The Continuing Debate about Safety in Numbers—Data from Oakland, CA

Judy Geyer, Noah Raford and David Ragland, Traffic Safety Center;
Trinh Pham, Department of Statistics, UC Berkeley

UCB-ITS-TSC-2006-3

April 2006

TRB Paper #06-2616

Title: The Continuing Debate about Safety in Numbers—Data From Oakland, CA

Submission Date: November 15, 2005

Word Count: 3,672
Number of Figures and Tables: 8

Authors:

Judy Geyer
Traffic Safety Center
University of California, Berkeley
140 Warren Hall #7360
Berkeley, CA 94709
510-643-5659
jgeyer@berkeley.edu

Noah Raford
Traffic Safety Center
University of California, Berkeley
140 Warren Hall #7360
Berkeley, CA 94709
510-643-5659
noahraford@gmail.com

David Ragland (corresponding author)
Traffic Safety Center
University of California, Berkeley
140 Warren Hall #7360
Berkeley, CA 94709
510-642-0655
davidr@berkeley.edu

Trinh Pham
Department of Statistics
University of California, Berkeley
tpham@stat.berkeley.edu

Abstract:
The primary objective of this paper is to review the appropriate use of ratio variables in the study of pedestrian injury exposure. We provide a discussion that rejects the assumption that the relationship between a random variable (e.g., a population X) and a ratio (e.g., injury or disease per population Y/X) is necessarily negative. In the study of pedestrian risk, the null hypothesis is that pedestrian injury risk is constant with respect to pedestrian volume. This study employs a unique data set containing the number of pedestrian collisions, average annual pedestrian volume, average annual vehicle volume, and physical intersection characteristics for 247 intersections in Oakland, California. We use a GLM to estimate the expected injury risk given average annual pedestrian volume and other explanatory variables. Consistent with studies by Leden, Ekman and Jacobsen, the null hypothesis is rejected. Indeed, the risk of collision for pedestrians decreases with increasing pedestrian flows, and it increases with increasing vehicle flows. We also find that pedestrians are more likely to be struck by motorists in commercial and mixed areas than in residential areas.

INTRODUCTION

In the urban planning and traffic safety research communities, there is an ongoing debate about whether a higher number of pedestrians correlates with lower risk of motor vehicle injury for pedestrians. Several researchers have found that areas with higher numbers of pedestrian exhibit lower pedestrian injury rates per pedestrian than areas with fewer pedestrians (*1, 2*). However, others are concerned that correlating collision rate (C/P) with pedestrian volume (P), (where C equals collisions and P equals pedestrian volume) will almost always yield a decreasing relationship due to the intrinsic relationship of the variable P and the fraction 1/P. This paper considers both sides of this debate and offers a model of pedestrian injury at roadway intersections in Oakland, California, in relation to both pedestrian volume and physical intersection characteristics.

Until recently, relatively little research has been conducted that looks at pedestrian-vehicle collisions in relation to pedestrian exposure (i.e, collisions per unit of pedestrian volume). This is because pedestrian volume data are necessary to calculate pedestrian risk, and since such data are not commonly collected by municipalities, pedestrian volume is difficult to measure (*3*). Having data on pedestrian volumes allows investigation into the impact of pedestrian volumes on risk; that is, if a higher pedestrian volume effects risk per pedestrian. Several recent studies have attempted to integrate pedestrian volumes into the risk-analysis process. Jacobsen analyzed ecological data from five studies in European cities and found that pedestrian risk, defined as the number of injuries per pedestrian km-traveled, decreased as a function of pedestrian km-traveled (*2*). Leden reported similar results in a detailed study of Hamilton, Ontario (*1*), but focused his research to the question of the effects of turning vehicles.

Modeling risk per pedestrian as a function of pedestrian volume is problematic in that pedestrian volume appears on both sides of the equation. In a critique of this type of modeling, Brindle published a paper showing that if variables C, P, and A are randomly generated, C/P and P/A produce a "spurious" association (*4*). In the case of pedestrians and intersections, let:

C = number of injuries
P = number of pedestrians or pedestrian km-traveled
A = a single intersection = 1 (i.e., because each intersection is a single unit)

Then, the correlation ratio between C/P and P/A will almost always be negative (*5*).

When expanding this analysis to multiple intersections, there is cause for concern when randomly generating values for C and P, and then studying the relationship between P and C/P. The number of pedestrians varies by intersection (A). This means that pedestrians in different intersections have, in effect, been assigned a different probability of being injured. In intersections with more pedestrians, randomly distributing injuries by intersections means that these pedestrians are assigned a lower probability of being injured. Conversely, in intersections with fewer pedestrians, randomly distributing injuries by intersection means that, in effect, these pedestrians actually are assigned a higher probability of being injured. In other words, the randomization exercise suggested by Brindle is producing a real association assumed implicitly in the context of the "stratified" randomization.

