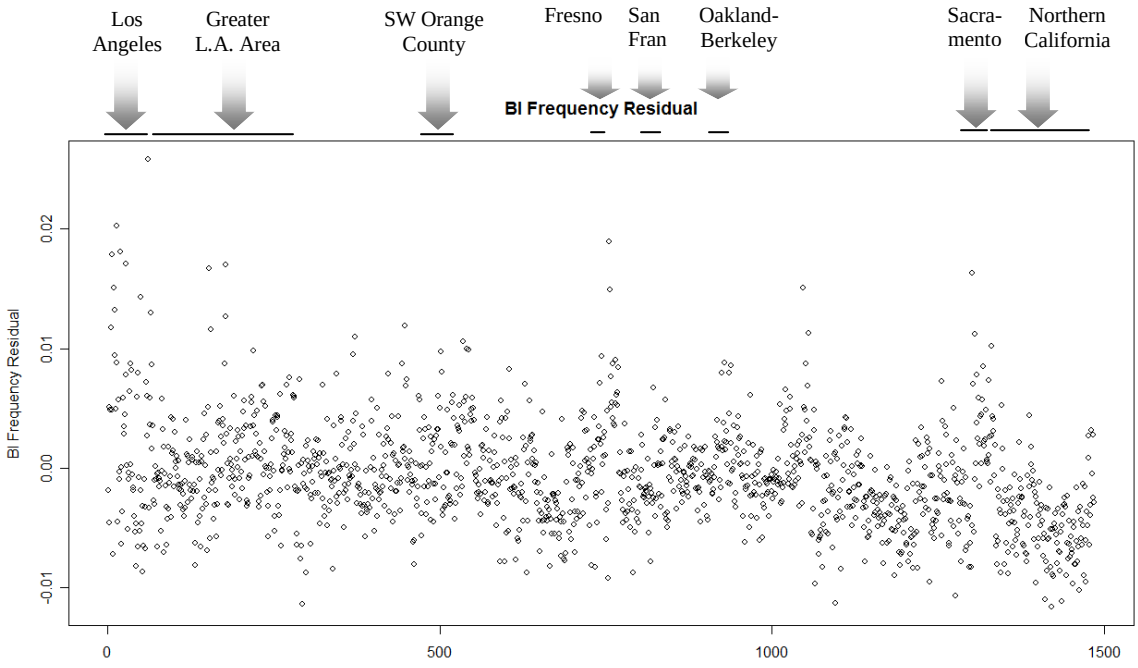

Los Angeles
Greater L.A. Area
SW Orange County
Fresno
San Fran
Oakland-Berkeley
Sacra-mento
Northern California
Observed PD Severity
Observed PD Severity
0
500
1000
1500
2000
2500
3000
0
500
1000
1500

