University of Alberta

Child Pedestrian Injury in an Urban Environment: Risk in Space and Time

by

Nikoloas William Yiannakoulias ©

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of Master of Arts

Department of Earth and Atmospheric Sciences

Edmonton, Alberta

Fall 2000

0-612-59744-X

University of Alberta

Library Release Form

Name of Author: Nikolaos William Yiannakoulias

Title of Thesis: Child Pedestrian Injury in an Urban Environment: Risk in Space and Time

Degree: Master of Arts

Year this Degree Granted: 2000

N. Yiannakoulias

#207 10625 81 Avenue

Edmonton Alberta

T6E 1Y1

August 1/2000

University of Alberta

Faculty of Graduate Studies and Research

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled *Child Pedestrian Injury in an Urban Environment: Risk in Space and Time* here submitted by Nikolaos William Yiannakoulias in partial fulfillment of the requirements for the degree of Master of Arts.

Dr. Karen E. Smoyer

Dr. M. John Hodgson

Dr. Donald W. Spady

July 25, 2000

Abstract

Child pedestrian injury remains an important concern both nationally and internationally. Although this topic has been of considerable interest among injury epidemiologists in the last decade, child pedestrian injury remains largely unexplored by geographers. Using both geographical and epidemiological techniques in data preparation, data transformation and data analysis, I attempt to isolate the spatial dimensions of risk while also considering temporal patterns in injury frequency. Results suggest that features of the social and physical environmental together put children living in the inner-city at greater risk than children living in suburban areas. The observed core-periphery gradient emphasises the importance of place in understanding human well being. The results also suggest that alteration of the times that children go to school, particularly in the morning hours, may reduce the frequency of collisions.

Dedication

For our robot overlords

(01101001010001010101001010100111101010100101011
1001010111010011001010101001010001001111001001)

Acknowledgement

My first acknowledgement goes to members of my supervisory and examining committees. Thanks to Dr. Karen Smoyer for her support and willingness to take me on as a student when better judgement might have suggested she do otherwise. Thanks to Dr. John Hodgson for sharing his vision of geography and to Dr. Don Spady for answering and asking many, many questions. Important thanks should also go to Dr. Don Voaklander who made it possible for me to do this project in the first place.

This section includes the names of some of the people who supplied important help along the way: Dr. Gordon Flowerdew, Dr. Mike Jerrett, Dr. Karen Kelly, Dr. Brian Rowe, Yvonne Urban, Chuan Kua, Laura Smith, T. Mazur, Leslie Gadowski, Sally Hampton, Larry Svenson, Dr. Ed Jackson and Dr. George Sitwell. Final thanks go to the two people who shared an office with me over the last two years: Jared Hewko and Narmantha Thanigasalam. Thanks for setting a high-standard of academic excellence and humility, and thanks for listening to me talk nonsense when you had important deadlines to meet.

Table of Contents

List of Tables

List of Figures

1. Introduction

1.1 The Problem of Child Pedestrian Injury

Child pedestrian injury has been studied in considerable detail in the last few decades. Research in New Zealand, Australia, the United Kingdom, the United States and Canada has resulted in a substantive body of literature describing behavioural, environmental and social factors of risk. Pedestrian injury rates have been dropping in many countries as well; initiatives in Sweden have reduced child pedestrian injury mortality to near zero.

Despite this evidence, child pedestrian injury remains a leading cause of death in many industrialised countries, including Canada. In 1996, 183 children aged 0-14 years died in Canada as a result of motor vehicle-related collisions, 63 of which involved pedestrians (Statistics Canada, 1996). When compared with 1996 data on other major causes of death within this age group (208 deaths from all forms of cancer, 79 deaths associated with respiratory illness and 126 deaths related to diseases of the nervous system and sense organs), pedestrian injury fatalities are clearly a significant public concern (Statistics Canada, 1996).

Eliminating all automotive transport would reduce the pedestrian-automobile collision injury rate to zero. Although most people would consider this to be an unreasonable solution to the problem of pedestrian injury, it does illustrate the role that society as a whole can play in influencing health. Our society makes a tacit decision that a certain amount of injury and death from pedestrian-automobile collisions is an acceptable consequence of the convenience of automotive transport. This research will not discuss the ethical conflicts inherent in this statement; the argument merely suggests that like many issues in health, child pedestrian injury is intertwined with social decision making. Therefore, the study of pedestrian injury is important not only from a public health and epidemiologic standpoint, but also from cultural, ethical, political and economic perspectives.

This topic also has important implications for the broader issue of child health. Recent research in Europe has suggested that environmental and educational interventions to reduce risk of pedestrian injury have contributed little to modern reductions in injury incidence (Hillman et al., 1990). Instead, evidence suggests that these reductions are more a function of reduced outdoor play activity in children, and greater restraints placed on independent activity among children in general. Concern about traffic safety remains a chief parental concern (Hillman et al., 1990). Although little research has explored the relationship between children's outdoor activity and health in detail, it is possible that this reduction in activity is related to decreases in overall child health, perhaps most notably with regard to the increased prevalence of childhood obesity (Shimai et al., 1993; Muller et al., 1999).

Recent technological and methodological advances offer new opportunities for investigating the spatial variability of health and well-being in urban communities. In recent years, geographic information systems (GIS) have become easier to use and have dropped considerably in price[1]. Combined with the increased availability of data, researchers are more able to integrate environmental and population data into geographic studies of health outcomes. Important statistical problems associated with mapping rare health outcomes have been addressed in the medical geography literature in recent years. In the past, studying rare health outcomes in the urban setting was problematic because there was no simple method for dealing with statistical uncertainty in incidence rates.

This challenge has been the object of considerable study in recent years, and methods have been devised to address these problems.

Child pedestrian injury is well suited to the application of geographic theory and methodology specific to the urban environment. Research in child pedestrian injury has shown that characteristics of the urban environment are important determinants of risk. The design of these urban environments—from the design of neighbourhoods to the routing of traffic—is the occupation of urban planners, many of whom have backgrounds in geography. Yet epidemiological and geographical research on the topic of child pedestrian injury remain disparate. The addition of geographic methods and theory into the epidemiology of child pedestrian injury is an important step in developing a complete understanding of this health outcome. The exploration of spatial relationships may reveal previously unconsidered factors of risk, and may offer some explanation of poorly understood relationships. Prevention programs will also benefit from this interdisciplinary approach. Knowing about areas where risk of injury is unusually high may improve the effectiveness of educational and environmental intervention programs.

1.2. Objectives

The goal of this study is to present comprehensive and useful information on the geographical epidemiology of child pedestrian injury in an urban environment. This requires a methodological and theoretical approach that reveals the spatial and non-spatial characteristics of a) the outcome itself and b) the factors of risk.

Three key objectives are addressed in order to meet the study goal. First, the spatial, temporal and demographic distributions of child pedestrian injuries are described. Secondly, determinants of risk are analysed in a multivariate framework where the separate influences of various factors can be compared. Finally, the spatial distribution of the relevant risk factors is investigated, and compared to the spatial distribution of injuries.

The first objective is met by using recent data (1995-1999) on child pedestrian injury in Edmonton to describe the differences in risk between genders, as well as patterns in the time and date of injury. Various graphical representations and descriptive statistics are used to identify temporal and demographic trends. A geographic information system is used to present and analyse spatial relationships in these data. In addition, a map of city-wide injury rates is developed which accounts for the methodological complications associated with mapping rare data in small areas.

Objective two is met by using previous literature on child pedestrian injury to choose variables for critical study. This study supposes that previous epidemiological research in child pedestrian injury has successfully identified the main factors of risk, but asserts that a geographical approach may add important knowledge to the understanding of child pedestrian injury. Therefore, this study analyses factors of risk similar to those explored in previous research, but uses a methodological and theoretical approach ignored in most literature on child pedestrian injury. The study variables are analysed using multivariate statistical techniques in order to understand the relative role of risk factors in explaining the spatial variability of injury rates. This approach is important since there are many factors of risk, and the relative importance of these factors can be understood only after controlling for separate effects.

The final objective is met by using the information from the statistical model to develop a spatial representation of risk factors. This map of combined risk factors is then compared to the map of injury occurrence to aid in interpretation. It is also compared to a map of residual error from the multivariate model.

1.3. Rationale

The results of this research do not offer information about risk to individual children. Instead, this research offers a perspective on how risk varies among neighbourhoods and over space[2]. This distinction is important for several reasons:

i) A study of neighbourhoods in an urban environment provides unique information about risk in a spatial context and at a spatial scale that has been generally unexplored in pedestrian injury research. In this way, this study is exploratory, even in light of the breadth of epidemiological research.

ii) This research provides information to local public health officials interested in variability of risk over space. An accurate spatial representation of risk can be useful for the development of prevention strategies aimed at education or environmental modification.

iii) The techniques used should contribute to the methodological literature in small-area research. As an analysis of rare events in a small area, this study of pedestrian injuries will face the many problems associated with analysing rare health outcomes. This study will explore the options for mitigating these problems, and if successful, demonstrate a method appropriate for extracting meaningful and representative results.

Notes

[1] Although base packages from major software vendors (ArcView and MapInfo) cost thousands of dollars, several new commercial packages offer full functioning geographic information systems under $200 (eg. Manifold, Mapsmith). Free geographic information systems have been developed as well. GRASS GIS can be downloaded for free from www.geog.uni-hannover.de/grass/welcome.html.

[2] Occasionally the term "neighbourhood" will be used in place of the term "census tract" which is the more precise term for the unit of study used in this research project. The neighbourhood subdivision as defined by the municipality of Edmonton is similar in size to the census tract as defined by Statistics Canada. As will be discussed in later sections, the census tract is assumed to be a region of geographic relevance to children. For these reasons, the term neighbourhood was deemed appropriate for periodic use.

2. Theoretical Overview

This section outlines and discusses the key methodological and theoretical issues related to this study. It includes a discussion of issues in geography and health research and an overview of pedestrian injury literature.

2.1 Issues in Geographical Epidemiology

Although the spectrum of academic pursuits under the heading of geography make simple definitions of the discipline difficult, in the strictest terms, geographic research concerns itself with places. Geographic data that describe places or how characteristics vary between places are an important resource in most quantitative geographic research. Researchers transform these data into useful information about causal relationships, spatial patterns, and changes over time.

The potential of geographical information to aid health researchers goes back as far as the origins of epidemiology. John Snow's research on cholera is a landmark study in epidemiology. Geographical recreations of Snow's study indicate a very strong spatial pattern in his data; the mean centre of cholera deaths is located very close to the notorious Broad Street pump (Meade, 1985). John Snow did not have the tools to observe detailed spatial relationships between exposure to this contaminated water pump and death from cholera, but it is clear that such knowledge could have aided his study, particularly if he were restricted to a smaller sampling frame. Geographical methodology also played an important role in understanding the causes of rickets and skin cancer (English, 1992).

Research on Lyme disease (Glass et al., 1995), exposure to radon (Kohli et al., 1997), groundwater vulnerabilities to pesticides (Teso et al., 1996) and exposure to high-voltage transmission lines (Wartenberg et al., 1993) are more recent examples of research that has used geographic data and approaches to understand health. In most of these examples, geographic information systems were important at every stage of study: data obtained from geographic information systems were used for the development of study hypotheses and geographic information systems were also necessary for data analysis and presentation of the results.

As interest in geographic information and spatial data analysis grows among health researchers, it is important to acknowledge the difficulties associated with such research. Arguably, the two most important topics relevant to modern geographical epidemiology involve the problem of analysing geographically aggregated data and the role of contextual effects on health. Any researcher using geographic data to understand the epidemiology of a health outcome must be familiar with these issues and how they affect the meaningfulness and interpretation of their research.

2.1.1 Ecological-Level Research

Analytical *ecological-level* studies—those studies that analyse relationships between geographically aggregated characteristics—have been the target of criticism inside and outside of health research since early examination of the practice in the Fifties (Susser, 1994). These are distinguished from *individual-level* studies that test associations between the characteristics of individuals themselves. The correlation between average income and cancer incidence in ten neighbourhoods is an ecological-level relationship. The correlation between income and a dichotomous variable (cancer vs. no-cancer) among ten individuals is an individual-level relationship. The term "ecological fallacy" has been used to argue that correlations (usually of means) based on aggregated data should not be used to make inferences about relationships at the level of the individual (Selvin, 1958). Research by Robinson (1950) showed how correlation between race and literacy varies considerably depending on whether the statistics are based on individual-level or aggregated data. In one case, a 0.773 correlation between race and literacy using aggregated data fell to 0.203 when based on individuals (Robinson 1950). Such systematic change in magnitude of correlation is explained by the fact that aggregated data suppress the variability inherent in individual-level data. Thus correlations based on such aggregations will generally show stronger associations. When the intent is to measure the relationship between individual-level factors, like race and literacy, ecological-level associations can provide misleading information.

Frequently researchers are forced to use ecological-level associations to understand individual-level relationships. In most jurisdictions, cost and confidentiality concerns prevent researchers from obtaining detailed information about individuals. The classical epidemiological approach typically considers studies based on ecological-level data as no more than inexpensive precursors to more detailed individual-level studies. This idea is based on the positivist scientific tradition which established an early standard for reasoning in most social sciences where cause-effect inferences are often harder to make compared to the natural sciences (Little, 1998). Epidemiologists have used this philosophical tradition not only as the basis for formulating hypotheses, but in assessing the legitimacy of hypothesised causal relationships. In their introductory epidemiology text, Henneckens et al. (1987) claim that there are three "positive" criteria fundamental to assessing the validity of a cause-effect relationship: strength of association, biological credibility and consistency of findings in other research. Put simply, if i) the magnitude of an association between the cause and effect is strong, ii) there is a reasonable physiological connection between the cause and effect, and iii) there is agreement among other studies of this relationship, then causality can be inferred.

Interpreting meaningful associations between ecological-level risk factors—like neighbourhood income—and broad, inclusive health outcomes—like neighbourhood mortality rate—is especially difficult given the requirements of cause-effect inference as it is defined in the epidemiological tradition. Mortality includes all outcomes of death, and therefore inferring clear causal links between this outcome and ecological-level risk factors is impossible. It is not feasible to study any biological relationship between neighbourhood deprivation and mortality because neither variable provides a context for clinical research, and therefore the second criteria can never be met. Even ecological-level studies that compare the effect of specific variables on specific health outcomes cannot satisfy these criteria. From this reasoning, it has been argued that although ecological-level studies can show statistical associations between possible determinants and outcomes, they cannot provide sufficient insight into these associations to infer causal relationships. This explains the view that ecological-level studies are meant for "quick and dirty" analysis preceding more detailed individual-level studies (Hennickens et al., 1987).

Bias is also a potential problem in ecological-level studies, particularly when the health outcome under study is non-specific. Ecological-level bias occurs when the effect of a risk factor differs among areas under study (effect modification) or when unconsidered risk factors are differentially distributed between areas (English, 1992). The former is less of a problem in small-area ecological studies (like the comparison of neighbourhoods) where geographic cultural variability is usually small. Multivariate analysis can also mitigate these problems when data are available, study areas are small and the outcome under study is specific. However, when the mathematical and theoretical problems relating to ecological-level associations are considered together, research based on ecological-level studies is a common target of criticism.

2.1.2 Contextual Effects on Health

Despite these arguments, the ecological-level study remains a frequently used method of analysis in the social sciences, particularly in sociology and geography. This is partly an issue of data availability; however, some proponents argue that this approach to research is a very important component of studies that investigate the role of the physical or social environment on health. Perhaps the most important function of ecological-level approaches is that they are effective tools for studying *contextual effects*—how characteristics of groups (geographic or otherwise) affect individuals (Susser, 1994). For example, the relationship between neighbourhood income and neighbourhood mortality rate is contextual if, after controlling for individual-level income, the income of the neighbourhood in which a person lives influences mortality risk. Associations between group characteristics and a health outcome require one to analyse characteristics of these groups specifically in addition to observing the relationships at the level of the individual. Indeed, the key criticism of the ecological approach—that ecological associations should be cautiously applied to individuals (the ecological fallacy)–also applies to studies of individuals–individual-level associations should be cautiously applied to groups (Susser, 1994). Thus, the "fallacies" are bi-directional- the latter of which has been termed the "atomistic fallacy". Using Robinson's analysis, when the inter-state relationships between literacy and race are of interest, the 0.773 correlation (based on aggregate data) is an appropriate measure, while the 0.203 is not since the study resolution is the individual.

Compositional effects represent the more traditional interpretation of ecological-level relationships (Shouls et al., 1996). Associations between ecological-level data are compositional when they measure individual-level relationships. The association between a community's smoking rate and the prevalence of emphysema is probably a compositional representation of an individual-level relationship. All other things being equal, a non-smoker's risk of emphysema is not dependent on whether or not he or she lives in a community in which the smoking-rate is high or low (ignoring the possible effects of second-hand smoke). Where there are many smokers, most of which have a higher-than-average risk of emphysema, we would expect more cases of emphysema. Thus it is the differential distribution of smokers that accounts for our ecological associations. It may be that the smoking habits of a community of people do exact an effect on the health of all individuals living there, regardless of personal smoking status. However, there are dominant associations based on the individual-level association between smoking status and health that make it difficult to understand possible contextual effects. Effectively separating these effects is both the chief purpose and greatest challenge of the contextual researcher.

Methods in the study of contextual effects (independent of compositional effects) continue to evolve, although there is a growing focus on multilevel modelling [1]. Shouls et al. (1996) used such techniques to investigate self-reported health status and Diez-Roux et al. (1999) to describe the role of the neighbourhood in influencing dietary patterns to support the hypothesis that contextual effects have a role on health.

However, these results are contrary to those found by Ecob (1996) who used similar techniques to show that area of residence had no association with most self-reported health measures. More recent research using multilevel statistical techniques has shown similar results. Contextual effects explained a very small proportion of spatial variation in health measures in a study using the Ontario Health Survey (Boyle and Willms, 1999).

The theoretical justification for research on contextual effects has been explored in detail in the past few decades (Iverson, 1991). The focus on the effects that group characteristics have on health was motivated by the realisation that physiology by itself does not explain all aspects of regional or temporal variation in health outcomes (Marmot, 1998). Rose (1991) acknowledges the role of the sociologist Emile Durkheim in establishing the idea of how societies themselves influence the health of people. Durkheim viewed suicide rates as functions of "collective characteristics" and that the very nature of societies influenced the occurrence of the phenomena in populations (Rose, 1991). Marmot provides a revealing example to illustrate this point. In a given modern industrialised region of the world, the rate of death from coronary heart disease is relatively constant over time. Even with improvements in health-care and changes in the population demographic, the rate remains relatively constant nationally (though perhaps declining slightly), but vary considerably internationally, even between countries with similar standards of health care. Marmot continues:

> Those of us trained in medicine, or in social psychology, tend to start with the characteristics of individuals—were the individuals who died smokers or did they have high cholesterol? But the individuals who died this year will not contribute to next year's death rate. Thus, the characteristics of societies, over and above the characteristics of individuals determine the death rate [...]. (Marmot, 1998).

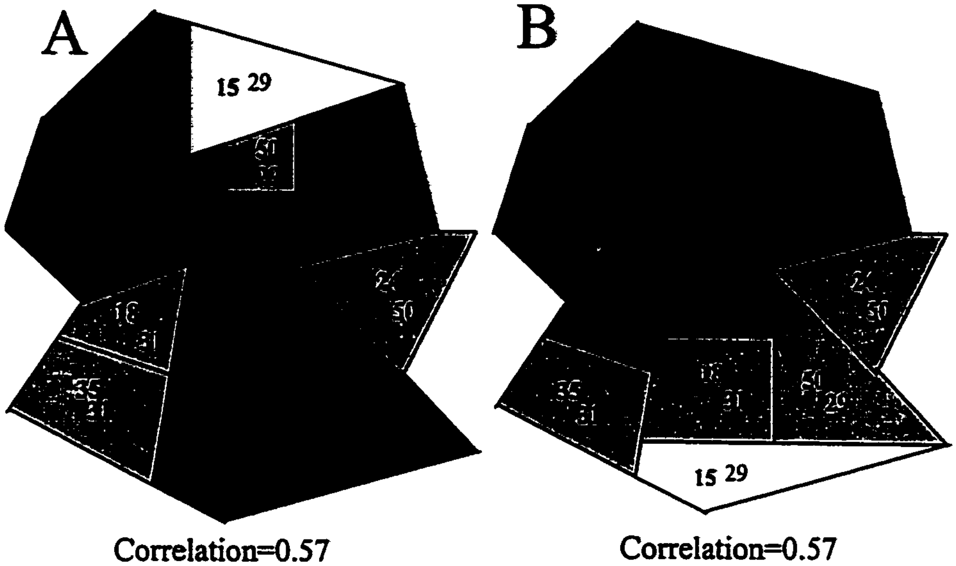
The fact that the rate of death from coronary heart disease is more or less constant over time in the U.K. suggests that something about the population and the places in which the population live play a role in shaping the rate of death from coronary heart disease. As a result, a focus on populations, and the places in which populations live provides specific insights into patterns of illness not typically visible in individual-level research.

2.2 Applying Geographic Knowledge in Health Research

2.2.1 Analytical applications

Spatial pattern detection is an important component of any urban health research. Geographical data analysis has shown that industrial centres possess higher rates of cancer than less industrial areas (Hoover, 1979). Environmental health researchers have also shown considerable interest in spatial data analysis. Geographic information systems have been used to study spatial relationships between air toxins and cancer (Gatrell and Dunn, 1995) and childhood leukaemia and nuclear power plants (Gardner, 1992). Patterns resulting from such analyses, though often not unexpected, can indicate that risk factors are unevenly distributed among the population.

Aggregate statistics can show a spatial or non-spatial pattern (Figure 2.1). These maps present different information. Map A indicates an association between pairs of moderately correlated variables, but the variable pairs are distributed randomly over space. Map B indicates the same degree of association, however, pairs of larger magnitude are located nearer to the centre of the map and pairs of smaller magnitude are located nearer to the periphery. This is a spatial relationship that requires a map, or at the very least a variable indicating proximity from the centre of the region, to impart the researcher with complete knowledge of the processes at work. On map A, the geographic



Dark areas indicate variable pairs with larger magnitudes representation of the relationship may be convenient for observing the spatial variation in the phenomena; however, this map does not impart the researcher with new analytical knowledge (except for the fact that there is no spatial pattern).

2.2.2 Geographic Knowledge and Prevention

In addition to helping researchers understand contextual influences and spatial patterns of disease, the geographic theoretical framework also directly affects methods of illness prevention. Obviously, clinical research and individual-level studies have done much to improve human health in the past century. However, intervention and prevention strategies that focus on populations and their neighbourhoods can also contribute significantly to improvements in public health. The social, cultural and physical environments exact a "nested" effect on the health of individuals: features at the micro, meso and macro environmental scale interact to influence the context of a person's life (Andrews, 1985). Poverty, for example, is a social, political and cultural (ie. macro-scale) process which affects the health of populations, but is often difficult, if not impossible, to address at the level of individuals. Widespread increases in individual wealth cannot occur without significant (if not revolutionary) changes in social and political structure.

It is beyond the scope of this research to discuss the possibilities of major political, and or social change as a strategy to improve health. However, the modification of social or physical environments in which people live does not have to involve major political or social upheaval. First, identifying characteristics of places that influence health provides a focal point for environmental modification. If environmental factor X increases the likelihood of undesirable outcome Y—independent of the immediate physiological causes of outcome Y—then X is a possible focus of intervention. Second, the process of "ecological screening"—the identification of the geographic setting in which risks are present—optimises the efficiency of prevention and promotion programs (Andrews, 1985). Although combating childhood poverty may be beyond the abilities of public health officials, poor communities can still be identified, and interventions can be focussed on these communities.

Causal associations based on geographic knowledge—whether the study design is based on points, small-areas or nations—should be made cautiously. Nevertheless, geographic data analysis is an important method of study in much of the research in medical geography and medical sociology and continues to be important in epidemiological research. It is an analytical tool useful for researchers interested in understanding spatial relationships between points or areas as well as researchers interested in the role of contextual effects.

2.3 Theory and Epidemiology of Child Pedestrian Injury

Until the mid-Eighties, most studies of injury were descriptive with limited analytical exploration of influences or causes of injuries (Rivara, 1990). Growing interest in the analysis of injury can be credited to changing perspectives on injury theory. In recent years, injury research has developed a conceptual framework that is similar to health outcomes more traditionally studied in epidemiology, like the agent-host-environment relationship observed in the study of communicable diseases (Guyer and Gallagher, 1985). In the case of burn injury, for example, the agent is the heat energy that damages the victim (host) and the environment is the physical and psychosocial setting in which the burn occurs—like the "child unfriendly" old kitchen-stove (physical) in a single-family household (psychosocial) (Guyer and Gallagher, 1985).

William Haddon was one of the first to discuss the need for developing research approaches and prevention strategies that treat injuries like other health outcomes (Haddon et al., 1964, Haddon, 1975). Haddon's argument is that knowledge about injury has advanced to the point where causal associations can be explored and identified. Like other health outcomes, all injuries have a unique cause—exposure to mechanical, chemical or thermal energy in a quantity or intensity that exceeds the ability of tissue to resist damage or destruction (Haddon, 1980). Specific forms of injury are usually subdivided into groups according to environmental settings or demographics of interest, but the underlying relationship remains of conceptual importance. Injuries have a causal structure which justifies the use of traditional prevention tools relating to the host-agent-environment model. Hosts can be separated from the causes of injury by education or behavioural modification, elimination of the sources of harm (the agents) and modification of the environments in which the injury agents and hosts interact.

The recognition that the causes of injury have the same structure as other health outcomes has justified more analytically oriented methods using the traditional epidemiological tools. In the last decade, numerous case-control and cohort studies have been conducted which have investigated the relationships between various factors and risk of injury in children (Larson and Pless, 1988; Pless et al., 1989; Mueller et al., 1990; Pless et al., 1995; Roberts et al., 1995; Narayan et al., 1997). International consistency in results between these and other studies suggests that such methodological approaches to injury research have the potential to offer useful insights into understanding the determinants of injuries.

In recent years there has been a growing interest in viewing the environment as a key component of injury prevention. This has been an important dimension of injury control in the last few decades because environmental modification does not require repeated deliberate action on the part of individuals. The continuum between *active* intervention strategies, which require persons to behave a certain way repeatedly (like putting on a seat-belts) and *passive* strategies, which are "put in place" without requiring the repeated action on the part of individuals (like air bags), is represented by a broad range of strategies in injury prevention (Wilson and Baker, 1987). In recent years, interest in passive strategies (which include environmental modification) has increased, most likely as a result of the varied success of educational prevention strategies which focus on modifying behaviour (Roberts, 1993a).

Risk of injury varies with the age, gender, socio-economic status and other characteristics of individuals. For example, intentional injuries (like homicide and suicide) are more likely to occur among young adult males (Powell and Tanz, 1999), whereas hip fractures are more likely to occur among elderly women (Rozycki and Mauli, 1991; Chan et al, 1997). Socio-economic status has also shown an important association with injury risk in many studies. Low income has been shown to be a risk factor for burn injuries (Warda L. et al., 1999), bicycle injuries in young children (Carlin et al., 1995) and several types of adolescent injury (Williams et al., 1996). Although high income has been shown to be a factor of risk for various sport-related injuries (ex. Shugerman et al., 1992), this is probably a function of exposure (children from wealthier families are more likely to be exposed to certain activities: like skiing and horse riding), and is overshadowed by the pattern of increased risk of injury in general to children of lower income (Dougherty et al., 1990; Rivara, 1990; Durkin et al., 1994; Sparks et al., 1994).

There are two common methods used in the study of child pedestrian injury: case-control/retrospective cohort studies (individual-level) and geographic correlation studies (ecological-level). Findings have varied over time; the general trend has been a change in focus from behavioural and familial influences to the increased interest in environmental and economic influences. An early example of the former can be found in Backett and Johnston (1964) who conducted a case-control study in which family characteristics, family income, and playground availability were studied in a cohort of children aged 5 to 14. The authors found that familial characteristics—in particular, health and occupation status of the mother—were strong determinants of pedestrian injury risk (Backett and Johnston, 1964). Child behaviour continues to be studied in injury research (Pless et al., 1995), often in combination with assessments of family character and stability (eg., Husband and Hinton, 1972; Matheny, 1987; Horowitz et al., 1988).

More recent studies have used similar methods to arrive at different conclusions, suggesting that behavioural factors may have a small effect on injury risk independent of socio-economic status (Rivara, 1990). In the past decade, research in Canada, the United States, New Zealand and Australia has shown significant associations between risk of pedestrian injury and socio-economic indicators. After controlling for parental status and ethnicity, family income has been shown to be an important determinant of risk in New

Zealand children (Roberts, 1994). Similar approaches have been used in other studies in Australia and New Zealand, which have offered perhaps the most complete studies on child pedestrian injury in the Nineties. Results have shown, fairly consistently, a combined role between socio-economic and environmental factors in predicting injury. For example, in an analysis of socio-economic status, availability of safe walking areas, traffic volume and traffic speed, Stevenson et al. (1995) found relationships between both traffic volume and low socio-economic status and increased risk of pedestrian injury. Pless et al. (1989), Dougherty et al. (1990) and Joly et al. (1991) showed that the relationship between increased risk of injury and low socio-economic status persists in Canada, while Rivara and Barber (1985) have found similar relationships in the United States.

In an attempt to explain the relationship between income and risk of pedestrian injury, some researchers have explored exposure to road environments. These researchers have asked whether variation in the number of roads children cross in their daily activity explains variation in injury risk, and in particular, the variation in risk between children of different socio-economic groups. Leading research in Australia and New Zealand shows the relationship between increased road exposure and increased risk of injury (Stevenson et al., 1996). Recent research in Canada supports these results (Macpherson et al., 1998). In an attempt to explain these results and the relationship between increased injury risk and low income, Roberts et al. (1996) found family car ownership to be a strong predictor of road exposure. This research supports the ideas of Hillman et al. (1990) who argue that drops in injury mortality over the last 30 years are more a function of changing frequency of pedestrian activity than improvements in education or environmental modification.

Whatever role individual behaviour and or socio-economic status may play, the fact that passive preventative strategies show considerably more promise in reducing injuries has pushed forward studies of environmental risk. Of the many environmental risk factors that have been investigated, the volume and density of traffic show the most consistent relationship of increased risk. Using a logistic regression model, Stevenson (1997) showed an independent effect of traffic volume (OR=1.12)[2]. Roberts et al. (1995) showed a strong relationship between traffic volume and injury risk, particularly in neighbourhoods where the number of vehicles per hour exceeded 750 (OR=13.00). Mueller et al., (1990) showed an increased risk at collision sites where traffic volume was high (OR=3.1). Other factors of environmental risk have been identified, including: unavailability of play-space (Mueller), presence of curb parking (Roberts; Agran et al., 1996), traffic speed (Mueller; Agran), and percentage of excessive speed traffic (Stevenson et al., 1995).

Although these studies have provided important insights into the relationships between child pedestrian injury and the urban environment, family income and road exposure, few offer insight into the geography or spatial distribution of risk. The only way to develop a geographical understanding of risk—for example to identify what neighbourhoods possess the greatest collection of risk factors associated with child pedestrian injury—is to test these associations at the ecological-level. Several recent studies of child pedestrian injury have investigated geographical aspects of risk. Dougherty et al. (1990) investigated the relationship between income and variability in injury mortality in Montreal and urban areas throughout Canada. The study found that death rates were consistently highest in census tracts with median family income in the lowest income quintile (Dougherty et al. 1990). A study by Joly et al. (1991) showed similar results. These studies suggest that low income status of a geographic area is a risk factor, and that income may not have a linear association with increased risk of child pedestrian injury. This conclusion is supported in several case-control studies (Stevenson et al., 1995; Roberts et al., 1995).

The failing of these and most ecological-level studies of pedestrian injury is that despite the indication from individual-level studies that environmental factors are important determinants of risk, these studies do not combine their investigations of socio-economic risk factors with assessments of the physical environment. This is not surprising since most of these studies come from the epidemiological literature, which typically regard such ecological-level studies as precursors to more rigorous individual-level studies. However, since such comprehensive studies of both social and environmental risk factors have not been conducted, the spatial properties of injury risk (as represented by the independent effects of the risk factors) remain unknown.

Wang (1994) conducted an important study of the association between environment and pedestrian injury in a comprehensive urban setting. This research used collisions resulting in pedestrian injury between 1982 and 1990 in the city of Edmonton. Results showed that in planned environments where pedestrian and automobile traffic are separated from each other the frequency of collision was reduced. However, Wang did not account for the effects of socio-economic confounding. Statistics from the Edmonton Household Travel Survey show that individuals from wealthier families make a smaller proportion of pedestrian trips than individuals from less wealthy families (City of Edmonton Transportation Department, 1995). Wealthier families tend to live in suburban neighbourhoods, which are generally planned. Without controlling for income variations among neighbourhoods, it is difficult to conclude that it is the physical structure and not the socio-economic composition of these neighbourhoods that influences the spatial variability of injury rates. Thus, although Wang's research studies environmental risk factors—something that is lacking in the ecological-level literature in epidemiology—it ignores the role of the social environment.