Several statisticians have debated this use of ratio variables (*5, 6*). Ratio variables are used in almost all types of scientific inquiry, from disease rates to crime rates. In many of these areas, the concern that C/P and P/A will almost always be negative ignores the fact that often C and P are related. Suppose, for example, that one measures an outcome for pedestrians (P)

traversing a crosswalk. Let C be the number of pedestrians who were involved in a motor vehicle collision while traversing a specific crosswalk , and let N be the number of pedestrians who traversed the crosswalk without incident (*6*). Clearly,

$$P = C + N$$

In this example, C/P and N/P are both ratios of random variables, but it is false to say that both are inversely related to P. One consequence of the equation P = C + N is that as one of the correlations between P and C/P or N/P decreases, the other must increase. Indeed, we repeated the Brindle randomization study but weighted the random number generator for injuries by the proportion of pedestrians in each intersection. The plot of C/P by P is then a straight line.

Based on this discussion, the null hypothesis for testing the association between pedestrian injury risk and pedestrian volume is that each pedestrian has an equal chance of being in a collision at an intersection, regardless of the pedestrian volume at the time of specific crossing. If each pedestrian has the same chance of an injury, then the observed relationship between C/P and P would be a straight line. We test this hypothesis on data gathered from 247 intersections in Oakland, California, controlling for various intersection variables such as traffic control devices, lane width, number of lanes, and other intersection fixtures.

ANALYSIS

Data

All research was conducted in Oakland, California, which is located directly across the San Francisco bay from the city of San Francisco. Oakland has an economically and racially diverse population of about 400,000 people. Pedestrian volume data were generated for Oakland using the pedestrian modeling process known as space syntax (*7*). This method combined Census 2000 population and employment densities with a network analysis of pedestrian routes in order to develop annual pedestrian volume estimates for each of the city's 670 individual intersections. First, annual pedestrian trips were estimated using density data and observed movement samples. These were then distributed for every street in the city using space syntax route choice algorithms. Estimated average annual pedestrian volume for the intersections ranged from 76,896 to over 3,058,752 pedestrians per year. The output of the model demonstrated a strong correlation between predicted and observed volume counts at a sample of 42 intersections throughout the city (r-squared = 0.7717, $p < 0.001$). The model was compared with 92 different observed counts at 42 different locations, which makes r-squared of 0.7717 a relatively robust finding. **Figure 1** displays the distribution of average annual pedestrian volumes in Oakland by intersection.

People choose routes based on aesthetics, noise, traffic volume and public transit. Space syntax route choice algorithms can include these criteria via multiple regression analysis. However, these specific variables were not included in the original Oakland study. That study found that route directness, residential and employment density were found to account for the 0.77 correlation. Additional factors such as those mentioned would likely increase this correlation. The original article on space syntax (7) noted that the model under-predicted volumes at several locations, most notably those which were by parks or other quiet, aesthetically pleasing areas. It also over-predicted some routes which were one high volume, noisy vehicular streets. This confirms that those variables mentioned are important aspects of

route choice, and these limitations were suggested as areas of future research. Numerous projects including these variables have been completed since the first publication of space syntax route choice algorithms.

Vehicle volumes were then procured for 455 street segments and intersections from the City of Oakland. The data were originally gathered by the City for on-going traffic studies and comprised observations over a period of three years, between January 1, 2000 and December 31st, 2002. This data set is based upon a sample of major arterials, secondary, and local streets. Counts were conducted using a combination of automatic vehicle counters and radar speed guns. Roadway geometrics, adjacent land use, street type, number of lanes, street width, signal presence and type, and average daily traffic were all recorded and converted to Geographic Information Systems (GIS) for analysis with the Crossroads Accident Analysis Software Suite. Average vehicular traffic ranged from 11,392 vehicles per year to over 19,282,384 per year. **Figure 2** displays the distribution of average annual vehicle volumes by intersection.

Annual pedestrian-vehicle collision data were gathered from the Statewide Integrated Traffic Reporting System (SWITRS), a database of reported motor vehicle collisions that is created and maintained by the California Highway Patrol (CHP). Three years of pedestrian-vehicle collision data on 247 intersections identified 185 incidents. Mid-block collisions were linked to the nearest intersection. Only 6 out of 247 intersections were cases of mid-block linked to the nearest intersection. Average annual collisions by intersection ranged from 0 to 5. The actual average annual collisions might be slightly underestimated, since SWITRS database does not include crashes involving pedestrians that are never reported.