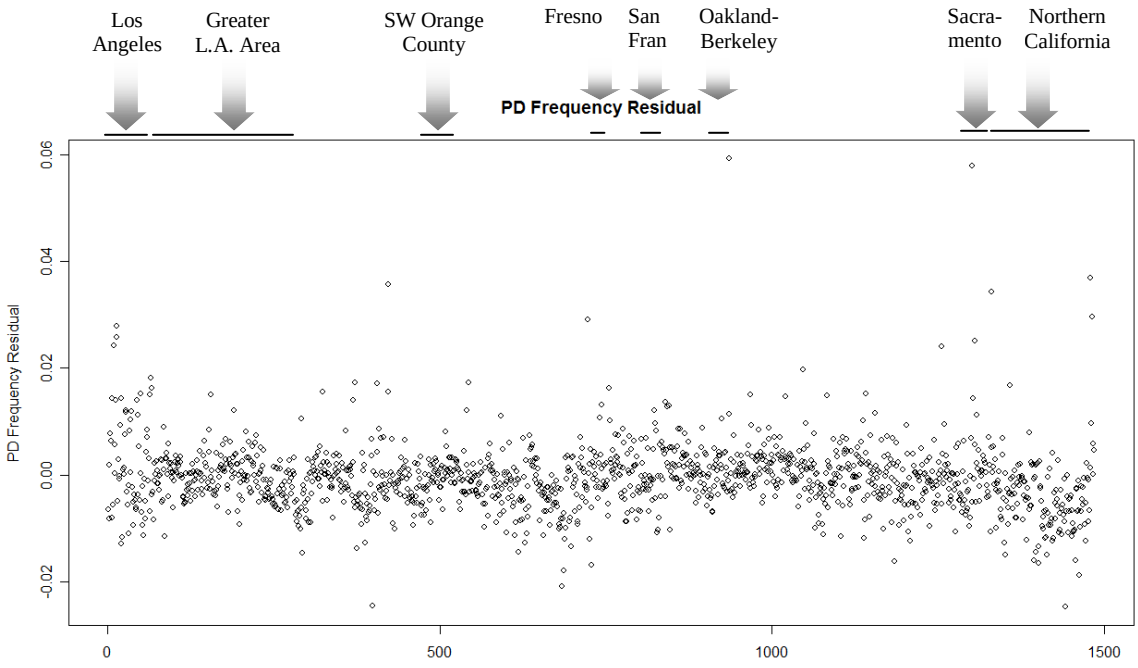

Los Angeles
Greater L.A. Area
SW Orange County
Fresno
San Fran
Oakland-Berkeley
Sacra-mento
Northern California
Model Predicted BI Frequency
Model Predicted BI Frequency
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0
500
1000
1500

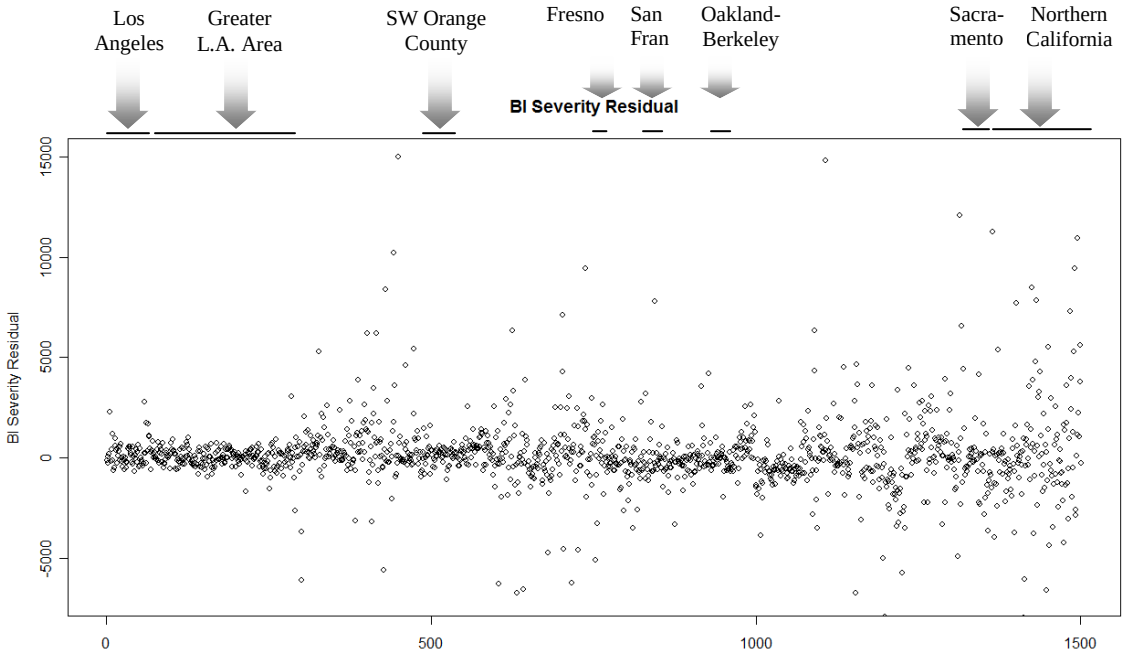

Los Angeles
Greater L.A. Area
SW Orange County
Fresno
San Fran
Oakland-Berkeley
Sacra-mento
Northern California
Model Predicted PD Frequency
Model Predicted PD Frequency
0.00
0.02
0.04
0.06
0.08
0.10
0
500
1000
1500