A further criticism of ecological-level pedestrian injury studies is that although they employ geographic data, they neglect the study of spatial relationships. Their analyses do not include the investigation of areal patterning of injuries or risk factors, for example. Complete understanding of the epidemiology of child pedestrian injury requires an investigation of such patterns. This may also aid the development of educational prevention programs or environmental modifications.

2.4 Conclusions

Research in health evolves with changes in the frequency of health outcomes, medical and computer technology, theoretical considerations and statistical techniques. Research in child pedestrian injury prevention has evolved from a focus on individual behaviour and family characteristics to a focus on the role of the physical and social environment. This is reflected in a change in emphasis of prevention strategies—a move from active to passive approaches. Understanding the role of the environment in injury risk requires an understanding of the physical and social environment, as well as techniques of analysis that effectively assess the associations between these factors.

Methods and theory in health and place, medical geography and geographic information systems provide an important context for the analysis and description of child pedestrian injury.

Although a few studies have given exploratory attention to the spatial dimensions of pedestrian injury (eg. Braddock et al., 1994; Raybould and Walsh, 1995) there has yet to be any comprehensive analysis in this area. In part this is due to a lack of detailed data on injury cases, methodological problems relating to rare data, and a lack of tools to integrate various data sources. Access to rich databases, the evolution of methods to resolve methodological issues and the use of a geographic information system to integrate data sources makes our study an important contributor to this area of research in child pedestrian injury.

Ecological-level analyses are often used as a fast, low-cost tools for preliminary analysis. However, ecological-level associations can provide information about contextual effects as well as describe how risk varies over space. Ecological analysis and child pedestrian injury have a natural conceptual partnership that has slowly developed out of growing concern about environment risk. Ecological methods of analysis present a meaningful picture of injury risk as a function of many factors, and not simply random events occurring to individuals. The ecological picture puts pedestrian injury into a context in which spatial processes–often undetectable at the individual level–can be observed.

Although ecological-level studies have a tradition in child pedestrian injury, previous research has lacked the detailed analytical exploration necessary to contribute significantly to understanding the geography of child pedestrian injuries. The primary failing is that none of these ecological-level studies have successfully shown independent associations of important risk factors. Furthermore, geographic research in this area has produced little information useful for injury prevention. Part of the problem has been the difficulty in producing reliable maps of risk—in the form of rates or the representation of risk factors themselves. Recently developed techniques in spatial statistics and geographic information systems allow researchers to explore fully these geographical issues in pedestrian injury. The re-exploration of this well-studied area of health research using new techniques and a new theoretical approach can strengthen our understanding of accepted risk factors and possibly offer new insights into patterns of risk.

Notes

[1] Multilevel modelling is a relatively new technique which is motivated by an interest in studying hierarchical relationships. Children in a given family possess individual characteristics as well as characteristics common to other family members. Both of these factors differentiate them from children in other families. In this example, multilevel modelling allows researchers to systematically differentiate individual differences from the familial differences.

[2] The odds ratio (OR) is a commonly used test of magnitude in epidemiology. Odds ratios above 1.00 suggest reflect increasing likelihood of an outcome occurring given an exposure. Odds ratios below 1.00 suggest a decreasing likelihood of an outcome occurring given an exposure.

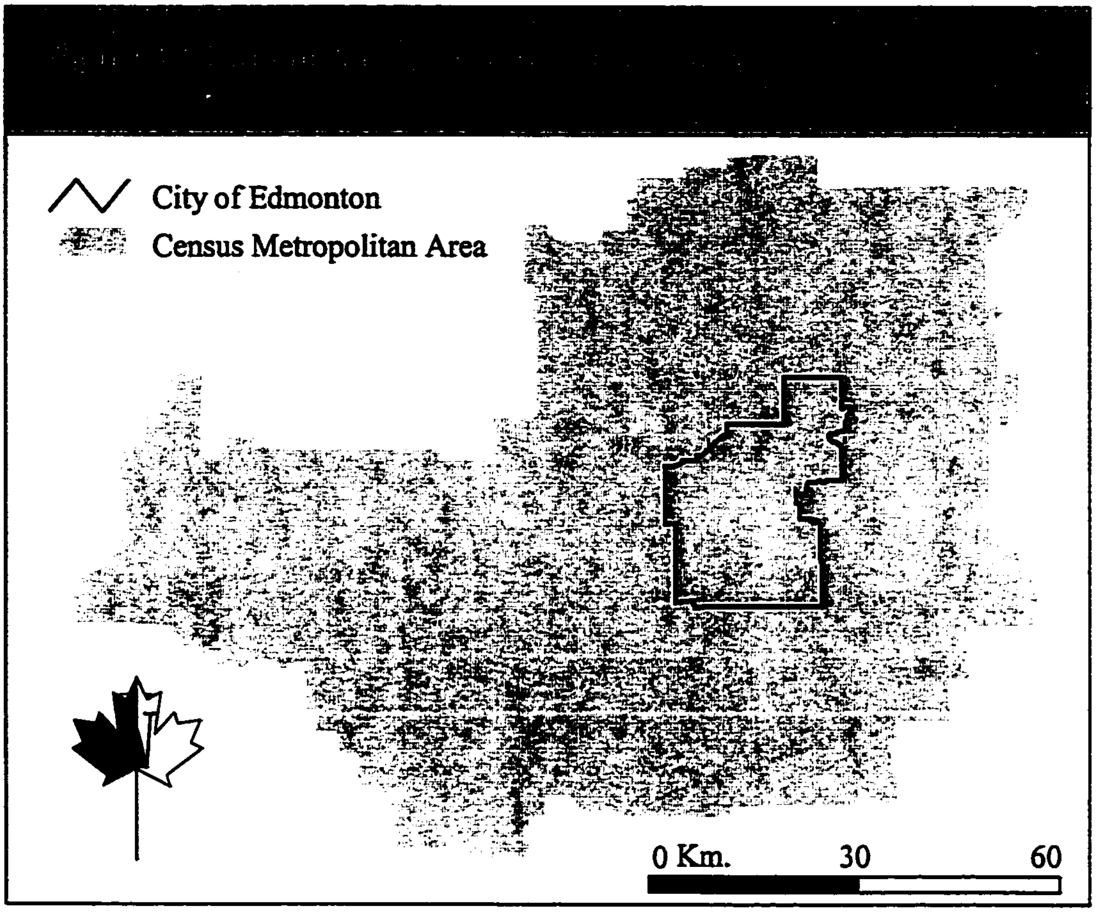
3. Methods

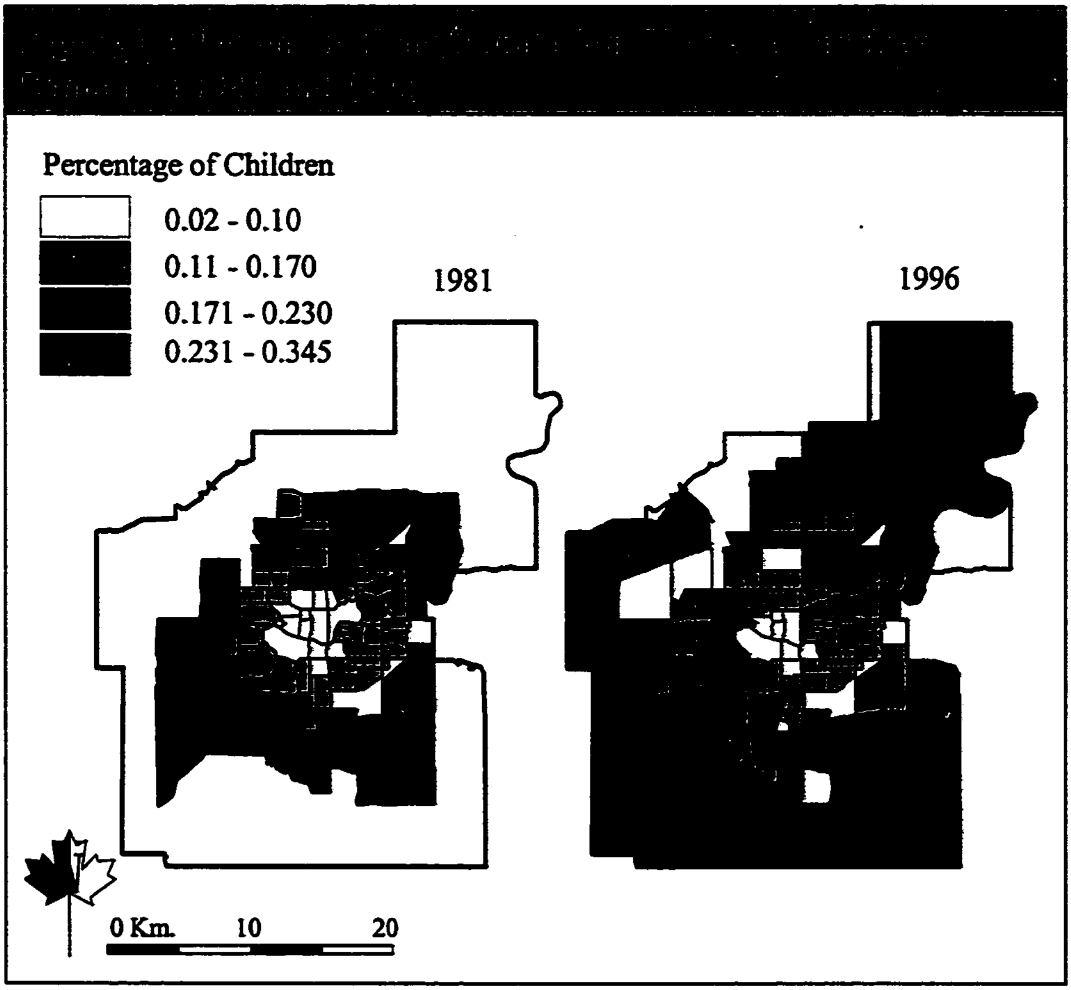
This chapter serves two purposes. It describes the data and techniques of data analysis used in this project, and it outlines and discusses issues of study design and techniques of analysis.

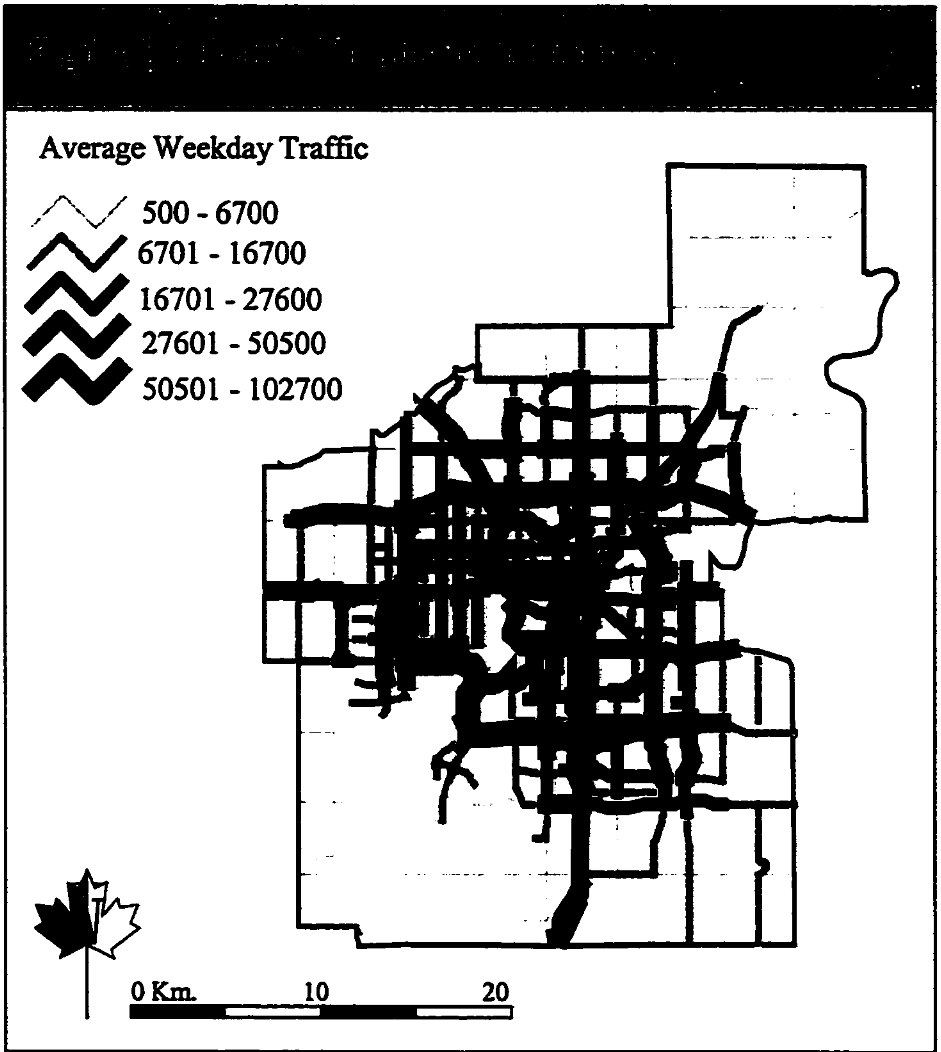
3.1. Study Area

The population of the Edmonton census metropolitan area has grown over 16% in the last 15 years—from 740 882 in 1981 to 862 597 in 1996 (Figure 3.1). Much of this growth in population has occurred in neighbourhoods along the city's perimeter. Changes in the population of children (0-14) are similar—from 102 880 in 1981 to 126 365 in 1996[1]. The highest proportions of children (0-14) can be found in the outlying areas of the city, a trend which persists over the last 20 years (Figure 3.2). Slight variation in this demographic can be found in select regions of the city, most apparently, a slight increase in the proportion of children in areas of Central and Central-West Edmonton, and small decreases in parts of Southwest Edmonton. Generally speaking, however, the spatial distribution of the children is relatively stable.

Traffic flow in Edmonton is shown on Figure 3.3. Notable traffic features are the main freeways: Calgary Trail North and South, the Yellowhead Trail, the Whitemud Freeway, St. Albert Trail and the main high traffic roadways like Stony Plain Road, Jasper Avenue, 97 Street and 170 Street. These arterials represent the main routes used by automotive commuters during weekday working hours. A more general assessment of the pattern of traffic flow identifies increasing traffic densities closer to the downtown core, and the Northwest industrial district, with lower densities in residential areas along the city's perimeter. A 1994 survey of all-purpose trips made by drivers also shows a


City of Edmonton
Census Metropolitan Area
0 Km.
30
60


Percentage of Children
0.02 - 0.10
0.11 - 0.170
0.171 - 0.230
0.231 - 0.345
1981
1996
0 Km.
10
20



Based on data obtained from 1996 survey of average weekday traffic (City of Edmonton Transportation Department, 1996).

strong core-periphery traffic movement pattern (City of Edmonton Transportation and Planning, 1995).

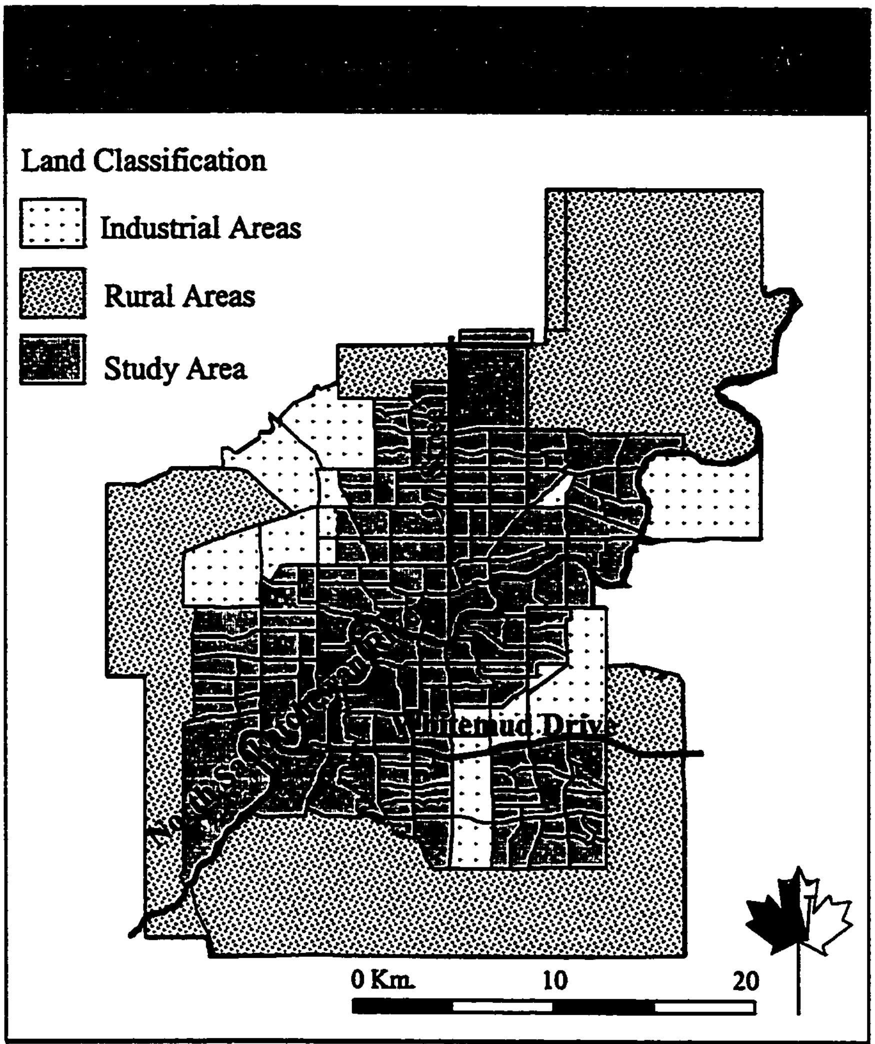
The physical geography of the city is dominated by the North Saskatchewan River Valley and to a lesser extent, its two main tributary valley systems: the Mill Creek and Whitemud Creek ravines. In addition to being physical barriers to pedestrian travel, their presence exhibits a relationship with the spatial distribution of various social and demographic indicators throughout the city. The 1996 average dwelling value in enumeration areas adjacent to the North Saskatchewan River Valley ($174 000) is larger than the average dwelling value of all other enumeration areas ($116 000) (Statistics Canada, 1996b). Average individual income shows a similar pattern—$36 000 for areas adjacent to the River Valley, $23 000 for all other areas.

The Edmonton transit system consists of 120 bus routes throughout the city, with a ten-station single light rail transit (LRT) route extending from the city's Northeast end to the University of Alberta (City of Edmonton, 1996). In addition, there is a small fleet of community buses that offer access to the main bus routes in low traffic residential neighbourhoods (City of Edmonton, 1996). Despite the availability of the transit service, its percentage of total daily person trips (8.6%) remains far behind automotive (77.7%) and slightly behind pedestrian (11.5%) modes of transport in Edmonton (City of Edmonton Transportation and Planning, 1995). The highest densities of bus routes corresponds with highest areas of commercial activity—the downtown area and major shopping malls (City of Edmonton, 1996). There are two peak hours of transit use: 8:30 a.m. and 4:00 p.m. (City of Edmonton Transportation and Planning, 1995). The LRT route provides important high-frequency service these same hours as well as brief peaks during hockey and football games.

3.2 Study Resolution and Child Mobility

One hundred and thirty-eight urban census tracts were chosen as the units for analysis. Census tracts were removed according to two criteria. First, tracts were excluded if their total population of children aged 0 to 15 was equal to or less than 100 (Figure 3.4). This removed non-residential, low population census tracts. Second, tracts were removed if they did not contain part of the main city road network. This removed rural census tracts, which probably represent different environments of risk to pedestrians, and deserve separate analysis.

Given the effects that different geographical aggregations can have on the results of a study, the choice to use the census tract as the unit of study resolution deserves special consideration. A thorough discussion of how resolution choices affect study results can be found in Openshaw (1984), who offers important insights into the modifiable areal units problem (MAUP). As discussed in Chapter 1, considerable research has shown that different levels of data aggregation will result in different degrees of associations, and most specifically, that increased aggregation often leads to increased statistical correlation. Attempts have been made to address this problem using weighting techniques, but disagreement in methods has discouraged intensive research in developing a purely statistical solution (Norcliffe, 1982).



Classification of Industrial and Rural Areas is based on neighbourhood subdivision categories (City of Edmonton Planning and Development, 1997). See Appendix A for census tract labels

Census tracts are formed in co-operation between Statistics Canada and local committees which define these geographic units based on four principles (Statistics Canada, 1991):

1- The boundaries of census tracts should correspond to existing, and when possible, permanent physical features (like rivers and major roads)
2- The population of non rural census tracts outside of business or industrial areas should have populations between 2500 and 8000
3- Census tracts are formed and/or split in a manner that corresponds as closely as possible to economically and socially homogenous areas
4- Census tracts are made to be as "compact as possible"

Census tracts are designed with the intention of representing homogenous and contiguous neighbourhood units useful for aggregated analysis, whereas enumeration areas are established by survey administrators to optimise the efficiency of the Census survey procedure (Statistics Canada, 1991). Census tracts are easier to compare over time than enumeration areas since their boundaries are usually defined by physical landmarks and the alteration of tract boundaries is generally discouraged (Statistics Canada, 1991). Enumeration areas usually represent smaller populations and smaller areas than census tracts, and are designed as building blocks for higher-order geographic divisions. Enumeration areas are not constructed to represent local regions or neighbourhoods with common characteristics.

Aside from the obvious difference in their spatial extents and the average population that they contain, the enumeration area and the census tract have geometric differences relevant to geographic research. Table 3.1 shows the differences in average area, average perimeter, the standard deviation of area, compactness and minimum and maximum sizes of each within the study area[2]. The average area of the census tract is

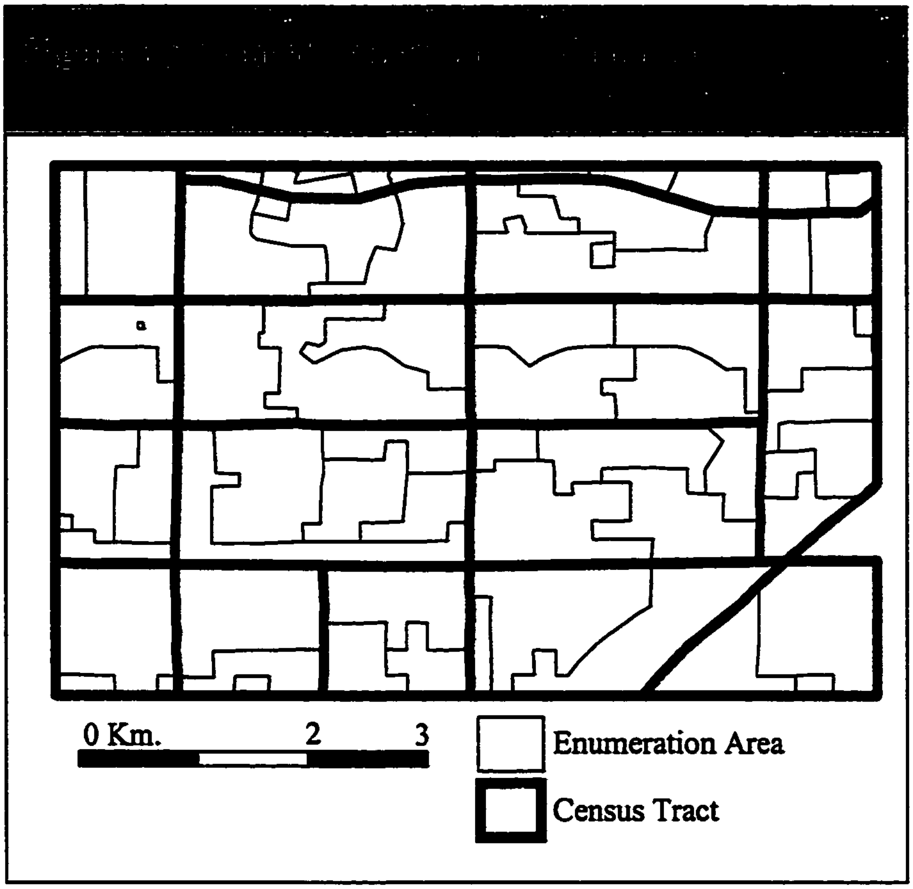
Table 3.1 The Geometry of Enumeration Areas and Census Tracts

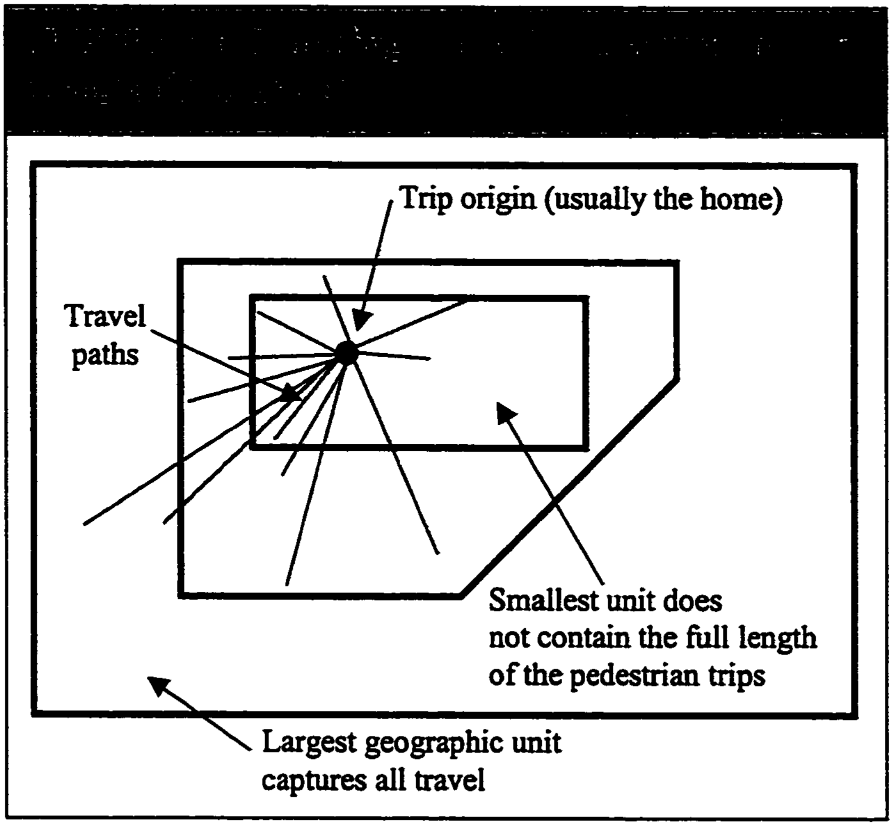
	Area ($Km.^2$)				Mean	Mean	95% CI Compactness	
	Mean	SD	Min.	Max.	Perimeter (Km.)	Compactness	Lower	Upper
CT	2.120	2.170	0.485	19.713	6.150	4.440	4.350	4.520
EA	0.413	1.780	0.001	15.641	3.100	5.310	5.230	5.390

larger than that of the enumeration area; however enumeration areas are both less compact in shape, and have a larger range of sizes, with an extreme value near in area to the largest census tract. Figure 3.5 shows a segment of the Edmonton map broken down into census tracts (thick lines) and enumeration areas (thin lines).

Several factors discourage the use of enumeration areas as study units in this project. Some Census data within the study area are suppressed in low-population enumeration areas due to confidentiality concerns. This is considerably less of a problem in census tracts; in fact, none of the 138 tracts which make up the study area have missing Census data on variables used in this research. Confidentiality of child pedestrian injury cases is better preserved in the census tract than the enumeration area as well. Census tracts represent larger populations of people, and therefore, are better at maintaining case anonymity.

The size of the study unit is also important. The larger the unit of study, the more pedestrian activity of child residents will be captured within that unit (Figure 3.6). This statement holds to the point of largest resolution where all urban pedestrian movement is contained—the city boundary itself. Obviously, as the resolution of study area increases,


0 Km.
2
3
Enumeration Area
Census Tract


Trip origin (usually the home)
Travel paths
Smallest unit does not contain the full length of the pedestrian trips
Largest geographic unit captures all travel

regional heterogeneity increases, and the meaningfulness of the ecological-level associations decrease. Therefore, a balance must be struck between capturing the pedestrian movement of children and preserving a reasonable degree of data homogeneity. The census tract best satisfies these conditions. It is also considerably more difficult to make statistical inferences about smaller geographic units (with smaller populations) than larger geographic units, when the data (in this case child pedestrian injury) are relatively rare. This issue will be discussed in more detail later.

Most Edmonton census tracts are bordered by major arterials. A further argument in favour of census tracts as study units is based on a possible boundary effect that these arterials have on the movement of children. Research in cognitive mapping suggests that the perceptions of neighbourhood in younger children are constrained by major roads (Gould and White, 1986). If indeed these roads act as boundaries to child movement, then the argument for using the census tract as the unit of study analysis is strengthened.

Another option would be to have geographically aggregated enumeration area data into larger units that better represent neighbourhoods within which children operate. Although such aggregation processes generally increase statistical associations (since they reduce within-area variability and increase between-area variability) they have been shown to bias results when the aggregation process is related to the dependent variable (Hannan and Burstein, 1974). This process also reduces the generalisability of the study results. The criteria by which Statistics Canada defines census tracts are not regionally differentiated, therefore the results from ecological-level analysis of tracts can carry meaning beyond Edmonton to similar cities in Canada. Finally, the aggregation process has no optimal statistical solution. A decision must be made to classify areas according to some criteria, and this decision will affect the statistical associations under study (Openshaw, 1984). Decisions on the classification of these areas according to the independent variables will influence the results of a multivariate model. Variables not used in the classification process will likely have more intra-group variability, and will therefore express weaker relationships with dependent variables. At best the aggregation of enumeration areas would slightly to moderately improve the statistical significance of true relationships; at worst, such aggregation would bias the results in favour of the variables used for grouping.

Yet another alternative would have been to use municipally defined *standard neighbourhoods* as the unit of study resolution. These geographic areas are on average both smaller than census tracts (though larger than enumeration areas) and more compact. The chief reason for choosing the census tract over the standard neighbourhood area is that tract data and tract geographic boundary files were immediately available and geographically referenced to road and postal code information.

3.3 Residential versus Collision Environments

Case data are aggregated to census tracts based on the location of child residence. A separate data source provides data on collision location, which could also have been used to aggregate data. One reason I choose to use the former data source as a basis for geographic analysis is theoretical: residential neighbourhoods represent places of exposure. Homes are the origin of the vast majority of pedestrian trips made by children, (Hillman and Whaley, 1979; City of Edmonton Transportation Department, 1995) and therefore, residential neighbourhoods represent an important environment of exposure to child pedestrians.

Research shows that increased exposure to road environments corresponds to increased risk of injury (Macpherson et al.., 1998). Most ecological-level and individual-level studies that consider environmental risk of the collision areas fail to account for spatial variability in pedestrian activity. In a study of environments around collision sites, Mueller et al., 1990 found that neither crosswalks nor sidewalks had a protective effect on pedestrian injury risk. However, this study did not account for the fact that crosswalks simply get more pedestrian activity than other areas of streets. Had they controlled for pedestrian traffic, their results may have confirmed the more sensible conclusion that crosswalks do offer protection to pedestrians.

Unfortunately there is little comprehensive work that explores how the pedestrian activity of children varies over space. One solution is to use information about family income as a surrogate for frequency of pedestrian travel. Evidence suggests that family income is inversely related to number of roads crossed (Roberts et al., 1994) and the proportion of pedestrian journeys made (Hillman and Whaley, 1979). The collision location database does not contain a socio-economic indicator for cases. One could attempt to control for income by using the family income measure of the neighbourhood in which the collision took place; however, this would assume that the cases live in these neighbourhoods. When this assumption is false, the relationship that income has with injury rate would be obscured, and in most cases, underestimated. Using collision location areas also complicates the calculation of injury rates. When calculating a rate using areas in which collisions occurred, the denominator is the population of children in that area and the numerator is the number of collision events, none of which necessarily involved the children in that area. Thus these rates are poor—if not meaningless—representations of injury risk.

Choosing to use the residential area as a unit of analysis is not without its own problems. Although evidence suggests that children are most often injured in their residential neighbourhoods (Mueller, 1990; Preston, 1974; Jolly 1991; Raybould and Walsh, 1995), a proportion are still involved in collisions away from home neighbourhoods. In these cases, the physical environment of the home neighbourhood is probably unimportant, although the environment of the area in which the collision occurred could still be a determinant of risk. As a result, any ecological-level analyses using the residential neighbourhood as a unit of study will probably underestimate the role of the physical environment in determining risk.

Both options offer challenges for conducting ecological-level analysis on pedestrian injury. Aggregating collision location may result in an underestimation of the importance of socio-economic factors, and prevents us from using rates as measures of risk. Using residential locations may underestimate the role of the physical environment in risk. This problem could be mitigated by linking the collision data to data on residential location, then removing cases in which injuries occurred outside of the residential neighbourhood. Unfortunately, results from a linkage of the health administrative database and collision-location resulted in a file with only 61% of cases linked to collision data. As no unique identifiers were available for linkage, probabilistic methods were used, which may have resulted in a number of false links. Combined with the fact that collision data obtained from police records are less up to date (1998/99 data were not available for linkage), and have been shown to be less reliable than data obtained through health surveillance systems (Walsh and Raybould, 1995), I choose to base the analysis on data aggregated from residential location.