In 2005 the following aspects of the 247 study intersections were recorded:

Number of lanes on the primary street [natural number]
Number of lanes on the secondary street [natural number]
Traffic signal present [Boolean]
Two-way stop sign [Boolean]
Four-way stop sign [Boolean]
Other traffic control [Boolean]
Marked crosswalks [natural number]
Unmarked crosswalks [natural number]
Median present on either primary or secondary street [Boolean]
Median present on both primary and secondary streets [Boolean]
Bike Lane present on either street [Boolean]
Residential area [Boolean]
Commercial area [Boolean]
Mixed-use area [Boolean]

Methodology

The first step was to accurately model the number of pedestrian collisions as a function of pedestrian volume and control variables, including vehicle volume and intersection attributes. Next, we used the results from our regression model to test the hypothesis, that the pedestrian-collision rate remains constant as pedestrian volume increases.

To model the data, we assumed that the number of collisions has a Poisson distribution. In the data, the variance was larger than the mean, which lead us to use the *quasipoisson* model in R, a statistical software package, to account for over-dispersion in Poisson Generalized Linear

Models. The *quasipoisson* model accounts for over-dispersion and also estimates the dispersion parameter. A GLM was used because it extends linear models to accommodate both non-normal response distributions and transformations to linearity. Moreover, GLMs allow a unified treatment of statistical methodology for several important classes of models, which also include the Poisson model (*8*).

For a Poisson distribution, we have:

$$g(E(C_i)) = X_i^T \beta$$

where $C_i \sim \text{Pois}(\lambda_i)$, is the number of collisions at a given intersection i, X_i^T is the transpose of the design matrix that contains the independent variables, and β is the vector of unknown parameters. The function below is the natural link function:

$$g(x) = \log x$$

So we now have:

$$\log(E(C_i)) = X_i^T \beta$$

and hence, the Poisson regression model is as follow:

$$E(C_i) = e^{x_i^T \beta}$$

Note: *log(e)* is an example of a GLM link function.
Since we observed that $\text{var}(C_i) > E(C_i)$ for all intersections i, we must use the *quasipoisson* model in R to account for over-dispersion in Poisson GLMs.

We performed backward elimination on the initial model with ten predictors at 247 intersections: annual vehicle volume, annual pedestrian volume, number of lanes on the primary street, number of lanes on the secondary street, a categorical variable to capture the geometry of the intersection (signalized or zero-way stop to 4 way-stop), number of marked cross-walk, number of unmarked cross-walk, a dummy variable to indicate whether the intersection has a median or not, a dummy variable for bike-lane, and a categorical variable for neighborhood type (residential, commercial, or mixed). Predictors with p-values higher than the test level of .05 were eliminated from the model.

One might consider whether backward elimination is an appropriate procedure for the selection of our final model. Both forward and backward selections have their own drawbacks. In forward selection, each addition of a new variable may render one or more of the already included variables non-significant. In backward selection, sometimes variables are dropped that would be significant when added to the final reduced models. Stepwise selection is a compromise between forward and backward selection methods. It allows for moves in either direction, dropping or adding variables at the various steps. Stepwise selection was not an option with the regression here using *quasipoisson* GLM, because in quasi models, there is no likelihood, and hence no Akaike Information Criterion (AIC) statistics (*8*). The first check we

made to confirm the appropriateness of using backward selection to select our final model was by running both backward and forward selections. Both directions resulted in the same model for our case. This means that stepwise selection will also result in the same model. There is no guarantee that both backward and forward selections will always yield the same results, however. The second check was by using stepwise selection in the Poisson GLM, since a Poisson distribution was reasonable because the dispersion parameter was not greatly different from 1, where 1 indicates that there is no over-dispersion in the data. The final model using stepwise selection in the Poisson GLM resulted in the same predictors. The two checks that we performed support the appropriateness of using backward elimination.

We hypothesized that the predictors mentioned above are significant variables that belong in our model. For example, the number of lanes on the primary street would be highly correlated with the number of collisions at an intersection since wider streets increase a pedestrian's exposure to vehicular traffic. Also, additional lanes are indications of heavier traffic, which might make a pedestrian less visible and a driver feel less accountable for yielding to a pedestrian. The categorical variable to capture the geometry of the intersection is another relevant variable. Besides capturing the geometry of the intersection, it also measures the safety of the intersections. For example, intersections with a signal or an increase in the number of stops are thought to be safer; hence we would expect there would be fewer collisions in those intersections.

RESULTS

Contrary to our beliefs, some of the variables that we believed to be relevant dropped out of the final model. The variables that dropped out of the regression during the process of backward elimination are number of lanes on the primary street, number of lanes on the secondary street, a categorical variable to capture the geometry of the intersection, number of marked cross-walks, number of unmarked cross-walks, presence of a median, and presence of a bike-lane. It was a surprising result that the number of lanes on both the primary and secondary streets dropped out of the final model. A possible explanation for this result is that the variable number of lanes is highly correlated to the neighborhood type.