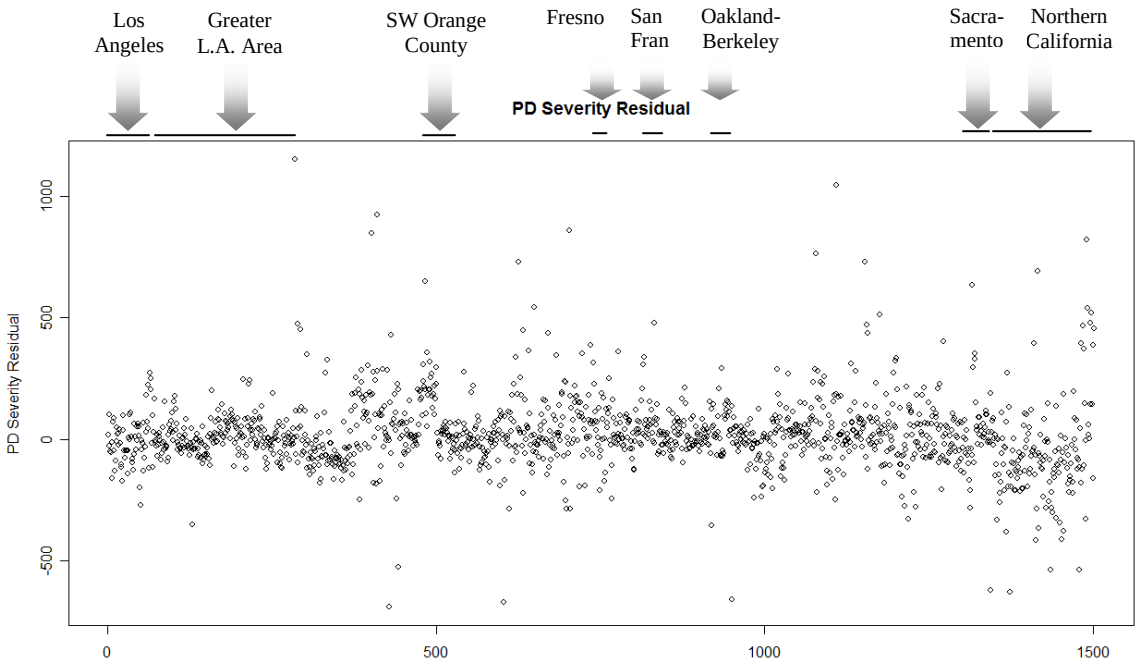

Los Angeles
Greater L.A. Area
SW Orange County
Fresno
San Fran
Oakland-Berkeley
Sacra-mento
Northern California
Model Predicted BI Severity
Model Predicted BI Severity
0
5000
10000
15000
0
500
1000
1500
Index

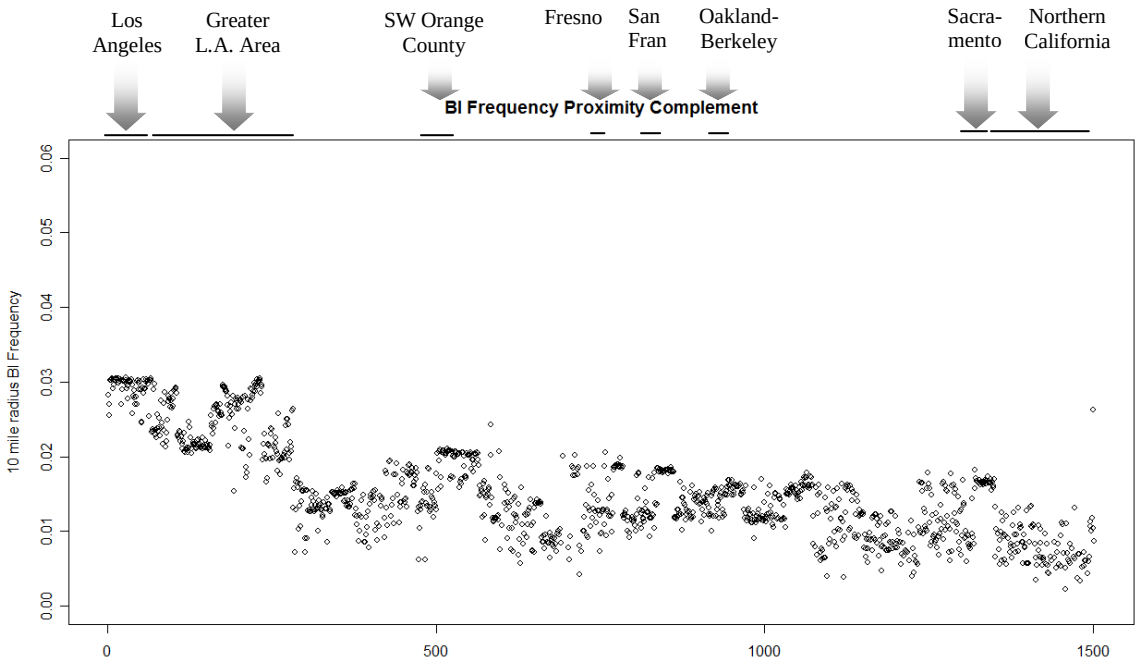

Los Angeles
Greater L.A. Area
SW Orange County
Fresno
San Fran
Oakland-Berkeley
Sacra-mento
Northern California
Model Predicted PD Severity
Model Predicted PD Severity
0
500
1000
1500
2000
2500
3000
0
500
1000
1500