3.4 Data Description

3.4.1 Child Pedestrian Injury Data

Data on injury cases were provided by the Alberta Centre for Injury Research and Control which receives data through a monitoring system reporting all injuries admitted to emergency rooms in the Capital Health Region since 1995. International classification of disease version 9 (ICD-9) code E814.7 was used to identify all cases of motor vehicle-related pedestrian injury reported in a four-year period (April 1995 - March 1999) for children under 16 years of age. This excludes pedestrian injuries involving off-road vehicles, bicycles and non-pedestrian (e.g. occupant) injuries resulting from motor-vehicle collisions with pedestrians. This same ICD-9 code has been used in previous studies of pedestrian injury in children (e.g. Roberts et al., 1992) and represents the most precise category of pedestrian injury. Pedestrian injuries not reported to Edmonton area emergency rooms could not be not included in the analysis. Injury severity was not easily determinable from the data set, nor was it inferred from the location of injury on the victim. Therefore there is no reliable manner with which to identify the most severe cases. Since vital statistics data were not incorporated into this data set, any deaths that occurred which were not processed through the emergency room are not included. In addition to standard administrative data , each injury record included date of birth, postal code of residence and gender of the injury victim. The postal codes of residence were aggregated to the census tract using a postal code conversion file developed and maintained by Statistics Canada. This allowed for a meaningful ecological analysis as well as preserving the confidentiality of injury victims.

Chart reviews of a sample of the cases did indicate some inconsistencies in the coding of data. In a subset of cases reported to the University of Alberta Hospital, several cases were miscoded as pedestrian-automobile collisions that were in fact pedestrian-bicycle collisions. Assuming the other Edmonton hospitals had a similar proportion of coding error, as many as a dozen cases (5%) may have been miscoded in the data set. Because charts from all hospitals were not reviewed, the data set was not modified to correct this miscoding.

3.4.2 Statistics Canada Census

This study uses several Census indicators obtained from the 1996 Statistics Canada Census of the Canadian Population. The Canadian Census is conducted every five years and collects data on approximately 98% of the households in Canada (Statistics Canada, 1991). Census representatives hand out one of two questionnaires to each household in Canada which are filled out by the household members themselves, and mailed back to Statistics Canada. The information contained on the short questionnaire is obtained from every household and contains information on household relationships, age, sex and mother tongue of residents. The information contained on the long questionnaire is obtained from one of every five households, and covers more detailed information including employment status, education, income and dwelling characteristics.

The Census data were made available through the Data Liberation Initiative, a dissemination program by which researchers at participating academic institutions can obtain Census data free of charge and in a timely fashion. The University of Alberta data librarian also released the 1996 Census geographic boundary files, street network files and the postal code conversion files in the summer of 1998.

Statistics Canada recognises five sources of potential error in Census data: coverage error, processing error, non-response error, response error and sampling error (Statistics Canada, 1991). Of these, the latter three are most likely to bias the results of this study, and therefore, each will be discussed briefly. It should be noted, however, that none of these potential errors in the Census data can be easily rectified in the context of this study. This brief overview only serves to show that interpretations related to these data must be made cautiously.

Non-response Error is the failure of the Census to gather responses from households. A significant majority of non-responses have been identified as coming from Indian-Reserves, which would not apply directly to this study. However, certain urban families may be less available during the Census period; for example, families that vacation in the spring may be underreported.

Response Error is the inability of a participant in the Census to understand parts of the questionnaire. Any associations between demographic or socio-economic status and the ability of a person to accurately answer questions in the Census may affect the survey results.

Sampling Error results when a 20% sample value in the long questionnaire is not equal to the value were 100% of the sample stratum asked the question. Below is a table with all Census data used in this study, and an indicator as to whether or not data are 100% or 20% samples (Table 3.2).

TABLE 3.[illegible]: Sample Size for Census Data Used in This Study		
Variable	**100%**	**20%**
Population of Children	X	
Proportion of Low Income Families		X
Proportion of Apartment Dwellings		X
Total number of Dwellings	X	

Occasionally, Census data are rounded (to five or zero) which makes certain variables less reliable (Statistics Canada, 1991). This is most often a problem for variables with low frequencies—like number of male Greek speakers in an enumeration area. Out of a concern for individual privacy, some data are suppressed altogether. By restricting the study to census tracts, one would expect that rounding error and data suppression do not pose a serious problem.

3.4.3 Child Population

The 1996 population of children in the study region aged 0-15 was used as the denominator in rate calculations. As this represents only one year, and injuries were gathered from four years of data, the population was multiplied by four to obtain the appropriate denominator for a rate calculation. This assumes that the population of children remained constant in all Census Tracts between 1995 and 1999. As mentioned previously, some localised variation in population can be found in Edmonton between 1981 and 1996, but over these four years, the author deems it reasonable to assume a static population.

3.4.4 Proportion of Low Income Families

Virtually every modern study of child pedestrian injury at the individual or ecological level has indicated the importance of family income (or an equivalent indicator) as a variable in determining risk. After controlling for behavioural, familial, ethnic and cultural factors, income has been shown to be a feature of importance. Perceptions of neighbourhood may influence pedestrian behaviour, and may vary dependent on individual-level and contextual factors.

As this study lacks income data on individual cases, multilevel analysis of the contextual versus compositional effects of income on child pedestrian injury is impossible, and therefore the relative contribution of these effects in describing the geography of risk cannot be investigated. As a result, the proportion of low income families probably represents a mixture of both contextual and compositional effects. In this study, statistical associations involving low income may be affected by cross-level bias. However, the alternative—not including a socio-economic indicator—would be to completely ignore the role of a well established determinant.

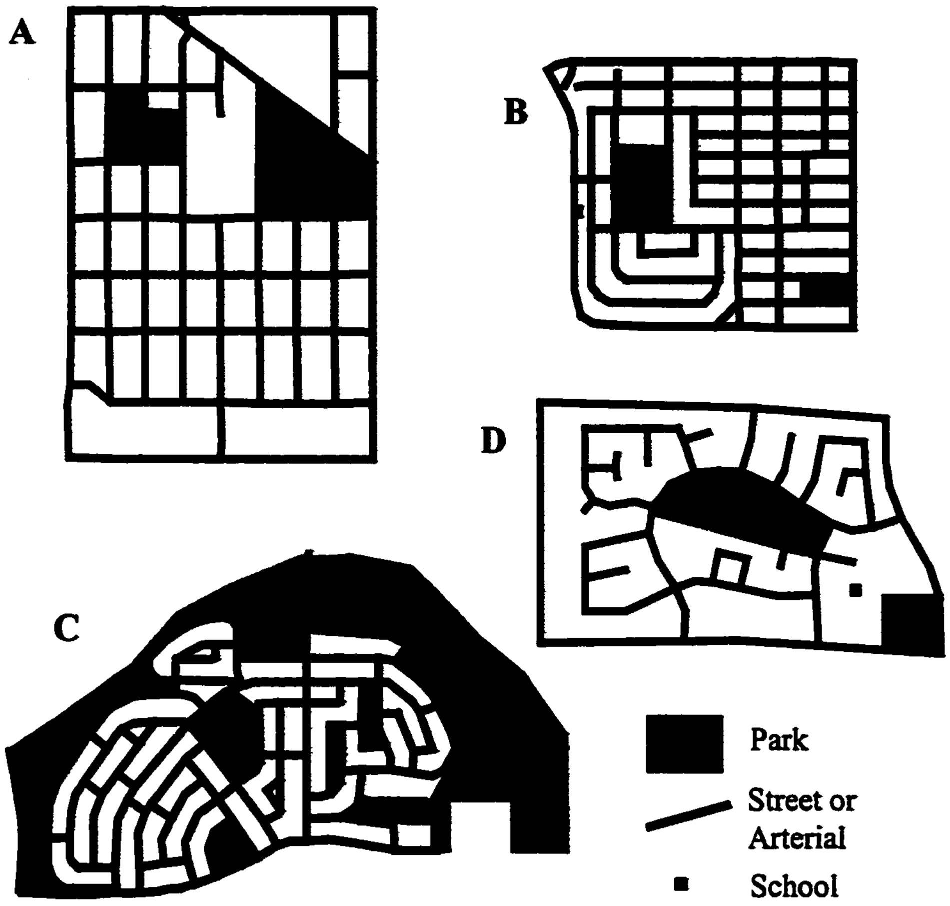
3.4.5 Proportion of Apartment Dwellings

The proportion of apartment dwellings in each census tract is obtained by dividing the total number of apartment dwellings (located in apartment buildings above and below five stories) by the total number of dwellings. Availability of play space in a child's neighbourhood may be a contributor to the risk of injury. Proportion of apartment dwellings gives an indication of the yard space available in the neighbourhood, which gives an indication of the availability of near-home safe play-space, especially for young children. Like proportion of low income families, there are potentially contextual and compositional factors at work, and therefore, this variable must be also interpreted carefully. Individual children may be at greater risk if they live in apartments—since they have no safe yard to play in on their own. Children living in a neighbourhood made up of mostly apartment dwellings may be at greater risk simply because there are not many yards to play in overall—independent of whether or not the children live in an apartment themselves.

3.4.6 Neighbourhood Planning Data

Planning data are derived from Wang's neighbourhood classification (Wang 1994). Wang identified 5 categories of neighbourhood planning in Edmonton, the main four of which will be used in this study: grid neighbourhoods, modified grid neighbourhoods, independently designed neighbourhood units, and hierarchically planned neighbourhoods. This study is expanding on Wang's general conclusions by using child pedestrian injuries in a multivariate statistical framework. This is best done by incorporating a planning variable as close as possible to the same variables used by Wang. The definitions below are based on Wang (1994) and Hodge (1991). Wang's research suggests that there is an order of increased risk from grid neighbourhoods, which are unplanned, to hierarchically planned neighbourhoods. For the purposes of this study, planning type is included in this study as a single variable on a scale of one to four—one being grid neighbourhoods, four being hierarchically planned (Figure 3.7). For more on the use of contrasts in favour of dummy variables, see Cohen and Cohen (1975).

Grid neigbourhoods are unplanned neighbourhoods formed previous to the widespread use of the car (Hodge,1991). These areas are considered unplanned because structures were built based on the availability of land and commercial interest, not on principles of pedestrian safety



A- Unplanned grid neighbourhood B- Modified grid neighbourhood
C- Independent neighbourhood unit D -Hierarchically planned
neighbourhood

or traffic-flow efficiency. Residential lots face each other in front, and are separated in the back by alleys.

Modified grid neighbourhoods consist of "partially structured" land use patterns. Planned elements –like centrally located facilities- are combined with unplanned, traditional street grids.

Independent neighbourhood units are influenced by the work of Clarence Perry, who promoted land-use planning around service points—like malls and schools. Separate neighbourhood areas were constructed around the service points, usually separated by major arterials. In Edmonton, few of these neighbourhood units are divided by major arterials (Wang, 1994). The idea of the neighbourhood unit was promoted in large part out of drastic increases in automobile use and automobile related fatalities (Hodge, 1991).

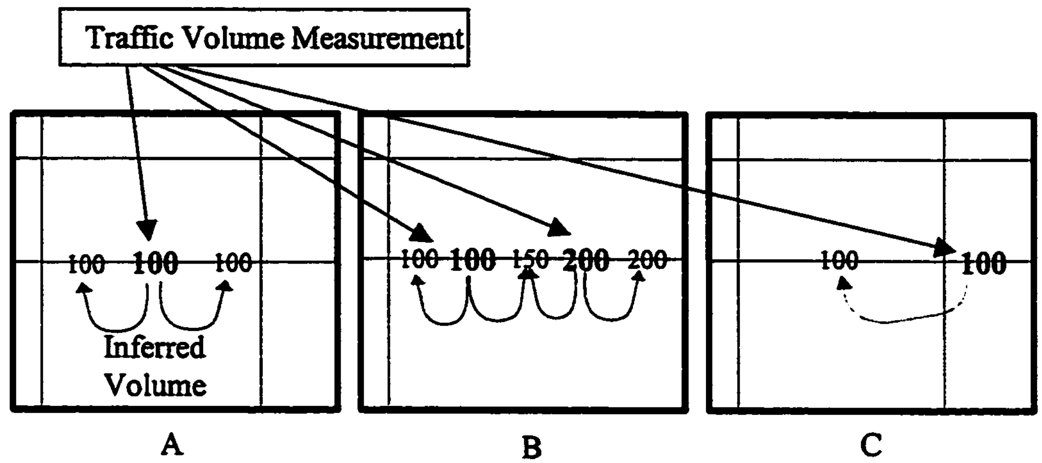
Hierarchical neighbourhoods are similar to the independent neighbourhood units and are designed around public service facilities—usually schools. Residential block patterns are structured to separate pedestrian routes to these facilities from automobile traffic. Parks are centrally located, and are also separated from main roadways (Hodge, 1991).

The municipal neighbourhood divisions Wang used to classify planning type do not match perfectly with census tracts. On occasions where the neighbourhood boundaries matched the census tract boundaries exactly, the census tract was assigned the planning characteristics of the associated neighbourhood area. When multiple neighbourhoods were contained in census tract, the census tract was assigned the planning characteristics of the largest neighbourhood contained within it.

3.4.7 Traffic Data

A traffic density index is calculated for each census tract using two sources of data. The purpose of developing this index is to develop integrated data of general volume characteristics of census tracts that accounts for the density of roadways and the volume of traffic in neighbourhoods. Wang (1994) showed that planning characteristics were important in explaining the variability of collisions between neighbourhoods; however, this may simply be a function of general city traffic flow, which shows a pattern of increased density from suburban planned areas to downtown unplanned areas. Wang did use traffic flow data to help classify Edmonton neighbourhoods into particular planning categories, but the variable was not systematically included in his analysis. Therefore it remains uncertain whether or not his findings only reflect the fact that there is less through traffic in suburban areas versus the actual environmental design that determines risk to pedestrians.

Mean daily traffic volume for 1996 is based on data collected by the Edmonton Transportation and Planning Branch. These data are a compilation of axle counts taken at measurement stations in order to make estimates of average weekday traffic. These stations are combined with a map of arterials generated by Statistics Canada. Because the volume data represent points, and arterials are represented by lines, an inference procedure was developed to integrate volume data with the arterials in a consistent manner. Figure 3.8 A-C shows the three most common cases of volume inference. An arterial segment with a lone measurement station adopted a volume equal to the measurement station, extending as far as the next intersection (case A). Arterial



Traffic volume measured at points (in bold) is used to infer volumes for complete road segments. Numbers in normal font indicate inferred volume. In case A, volume of the road segment is based on the single measurement of the segment. In case B, volume is based on the average of two measurements (for sections between measurements) or on a single measurement (for sections between the measurement point and the end of the segment). In cases where the segment has no measurement, volume is inferred based on the neighbouring segment (case C).

segments with two or more measurements were assigned a volume to their sub-segments according to case two: sub-segments between the points were given a value equal to the average of the two points (case B). When segments on a major arterial lacked road volume measurements, the nearest measurements on the same arterial were used as estimates (case C). The majority of the estimates were based on cases A and B. Several arterials on the street network map lacked any measurement altogether—in these cases, the arterial is not included in the data set.

These traffic data were aggregated into census tracts by multiplying the length of the road segment by the volume of the road segment and then summing this value for each arterial segment in each census tract. Roads were counted twice if they represented a border between two census tracts. This variable, henceforth referred to as the *traffic density index*, describes the density of traffic in census tracts independent of major freeways (which do not really represent a hazard to pedestrians since they are usually separated by barriers, and walkways are above the roads themselves). These freeways were identified in consultation with experts in the planning department. This index gives an indication of the density of threatening traffic. As a measure of an environmental characteristic, any direct relationship that this variable has on injury risk is probably contextual. It is possible that traffic volume is overestimated on certain arterial segments. The traffic measurement equipment is based on axle count, and therefore roadways with considerable quantity of multiple-axle (three or more) traffic may have inflated index values not reflective of the actual number of automobiles that travel the roadway, although this is most likely an issue on freeways, most of which have been excluded.

3.4.8 Park Data

Park data were digitised using several real estate maps and a collection of air-photos. Three kinds of parks were identified in the city of Edmonton according to criteria related to their use by children: i) river valley parks which include the N. Saskatchewan River Valley, the Whitemud Creek Ravine and the Mill Creek Ravine, ii) school parks (defined as parks within 50m of a school) and iii) city parks (defined as parks more than 50m from a school) that are not already classified as River Valley parks. The proportion of park space within a census tract was calculated by dividing the total area of school park space by the total area of the census tract. School parks were selected because they represent the parks that children are most likely to frequent independent of automotive transport and parental supervision. As a measure of an environmental characteristic, the park data represent a contextual factor of risk.

3.5 Mapping Methods

3.5.1 Injury Incidence and Rates

Techniques used to map incidents of injury are comparable to the mapping of mortality as both represent discrete and relatively rare events. These techniques are not only the basis for presenting disease frequency on a map, but often generate the dependent variables used in ecological-level regression models. However, point maps themselves have limited informational value with regards to relative risk since they give no indication of injury events in terms of population at risk. For this reason, rate maps are frequently used to portray risk.

Mapping mortality rates and disease incidence rates has been a frequent practice among epidemiologists, geographers and public health officials. In addition to identifying regions with unexpectedly high rates of a particular health concern, mapping health outcomes can provide information about sources of exposure. While the creation of cancer atlases and the mapping of malaria and other vector borne diseases are perhaps the most common recent application of mapping techniques in health, researchers have expanded their scope into the mapping of injury (e.g. Raybould and Walsh, 1995; Braddock, 1994).

The mapping of crude rates—health outcome in the numerator, population at risk in the denominator—is a simple method of mapping variations in a health outcome over space. When specific age groups are studied, the population of these age groups replace total population in the rate equation, and the rates are generally more informative since they are unconfounded by variations in age or other population structures between areas (Meade et al., 1988). A map of these incidence rates identifies areas where the rates are high or low in relation to other areas on the map, however, the map does not give an indication as to whether or not the observed rates are above or below an expected value.

Mortality/morbidity ratios expand on the simple rate map by describing the degree to which observations depart from a global (total study area) value. The following formula illustrates:

$$\boxed{E_i = P_i * GR} \qquad 3.1$$

$$\boxed{MR_i = O_i / E_i} \qquad 3.2$$

The expected number of cases E in an area is determined by multiplying the population at risk in the area (P_i) by the global rate of occurrence (GR) (Equation 3.1). This expected value is then used in the denominator of Equation 3.2 where the morbidity ratio for a given area (MR_i) equals the total number of outcomes in that area (O_i) divided by the expected number of cases (E_i). The MR is usually multiplied by 100. For areas in which the observed number of cases is close to the expected number of cases, MR will be close to or equal to 100. When the number of cases in an area is below the global expected, MR will be less than 100; when cases are greater than the global expected, MR is greater than 100.

3.5.2 Small Numbers Problem

The small numbers problem is perhaps the most important methodological issue in disease mapping. Put simply, when population counts in areas are small and health outcomes under observation are rare, rates in some areas may be "unstable", and therefore demonstrate an exaggerated spatial variation (Clayton and Kaldor, 1987). This data instability is a function of sampling error that is typically observed in data described by the Poisson distribution, where small changes in the number of events in some area results in large changes in rates (Table 3.3). A one case increase in area *B* has a small effect on the rate of the outcome; an equivalent change in area *A*

Table 3.3 Example of Rate Instability[1]

Area	Case	Populatio	Rate	Cases +	New Rate
A	1	100	0.01	2	0.02
B	10	1000	0.01	11	0.011

[1] Adapted From Pringle (1996)

results in a doubling of the rate of outcome occurrence. Neither morbidity ratios nor morbidity rates are equipped by themselves to account for this variability. What results in many circumstances is a map of misleading information—showing higher rates in areas where the population counts are smaller, but where the inflated rates are probably a product of expected variability in a Poisson process. For data in which case counts are high, such variability has little effect on rates. However, in the study of meaningful geographic units—which are almost always small and contain low population counts—this problem is likely to persist for almost every health outcome.

The mapping of empirical Bayes estimates offers an alternative to the classical disease mapping techniques. This method has been designed to specifically address the small numbers problem, offering a smoothing process to stabilise rates of rare health outcomes in areas with low population counts. This method of data adjustment is based on the Bayesian statistical tradition. Bayesian statistics address a perceived shortcoming of classical statistical techniques which fail to account for "prior belief" in the study of probability. The heart of Bayesian inference is the idea that scientific decision making should be able to build on itself. Using prior knowledge from previous statistical analysis or other sources of information, a researcher should be able to modify classical inference measures to make more intuitive inferences about phenomena (Iverson, 1984).

In classical statistics, before a statistical inference is made, criteria are established for accepting or rejecting a null hypothesis that *in the long run*, the observed phenomena would not be unexpected. For example, experience suggests that in an unbiased coin-flipping experiment of ten tosses, one might expect an equal number of heads and tails. To test this assumption using the classical statistical method, we establish a criteria for rejecting this claim that takes into account the size of the sample and the distribution of the population. If the criteria for rejecting the null are obtained, then we proceed to describe the population in terms of the sample. This procedure does not take into account what we all know about physics, or the years of observations we've had flipping coins.

Nor does this procedure of analysis easily incorporate new data—the statistical significance of a second experiment is determined using the same criteria, and without incorporating the previous experiment into the calculation. The Bayesian statistical tradition offers a method of incorporating prior information—like what we know from other statistical tests or philosophical inquiry—into a statistical inference. The philosophy behind Bayesian inference is flexible enough for one to incorporate virtually any previous knowledge into statistical decision making. This may explain why Bayesian inference procedures remain rare outside of statistical journals (Moore, 1997). A more thorough discussion of Bayes' Theorem and the strengths and weaknesses of Bayesian approaches to statistical decision making can be found in Iverson (1984) or Winkler (1972).

Empirical Bayes estimates are adjusted rates which take into account prior knowledge about properties of the distribution of the observed data over space. There are several methods with which one can calculate the prior distribution, but the gamma distribution is most common (Pringle, 1996). Unlike the Poisson, which describes the probability of n discrete events in a given space, the gamma distribution describes a continuous variable, in this case, the rate. The two parameters that describe the properties of the gamma distribution—α (scale) and v (shape)—are used to adjust the incidence rate where O_i, P_i and GR are the observed cases, population and global incidence rate respectively (Equation 3.3). A recursive formula is required to obtain

$$\boxed{\frac{(O_i + \nu) * GR}{P_i * GR + \alpha}} \qquad 3.3$$

values of α and ν^3. Although there are several methods with which to derive the prior distribution parameters, this iterative calculation is the simplest. When the number of outcomes in a given area is high and the population is large, this adjustment procedure has a minimal effect, and the incidence rate is stable. Alternatively, when the number of events and the population are small, the adjustment procedure has a more significant smoothing effect.

As a general method for dealing with the small numbers problem, empirical Bayes estimates offer an improvement over normal rates because they contain more intuitive information than standard smoothing techniques (like log transformation). Clayton and Kaldor (1987) used empirical Bayes estimates to stabilise data on lip cancer in England. In their work they compared various methods of obtaining prior knowledge (including gamma, log-normal and several other distributions). Langford (1994) made use of empirical Bayes estimates to adjust relative risk of leukaemia and Saunderson and Langford (1996) showed the usefulness of these adjustments in mapping suicide rates in the United Kingdom. Pringle (1996) showed that empirical Bayes estimates show a representative and intuitive picture of cancer deaths in Ireland. Perhaps the greatest indication that the technique is gaining widespread acceptance among geographers is the chapter devoted to it in Bailey and Gatrell (1995).

3.5.3 Distance Calculations

Both collision location (by closest street intersection) and location of residence (by postal code) are available for many cases in the data set. Measuring the distance between these points gives a general indication of the range of distances in which children are hit by automobiles. This measurement is the best available indicator of distance between collision location and residence. Location of residence by street address would result in more accurate measurements, but these data were not available and would represent an unreasonable risk to case confidentiality. As the regions which the postal code points represent are larger in outlying areas of the city, distances may be biased—distances will tend to be larger for cases in which location of residence is in outlying regions.

3.6. Statistical Methods

The primary analytical tool in this study is multiple regression modelling. Given the theoretical and methodological issues complicating the use of ecological-level analysis, it is particularly important that this study critically assesses both the meaningfulness and robustness of the modelling procedure. All methodological issues discussed below represent the main statistical concerns related to the statistical modelling of these cases of child pedestrian injury to social and environmental risk factors.

3.6.1 Tests of Correlation

Correlation techniques measure coincident variation between variables, and are therefore inappropriate tools for studying causal relationships. Despite this limitation, as preliminary statistical techniques, tests of correlation are simple and easily applied methods for isolating variables to be involved in more detailed investigation. The most popular technique of correlation analysis is probably the product-moment correlation, which measures relationships between normally distributed continuous variables (Ebdon, 1985).

A popular alternative among geographers is the Spearman's rank correlation, a non-parametric test of the association between ranked data (Shaw and Wheeler, 1994). The Spearman's rank correlation can be used to test naturally ranked data or ranked continuous data, and is less sensitive to extreme values than the product-moment correlation (Ebdon, 1985). This insensitivity can be beneficial when the goal of the

$$\boxed{r_s = 1 - \frac{6\Sigma D^2}{N(N^2-1)}} \qquad 3.4$$

correlation procedure is to identify variables of importance for more advanced statistical testing. Equation 3.4 can be used to calculate the Spearman's rank correlation, where N is the number of rank pairs, and D is the difference between ranks for given observations. Neither correlation technique is without its limitations: the former is more affected by extreme values in the data, the latter ignores extreme values altogether, and therefore, both will be used where appropriate. Their purpose in this study is simply to suggest which variables are most relevant for inclusion in the regression model.

3.6.2 Choice of Regression Techniques

The primary challenge of a study in small areas is dealing with data variability and identifying areas where incidence rate is a reflection of real risk rather than a function of random (and not unexpected) variation. In this study, the data are rare and the population of study areas is small, and therefore, a method must be chosen that takes into account these uncertainties.

Traditionally, modelling rare data involves Poisson regression techniques. Ordinary least squares regression of rare incidence data is often inappropriate since models may predict nonsense values (values below zero) and in most cases violate assumptions of constant error (Lovett and Flowerdew, 1989). The reweighted least squares procedure used to obtain parameter estimates in a Poisson regression model does not assume constant variance, and generates intuitive model estimates. Unfortunately, there are very few established model diagnostic tools developed and tested for Poisson regression modelling, making critical model assessment difficult. An alternative method for dealing with non-constant variability in the incidence data is a weighted least squares approach which adjusts each rate by the inverse of its variance. In this case the most logical weights would be the population of children in each tract.

Both of these methods offer solutions that are embedded in the modelling procedure itself, making adjustments when the statistical model is executed (sometimes called *run-time* adjustments). The data incorporated into the model are not smoothed to adjust for uncertainty, but instead, these issues are dealt with the through the modelling process itself. As a fundamental goal of this study is to understand the variability of rates, it is desirable to have a method which allows for the easy comparison of observed and estimated values of the outcome. This is best achieved by first applying a smoothing technique to the data and then running a statistical model using these smoothed rates. By running a regression model of these smoothed rates, one can then observe differences in the model estimates and the observed values. This is difficult to do using either of the methods mentioned above. Although the estimates generated from a Poisson model are adjusted for data heteroscedasticity, the observed data are not. Therefore comparisons between the observed and predicted values are difficult.

The best method for achieving the study goals is to generate empirical Bayes estimates of the rates *before* the modelling procedure, then generate an ordinary least squares model that predicts rates without applying a run-time smoothing process. This way, observed and predicted rates are comparable, and one can gain both an effective understanding of the spatial variability of risk and critically analyse the spatial characteristics of model residual error. In order to substantiate this method, the model parameters from the ordinary least squares model are compared to a run of the incidence data using a Poisson regression model. If the same parameter estimates show importance in both models, then this helps to validate our decision.

3.6.3 Linear Regression

Regression techniques are frequently used to understand relationships between phenomena in the social and physical sciences. In the social sciences, experimental conditions are rarely strictly controlled, and definitive conclusions are hard to make. Nevertheless, regression methods remain very popular techniques for data analysis. If the intention is to understand the degree to which vector x_i is a linear predictor of vector y_i, the basic formula of a linear regression equation takes the form:

$$\boxed{\mathbf{y}_i = \mathbf{B}_0 + \mathbf{B}_1\mathbf{x}_{1i} + \mathbf{e}_i} \qquad \mathbf{3.5}$$

In this bivariate regression formula, B_0 is the intercept or model constant, B_1 is the coefficient associated with observations of x_i and e_i is the variability in y_i that remains "unexplained" following the modelling process. The multiple linear regression equation possesses the same properties as the bivariate case, the key difference being the number of predictors. As in the bivariate case, the model R^2 gives an indication of the amount of variance in y that is explained by the predictor variables $x_1, x_2, x_3 \ldots x_n$. However, each coefficient now represents a unit change in y given a unit change in x_n while holding the other coefficients constant. Independent or *partialled* effects of the independent variables are calculated by examining the independent relationships between the dependent variables and the residual error among the predictors. The method by which multiple predictor coefficients are obtained is well summarised by Von-Eye and Shuster (1998).

There are three general assumptions in least squares regression models: absence of specification error, absence of measurement error, and a "well-behaved" error term (Lewis-Beck, 1980). Of these, the third can be subdivided into five assumptions: all error terms have an expected value of zero, error terms are homoscedastic (constant variance over all values of the predictors), error terms are not autocorrelated, error terms are not correlated with the independent variable (s), and the error terms are normally distributed (Lewis-Beck, 1980). The meaningfulness of a model's parameter estimates is dependent largely on the degree to which these assumptions are valid; therefore, careful consideration of these assumptions is required before a modelling procedure can be considered valid.

Controversy persists regarding the effects of assumption violations (particularly with regard to the residuals) on multiple regression. Extreme violations of regression assumptions probably do affect the likelihood that a given model represents something meaningful. Generally speaking, asymptotic statistical theory (i.e., the central limit theorem) suggests that techniques relying on these assumptions (like *t* scores) are statistically robust when sample sizes are larger (Statsoft, 2000). Unfortunately, there are no obvious boundaries that a researcher can use to identify when a sample of observations is large enough. Therefore, we will consider the affects of assumption violations in detail.

3.6.4 Methods of Model Appraisal

Evaluating the ability of a modelling procedure to fit a given data set is dependent on both theoretical and statistical measures. The theoretical discussion will be reserved for Chapter 5. Below are the main statistical considerations for evaluating the fit and suitability of the modelling procedure. Each model diagnostic will be addressed specifically in Chapter 4. The software package *Stata* has a strong set of tools to assess model fit, and will be relied on for most model diagnostics. Although not all methods used are the same, the guideline for model assessment is based on Jerrett (1997).

When variables lacking meaningful explanatory power correlate with the relevant variable(s) in the model, model coefficients may be biased (Berry and Feldman, 1985). Similar problems occur when relevant variables are excluded from the model. Specifically, variables in the model will correlate with the model error term, thereby violating one of the necessary conditions of multiple regression. These are problems associated with the mis-specification of the model, frequently termed *specification error*. The relationship between relevant variables that are excluded from the model and variables that are not excluded is an important determinant of this biasing process—the stronger the correlation between included and excluded variables, the greater the biasing effect (Berry and Feldman, 1985).

Models with a strong theoretical basis are less likely to contain these types of errors. These models are more likely to include all relevant predictors and exclude all irrelevant predictors. A model with little or no specification error is more likely to be representative of reality, and more likely to show meaningful relationships between explanatory and outcome variables. It is particularly difficult, however, to systematically identify missing model variables, or clearly identify a case of specification error. Although the R^2 value is a general measure of explained variance, other methods can offer additional insight. In cases when theory fails to identify mis-specification, there are statistical tools that can aid the researcher. In a study of the effects of mis-specification, Ramsey (1969) developed a method for identifying the absence of missing variables. The regression specification error test (RESET) approximates a missing parameter by adding the powers of the predicted value (y^2,y^3,y^4) to the model. An F statistic is used to determine whether the differences between these approximations is significant enough to indicate that an important variable was omitted (Stata, 1997).

Outlying observations can exact a strong influence on some regression models. Although detection of outliers is sometimes possible with scatterplots of the dependent variable and independent variables, understanding their role in the model often requires an understanding of residual error. This is of critical importance since outliers do not always have strong effect (also known as *leverage*) on a given model. Leverage versus residual plots can give an indication of the role outlying observations have on the model. A commonly employed method used to identify such observations is to observe changes in the model following the temporary removal of observations from the data set. This can give an indication of the relative effect of particular outliers. In cases where one outlying observation is responsible for making one of the explanatory variables change from a significant to an insignificant contributor to the model, caution must be applied to the interpretation of that variable. Since issues relating to the stability of outlying observations relates to issues of data stability discussed later, nothing more will be said here except that outliers should be judged based on both the likelihood that they represent real measurements and their effect on the overall model. Accurate observations should not be removed from the model simply because they affect the results in an unexpected manner. As the child pedestrian injury data are derived from an administrative database, and not collected by the researchers, there are no simple ways to identify mistakes in the data.