Three statistically significant variables remained in the model. These three variables are annual pedestrian volume, annual vehicle volume, and neighborhood type (residential, commercial, or mixed-used areas). The dispersion parameter of collisions, as a *quasipoisson* random variable, is 1.169. This is not greatly different from 1, where 1 indicates that there exists no over-dispersion. Hence, a Poisson distribution is sufficiently reasonable to describe the distribution of collisions. Moreover, when the results from the *quasipoisson* regression are compared with the results from the Poisson regression, the parameter estimates are precisely the same; the standard errors corresponding to the estimates are slightly different, but the difference is not significant. **Tables 1 and 2** provide the parameter estimates, standard errors, t-values or z-values, and p-values for the *quasipoisson* model and Poisson model, respectively.

We also looked at Cook's distance plot, which showed three potential influential observations that might require further investigation. In R, there is a function called robust linear model (*rlm)*, which fits a linear model by robust regression. The original model was run using this function to determine whether the potential influential points actually had a great impact on our results. If the estimates and their corresponding statistics were not significantly different, then we could conclude that those observations were not influential enough to change our

estimates. Since this was the case, the three significant variables in the original model were retained in the model.

In the categorical variable for neighborhood type, the category 'residential area' was omitted. The estimates for the categories 'commercial area' and 'mixed area' were positive, which indicated that there were more collisions in these areas. The interpretation of the estimates in GLMs for *quasipoisson* models is slightly different than in ordinary linear models. The value of 0.4977 for commercial areas means that the number of collisions will be higher by e^0.4977 in commercial area where all other variables are held constant. The interpretation for all of the other predictors is similar to the one just described.

The estimates for annual pedestrian volume and annual vehicle volume were very close to zero, indicating that for each additional pedestrian or car on the street, the number of pedestrian-vehicle collisions increases only slightly. If the number of pedestrians is increased by 100,000 pedestrians at a particular intersection in a given year, the number of collisions will increase by only 1.06 where all other variables are held constant. **Figure 3** plots the number of pedestrians against the number of collisions. It shows that the rate of the number of collisions increases, but very slowly and the increase is not consistent over each interval of pedestrian volume. In other words, this curve lies deeply below the line of x = y on this graph. In fact, the collision rate per pedestrian decreases as the number of pedestrians increases. **Figure 4** plots the number of pedestrians against the rate of collisions over pedestrians using the fitted values from our regression model as estimates for the true number of collisions. However, the rate of collisions per pedestrian increases as the number of vehicles increases. **Figure 5** plots the number of vehicles against the rate of collisions per pedestrian.

The earlier argument in the introduction that collisions (C) and pedestrian volume (P) are related implies that the negative relationship between C/P and P is not spurious. As stated earlier, if P = C + N where N is the number of pedestrians who traversed the crosswalk without incident, then the ratios C/P and N/P cannot both decrease as P increases.

Figure 6 plots the number of pedestrians against the rate N/P, where N is the difference between the average annual pedestrians and the average annual collisions from our data set. The reason why there are many points at N/P = 1 is because we have many intersections where the number of collisions is zero. Therefore, the *lowess* smooth line that summarizes the trend of the rate N/P as a function of P is pulled up in the first 1,000,000 pedestrians range. If we ignore the intersections with no incidents, then the plot clearly shows that N/P increases as the number of pedestrians increases. This implies that the rate of C/P and P has a negative relationship is a valid result.

DISCUSSION

Our results suggest that the risk for pedestrian-vehicle collisions (where risk is calculated as collisions per pedestrian) is smaller in areas with greater pedestrian flows and greater in areas with higher vehicle flows. These results from 247 Oakland intersections are consistent with previous studies in Europe, Canada, and other California cities (*1,2*).

These findings may have important implications for pedestrian safety planning and transportation policy. Policy makers and traffic engineers are often reluctant to make modifications to the roadway network that will increase pedestrian traffic for fear that the change might increase the total number of pedestrian injuries per year. This research suggests that the risk of collision for individual pedestrians is significantly lower in areas with higher pedestrian

volume. The estimated rate of total collisions grows slowly as well. This finding, combined with the finding that risk for pedestrian-vehicle collisions is higher in areas with more vehicles, suggests that these fears may be unwarranted. Indeed this research suggests that the opposite may be true, and that the more pedestrians on the street, the safer for everyone.

Posted speed limits are an important variable that we have not incorporated in our model selection and analysis as these data were not available. Injury severity is not studied here, which would be effected by speed limits. Slower vehicle speeds would be expected to contribute to pedestrian safety since pedestrians would have more time to detect and avoid motor vehicles. If this implication were valid, we would conclude that injuries per pedestrian volume can be explained both by pedestrian volume and by vehicle speed.