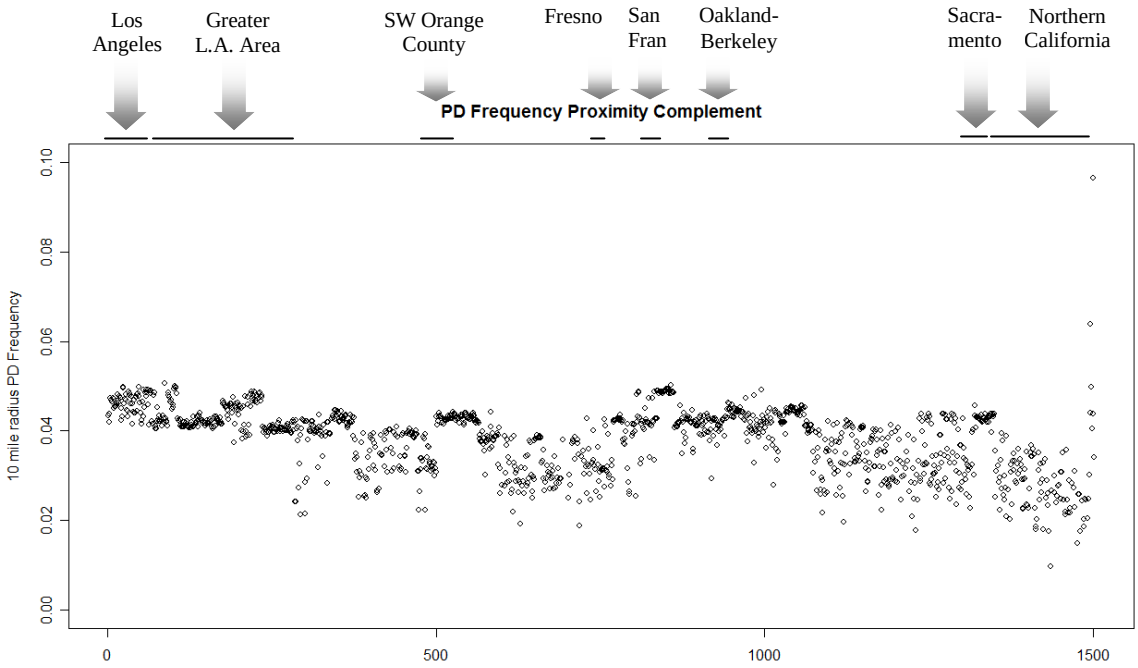

Los Angeles
Greater L.A. Area
SW Orange County
Fresno
San Fran
Oakland-Berkeley
Sacra-mento
Northern California
BI Frequency Residual
BI Frequency Residual
0.02
0.01
0.00
-0.01
0
500
1000
1500