Ordinary least squares estimation produces regression parameters (B_0, B_1, $\ldots B_n$) that are the best linear unbiased estimates when the regression assumptions are valid. The combination of a non-normally distributed dependent variable, small sample size and small R^2 virtually guarantees a non-normal distribution of residuals. As mentioned previously, the effect of such non-normality is unclear. We shall use a combined test for skewness and kurtosis to assess the normality of residual distribution (Stata, 1997). However, we will not concern ourselves with minor departures from normality in the residuals.

It is common in the social sciences to generate models in which the residual error is not constant over values of a predictor, thus violating the assumption of homoscedasticity. Regression parameter estimates in a heteroscedastic model are no longer the most efficient, and are more likely to differ from the true (population) estimates (Berry and Feldman, 1985). Ideally, this problem should be addressed by revisiting theory—adding an important variable or obtaining better measurements. Frequently, however, a statistical solution must be sought. One of the simplest solutions for dealing with heteroscedastic dispersion of residuals is to transform one or more of the model variables. Unfortunately, models containing transformed variables can be more difficult to interpret, and also assume a functional form different from the non-transformed model (Johnston, 1978). *Stata* offers a tool for the calculation of standard error that is less prone to the effects of heteroscedastic residuals (Stata, 1997). If the non-robust standard errors are similar to the robust, there is generally good reason to think the parameter estimates are efficient, even in the presence of heteroscedasticity.

Multicollinearity is present in a regression model when the predictors are strongly related to each other, such that a well fitting model contains few or no predictor variables that have statistically significant partial coefficient values. In an explanatory model such as the one used in this study, high multicollinearity can be particularly problematic, in large part because theory alone is incapable of separating the effects of one independent variable on the model while controlling for the effects of other strongly correlated variables. The result is larger standard errors of the partialled coefficients, and lower significance values (*t scores*) (Berry and Feldman, 1985). One of the easiest methods for eliminating the effects of multicollinearity is to remove predictors which have a correlation coefficient greater than .80 (Berry and Feldman, 1985).

Additivity is a term referring to the assumption that in a given multiple regression model, the change in the dependent variable, given a change in one of the independent variables, is unrelated to the value of other variables in the model. The validity of this assumption is important for the results of the model to be meaningful—the model parameters are meant to represent the effects of change in one of the parameters while all other independent variables are held constant. If one of the independent variables demonstrates a non-additive or *interactive* relationship with another variable, then the interpretation of the model becomes more difficult. One method of identifying the effects of non-additive relationships is to run the regression model with mutiplicative terms (predictors multiplied with each other). If a multiplicative term shows statistical importance in the model, then the relationship between a given pair of variables may not be additive. In these cases, the multiplicative term should remain in the model and theoretical explanations for the non-additive relationship should be sought.

Nonlinearity refers to cases where change in values of the independent variable correspond to changes in the dependent variable that are not the same for all values of the independent variable. In bivariate regression, scatterplots can be useful indicators of non-linearity in relationships. Quadratic terms (like the square or cube of a predictor) can be added to a regression model to detect non-linear relationships. When the square of a predictor provides a better fit than a normal predictor, then a non-linear relationship is probably present.

Autocorrelation describes spatial or temporal associations between variables (Kendall and Buckland, 1971). Autocorrelation of residuals in a regression model is an important concern since least squares regression assumes that residuals are independent. In ecological-level studies, spatial autocorrelation is a common concern. Spatial autocorrelation refers to the clustering of variables in space. In ecological-level models, model residuals are frequently tested for spatial autocorrelation in order to identify important missing variables. In a study of precipitation in California, a clustering of negative residuals on the lee side of the mountains indicated the role of these areas in influencing the amount of rainfall, a factor unaccounted for the original model (Bailey and Gatrell, 1995). *Moran's I* is a commonly used measure of spatial autocorrelation, and is calculated using Equation 3.6.

$$\boxed{I=\frac{n\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i-\bar{x})(x_j-\bar{x})}{\left(\sum_{i=1}(x_i-\bar{x})^2\right)\left(\sum\sum_{i\neq j} w_{ij}\right)}} \qquad 3.6$$

Where x_i and x_j are the values of the variable x in areas i and j respectively, and w is the spatial neighbour weight. In the calculations employed in *S-plus*, w is equal to 1 when areas *i* and *j* are identified as neighbours, and 0 otherwise. The meaning of a given calculation of *Moran's I* is dependent on the spatial arrangement of the data; however, generally speaking, positive values indicate clustering, negative indicate dispersion, and values near zero indicate a random arrangement (Ebdon, 1985).

In the study of spatial patterns, scale of analysis and the structure of the spatial relationships can drastically alter results. In this case, the system which determines neighbouring status is an important consideration. *S-Plus* offers a number of options for calculating spatial relationships; among them *first order* (all adjacent areas are considered neighbours) and *adjusted first order*. The adjusted first order neighbouring technique involves two steps. First, all neighbouring areas fulfilling the first order criteria are assigned a connective weight of one. Second, the average distance between the centroids of area i and all neighbouring areas j is calculated (Splus, 1998). Any area with a centroid within this average distance is also assigned a weight of one. Second order calculations will not greatly affect the number of joins in cases when the tracts are rectangular and of similar size. Instead, it will add joins in cases of oddly shaped tracts, and in the cases where the size of tracts in a region are variable. In this study, *Moran's I* is calculated using both neighbour weighting techniques and results will be compared to determine the best measure.

3.6.5 Modelling Smoothed Rates

The empirical Bayes estimates are adjusted for uncertainty before modelling. Thus the dependent variable is the adjusted rate. As mentioned earlier, Poisson regression involves a run-time adjustment, and therefore does not generate predicted values useful for data mapping. A model of smoothed rates allows for the development of a map of predicted rates in units comparable to the observed data. Recent work has suggested that empirical Bayes estimates, while an improvement over unadjusted rates, tend to present an "over-shrunken" (that is, less variable) result in comparison to the true rate (Devine et al. 1996). A constrained empirical Bayes estimation procedure was developed to address this problem (Louis 1984). To address these possible complications of the technique, we compare results of the least squares model to the results of a Poisson regression model. Significant differences in results could suggest that constrained empirical Bayes estimates should be used in favour of standard empirical Bayes estimates.

3.7 Software Tools

Maps were developed using the *ArcView* geographic information system and several of its extensions. Deterministic data linkage was conducted in *ArcView*—for example, the joining of postal codes to their respective census tracts. *ArcView's* object oriented programming language *Avenue* was used to program some of the functions used in the analysis. Several statistical applications were used to analyse and present data for analysis. *Stata* was used primarily to perform the analysis and describe the data. *SPlus* was used to calculate several of the spatial statistics. *SPSS* was used to create most tables and graphs.

Note

[1] The 1981 Canadian Census uses different age-grouping categories than the 1996 Census. In order to allow for meaningful comparisons between these Census years, the 0-14 year age-group had to be used in this particular comparison, even though this study includes children 15 years and younger in the analysis.

[2] Compactness is calculated by dividing the perimeter of the tract by the square root of its area.

[3] This formula can be used to obtain values of α and ν:

$$\frac{\nu}{\alpha} = \frac{1}{N}\sum_i \frac{O_i + \nu}{E_i + \alpha} = \frac{1}{N}\sum_i \theta_i$$

$$\frac{\nu}{\alpha^2} = \frac{1}{N-1}\sum_i \left(1 + \frac{\alpha}{E_i}\right)\left(\theta_i - \frac{\nu}{\alpha}\right)^2$$

O_i is the observed events in area i, E_i is the expected number of events in area i (Ei=Population$_i$ * Overall Injury Rate). θi is the standardised incidence rate for area i. Appendix C contains code that can be used generate the values. For the first iteration, values of α and ν should be equal to one. First we solve for ν/α (mean) then ν/α^2 (variance). A value of α is obtained by dividing the mean by the variance which is then used to solve for ν. These values are then substituted back into the equation. This process is repeated until values of α and ν converge. If the program does not converge (default setting is to seven decimal places) then an index of two should be tried. Index values of three, four and so on should be entered until the data converge.

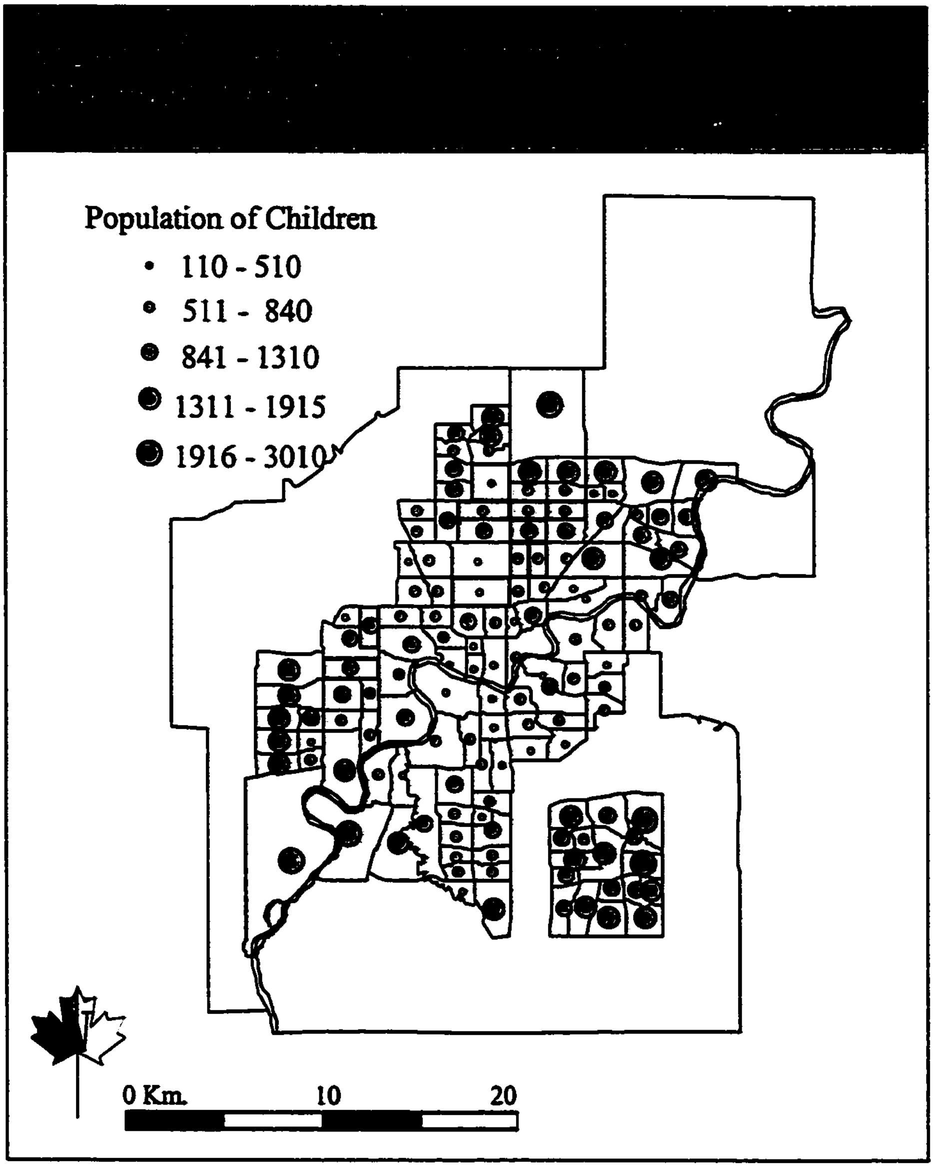
4. Results

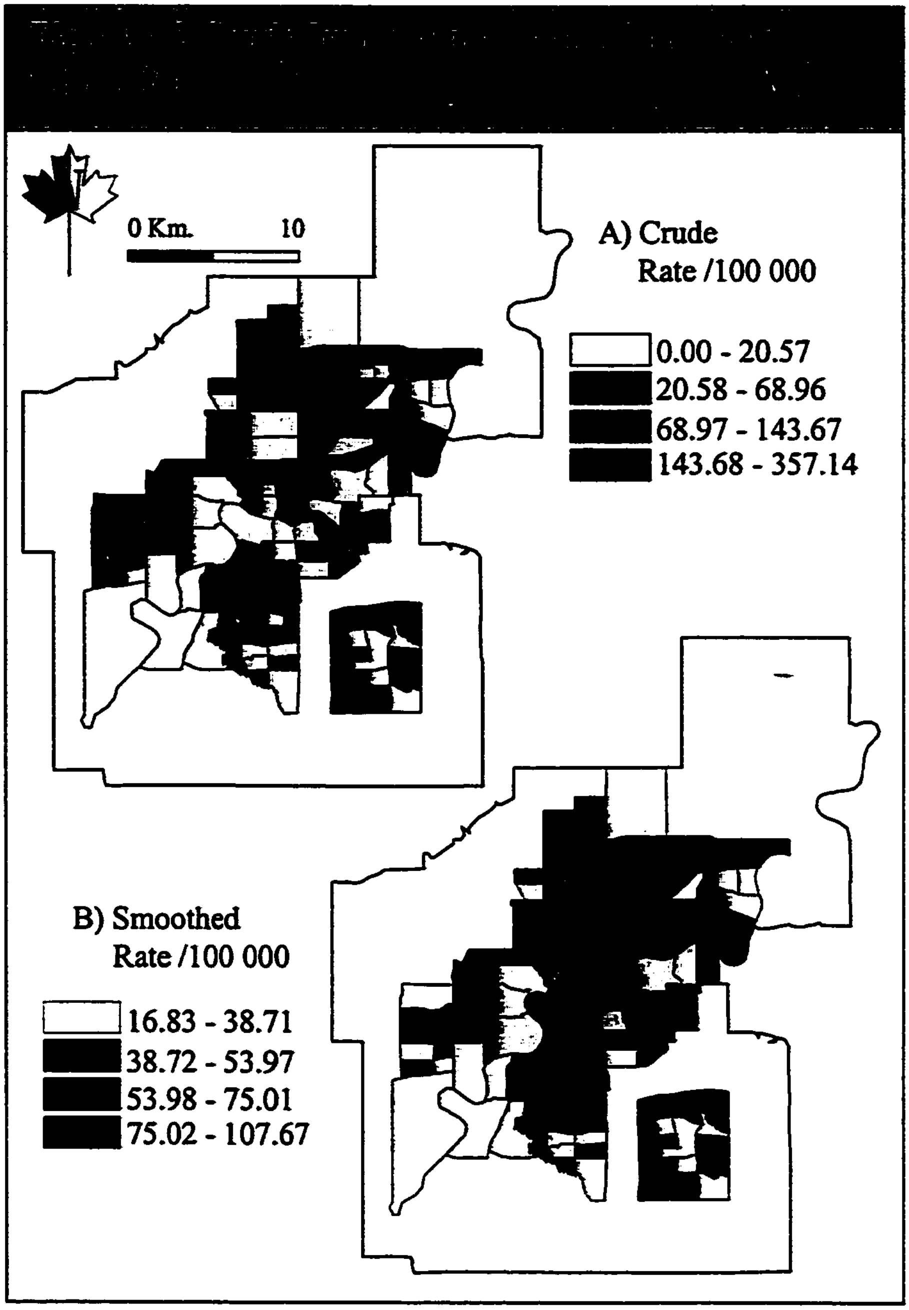
4.1 Univariate and Bivariate Analysis

4.1.1 Child Pedestrian Injury in Edmonton

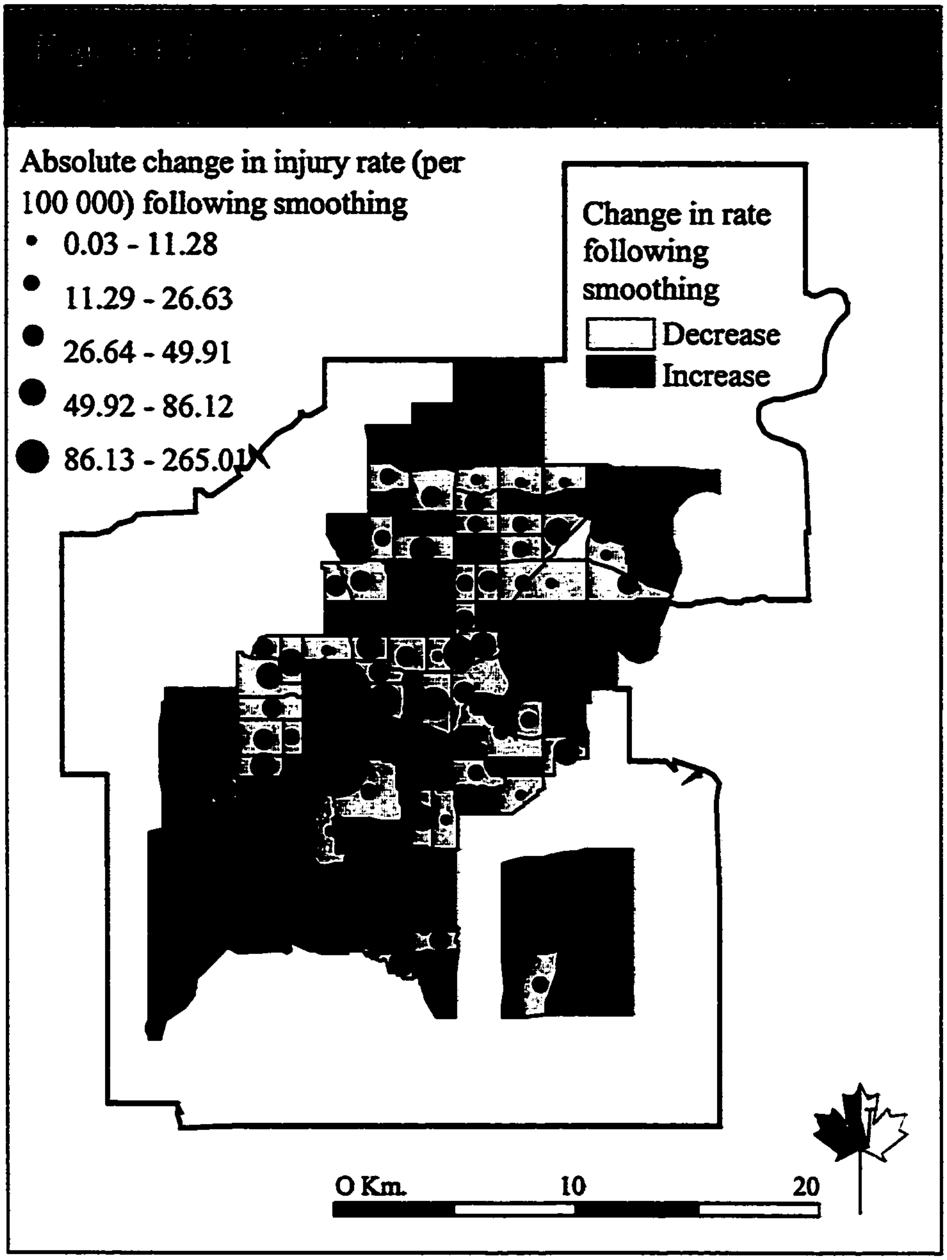
The largest population of children (0-15) in the study area is located in the outlying regions of the city (Figure 4.1). Within the study area, a total of 258 cases (0-15 years) were identified. Cases are geographically referenced based on postal code. Records with missing or incomplete postal code data are not included in this count. Within the database (which covers the whole Capital Health Region), a total of 22 child pedestrian injury records have missing postal code information and therefore the case count is probably an underestimate of child pedestrian injury frequency in the study area. The rate of emergency-room reported child pedestrian injury in the study area during this study period is 50.23 cases per 100 000 persons. In comparison maps of the crude and smoothed incidence rates of child pedestrian injury over the study period, the map of smoothed rates shows considerably less spatial variability than the crude rate map (Figure 4.2). In this four-year period, several of the most populated suburban census tracts in North and Southwest Edmonton recorded no injuries.

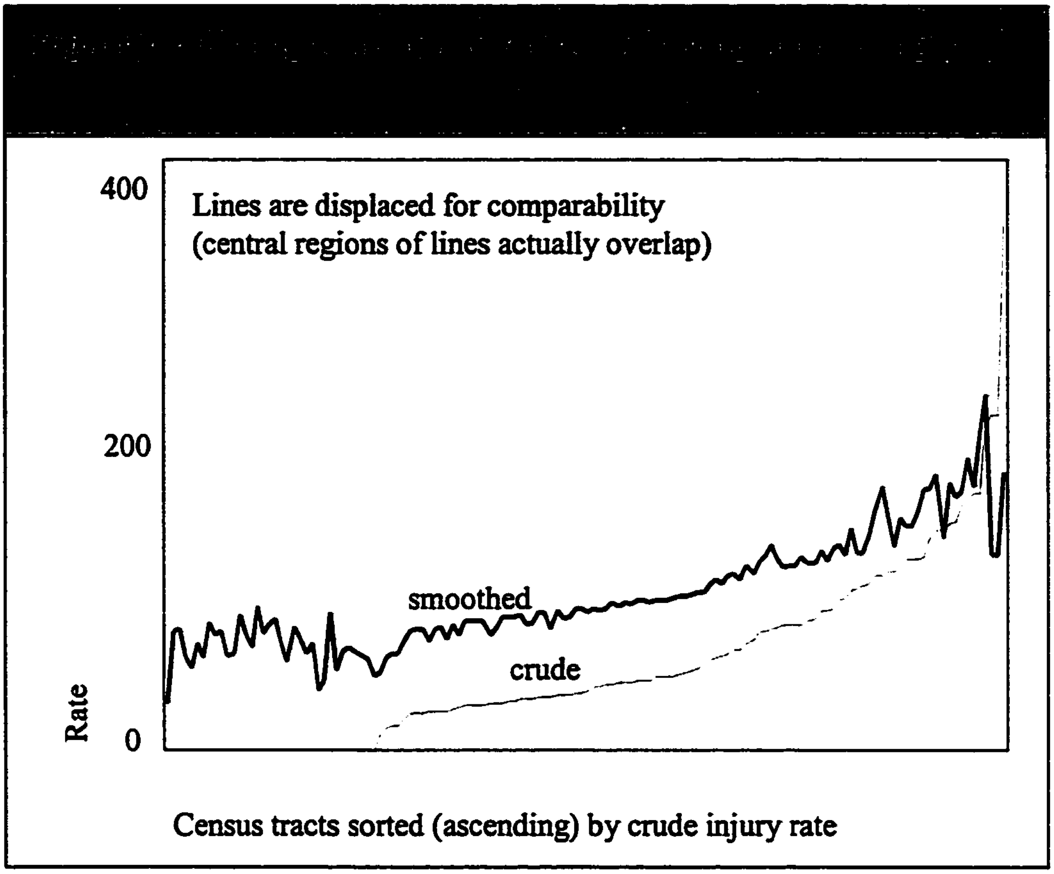
The greatest absolute change in injury rates after smoothing is in areas where the population of children is small (Figure 4.3). In these regions, rates drop. Moderate increases in injury rates occur in outlying areas where no injuries had occurred in the period. Figure 4.4 shows how the crude incidence rates and empirical Bayes estimates vary over space. The peak areas on the right correspond with the darkest areas on rate maps in Figure 4.2. Although the empirical Bayes estimates show a similar overall pattern as the crude rates, several features of the graph are of notable interest. First,


Population of Children
110 - 510
511 - 840
841 - 1310
1311 - 1915
1916 - 3010
0 Km.
10
20



0 Km.
10
A) Crude
Rate /100 000
0.00 - 20.57
20.58 - 68.96
68.97 - 143.67
143.68 - 357.14
B) Smoothed
Rate /100 000
16.83 - 38.71
38.72 - 53.97
53.98 - 75.01
75.02 - 107.67


Absolute change in injury rate (per 100 000) following smoothing
0.03 - 11.28
11.29 - 26.63
26.64 - 49.91
49.92 - 86.12
86.13 - 265.01
Change in rate following smoothing
Decrease
Increase
O Km.
10
20



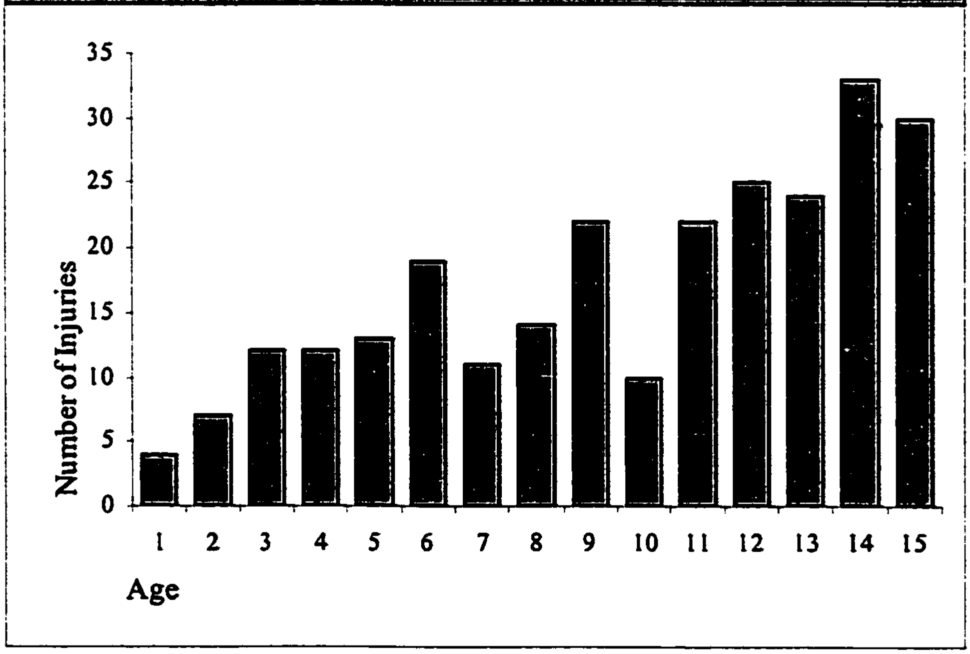
400
200
0
Rate
Lines are displaced for comparability
(central regions of lines actually overlap)
smoothed
crude
Census tracts sorted (ascending) by crude injury rate

estimates are made in census tracts where the crude incidence rate is zero. This strongly affects the description of events in space; an observation of "0" events in a census tract with a large population does not have the same meaning as a "0" in a census tract with a small population. Empirical Bayes estimates give an indication of risk in areas where no events occurred by taking into account the population of the area, the properties of the distribution, and the overall incidence rate. Second, the range of estimates is smaller. The empirical Bayes estimates have collapsed maximum and minimum values, although the general shape of the distributions is similar. Finally, the extreme values associated with the lowest population census tracts have been dragged towards the overall rate.

The demographic characteristics of the cases show several interesting patterns. Although the age group identified for study consists of all children under 16 years of age, no children less than one year of age were recorded in the data set. The youngest child reported in the data was one year and four months. The distribution of child pedestrian injury by age shows a variable pattern, with an increasing trend with age (Figure 4.5). The two spikes at ages six and nine break up what would appear to be a stabilisation from age three to an increasing trend starting at age eleven. Overall there is a general pattern of increase in number of injuries in progressively older age groups. The incidence ratio of injury in three age strata (0-4, 5-10, 11-15) is calculated using the population counts available in the census (Table 4.1). These ratios show the same trend as injury counts, with frequency of injury increasing with age. The incidence ratios also show that the largest proportion of injuries occurred in the oldest age stratum.

Age Group	Strata Counts						Incidence Ratio[1]		
	Male	Pop.	Female	Pop.	Total	Pop.	Male	Female	Total
0 to 4	26	20225	9	19830	35	40055	63.98	22.59	43.49
5 to 9	53	21005	36	20050	89	41055	125.58	89.36	107.9
10 to 15	70	24265	64	22975	134	47240	143.58	138.64	141
0 to 15	149	65495	109	62855	258	128350	113.23	86.31	100

[1] Incidence ratio equals observed cases per stratum divided by expected cases per stratum x 100 (expected cases = stratum population * overall rate)



Boys were involved in more collisions than girls; however, the gender differences between age strata show a variable pattern. In the first age stratum, boys receive by far the majority of injuries, with a smaller majority in the middle stratum. The pattern changes in the oldest stratum where almost as many girls were injured as boys. The highest incidence ratio is among boys in the 10 to 15 year-old stratum; the lowest among girls in the 0 to 4 year-old stratum (Table 4.1). There is a six-fold increase in the incidence ratio among girls from the youngest to oldest age stratum. This is a more drastic increase than among boys, for whom the incidence ratio merely doubles from the youngest to the oldest strata.