In future work, this research team aspires to create a generalized model for the risk of pedestrian-vehicle collisions. Such a model would include significant variables from the current model as well as other relevant variables as our research continues. A question of interest is whether it is more important to reduce the rate of collisions (i.e., collisions per pedestrian) or to reduce the number of collisions overall. Also, how might the increase in the number of pedestrians affect traffic? If in response, vehicle traffic is slowed, then the average motorist trip time might increase. We might want to develop a way to quantify the costs and benefits of such changes to analyze this trade-off between reduced risk for pedestrians, and increase in vehicle traffic and average motorist trip time.

ACKNOWLEDGEMENTS

The authors would like to thank Rasha Aweiss, who helped us in the process of collecting data. We would also like to thank Jamie Hollander and Frances Tong, who are graduate students from the department of statistics at University of California, Berkeley, for their contribution.

REFERENCES

1. Leden, L. *Pedestrian risk decrease with pedestrian flow. A case study based on data from signalized intersections in Hamilton, Ontario*. Accident Analysis and Prevention, Volume 34, pp. 457-464. 2002.

2. Jacobsen, PL. *Safety in numbers: more walkers and bicyclists, safer walking and bicycling*. Injury Prevention, Volume 9, pp. 205-209. 2003.

3. NHTSA/FHWA Pedestrian and Bicycle Strategic Planning Research Workshops, Final Report, April, 2000.

4. Brindle, R. *Lies, damn lies, and "automobile dependence"- some hyperbolic reflections*. Australasian Transport Research Forum, Vol. 19, pp 117-131. 1994.

5. Schuessler, K.F. *Analysis of ratio variables: Opportunities and pitfalls*. American Journal of Sociology, 1974. 80:379-396.

6. Long, Susan B. *The Continuing Debate over the Use of Ratio Variables: Facts and Fiction*. Sociological Methodology, Vol. 11, 37-67. 1980.

7. Raford, N. and Ragland, D. *Space Syntax: An innovative pedestrian volume modeling tool for pedestrian safety*. Annual Meeting of the Transportation Review Board, 2003.

8. Venables, W.N. and Ripley, B.D. *Statistics and Computing: Modern Applied Statistics with S*. Fourth Edition. 2002 Springer-Verlag New York, Inc.

TABLES AND FIGURES

Table 1.

Model: $Collision = e^{\alpha+\beta_1(Vehicle)+\beta_2(Pedestrian)+\beta_3(Neighborhood)}$,

where neighborhood is a categorical variable with 0 = residential, 1 = commercial, 2 = mixed

Family = *quasipoisson*

Variable	Estimate	Std. Error	t-value	p-value
Intercept	-1.712	2.694e-01	-6.354	1.03e-09
Vehicle	6.027e-08	2.424e-08	2.487	0.0136
Pedestrian	5.942e-07	1.267e-07	4.692	4.54e-06
Neighborhood1	4.977e-01	2.349e-01	2.119	0.0351
Neighborhood2	4.933e-01	2.020e-01	2.443	0.0153

Table 2.

Model: $Collision = e^{\alpha+\beta_1(Vehicle)+\beta_2(Pedestrian)+\beta_3(Neighborhood)}$

where neighborhood is a categorical variable with 0 = residential, 1 = commercial, 2 = mixed

Family = *Poisson*

Variable	Estimate	Std. Error	z-value	p-value
Intercept	-1.712	2.491e-01	-6.872	6.35e-12
Vehicle	6.027e-08	2.241e-08	2.689	0.00717
Pedestrian	5.942e-07	1.171e-07	5.073	3.91e-07
Neighborhood1	4.977e-01	2.172e-01	2.291	0.02196
Neighborhood2	4.933e-01	1.868e-01	2.641	0.00826

Figure 1. Estimated Annual Pedestrian Volumes at Intersections Vary by Different Colors