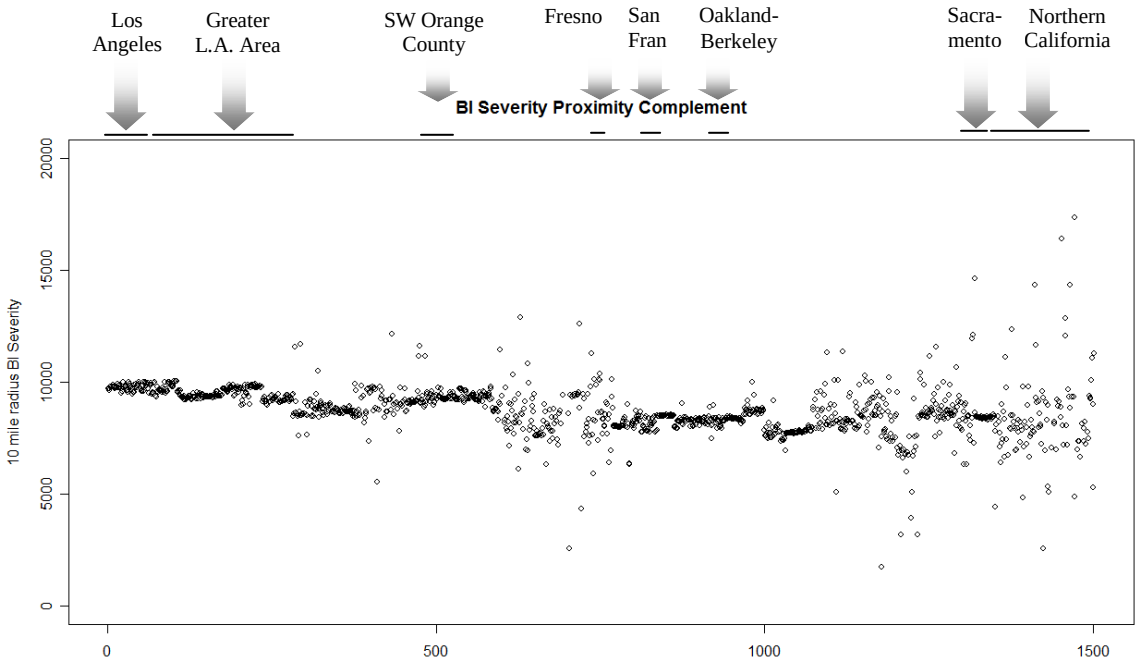

Los Angeles
Greater L.A. Area
SW Orange County
Fresno
San Fran
Oakland-Berkeley
Sacra-mento
Northern California
PD Frequency Residual
PD Frequency Residual
0.06
0.04
0.02
0.00
-0.02
0
500
1000
1500

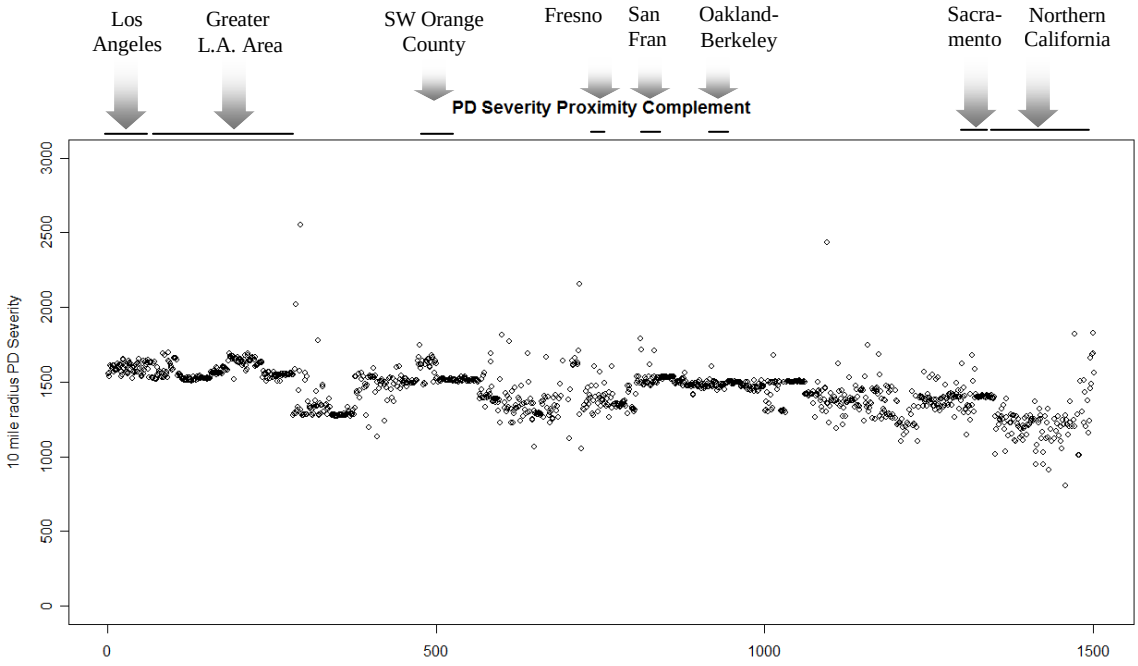

Los Angeles
Greater L.A. Area
SW Orange County
Fresno
San Fran
Oakland-Berkeley
Sacra-mento
Northern California
BI Severity Residual
BI Severity Residual
15000
10000
5000
0
-5000
0
500
1000
1500


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.