In a monthly breakdown of injury events by age group, the youngest age group shows the greatest degree of monthly variability with the fewest number of events occurring in the winter months, and more injury events occurring in the summer (Figure 4.6). The two older age groups show a more constant monthly frequency of injury, with the oldest group experiencing roughly the same number of injuries each month. The aggregate pattern (all three age groups combined) shows that with the exception of January and March, the monthly frequency of injury is fairly constant. Injuries by age group also exhibits a variable pattern over the days of the week, with Mondays and Fridays having the highest frequencies, and Sundays having the lowest (Figure 4.7). There is a notable mid-week dip in the number of injury events for the first age stratum, while the two older strata possess a less variable pattern. The first age stratum also reveals a higher proportion of injury on Saturdays than the other groups. A close look at the monthly pattern of injuries against the days of week shows that the frequency of accidents on a given day is dependent on the month in which the accident occurs (Figure

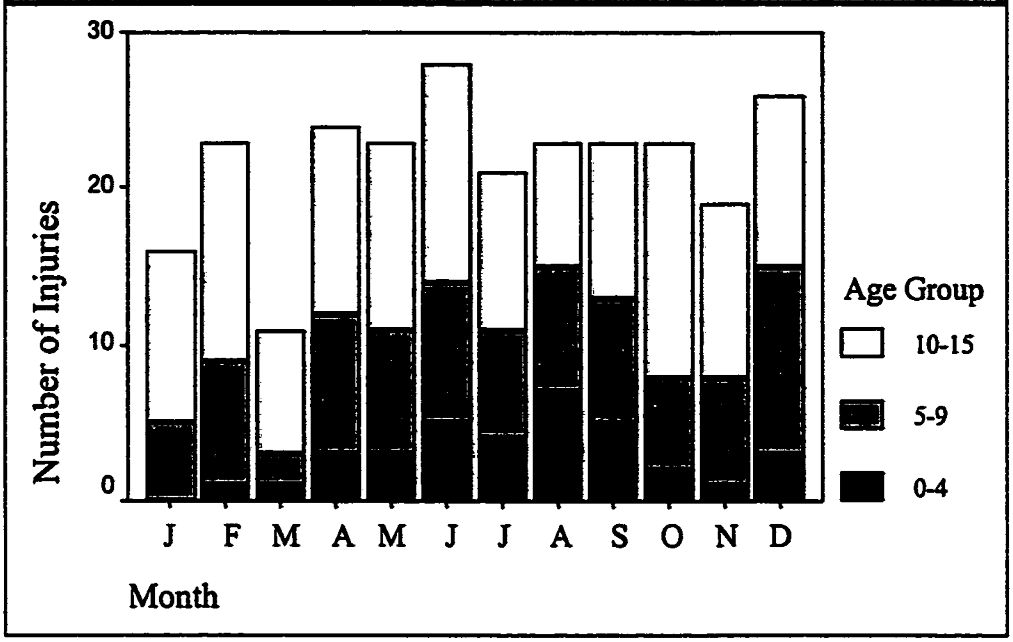
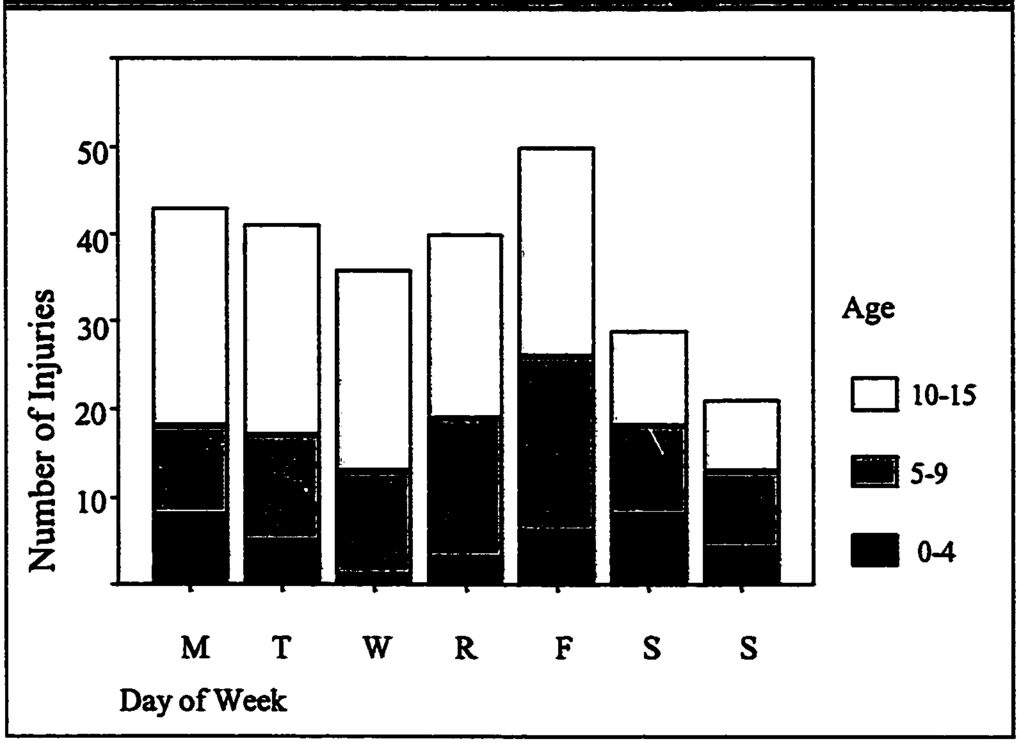


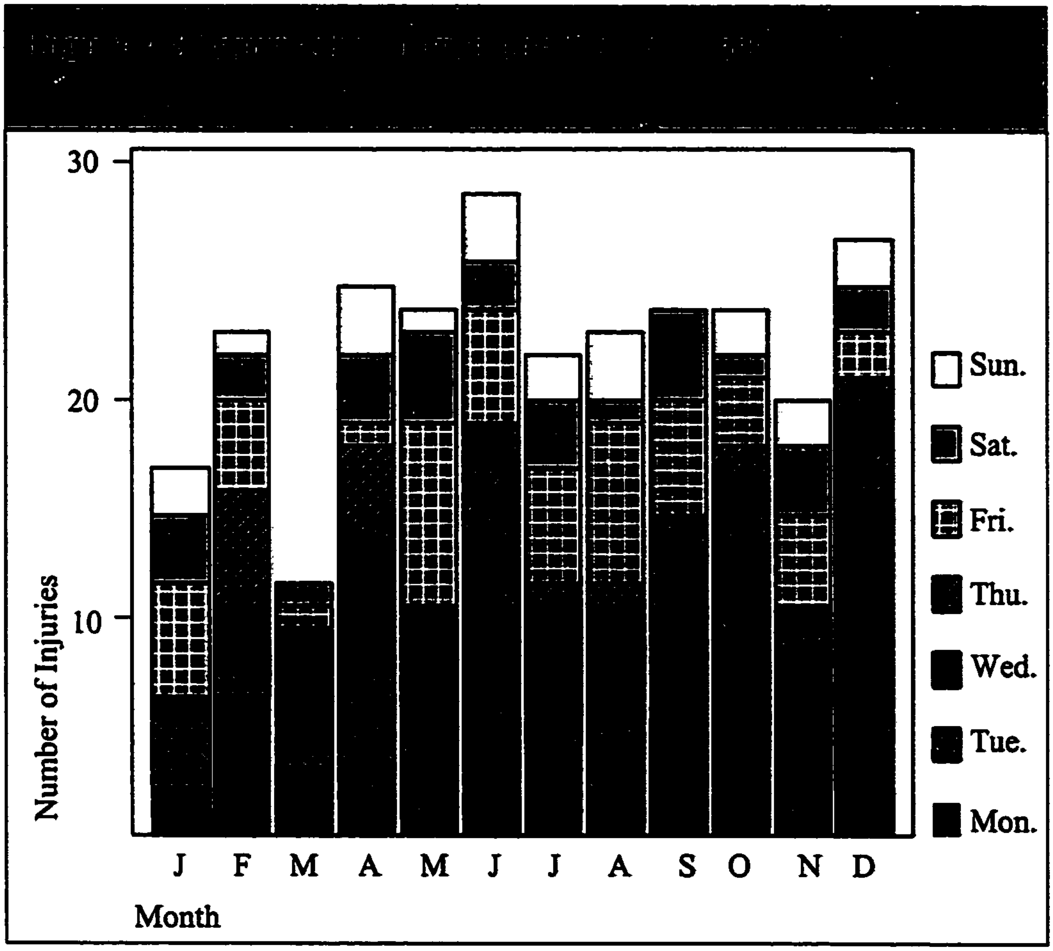
Figure 4. Injuries by Day and Age Group

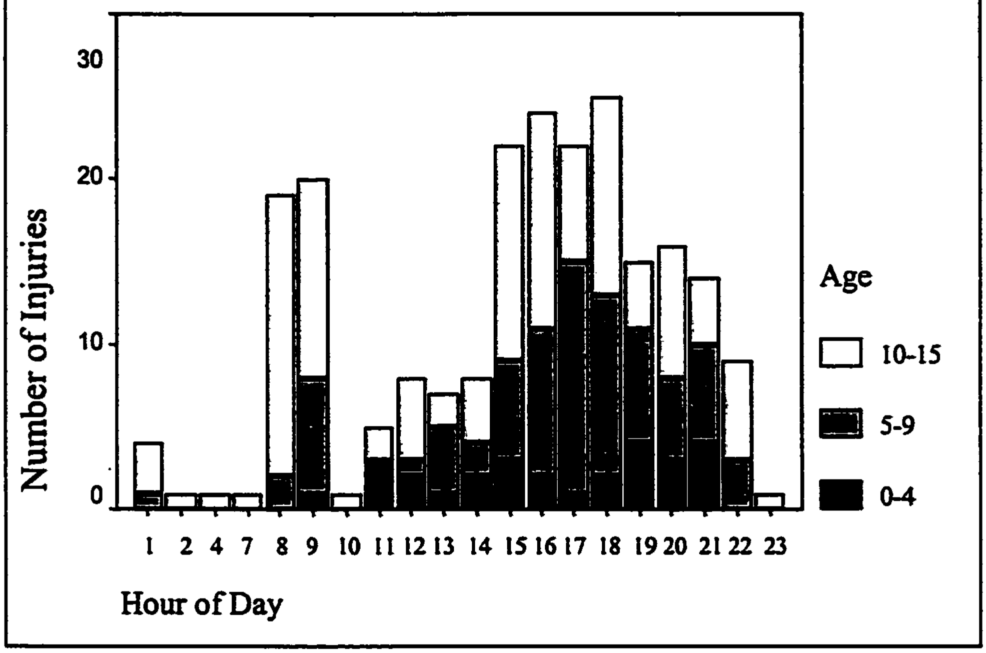


4.8). In February and October, Mondays are frequently the days upon which collisions occur. In March, April and December, collisions appear to occur less frequently on Friday than on other months.

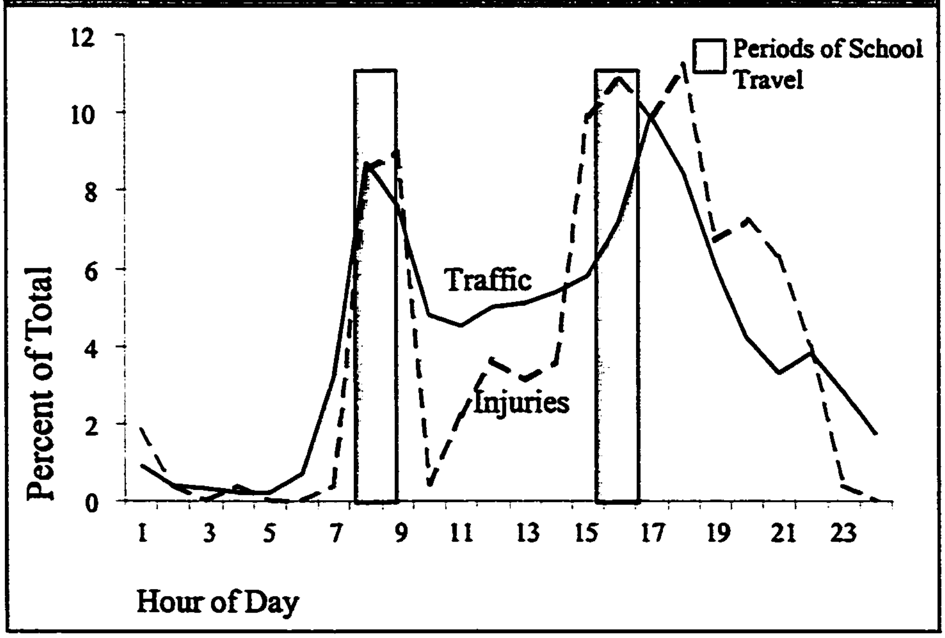
Time of injury by age group reveals a notable pattern (Figure 4.9). Numbers along the horizontal axis represent the hours at which the case was reported to the emergency room, and do not provide a perfect representation of the times at which a collision event occurred. For many of the cases, time of admittance was recorded as "0.00". These cases are not included in the graph since they most likely represent unknown time of admittance data and not admittance at midnight. The two oldest age strata possess the largest degree of variation in the hourly frequency of injury, with the highest frequencies in the mid-morning hours and mid-afternoon. The bi-modality of the graph is mainly the result of high collision frequency in the oldest age strata. The oldest stratum reveals the widest range of times of injury, with events occurring in morning and late night hours. The middle stratum's peak period of injury occurs in the mid afternoon, with a smaller but noticeable peak in the morning. The youngest age stratum shows a relatively constant spread of injury over daylight hours.

A graph of time of injury and peak traffic hours is presented on Figure 4.10. School opening and closing hours are indicated by vertical bands. In Edmonton, each school sets its own hours of operation, but for the majority of schools, classes begin between 8:30 and 9:00 a.m. and end between 3:30 and 4:00 p.m. The vertical bands represent these two time periods. The temporal correlation between these three factors is strong, especially during the morning hours. However, since the times of injury


30
20
10
Number of Injuries
J F M A M J J A S O N D
Month
Sun.
Sat.
Fri.
Thu.
Wed.
Tue.
Mon.



30
20
10
0
Number of Injuries
Age
10-15
5-9
0-4
1 2 4 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Hour of Day



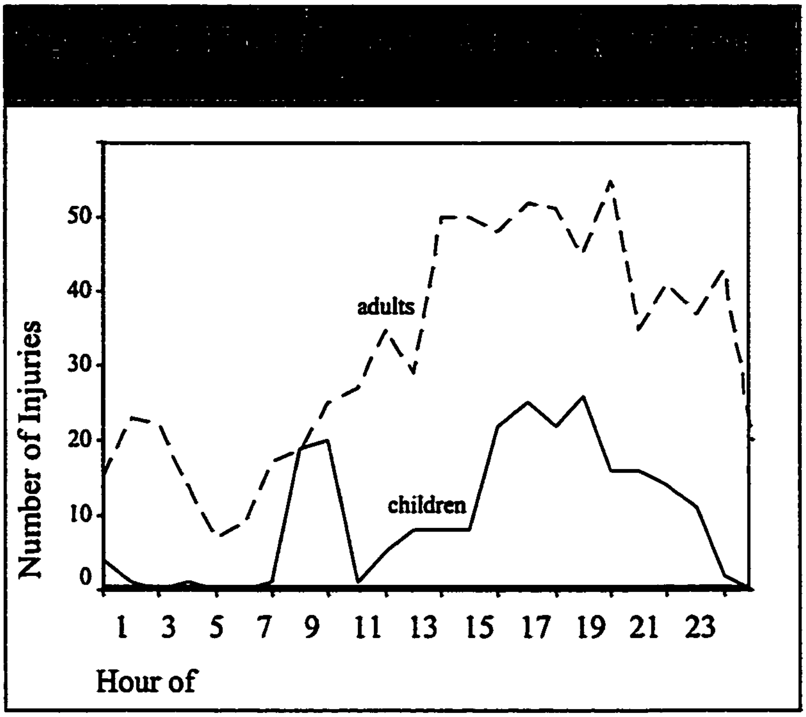
12
10
8
6
4
2
0
Percent of Total
Periods of School Travel
Traffic
Injuries
1 3 5 7 9 11 13 15 17 19 21 23
Hour of Day

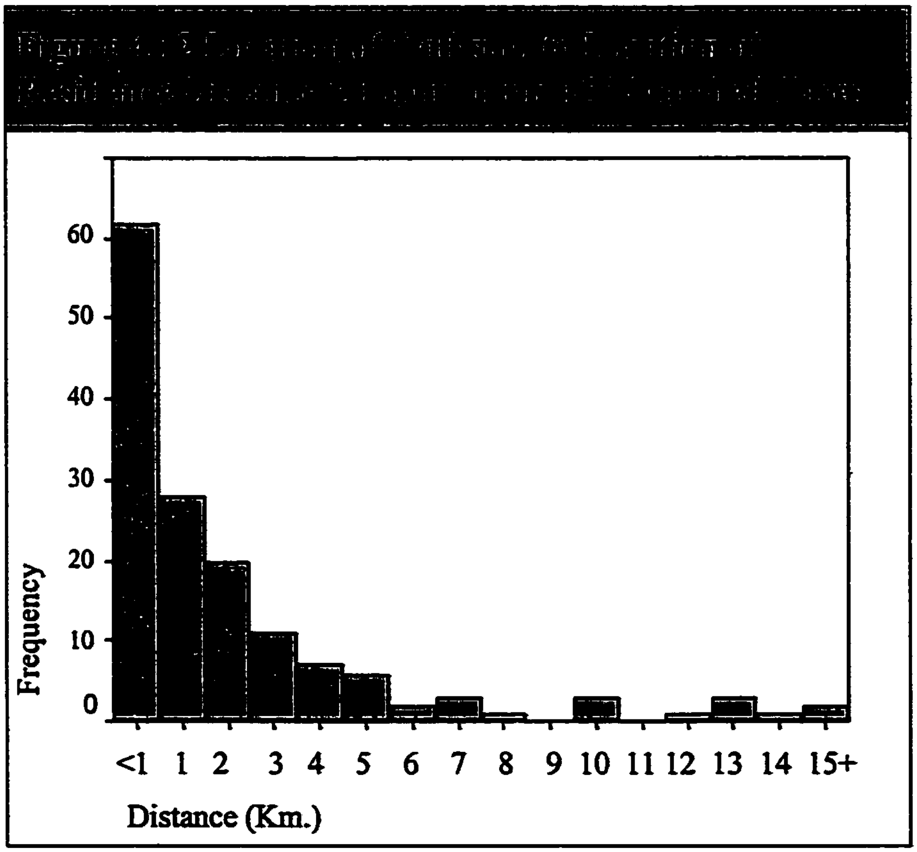
occurrence represent hour of admittance to an emergency room and not the exact time of collision, conclusions must be drawn cautiously.

In the four year period, 890 persons over 15 years of age were involved in a pedestrian-automobile collision in the study area. Again this may be a slight underestimate since cases without a recorded geographic identifier were not included. The distribution of injury by hour of day among persons over 15 years of age is compared to the distribution of injury among children age 15 or under in Figure 4.11. The mid-afternoon is a peak period of injury occurrence for both adults and children. However, a significant difference between the groups is observed in the earlier times of the day—children are frequently injured in the 8:00 to 9:00 a.m. range, whereas adults are not. There is also a peak period of injury for adults in the early morning hours that is not paralleled among children.

4.1.2 Collision locations

Two measures can give an indication of the spatial relationship between residential locations and sites of collisions: 1) distance (Euclidean or according to the street network) and 2) whether or not the collision events occurred in the tracts of residence. Data on collision location were obtained by matching case information with a database of collision data maintained by Edmonton's Traffic and Planning department. Records were linked on age, date of collision occurrence and gender[1]. One-hundred and fifty-seven records from the case file were successfully linked to the traffic department data, and the distribution of distance is described in Figure 4.12. This graph gives a strong indication that children are involved in these collision events near to home.


adults
children
Number of Injuries
0
10
20
30
40
50
1 3 5 7 9 11 13 15 17 19 21 23
Hour of


Frequency
0
10
20
30
40
50
60
<1 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15+
Distance (Km.)

Nearly 123 (78%) of the matched cases were injured within three kilometres of home, and 109 (69%) cases were injured within two kilometres However, of the 157 matched cases, only 66 (42%) were injured in their census tract of residence.

Collision location to residential location distance is compared across four age strata (Tables 4.2 and 4.3). Although the oldest and youngest age groups appear to have large average distances, statistical assessment indicates that the differences are not significant. The 2 x 4 contingency table shows a similar trend. The two middle age groups show shorter distances when the data are grouped by distance categories (less than/greater than 2 km). Collision location to residential location distance does not vary significantly between boys and girls (χ^2,1 d.f.=0.118,(p = 0.731))

Table 4.2 Distance (Km.) by age group

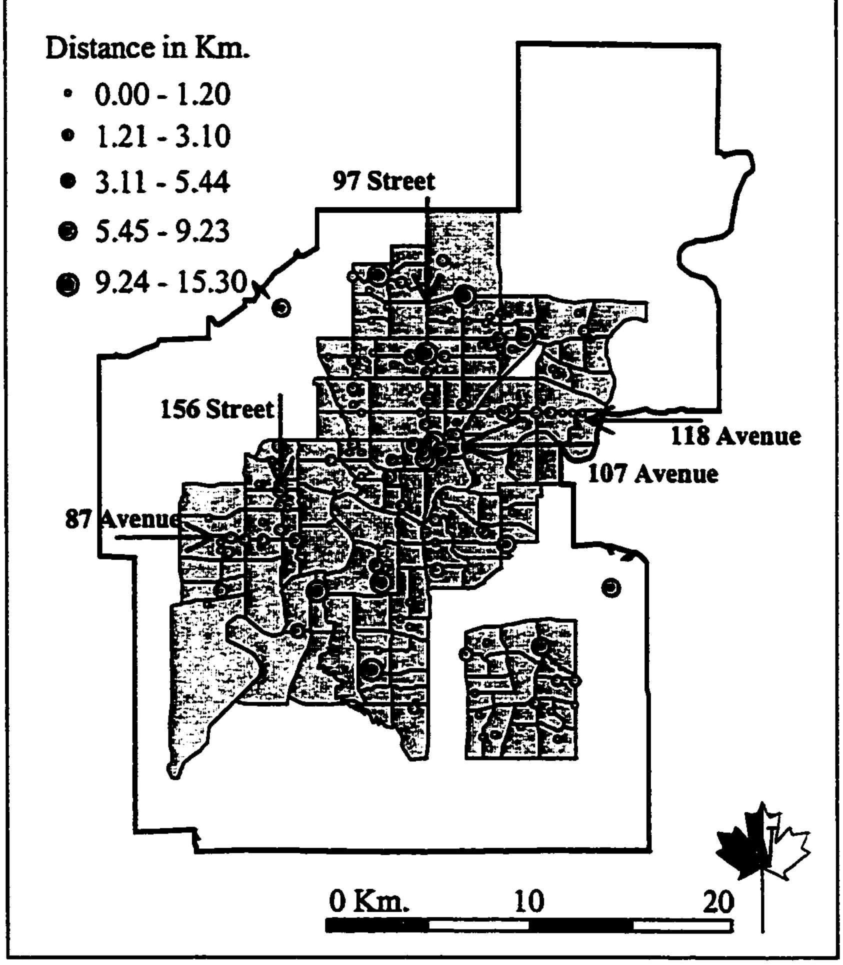
Age Group	N	Mean	SD	Median
0 to 4	17	2.74	3.64	1.69
5 to 7	31	1.88	2.67	0.67
8 to 10	29	1.83	3.55	0.37
11 to 15	80	2.27	3.27	0.92
	F (152,4 d.f)=1.820 (p=0.295)			
Age categories based on Preston, 1972				

Figure 4.13 presents the adjusted rate map overlain by a graduated point map. The points represent locations at which collisions occurred. The size of the points corresponds to distance from home—larger points indicate that the child was farther from

Table 4.3 Distance Greater than/Less than 2km

	Age Group				
	0 to 4	5 to 7	8 to 10	11 to 15	Total
<= 2km	9	22	22	56	123
> 2km	8	9	7	24	34
Total	17	31	29	80	157
	χ^2,3 d.f.=2.790 (p=0.425)				

Figure 4.13 Collision Sites by Distance from Home (n=157)



home when involved in a collision. Although the map suggests that collisions occurring closer to the city core tend to occur farther from home than collisions that take place in more peripheral areas, the values do not correlate significantly ($r=-0.068$, $p=0.394$). It is apparent from this map that several Edmonton arterial segments are common sites of collision occurrence. In particular, 118 Avenue (East Edmonton), 107 Avenue (Central Edmonton), 87 Avenue (West Edmonton), 156 Street (West Edmonton) and 97 Street (Central/North Edmonton) are arterials upon which a number of collision incidents occurred over the study period. Ninety-one of the 157 collisions took place on a major arterial. Eleven of the 91 took place at the intersection of two major arterials. Remaining cases occurred on either residential or collector roads, or an arterial for which no volume measurement was obtained. Of the 91 cases that occurred on a major arterial, most occurred on arterials with an average daily traffic volume between 10 000 and 20 000 vehicles per day (Table 4.4).

Table 4.4 Volume of Arterials at Collision Sites

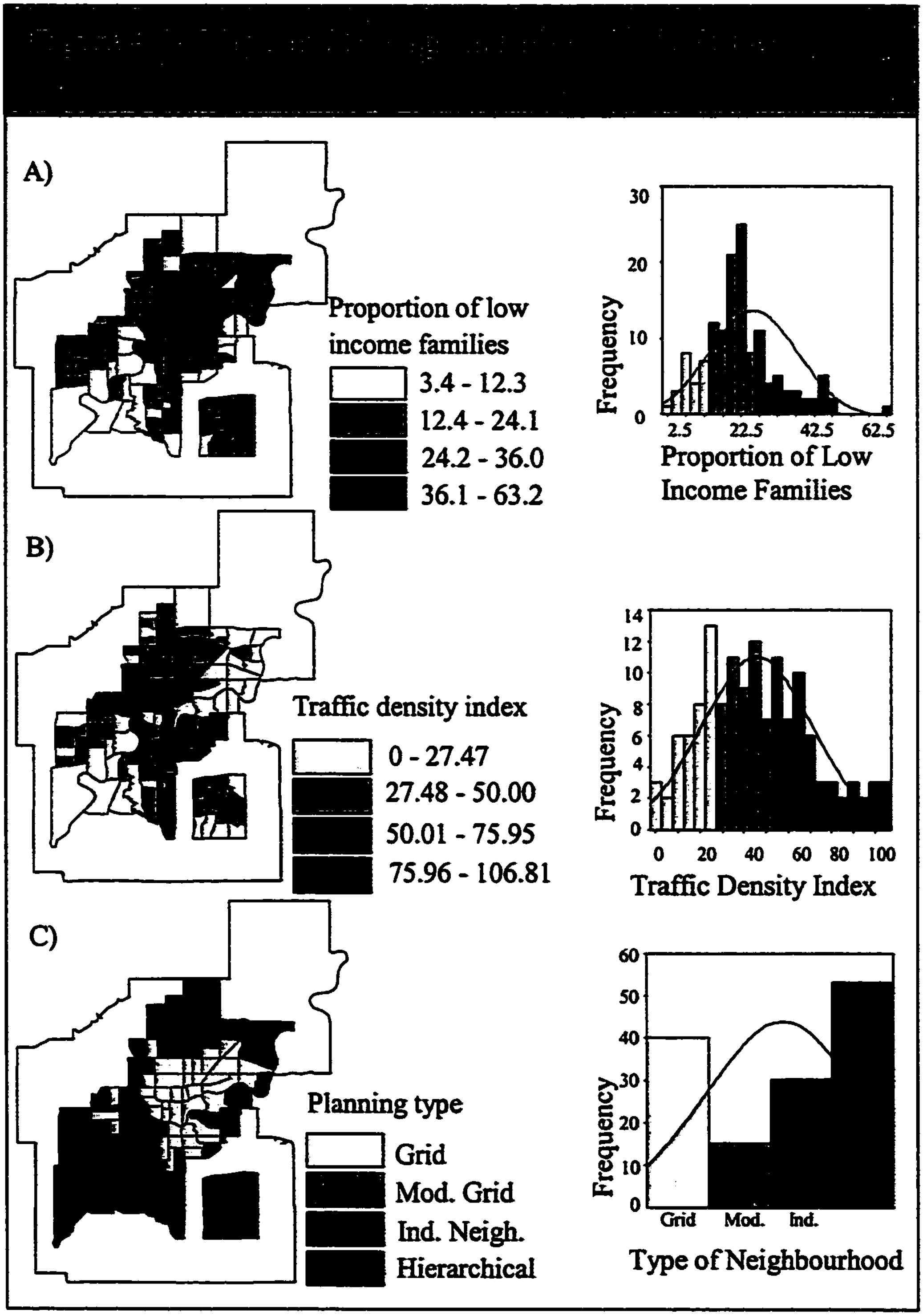
Volume (mean daily vehicles)	Collisions
Less than 10 000	15
10-20 000	55
20-30 000	29
30-40 000	9
40-50 000	5
50 000 and up	0

There are 113 collisions in this table because 11 cases occurred at an intersection of two arterials, both of which receive a measurement of volume in this table.

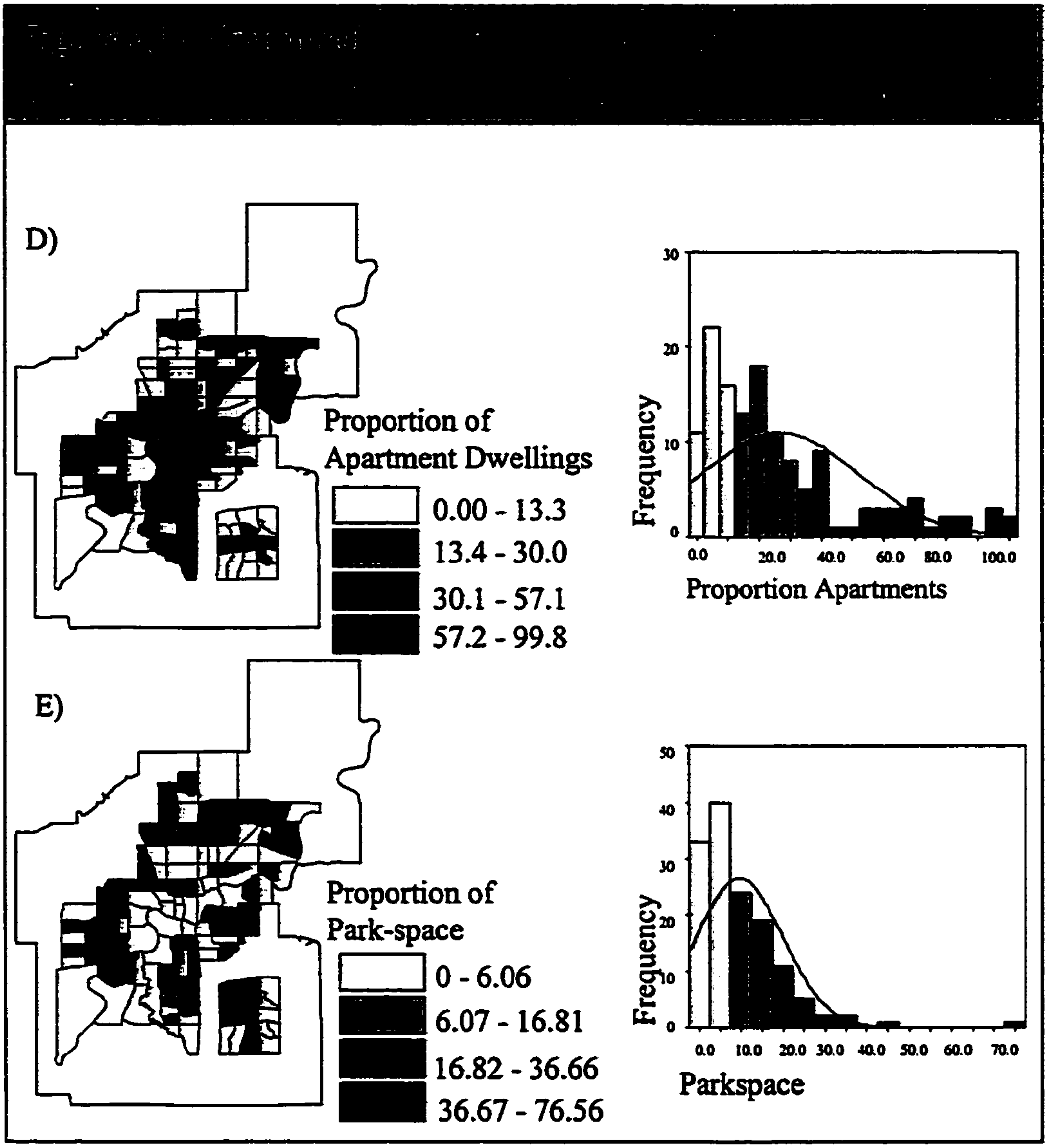
4.1.3 Independent Variables

Proportion of low income families and road density possess the distributions closest to the normal, while proportion of neighbourhood park-space and proportion of apartment dwellings have a positive skew (Figure 4.14 A-E). The corresponding maps give a spatial representation of the range of variability observed in the histograms. The smallest proportion of low income families live in the Southwest portions of the city. The largest proportion of low income families live in Central Edmonton, with pockets in both the East and West ends of the city (Figure 4.14 A). Road density has an approximately normal distribution, with various small clusters of high and low density values throughout the city (Figure 4.14 B). The map of planning type shows the least amount of spatial variation and the most notable pattern of clustering. Planning types cluster in form, from higher order planning structures in Southeast and Southwest Edmonton, to grid-planned in the central areas of the city. As can be judged from the histogram, grid and hierarchically planned neighbourhoods are the dominating planning structures at the level of the census tract (Figure 4.14 C). Percent low family income and the road density index also show reasonably strong spatial clustering patterns. The distribution of apartment dwellings is fairly variable throughout the city, with an important exception in the city centre, where the proportion of apartment dwellings is quite high (Figure 4.14 D). The proportion of park space has a similar skew; however, there is no noticeable spatial clustering (Figure 4.14 E).

Visual comparison of the maps indicates that Central areas are typically grid-planned neighbourhoods of lower income, higher traffic density and higher proportion of apartment dwellings. Tracts surrounding the River Valley, particularly in the South end



A)
Proportion of low income families
3.4 - 12.3
12.4 - 24.1
24.2 - 36.0
36.1 - 63.2
Frequency
Proportion of Low Income Families
B)
Traffic density index
0 - 27.47
27.48 - 50.00
50.01 - 75.95
75.96 - 106.81
Frequency
Traffic Density Index
C)
Planning type
Grid
Mod. Grid
Ind. Neigh.
Hierarchical
Frequency
Grid
Mod.
Ind.
Type of Neighbourhood


D)
Proportion of
Apartment Dwellings
0.00 - 13.3
13.4 - 30.0
30.1 - 57.1
57.2 - 99.8
Frequency
Proportion Apartments
E)
Proportion of
Park-space
0 - 6.06
6.07 - 16.81
16.82 - 36.66
36.67 - 76.56
Frequency
Parkspace

of the city, possess the opposite pattern—fewer low income families, lower traffic densities and fewer apartments. A North-South contrast is evident in the spatial variation of most study variables. Using the geometric centre of the study area as a dividing point, the North End of Edmonton has a higher rate of injury, higher percent low income families and a higher density of traffic.

4.1.4 Correlation Matrices

With the exception of the proportion of parkspace (parks), all variables show a correlation to injury rate as well as a degree of inter-correlation between the independent variables (Table 4.5 A). The largest correlation coefficients are found between proportion of low income families (low_inc) and planning type (plan_typ), proportion of low income families and proportion of apartments (apt), and planning type and the smoothed child pedestrian injury incidence rate (inj_rate). Overall, the range of statistically significant coefficients is fairly small—absolute values of the coefficients range from a low of 0.206 to a maximum of 0.454. Although the degree of collinearity is difficult to confirm from a correlation matrix, these results suggest that multicollinearity may present a problem in the model as correlation between independent variables is of a similar magnitude to correlations between independent variables and injury rate.

Proportion of low income families shows a positive relationship with all of the variables except for planning type. An inspection of the maps would suggest that this is the case—areas in which fewer low income families live seem to coincide with areas of higher order planning. Alternatively, the lower income areas of the city—in particular, Central Edmonton—coincide with the unplanned areas. Proportion of park space shows a statistically significant correlation with planning type only.

Table 4.5 A Product Moment Correlation

	inj_rate	apt	roadden	plan_typ	parks	low_inc
inj_rate	1					
lnapt	0.206*	1				
roadden	0.381*	0.274*	1			
plan_typ	-0.410*	-0.343*	-0.343*	1		
lnparks	0.051	-0.038	0.128	0.208*	1	
low_inc	0.398*	0.454*	0.337*	-0.421*	0.025	1

Table 4.5 B Spearman's Rank –Planning Type

	R_s	#obs	
inj_rate	-0.400*	138	* indicates statistically Significant at 0.05 level
apt	-0.332*	133	
roadden	-0.359*	138	
parks	0.166	112	
low_inc	-0.337*	138	

Since planning type is not a continuous variable, but is ordered into four categories, the Spearman's rank is a more appropriate test of concomitant variation (table 4.5 B). Three differences are worthy of note. First, the correlation between planning type and the rate of injury is marginally weaker (ρ=-0.400) than calculated using the product moment coefficient. Second, the correlation between planning type and low family income is weaker (ρ=-0.337). Finally, proportion of park space went from a moderate association to a non-significant association. Since planning type should not be used in calculations of product moment correlation (it is neither continuous nor normally distributed), these values probably give a better indication of correlation.