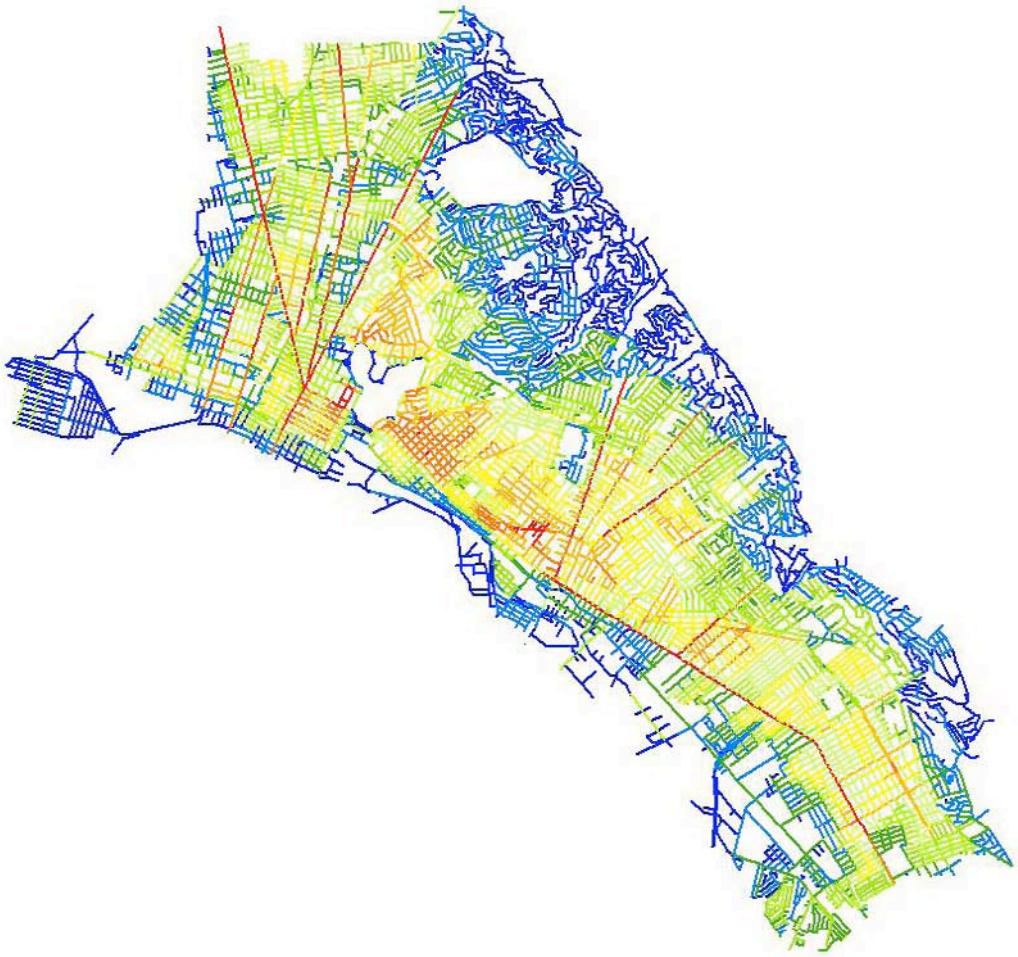

Figure 2. Vehicle Volume Counts at Intersections Vary by the Size of the Circles

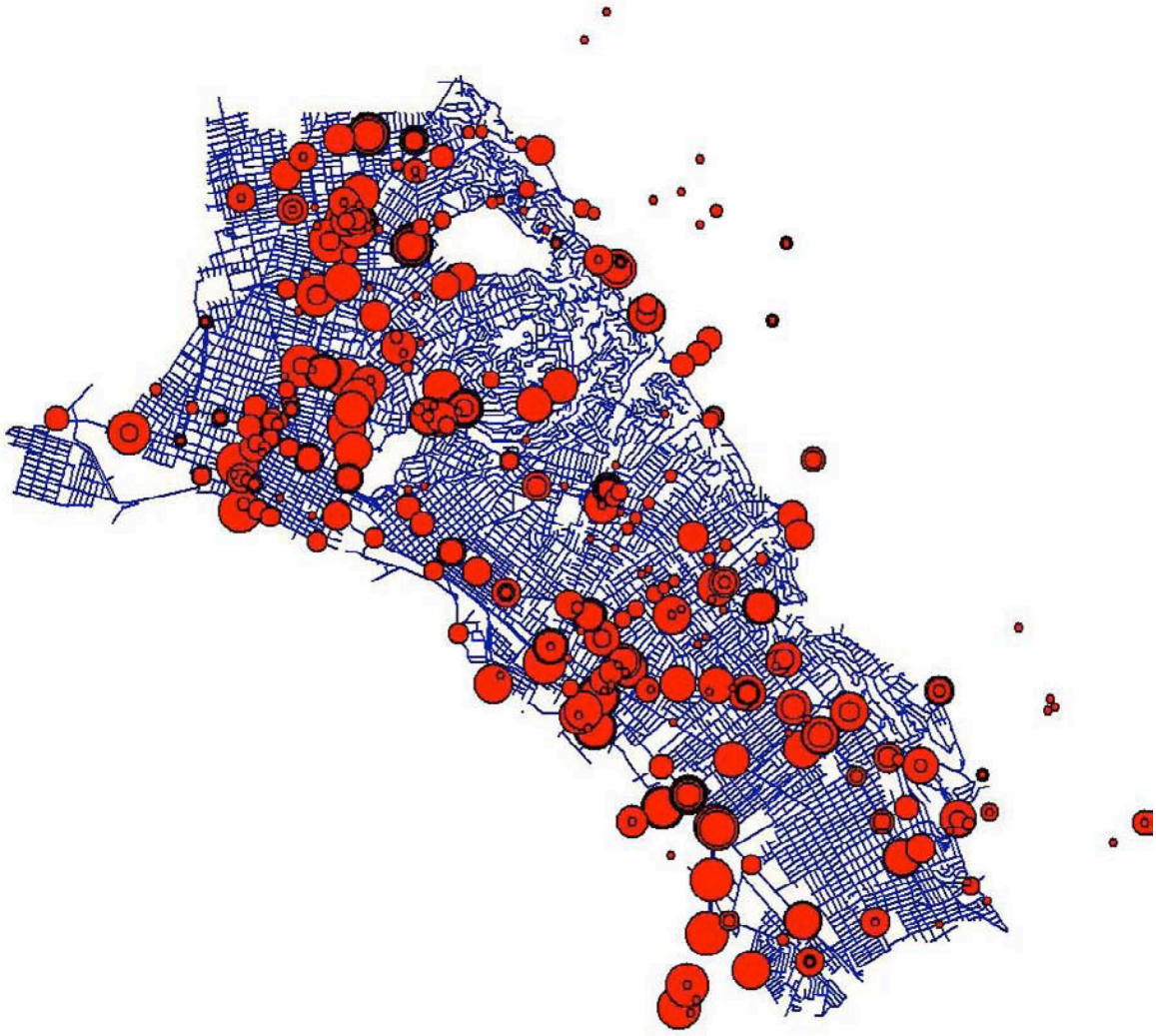

Figure 3. Average Annual Pedestrians by Number of Collisions at 247 intersections in Oakland, California (01/01/2000 to 12/31/2002)

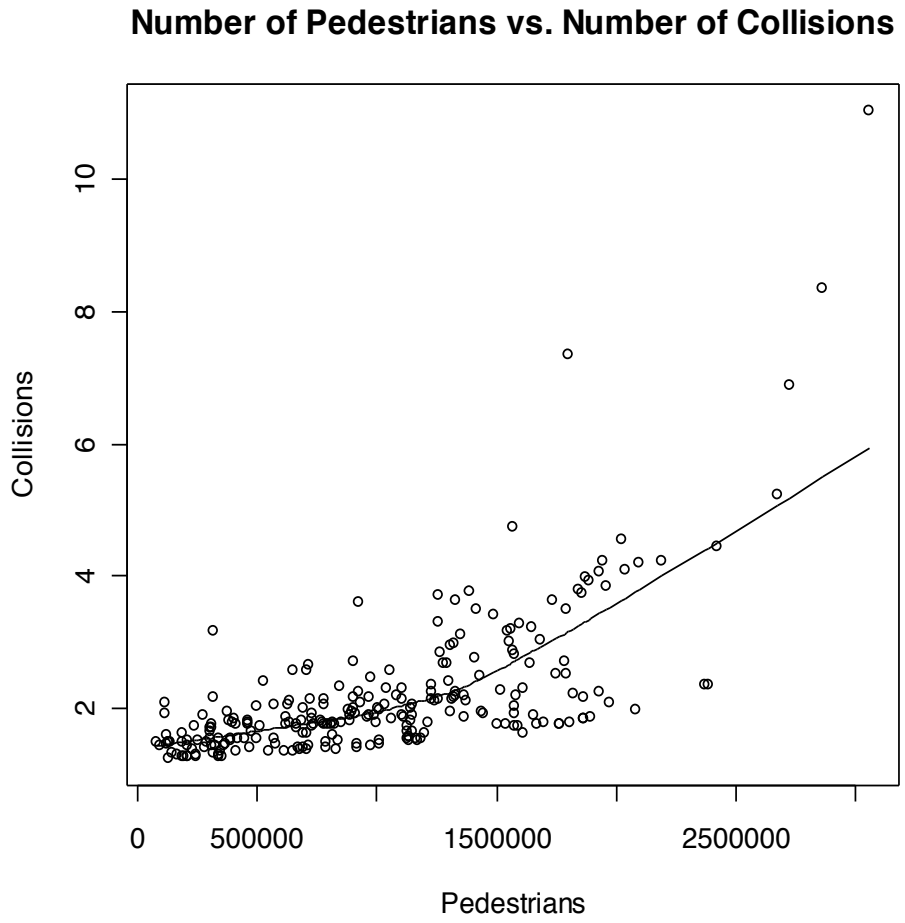

Figure 4. Rate of pedestrian-vehicle collisions per pedestrian by average annual number of pedestrians at 247 intersections in Oakland, California (01/01/2000 to 12/31/2002)