4.1.5 Bivariate Regression

Bivariate regression models of the smoothed injury rate were performed for all the independent variables individually (Table 4.6). The statistical significance of each bivariate coefficient was used to indicate which independent variables are appropriate for inclusion in further analysis. For each regression run, the confidence intervals of the regression coefficients exclude the value of zero.

Table 4.6 Bivariate Regression: Smoothed Rate Modeled Against Independent Variables

Variable	COEF	SE	95% Conf. Int.(L,H)		T	P val.	R^2
Low Income	0.695	0.138	0.423	0.967	5.053	0.000	0.152
Constant	37.060	3.403	30.331	43.788	10.892	0.000	
Apartments	0.148	0.060	0.029	0.268	2.453	0.015	0.035
Constant	48.743	2.207	44.378	53.108	22.083	0.000	
Road Density	0.275	0.057	0.162	0.388	4.808	0.000	0.139
Constant	40.073	2.987	34.167	45.980	13.417	0.000	
Planning Type	-5.886	1.123	-8.107	-3.665	-5.240	0.000	0.162
Constant	68.571	3.337	61.972	75.170	20.548	0.000	

In no case do the adjusted R^2 values exceed 0.20 and the shapes of the scatterplots do not suggest strong relationships (Figure 4.15). The scatterplots of the smoothed rate offer an interpretive advantage over the scatterplots of the raw rates since they do not show the row of zero values typical in unadjusted plots of rare data (Figure 4.16). Assuming the smoothing process is valid, then this is a far more desirable method for presenting the data, since zero values give limited information about risk, and at the very least make meaningful interpretation of the data more difficult.

Often scatterplots can help identify non-linear relationships between variables. In this case, however, the broadly shaped distribution of points makes it difficult to identify

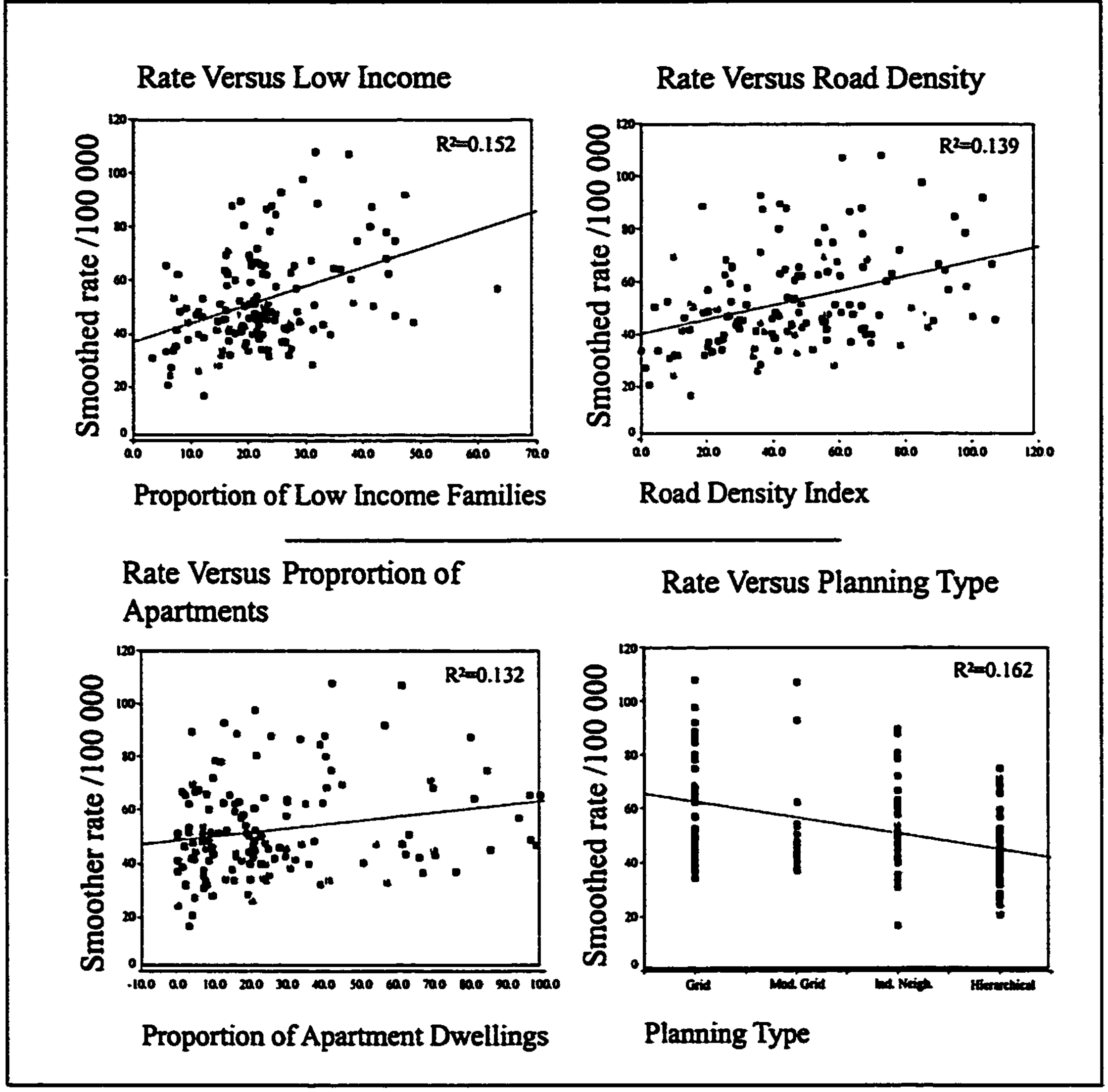
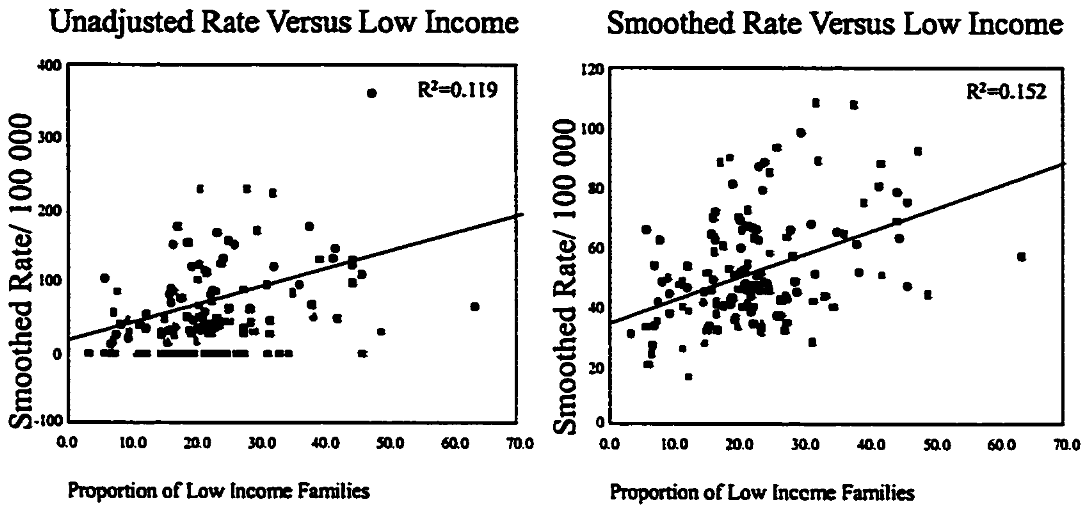

Rate Versus Low Income
Smoothed rate /100 000
R²=0.152
Proportion of Low Income Families
Rate Versus Road Density
Smoothed rate /100 000
R²=0.139
Road Density Index
Rate Versus Proprortion of Apartments
Smoother rate /100 000
R²=0.132
Proportion of Apartment Dwellings
Rate Versus Planning Type
Smoothed rate /100 000
R²=0.162
Grid
Mod. Grid
Ind. Neigh.
Hierarchical
Planning Type

Figure 4.16 Smooth and Unsmoothed Child Pedestrian Injury Rates



any non-linear patterns if they did exist in these data. On the scatterplot of planning type, tracts characterised as having the modified grid planning form appear to break from the pattern with the remainder of the data. In particular, there is a gap between a cluster of points at the low and high rates. Although this may suggest some curious relationships in the model, this apparently anomalous series of observations is based on significantly fewer observations than present in the other planning groups.

Together, the correlation analysis and bivariate regression models indicate that four of the five variables of interest have an association with injury rate, and that there are no obvious non-linear relationships. The explained variance is low for each of the variables independently, and the multicollinearity among the independent variables indicates that one should not expect total explained variance in the multiple regression model to be high. The next stage of analysis includes a multivariate model of the smoothed injury rate based on the four variables that remain after bivariate and univariate analysis: proportion of low income families, proportion of apartment dwellings, planning type and road density.

4.2 Multivariate Analysis

The first stage of the modelling process is a fully inclusive model of variables that remained following the univariate and bivariate analyses. Formalised modelling procedures (eg. stepwise, backwards inclusion) were deemed inappropriate given the breadth of research in child pedestrian injury. Such procedures solve models based on statistical and computational criteria, and are better suited for exploratory research or models with numerous independent variables of unknown effects (Von Eye and Schuster, 1998). Although multivariate ecological-level models are rare in pedestrian injury, the importance of the variables of interest is supported by theory as discussed in Chapter 2.

Table 4.7 Multivariate Model – Four Predictors

Obs.	138				**F(4,133)**	12.690
R^2	0.276				**P val.**	0.000
Adj. R^2	0.254					
Variable	**Coeff.**	**SE**	**95% Conf.Int (L,H)**		**T**	**P val.**
Low Income	0.430	0.154	0.126	0.734	2.796	0.006
Apartments	-0.057	0.062	-0.180	0.067	-0.910	0.364
Road Density	0.166	0.058	0.051	0.282	2.850	0.005
Planning Type	-3.742	1.245	-6.204	-1.279	-3.005	0.003
Constant	46.991	6.258	34.614	59.369	7.509	0.000

4.2.1 Multiple Regression Model

Three of the four predictors in the model have statistically significant partialled regression coefficients (Table 4.7). Of equivalent importance is the value of the adjusted R^2 in the multivariate model. This value is considerably larger than the R^2 of any single bivariate regression run, which suggests that the multivariate model fits the data better than the bivariate models. Although most of the data are on similar numerical scales (road density, low income and proportion of apartments are roughly between 0 and 100), planning type is on a different numerical scale, and therefore the coefficients are not a good indicator of relative magnitude of effect on the rate of injury in census tracts.

Proportion of apartment dwellings does not have a statistically significant partialled regression coefficient. When this variable was removed from the model, the results are virtually identical to the four variable model (Table 4.8). As this model includes fewer variables (all of which have non-zero coefficients) it will be used for future steps of analysis.

Table 4.8 Multivariate Model - Three Predictors						
Obs.	138				**F(3,134)**	16.660
R^2	0.2717				**P val.**	0.000
Adj. R^2	0.255					
Variable	**Coeff.**	**SE**	**95% Conf.Int(L,H)**		**T**	**P val.**
Low Income	0.386	0.146	0.097	0.674	2.646	0.009
Road Density	0.162	0.058	0.047	0.277	2.786	0.006
Planning Type	-3.446	1.201	-5.821	-1.070	-2.869	0.005
Constant	45.866	6.130	33.741	57.990	7.482	0.000

Table 4.9 Comparison of One and Two-Predictor R^2 Against Three-Predictor Adjusted R^2 ($3PR^2$ =0.255)		
Variable Combination	**R^2**	**$3PR^2$-R^2**
Low Income	0.152	0.103
Low Income+Planning Type	0.218	0.037
Planning Type	0.162	0.093
Planning Type+Road Density	0.222	0.033
Road Density	0.139	0.116
Road Density+Low Income	0.216	0.040

Using the bivariate models as the initial guide, changes in planning type exacts the greatest influence, and road density the weakest (Table 4.9). In the three two-predictor models, the models with planning type included show the greatest associations with the rate of injury. The smallest change between the three-predictor and two predictor models comes with the addition of the low income variable to the planning type + road density model. An alternative method of judging relative importance of variables is to compare the values of the coefficients that are standardised to a mean of zero and a variance of one (standardised coefficients). Such standardised coefficients are calculated for planning type (-0.240), road density (0.224), and proportion of low income families (0.221). However, such standardisation procedures are most appropriate for normally distributed data (planning type violates this property), and even then only indicate "explained" variance, and not relative importance in any real-world application (Achen, 1982). Nevertheless, evidence from Table 4.9 and the standardised coefficients do give some indication of the relative role of each variable in the model.

4.3 Model Diagnostics

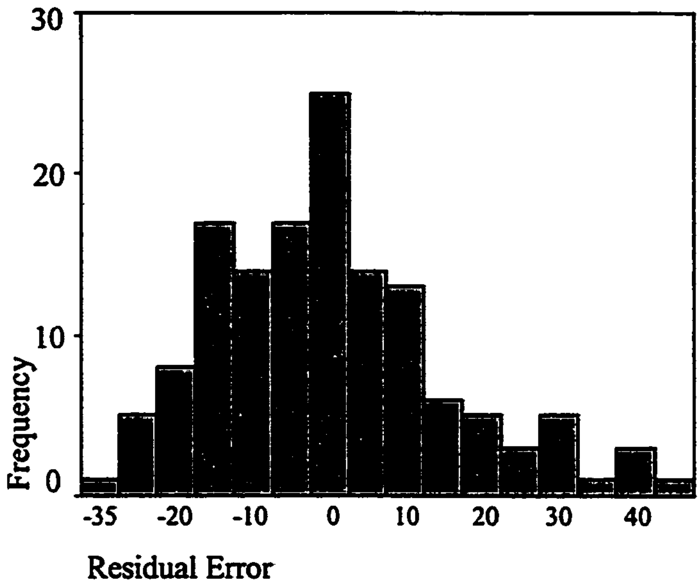
4.3.1 Residual Analysis

An important concern in any model building procedure is whether or not the methodological assumptions required in regression analysis are valid. As most assumptions are based on the behaviour of the residuals, these will be examined first. The frequency distribution of the residual is moderately skewed, but the observations are fairly evenly clumped about the centre of the histogram—suggesting little kurtosis (Figure 4.17). The properties of the residual error are also described Table 4.10. Tests for skewness/kurtosis indicate that the residuals depart slightly from the normal distribution. Although the mean of the residuals is approximately zero, the median is not.

Table 4.10 Residual Error	
Mean	0.000
Std.Dev.	15.361
Median	-1.745
Max.	43.805
Min.	-29.637
Skewness[1]	0.653
Kurtosis[2]	3.208
Ho: (Normal Distribution)	
P value	0.013

[1] Normal distribution, skewness= 0.00

[2] Normal distribution, kurtosis = 3.00

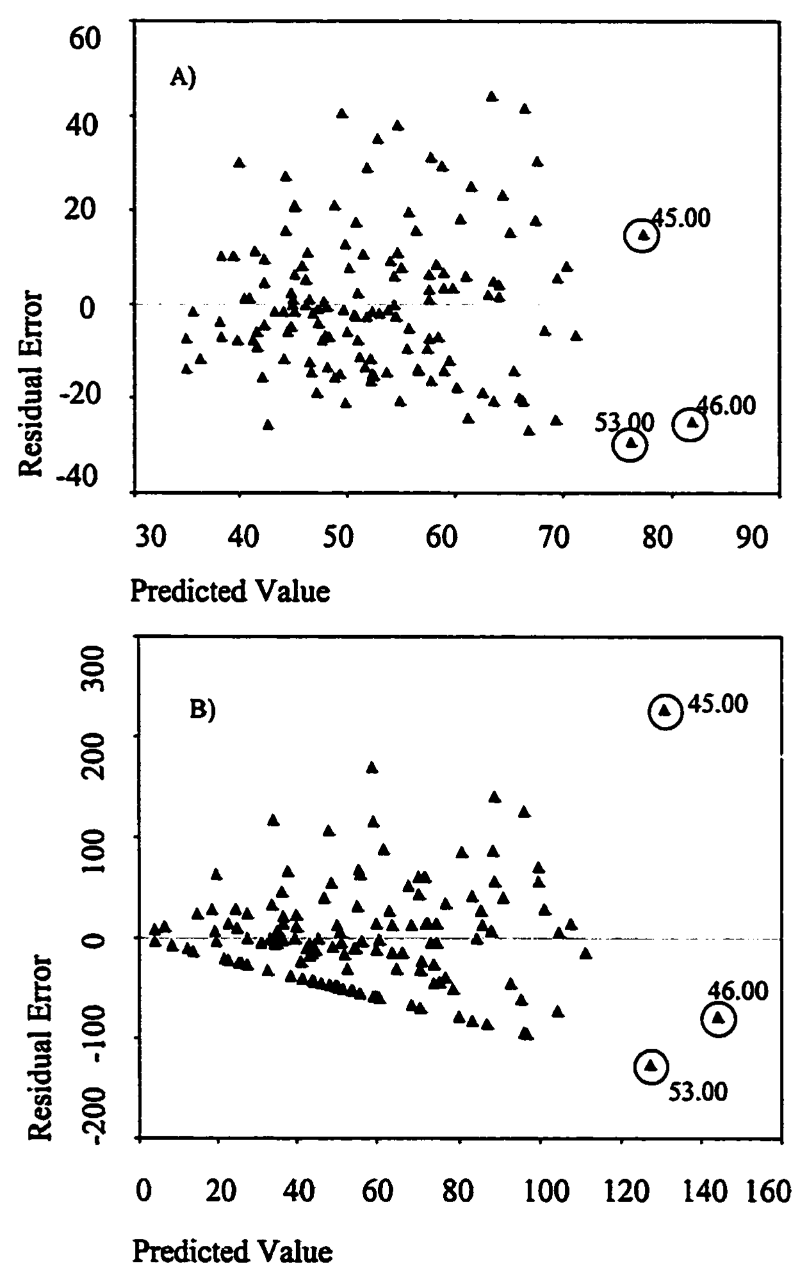

30
20
10
0
Frequency
-35
-20
-10
0
10
20
30
40
Residual Error

The Cook-Weisberg test for heteroscedasticity indicates that error is not constant, and may pose a problem in the model (Ho: Constant variance: 12.36, p=0.000). Residual versus predictor plots give a visual representation of this heteroscedasticity for smoothed (Figure 4.18 A) and crude (Figure 4.18 B). The conical shape is of common form—showing increased residual error with larger values of the predictors.

The effect of heteroscedasticity on the parameter estimates can be determined by calculating a more robust standard error and then comparing it to the normal standard errors. Using *Stata's robust* option, the robust estimators of variance did not differ considerably from the normal standard errors (Table 4.11). Under the robust option, the same parameter estimates reached statistical significance. This suggests that whatever heteroscedasticity may exist in the model error, the parameter estimates remain the best unbiased linear predictors of the population parameters. Further transformation of the dependent variable is therefore considered unnecessary.

Table 4.11 Standard Errors of Parameter Estimates

	Low Income	**Planning Type**	**Road Density**
Normal Standard Errors	0.146	1.201	0.058
Robust Standard Errors	0.150	1.123	0.057



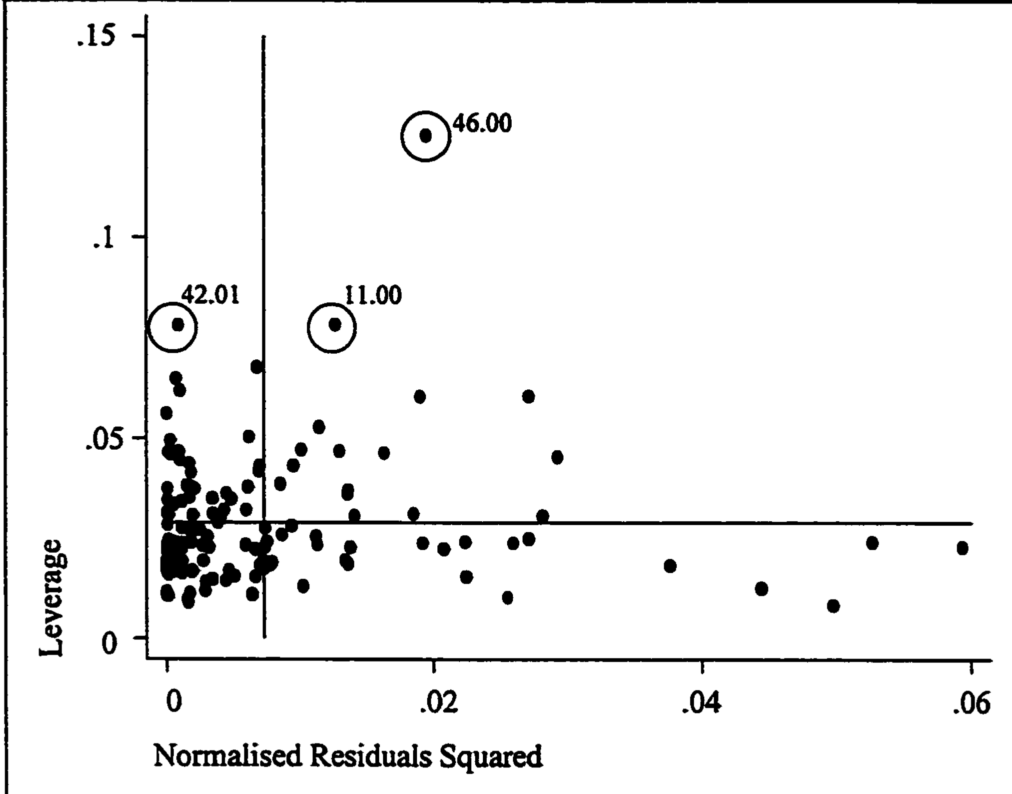
Predicted values for regression models of injury rate are largest in census tracts 45.00, 46.00 and 53.00. However, the predicted values are more extreme (relative to other predicted values) in the regression of crude rates.

4.3.2 Leverage Analysis

The scatterplots of residuals and predicted injury rates indicates that several observations may lie outside the main body of observations. (Figure 4.18). Tracts 45.00 and 46.00 are both low income areas in Central Edmonton. The leverage versus residual plot shows a cluster of points in the bottom left-hand corner, suggesting that most points have both similar leverage and residual error (Figure 4.19). Of most concern is any observation with large leverage and large residual error, but all points with considerably higher than average leverage should be noted. In this model, one point of high leverage clearly stands out, and several are worthy of note. Census tract 46.00 is an urban census tract located in Central Edmonton (population 0-15=780), census tract 42.01 is on the Eastern perimeter of the city (population 0-15=1055), and tract 11.00 is in South-Central Edmonton (population 0-15=765). The tracts have no outstanding similarities with respect to their demographic or physical characteristics, although tracts 46.00 and 42.01 both have high proportions of low income families. The effects of these observations on the parameter estimates can be tested by running the model without them (Table 4.12). The model R^2 increases, as do the parameter estimates. Most of the change is in the proportion of low income families parameter estimate. Nevertheless, I was not inclined to remove these observations from the model since in either case their effect is fairly small.

Table 4.12 Large Leverage Observations Removed

	R^2	Low Income	Planning Type	Road Density
N=138	0.255	0.386	-3.446	0.162
N=135	0.281	0.446	-3.529	0.180



Census tracts 46.00, 42.01 and 11.00 all show noticeably large leverage on the regression model. Of greatest concern is tract 46.00 which has one of the largest predicted values in the model.

4.3.3 Non-linearity, Non-additivity and Predictor Specification

Given the design of this study, there is no immediate reason to expect non-linear or non-additive relationships in the model. Although income measures frequently involved show non-linear associations in the regression models (Achen, 1982), in this particular case, the low income variable is a surrogate for many factors that may have complex associations with risk of injury—like car ownership and child behaviour. The functional form of these complex relationships would be difficult to detect given their mixture into a single variable. The bivariate scatterplots did not show any noticeable non-linear patterns. Regressions were performed using quadratic and cubic terms of each predictor. In the presence of a non-linear relationship, these terms should show stronger relationships with the dependent variable than normal predictors. Results give no indication of such non-linear variation (Appendix B, pt 1).

Multiplied predictor pairs ($X1_i * X2_i$) were added into the model (Appendix B, pt 2). In an additive model, the multiplied pairs should not improve model fit, and should not obtain statistically significant parameter estimates. In no cases do the multiplied predictors gain statistically-significant regression coefficients, or substantially modify the model R^2.

Ramsey's (1969) test for missing predictors can indicate that important model variables are missing. Results do not indicate that there are any missing variables in the model (Ho: missing variables ($F=1.38$, $p=0.25$)). However, missing variables may also be identified by looking for spatial patterns in the residual error. Obvious patterns in residuals over space may indicate a missing variable, or error in one of the existing variables.

4.3.4 Spatial Autocorrelation of Residuals

Table 4.13 shows the degree of spatial autocorrelation for the independent variables, injury count, crude injury rate smoothed injury rate and the model residuals using first order and adjusted first order neighbouring techniques. The results from the neighbour weighting techniques differ—particularly for the model residuals which show no spatial autocorrelation in the first order case and statistically significant spatial autocorrelation in the second order case. The smoothed rates show spatial autocorrelation in both cases, whereas injury rate and injury counts show positive spatial autocorrelation in the second case only. All independent variables exhibit moderate to strong spatial autocorrelation for both neighbouring methods.

Table 4.13 Spatial Autocorrelation

	First Order	**Second Order**
	(Ho: no SA)	
A) Study Variables	**Correlation**	**Correlation**
Low Income	0.485*	0.596*
Planning Type	0.763*	0.880*
Road Density	0.362*	0.489*
Injuries	0.061	0.247*
Injury Rate (unsmoothed)	0.079	0.243*
Injury Rate (smoothed)	0.131*	0.306*
Residuals	-0.022	0.175*
Fitted Values	0.660*	0.774*
B) Pseudo-Random Values, Normal Distribution		
Random1	0.053	0.188*
Random2	-0.054	0.113*
Random3	0.096*	0.274*
Random4	-0.010	0.169*
* Statistically Significant (p=0.05)		

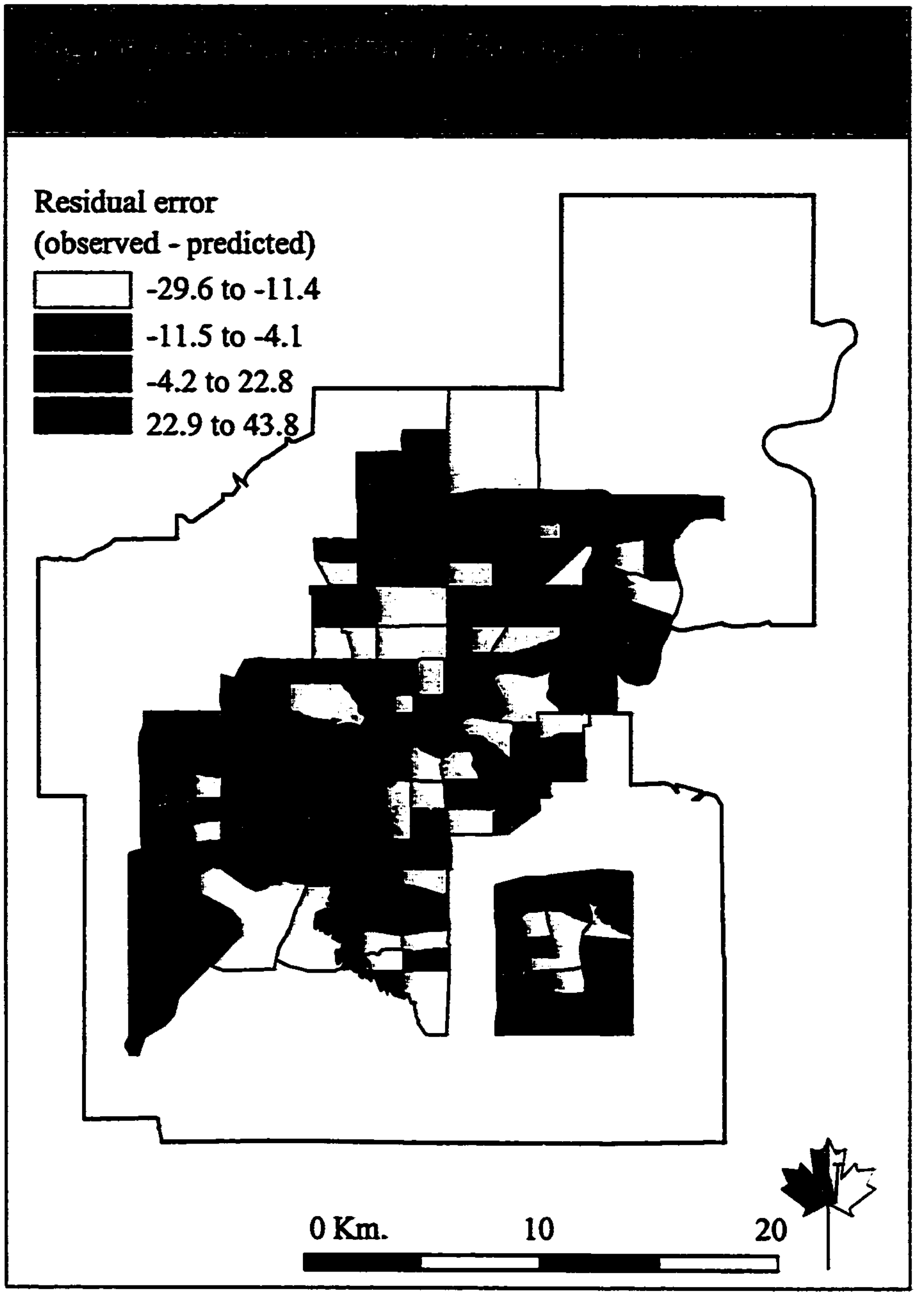
Of more critical concern is the spatial autocorrelation in residual error. According to the first order case, the residual error is not spatially autocorrelated, but in the second order case, it shows spatial autocorrelation. Among all variables tested, the residuals show the weakest spatial pattern. All variables become significantly spatially autocorrelated in the second order neighbouring case. There is no clear answer as to which technique best represents spatial patterning. However, to help make a decision as to which technique is most appropriate, four randomly generated normally distributed numbers were calculated for each tract, and then tested for spatial autocorrelation. One would not expect randomly generated numbers to show spatial autocorrelation, and only one of the four random variables showed spatial autocorrelation in the first order neighbouring case. However, all four random variables showed spatial autocorrelation in the second order neighbouring case. This suggests that the second-order neighbouring technique probably gives a poor indication of spatial autocorrelation in this study area.

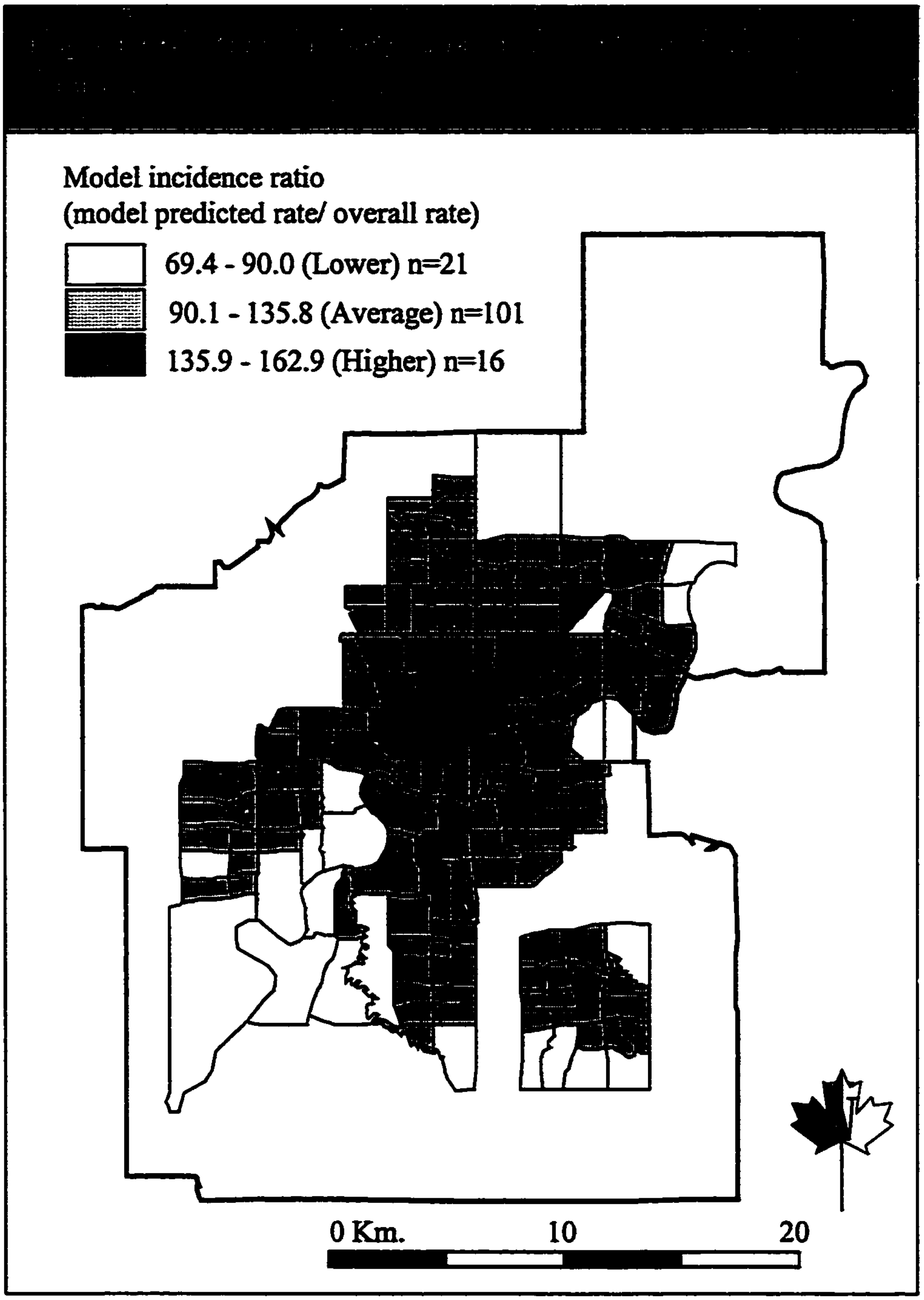
Although *Moran's I* gives no indication of spatial autocorrelation in residual error overall, a map of residual error shows notable cluster of model under-prediction in West Edmonton (Figure 4.20). Rates of injury were higher in these tracts which seem to form a strip running South from the Northwest industrial region of Edmonton.

4.4 Risk Surface

Predicted values from the multiple regression model are used to generate a map of risk (Figure 4.21). The results show predicted rates of child pedestrian injury as described by Equation 4.1.

$$\boxed{Y_i = B_0 + B_1X_{1i} + B_2X_{2i} + B_3X_{3i}} \qquad \textbf{4.1}$$


Residual error
(observed - predicted)
-29.6 to -11.4
-11.5 to -4.1
-4.2 to 22.8
22.9 to 43.8
0 Km.
10
20


Model incidence ratio
(model predicted rate/ overall rate)
69.4 - 90.0 (Lower) n=21
90.1 - 135.8 (Average) n=101
135.9 - 162.9 (Higher) n=16
0 Km.
10
20

Observations of each variable (X_{1i}= low income, X_{2i}= road density and X_{3i}= planning type) were added to the formula to obtain predicted rates for each census tract (Y_i). Compared to the map of residuals, which shows weak spatial patterns, the map of predicted rates shows a strong core-periphery gradient in which the predicted injury rate is generally higher in central areas of the city (Figure 4.21). This is consistent with the pattern observed in maps of the independent variables earlier in this section. When compared to the map of smoothed rates (Figure 4.2 A) the predicted map also shows a greater degree of spatial clustering.

4.5 Alternative Regression Models

Appendix B (pt. 3) shows the results of a Poisson regression model. Although the comparison of coefficients is difficult (as the predictor is a count in the Poisson model, and a smoothed rate in the least squares model) in both cases, the same three variables remain in the model. Appendix B (pt. 3) shows results from least squares regression on subsets of case data. The 66 case subset includes only the cases that were successfully linked to collision data and were injured in their tracts of residence. The 91 case subset includes only the cases that were successfully linked to collision data and were not injured in their tracts of residence. The 109 case subset includes cases that were successfully linked to collision data and were injured within 2km of their residence and the 123 case subset includes cases that were successfully linked to collision data and were injured within 3km. The 157 case subset includes all cases which were successfully linked to collision data.

Of these subsets, the 123 case set generates results most similar to the full data set. In this subset, the same variables are statistically significant predictors of risk. Low income is not a statistically significant predictor in the 157 case subset. The 66 case subset shows a statistically significant road density parameter estimate, while low income and planning type are non-significant. The 109 case subset shows a statistically significant planning type parameter estimate. Finally, the 91 case subset shows significant parameter estimates for low income and planning type, but not for road density. Poisson regression was performed on these subsets using count data. For each subset, the same parameters attained statistical significance as in the least squares regression of smoothed rates.

The main purpose of these analytical procedures was to separate the spatial variability of injury rates from the noise produced by non-spatial factors of risk. The chapter also discussed the characteristics of the injury cases. The final map of predicted rates shows a strong spatial pattern not immediately visible from the descriptive maps of rates or point data. The next chapter will discuss the validity of the model as well as the interpretation of the results as a whole.

Notes

[1] The linking procedure involved clerical (non-computerised) comparisons of the three match criteria in the two databases. Records which matched on gender, matched within one year of age and matched on date of injury were assumed linked. This resulted in 157 successful links. This procedure results in a 0.03% probability of false links: (p(false link) = p(false link on gender (1/2)) * p(false link on date(1/365)) *p(false link on age +/- one year (1/5)) = 2.8×10^{-4}.

5. Discussion

Discussion of this study shall focus both on the ability of the results to fulfil the objectives set out in Chapter 1, and on interpreting the results in light of the research questions themselves. Step one is a detailed discussion of the theoretical issues involving the regression model. Step two is a discussion of the predictors and their relationship with injury. In step three I discuss the epidemiology and geography of risk to children followed by a discussion of limitations of this project with final comments on preventative strategies and opportunities for future research.

5.1 Model Evaluation

The interpretive value of the regression model is particularly important in this study. Of primary concern are 1) the spatial variation in injury rates that the model explains and 2) the degree to which the parameter estimates in the model are reliable and meaningful ecological-level predictors of injury rate. The coefficient of determination, R^2, is the measure most frequently used to assess explained variance. In the final three-variable model, R^2=0.255. Although this indicates that a significant proportion of the variation in injury rates over space remains unexplained by the model, this was not unexpected. An ecological-level model describes variation among aggregate geographic units, not among people. Many individual-level factors influence the probability of an injury event but cannot be detected by models based on aggregate data. In a properly specified ecological-level model, unexplained variation appears as *random noise*, the existence of which should not have a systematic effect on the model parameters. Throughout this work, this noise has been referred to as residual error.

Examining the properties of this residual error is one of the easiest methods of deciding whether or not the parameter estimates obtained in the regression model are accurate and meaningful representations of independent associations with the dependent variable. In an idealised regression model, the residual error is un-patterned. Whatever the causes of the error, if the variation in the error is un-patterned there is good reason to believe that the model (assuming it has non-zero parameter estimates) describes a meaningful pattern in the dependent variable.

By calculating robust standard errors for the regression parameters, it would appear that heteroscedasticity does not seriously affect the parameter estimates in the model. Robust standard error calculations were similar to the standard error calculations from the ordinary least squares model.

Examination of the leverage that certain observations have on the model also suggests that residual error does not have a patterned effect on the reliability of the parameter estimates. However, some of the observations with greatest leverage do have common characteristics—two of the lowest income tracts were among the observations with the greatest leverage on the model. Proportion of low income families was obtained by Statistics Canada based on a 20% sample of families in each tract. It is possible that the leverage these tracts represent in the model is influenced by sampling error. The proportion of low income families in these tracts may be higher or lower than what would be expected had a larger sample been collected. Little work has been done to examine the effects of sampling error in the Canadian Census and its effects on regression models; therefore, it is difficult to quantify the effect in this particular case. Observations with higher than average leverage were temporarily removed from the model to judge the magnitude of their effect. The result was an increase in R^2 and a moderate decrease in the low income parameter estimate. The overall effect of these observations is small however, and therefore their presence was not deemed to invalidate the model.

Tests of spatial autocorrelation were performed to identify patterns of residual error in space. Ecological-level regression models that show considerable spatial autocorrelation of residual error may require a revisiting of theory, as such patterns may indicate missing variables. In this study, *Moran's I*, a measure of study-wide spatial autocorrelation, does not indicate a strong pattern of spatial autocorrelation. The map of residual error does show a group of communities in West-Central Edmonton that possess slightly to moderately higher injury rates than the model predicts (4.20). This region of Edmonton is given special attention in a later section of this chapter.

The distribution of residual error was examined for normality by visually inspecting the histogram as well as a statistical test for normality. Although there is evidence of a non-normal distribution of the residual error, it is unlikely to greatly affect the significance tests of the parameter estimates. Such tests of normality in residual error are more important in cases where there are extreme departures from the normal distribution, or in small sample sets. In most other cases where residual error is non-normal, tests of significance (like F tests for significant partialled effects and t statistics for non-zero regression parameters) are robust (Berry and Feldman, 1985).

The sum of these observations suggests that the parameter estimates are the best linear unbiased estimates, and that the model represents a meaningful and repeatable prediction of injury rates over space. Although the R^2 is not large, the model does explain a notable proportion of variation in injury rates between census tracts in

Edmonton. As this model is simple to interpret, and provides useful information about risk over space, it is reasonable to accept it as a suitable analytical tool for describing variability in injury rates between communities in Edmonton.

5.2 Predictors of Child Pedestrian Injury Rates

The degree to which a regression model can provide a researcher with useful information is dependent on the theory underlying it, much of which determines what variables should be included in the modelling procedure. The variables in ecological-level regression models are often complicated by the fact that they may represent compositional and/or contextual relationships. Five variables were chosen for study. Of these five, three were shown to have independent effects in the model: planning type, low income and traffic density. The two variables which did not remain important in the model are discussed first.

Proportion of park area in census tracts shows no statistically significant association with injury rate. The simple interpretation is that the availability of parks is not associated with variability in injury rates over space. This may be because parks are not important destinations for children. Parental decision-making may also affect the role parks play in influencing risk; children who do not have easy access to parks may be restricted from travelling to them altogether, thereby reducing their exposure to the road environment, and reducing risk of pedestrian injury. Unfortunately, a general lack of quantitative knowledge about child mobility—particularly in residential environments—makes the interpretation of such a variable very difficult.

Proportion of park space does correlate with planning type; more sophisticated planning structures possess more park space, and this park space is usually separated from high volume roads. However, this study suggests that park space does not itself have an effect on injury rates. Individual-level studies have shown significant relationships between play/park space and risk of injury (e.g. Mueller et al., 1990; von Kries et al., 1998). However, none of these studies investigated the role of neighbourhood planning on risk. Therefore, it is possible that the associations these studies observed were relationships between the neighbourhood plan and not the quantity of play area available to children. At the very least, these results suggest that the relationship between play space and planning structure deserves more direct consideration in case-control studies which explore environmental risk factors.

Although proportion of apartment dwellings shows a correlation with injury rate, it fails to show an independent effect in the regression model. All three independent variables with non-zero parameter estimates in the multiple regression model do show moderate correlation with proportion of apartment dwellings. After controlling for the effects of these variables, proportion of apartment dwellings does not explain significant variance in the dependent variable. Since this statistical control is based on aggregated data, it does not exclude the possibility of significant individual-level effects. However, as was the case for proportion of park space, previous studies which considered apartment dwellings as factors of risk failed to consider all features of the neighbourhood environment related to this variable. Results indicate that proportion of apartment dwellings is redundant after including any of the other three predictors. This may also suggest that proportion of apartment dwellings does not have a contextual effect on injury risk. If a significant contextual effect were present (i.e., if living in a neighbourhood with many apartments was a factor of risk, regardless of whether or not a given child lives in an apartment), then one would have expected to see an ecological-level correlation. However as previously stated, it is difficult to speak conclusively on this matter since data were not available to conduct robust contextual analysis.

The model indicates that the spatial variability in risk of injury is not related to the presence or absence of public or private play-areas. This suggests that the relationship between injury risk and availability of outdoor play areas in a neighbourhood—private or public—may need to be re-evaluated in the case-control setting. This comment is made cautiously, however, as individual-level relationships are not directly addressed by this model. These observations may also be study-site specific. Edmonton has long winters which probably discourage prolonged outdoor play by children. In cities with milder winters, outdoor play-space may have a more protective effect since these resources would be more frequently used.

Wang's (1994) findings about planning type and injury frequency are supported by this research despite the considerable differences in study design. After controlling for low income and traffic density, tract injury rate is inversely proportional to the sophistication of neighbourhood planning structure. The exact design features responsible for the association are unclear, but since planning type has an effect after controlling for traffic density, features other than the quantity of traffic on arterials must be important. This could relate to traffic speed or fundamental properties of the plan itself—like the degree to which pedestrian and automobile traffic are separated, and the number and type of intersections.

Proportion of low income families shows an independent role in explaining variability in rates between communities, however as a compositional effect, the exact properties of this relationship is unclear. The strength of the individual-level research done in New Zealand and Australia suggests that the relationship may be related to proportion of car ownership in these communities. As my model was incapable of accurately isolating individual-level relationships, it is difficult to compare findings in these studies. Family car ownership is probably lower in regions of the city with more low-income families. All else being equal, children in these areas are probably making more pedestrian trips. The relationship between car ownership and pedestrian trips taken is supported both in the pedestrian injury literature (Roberts et al. 1996) and in the Edmonton Household Travel Survey (City of Edmonton transportation and Planning 1995). Increased exposure to the traffic environment results in increased risk of injury, which may account for the ecological-level association between low income and injury rate.

The traffic density index is a measure of traffic volume in Edmonton census tracts. The model indicates that traffic density remains a significant predictor of injury rates even after controlling for planning type. Since the indicator is a measure of cars per kilometre, it is unclear whether the association detected is the result of the quantity of arterials in tracts, the volume of cars, or both. Although the traffic density index gives no indication of traffic density on residential streets, it does give a general indication of the quantity of road danger to children. Of the cases for which collision location was obtained, the majority (~60%) were injured on arterial road segments. The presence of these roads and their volume are important predictors of injury rates after controlling for income and planning structure.

Standardised coefficients indicate that the three variables "explain" similar amounts of variability in the dependent variable. In Chapter 4, Table 4.9 shows that the addition or removal of any of the three variables results in a similar change in R^2. Despite the similarities between the three variables, the low income parameter estimate shows less stability. When the highest leverage tracts are removed from the model, the low income parameter experiences the greatest change. Although traffic density has the largest range of observations (0-112) low income possesses a notable cluster of extreme values, particularly in the high end. Together this evidence suggests that among the significant model variables, low income has the most tenuous association with injury rate. This may be related to the fact that proportion of low income families is compositional while the other variables are contextual. It also may reflect the complex role that income has on risk. Income is likely a surrogate for many variables—compositional and contextual—some of which may have non-linear or otherwise complicated relationships with pedestrian injury risk.

The regression model provides information consistent with most research on pedestrian injury. Elements of both the physical and social environment show a relationship with injury risk. Although this model does not contribute any new information to the understanding of individual-level factors of risk, this was never the intent of this study. However, it does provide an adequate framework for understanding the spatial variability of child pedestrian injury rates.

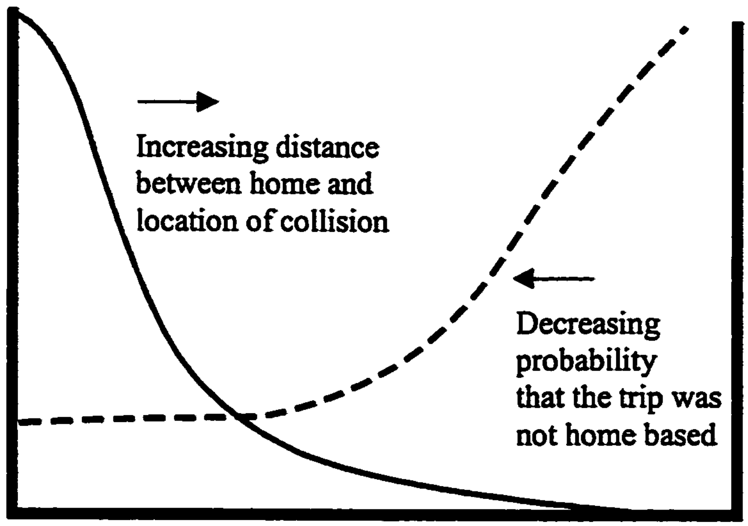
5.3 Features of Child Pedestrian Injury in Edmonton

5.3.1 Population Risk

Many of these results support findings in previous research on child pedestrian injury. In the study period, more boys were injured than girls. Other studies show similar findings (Rivara, 1990; Scheidt et al., 1995). However, the causes of this difference are unclear. Previous research failed to show a relationship between gender and the number of roads crossed (Roberts et al., 1996). The difference may be related to the ways in which boys and girls use the road environments. These data reveal a trend indicating that the gender difference in injury frequency varies less in older age groups (Table 4.1). Although it is difficult to identify what factor is responsible for the variability in risk between boys and girls, it would appear from evidence in this study that this factor is less important as a child grows older. There is also a pattern in the time of day when children are injured that is consistent with the literature. Children are most often injured during two time periods: in the 8-9 am range and in a two to three hour period in the mid to late afternoon. This pattern has appeared consistently over the last twenty years (Wade et al., 1982; Calhoun et al., 1998) although the magnitude of the 9 am spike in these data is uncharacteristically large. Injuries occurring in the latter period of the day represent a broader period of risk which also appears in many studies of pedestrian injury in children. Lapses in driver and pedestrian attention at this time of day may partially account for the pattern.

Despite the fact that only 157 of the 258 cases could be linked to collision location data, the results reveal a pattern consistent in child pedestrian research—that most children are injured near home. The exponential drop in distance between residence and collision location confirms the importance of neighbourhood environments as places of pedestrian activity (Table 4.11). The shorter distances probably represent collisions that occurred on home-based trips. The longer distances likely represent collisions that occurred on pedestrian trips where the home was not a point of origin. There were no data available indicating whether injuries occurred on home-based or non home-based trips. However, it seems reasonable to suggest that in cases where the distance between residence and collision location is small, trips are more likely to be home based (Figure 5.1). As the distance increases, the probability that a given trip is not home-based increases.

The findings of this study do vary considerably with those of Joly et al. (1991) which show that 96% of injury cases were injured in the census tract of residence. Among cases linked to collision data, this study indicates that less than 50% of the child pedestrians were injured in the tract of residence. Their study included more cases (n=1233), involved a slightly different age group (0-14), and involved both bicyclist and pedestrian injuries. However, it is unlikely that these differences alone can account for the drastic contrast in findings. Although Montreal and Edmonton are considerably different in terms of their urban design and history, since Statistics Canada uses similar criteria for constructing tracts in Edmonton and Montreal, it is unlikely that tract size accounts for the differences between the studies. Differences in pedestrian behaviour between the cities may offer some explanation; however, little research has made inter-urban comparisons of pedestrian travel habits. The most substantive difference between study approaches is in the data collected—their study included both emergency-room reported and police-reported injuries whereas this study is restricted to emergency-room



Increasing distance
between home and
location of collision
Decreasing
probability
that the trip was
not home based

reported injuries. How this may account for the difference in findings is unclear. It is possible that police reported injuries tend to occur closer to home than emergency-room reported injuries. A direct comparison of the databases could explain this considerable discrepancy. At the very least, this evidence emphasises 1) the importance of regionally-based research in pedestrian injury and 2) that inter-regional comparisons are needed that standardise data and study design.

Other results are also inconsistent with literature on pedestrian injury. These results show that the frequency of injury increases with age. Several studies have found that the highest frequency of injury occurs in five to nine year olds (Rivara and Barber, 1985; Pless et al., 1987; Mueller et al., 1987; Rivara 1990). Yet other research suggests that younger children (under 6 years) are at highest risk (Read et al., 1963). As this research examines more recent data than these older studies, results may represent a late stage in a long-term trend of changing risk among children. It is possible that past and current efforts to reduce risk—focussed most closely on pedestrian education—have done much to improve child safety at younger ages. These data may indicate the importance of developing alternative methods of injury prevention in the 10 to 15 year old age group. This may be a function inter-regional differences. The social and/or environmental context of Edmonton may put children between 10 and 15 at greater risk than children of this age group in other study areas. These contrary findings stress the importance of studying outcomes in their spatial and temporal context.

5.3.2 Geography of Risk

Together, the smoothed rate map, the map of residual error and the map of predicted rates derived from the regression model are the basis from which one can understand the geography of child pedestrian injury. The map of smoothed rates shows a variable pattern of injury between Edmonton communities, though not nearly as extreme as the variability represented in the map of crude rates (Figure 4.2b). Communities in the West-Central areas of the city and near the downtown show the highest rates. Injury rates are lowest in the Southwest tracts of the city, which are made up of mostly higher income, single dwelling households. Although a weak spatial pattern of rates was detected using *Moran's I*, it is not obvious from the map where this clustering in rates is concentrated. The map of injury rates alone provides little information about risk or spatial processes that influence risk of injury. An attempt to base projections of future injury rates using this information alone would prove difficult.

The mâp of residual error, although not spatially autocorrelated, does show a cluster of slight under-prediction in West-Central areas of the city (Figure 4.19). This under-prediction may be due to some factor of risk common to these areas. The linear pattern of the under-prediction error suggests a cause related to traffic that is otherwise unaccounted for by traffic density of major arterials. These tracts are located between three regions of the city which attract significant traffic volume: the Industrial Northwest, the City Centre and the West Edmonton Mall. It is possible that residential streets in these areas experience high traffic volume. When considering cases for which collision and residential location are known, this hypothesis seems unlikely as most cases were injured on a major arterial, and not on a residential roadway (Table 4.4). Patterns of driver behaviour or regional-specific environmental factors unrelated to neighbourhood planning or traffic density may also explain this pattern. A detailed study of these neighbourhoods may be required in order to reveal the cause of this pattern and provide explanations for this anomaly.

The map of predicted rates generated from the regression model shows a strong spatial pattern and is probably the most informative map generated in this study (Figure 4.21). This map shows pedestrian injury risk after removing the variability in rates unaccounted for in the ecological-level model. Although there are communities with high predicted rates scattered in various regions of the city, the map of risk factors clearly predicts higher rates in Central areas of the city and lower rates in outlying regions. Communities in the Central areas of the city have high proportions of low income families, unplanned neighbourhood structure and higher traffic density. Other research in medical geography has suggested similar patterns. For example, the core-periphery gradient has been shown to be relevant in explaining distribution of infant mortality in Finland (Vuorinen, 1987).

Although the range of rates is not large, the risk surface offers unique, although not entirely unexpected, information about the spatial patterns of injury risk. The core-periphery risk gradient is not visible in a map of rates alone. On this map, spatial patterns are obscured by non-spatial individual-level processes that otherwise dominate risk of injury. Unfortunately, the risk surface is also of limited value from a descriptive standpoint. It would be inappropriate, for example, to view the risk surface as a predictor of future trends child pedestrian injury. The model explains a relatively small proportion of the variability in rates, and therefore would provide a poor estimation of future patterns. However, after developing an ecological-level model which effectively predicts some of the spatial variability in rates, the risk surface indicates that areas in the city centre have a greater combination of risk factors than outlying areas. Although this generalisation is not perfect (as can be seen by the map of residuals), it does indicate a spatial process at work that has not been widely discussed in the child pedestrian injury literature.

5.4 Limitations of Research Approach

5.4.1 Residential Census Tract as the Study Unit

One of the key assumptions discussed in Chapter 3 is that children involved in collisions are most frequently injured near home. This assumption is well supported by the child pedestrian injury literature. This assumption is a very important element of this study's design and the main theoretical argument for choosing the census tract as the unit of analysis. Evidence in Chapter 4 suggests that children are most frequently injured near the home (~70% were injured within 2 Km. of their home). However, of the 157 cases for which collision location data were obtained, less than 50% were actually injured in their tract of residence.

It is possible to estimate the effects of this failed assumption on the model. As a compositional effect, the role of low income is unlikely to change. This variable is a surrogate for the relationship between family income and injury risk at the individual-level. As such, the location of the injury in relation to the home is not immediately relevant. However, the other variables—planning type and road density—are contextual effects. They describe a property of the environment of risk—in this case, the residential census tract. For cases that were injured in an area outside the tract of residence, the characteristics of the home environment may have no relevance on risk. Thus the resultant model may under-represent the association between the physical environment and injury.

The degree to which this problem affects the model may be small, however. All independent variables are highly spatially autocorrelated. Since neighbouring regions tend to share similar characteristics, it is likely that neighbouring residential areas possess similar environmental characteristics. Thus these particular variables probably still describe—for many cases, by proxy—the environment in which the collision took place.

The choice to use residential areas allows me to generate meaningful rates such that the numerator (number of injuries) and denominator (population of children at risk) are consistent and intuitive. Researchers studying the region of collision use the same variables to calculate rates, however, such calculations result in considerably less intuitive results. In these cases, the denominator represents the number of children living in the area of collision. This is meaningfully only if the assumption that children live in the same regions in which they are injured is true. Otherwise this denominator does not relate directly to the population at risk of injury. Until reliable links can be made between data on collisions sites (usually collected by the police) and data on residential locations (usually collected through health administrators) most ecological-level research on child pedestrian injury will be forced to make the assumption that children live in the same tracts that they are injured.

5.4.2 Roadway versus Non-Roadway Injury

Previous studies using data obtained through similar injury surveillance systems have indicated that most (>80%) child pedestrian injuries in the 0-15 year old stratum occur on roadways (Calhoun et al., 1998). Nevertheless, no reliable data could be obtained which identified whether or not a particular case was injured while crossing a road, or walking in a driveway. The data set does include an ICD-9 Ecode that classifies the location of injury (to distinguish road from driveway injuries) but this is very poorly coded in the emergency room database.[1] Research on pedestrian injury in children indicates that the aetiology of driveway injuries differs from those involving on-road collisions. For one, it is unlikely that factors like traffic density or neighbourhood design would have an equivalent relationship with injury risk as on-street collisions. This difference is unaccounted for in this model, and may affect findings.

5.4.3 Smoothed Rates

The smoothed injury rates were used in order to generate more reasonable estimates of injury risk. Typically, regression models of rates based on rare data are weighted by the inverse variance of the dependent variable (in this case child population) or rates are rejected altogether in favour of Poisson regression modelling. The choice to model empirical Bayes estimates has not been examined in detail, however, and problems associated with this technique may have affected the model. I addressed this somewhat by comparing the least squares model of smoothed rates to a Poisson regression model of injury counts. The results show a strong similarity between the models. However, until more research is done to explore the modelling of these rates, the degree to which this technique can be applied is unclear and should probably be accompanied by more widely applied methods.

5.5 Child Pedestrian Injury –Prevention in Time and Space

5.5.1 Regionally Specific Prevention

The findings indicate that risk of child pedestrian injury probably varies between neighbourhoods. In fact, evidence suggests a general city-wide pattern in risk. Prevention programs—educational or environmental—should consider this pattern to improve the efficiency of resource allocation. Traffic calming measures are probably unnecessary in neighbourhoods of Southwest Edmonton where residential traffic volume is low and where most trips by children involve automotive transport. On the other hand, environmental modification may be more important in central areas of the city where through traffic on residential streets is high and where more children rely on pedestrian transportation.

5.5.2 Space-time Interaction

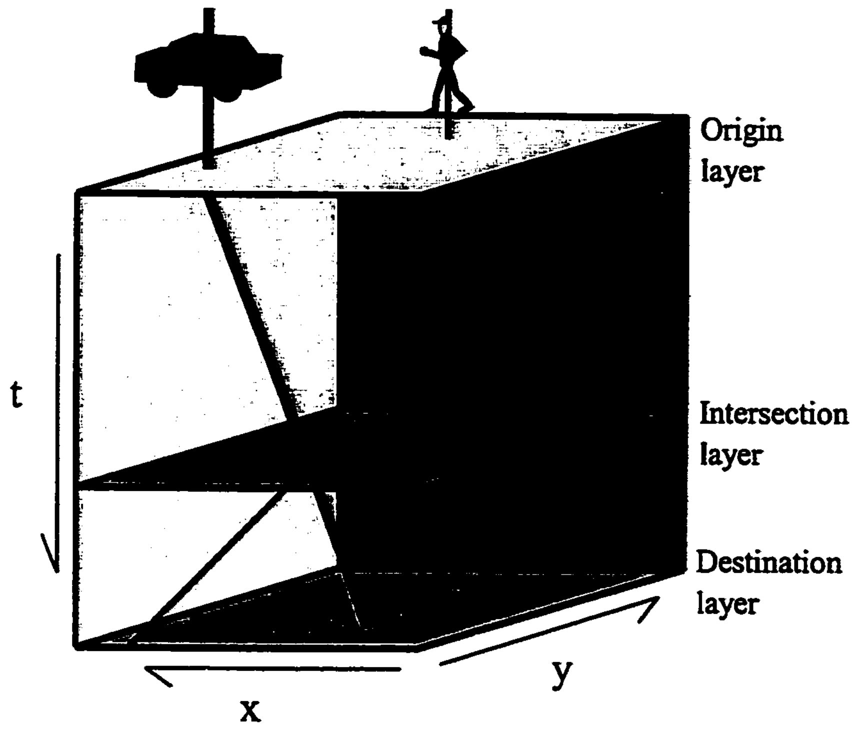
The convergence of pedestrians and automobiles is the ultimate first-cause of collision events. When an automobile and a pedestrian converge in time and space and the sum of the converging forces exceeds the ability of the pedestrian's body to avoid damage, an injury occurs. Therefore it is of little surprise that most prevention strategies aim at reducing this convergence.

Behaviourally focused preventative strategies direct their attention to children; children are taught to behave in a manner that reduces this dangerous interaction with automobiles. This involves teaching children to cross roads more carefully and to avoid roads with high traffic volume. Driver behaviour can also be modified. Changes in driver behaviour—like driving slowly in areas where children are active—increases the driver's ability to react to a situation in which a child is crossing the road unsafely. Both of these approaches represent active prevention approaches that require individuals to make appropriate decisions in order to reduce the convergence of children and automobiles in space and time. A mistake on the part of a driver or child in a given situation can result in a collision. Although such education programs which aim to improve the habits of drivers and pedestrians are critical components of risk reduction, they are limited by the fact that all humans, regardless of education and age, are bound to make mistakes.

Other methods of prevention attempt to reduce the convergence by changing the environment. This can be large scale—as is seen in super-block neighbourhood planning that completely separates pedestrian and automobile traffic— or small scale—like speed bumps and one way streets. These modifications are designed to reduce automobile and pedestrian convergence independent of individual habits. Physical barriers that entirely restrict interaction between pedestrians and automobiles necessarily reduce the frequency of collisions to zero. However, the lack of feasible methods of environmental modification without affecting the efficient traffic flow of automobiles has hampered the widespread introduction of these methods of prevention.

The most desirable strategy for reducing the frequency of child pedestrian injury would be to reduce risk without reducing the opportunities for children to travel independently in their neighbourhoods. Roberts (1993) claims that greater financial equality is an important component for reducing risk of injury and may explain the strides that Sweden has made in reducing child pedestrian injury rates. This argument rests on the assumption that increased wealth among the poor will result in an increase in the number of automobile trips, a decrease in the number of pedestrian trips, and therefore a reduction in pedestrian injury among poor children—a high risk group. Although there is evidence to support such arguments, this approach would further restrict the independent travel of children, an important component of their mental and physical development and well-being. Furthermore, increased automobile use would likely result in increased automobile to automobile collisions. Although this approach would probably reduce the frequency of pedestrian injury among children, it may have consequences that make the treatment worse than the disease.

A rarely discussed method for reducing the space-time convergence of pedestrians and automobiles would be to change the frequency with which the paths of children and automobiles cross in time. Torsten Haagerstrand, an important figure in space-time geographical research, developed a visual model useful for understanding the manner in which individuals interact in space and time (Figure 5.2). These results suggest that times of peak child pedestrian activity coincide with periods of the day when automobile traffic density is high. As long as these times of day are coincident, and frequencies of pedestrian and automobile traffic do not change, collision events will continue to occur. If these peak periods of traffic flow were staggered, the frequency of spatial convergence between pedestrian and automobiles would probably drop. This time change may offer a conceptually and financially feasible method for reducing rates of injury. One option is to consider changing the times in which children go to school. At present, children travel to and from school at times of the day that coincide with peak periods of car travel. Staggering school times so that classes start and end at times less coincident with peak traffic flow could reduce the frequency of collisions between automobiles and child pedestrians. The time staggering need not be considerable if done while incorporating



Two movement paths collide only when there is spatial (x,y) and temporal (t) intersection. Most passive approaches to pedestrian injury prevention focus on reducing spatial interaction.

knowledge about local traffic flow times. The map of traffic flow suggests areas of the city in which changing school hours might be a particularly useful strategy (Figure 3.3). There is a strong core-periphery gradient in which traffic flows from the outer regions of the city to the downtown. School hours could be modified while considering distance from the city core.

Admittedly, the times that schools operate are probably at least in part related to the times of peak traffic flow. Parents who drive their children to and from school will continue to generate considerable traffic during the school commuting hours, whatever the school hours of operation. Furthermore, the inconvenience associated with staggering these times may introduce new problems. Obviously, such changes would require community-level consultation and the consideration of particular neighbourhood concerns. However, this idea offers a new and feasible approach for reducing risk of injury to children.