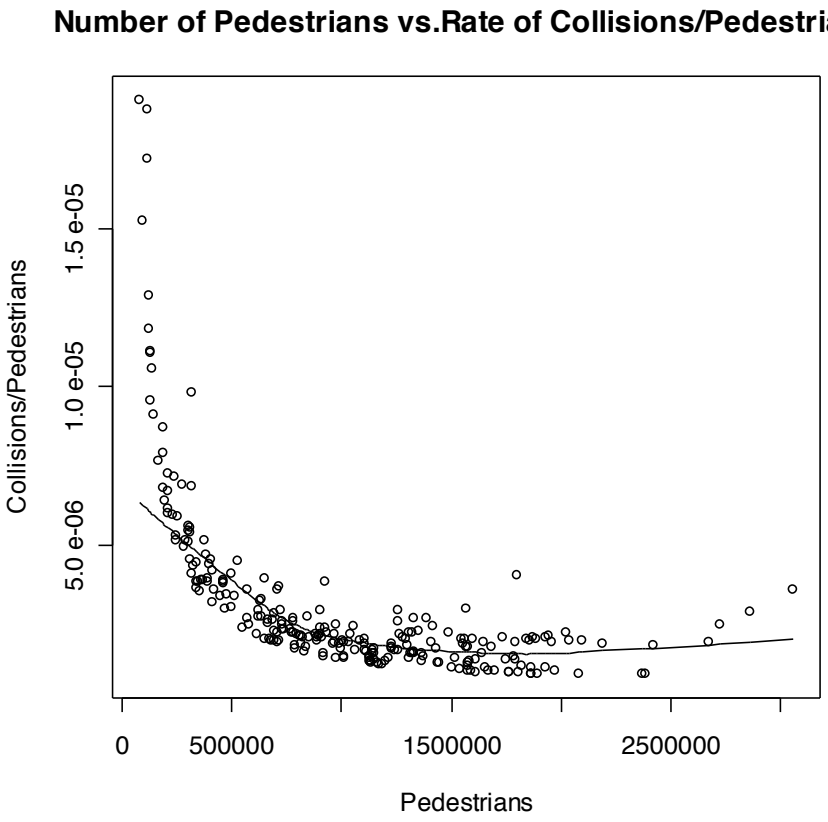

Figure 5. Rate of pedestrian-vehicle collisions per pedestrian by average annual number of vehicles at 247 intersections in Oakland, California (01/01/2000 to 12/31/2002)

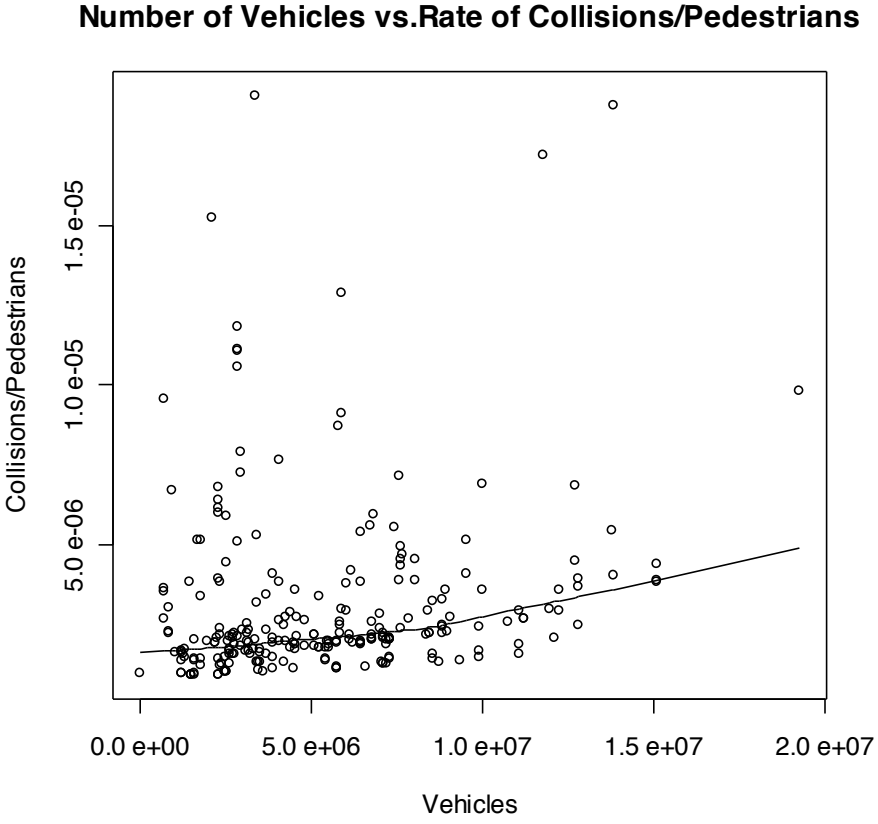

Figure 6. Rate of non-collisions (N=P-C) per pedestrian by average annual number of pedestrians at 247 intersections in Oakland, California (01/01/2000 to 12/31/2002)