The core-periphery gradient in risk over space combined with knowledge about peak periods of traffic flow may provide a simple starting point for intervention by altering school times. This may be a particularly important step for those who are interested in reducing injury risk among children in the older age group. Surveys of peak traffic times in each of these areas would offer information about the degree of temporal staggering necessary to effectively reduce pedestrian and automobile interaction. At the very least, future research should consider studying relationships between the times of school attendance and the risk of injury.

Notes

[1] In our data set the vast majority of cases were assigned an "undetermined" setting of injury location.

6. Conclusions

From the outset I have emphasised that this research is limited to describing risk in terms of neighbourhoods rather than individuals. Such ecological-level research is important to understand the geographic properties of health outcomes as well as offering information useful for prevention and resource allocation. These methods are especially important at the scale of this study. Most people in the world live in urban environments, and each city offers a unique context of observation. Methods that explore differences between neighbourhoods can be useful for understanding how to improve these environments and as a result, the health, safety and overall quality of life.

The results of the regression model, although complicated by several methodological issues, do not depart considerably from most research in pedestrian injury. The description of the demographic spatial and temporal trends does add to the understanding of child pedestrian injury for both epidemiologists and geographers. By isolating the spatial variability in injury rates, it is now possible to identify a notable pattern in the spatial variability of risk. This emphasises the importance of considering geographic location in the design of case-control studies. First and foremost, research that does not focus on environmental risk factors should still incorporate the geographical location of cases and controls into the study design—either by matching on location itself or by including location as a variable of study.

Two important steps would do much to further the understanding of pedestrian injury. First, there must be more work on the travel habits of child pedestrians in the urban environment. Some research has studied the frequency of roads crossed and the total number of trips taken; however, these studies have offered little to develop a good generalised idea of pedestrian travel in children. What are the destinations of choice among child pedestrians during after-school hours and the weekends? What distances are children likely to walk to school and other destinations at various ages? Do children from poorer families have different travel habits than children from wealthier families? What role does parental concern play in affecting the pedestrian activities of children? In addition to expanding our knowledge about injury risk, such research may help us understand the relationship between physical and mental health and independent child activity.

Second, researchers should consider investigating the links between the two main sources of information on pedestrian injury—hospital records (which typically contain information about residence) and police records (which typically include information about collision location). A comparison of the cases reported in the study period, and the identification of systematic differences between these sources of information should be investigated. The collection of residential data—like postal codes—on police records could improve the linkage process as well. Although ambulatory records often contain information on injury cases that can be linked to hospital records (in the Capital Health Region an ambulatory record is attached to the patient chart) the records are frequently incomplete and are not always specific. The formal linkage of these three data sources could prove effective in identifying collision locations while avoiding complications associated with database merging.

At the heart of this research is an emphasis on bringing together methods in geography and epidemiology. In this study and many other applications in health, researchers in these fields must consider methods and theory derived from both disciplines. The clinical and population based epidemiological studies tell us about the factors of immediate risk. The ecological-level approach emphasises the role of the society and the environment in determining risk. Both research areas are important for understanding the complex issues underlying human well being.

References

Achen CH (1982). Interpreting and Using Regression. Series No. 29. Beverly Hills: Sage.

Andrews HF (1985). The ecology of risk and the geography of intervention: from research to practice for the health and well-being of urban children. Annals of the Association of American Geographers 75(3):370-382.

Agran PF, Winn DG, Anderson CL, Tran C, Del Valle CP (1996). The role of the physical and traffic environment in child pedestrian injuries. Pediatrics 98(6):1096-1103.

Backett EM, Johnston AM (1964). Social patterns of road accidents to children. *In* Accident Research: Methods and Approaches (ed. Haddon, Suchman, Klein). New York: Harper and Row. 483-491.

Bailey TC, Gatrell AC (1995). Interactive Spatial Data Analysis. New York: J. Wiley.

Berry WD, Feldman S (1985). Multiple Regression in Practice. Series No.50. Beverly Hills: Sage Publications.

Boyle MH, Willms JD (1999). Place effects for areas defined by administrative boundaries. American Journal of Epidemiology 149:577-585.

Braddock M, Lapidus G, Cromley E, Cromley R, Burke G, Banco L (1994). Using a Geographic Information System to Understand Child Pedestrian Injury. American Journal of Public Health. 84(7):1158-1161.

Calhoun AD, McGwin G, King WD, Rousculp MD (1998). Pediatric pedestrian injuries: a community assessment using a hospital surveillance system. Academic Emergency Medicine. 5(7): 685-690.

Carlin JB, Taylor P, Nolan T (1995). A case-control study of child bicycle injuries: relationship of risk to exposure. Accident Analysis & Prevention. 27(6):839-44.

Chan KM, Pang WS, Ee CH, Ding YY, Choo P (1997). Epidemiology of falls among the elderly community dwellers in Singapore. Singapore Medical Journal. 38(10):427-431.

City of Edmonton Transportation Department (1995). Household Travel Survey Project Report. Edmonton: City of Edmonton.

City of Edmonton Transportation Department (1996). Average Annual Weekday Traffic Flow. Edmonton: City of Edmonton.

City of Edmonton (1996). Edmonton Transit System Review: Ten Year Transit Service and Fare Strategy Plan. Edmonton: City of Edmonton.

City of Edmonton Planning and Development (1997). Wards and Standard Neighbourhoods. Edmonton: City of Edmonton.

Clayton D, Kaldor J (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. Biometrics **43**:671-681.

Cohen J, Cohen, P (1975), Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences, Hillsdale: Lawrence Erlbaum.

Devine OJ, Louis TA, Halloran ME (1996). Identifying areas with elevated disease incidence rates using empirical Bayes estimators. Geographical Analysis **28**(3):187-199.

Diez-Roux AV, Nieto FJ, caulfield L, Tyroler HA, Watson RL, Szklo M (1999). Neighbourhood differences in diet: the Atherosclerosis Risk in Communities (ARIC) Study. Journal of Epidemiology and Community Health. **53**:55-63.

Dougherty G, Pless IB, Wilkins R (1990). Social class and the occurrence of traffic injuries and deaths in urban children. Canadian Journal of Public Health. **81**:204-209.

Durkin MS, Davidson LL, Kuhn L, O'Connor P, Barlow B (1994). Low-Income Neighbourhoods and the Risk of Severe Pediatric Injury: A Small-Area Analysis in Northern Manhattan. American Journal of Public Health. **84**(4): 587-592.

Ebdon D (1985). Statistics in Geography, 2nd Edition. New York: Basil Blackwell Ltd.

Ecob R (1992). Multilevel modelling approach to examining the effects of area of residence on health functioning. Journal of the Royal Statistical Society **159**(1):61-75.

English D (1992). Geographical epidemiology and ecological studies. *In* Geographical and Environmental Epidemiology: Methods for Small-Area Studies (ed. Elliot, Cuzick, English and Stern). New York: Oxford University Press: 3-13.

Gardner MJ (1992). Childhood leukaemia around the Sellafield nuclear plant. *In* Geographical and Environmental Epidemiology: Methods for Small-Area Studies (ed. Elliot, Cuzick, English and Stern). New York: Oxford University Press: 291-309.

Gatrell AC, Dunn CE (1995) Geographical information systems and spatial epidemiology: modelling the possible association between cancer of the larynx and incineration in North-West England. *In* The Added Value of Geographic Information Systems in Public and Environmental Health (ed. De Lepper, Scholten and Stern). Dordrecht: Kluwer Academic Publishers. 215-233.

Glass GE, Schwartz BS, Morgan JM, Johnson DT, Noy PM, Israel E (1995). Environmental Risk Factors for Lyme Disease Identified with Geographic Informations Systems. American Journal of Public Health. **85**:944-948.

Gould P, White R (1986). Mental Maps. Boston: Allen & Unwin Inc.

Guyer B, Gallagher SS (1985). An Approach to the Epidemiology of Childhood Injuries. Pediatric Clinics of North America. 32(1):5-15.

Guyer B, Gallagher SS (1985). An approach to the epidemiology of childhood injuries. Pediatric clinics of North America. 32(1):5-15.

Haddon W, Suchman EA, Klein D (1964). Accident research methods and approaches. New York: Harper & Row.

Haddon W (1975). Perspective on a current public health controversy. American Journal of Public Health. 65:1342-1344.

Haddon W (1980). Advances in the epidemiology of injuries as basis for public policy. Public Health Reports. 95(5):411-421.

Hannan MT, Burstein L (1974). Estimation from grouped observations. Amercan Sociological Review. 39(6):374-392.

Henneckens CH, Buring JE, Mayrent SL (1987) Epidemiology in Medicine. Boston: Little, Brown.

Hillman M, Adams J, Whitelegg J (1990) One False Move... A study of children's independent mobility. London: Policy Studies Institute.

Hillman M, Whalley A (1979). Walking is Transport. London: Policy Studies Institute.

Hodge G (1991). Planning Canadian Communities: An Introduction to the Principles, Practice and Participants. Scarborough: Nelson Canada.

Hoover R (1979). Environmental Cancer. Annals of the New York Academy of Science. 329:50-60.

Horowitz SM, Morgenstern H, DiPietro L, Morrison CL (1988). Determinants of Pediatric Injuries. American Journal of Disease in Children. 142:605-611.

Husband P, Hinton PE (1972). Families of children with repeated accidents. Archives of Diseases in Childhood. 47:396-401.

Iverson GR (1984). Bayesian Statistical Inference. Series No.43 . Beverly Hills: Sage Publications.

Iverson GR (1991). Contextual Analysis. Series No. 81.Thousand Oaks: Sage Publications.

Jerrett M, Cole D, Reader S (1997). Environmental equity in Canada: an empirical investigation into the income distribution of pollution in Ontario. Environment and Planning A. **29**:1777-1800.

Johnston RJ (1978). Multivariate Statistical Analysis in Geography: a primer on the general linear model. London: Longman

Joly MF, Foggin PM, Pless IB (1991). Geographical and socio-ecological variations of traffic accidents among children. Social Science and Medicine. **33**(7):765-769.

Louis TA (1984). Estimating a population of parameter values using Bayes and empirical Bayes methods. Journal of the American Statistical Association. **79**:393-398.

Kendall MG, Buckland WR (1971). A Dictionary of Statistical Terms. London: Hafner Publishing.

Kohli S, Sahlen K, Lofman O, Sivertun M, Trell E, Wigertz O (1997). Individuals living in areas with high background radon: a GIS method to identify populations at risk. Computer Methods and Programs in Biomedicine. **53**: 105-112.

Langford JH (1994). Using empirical Bayes estimates in the geographical analysis of disease risk. Area **26**(2):142-149.

Larson CP, Pless IB (1988). Risk factors for injury in a 3-year old birth cohort. American Journal of Diseases of Children. **142**: 1052-1057.

Lewis-Beck MS (1980). Applied Regression: an introduction. Series No. 22 . Beverly Hills: Sage Publications.

Little D (1998). Microfoundations, Method, and Causation : on the Philosophy of the Social Sciences. New Brunswick: Transaction Publishers.

Lovett A, Flowerdew R (1989). Analysis of count data using Poisson regression. Professional Geographer **41**(2):190-198.

Macpherson A, Roberts IG, Pless IB (1998). Children's exposure to traffic and pedestrian injuries. American Journal of Public Health. **88**(12):1840-1845.

Matheny AP (1987). Psychological characteristics of childhood accidents. Journal of Social Issues. **43**(2):45-60.

Meade MS (1988). Medical Geography. New York: Guilford Press.

Marmot MG (1998). Improvement of social environment to improve health. The Lancet. **351**:57-60.

Moore DS (1997). Bayes for beginners? Some reason to hesitate. American Statistician. **51**(3):254-262.

Mueller BA, Rivara FP, Lii S, Weiss NS (1990). Environmental factors and the risk for childhood pedestrian-motor vehicle collision occurrence. American Journal of Epidemiology. **132**(2):550-560.

Muller MJ, Koertringer I, Mast M, Languix K, Frunch A (1999). Physical activity and diet in 5 to 7 years old children. Public Health Nutrition. **2**(3A):443-444.

Narayan KM, Ruta D, Beattie T (1997). Seat restraint use, previous drinking history and non-fatal injury: quantifying the risks. Archives of Diseases in Childhood. **77**(4): 335-338.

Norcliffe GB (1982). Inferential Statistics for Geographers: An Introduction. Hutchinson & Co.

Openshaw S (1984). The modifiable areal unit problem. Concepts and Techniques in Modern Geography No.38. Norwich: Geo Books.

Pless IB, Verrault R, Arsenault L, Frappier J, Stulginskas J (1987). The epidemiology of road accidents in childhood. American Journal of Public Health. 77(3):358-360.

Pless IB, Verreault R, Tenina S (1989). A case-control study of pedestrian and bicyclist injuries in childhood. **79**(8):995-998.

Pless IB, Taylor HG, Arsenault L (1995). The relationship between vigilance deficits and traffic injuries. Pediatrics **95**:219-224.

Powell EC, Tanz RR (1999). Child and adolescent injury and death from urban firearm assaults: association with age, race, and poverty. Injury Prevention. **5**(1):41-7.

Preston B (1972). Statistical analysis of child pedestrian accidents in Manchester and Salford. Accident Analysis and Prevention. **4**:323-332.

Pringle DG (1996). Mapping disease risk estimates based on small numbers: an assessment of empirical Bayes techniques. The Economic and Social Review **27**(4):341-363.

Ramsey JB (1969). Tests for specification errors in classical linear least-squares regression analysis. Journal of the Royal Statistical Society. 31(2):350-371.

Raybould S, Walsh S (1995). Road traffic accidents involving children in North-East England. *In* The Added Value of Geographic Information Systems in Public and Environmental Health (ed. De Lepper, Scholten and Stern). Dordrecht: Kluwer Academic Publishers. 181-188.

Read JH, Bradley EJ, Morison JD, Lewall D, Clarke DA (1963). The epidemiology and prevention of traffic accidents involving child pedestrians. Canadian Medical Association Journal. 89(10):687-701.

Rivara FP, Barber M (1985). Demographic Analysis of Childhood Pedestrian Injuries. Pediatrics. **76**(3): 375-381.

Rivara FP (1990). Child Pedestrian Injuries in the United States. American Journal of Diseases of Children. **144**:692-696.

Roberts IG, Marshall R, Norton R, Borman B (1992). An area analysis of child injury morbidity in Auckland. **28**:438-441.

Roberts IG (1993a). International trends in pedestrian injury mortality. Archives of Diseases in Children **68**:1990-1992.

Roberts IG (1993b). Why have child pedestrian death rates fallen? BMJ **306**(26): 1737-1739.

Roberts IG (1994). Sole parenthood and the risk of child pedestrian injury. Journal of Paediatrics and Child Health. **30**:530-532.

Roberts IG, Norton R, Jackson R, Dunn R, Hassall I (1995). Effect of environmental factors on risk of injury of child pedestrians by motor vehicles: a case-control study. BMJ **310**:91-94

Roberts IG, Norton R, Taua B (1996). Child pedestrian injury rates: the importance of "exposure to risk" relating to socioeconomic and ethnic differences, in Auckland, New Zealand. **50**:162-165.

Robinson WS (1950). Ecological correlations and the behaviour of individuals. American Sociological Review. **15**:351-357.

Rose (1992). The Strategy of Preventive Medicine. Oxford: Oxford University Press.

Rozycki GS, Mauli KI (1991). Injuries sustained by falls. Archives of Emergency Medicine. **8**(4): 245-252.

Saunderson TR, Langford IH (1996). A study of the geographical distribution of suicide rates in England and Wales 1989-92 using empirical Bayes estimates. Social Science and Medicine. **43**(4):489-502.

Scheidt PC, Harel Y, Trumble AC, Jones DH, Overpeck MD, Bijur PE (1995). The epidemiology of nonfatal injuries among US children and youth. Americal Journal of Public Health. **85**(7): 932-938.

Selvin HC (1958). Durkheim's suicide and problems of empirical research. American Journal of Sociology. **64**:607-619

Shaw G, Wheeler D (1994). Statistical Techniques in Geographical Analysis. New York: Halsted Press.

Shimai S, Yamada F, Masuda K, Tada M (1993). T.V. game play and obesity in Japanese School Children. Perceptual and Motor Skills **73**(3 pt. 2): 1121-1122.

Shouls S, Congdon P, Curtis S (1996). Modelling inequality in reported long term illness in the UK: combining individual and area characteristics. Journal of Epidemiology and Community Health. **50**:366-376.

Shugerman RP, Rivara FP, Wolf ME (1992). Schneider CJ. Risk factors for childhood sledding injuries: a case-control study. Pediatric Emergency Care. **8**(5):283-6.

Sparks G, Craven MA, Worth C (1994). Understanding differences between high and low childhood accident rate areas: the importance of qualitative data. Journal of Public Health Medicine. **16**(4):439-446.

Splus (1998). Splus for ArcView GIS User's Guide. Cambridge: Mathsoft Inc.

Stata (1997). Stata Reference Manual 5. College Station: Stata Press.

Statistics Canada (1991) Census Dictionary. Statistics Canada No. 92351XPE. Ottawa: Government of Canada

Statistics Canada (1996a) Causes of Deaths, 1996. Statistics Canada No. 84F0208XPB. Ottawa: Government of Canada

Statistics Canada (1996b) Canadian Census, 1996. Ottawa: Government of Canada

Statsoft (2000). Electronic Statistics Textbook. Tulsa: Statsoft.

Stevenson MR, Jamrozik KD, Spittle J (1995). A case-control study of traffic risk factors and child pedestrian injury. International Journal of Epidemiology.**24**(5):957-964.

Stevenson M, Jamrozik K, Burton P (1996). A case-control study of childhood pedestrian injuries in Perth, Western Australia.**50**:280-287.

Stevenson M (1997). Childhood pedestrian injuries: what can changes to the road environment achieve? Australian and New Zealand Journal of Public Health.**21**(1):33-37.

Susser M (1994) The Logic in Ecological: I. The Logic of Analysis. American Journal of Public Health. **84**(5):825-829.

Teso RR, Poe MP, Younglove T, McCool PM (1996). Use of Logistic Regression and GIS Modeling to Predict Groundwater Vulnerability to Pesticides. Journal of Environmental Quality. **25**: 425-432.

Von Eye A, Schuster C (1998). Regression Analysis for Social Sciences. Academic Press.

Von Kries R, Kohne C, Bohm O, Von Voss H (1998). Road injuries in school aged children: relationship to environmental factors amenable to interventions. Injury Prevention. **4**(2):103-105.

Vuorinen HS (1987). Core-periphery differences in infant mortality. Social Science and Medicine. **24**(8): 659-667.

Wang S (1994). Residential Planning for Pedestrian Safety. Ph.D. Dissertation

Warda L, Tenenbein M, Moffatt ME (1999). House fire injury prevention update. Part I. A review of risk factors for fatal and non-fatal house fire injury. Injury Prevention. **5**(2):145-50.

Wartenberg D, Greenberg M, Lathrop R (1993). Identification and Characterization of Populations Living Near High-Voltage Transmission Lines: A Pilot Study. Environmental Health Perspectives. **101**(7): 626-632.

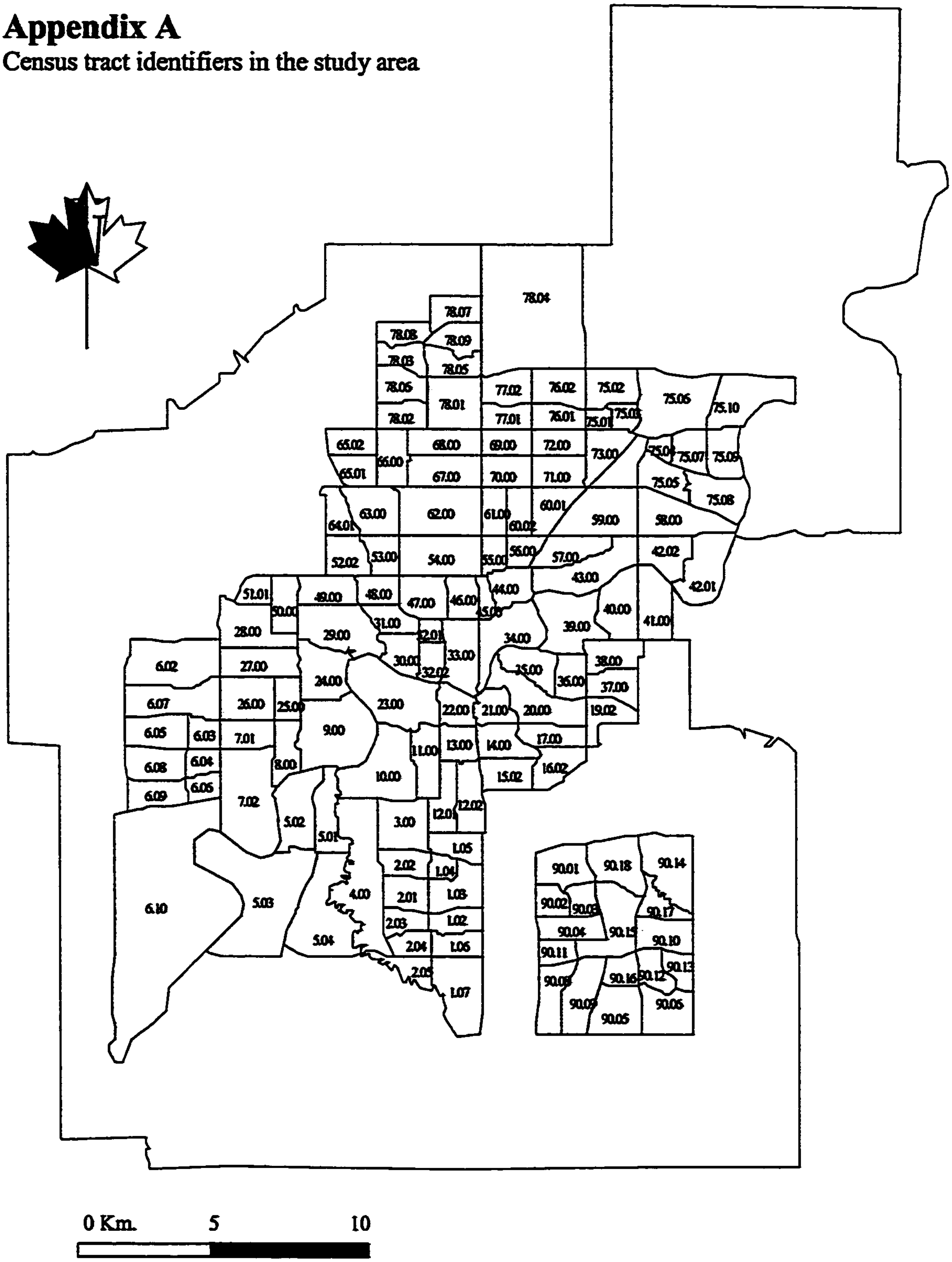
Williams JM, Currie CE, Wright P, Elton RA, Beattie TF (1996). Socioeconomic status and adolescent injuries. Social Science and Medicine. **44**(12):1881-1891.

Wilson M, Baker S (1987). Structural approach to injury control. Journal of Social Issues. **43**(2):73-86.

Winkler RL (1972). Introduction to Bayesian Inference and Decision. New York: Holt, Rinehart and Winston Inc.

Appendix A

Census tract identifiers in the study area



Appendix B

Additional Regression Tables

1) Regression models of squared and cubed terms

Squared Term Low Income				
Obs.	138		F(3,134)	13.260
R^2	0.285		p val.	0.000
Adj. R^2	0.264			
Variable	Coeff.	SE	T	sig.lvl.
Low Income*	1.060	0.451	2.351	0.020
Traffic Density*	0.154	0.058	2.653	0.009
Planning Type*	-3.622	1.200	-3.019	0.003
Low Income ^{2}	-0.012	0.008	-1.579	0.117
Constant*	39.172	7.426	5.275	0.000

Cubed Term Low Income				
Obs.	138		F(3,134)	13.460
R^2	0.288		p val.	0.000
Adj. R^2	0.267			
Variable	Coeff.	SE	T	sig.lvl.
Low Income*	0.778	0.266	2.934	0.004
Traffic Density*	0.157	0.058	2.719	0.007
Planning Type*	-3.583	1.194	-3.000	0.003
Low Income ^{3}	0.000	0.000	-1.756	0.081
Constant*	40.608	6.781	5.989	0.000

Squared Term Traffic Density				
Obs.	138		F(3,134)	12.640
R^2	0.275		p val.	0.000
Adj. R^2	0.254			
Variable	Coeff.	SE	T	sig.lvl.
Low Income*	0.378	0.146	2.585	0.011
Traffic Density*	0.313	0.192	1.632	0.105
Planning Type*	-3.419	1.203	-2.843	0.005
Traffic Density ^{2}	-0.001	0.002	-0.828	0.409
Constant*	43.026	7.031	6.119	0.000

Cubed Traffic Density				
Obs.	138		F(3,134)	12.610
R^2	0.275		p val.	0.000
Adj. R^2	0.253			
Variable	Coeff.	SE	T	sig.lvl.
Low Income*	0.380	0.146	2.601	0.010
Traffic Density*	0.243	0.120	2.025	0.045
Planning Type*	-3.421	1.203	-2.843	0.005
Traffic Density ^{3}	0.000	0.000	-0.774	0.440
Constant*	43.803	6.692	6.545	0.000

* Significant parameter estimate (0.05)

2) Regression models of interaction variables

Interaction 1: Low Income (LI) and Traffic Density (TD)				
Obs.	138		F(3,134)	13.430
R^2	0.288		p val.	0.000
Adj. R^2	0.266			
Variable	Coeff.	SE	T	sig.lvl.
Low Income*	0.773	0.267	2.898	0.004
Traffic Density*	0.331	0.114	2.915	0.004
Planning Type*	-3.475	1.192	-2.915	0.004
LI * TD	-0.008	0.004	-1.728	0.086
Constant*	37.978	7.607	4.993	0.000

Interaction 2: Low Income (LI) and Planning Type (PT)				
Obs.	138		F(3,134)	12.410
R^2	0.272		p val.	0.000
Adj. R^2	0.250			
Variable	Coeff.	SE	T	sig.lvl.
Low Income	0.363	0.285	1.273	0.205
Traffic Density*	0.161	0.059	2.750	0.007
Planning Type	-3.700	3.013	-1.228	0.222
LI * PT	0.011	0.116	0.092	0.927
Constant*	46.597	9.198	5.055	0.000

Interaction 3: Planning Type (PT) and Traffic Density (TD)				
Obs.	138		F(3,134)	12.900
R^2	0.280		sig.lvl.	0.000
Adj. R^2	0.258			
Variable	Coeff.	SE	T	sig.lvl.
Low Income*	0.400	0.146	2.737	0.007
Traffic Density	0.020	0.132	0.149	0.882
Planning Type*	-5.909	2.372	-2.492	0.001
LI * PT	0.054	0.045	1.200	0.231
Constant*	52.590	8.286	6.347	0.000

* Significant parameter estimate (0.05)

3) Poisson regression model

[illegible]				
Obs.	138		X^2,3	62.073
Pseudo R^2	0.123		p val.	0.000
GOF X2,136	169.685			
sig.lvl.	0.020			
Variable	**Coeff.**	**SE**	**Z**	**sig.lvl.**
Low Income*	0.018	0.006	2.876	0.004
Road Density*	-0.181	0.058	-3.146	0.002
Planning Type*	0.008	0.003	2.906	0.004
Constant*	-7.859	0.300	-26.238	0.000

* Significant parameter estimate (0.05)

2) Regression model of subsets based on location of collision and residence

[illegible]				
Obs.	138		F(3,134)	3.984
R^2	0.082		p val.	0.009
Adj. R^2	0.061			
Variable	**Coeff.**	**SE**	**T**	**sig.lvl.**
Low Income	0.030	0.071	0.429	0.669
Road Density*	0.060	0.028	2.133	0.035
Planning Type	-0.798	0.584	-1.368	0.174
Constant*	12.003	2.980	4.029	0.000

Note: When the same subsets were run using a Poisson regression model of injury counts, the same predictors reached statistical significance for each subset.

[illegible]				
Obs.	138		F(3,134)	7.920
R^2	0.151		p val.	0.000
Adj. R^2	0.132			
Variable	**Coeff.**	**SE**	**T**	**sig.lvl.**
Low Income*	0.212	0.103	2.066	0.041
Road Density	0.040	0.041	0.984	0.327
Planning Type*	-2.090	0.846	-2.470	0.015
Constant*	17.975	4.319	4.162	0.000

[illegible]				
Obs.	138		F(3,134)	7.260
R2	0.140		p val.	0.000
Adj. R2	0.121			
Variable	**Coeff.**	**SE**	**T**	**sig.lvl.**
Low Income	0.1011406	0.087	1.167	0.140
Road Density	0.060	0.035	1.723	0.087
Planning Type*	-1.782	0.714	-2.497	0.014
Constant*	21.834	3.644	5.992	0.000

[illegible]				
Obs.	138		F(3,134)	11.690
R^2	0.207		p val.	0.000
Adj. R^2	0.190			
Variable	**Coeff.**	**SE**	**T**	**sig.lvl.**
Low Income*	0.194	0.092	2.117	0.036
Road Density*	0.094	0.037	2.568	0.011
Planning Type*	-1.725	0.756	-2.282	0.024
Constant*	21.086	3.858	5.466	0.000

[illegible]				
Obs.	138		F(3,134)	10.998
R^2	0.198		p val.	0.000
Adj. R^2	0.180			
Variable	**Coeff.**	**SE**	**T**	**sig.lvl.**
Low Income	0.246	0.137	1.797	0.075
Road Density*	0.120	0.055	2.192	0.030
Planning Type*	-3.062	1.127	-2.716	0.007
Constant*	29.534	5.753	5.133	0.000

* Significant parameter estimate (0.05)

Appendix C.

```
/*-----------------Niko Yiannakoulias September 27, 1999----------------*/
/*Code may be used and modified without permission of author. However, the
author assumes no responsibility for problems associated with this code. It
was originally compiled to run on an WinNT platform and should run
on and system running a version of DOS older than 3.3. This program uses standard
functions and therefore is portable to a Unix/Linux environment.
/*---------------------------------------------------------------------------*/

#include <stdio.h>
#include <math.h>
#define TRUE 1
void main()
{
        FILE *input;


FILE *output;

        float O[1000];
        float E[1000];
        float theta[1000];
        float var[1000];
        float ebe[1000];
        float SMR[1000];
        float talpha;
        float tnu;
        char answer[];
        float ctuid[1000];

        float nu,alpha,sumtheta,sumvar,mean,variance,indox,successes;
        int m,i,N,iterations,temp;
        output=fopen("output.txt","w");

        printf("      ---------------------------------\n");
        printf("      Calculates empirical Bayes estimates.\n");
        printf("         Max number of records:1000\n");
        printf("    Required fields: 2 numeric (observed and expected)\n");
        printf("      ---------------------------------\n");

        printf("Enter Filename (include extension):\n");
        scanf("%s",answer);
        input=fopen(answer,"r");
        for(i=1;i<=1002;++i)
        {
                fscanf(input,"%f      %f          %f\n",&ctuid[i],&O[i],&E[i]);
                if(O[i]==0&&E[i]==0)
                {
                        printf("End of File\n");
                        break;
                }
                ++N;
                if(N>=1001)
                {
                        printf("Too many records\n");
                        exit(1);
                }
        }
        printf("%d Records\n",N);
        if(N==0)
        {
                printf("File doesn't exist\n");
                exit(2);
        }
        printf("Enter index value:\n");
        scanf("%f",&indox);
```

```
        alpha=indox;nu=indox;talpha=indox;tnu=indox;

        while(TRUE)
        {
                sumtheta=0;sumvar=0;
                for(i=1;i<=N;++i)
                {
                        theta[i]=((O[i]+nu)/(E[i]+alpha));
                        sumtheta=sumtheta+theta[i];
                        var[i]=(1+alpha/E[i])*((theta[i]-(nu/alpha))*(theta[i]-(nu/alpha)));
                        sumvar=var[i]+sumvar;
                }
                mean=sumtheta/N;
                variance=sumvar/(N-1);
                alpha=mean/variance;
                nu=mean*alpha;
                if(((talpha-alpha)*(talpha-alpha))<0.0000001&&((tnu-nu)*(tnu-nu))<0.0000001)
                {
                        break;
                }
                talpha=alpha;
                tnu=nu;
                ++iterations;
                if (iterations==1000)
                {
                        printf("No convergence, try a new index\n");
                        break;
                }
        }
        fprintf(output,"areaid obs        exp        smr        ebe\n");
        for(i=1;i<=N;++i)
        {
                SMR[i]=100*((O[i])/(E[i]));
                ebe[i]=100*((O[i]+nu)/(E[i]+alpha));
                fprintf(output,"%3.2f,%f,%f,%f,%f\n",ctuid[i],O[i],E[i],SMR[i],ebe[i]);
        }
        printf("%d Iterations\n",iterations);
fclose(input);
fclose(output);
}
```